



Université Paris 1 Panthéon - Sorbonne

Master 2 TIDE

Projet de scoring : prédition du risque de résiliation des contrats auto

Auteurs :

ISSA Karen

ADRIEN Davidson

TAVSIEV Askhab

Sommaire

1. Introduction

2. Chargement et préparation des données

2.1 Focus sur les variables sans définition

3. Analyse exploratoire des données (EDA)

3.1 Analyse descriptive univariée et bivariée

3.1.1 Variables quantitatives - Lien avec la cible (résiliation)

3.1.2 Variables qualitatives - Lien avec la cible (résiliation)

4. Analyse des interactions complexes

4.1 Croisement de variables numériques avec le taux de résiliation

5. Création de nouvelles variables

5.1 Indicateur de responsabilité des sinistres

5.2 Écart entre prime réelle et prime de référence

5.3 Résumés de contrats résiliés

6. Analyse statistique de la dépendance des variables avec la cible

6.1 Lien des variables catégorielles avec la cible via test du chi²

6.2 Lien statistique entre les variables numériques et la variable cible

6.3 Matrices de corrélation des variables

6.3.1 Variables qualitatives (V de Cramér)

6.3.2 Variables quantitatives (corrélation de Pearson)

7. Quels modèles ont du potentiel ?

8. Data Preprocessing

- 8.1 Gestion des Valeurs Aberrantes
- 8.2 Gestion des Valeurs Manquantes

9. Modélisation

- 9.1 Encodage et Normalisation
- 9.2 Modèle 1 : Logistic Regression
- 9.3 Modèle 2 : Random Forest
- 9.4 Modèle 3 : XGBoost

10. Modélisation sans la variable DTEFTMVT

- 10.1 Contexte
- 10.2 Modèle 1 : Logistic Regression
- 10.3 Modèle 2 : Random Forest
- 10.4 Modèle 3 : XGBoost

11. Comparaison des modèles – Performances

12. Conclusion

1. Introduction

Dans le contexte actuel, la résiliation des contrats d'assurance automobile représente un enjeu majeur pour les compagnies d'assurance. Elle peut affecter directement la rentabilité et la stabilité du portefeuille de clients. Mieux comprendre les comportements de résiliation permet aux assureurs d'agir en amont, d'adapter leurs offres et d'améliorer leurs actions de fidélisation en ciblant les profils les plus fragiles.

Ce rapport a pour objectif de construire un modèle de scoring capable de prédire la probabilité qu'un contrat auto soit résilié. Pour ce faire, nous utiliserons un ensemble de données contenant des informations sur les assurés, les contrats, les sinistres, les primes ainsi que l'historique des résiliations. L'analyse visera à identifier les caractéristiques les plus fortement associées à la résiliation d'un contrat.

Nous mobiliserons différentes méthodes statistiques et de machine learning, telles que la régression logistique, les algorithmes d'arbres de décision et les techniques de boosting, afin d'évaluer le pouvoir prédictif des variables. Le travail se déroulera en plusieurs étapes : une description détaillée des données, une analyse exploratoire, le traitement des données, puis la modélisation proprement dite.

Tout au long du processus de modélisation, une attention particulière sera portée à la simplicité du modèle final, en cherchant à réduire autant que possible la complexité du problème et du modèle. Les variables, et donc les données, représentent un coût non négligeable ; c'est pourquoi proposer un modèle à la fois simple et parcimonieux constitue un objectif central de cette étude.

Un point d'attention portera sur la multiplicité des variables proposées et l'absence d'un contexte clairement défini, ce qui rend nécessaire d'effectuer la démarche tout en la reliant aux objectifs métiers. Un notebook accompagne ce rapport, dans lequel toutes les analyses sont clairement décrites, ainsi que les différents choix méthodologiques retenus et leur pertinence au regard de la problématique étudiée.

Enfin, une remarque sur l'aspect organisationnel de notre groupe : des outils collaboratifs tels que Kaggle et Google Colab ont été utilisés afin de faciliter le travail à plusieurs sur le code Python. Grâce à des points réguliers et à des échanges constants, les choix méthodologiques ont été discutés et validés collectivement par l'ensemble des membres du groupe.

2. Chargement et préparation des données

Le jeu de données utilisé pour cette analyse regroupe des caractéristiques sur 90 247 contrats auto issus d'une compagnie d'assurance. Après suppression des doublons, en donnant la priorité aux contrats résiliés, le jeu est réduit à 90 243 lignes. Chaque ligne représente un contrat, et les colonnes fournissent diverses informations pertinentes pour l'évaluation du risque de résiliation. Au total, 58 variables sont disponibles, et la variable NO_AFR joue le rôle d'identifiant unique pour chaque contrat.

2.1 Focus sur les variables sans définition

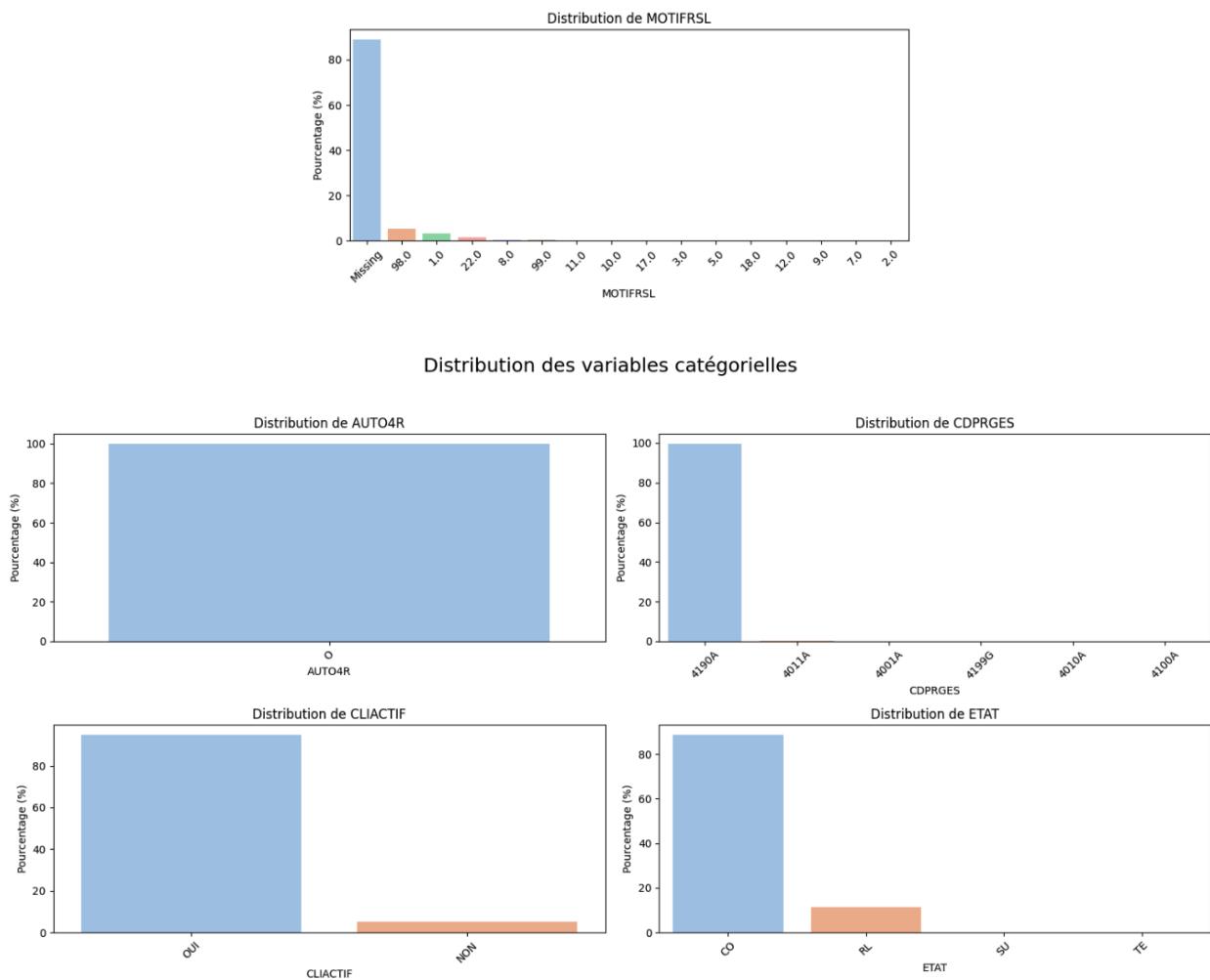
En comparant les colonnes du fichier de données principal avec celles du fichier Excel fourni à titre descriptif, nous identifions plusieurs variables supplémentaires non documentées. L'objectif est de formuler des hypothèses sur leur signification et d'évaluer leur utilité dans notre contexte.

Parmi ces variables, cinq sont de type numérique : CRM, COEFFCOMM, COEFFPLT, NOCLIGES et U :

- Aucune hypothèse n'a été formulée sur la variable U pour l'instant. Son interprétation ou son utilité reste à définir dans le cadre de l'analyse.
- NOCLIGES semble être un identifiant unique (beaucoup de valeurs sous forme de code). En raison de sa nature, il est peu pertinent pour l'analyse statistique ou la modélisation. Cette variable sera donc exclue.
- CRM correspond au coefficient bonus-malus. Un CRM inférieur à 100 indique un bonus, tandis qu'un CRM supérieur à 100 indique un malus. Ce coefficient dépend des antécédents du client, notamment en lien avec les sinistres.
- COEFFCOMM pourrait représenter un coefficient commercial. Il pourrait s'agir d'une note interne attribuée selon un processus spécifique, basée sur des critères d'appétence client. Cette note pourrait être utilisée dans des actions de relance ou pour estimer un risque de résiliation.
- Il est difficile de formuler une hypothèse claire sur la variable COEFFPLT pour le moment. Elle semble être un coefficient, proche de COEFFCOMM, car ses valeurs sont centrées autour de 1, alors que celles de COEFFCOMM sont centrées autour de 100.

L'analyse à l'aide de la densité de distribution, des valeurs uniques et de la forme des valeurs a permis ces hypothèses. Nous resterons prudents quant à leur utilité, en vérifiant si l'information n'est pas déjà présente dans les variables définies dans le descriptif (forme de corrélation).

Les variables catégorielles concernées sont MOTIFRSL, CLIACTIF, ETAT, CDPRGES et AUTO4R :



On peut supposer que ETAT correspond à l'état du contrat, avec par exemple la modalité "SU" signifiant "suspendu". Les autres modalités restent à éclaircir. Cette variable semble très proche de la cible (CONTRAT), ce qui rend pertinent de la supprimer.

Après étude, les deux variables sont très similaires : la modalité "CO" (la plus fréquente) semble correspondre à "COURANT", et "RL" à "RÉSILIÉ". Or, avoir deux cibles dans le modèle fausserait les résultats. On préfère donc conserver uniquement la variable CONTRAT, qui présente clairement deux modalités.

La variable CLIACTIF semble indiquer le statut actif/inactif du client ("OUI" / "NON"), ce qui paraît pertinent à conserver. Toutefois, actif au sens de quoi ? Cela reste à définir plus précisément.

En revanche :

- AUTO4R ne contient qu'une seule modalité → peu informatif, à supprimer.
- CDPRGES présente 6 codes différents dont la signification est inconnue (à clarifier avec les équipes métier), mais une modalité représente plus de 99 % des occurrences → peu informatif, à supprimer.
- MOTIFRSL contient environ 90 % de valeurs manquantes → variable peu pertinente, à supprimer.

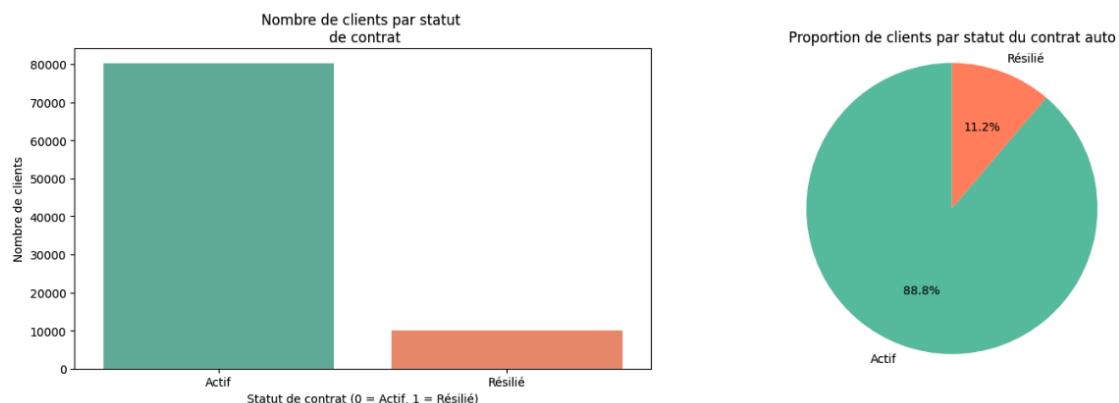
Une autre hypothèse serait que la majorité des clients ne résilient pas. Pour cette population, on ne dispose donc pas d'information. En revanche, ceux qui résilient ont un motif spécifique, ce qui expliquerait le déséquilibre.

Toutefois, nous n'avons toujours pas d'information sur la nature exacte des résiliations (motifs 98, 1, 22, etc.). De toute façon, MOTIFRSL n'est connu qu'après la résiliation.

Enfin, nous pourrions conserver les variables suivantes dans le processus de modélisation :

- CRM,
- COEFFCOMM,
- COEFFPLT,
- CLIACTIF.

3. Analyse exploratoire des données (EDA)

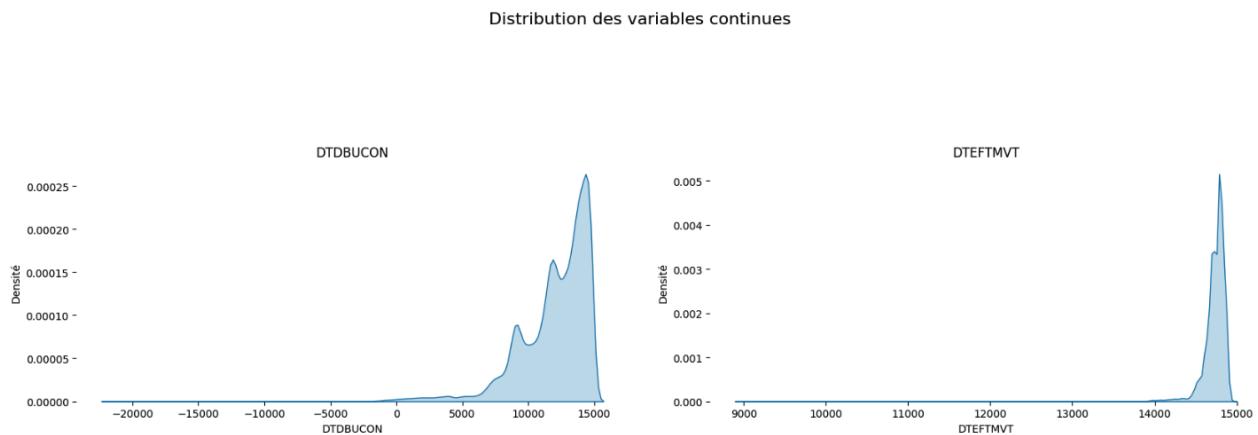


Nous observons un déséquilibre marqué entre les classes, avec une large majorité de contrats auto actifs (CONTRAT = 0). Le ratio est d'environ 90 % de contrats actifs contre 10 % de contrats résiliés.

3.1 Analyse descriptive univariée et bivariée

3.1.1 Variables quantitatives - Lien avec la cible (résiliation)

À l'aide des courbes de densité et des statistiques descriptives (moyenne, médiane, écart-type ainsi que les quantiles d'ordre 1 %, 5 % à 99 %), nous formulons un premier avis sur la définition et nature des variables quantitatives et leur distribution.



La variable DTDBUCON correspond à la date de début du contrat, c'est-à-dire la date de souscription par le client. Elle semble être encodée sous forme numérique, probablement en nombre de jours depuis une date d'origine, comme le 01/01/1900. Il est utile de noter que la table date de février 2006, ce qui est important pour l'interprétation et la notion de “contrats récents”.

La variable DTEFTMVT représente la date du dernier mouvement sur le contrat, comme une modification, une mise à jour ou un avenant. Elle est aussi encodée sous forme numérique, et ses valeurs sont centrées autour de 14 730, ce qui indique des dates récentes.

La variable MMJECHPP indique la date d'échéance annuelle du contrat. Elle est exprimée au format Mois/Jour (par exemple 101 = 1er janvier, 1231 = 31 décembre). Cette variable peut être utile pour étudier une éventuelle saisonnalité dans les résiliations ou les renouvellements.

La variable MTPAATTC désigne le montant total payé par le client pour son contrat d'assurance. C'est une variable continue, probablement exprimée en euros. Elle peut être utilisée comme variable explicative dans certaines analyses.

La variable DTOBTPDC correspond à la date d'obtention du permis de conduire du client. Elle est encodée comme DTDBUCON en nombre de jours. Certaines valeurs sont négatives, ce qui nécessite un contrôle. Cette variable est un bon indicateur de l'expérience du conducteur.

La variable DTPMRMCI indique la date de première mise en circulation du véhicule. Elle est encodée en format numérique (jours). La présence de valeurs négatives suggère qu'un nettoyage ou une conversion est nécessaire avant utilisation.

Les variables liées aux sinistres (S_0_O à S_3_O pour les sinistres responsables, et S_0_N à S_3_N pour les non responsables) indiquent le nombre de sinistres déclarés par le client au cours des quatre dernières années, année par année. Ces variables, bien que discrètes et faiblement dispersées, permettent d'identifier les clients ayant accumulé plusieurs sinistres. Cela peut refléter un comportement à risque. Leurs moyennes sont autour de 1. Elles peuvent être pertinentes pour l'analyse.

La variable NIVBM représente le niveau de bonus–malus. Elle est codée et présente une distribution très asymétrique à droite, avec un pic massif autour de 50–60. Cela suggère une échelle normalisée (par exemple 50 = bonus de référence, 100 = malus). La fréquence très élevée autour d'une valeur centrale peut indiquer un effet plancher ou plafond. Les valeurs supérieures à 100 doivent être vérifiées.

La variable MTPAAREF correspond au montant de prime ou de référence tarifaire. Sa distribution est très étalée avec une longue queue à droite. La majorité des montants se situe entre 1 000 et 10 000 euros, mais certains dépassent 80 000 euros, ce qui sont probablement des valeurs aberrantes ou des cas particuliers (multi-contrats, risques élevés). Il est recommandé de les contrôler avant modélisation.

La variable NBCTRES correspond au nombre de contrats résiliés dans l'historique du client ou du foyer. Elle est fortement concentrée sur les valeurs 0 et 1. Les valeurs supérieures à 5 sont rares, et au-delà de 10 ou 20, une revue qualité est recommandée. Cela montre que la plupart des clients résilient peu. Cette variable peut être binarisée (0 vs ≥ 1) ou capée.

La variable ANCCLI représente l'ancienneté du client en jours. La présence de valeurs négatives indique un problème de calcul ou de référence temporelle. La majorité des valeurs se situe entre 5 000 et 15 000 jours (soit environ 14 à 41 ans), ce qui est cohérent si la base couvre des relations de long terme. Il faut exclure les valeurs négatives et vérifier la date de référence. Une conversion en années peut être utile.

La variable DT_NAI correspond à la date de naissance transformée, exprimée en jours relatifs à une date de référence non connue. Elle est centrée autour de 20 000 jours, ce qui correspond à un âge moyen de 55 ans si la référence est 2006. Les valeurs inférieures à -100 000 ou supérieures à 0 (naissance future) sont aberrantes. Il est nécessaire de reconvertis en âge et de contrôler les bornes raisonnables (par exemple entre 16 et 100 ans).

La variable DI indique le nombre de contrats divers actifs détenus par le client. La moyenne est de 0,27, la médiane est à 0, et le maximum à 16. La distribution est très concentrée sur 0. Cette variable est discrète et très asymétrique. Une valeur de 0 peut indiquer un faible engagement. Elle est intéressante pour modéliser la fidélité.

La variable IV indique le nombre de contrats individuels-vie actifs. La moyenne est très faible (0,065), la médiane est à 0 et le maximum à 26. La distribution est ultra concentrée, avec très peu de clients concernés. Elle a un faible pouvoir explicatif sauf si elle est réagréée en binaire (présence ou non).

La variable MH correspond au nombre de contrats multirisques habitation actifs. Sa moyenne est de 0,69, la médiane est à 0, et le maximum à 19. La majorité des clients n'en a pas, mais certains en ont plusieurs. Elle est utile pour mesurer l'engagement dans d'autres produits.

La variable SA indique le nombre de contrats santé actifs. Sa moyenne est de 0,10, la médiane à 0 et le maximum à 8. Comme pour IV, la distribution est très concentrée sur 0. Elle peut être utilisée en version binaire (0 vs ≥1).

La variable NBCTACT représente le nombre total de contrats actifs, toutes branches confondues. Sa moyenne est de 2,93, la médiane de 2, et le maximum de 43. La variable est bien étalée et utile pour mesurer le degré d'engagement du client. Elle est pertinente pour modéliser la fidélité.

La variable RESAU4R indique le nombre de contrats auto résiliés. La moyenne est élevée (0,90) sur une sous-population de 26 506 clients. La distribution est concentrée entre 0 et 2, avec un maximum de 12. Elle reflète un historique de ruptures dans l'assurance auto, ce qui est très pertinent.

La variable RESDI indique le nombre de contrats divers résiliés. La moyenne est de 0,16, la médiane est à 0, et le maximum à 7. Elle est complémentaire de RESAU4R et permet d'identifier des profils instables sur plusieurs produits.

La variable RESIV correspond au nombre de contrats individuels-vie résiliés. La moyenne est très faible (0,0017), la médiane est à 0, et le maximum à 2. Cette variable est très rarement renseignée, donc peu exploitable, sauf sous forme binaire.

La variable RESMH indique le nombre de contrats multirisques habitation résiliés. Sa moyenne est de 0,27, la médiane à 0, et le maximum à 7. Elle peut être intéressante combinée avec la variable MH pour évaluer la stabilité du client sur ce produit.

La variable RESSA correspond au nombre de contrats santé résiliés. Sa moyenne est de 0,063, et le maximum est de 5. La distribution est rare, mais elle peut servir à caractériser des foyers instables.

La variable NBCTRES donne le nombre total de contrats résiliés toutes branches confondues. La moyenne est de 0,47, la médiane à 0, et le maximum à 26. Elle est bien étalée et très discriminante pour modéliser le risque de résiliation. Elle peut être transformée en classes (0 / 1–3 / >3) ou capée.

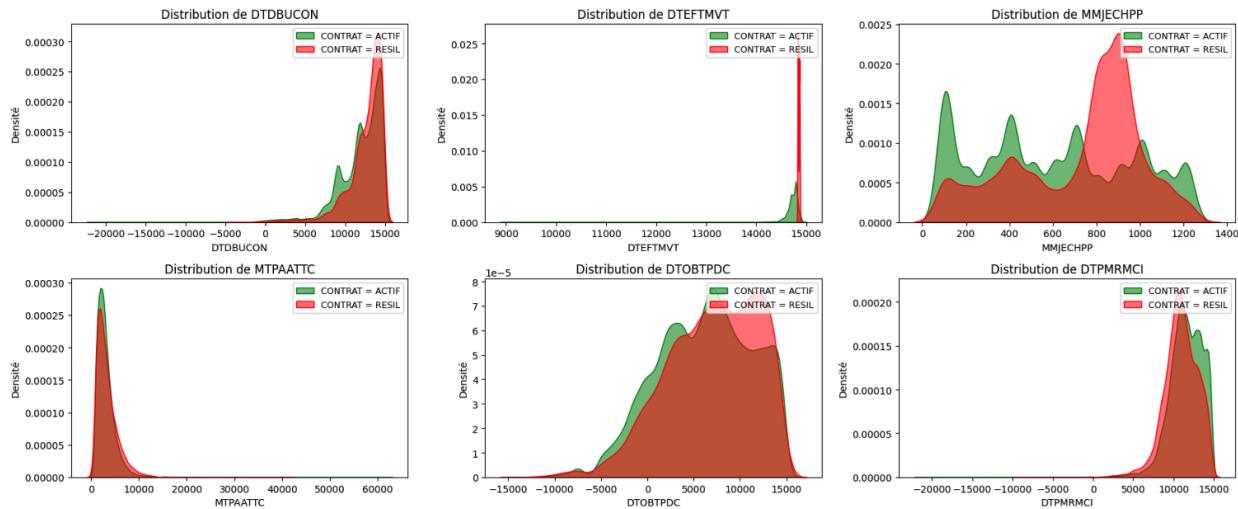
La variable COEFFCOMM est un coefficient de majoration, appliqué en lien avec la tarification ou le profil du client. Sa moyenne est de 97,56, la médiane de 100, avec des valeurs entre 30 et 595. La distribution est très asymétrique, avec un effet plafond autour de 100. Les valeurs supérieures à 200 doivent être vérifiées, car elles peuvent refléter une tarification élevée.

