IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Yini LIU et Mayeul COTHENET

Titre : Prédiction de la résiliation des contrats d'assurance Santé Individuelle

Confidentialité : ☐ NON ☐ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

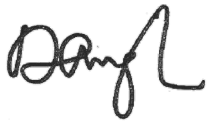
Secrétariat :

Bibliothèque :

Entreprise : AXA France

Nom : DANGIBEAUD Stéphane

Signature et Cachet :



Directeur de mémoire en entreprise :

Nom : LECLERC Olivier

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature(s) du candidat(s)

Prédiction de la résiliation des contrats d'assurance Santé Individuelle

Remerciements

Nous tenons à remercier monsieur Olivier Leclerc, responsable de la direction technique santé individuelle d'AXA France qui nous a permis d'effectuer la formation CEA, et qui nous a donné l'opportunité de travailler sur ce mémoire avec les bases de données d'AXA.

Merci à Florence Pasquié-Dussoubs et son successeur Stéphane Dangibeaud, Directrice et Directeur DECACGSI qui nous ont autorisé à continuer à travailler sur ces données après nos départs respectifs de l'équipe Santé individuelle d'AXA France.

Nous souhaitons également remercier l'ensemble des personnes qui nous ont soutenu pendant la rédaction, que ça soit pour le côté technique ou moralement.

Merci aux équipes de l'Institut du Risk Management et aux professeurs