La variable COEFFPLT est un coefficient de plafonnement ou d'ajustement tarifaire. La moyenne est de 0,95, les valeurs vont de 0,29 à 5,41, et la médiane est proche de 0,96. La distribution est resserrée autour de 1, mais des valeurs extrêmes sont présentes. Cela peut indiquer des ajustements spécifiques.

Enfin, la variable CRM correspond au coefficient de réduction-majoration, aussi appelé bonus-malus. Sa moyenne est de 59,1, la médiane de 50, et le maximum de 289. La distribution est très asymétrique, avec un pic sur 50 et une longue traîne. C'est une variable importante car elle est liée à la sinistralité passée et au niveau de tarification.

Nous croisons ensuite cette distribution en distinguant les observations selon le statut de la variable cible CONTRAT, afin d'identifier les différences de comportements à l'aide d'une lecture visuelle des graphiques suivants. Étant donné le nombre important de variables, les graphiques restants seront présentés en annexe.

Distribution des variables numériques selon la résiliation du contrat



Hypothèses initiales sur le lien entre les variables numériques et la variable cible :

Les contrats les plus récents semblent présenter un risque plus élevé de résiliation : la variable DTDBUCON (date de début du contrat) montre que les contrats actifs, représentés en vert sur la courbe, sont généralement plus anciens, tandis que les contrats résiliés, en rouge, ont été souscrits plus récemment.

Il est possible qu'un événement particulier, tel qu'une modification du contrat ou un sinistre, précède directement la résiliation. La variable DTEFTMVT (dernier mouvement sur le contrat) révèle en effet que presque tous les contrats résiliés ont connu un mouvement très récent, visible sous forme d'un pic soudain, alors que les contrats actifs affichent des mouvements plus étalés dans le temps.

Certains moments clés de l'année semblent propices aux résiliations, peut-être en raison de l'effet saisonnier ou des politiques de renouvellement : la variable MMJECHPP (date d'échéance) indique que les contrats résiliés présentent des échéances très concentrées autour de quelques dates, tandis que les contrats actifs sont répartis de façon plus uniforme.

Les clients qui paient une prime annuelle faible paraissent plus sensibles au prix et changent plus facilement d'assureur. Bien que les courbes soient proches pour la variable MTPAATT (montant de la prime), on observe une légère surreprésentation des contrats résiliés parmi les primes les plus faibles.

Les conducteurs expérimentés sont en général plus stables : avec la variable DTOBTPDC (date du permis), on constate que les contrats actifs appartiennent plus souvent à des conducteurs dont le permis est ancien, même si la distinction reste subtile.

Le changement de véhicule peut favoriser la résiliation. La variable DTPMRMCI (mise en circulation du véhicule) montre que les contrats résiliés sont légèrement liés à des véhicules plus récents, suggérant qu'un changement d'auto est parfois l'occasion de comparer les offres ou de résilier.

Les graphiques des variables S_0_N à S_3_N indiquent d'abord une prédominance numérique des contrats actifs : par exemple, pour S_0_N = 1, on dénombre 8 540 actifs contre 965 résiliés. Toutefois, le rapport résiliés/actifs augmente régulièrement avec le nombre de sinistres non responsables dans chaque période, signe qu'une sinistralité même non fautive accroît la probabilité de résiliation. Enfin, le volume total diminue d'une année à l'autre, ce qui reflète la mémoire limitée des sinistres ou le recul du nombre de clients historiques.

Pour les variables S_0_O à S_3_O, la corrélation entre le nombre de sinistres responsables et la résiliation est encore plus marquée : dans S_0_O, le ratio résiliés/actifs passe d'environ 17 % chez les clients ayant un sinistre à plus de 42 % chez ceux ayant deux. Sur la période S_3_O, la majorité des sinistres responsables concerne même des contrats résiliés (2 450 résiliés contre 381 actifs). Ce basculement suggère un seuil au-delà duquel la répétition des sinistres devient déterminante pour la résiliation, conférant à ces variables une forte valeur prédictive.

Les clients affichant un niveau de bonus-malus plus élevé sont légèrement plus exposés à la résiliation. La variable NIVBM montre des courbes similaires pour actifs et résiliés, mais la densité des résiliés augmente dans les valeurs élevées, ce qui peut traduire une sinistralité plus importante ou un comportement défavorable à l'assureur.

Les primes annuelles faibles sont également surreprésentées parmi les résiliés : avec MTPAAREF, la densité des résiliés est plus forte en dessous de 5 000 €, suggérant une plus grande sensibilité au prix ou une rentabilité insuffisante ressentie par l'assureur.

Un historique de résiliations accroît nettement le risque futur. À mesure que le nombre total de contrats résiliés (NBCTRES) augmente, la densité des résiliés devient dominante, confirmant que les clients ayant déjà résilié ailleurs sont particulièrement à risque.

L'ancienneté du client (ANCCLI) joue aussi un rôle : les contrats actifs correspondent à des clients plus anciens tandis que les résiliés se concentrent sur des anciennetés intermédiaires, ce qui indique des ruptures précoces de la relation.

L'âge du client estimé par DT_NAI n'apparaît pas fortement discriminant, bien que l'on constate une légère densité supplémentaire de résiliés parmi les plus jeunes, point qui mériterait une analyse plus poussée.

Pour les variables liées au nombre de contrats actifs (DI, IV, MH, SA, NBCTACT), les résiliés se concentrent sur les clients ayant zéro ou très peu de contrats. Les clients faiblement engagés — un seul contrat et peu ou pas de multirisques, vie ou santé — sont plus exposés au départ, alors que ceux disposant de plusieurs produits restent plus fidèles.

Les variables RESAU4R, RESDI, RESIV, RESMH, RESSA et NBCTRES montrent que les clients résiliés ont beaucoup plus souvent résilié dans d'autres branches. Un client ayant déjà rompu un contrat auto ou autre présente un risque bien plus élevé de résiliation ultérieure, ce qui traduit une forte corrélation négative avec la fidélité.

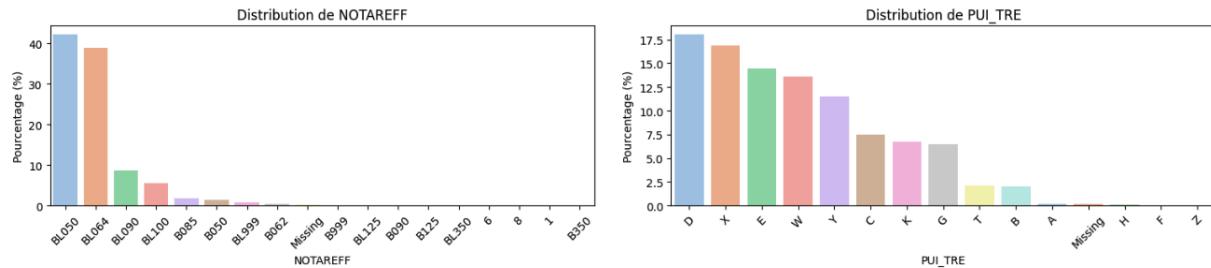
Les coefficients COEFFCOMM et COEFFPLFT ainsi que le bonus-malus CRM diffèrent sensiblement entre actifs et résiliés. Les résiliés présentent un COEFFCOMM plus élevé, laissant penser à une tarification défavorable ou mal perçue ; un COEFFPLFT légèrement supérieur, signe de plafonnements tarifaires plus fréquents ; et un CRM plus élevé, reflet d'une sinistralité passée. Ces indicateurs décrivent un profil à risque ou une prime élevée, tous deux susceptibles d'expliquer la résiliation.

3.1.2 Variables qualitatives - Lien avec la cible (résiliation)

Nous effectuons, pour les variables catégorielles, le même travail d'analyse. Il s'agit d'examiner les différentes modalités que ces variables contiennent, leur fréquence, le nombre de modalités distinctes, le mode, ainsi qu'une première idée de leur lien avec la variable cible, afin de détecter d'éventuelles différences de comportement selon le statut ACTIF ou RÉSILIÉ. Les valeurs manquantes seront remplacées par la modalité « missing » afin de pouvoir observer leur présence et leur poids dans l'analyse.

	count	unique	top	freq
CD_ACTIF	90247	1255	A00676	524
CD_FML	89702	25	T	37944
CDPRGES	90247	6	4190A	89973
AUTO4R	90247	1	O	90247
CDMARVEH	90160	256	RENAULT	27666
LBMDLVH	90137	6806	205	4626
NOTAREFF	89980	17	BL050	38096
PUI_TRE	90081	14	D	16303
RN_VL_VH	90077.00	18.00	7.00	10499.00
CDUSGAUT	90231.00	71.00	611.00	38575.00
CDMCE	90247	2	P	77096
CD_SEX	90147	2	M	67342
CDSITFAM	89887	7	M	64552
DEPT	90048.00	196.00	59.00	3374.00
REGION	90048	11	NE	20170
CD_CSP	50585.00	234.00	5.00	12697.00
CLIACTIF	90247	2	OUI	85675

Distribution des variables catégorielles



La variable AUTO4R sera supprimée, comme justifié précédemment. Elle contient une seule modalité, "O", ce qui signifie qu'elle n'apporte aucune variabilité et n'est donc pas utile.

La variable CD_AGT présente un très grand nombre de modalités. Elle ne peut pas être utilisée en l'état. Il faudra regrouper certaines modalités (par exemple en utilisant la méthode du Chi-2 ou par fusion de catégories) pour former des groupes de risque exploitables. Elle ne contient pas de valeurs manquantes. Il s'agit de l'agent ayant vendu le contrat. Les comportements commerciaux peuvent varier selon l'agent. On recense 1255 agents. L'agent le plus fréquent est A00676, présent dans 524 contrats.

La variable CD_FML correspond au code formule du contrat auto (comme tiers, tous risques, etc.). Elle contient 25 modalités. La modalité majoritaire est T, présente dans 37 944 contrats sur 90 247. Elle contient très peu de valeurs manquantes. Il pourrait être utile de regrouper certaines modalités.

La variable CDPRGES est une variable dont la signification exacte n'est pas connue. Elle est bien renseignée, mais la modalité 4190A représente 99 % des cas. Cette forte concentration la rend peu pertinente, donc elle sera supprimée. Elle contient 6 modalités au total, et 4190A apparaît environ 89 000 fois.

La variable CDMARVEH indique la marque du véhicule. C'est une variable intéressante. Elle contient 256 modalités. Il pourrait être pertinent de regrouper certaines marques. Elle contient très peu de valeurs manquantes. Il s'agit du type de constructeur (Renault, Peugeot, etc.). La marque la plus fréquente est RENAULT, avec 27 600 occurrences, ce qui suggère qu'il s'agit probablement d'un assureur français.

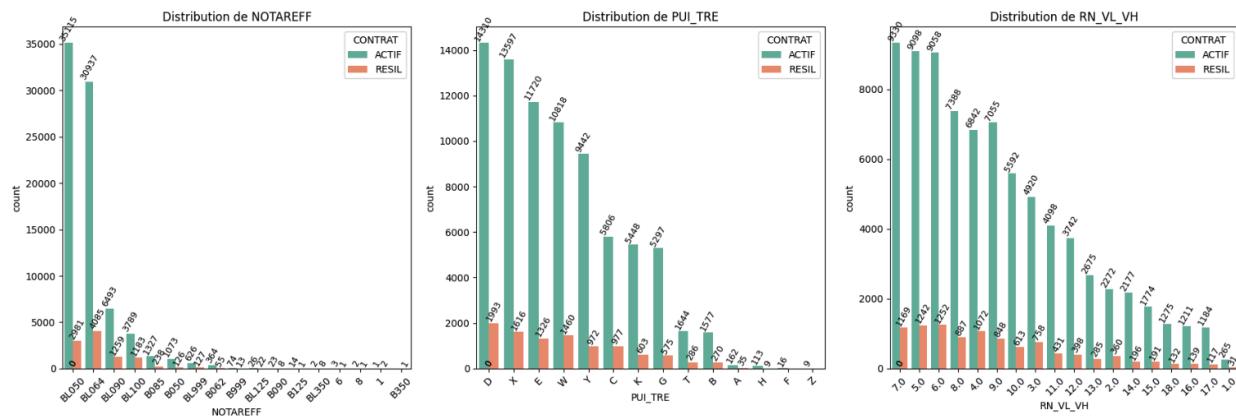
La variable LBMDLVH semble représenter le modèle du véhicule. Elle est souvent liée à la variable CDMARVEH. Elle contient un grand nombre de modalités (6 806). Il y a très peu de valeurs manquantes. La modalité la plus fréquente est "205" (Peugeot 205 ou Renault ?).

La variable NOTAREFF correspond au numéro tarifaire, probablement en lien avec le système de bonus-malus. Elle est utilisée dans la tarification. Elle peut aider à identifier des cas de sur-tarification. Elle contient très peu de valeurs manquantes. Elle contient 17 modalités. La modalité la plus fréquente est BL050, présente dans 38 000 contrats.

La variable PUI_TRE correspond à la puissance fiscale du véhicule. C'est une variable réglementaire importante en assurance auto. Elle est bien renseignée, avec très peu de valeurs manquantes. Elle comprend 14 modalités. La plus fréquente est D, présente dans 16 000 cas.

La variable RN_VL_VH indique le rang ou niveau de valeur du véhicule, souvent en lien avec sa cotation Argus. Elle est utile pour estimer le coût d'un sinistre. Elle contient très peu de valeurs manquantes. Il y a 18 niveaux, et la valeur la plus fréquente est 7, avec environ 10 000 occurrences.

La variable CDUSGAUT est le code d'usage du véhicule (usage personnel, professionnel, etc.). Elle contient 71 modalités. La modalité la plus fréquente est 611, présente dans 38 000 cas.



Hypothèses initiales sur le lien entre les variables catégorielles et la variable cible :

Pour la variable CD_FML, la distribution est fortement concentrée autour de quelques modalités dominantes, notamment T, C et S. Les résiliations sont présentes dans toutes les modalités, mais elles suivent la même répartition que les contrats actifs. Cela pourrait réduire la capacité discriminante de cette variable.

La variable CDPRGES présente une écrasante majorité sur une seule modalité, 4190A, aussi bien pour les contrats résiliés que pour les actifs. Cette faible variabilité rend la variable peu informative pour le scoring en l'état.

Concernant AUTO4R, la variable est quasi constante : la majorité des clients possèdent un contrat auto 4R. Elle offre donc peu de pouvoir explicatif.

Pour NOTAREFF, on observe quelques modalités fréquentes comme BL050 et BL064, avec une structure globale similaire entre les résiliés et les actifs. Toutefois, certaines modalités sont davantage associées aux contrats résiliés, ce qui pourrait la rendre utile.

La variable PUI_TRE montre une distribution des contrats résiliés relativement similaire à celle des actifs. Cependant, des écarts dans certaines classes peuvent justifier son usage dans un modèle prédictif.

La variable RN_VL_VH semble plus granulaire, avec une distribution dégressive. Les contrats résiliés apparaissent légèrement plus fréquents dans certaines valeurs élevées (supérieures à 12), ce qui peut indiquer une corrélation avec le risque de résiliation.

La variable CDMCE présente deux grandes modalités clairement identifiables, avec une répartition plus équilibrée entre contrats actifs et résiliés. Cela suggère que ce critère de segmentation client pourrait influencer la fidélité.

La variable CD_SEX montre une majorité d'hommes dans l'échantillon. Les proportions de résiliés sont comparables entre les genres, ce qui limite le pouvoir discriminant de cette variable prise seule.

Concernant CDSITFAM, la modalité M (probablement "marié") domine largement. Toutefois, les résiliations sont proportionnellement plus fréquentes dans certaines situations familiales comme C ou D, ce qui pourrait signaler des profils moins stables.

La variable REGION offre une bonne variabilité géographique avec des disparités notables (NE > SO > MM...). Certaines régions présentent des taux de résiliation plus élevés, ce qui peut révéler des comportements régionaux spécifiques à modéliser.

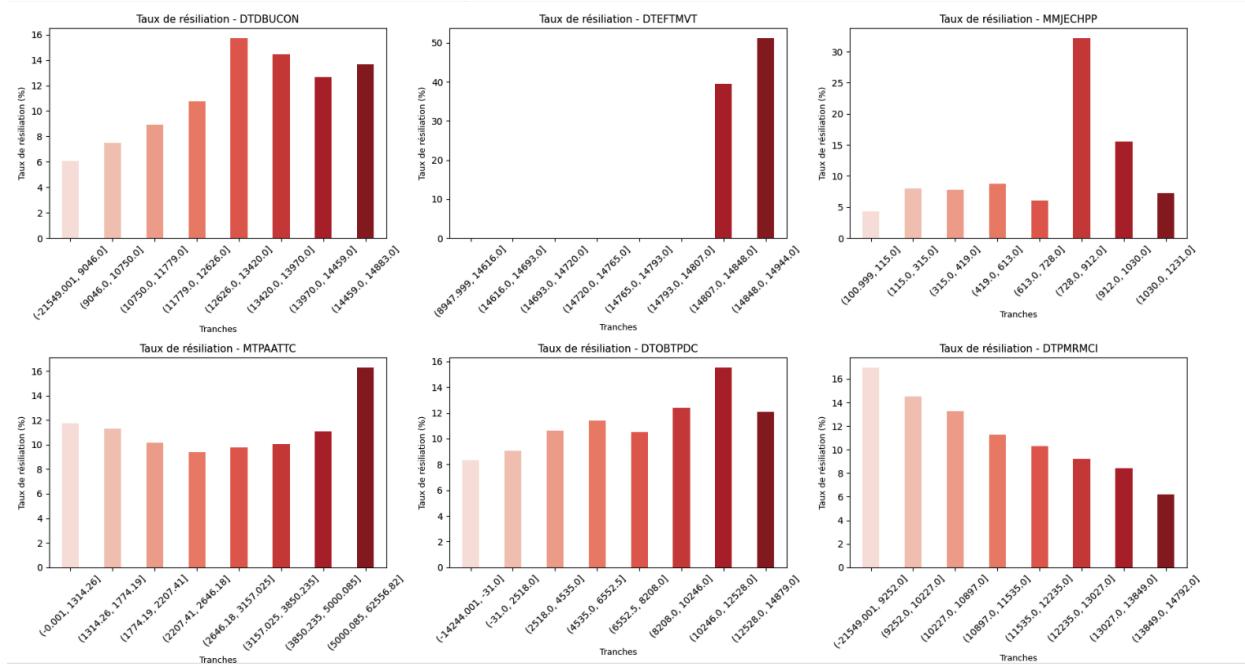
Enfin, la variable CLIACTIF est hautement discriminante : tous les contrats actifs sont marqués OUI, tandis que la modalité NON n'existe que chez les résiliés. Elle est donc fortement corrélée à la cible et indique clairement un risque de fuite. Toutefois, son usage doit être encadré pour éviter le data leakage.

4. Analyse des interactions complexes

4.1 Croisement de variables numériques avec le taux de résiliation

Cette analyse repose sur des graphiques représentant le taux de résiliation par quantiles (8, dans notre cas) pour plusieurs variables explicatives. L'objectif est d'interpréter le comportement de la variable cible (CONTRAT) en fonction de ces variables.

Au niveau des variables quantitatives (présentant une grande variabilité et un grand nombre de données) :



DTDBUCON – Date de début du contrat

Le taux de résiliation augmente nettement avec la date. Les contrats plus récents sont davantage concernés, ce qui peut s'expliquer par un effet d'essai ou une insatisfaction dès les premiers mois.

DTEFTMVT – Date du dernier mouvement sur le contrat

Les contrats anciens présentent un très faible taux de résiliation. En revanche, un pic brutal (environ 50 %) apparaît dans les dates les plus récentes. Cela indique qu'une activité récente sur le contrat est fortement corrélée à une résiliation prochaine.

DTOBTPDC – Date d'obtention du permis de conduire

Le taux de résiliation augmente avec la récence du permis. Les jeunes conducteurs, souvent plus sensibles aux prix ou aux offres concurrentes, se révèlent plus volatils.

DTPMRMCI – Date de mise en circulation du véhicule

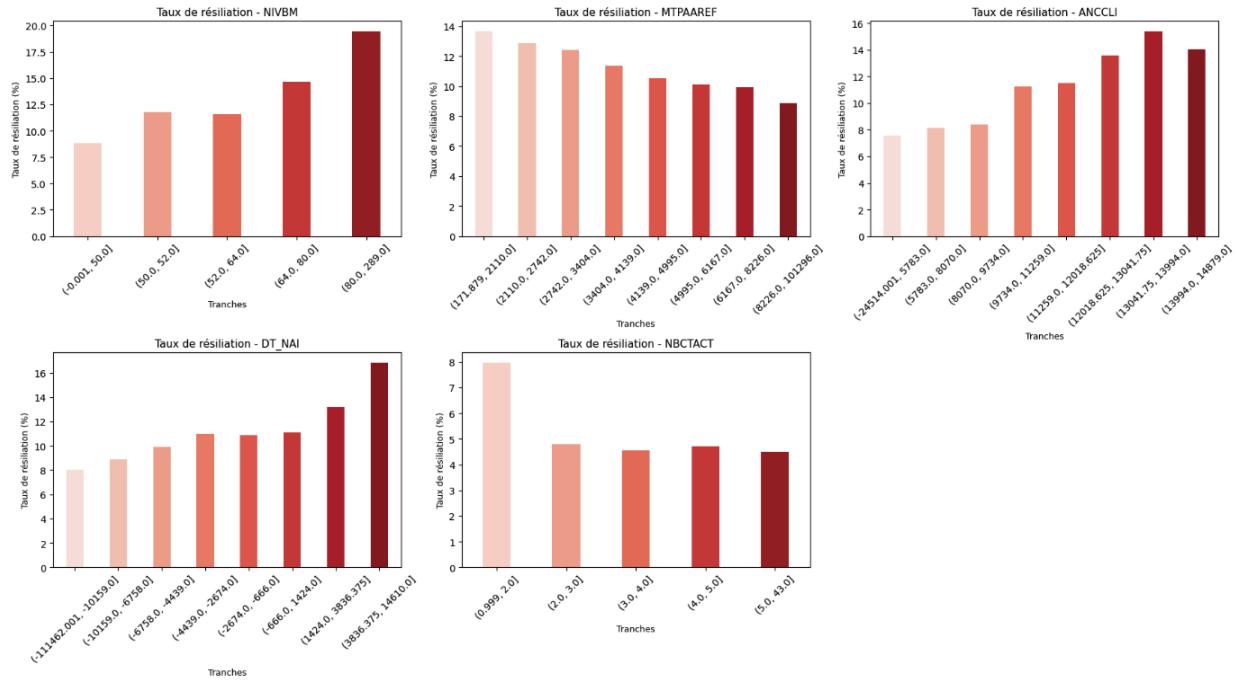
La résiliation diminue lorsque le véhicule est plus récent. Les véhicules anciens sont plus souvent associés à une résiliation, possiblement en lien avec des primes plus élevées ou un véhicule en fin de cycle.

MMJECHPP – Date d'échéance du contrat (MMJJ)

Un pic net de résiliation apparaît dans un seul intervalle (entre 728 et 912). Cela suggère que certaines périodes de l'année (éventuellement l'été) sont plus propices à la résiliation.

MTPAACCC – Montant annuel de la prime

La courbe présente une forme de U inversé. Le taux de résiliation est relativement stable jusqu'à un certain niveau, puis il augmente nettement dans les tranches les plus élevées. Les clients avec une prime importante sont donc plus enclins à résilier.



NIVBM – Niveau de bonus-malus

Le taux de résiliation augmente avec le niveau de malus. Les clients bénéficiant d'un bonus élevé (valeurs inférieures à 50) résilient beaucoup moins. À l'inverse, ceux ayant un malus (valeurs supérieures à 80) présentent un taux de résiliation nettement plus élevé, atteignant près de 20 %. Cela confirme un lien entre le comportement routier (ou la tarification qui en découle) et la probabilité de résiliation.

MTPAAREF – Montant de prime annuelle de référence

On observe une tendance décroissante : les clients dont la prime est faible (inférieure à 2 000 euros) résilient davantage. Cela reflète une sensibilité au prix de la part des assurés.

ANCCLI – Ancienneté du client

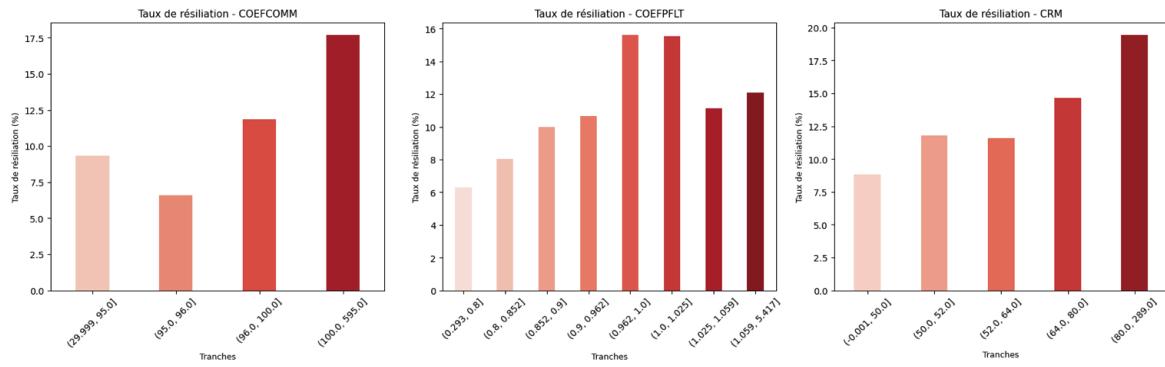
Il existe une corrélation positive : le taux de résiliation augmente avec l'ancienneté du client, jusqu'à un certain seuil. Une hypothèse possible est que les clients ayant une ancieneté moyenne sont les plus stables, tandis que les très anciens finissent parfois par partir.

DT_NAI – Date de naissance (en jours)

La résiliation est plus fréquente chez les clients les plus jeunes (dates les plus récentes). À l'inverse, les assurés plus âgés résilient moins. Cela confirme l'idée que les jeunes profils sont plus mobiles, plus sensibles au prix, et plus enclins à changer d'assureur. Cette variable est donc utile à conserver, notamment après transformation en âge.

NBCTACT – Nombre de contrats actifs

Le taux de résiliation diminue clairement avec le nombre de contrats détenus. Les clients ayant seulement un ou deux contrats sont les plus exposés à la résiliation.



COEFCOMM – Coefficient de majoration (lié à la tarification ou au profil)

On observe une forte augmentation du taux de résiliation pour les valeurs supérieures à 100. Les clients ayant un coefficient compris entre 100 et 595 présentent un taux de résiliation nettement plus élevé, atteignant près de 18 %.

COEFPLT – Coefficient de plafonnement tarifaire

Le taux de résiliation augmente progressivement jusqu'à atteindre un pic dans les tranches [0.9, 0.962] et [1.0, 1.025]. On note ensuite un léger recul, suivi d'une reprise modérée dans la tranche la plus élevée. Cela montre que certains ajustements tarifaires sont perçus négativement à partir d'un certain seuil, en particulier autour de la valeur 1. Cette variable peut donc servir d'indicateur d'insatisfaction tarifaire.

CRM – Coefficient de réduction majoration (bonus-malus)

Il existe une hausse nette du taux de résiliation avec l'augmentation du malus. Les clients avec un CRM supérieur à 80 présentent un taux proche de 20 %, contre moins de 10 % pour ceux bénéficiant d'un bonus (CRM < 50). Cela confirme que le malus est un facteur de résiliation, que ce soit pour des raisons de coût ou de perception négative du client par l'assureur.

5. Création de nouvelles variables

Trois nouvelles variables ont été créées pour enrichir l'analyse et mieux capter les comportements à risque.

5.1 Indicateur de responsabilité des sinistres

D'abord, un indice de responsabilité (INDICE_RESPONSABILITE) a été créé pour capturer la répartition entre sinistres responsables et non responsables. Toutes les variables sinistres ont d'abord été nettoyées (remplacement des valeurs manquantes par 0), puis le total de sinistres responsables (TOTAL_SIN_RESP) et non responsables (TOTAL_SIN_N_RESP) a été calculé. L'indice est obtenu en divisant les sinistres responsables par la somme des deux types. Cet indicateur permet de synthétiser le comportement du client en matière de sinistralité et peut être très pertinent (hypothèse initiale) dans un modèle prédictif.

5.2 Écart entre prime réelle et prime de référence

Ensuite, les variables liées à la prime payée (MTPAATTC) et à la prime de référence (MTPAAREF) étant fortement corrélées, il a été jugé plus pertinent de créer une nouvelle variable : le ratio MTPAATTC / MTPAAREF. Ce ratio permet de mesurer l'écart entre le tarif payé et le tarif de référence, ce qui est interprétable comme un indicateur de remise ou de surcoût. Il peut révéler un traitement tarifaire particulier, potentiellement lié à des risques ou à des politiques commerciales.

5.3 Résumés de contrats résiliés

Enfin, pour évaluer plus finement le risque lié à l'historique de résiliations, un score pondéré de résiliation (SCORE_RESIL_PONDERE) a été construit. Il prend en compte le type de contrat résilié avec des poids différents selon la nature du produit (par exemple 3.0 pour l'auto, 2.5 pour l'habitation, etc.). Cela permet de mieux différencier les clients ayant résilié des contrats critiques de ceux ayant résilié des produits moins stratégiques. Une variable dérivée, TAUX_RESIL_PONDERE_TOTAL, est ensuite calculée pour rapporter ce score pondéré au nombre total de contrats, ce qui donne une mesure normalisée du risque de résiliation. Les pondérations avec des coefficients fixes comme 3.0 ou 2.5 sont établies de manière à attribuer plus de poids aux contrats auto, et décroissantes pour les autres types de contrats en fonction de leur impact sur le budget de l'individu. Des hypothèses sont donc formulées ici concernant l'effet de la détention de plusieurs contrats sur le comportement de résiliation, et plus particulièrement sur le contrat auto que l'on cherche à modéliser.

6. Analyse statistique de la dépendance des variables avec la cible

6.1 Lien des variables catégorielles avec la cible via test du chi²

Afin de tester le lien statistique de nos variables catégorielles avec la cible, nous avons utilisé le test du Chi² qui permet d'évaluer s'il existe une dépendance entre deux variables qualitatives. Dans notre cas, il sert à vérifier si une variable catégorielle est liée de manière significative à la variable cible (CONTRAT). On interprète le résultat à partir de la p-value :

- Si p-value < 0,05, on rejette l'hypothèse d'indépendance donc la variable est significativement liée à la cible.
- Si p-value > 0,05, on ne peut pas conclure à une dépendance donc la variable est peu utile pour prédire la cible.

Suite aux tests, voici les résultats :

La majorité des variables présentent une forte dépendance statistique avec la cible (p-value = 0.000) :

- CD_AGT, CD_FML, LBMDLVH, NOTAREFF, PUL_TRE, RN_VL_VH, CDUSGAUT, CDMCE, CD_SEX, CDSITFAM, DEPT, REGION, CD_CSP, CLIACTIF
- Ces variables ont une influence significative sur la probabilité de résiliation.

D'autres variables ne montrent aucun lien significatif et confirment nos hypothèses initiales :

- AUTO4R (p = 1.000), CDPRGES (p = 0.149), CDMARVEH (p = 0.030, significatif à 5%)

En résumé, la majorité des variables catégorielles testées présentent une association significative avec le statut du contrat, et pourront être retenues ou transformées pour améliorer la performance d'un modèle de prédiction.

6.2 Lien statistique entre les variables numériques et la variable cible

Le test de Student (t-test) va permettre de vérifier si la moyenne d'une variable numérique est significativement différente entre deux groupes ici :

- Contrats résiliés (RESIL = 1)
- Contrats actifs (RESIL = 0)

Une p-value < 0.05 indique une différence significative entre les deux groupes et que donc la variable est potentiellement utile pour expliquer la résiliation.

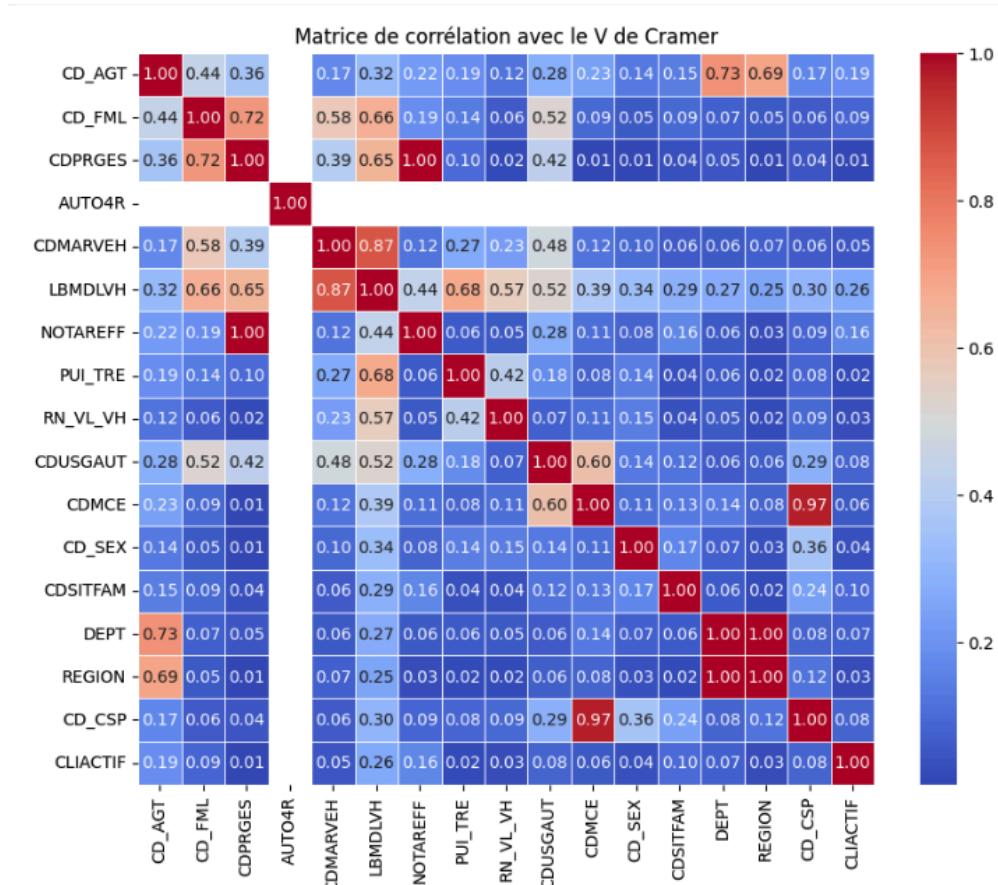
Toutes nos variables sont significatives à 5% sauf quelques variables qui n'affichent aucune différence statistiquement significative entre résiliés et actifs :

- S_2_N (sinistres non responsables en année 2),
- DI (nombre de contrats divers actifs),
- RESIV (nombre de contrats individu-vie résiliés),
- RESSA (nombre de contrats santé résiliés).

6.3 Matrices de corrélation des variables

6.3.1 Variables qualitatives (V de Cramér)

Sur nos variables qualitatives, un moyen d'identifier l'autocorrélation est de calculer le V de Cramér entre deux variables de type catégoriel. L'objectif de ce calcul est, d'une part, d'identifier les variables fortement corrélées entre elles, donc apportant la même information, et par conséquent d'en retirer une afin de simplifier le modèle. D'autre part, cela permet d'éviter la redondance et facilite la mise en place de l'algorithme de régression logistique en réduisant le risque de multicolinéarité. Cette dernière sera également testée par d'autres méthodes, telles que le Lasso ou la sélection pas à pas (stepwise selection). Mais dès cette étape, il est important de se poser la question d'une éventuelle réduction de dimension.



Le V de Cramér est une mesure de l'association entre deux variables catégorielles. Il évalue la force de la relation entre ces variables, avec une valeur comprise entre 0 (aucune association) et 1 (association parfaite). Plus le V de Cramér est élevé, plus les variables sont corrélées, ce qui peut donc indiquer une redondance d'information dans un modèle statistique.

Lors de l'analyse des variables catégorielles, on remarque que certaines sont très proches les unes des autres. Cela signifie qu'elles contiennent presque la même information. Pour éviter les doublons et ne pas compliquer le modèle, il est important de supprimer certaines variables redondantes. Le V de Cramér nous permet justement de mesurer à quel point deux variables sont liées.

Les variables CDPRGES (type de produit) et AUTO4R (nombre de contrats 4 roues actifs) ont été supprimées car elles contiennent une seule modalité très dominante. Dans ce cas, elles n'aident pas à différencier les clients, ce qui les rend peu utiles pour la suite de l'analyse. De plus, le V de Cramér ne peut pas être calculé correctement dans ces cas-là, car il faut de la variabilité dans les données.

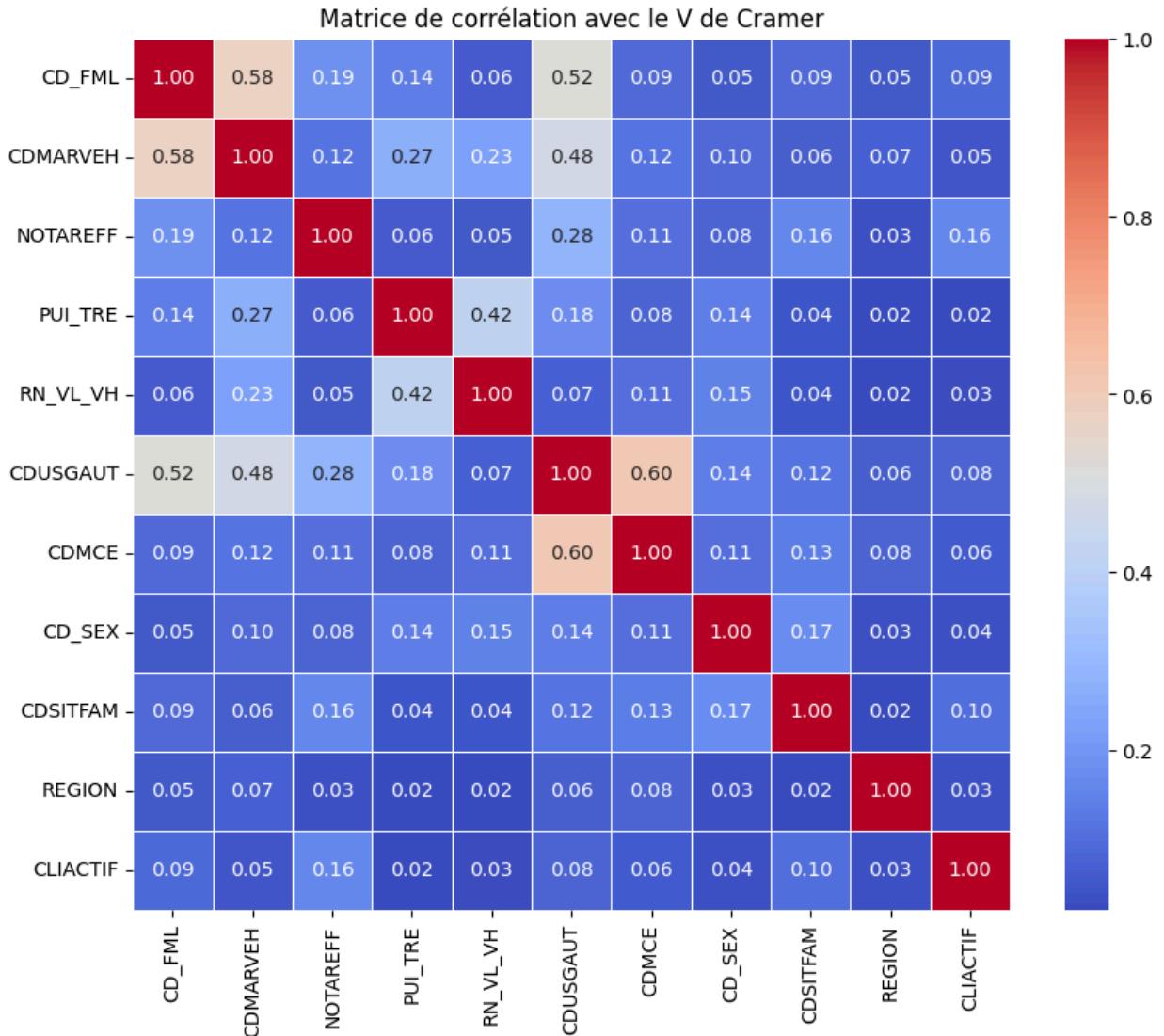
La variable DEPT (département) est très proche de REGION, avec une valeur du V de Cramér de 0,69. Comme les deux donnent une information géographique similaire, nous avons gardé REGION, qui contient moins de modalités et est donc plus simple à utiliser dans un modèle. Supprimer DEPT permet de gagner en simplicité sans perdre d'information importante.

Pour la variable LBMDLVH (rang de valeur du véhicule), on observe qu'elle est très liée à CDMARVEH (marque du véhicule), avec un V de Cramér de 0,87. Nous avons choisi de garder la marque du véhicule, car c'est une information plus facilement compréhensible et exploitable, notamment pour les équipes marketing ou commerciales. Supprimer LBMDLVH permet de réduire la redondance dans le jeu de données.

La variable CD_CSP (catégorie socio-professionnelle) est aussi très proche de CDMCE (marché pro ou particulier), avec un V de Cramér de 0,97. Comme CDMCE est plus simple et contient moins de modalités, elle a été conservée. Supprimer CD_CSP permet d'éviter la multicolinéarité et d'alléger le modèle.

Enfin, la variable CD_AGT (agent ayant vendu le contrat) a été supprimée car elle contient trop de modalités différentes (un identifiant unique pour chaque agent). Cela n'apporte pas de vraie valeur ajoutée au modèle, et rend les calculs plus lourds. En la supprimant, on évite d'introduire du bruit et on améliore la stabilité du modèle.

Après simplification, nous obtenons la matrice de corrélation suivante, que nous jugeons acceptable en termes d'autocorrélation. Certaines valeurs restent modérément élevées, comme 0,58 entre CDMARVEH et CD_FML, ou 0,60 entre CDUSGAUT et CDMCE, mais elles ne justifient pas, à ce stade, une suppression supplémentaire.

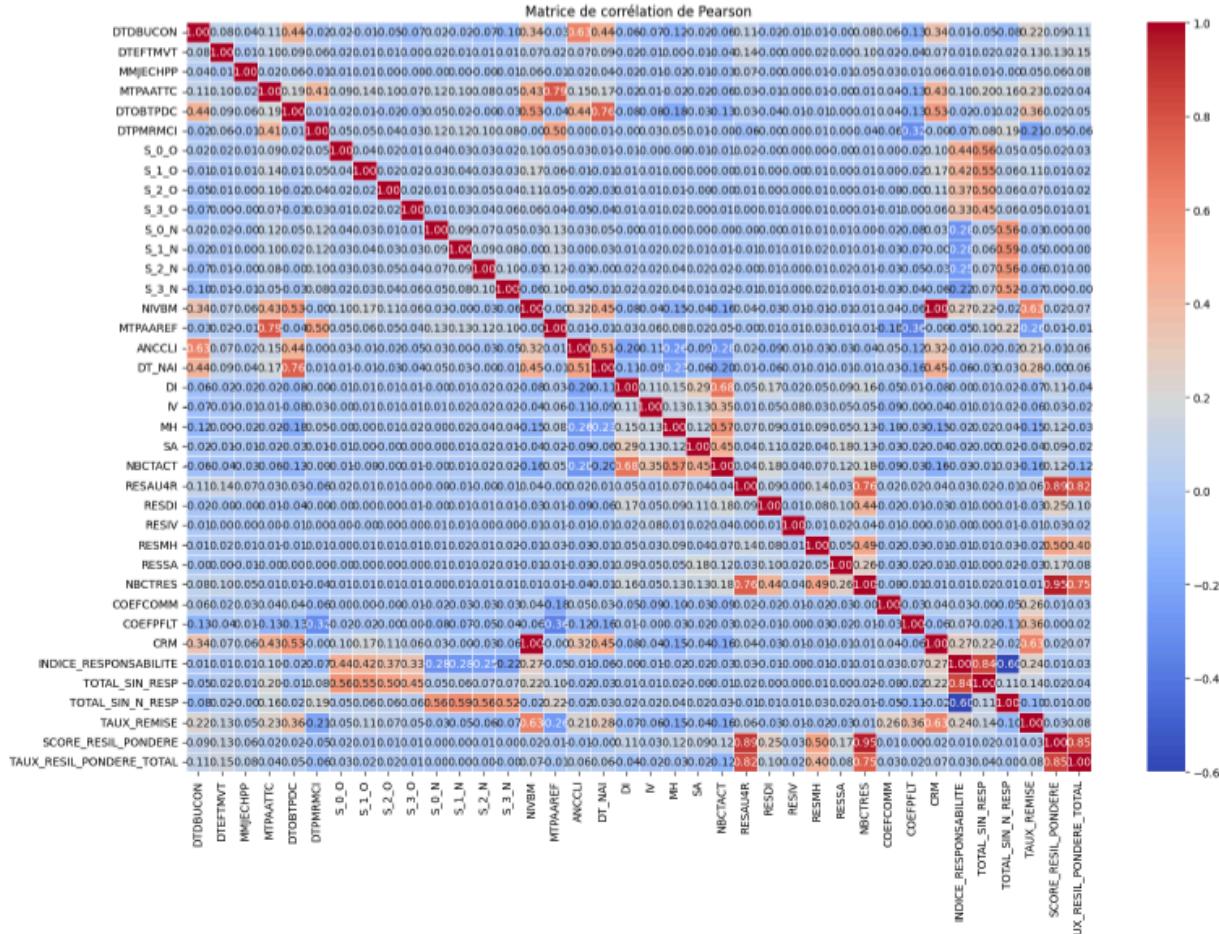


6.3.2 Variables quantitatives (corrélation de Pearson)

Nous traçons aussi la matrice de corrélation des variables continues à l'aide du coefficient de Pearson. L'objectif est d'identifier les redondances. Si deux variables décrivent pratiquement la même chose, nous n'en conservons qu'une seule. Nous créons par ailleurs des indicateurs synthétiques pour réduire encore le nombre de variables. Par exemple, toutes les informations sur les sinistres (responsables et non responsables) sont regroupées dans un unique INDICE_RESPONSABILITE.

La forte corrélation observée entre la prime actuelle (MTPAATTC) et la prime de référence (MTPAAREF) est capturée par le TAUX_REMISE. Nous supprimons donc les deux premières variables, trop corrélées. Pour les modèles non linéaires issus du boosting ou du bagging, la multicolinéarité pose moins de souci, mais nous analyserons tout de même l'importance des variables afin de savoir lesquelles expliquent le mieux la résiliation et distinguent les contrats actifs des résiliés.

Enfin, plusieurs variables restent mal documentées et nous ne connaissons pas toujours leur mode de calcul. Il se peut donc que des informations utiles nous échappent, ce qui pourrait biaiser les résultats et les performances des modèles à venir.



La variable CRM (coefficients de réduction-majoration) a été supprimée car elle fournit la même information que NIVBVM (niveau de bonus-malus), qui a été conservée. Pour éviter la redondance et la multicolinéarité, une seule des deux est gardée.

La variable DT_NAI (date de naissance) a été supprimée car elle est très corrélée à DTOBTPDC (date d'obtention du permis). Cette dernière est jugée plus pertinente en assurance auto car elle informe à la fois sur l'âge et l'ancienneté de conduite.

7. Quels modèles ont du potentiel ?

En pratique, le choix des modèles constitue l'une des étapes les plus délicates. Elle requiert non seulement une expertise technique, mais aussi une bonne compréhension du contexte métier. Déterminer quel algorithme utiliser de manière éclairée est donc un vrai défi. Toutefois, avec l'évolution des bibliothèques disponibles en Python, il est aujourd'hui possible d'évaluer rapidement plusieurs modèles candidats selon une métrique de performance choisie.

Dans cette optique, nous avons utilisé la bibliothèque LazyPredict, qui permet d'obtenir une comparaison rapide et automatisée de différents modèles. Cet outil est très utile pour identifier les algorithmes les plus prometteurs dès les premières étapes, sans avoir à construire manuellement chaque pipeline.

Un autre facteur important dans le choix d'un modèle est le temps d'exécution. Cet aspect est souvent mis de côté dans un cadre académique, mais il est crucial en entreprise. En effet, les entreprises préfèrent souvent un modèle rapide, stable et facile à déployer, quitte à sacrifier légèrement la performance pure. Le modèle final retenu doit donc trouver un équilibre entre performance et efficacité opérationnelle.

Une fois les modèles candidats identifiés via LazyPredict, nous approfondissons leur optimisation pour améliorer les résultats, en testant notamment plusieurs stratégies de traitement des données et de réglage des hyperparamètres.

Attention à l'interprétation des résultats de LazyPredict

Il est important de noter que LazyPredict utilise un prétraitement des données très simplifié. Le but est uniquement de permettre l'exécution des modèles sans crash, sans viser une réelle optimisation.

En pratique :

- Les variables numériques sont traitées via deux étapes :
SimpleImputer(strategy="mean") pour remplacer les valeurs manquantes par la moyenne, puis StandardScaler pour normaliser les données.
- Les variables catégorielles à faible cardinalité (moins de 11 modalités distinctes) sont prétraitées avec :
SimpleImputer(strategy="constant", fill_value="missing") suivi de OneHotEncoder.
- Les variables catégorielles à haute cardinalité (11 modalités ou plus) sont traitées avec :
SimpleImputer(strategy="constant", fill_value="missing") puis OrdinalEncoder.

Cependant, ce traitement présente une limite importante : aucune distinction n'est faite entre les variables nominales (sans ordre) et ordinaires (avec ordre). Cela peut entraîner une perte de pertinence dans la représentation des données et impacter les performances des modèles.

Pour toutes ces raisons, les résultats de LazyPredict doivent être vus comme une première base de comparaison. Ils permettent de repérer les modèles prometteurs, avant de les réentraîner avec un prétraitement manuel mieux adapté, ce qui permet souvent de gagner en précision et en robustesse.

Lien vers la bibliothèque LazyPredict : <https://github.com/shankarpandala/lazypredict>

Remarque :

Le temps de calcul nécessaire pour tester l'ensemble des modèles de classification (environ 30 algorithmes) sur la totalité de notre base, soit près de 90 000 contrats auto, serait particulièrement long.

Afin d'accélérer cette première phase d'exploration, nous avons donc choisi de sélectionner aléatoirement un échantillon de 10 000 lignes. Cela nous permet d'obtenir un aperçu rapide et représentatif des modèles les plus adaptés à notre problématique, tout en réduisant le temps de traitement.

Voici les premiers résultats obtenus après avoir mis les données au bon format afin de permettre l'exécution du code :

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	recall_score	Time Taken
<code>RandomForestClassifier</code>	0.99	0.99	0.99	0.99	0.99	0.67
<code>AdaBoostClassifier</code>	0.99	0.98	0.98	0.99	0.98	0.82
<code>LGBMClassifier</code>	0.99	0.98	0.98	0.99	0.97	0.27
<code>BaggingClassifier</code>	0.99	0.98	0.98	0.99	0.97	0.23
<code>XGBClassifier</code>	0.99	0.98	0.98	0.99	0.96	0.23
<code>BernoulliNB</code>	0.93	0.96	0.96	0.94	1.00	0.03
<code>DecisionTreeClassifier</code>	0.98	0.95	0.95	0.98	0.91	0.06
<code>ExtraTreesClassifier</code>	0.98	0.94	0.94	0.98	0.88	0.52
<code>LinearSVC</code>	0.97	0.94	0.94	0.97	0.89	0.32
<code>SGDClassifier</code>	0.97	0.94	0.94	0.97	0.89	0.10
<code>SVC</code>	0.97	0.93	0.93	0.97	0.89	0.73
<code>Perceptron</code>	0.95	0.93	0.93	0.96	0.90	0.03
<code>CalibratedClassifierCV</code>	0.97	0.93	0.93	0.97	0.88	0.99
<code>LogisticRegression</code>	0.97	0.93	0.93	0.97	0.87	0.40
<code>NearestCentroid</code>	0.88	0.92	0.92	0.89	0.97	0.05
<code>GaussianNB</code>	0.85	0.91	0.91	0.87	0.99	0.03
<code>LinearDiscriminantAnalysis</code>	0.94	0.91	0.91	0.94	0.87	0.06
<code>QuadraticDiscriminantAnalysis</code>	0.90	0.89	0.89	0.91	0.88	0.03
<code>PassiveAggressiveClassifier</code>	0.94	0.87	0.87	0.94	0.77	0.04
<code>ExtraTreeClassifier</code>	0.94	0.87	0.87	0.94	0.77	0.03
<code>KNeighborsClassifier</code>	0.93	0.83	0.83	0.93	0.71	0.17
<code>LabelSpreading</code>	0.93	0.82	0.82	0.93	0.69	7.23
<code>LabelPropagation</code>	0.93	0.82	0.82	0.93	0.69	6.50
<code>RidgeClassifier</code>	0.91	0.73	0.73	0.90	0.50	0.03
<code>RidgeClassifierCV</code>	0.91	0.73	0.73	0.90	0.50	0.06
<code>DummyClassifier</code>	0.88	0.50	0.50	0.83	0.00	0.02

Les résultats obtenus sont très encourageants. Il est toutefois important de rappeler que cette première évaluation a été réalisée sans traitement des valeurs manquantes ni des valeurs aberrantes, et que toutes les variables ont été utilisées sans sélection ni encodage spécifique. Le risque de surapprentissage est donc élevé à ce stade. La gestion fine des encodages et du nettoyage des données constituera un enjeu majeur pour les prochaines étapes.

Cette analyse rapide nous permet néanmoins d'avoir un premier aperçu du comportement des modèles sur notre jeu de données. Rappelons également que seuls 10 000 contrats ont été utilisés pour cette phase de test, afin de limiter le temps de calcul.

Dès à présent, on constate que les méthodes de type Bagging et Boosting (comme Random Forest, XGBoost) montrent de très bonnes performances. Nous déciderons donc de les approfondir lors des étapes suivantes. Par ailleurs, nous conserverons la régression logistique comme modèle de référence, car elle reste une méthode simple, rapide, et surtout largement acceptée dans de nombreuses institutions pour sa lisibilité et son interprétabilité.

8. Data Preprocessing

8.1 Gestion des Valeurs Aberrantes

	Q1	Q3	IQR	borne_min	borne_max	nb_outliers	%_outliers	%_negatives
DTDBUCON	10766.00	13970.00	3204.00	5960.00	18776.00	2197.00	3.04	0.26
DTEFTMVT	14693.00	14807.00	114.00	14522.00	14978.00	3488.00	4.83	0.00
MMJECHPP	316.00	912.00	596.00	-578.00	1806.00	0.00	0.00	0.00
MTPAATTC	1772.16	3852.92	2080.76	-1348.98	6974.06	2933.00	4.06	0.00
DTOBTPDC	2524.00	10279.00	7755.00	-9108.50	21911.50	147.00	0.20	12.61
DTPMRMCI	10196.00	13057.00	2861.00	5904.50	17348.50	1097.00	1.54	0.22

Une étude a été menée sur chaque variable quantitative afin de détecter d'éventuelles valeurs aberrantes susceptibles d'affecter la robustesse et la stabilité des modèles. L'objectif est de repérer des erreurs de saisie ou des valeurs anormalement élevées ou faibles, notamment des valeurs négatives non cohérentes, comme pour les montants en unités monétaires (ex. : primes d'assurance).

Pour certaines variables calculées à partir de dates (ex. : nombre de jours depuis un événement), il est normal d'obtenir des valeurs négatives si la date de référence est postérieure à la date d'observation. Ces cas ne sont donc pas considérés comme aberrants par défaut.

Afin de détecter statistiquement les valeurs extrêmes, nous avons utilisé la méthode de Tukey, basée sur l'écart interquartile (IQR). Elle permet de définir des bornes minimales et maximales à partir desquelles une valeur est considérée comme hors norme.

Malgré cela, un nombre important de valeurs détectées comme aberrantes est présent dans notre base. Toutefois, à l'échelle individuelle, beaucoup de ces valeurs sont plausibles dans le contexte métier. Par exemple, des montants de primes élevés ou des réductions très importantes peuvent parfaitement exister dans certains cas spécifiques. Les supprimer reviendrait à éloigner notre jeu de données de la réalité terrain.

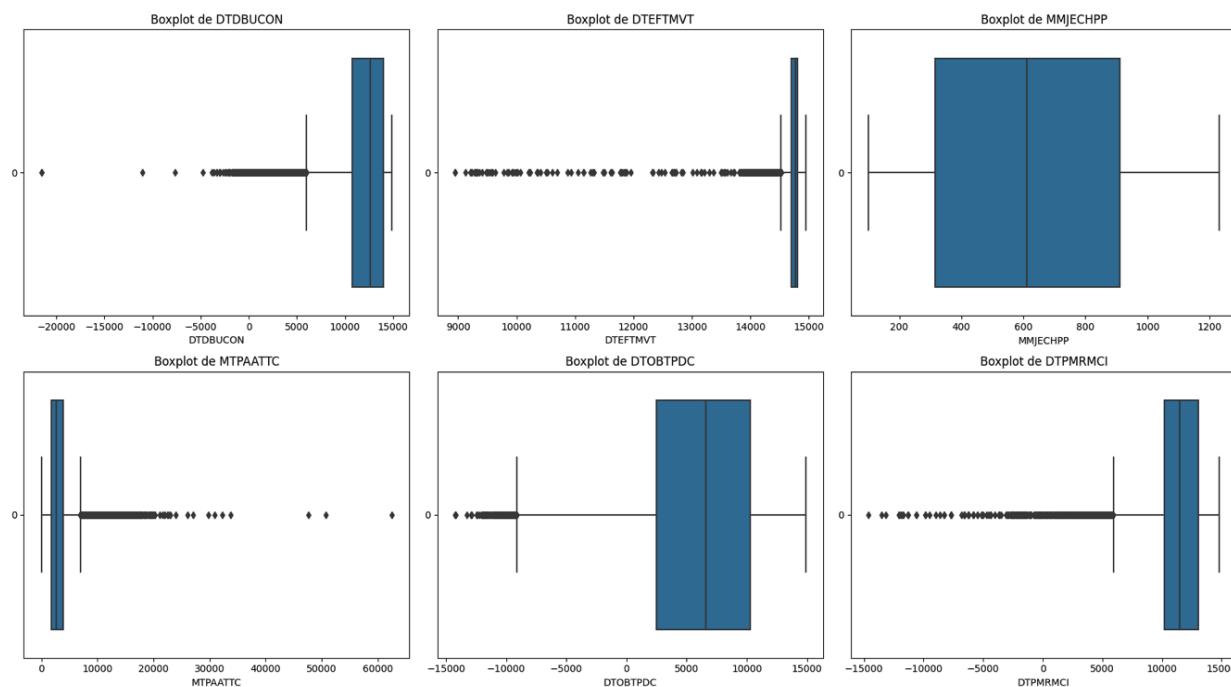
D'ailleurs, la notion même de « valeur aberrante » soulève plusieurs interrogations :

- S'agit-il d'une erreur ou d'un cas extrême mais légitime ?
- Faut-il traiter chaque variable isolément ou adopter une approche globale ?

Supprimer ces observations une à une pourrait entraîner une perte d'information importante et réduire significativement la taille du jeu de données.

Une alternative pertinente consiste à utiliser des techniques de normalisation robuste, comme le RobustScaler, qui limite l'impact des valeurs extrêmes. Cela est particulièrement utile pour les modèles sensibles aux outliers, comme la régression logistique. En revanche, certains modèles comme les algorithmes de bagging (ex. : Random Forest) sont naturellement plus robustes face aux valeurs extrêmes.

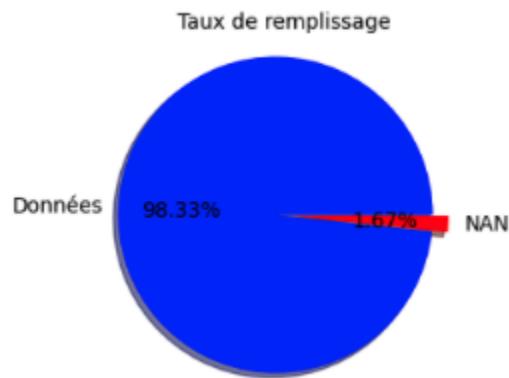
Enfin, il est aussi possible d'intégrer la gestion des valeurs aberrantes dans une démarche de discréétisation des variables continues. Cela peut renforcer la stabilité des modèles, faciliter leur interprétation (notamment pour la régression logistique) et aider à rendre les variables plus explicites.



8.2 Gestion des Valeurs Manquantes

Le dataset a été divisé en deux parties : 80 % pour l'entraînement et 20 % pour le test, dès l'étape de traitement des valeurs aberrantes. Cette séparation anticipée permet de calculer les statistiques (comme les bornes de l'IQR ou les valeurs de normalisation) uniquement sur l'échantillon d'entraînement, puis de les appliquer de manière cohérente à l'échantillon de test. La même logique est utilisée pour le RobustScaler, afin d'éviter toute fuite d'information entre les deux ensembles.

Le dataset contient 1.67% de valeurs manquantes



Compte tenu du faible nombre de valeurs manquantes, une suppression directe pourrait être envisagée. Toutefois, nous avons privilégié des techniques d'imputation pour conserver un maximum d'observations.

- Concernant la variable INDICE_RESPONSABILITE, les valeurs manquantes correspondent en réalité à des situations sans sinistres, donc équivalentes à des zéros. Elles seront traitées comme telles.
- Pour les variables numériques, les valeurs manquantes seront remplacées par la médiane, calculée sur l'échantillon d'entraînement. Cette méthode est plus robuste que la moyenne en présence de valeurs extrêmes.
- Pour les variables catégorielles, nous utiliserons le mode (la modalité la plus fréquente) comme stratégie d'imputation.

9. Modélisation

Dans cette section, nous allons tester plusieurs versions de modèles de classification afin de modéliser la probabilité de résiliation d'un contrat auto, en nous basant sur les caractéristiques sélectionnées parmi l'ensemble des données disponibles. L'objectif est d'identifier les modèles les plus performants pour prédire ce phénomène, en tenant compte à la fois des performances globales et des spécificités du problème.

Pour évaluer ces modèles, nous utiliserons plusieurs métriques de performance. Cela nous permettra d'avoir une vision plus complète de leurs résultats, en allant au-delà d'une simple mesure de précision. Afin de garantir la pertinence des évaluations sur des données nouvelles, un jeu de test fixe a été défini. Il représente 20 % du jeu de données et est strictement exclu de toute phase d'entraînement, de validation ou d'optimisation des modèles. Ce choix permet de conserver un jeu de test non biaisé, identique pour tous les modèles comparés.

Parmi les principales métriques utilisées dans ce projet, nous retrouvons notamment :

- Le score AUC (Area Under the Curve) : il mesure la capacité du modèle à discriminer entre les classes. L'AUC correspond à la surface sous la courbe ROC, qui trace le taux de vrais positifs contre le taux de faux positifs. Plus l'AUC est proche de 1, meilleure est la capacité de discrimination du modèle. À l'inverse, un score proche de 0,5 indique que le modèle ne fait pas mieux qu'un tirage aléatoire.
- Le F1-score : il s'agit de la moyenne harmonique entre la précision et le rappel. Cette métrique est particulièrement utile dans les cas de déséquilibre de classes, comme dans notre problématique, où la classe "résilié" est moins fréquente. Nous nous intéressons à deux variantes :
 - Le F1 macro-averaged, qui donne un poids égal à chaque classe, quelle que soit leur fréquence.
 - Le F1-score de la classe 1 (résilié), qui mesure spécifiquement la performance sur la classe positive, souvent la plus importante dans des contextes métiers où l'on cherche à détecter un comportement particulier.

Ces métriques nous guideront dans le choix du modèle final, en tenant compte à la fois de la qualité globale des prédictions et de la capacité à détecter efficacement les cas de résiliation.

9.1 Encodage et Normalisation

Avant d'entraîner les modèles de classification, un travail rigoureux de préparation des données a été effectué. Cette phase est essentielle pour garantir des résultats fiables, comparables et exploitables.

Les variables catégorielles ont été traitées différemment selon leur nombre de modalités :

- Celles avec 5 modalités ou moins ont été encodées via One-Hot Encoding, méthode qui crée une colonne par modalité (en supprimant une modalité qui sera la référence). Les colonnes concernées sont : CDMCE, CD_SEX et CLIACTIF.
- Les variables à cardinalité plus élevée ont été encodées par Target Encoding. Cette méthode consiste à remplacer chaque modalité par la moyenne de la variable cible (CONTRAT) pour cette modalité. Elle est appliquée à des variables telles que CD_FML, CDMARVEH, NOTAREFF, PUI_TRE, RN_VL_VH.

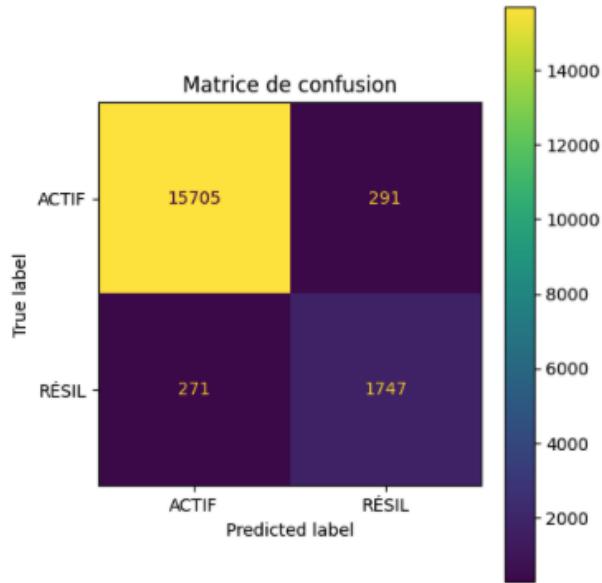
Cette approche permet d'éviter l'explosion du nombre de colonnes, tout en conservant une information utile sur la relation entre la variable et la cible.

Les variables numériques ont ensuite été normalisées à l'aide du RobustScaler. Cette méthode est robuste face aux valeurs extrêmes (outliers), contrairement à une normalisation classique comme le MinMaxScaler. Le scaler est ajusté (fit) uniquement sur les données d'entraînement, puis appliqué aux données de test afin de prévenir toute fuite de données.

9.2 Modèle 1 : Logistic Regression

Generalized Linear Model Regression Results						
Dep. Variable:	CONTRAT	No. Observations:	72194			
Model:	GLM	Df Residuals:	72176			
Model Family:	Binomial	Df Model:	17			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-5468.9			
Date:	Fri, 18 Jul 2025	Deviance:	10938.			
Time:	22:47:19	Pearson chi2:	7.35e+04			
No. Iterations:	10	Pseudo R-squ. (CS):	0.4237			
Covariance Type:	nonrobust					
	coef	std err	z	P> z	[0.025	0.975]
const	-11.5459	0.305	-37.843	0.000	-12.144	-10.948
ANCLLI	0.2087	0.040	5.250	0.000	0.131	0.287
CDMARVEH	8.1421	1.426	5.708	0.000	5.347	10.938
COEFCOMM	0.0613	0.013	4.838	0.000	0.036	0.086
COEFPLT	0.4178	0.036	11.593	0.000	0.347	0.488
DI	0.6945	0.052	13.256	0.000	0.592	0.797
DTEFTMVT	5.9315	0.100	59.211	0.000	5.735	6.128
DTOBTPDC	-0.3764	0.051	-7.438	0.000	-0.476	-0.277
DTPMRMCI	-0.4351	0.035	-12.449	0.000	-0.504	-0.367
MH	1.1332	0.041	27.413	0.000	1.052	1.214
MMJECHPP	0.1659	0.048	3.480	0.001	0.072	0.259
NBCTACT	-0.3793	0.052	-7.349	0.000	-0.480	-0.278
NBCTRES	-0.1396	0.033	-4.190	0.000	-0.205	-0.074
NOTAREFF	5.3675	0.612	8.768	0.000	4.168	6.567
RN_VL_VH	15.4221	1.821	8.467	0.000	11.852	18.992
SA	0.5729	0.075	7.669	0.000	0.427	0.719
TAUX_RESIL_PONDRE_TOTAL	2.1877	0.037	59.487	0.000	2.116	2.260
TOTAL_SIN_RESP	0.2731	0.051	5.347	0.000	0.173	0.373

```
Évaluation du modèle :
AUC : 0.9908
Gini (2*AUC - 1) : 0.9816
F1-score macro : 0.9219
F1-score classe 1 : 0.8614
```



Pour la régression logistique une sélection stepwise a été réalisée en n'acceptant dans le modèle que les variables dont la p-value est inférieure à 0.01.

- Ce seuil a été volontairement fixé afin de ne conserver que les variables les plus robustes et statistiquement significatives (principe de parcimonie).
- Lorsqu'une variable provoque une matrice singulière (problème de multicolinéarité), elle est automatiquement exclue et ignorée pour la suite de la sélection.
- Pour contrôler la multicolinéarité :
 - Les variables trop corrélées entre elles sont naturellement repérées et écartées grâce à la gestion automatique des erreurs de type "singular matrix".
 - Une régression LASSO sera ensuite appliquée pour affiner davantage la sélection. Le LASSO permet de réduire à zéro les coefficients de certaines variables redondantes, ce qui complète efficacement la sélection par p-value.

Résumé du modèle :

- AUC proche de 0.991 : très bonne séparation entre les classes.
- Gini proche de 0.982 : très élevé, ce qui indique une discrimination presque parfaite.
- F1-score macro d'environ 0.92 : bon équilibre global entre les classes.
- F1-score classe 1 d'environ 0.86 : bonne performance sur les cas de résiliation.

La matrice de confusion montre un bon équilibre entre faux positifs et faux négatifs. Le modèle identifie correctement les clients à risque de résiliation, tout en conservant une bonne précision globale.

Les coefficients estimés sont tous significatifs (p-value très faibles), et leur signe permet d'interpréter leur influence sur la probabilité de résiliation (sans toutefois indiquer directement l'ampleur de l'effet).

Interprétation de quelques variables / coefficients :

- TAUX_RESIL_PONDERE_TOTAL (positif) : plus le score de résiliation est élevé, plus le risque de résiliation est fort.
- RN_VL_VH (positif) : les véhicules haut de gamme ou très cotés sont associés à une volatilité plus élevée.
- DI (positif) : un grand nombre de contrats divers pourrait signaler un profil instable ou opportuniste.
- MH (négatif) : avoir un contrat habitation en plus d'un contrat auto peut refléter une fidélité plus forte.
- DTEFTMVT (positif) : un mouvement récent sur le contrat est associé à un risque de résiliation plus élevé.

Nature des variables douteuses :

- DTEFTMVT : variable incertaine, car le dernier mouvement peut justement indiquer une résiliation. L'hypothèse serait que la colonne DTEFTMVT est mise à jour au moment où le passage en statut RÉSIL est enregistré. De manière générale, cette variable est concentrée sur les dates récentes (valeurs élevées par rapport à la référence), ce qui peut sembler normal puisque les contrats sont revus régulièrement (au moins semestriellement).
- Autres variables douteuses : COEFCOMM et COEFPLT. Nous ne disposons pas d'une définition claire pour ces variables, mais elles semblent néanmoins pertinentes pour le modèle.
- Interrogation également sur les variables liées aux nombres de contrats (NBCTACT, NBCTRES, ...) : on ne sait pas si le contrat auto en question est compris dans ces comptes.
- Une étude approfondie avec les équipes métier est indispensable pour analyser ces variables en détail. Sinon, il existe un risque de mauvaise construction du modèle, notamment par utilisation de variables potentiellement causales ou corrélées avec la cible.

Variables non sélectionnées par le LASSO :

Le LASSO n'a pas sélectionné certaines variables, leurs coefficients étant réduits à zéro dans le modèle final. Parmi elles, on trouve notamment :

- CD_SEX_M
- DTDBUCON
- CDMCE_T
- NIVBM
- TOTAL_SIN_N_RESP
- COEFCOMM
- ANCCLI

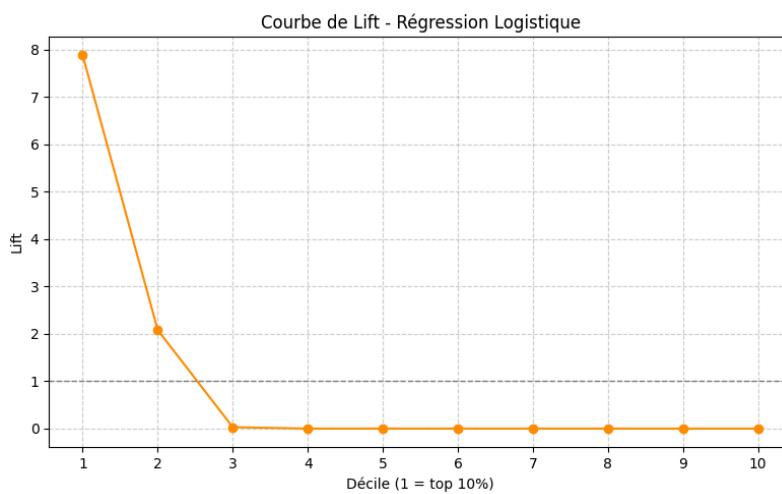
Cela signifie que, selon la logique de régularisation du LASSO, ces variables sont jugées comme redondantes ou peu contributives une fois les autres variables du modèle prises en compte.

En revanche, les autres variables ont été conservées avec des coefficients non nuls, ce qui indique qu'elles apportent une valeur ajoutée au modèle prédictif.

Mais cette méthode repose sur une contrainte mathématique (minimisation d'une fonction de coût avec pénalisation L1), et non sur la significativité statistique individuelle de chaque variable.

La sélection stepwise, quant à elle, repose sur un critère de p-value (ici < 0.01), ce qui permet de ne conserver que les variables ayant une contribution statistiquement significative à la modélisation. Cette approche est généralement plus interprétable et plus conforme aux exigences (métier), notamment dans un cadre assurantiel ou réglementaire (risque de crédit, défaut bâlois, ...).

Courbe de Lift :



	bin	count	positives	rate	lift	cumulative_positives	cumulative_rate
0	(0.635, 1.0]	1802	1592	0.88	7.89	1592	0.88
1	(0.0246, 0.635]	1801	420	0.23	2.08	2012	0.56
2	(0.00282, 0.0246]	1801	6	0.00	0.03	2018	0.37
3	(0.000826, 0.00282]	1802	0	0.00	0.00	2018	0.28
4	(0.000292, 0.000826]	1801	0	0.00	0.00	2018	0.22
5	(6.5e-05, 0.000292]	1801	0	0.00	0.00	2018	0.19
6	(1.08e-05, 6.5e-05]	1802	0	0.00	0.00	2018	0.16
7	(1.9e-06, 1.08e-05]	1801	0	0.00	0.00	2018	0.14
8	(8.11e-08, 1.9e-06]	1801	0	0.00	0.00	2018	0.12
9	(-0.001, 8.11e-08]	1802	0	0.00	0.00	2018	0.11

D'après les résultats de la courbe de lift, nous pouvons observer qu'en prenant le 1er décile (top 10 % des clients selon le modèle), le modèle identifie 88 % des clients résiliés dans ce groupe. Autrement dit, cela

donne un lift de 7,89, ce qui signifie que ce groupe est près de 8 fois plus riche en résiliés que si l'on avait sélectionné 10 % des clients au hasard.

Pour le 2e décile, le lift tombe à 2,08, ce qui reste intéressant, mais beaucoup moins puissant.

À partir du 3e décile, le lift chute presque à 0, ce qui signifie que les groupes suivants contiennent très peu, voire aucun client résilié.

Cela montre que le modèle distingue très bien les clients les plus à risque uniquement dans les premiers groupes.

9.3 Modèle 2 : Random Forest

9.3.1 Modèle général

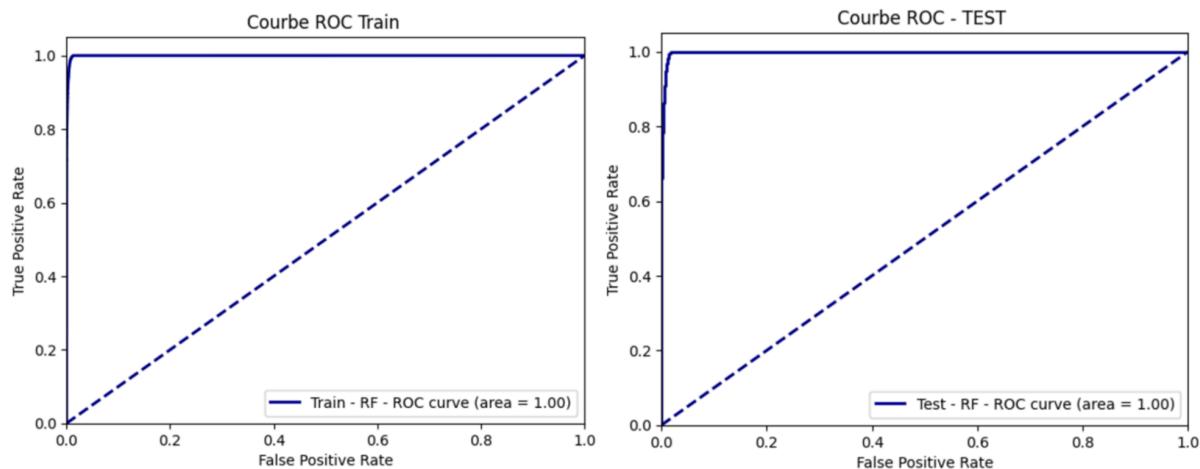
Dans cette étape, nous avons exploré une nouvelle famille de modèles d'ensemble basée sur la méthode du bagging, en particulier l'algorithme Random Forest. L'objectif était d'évaluer si cette approche, qui consiste à combiner plusieurs arbres de décision pour réduire la variance et améliorer la robustesse, permettrait d'obtenir de meilleurs résultats que les modèles précédemment testés.

Caractéristiques du modèle :

Le modèle Random Forest a été optimisé à l'aide d'une recherche de Bayes (BayesSearchCV) visant à maximiser l'AUC ROC moyen en validation croisée. Les hyperparamètres retenus pour la meilleure configuration sont les suivants : le critère de division utilisé est l'entropie, la profondeur maximale des arbres est limitée à 10, le nombre d'arbres (n_estimators) est fixé à 70, et l'échantillonnage avec remise (bootstrap) est activé. Cette configuration a permis d'atteindre un score ROC AUC moyen de 0.9987, témoignant d'une excellente capacité de discrimination du modèle sur l'ensemble des plis de validation croisée.

Principaux résultats :

Figure: Courbes ROC - Ensembles Train et Test - RF

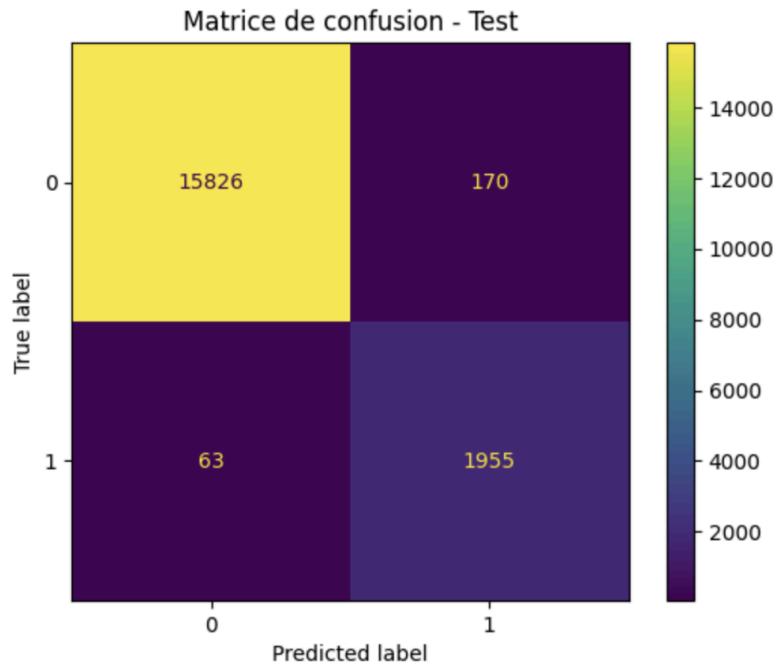


Le modèle Random Forest optimisé à l'aide d'une recherche par validation croisée a obtenu des résultats exceptionnels. Sur les courbes ROC, on observe une AUC de 1.00 à la fois sur les jeux d'entraînement et de test. Cela indique une capacité de discrimination parfaite, ce qui est très rare et peut évoquer un risque de surapprentissage. La forme de la courbe montre une montée quasi verticale suivie d'un plateau, signe que le modèle distingue très bien les classes positives et négatives.

Analyse de la matrice de confusion et des métriques de classification :

Figure: Classification report et Matrice de confusion - RF

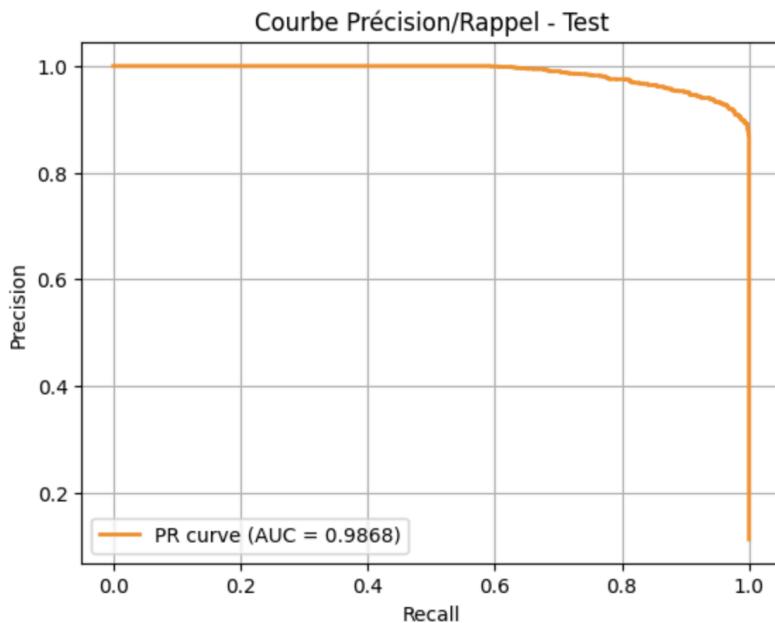
	precision	recall	f1-score	support
0	0.9960	0.9894	0.9927	15996
1	0.9200	0.9688	0.9438	2018
accuracy			0.9871	18014
macro avg	0.9580	0.9791	0.9682	18014
weighted avg	0.9875	0.9871	0.9872	18014



La matrice de confusion pour les données de test confirme la très bonne performance du modèle. Sur 2 018 cas réels de résiliation, le modèle en a correctement prédit 1 955 (vrais positifs), n'en manquant que 63 (faux négatifs). Il a par ailleurs très peu confondu des actifs avec des résiliés : seulement 170 faux positifs sur 15 996 clients actifs. Ces résultats se traduisent par un recall de 0.9688 et une precision de 0.9200 pour la classe minoritaire, avec un F1-score de 0.9438. L'accuracy globale est de 0.9871, avec une moyenne pondérée des scores de F1 très élevée également.

Courbe précision-rappel :

Figure: Courbe Précision / Rappel - RF



La courbe précision-rappel confirme cette excellente performance. L'AUC PR est de 0.9868, ce qui est extrêmement élevé. Cela montre que même en présence de déséquilibre entre les classes, le modèle conserve une capacité fiable à identifier correctement les clients résiliés, tout en maintenant une bonne précision. Ce graphique met en lumière que le modèle parvient à équilibrer très efficacement les compromis entre les faux positifs et les vrais positifs.

Importance des variables :

Tableau: Importance des variables

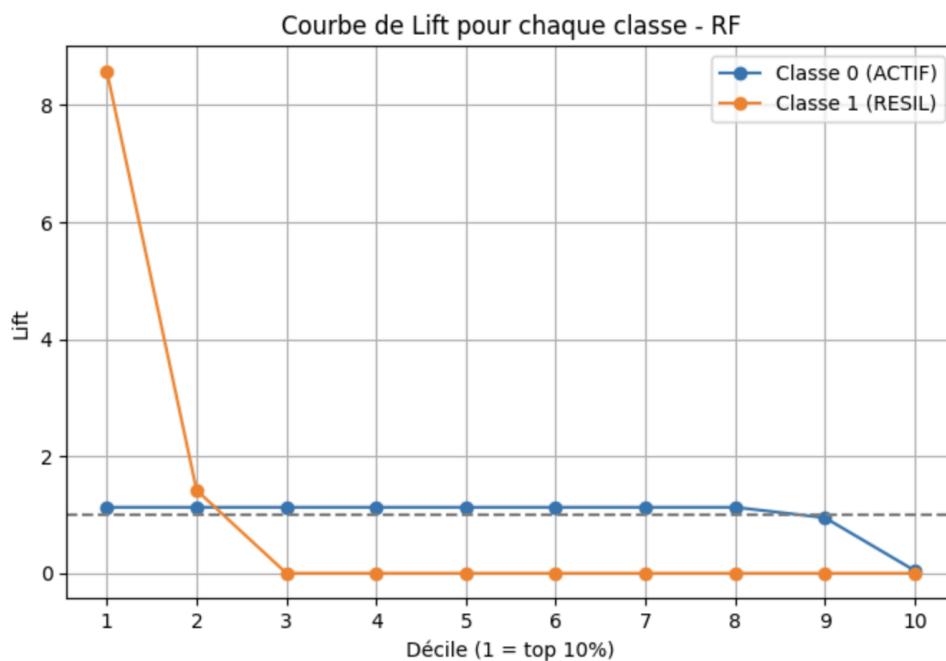
Variable	Feature Importance
DTEFTMVT	0.37
TAUX_RESIL_PONDERE_TOTAL	0.26
CLIACTIF_OUI	0.13
NBCTRES	0.12
NBCTACT	0.3

Variable	Feature Importance
MH	0.2
MMJECHPP	0.1
TAUX_REMISE	0.1
COEFPFLT	0.1
DTBUCON	0.1

En termes d'explicabilité, les variables les plus influentes dans la prédiction sont DTEFTMVT (date du dernier mouvement du contrat), TAUX_RESIL_PONDERE_TOTAL, CLIACТИF_OUI, NBCTRES, et NBCTACT. La variable DTEFTMVT à elle seule représente 37% de l'importance relative attribuée par l'algorithme, ce qui suggère qu'un contrat qui n'a pas été actif récemment est fortement associé à une probabilité plus élevée de résiliation. Ce résultat semble cohérent avec une logique métier : l'inactivité est un indicateur fort de désengagement client.

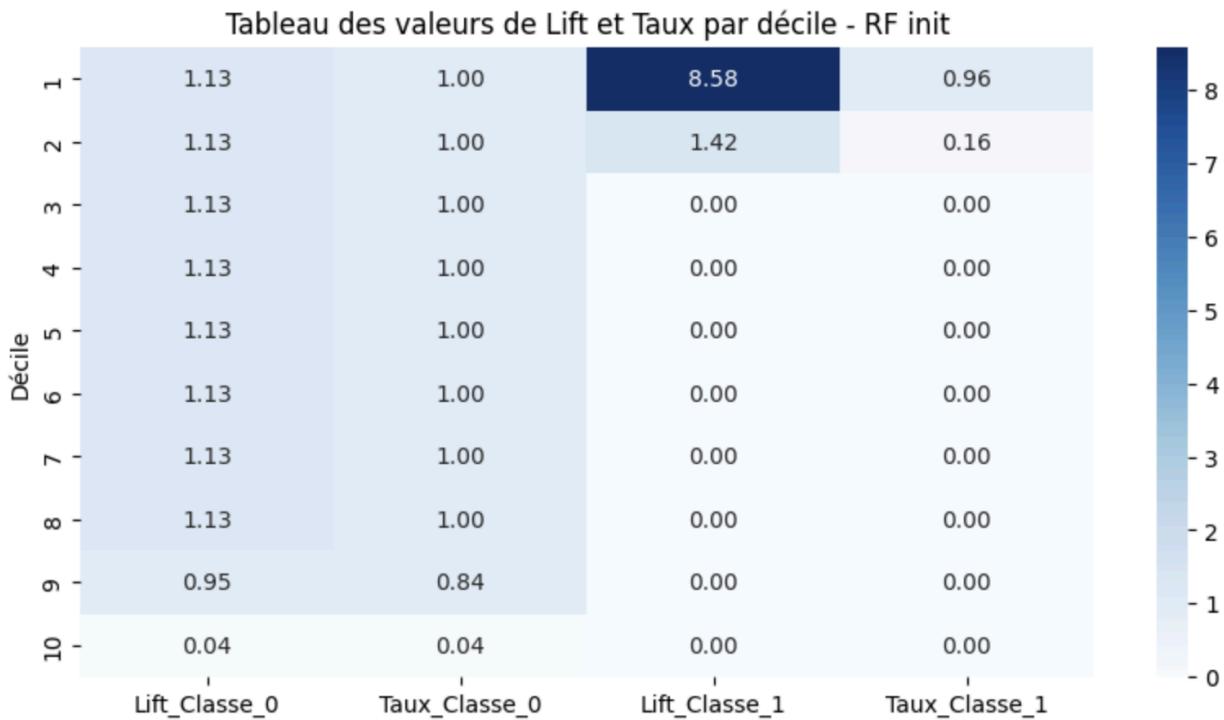
Analyse de la courbe de Lift :

Figure: Courbe LIFT - RF initial



La courbe de lift révèle une performance discriminante particulièrement marquée pour la classe 1 (résiliés). On constate que le premier décile (les 10% des individus ayant la probabilité prédictive la plus élevée) contient environ 8,5 fois plus de résiliés que la moyenne de l'échantillon. Cela suggère que le modèle est extrêmement efficace pour concentrer les vrais résiliés dans les tout premiers segments. Les déciles suivants voient ce lift chuter rapidement, ce qui est typique d'un bon modèle discriminant.

Tableau: Valeurs LIFT et Taux par décile



Ce tableau confirme les résultats observés sur la courbe de lift : le premier décile concentre à lui seul un lift de 8,58 pour la classe 1 (résiliés), soit une surreprésentation très marquée par rapport à la moyenne. Dès le troisième décile, le lift pour cette classe chute à zéro, ce qui illustre parfaitement la forte capacité du modèle à identifier les individus les plus à risque de résiliation dans les tout premiers segments.

Dans l'ensemble, ces résultats montrent que le modèle Random Forest, utilisant l'ensemble des variables, offre une performance de prédiction remarquable, avec un AUC de 1.00 et d'excellents scores de précision et de rappel. Toutefois, une telle perfection apparente, combinée à une très forte concentration du lift dans les premiers déciles, invite à la prudence et suggère un risque potentiel de surapprentissage (overfitting). Afin d'améliorer la robustesse du modèle et de favoriser une meilleure généralisation, une sélection de variables sera entreprise pour ne conserver que les plus pertinentes dans la construction du modèle.

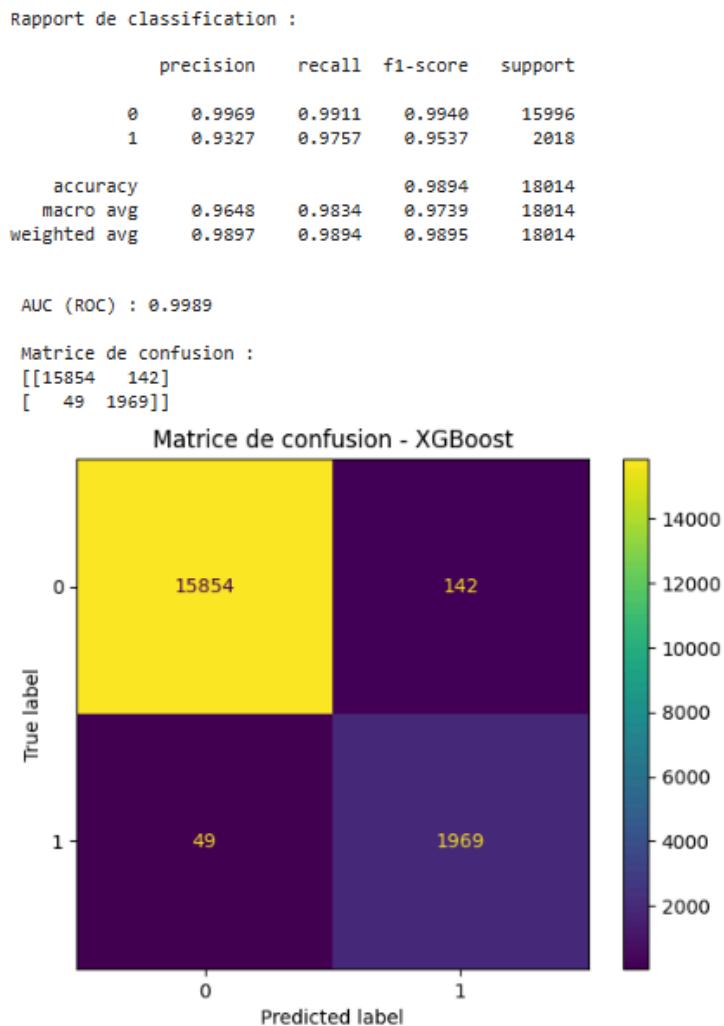
9.4 Modèle 3 : XGBoost

L'un des plus performants s'est avéré être XGBoost, un algorithme de boosting d'arbres de décision réputé pour son efficacité et sa robustesse sur des données structurées. Ce modèle a été entraîné pour prédire la probabilité de résiliation d'un contrat.

Optimisation des hyperparamètres :

Une optimisation bayésienne des hyperparamètres a été menée avec Optuna, sur 200 itérations. Cette approche n'a pas permis de dépasser les performances obtenues avec les paramètres standards du modèle. Ainsi, nous avons conservé une configuration par défaut, qui offrait déjà une qualité de prédiction remarquable.

Performances du modèle :



À l'issue de l'entraînement, le modèle a été évalué sur un jeu de test de 18 014 individus. Les performances observées sont les suivantes :

- Précision globale : 98,9 %
- Score F1 (classe résiliée) : 0,9537
- Score F1 (classe active) : 0,9911
- AUC (ROC) : 0,9989, soit une quasi-parfaite capacité à discriminer les deux classes.

Ces résultats indiquent une excellente qualité de classification, avec un équilibre très satisfaisant entre précision et rappel pour les deux classes.

L'analyse de la matrice de confusion renforce ce constat :

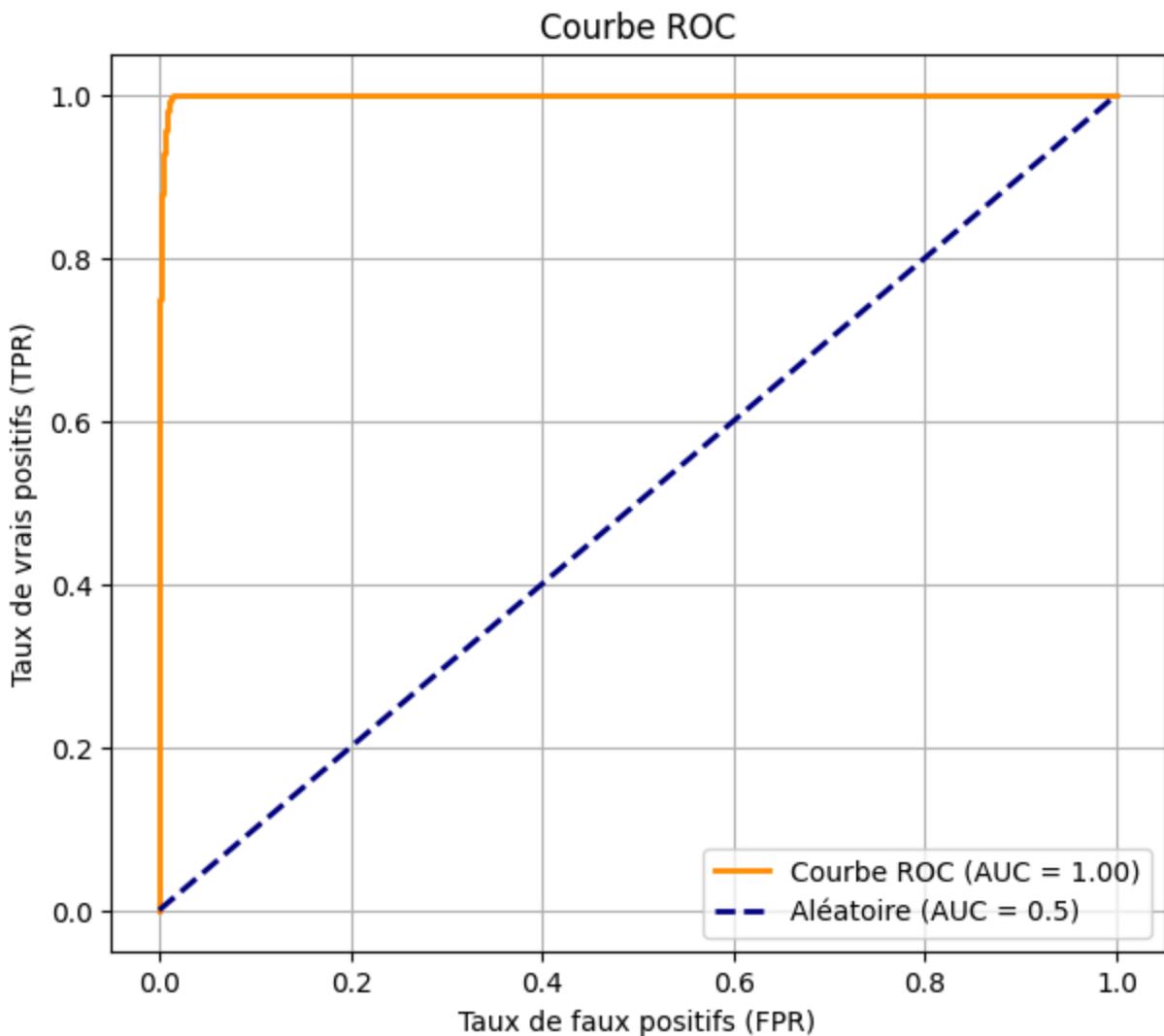
- 1 969 clients résiliés correctement identifiés sur 2 018.
- Seulement 49 résiliés classés à tort comme actifs (faux négatifs).
- 15 854 clients actifs correctement prédis sur 15 996.
- Seulement 142 actifs confondus avec des résiliés (faux positifs).

Le modèle permet donc de détecter efficacement les clients à risque tout en conservant une faible marge d'erreur, ce qui en fait un outil fiable pour la prise de décision.

Courbe ROC :

La courbe ROC (Receiver Operating Characteristic) permet d'évaluer la capacité de classement du modèle entre les deux classes : clients actifs (0) et clients résiliés (1). La courbe obtenue est très proche de l'axe supérieur gauche du graphique, ce qui traduit une très bonne performance de classification.

L'aire sous la courbe (AUC) est de 1.00, ce qui signifie que le modèle est quasiment parfait dans sa capacité à distinguer les deux classes. En comparaison, une prédiction aléatoire aurait conduit à une courbe diagonale avec une AUC de 0.5.

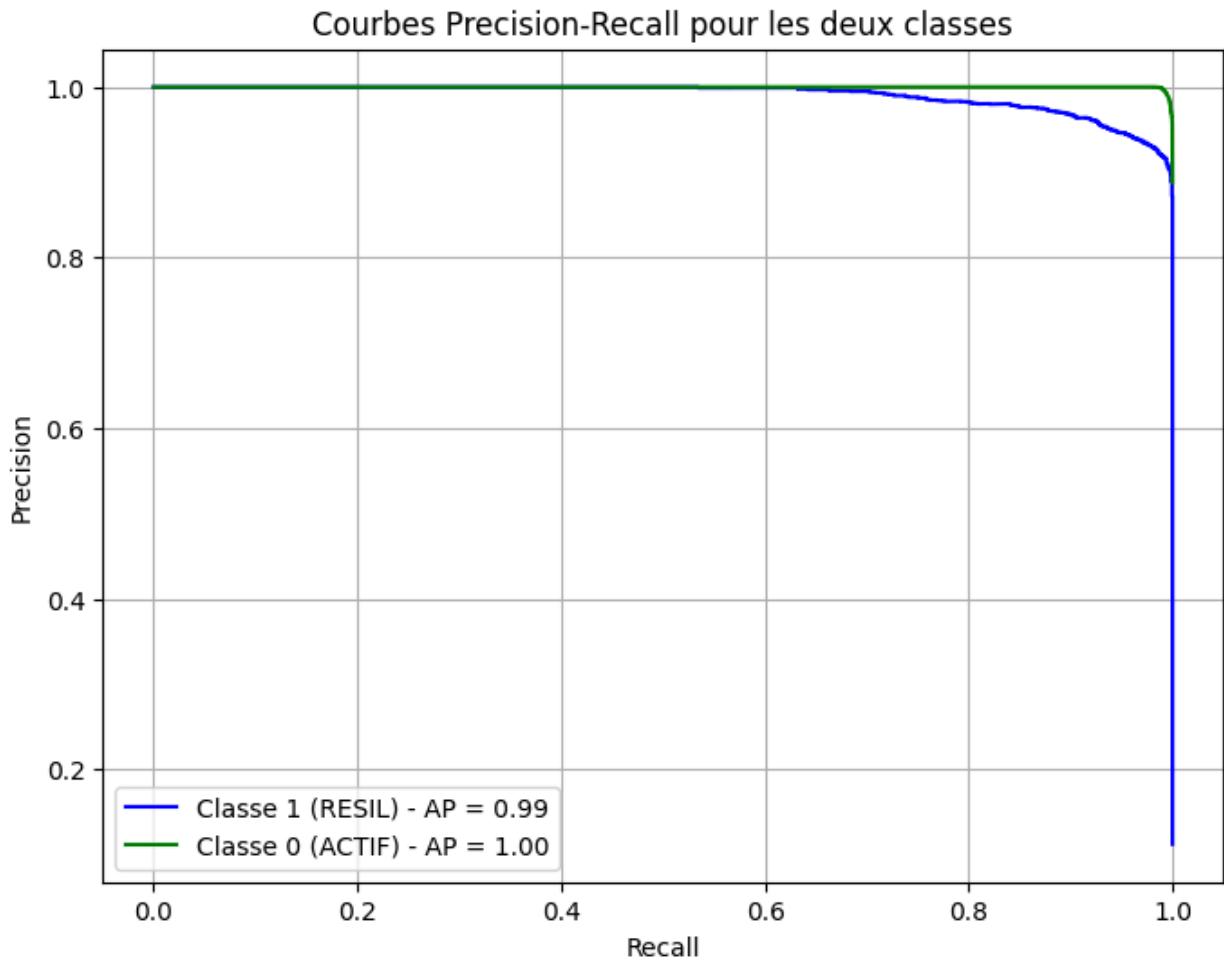


Analyse des courbes Precision-Recall :

Pour évaluer plus finement la performance du modèle sur chaque classe, les courbes Precision-Recall ont été tracées pour les deux classes (résiliés et actifs). Ce type de courbe est particulièrement pertinent dans le cadre d'un problème déséquilibré, comme c'est le cas ici, où la classe 1 (résiliation) est minoritaire.

- Classe 0 (clients actifs) : le modèle atteint une aire sous la courbe (AP) très élevée, égale à 1.00, traduisant une capacité quasi-parfaite à prédire correctement les clients qui ne résilient pas.
- Classe 1 (clients résiliés) : la précision se maintient au-dessus de 95 % jusqu'à un rappel de 0.9, avec une aire moyenne sous la courbe (AP) de 0.99. Le modèle est donc capable d'identifier avec une grande fiabilité les clients susceptibles de résilier, tout en limitant les faux positifs.

Le fait que les deux courbes soient situées très proches de la valeur optimale (précision = 1 pour rappel = 1) témoigne d'un excellent compromis entre sensibilité (recall) et précision, et confirme la robustesse du modèle, en particulier sur la classe minoritaire qui est souvent plus difficile à détecter correctement.



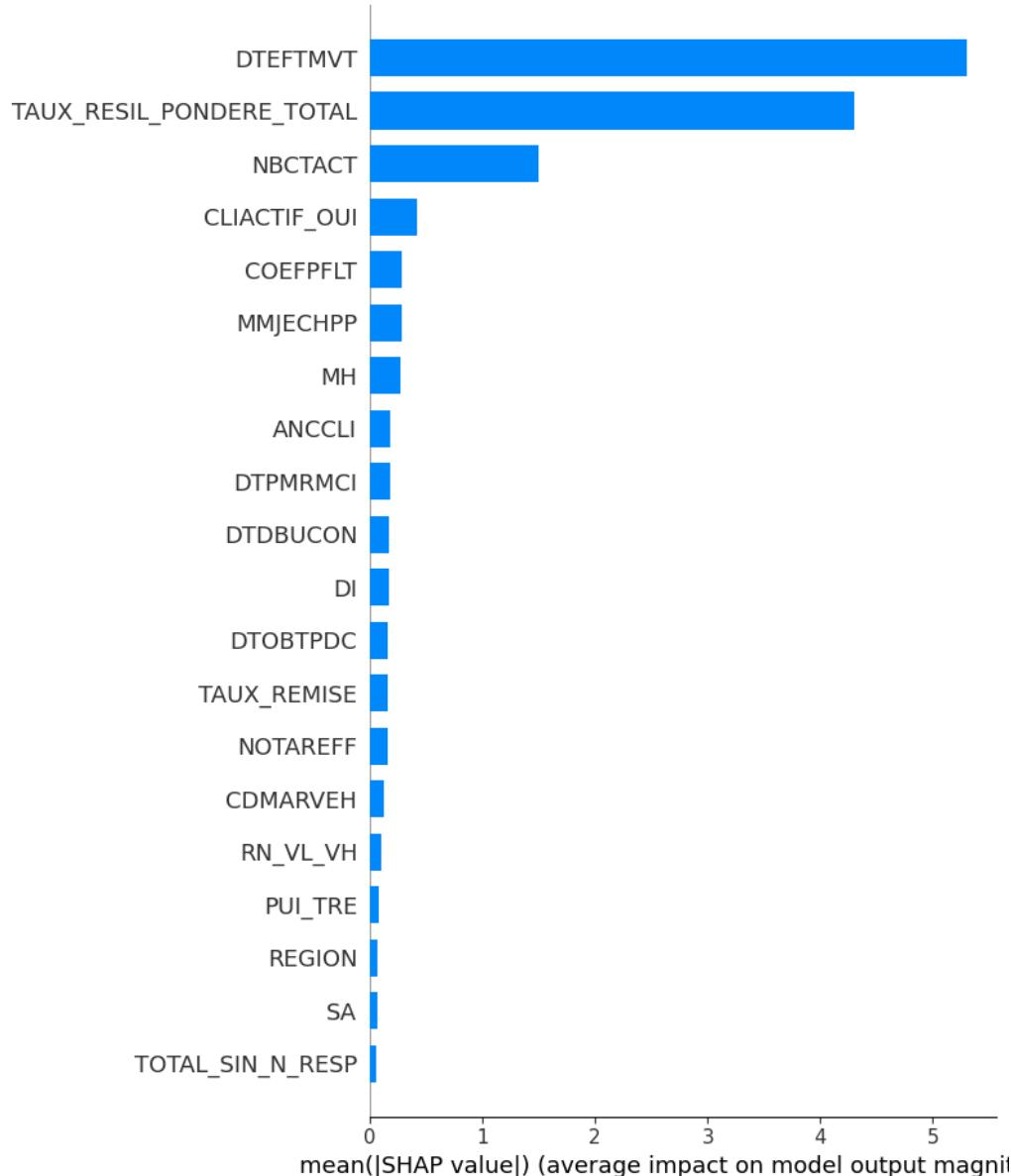
Importance globale des variables :

Pour mieux comprendre le comportement du modèle XGBoost, nous avons eu recours à l'interprétabilité via les valeurs SHAP (SHapley Additive exPlanations). Ces valeurs permettent d'identifier l'impact de chaque variable sur les prédictions du modèle, en tenant compte de leurs interactions.

Le graphique ci-dessous présente l'importance moyenne absolue des variables. Il en ressort que :

- DTEFTMVT (durée écoulée depuis le dernier mouvement) est la variable la plus influente, avec un impact moyen supérieur à 5.
- TAUX_RESIL_PONDERE_TOTAL (taux de résiliation pondéré) vient en deuxième position avec une importance moyenne avoisinant 4,5.
- NBCTACT (nombre de contrats actifs) complète le podium avec un impact autour de 1,5.

Les autres variables comme CLIACTIF_OUI, COEFPFLT, ou encore MMJECHPP ont également une contribution, bien que plus modérée



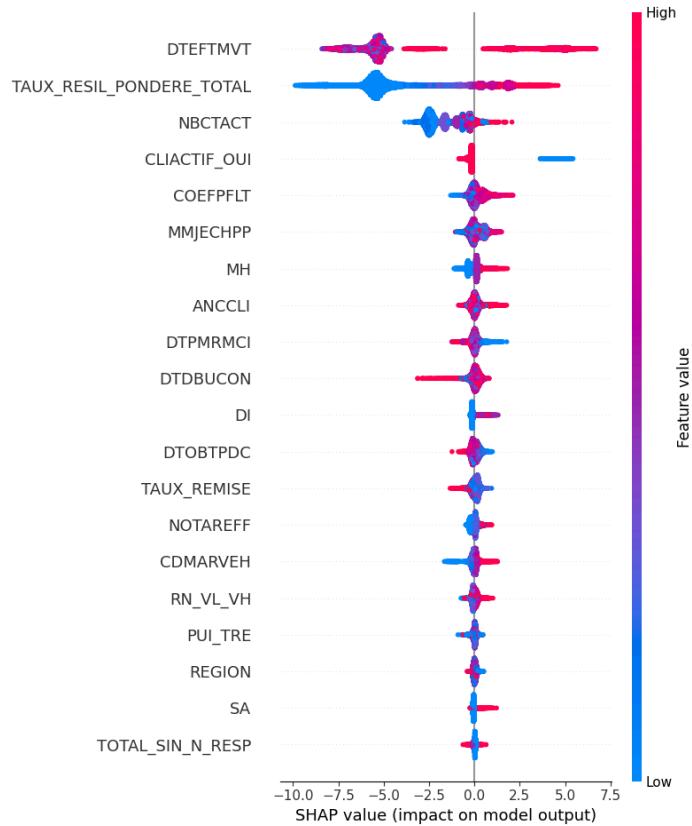
Analyse locale : effet directionnel des variables

En complément, le graphique SHAP en dispersion permet de visualiser comment les valeurs de chaque variable influencent la probabilité de résiliation (classe 1).

Quelques observations majeures :

- Pour DTEFTMVT, des valeurs élevées (en rose) augmentent fortement la probabilité de résiliation, ce qui est intuitif : plus le temps passe sans mouvement, plus le risque de résiliation croît.

- À l'inverse, un TAUX_RESIL_PONDRE_TOTAL élevé tend également à favoriser la classe 1, soulignant que les clients ayant un historique de résiliation important sont plus susceptibles de résilier à nouveau.
- Pour NBCTACT, les valeurs faibles (en bleu) sont associées à une plus forte probabilité de résiliation, suggérant que les clients avec peu de contrats actifs sont plus instables.



Analyse individuelle : cascade SHAP d'un client

Afin d'approfondir l'interprétabilité du modèle à l'échelle individuelle, une analyse locale a été menée à l'aide d'un graphe en cascade (SHAP waterfall plot), centré sur un exemple de client spécifique.

Ce graphique décompose la prédiction du modèle en montrant comment chaque variable contribue à l'écart entre la valeur de base (valeur moyenne du modèle sur l'ensemble des données) et la prédiction finale pour cet individu.

Quelques éléments notables :

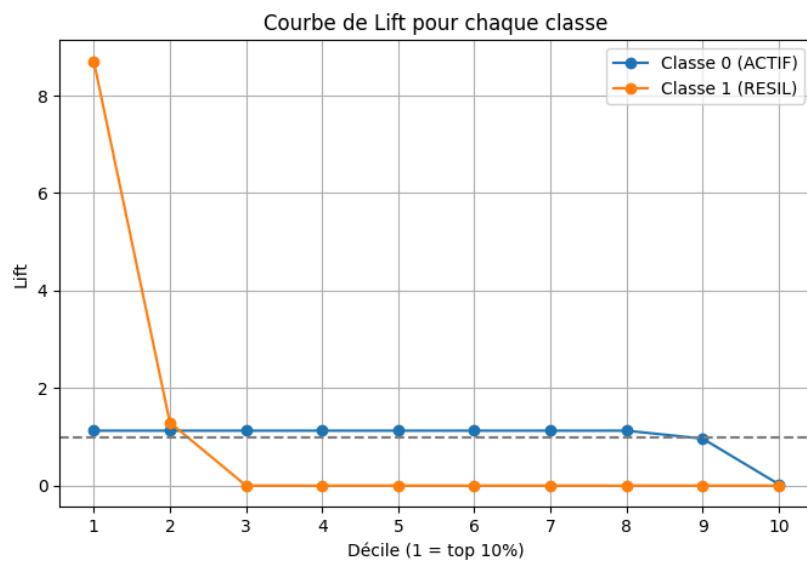
- La variable DTEFTMVT, avec une valeur de 0.237, réduit fortement la probabilité de résiliation, avec un effet négatif de -6.82 sur la sortie du modèle. Cela suggère qu'un faible temps écoulé

depuis le dernier mouvement du client est fortement rassurant pour le modèle.

- À l'inverse, TAUX_RESIL_PONDERE_TOTAL, avec une valeur de 2.333, contribue positivement à la prédiction (valeur SHAP : +0.98), indiquant que le modèle considère ce client à risque car il présente un historique élevé de résiliation.
- D'autres variables comme NBCTACT, NOTAREFF, ou DTD_BUCON ont également un effet modérément négatif (entre -0.2 et -0.5), renforçant la confiance du modèle dans la non-résiliation pour ce client.
- À l'inverse, des variables comme PUI_TRE (+0.20) ou COEFPFLT (+0.19) participent à modérer cet effet en apportant une légère incertitude.

Au final, la prédiction du modèle pour ce client reste fortement orientée vers la classe 0 (non-résiliation), comme en témoigne la somme totale des contributions menant à une valeur de sortie très négative (-10.099).

Courbe de Lift :



La courbe de Lift permet d'évaluer la capacité du modèle à discriminer les classes en comparant la concentration des observations cibles dans les différents déciles d'une population triée selon les scores de probabilité. Le Lift indique le gain obtenu par rapport à une sélection aléatoire : une valeur de 1 signifie que le modèle n'apporte aucune amélioration, tandis qu'un Lift supérieur à 1 traduit une meilleure performance.

Classe 1 (Résiliés) :

Le modèle affiche un Lift maximal de 8,70 dans le premier décile, indiquant que les 10 % de clients ayant les scores les plus élevés contiennent environ 8,7 fois plus de résiliés qu'une sélection aléatoire. Le deuxième décile présente également une performance significative avec un Lift de 1,29, ce qui montre que le modèle conserve une certaine capacité de détection dans les 20 % supérieurs de la population.

À partir du troisième décile, le Lift chute brusquement à zéro. Cela signifie que le modèle ne détecte plus aucun résilié dans les 80 % restants. Il se focalise donc uniquement sur une minorité bien définie, ce qui peut s'avérer efficace pour des stratégies de ciblage très sélectives, mais constitue une limite si l'objectif est d'obtenir une couverture plus large des cas de résiliation.

Classe 0 (Actifs) :

Pour les individus de la classe 0 (actifs), le modèle présente une stabilité du Lift autour de 1,13 sur les huit premiers déciles, indiquant une capacité légèrement supérieure au hasard pour identifier les actifs. Les neuvième et dixième déciles montrent un repli progressif, avec un Lift de 0,96 et 0,03 respectivement, traduisant une baisse d'efficacité marginale sur ces segments extrêmes.

Ce modèle montre une forte capacité à identifier les résiliés les plus probables, avec un pouvoir discriminant élevé dans les deux premiers déciles. Toutefois, cette performance s'accompagne d'une perte totale de détection au-delà du deuxième décile, limitant la couverture à une portion restreinte de la population. En fonction des objectifs opérationnels, des ajustements comme le rééquilibrage des classes, l'ajustement du seuil de décision, ou l'intégration de modèles complémentaires pourraient être envisagés pour améliorer la couverture sans sacrifier la précision.

10. Modélisation sans la variable DTEFTMVT

10.1 Contexte

Après l'analyse des premières performances obtenues sur chaque modèle, des doutes ont émergé quant à la qualité des données. En particulier, nous soupçonnons la présence d'une fuite de données (data leakage), notamment via la variable DTEFTMVT, qui correspond à la date du dernier mouvement enregistré sur le contrat.

Dans la suite de l'étude, nous avons donc décidé de reconstruire les modèles sans inclure cette variable, afin de vérifier si elle influence anormalement les résultats et d'identifier le modèle réellement le plus performant.

En effet, une analyse plus approfondie de DTEFTMVT suggère que cette variable pourrait être mise à jour systématiquement juste avant une résiliation. Cela impliquerait qu'elle n'est disponible qu'après la survenue de l'événement que l'on cherche à prédire, ce qui constitue une fuite d'information typique. Pour le confirmer, nous avons comparé l'écart entre DTEFTMVT et la date d'échéance MMJECHPP. Les résultats montrent que, chez les résiliés, cet écart est très concentré autour d'une valeur fixe, tandis que pour les contrats actifs, il est bien plus dispersé.

Ces éléments renforcent notre suspicion d'un biais de fuite. Toutefois, un échange avec les équipes métier serait indispensable pour confirmer la nature exacte de cette variable et déterminer si elle peut être conservée ou non dans le cadre prédictif.

10.2 Modèle 1 : Logistic Regression

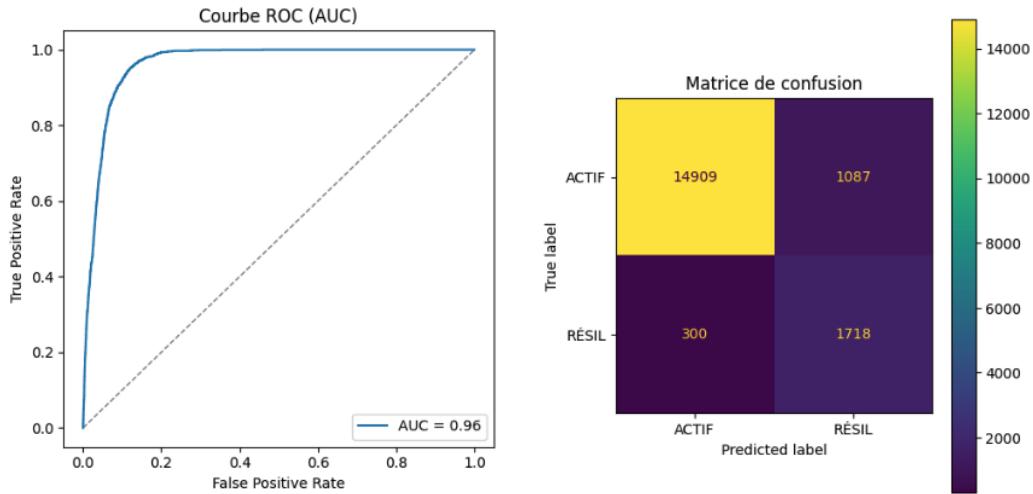
Le modèle de régression logistique a été reconstruit après plusieurs ajustements visant à renforcer sa robustesse et à corriger les éventuelles incohérences liées à un nombre élevé de variables. Par rapport à la première version du modèle, plusieurs variables ont été supprimées afin de limiter les problèmes de multicolinéarité et d'améliorer la lisibilité globale du modèle.

Surtout, la variable suspectée d'être à l'origine d'une fuite de données (DTEFTMVT) a été retirée du jeu de données pour éviter tout biais dans l'apprentissage.

Au final, par rapport à la version initiale, les variables suivantes ont été supprimées :

- DTEFTMVT
- TOTAL_SIN_RESP
- NBCTACT
- IV

Ces variables sont, pour la plupart, redondantes avec d'autres déjà présentes dans le modèle, notamment via des indicateurs synthétiques comme INDICE_RESPONSABILITÉ ou TAUX_RESIL_PONDERE_TOTAL.



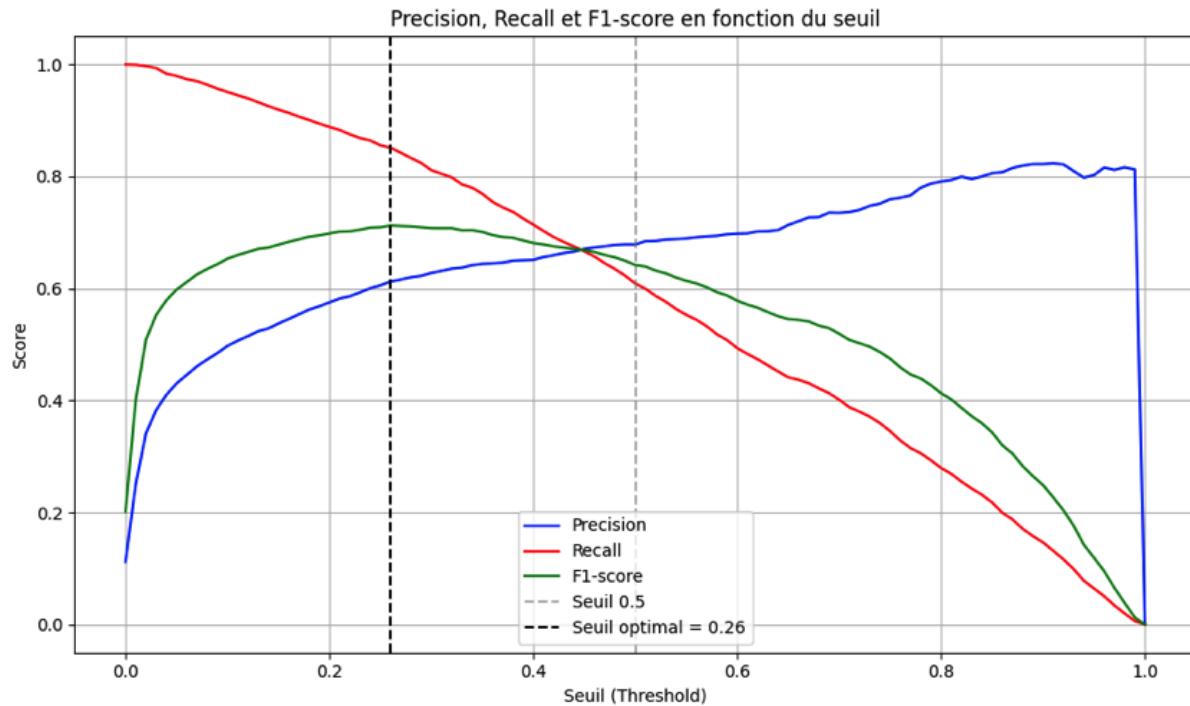
Generalized Linear Model Regression Results						
Dep. Variable:	CONTRAT	No. Observations:	72194			
Model:	GLM	Df Residuals:	72175			
Model Family:	Binomial	Df Model:	18			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-11165.			
Date:	Sat, 19 Jul 2025	Deviance:	22330.			
Time:	17:56:16	Pearson chi2:	7.85e+04			
No. Iterations:	8	Pseudo R-squ. (CS):	0.3252			
Covariance Type:	nonrobust					
	coef	std err	z	P> z	[0.025	0.975]
const	-9.5025	0.262	-36.258	0.000	-10.016	-8.989
ANCLLI	0.3297	0.029	11.238	0.000	0.272	0.387
CDMARVEH	8.1466	0.897	9.077	0.000	6.388	9.906
CDSITFAM	9.4045	1.284	7.324	0.000	6.888	11.921
CDUSGAUT	3.3001	0.790	4.177	0.000	1.751	4.849
CD_FML	1.4568	0.594	2.454	0.014	0.293	2.620
COEFCOMM	0.0353	0.008	4.420	0.000	0.020	0.051
COEFPFLT	0.2141	0.026	8.257	0.000	0.163	0.265
DI	0.1694	0.026	6.415	0.000	0.118	0.221
DTOBTPDC	-0.2962	0.036	-8.125	0.000	-0.368	-0.225
DTPMRMCI	-0.2312	0.027	-8.412	0.000	-0.285	-0.177
INDICE_RESPONSABILITE	0.2663	0.057	4.637	0.000	0.154	0.379
MH	0.7920	0.020	40.469	0.000	0.754	0.830
MMJECHP	0.5551	0.032	17.598	0.000	0.493	0.617
NOTAREFF	6.8511	0.485	14.115	0.000	5.900	7.802
RN_VL_VH	11.4772	1.270	9.034	0.000	8.987	13.967
SA	0.1893	0.046	4.106	0.000	0.099	0.280
TAUX_REMISE	0.0619	0.015	4.009	0.000	0.032	0.092
TAUX_RESIL_PONDERE_TOTAL	1.8220	0.018	101.602	0.000	1.787	1.857

Évaluation du modèle :
AUC : 0.9620
Gini (2*AUC - 1) : 0.9240
F1-score macro : 0.8340
F1-score classe 1 : 0.7124

Performances globales du modèle :

Le modèle présente une très bonne capacité de discrimination entre les classes (résilié vs actif). Une AUC aussi élevée (> 0.95 et par conséquent, Gini = 0.9240 très élevé) montre que le modèle est capable de classer correctement la majorité des observations.

On a aussi un bon équilibre global entre les deux classes, même si le dataset est déséquilibré. Le modèle ne favorise pas excessivement une classe au détriment de l'autre.



Le seuil de décision a été optimisé grâce à une méthode d'ajustement adaptée à la classification binaire.

En effet, l'optimisation des seuils de prédiction consiste à ajuster le seuil de décision qui permet de classifier une instance comme positive, en fonction de la probabilité prédictée par le modèle. Par défaut, ce seuil est fixé à 0,5, mais il peut être modifié afin de mieux répondre aux objectifs métier.

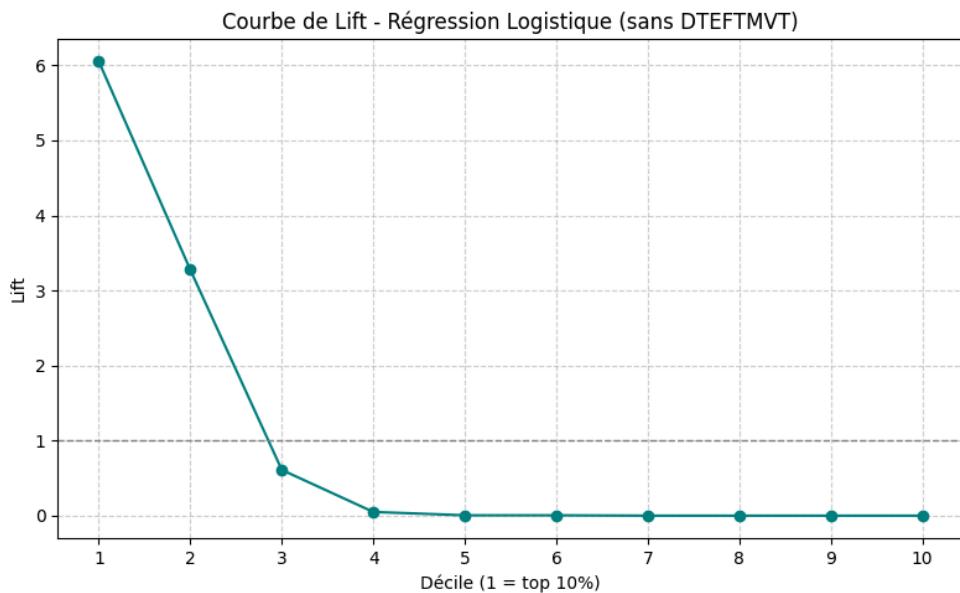
Selon le contexte, il peut être plus pertinent de maximiser certaines métriques comme la précision, le rappel ou encore le F1-score. Le graphique des *cutoffs* permet de visualiser l'évolution de ces performances en fonction du seuil choisi. Ce type de représentation aide à identifier un seuil optimal selon les priorités métier :

- un seuil plus bas favorise la détection (meilleur rappel), au prix d'un risque plus élevé de faux positifs ;
- un seuil plus élevé améliore la précision, mais peut réduire la capacité de détection.

Ainsi, l'optimisation du seuil permet d'adapter le modèle aux besoins spécifiques de l'analyse, en fonction de l'équilibre souhaité entre précision et rappel.

Dans notre cas, le seuil optimal a été fixé à 0,26.

Courbe de Lift :



	bin	count	positives	rate	lift	cumulative_positives	cumulative_rate
0	(0.502, 0.999]	1802	1222	0.68	6.05	1222	0.68
1	(0.129, 0.502]	1801	661	0.37	3.28	1883	0.52
2	(0.0266, 0.129]	1801	123	0.07	0.61	2006	0.37
3	(0.0124, 0.0266]	1802	10	0.01	0.05	2016	0.28
4	(0.00822, 0.0124]	1801	1	0.00	0.00	2017	0.22
5	(0.00584, 0.00822]	1801	1	0.00	0.00	2018	0.19
6	(0.00433, 0.00584]	1802	0	0.00	0.00	2018	0.16
7	(0.00317, 0.00433]	1801	0	0.00	0.00	2018	0.14
8	(0.00214, 0.00317]	1801	0	0.00	0.00	2018	0.12
9	(-0.000721000000000001, 0.00214]	1802	0	0.00	0.00	2018	0.11

Sur cette nouvelle courbe de lift, construite après la suppression de la variable DTEFTMVT (sujette à un possible data leakage), on observe que la performance reste correcte, bien qu'elle ait légèrement diminué.

Le modèle parvient toujours à capturer 68 % des résiliés dans le premier décile (les 10 % de clients les plus à risque), avec un lift de 6,05. Cela signifie que ce groupe est 6 fois plus riche en résiliés que si l'on sélectionnait un groupe de manière aléatoire.

Le deuxième décile contient 33 % de résiliés avec un lift de 3,28, ce qui reste un résultat intéressant. À partir du troisième décile, la capacité de détection chute fortement, le lift tombant à 0,61. Les déciles suivants n'apportent quasiment plus d'information, avec un nombre très faible voire nul de clients résiliés.

En résumé, même sans DTEFTMVT, le modèle continue de bien hiérarchiser les clients selon leur risque de résiliation. Les groupes les plus à risque sont bien identifiés, ce qui est essentiel pour cibler des actions de rétention.

10.3 Modèle 2 : Random Forest - Modèle simplifié

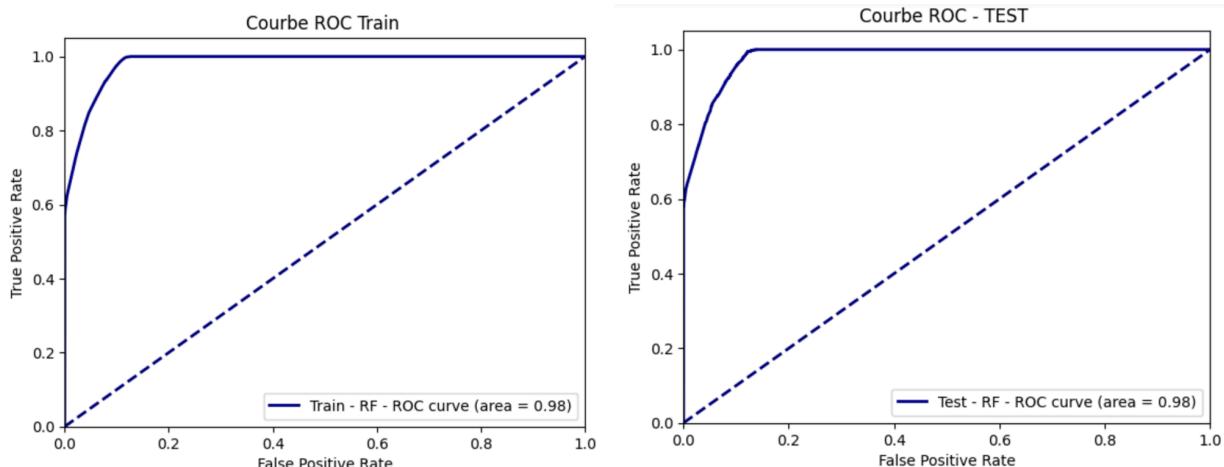
Dans cette partie, nous avons procédé à la sélection de variables en nous basant uniquement sur l'importance des variables dans notre modèle de base. Ces variables sont les suivantes: TAUX_RESIL_PONDERE_TOTAL, CLIACTIF_OUI, NBCTRES, NBCTACT, MH - chacune présentant un niveau d'importance (moyenne) substantiel.

Caractéristiques du modèle :

Le modèle Random Forest a été optimisé toujours à l'aide d'une recherche de Bayes (BayesSearchCV). Nous avons obtenu les mêmes niveaux et catégories pour les hyperparamètres avec un score ROC AUC moyen de 0.981, témoignant d'une excellente capacité de discrimination du modèle sur l'ensemble des plis de validation croisée.

Principaux résultats :

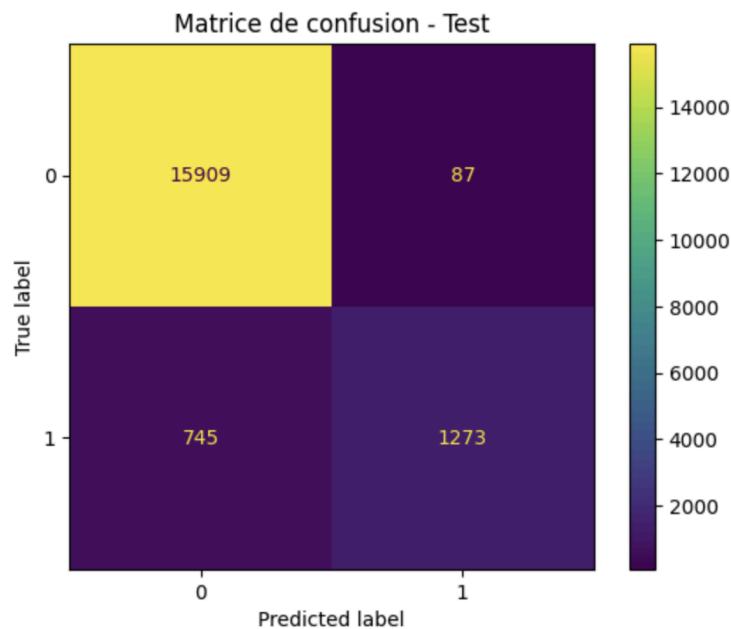
Figure: Courbes ROC - Ensembles Train et Test - RF



L'évaluation du modèle Random Forest réduit à ses cinq variables les plus importantes met en évidence une performance globale remarquable. La courbe ROC affiche une aire sous la courbe (AUC) de 0.98 aussi bien en entraînement qu'en test, traduisant une excellente capacité de discrimination du modèle, même après réduction dimensionnelle. Cela suggère que ces cinq variables concentrent l'essentiel de l'information pertinente pour prédire la résiliation.

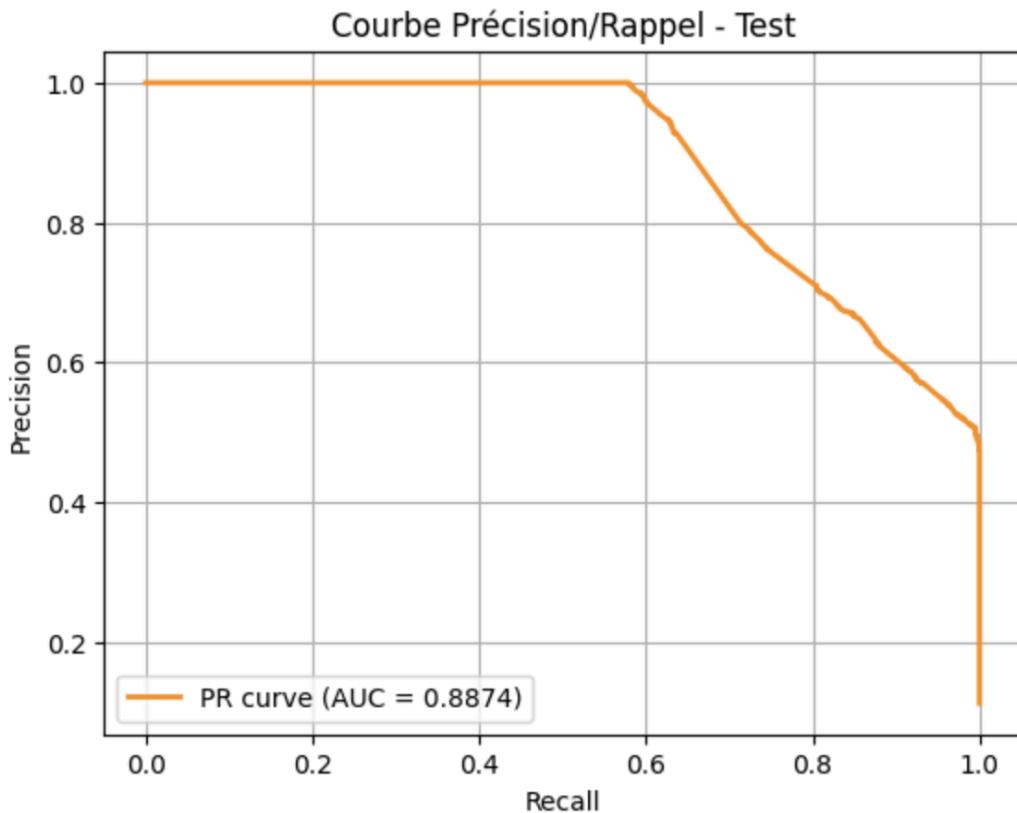
Figure: Classification report et Matrice de confusion - RF

	precision	recall	f1-score	support
0	0.9553	0.9946	0.9745	15996
1	0.9360	0.6308	0.7537	2018
accuracy			0.9538	18014
macro avg	0.9456	0.8127	0.8641	18014
weighted avg	0.9531	0.9538	0.9498	18014



La matrice de confusion pour les données de test révèle cependant un déséquilibre dans la capacité de prédiction selon les classes. Le modèle identifie très bien les contrats actifs avec 15 909 prédictions correctes sur 15 996, mais il peine à détecter tous les cas de résiliation : seuls 1 273 des 2 018 cas sont identifiés comme tels, soit un rappel de 63 %. Ce compromis se reflète dans le F1-score de la classe 1, qui s'établit à 0.75, contre 0.97 pour la classe 0.

Figure: Courbe Précision / Rappel - RF



La courbe précision-rappel confirme ce constat. Avec une AUC de 0.8874 pour la classe résiliée, le modèle maintient un bon équilibre global entre précision et rappel, même si une partie des cas à risque est encore ignorée. Ce niveau de performance est tout à fait satisfaisant dans un contexte où la résiliation est un phénomène minoritaire mais coûteux à anticiper.

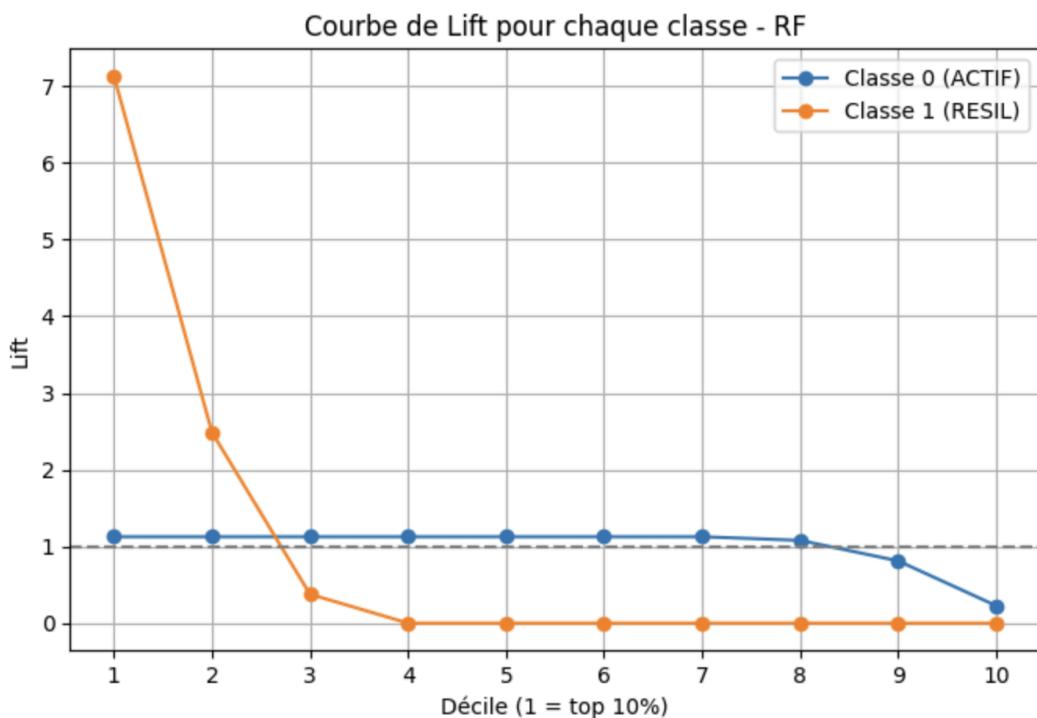
Tableau: Importance des variables - RF affiné

Variable	Feature Importance
TAUX_RESIL_PONDERE_TOTAL	0.46
NBCTRES	0.24
CLIACTIF_OUI	0.19
NBCTACT	0.06

Variable	Feature Importance
MH	0.04

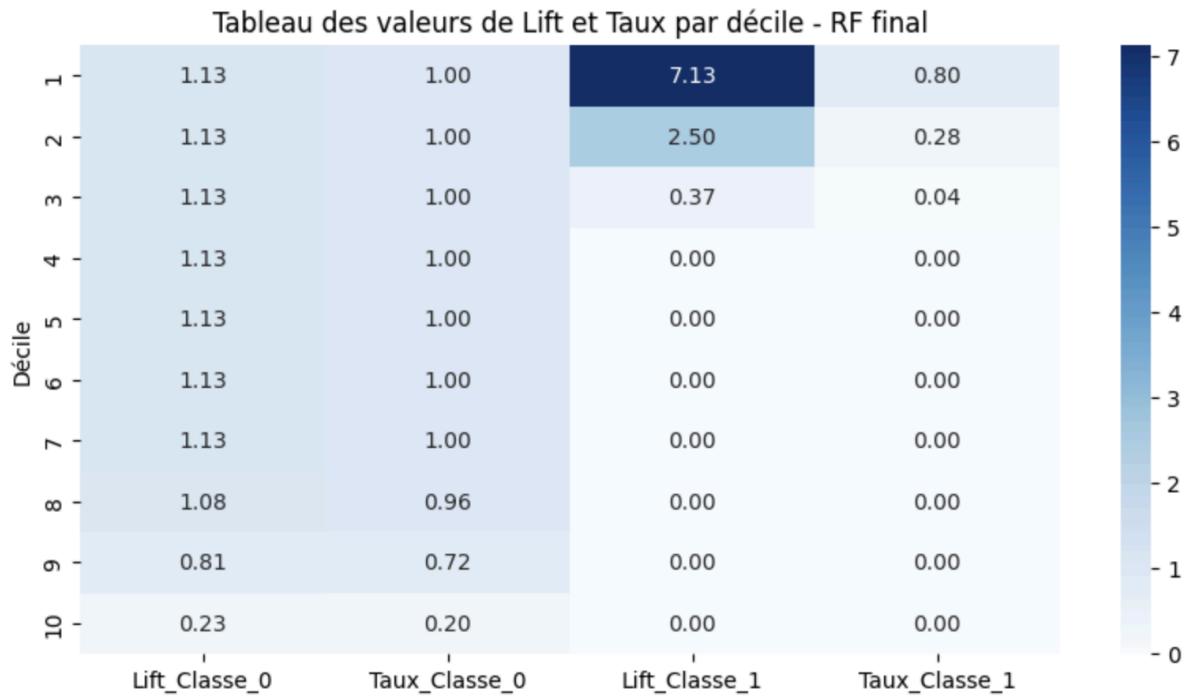
En termes d'explicabilité, l'analyse des importances de variables montre que TAUX_RESIL_PONDERE_TOTAL est de loin la variable la plus influente (poids de 0.46), suivie de NBCTRES et CLIACTIF_OUI. Les diagrammes SHAP permettent d'aller plus loin : une valeur élevée du taux de résiliation pondéré augmente fortement la probabilité prédictive de résiliation, tandis qu'un statut client encore actif joue en faveur d'un maintien. Le nombre de transactions, qu'elles soient résiliées ou actives, agit également dans ce sens.

Figure: Courbe Lift - RF affiné



Enfin, la courbe de Lift démontre l'efficacité opérationnelle du modèle. Le premier décile (10 % des clients les plus à risque selon le modèle) contient à lui seul environ 7 fois plus de résiliés que ce que l'on attendrait dans une répartition aléatoire. Cela en fait un excellent outil de ciblage marketing ou de prévention, car il permet de prioriser efficacement les interventions sur la population la plus vulnérable.

Tableau: Valeurs LIFT et Taux par décile - RF Final



En résumé, ce modèle Random Forest affiné ou simplifié conserve des performances élevées tout en améliorant la lisibilité et l'interprétabilité. Il constitue une base robuste pour des applications concrètes en gestion de la résiliation client.

10.4 Modèle 3 : XGBoost

Rapport de classification :

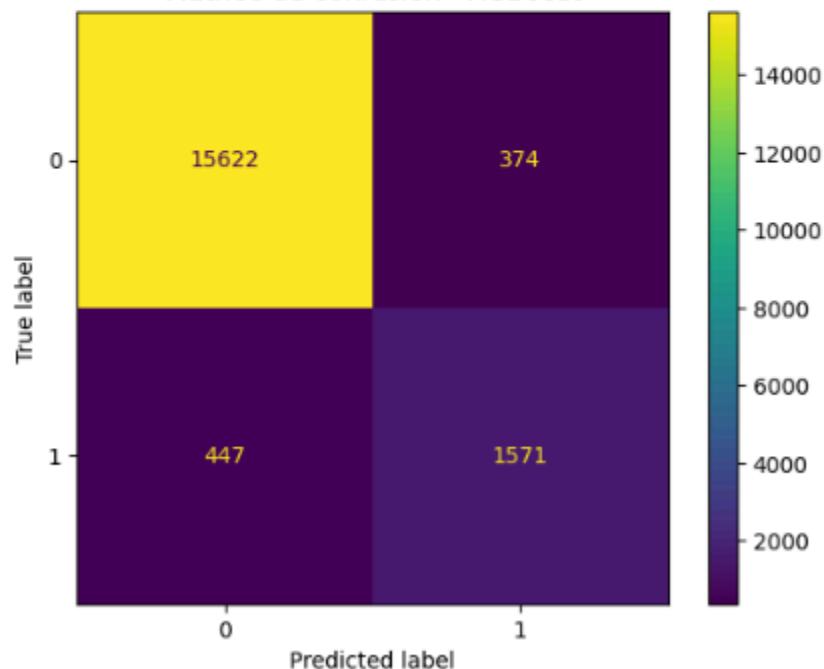
	precision	recall	f1-score	support
0	0.9722	0.9766	0.9744	15996
1	0.8077	0.7785	0.7928	2018
accuracy			0.9544	18014
macro avg	0.8899	0.8776	0.8836	18014
weighted avg	0.9538	0.9544	0.9541	18014

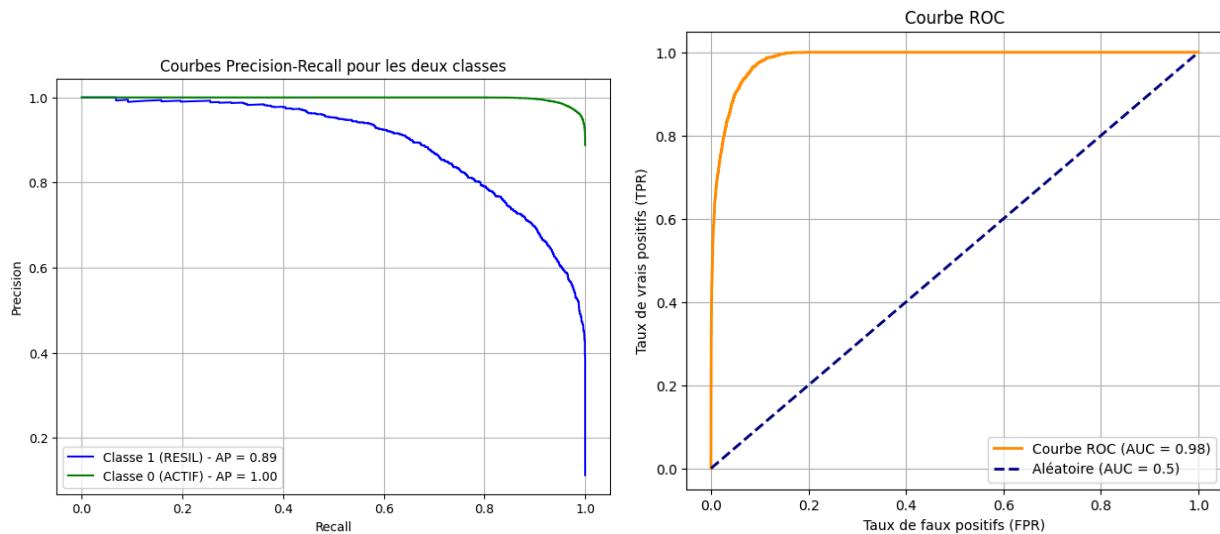
AUC (ROC) : 0.9842

Matrice de confusion :

```
[[15622  374]
 [ 447 1571]]
```

Matrice de confusion - XGBoost





Performances :

De manière globale, le modèle XGBoost offre des résultats très satisfaisants, avec une accuracy de 95,4 % sur les données de test. L'aire sous la courbe ROC (AUC) atteint 0,984 (et par conséquent, un Gini très élevé, 0,968), ce qui indique une excellente capacité de discrimination entre les clients actifs et ceux qui ont résilié leur contrat.

Pour la classe des clients actifs (classe 0), le modèle atteint une précision de 97,2 %, un rappel de 97,6 % et un F1-score de 97,4 %.

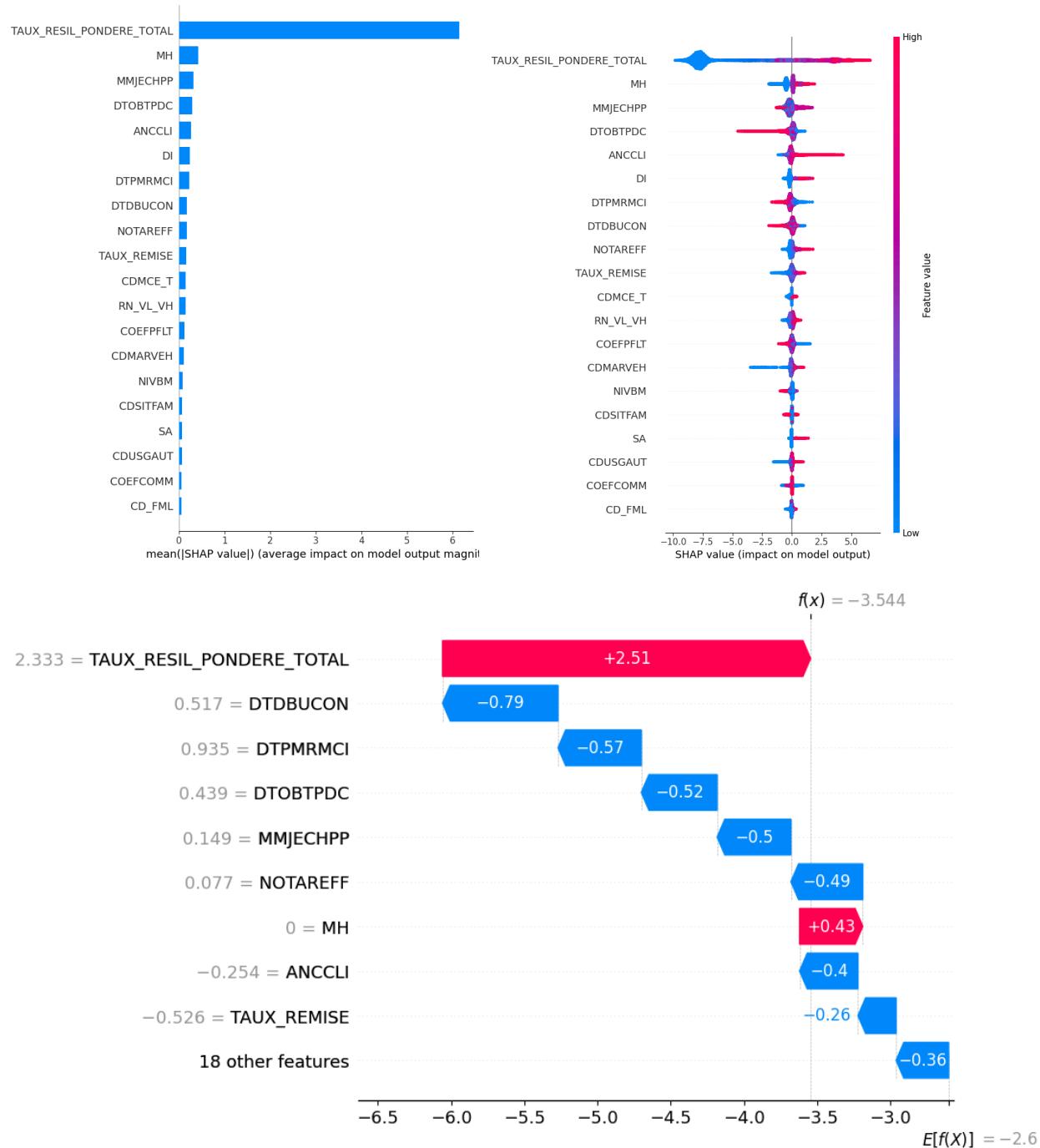
Pour la classe des clients résiliés (classe 1), les performances sont plus modestes avec une précision de 80,8 %, un rappel de 77,9 % et un F1-score de 79,3 %. Cette différence s'explique en partie par le déséquilibre entre les deux classes, la majorité des clients étant actifs.

De plus, nous n'avons pas voulu rééquilibrer les poids des classes, car une répartition 90 % vs 10 % nous semblait « assez » pour permettre au modèle de prédire correctement, tout en restant fidèle à la réalité observée. Un rééquilibrage artificiel aurait pu fausser les résultats en s'éloignant du contexte réel de l'entreprise.

Sur les 18 014 observations, le modèle a correctement identifié 15 622 clients actifs et 1 571 clients résiliés. En revanche, 447 clients résiliés ont été prédits à tort comme actifs, et 374 clients actifs ont été classés à tort comme résiliés. Cela montre une légère difficulté à détecter la classe minoritaire, mais les performances restent solides.

La courbe ROC met en évidence une très bonne séparation entre les deux classes, avec une zone AUC de 0.98. La courbe s'élève rapidement vers le coin supérieur gauche, ce qui montre que le taux de faux positifs est faible pour un taux élevé de vrais positifs.

Et enfin, les courbes Precision-Recall indiquent un AP (Average Precision) de 1.00 pour la classe active et de 0.89 pour la classe résiliée. Le modèle parvient donc à maintenir une bonne précision même pour la classe minoritaire, tout en assurant un bon rappel.



Analyse SHAP – Interprétabilité du modèle :

Afin d'interpréter les prédictions de notre modèle de manière globale et locale, nous avons utilisé les valeurs SHAP (SHapley Additive exPlanations). Cette méthode permet de mesurer l'impact de chaque variable sur la prédiction, en moyenne sur l'ensemble des observations ou pour une observation spécifique.

Pour l'importance moyenne des variables (SHAP - bar plot), on en conclut que :

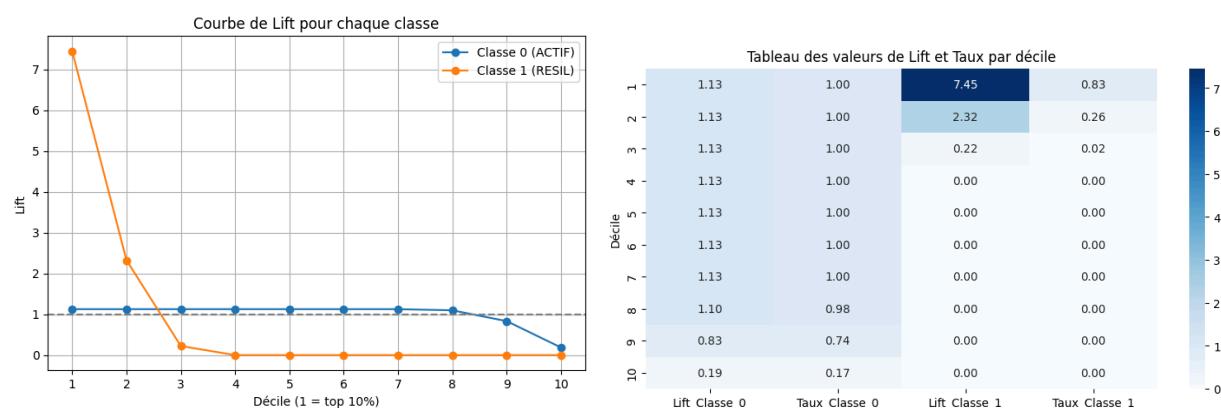
Ce graphique montre que TAUX_RESIL_PONDRE_TOTAL est, de loin, la variable la plus importante pour prédire la résiliation. D'autres variables ont un effet plus faible, comme MH, MMJECHPP, DTOBTPDC ou ANCCLI.

Concernant l'impact des valeurs des variables (SHAP - beeswarm plot) :

Ce graphique va plus loin : il montre comment des valeurs faibles (en bleu) ou élevées (en rose) de chaque variable influencent les prédictions. Par exemple, une valeur élevée de TAUX_RESIL_PONDRE_TOTAL augmente fortement la probabilité de résiliation. À l'inverse, une valeur faible réduit cette probabilité.

Enfin, une prédiction et son explication :

Ce dernier graphique montre en détail ce qui explique une prédiction individuelle. Dans l'exemple, la résiliation est surtout expliquée par une valeur élevée de TAUX_RESIL_PONDRE_TOTAL. D'autres variables, comme DTDBUCON, DTPMRMCI ou DTOBTPDC, tirent plutôt la prédiction vers une non-résiliation. La prédiction finale est $f(x) \approx -3.54$, bien en dessous de la moyenne du modèle, $E[f(X)] \approx -2.6$.



Analyse de la courbe de Lift :

La courbe de Lift permet de quantifier la performance du modèle en matière de détection des individus appartenant à une classe cible, en comparant leur concentration dans les différents déciles d'une population triée selon le score de prédiction. Un Lift supérieur à 1 indique que le modèle est plus performant qu'un tirage aléatoire.

Classe 1 (Résiliés) :

Le modèle atteint un Lift maximal de 7,45 dans le premier décile, ce qui signifie que les 10 % de clients ayant les scores les plus élevés contiennent environ 7,45 fois plus de résiliés qu'une sélection aléatoire. Le deuxième décile affiche également une performance notable avec un Lift de 2,32, traduisant encore une bonne capacité de ciblage.

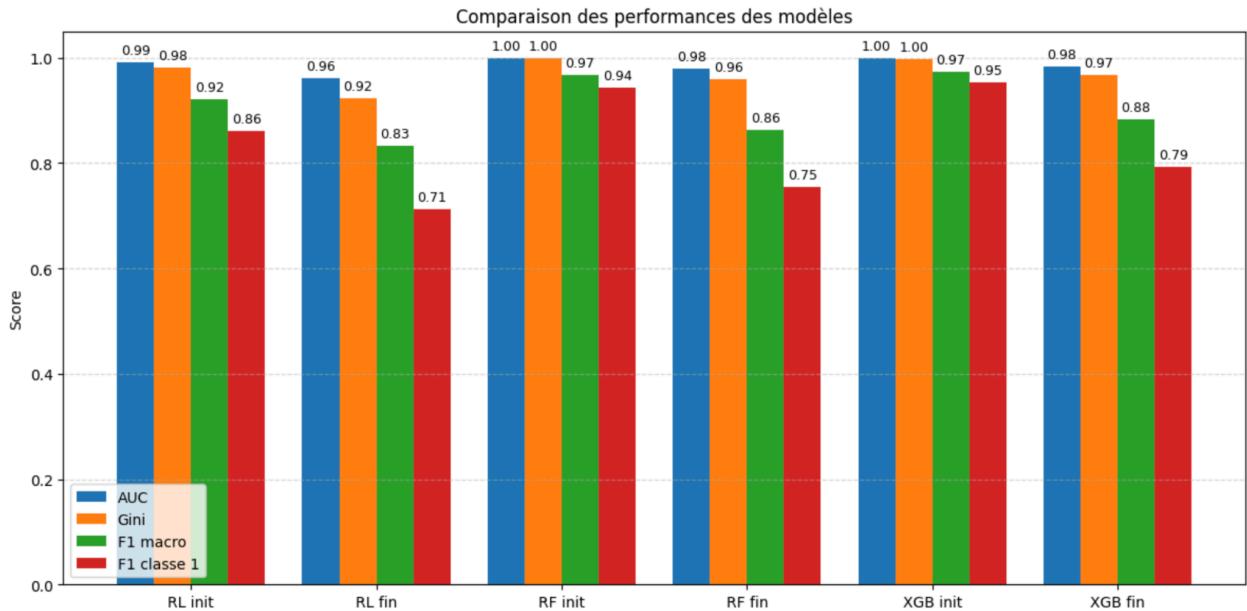
À partir du troisième décile, le Lift chute très rapidement : il descend à 0,22 puis devient nul dès le quatrième décile. Cela montre que la détection des résiliés est concentrée exclusivement dans les 20 % supérieurs de la population. Le modèle ne repère plus aucun individu de la classe 1 dans les 80 % restants. Il s'agit donc d'un modèle très sélectif, performant pour détecter les cas les plus extrêmes, mais avec une couverture limitée.

Classe 0 (Actifs) :

Concernant les individus actifs (classe 0), le modèle présente un Lift relativement stable autour de 1,13 sur les huit premiers déciles, ce qui indique une détection légèrement meilleure que le hasard. Les neuvième et dixième déciles révèlent une baisse progressive du Lift à 0,83 puis 0,19, ce qui traduit une perte d'efficacité sur cette fraction de la population.

Cette courbe traduit un modèle particulièrement efficace pour identifier les résiliés les plus probables, avec un pouvoir discriminant très élevé dans les deux premiers déciles. Cependant, cette performance se fait au détriment de la couverture globale, car aucun résilié n'est détecté au-delà du deuxième décile. En fonction des objectifs opérationnels (par exemple, maximiser la couverture ou limiter les faux positifs), des ajustements du modèle peuvent être envisagés : rééquilibrage des classes, réglage du seuil de probabilité, ou utilisation de méthodes complémentaires pour élargir la détection.

11. Comparaison des modèles - Performances



Nous arrivons au terme de notre étude et nous présentons le graphique qui compare les performances des trois modèles (régression logistique, Random Forest et XGBoost) dans deux versions : la version initiale (avec la variable DTEFTMVT) et la version finale (sans la variable suspectée de data leakage).

Les versions initiales des trois modèles affichent des scores exceptionnellement élevés, notamment en AUC et Gini (souvent à 0.99 ou 1.00), ainsi que des F1-scores très élevés, y compris pour la classe 1. Cela suggère un surenrichissement artificiel (ce qui a éveillé nos doutes), probablement dû à l'utilisation de DTEFTMVT, une variable qui encode indirectement la résiliation après qu'elle ait eu lieu (data leakage).

Après suppression de cette variable, les scores baissent légèrement mais deviennent plus crédibles et représentatifs d'un vrai contexte. On passe par exemple :

- Pour la régression logistique, d'un F1-score classe 1 de 0.86 à 0.71
- Pour XGBoost, de 0.95 à 0.79

Quel modèle choisir ?

- Régression Logistique (RL) :

Bien qu'il s'agisse du modèle le plus simple, il affiche des résultats solides, notamment un AUC de 0.96 et un F1-score classe 1 de 0.71 en version finale.

Son principal avantage est qu'il est interprétable, rapide à entraîner, facile à déployer et à expliquer aux équipes métiers. Il sera recommandé si la transparence et l'explicabilité sont prioritaires.
- Random Forest (RF) et XGBoost :

Ces modèles plus complexes offrent de meilleures performances, en particulier sur le rappel et le

F1-score de la classe minoritaire. Ils ont une meilleure capacité à capter les non-linéarités et les interactions.

Ils seront recommandés si la priorité est la performance pure, notamment en détection des résiliés.

D'un point de vue de performance pure (prédiction brute, sans contrainte d'interprétabilité), les modèles Random Forest et XGBoost se démarquent très nettement dans leur version initiale parce qu'ils atteignent des scores presque parfaits sur l'ensemble des métriques : AUC, Gini et F1-score, notamment pour la classe 1 (les clients résiliés), avec un F1-score de 0.86 pour RF et 0.88 pour XGBoost. Cela montre une très forte capacité de détection, bien supérieure à celle de la régression logistique. Néanmoins, ces performances irréalistes sont très probablement dues à une fuite d'information liée à la variable DTEFTMVT, ce qui biaise l'évaluation.

Après retrait de cette variable (version “finale”), les scores deviennent plus cohérents et permettent une meilleure comparaison. XGBoost reste le modèle le plus performant avec un AUC de 0.98 et un F1-score classe 1 de 0.79, suivi par la Random Forest avec un F1-score de 0.75. La régression logistique, bien que plus simple, conserve un bon niveau de performance avec un F1-score de 0.71, ce qui reste compétitif. XGBoost apparaît donc comme le meilleur modèle pour maximiser la détection des résiliations, même en contexte réaliste, tandis que la régression logistique constitue un excellent compromis pour des usages nécessitant de la transparence.

12. Conclusion

En conclusion, nous souhaitons d'abord évoquer la multiplicité des variables proposées pour la mise en place des différents algorithmes de prédiction et de classification. L'absence de contexte clair, ainsi que le manque d'explications sur chaque champ de la table, ont rendu cette mise en œuvre difficile. Cela nous a obligés à formuler des hypothèses sur le sens des variables, et une fois cumulées, ces incertitudes peuvent biaiser fortement le modèle.

Dès le début du projet, de nombreuses interrogations ont émergé concernant une possible fuite de données. Une partie importante du travail a donc été consacrée à l'analyse des variables susceptibles de poser ce type de problème.

Un autre objectif majeur était de simplifier le modèle. En effet, nous savons que l'usage intensif de données a un coût : plus l'utilisation est importante, plus les coûts augmentent. En parallèle avec un contexte métier réel, nous avons donc cherché à réduire la dimensionnalité du modèle, en supprimant autant que possible les variables redondantes ou peu utiles.

Ce projet a été très enrichissant, car il nous a confrontés à une situation proche du réel, notamment en matière de qualité de données. Cela a nécessité beaucoup de temps, d'énergie et d'itérations pour corriger les incohérences potentielles, affiner les modèles et surtout assurer une bonne interprétation des résultats.

Une piste intéressante pour aller plus loin serait de créer des groupes de risque ou classes de risque, permettant d'attribuer une note aux assurés, du moins risqué au plus risqué. Cela permettrait ensuite de mieux cibler les actions à mener pour maintenir et fidéliser les clients.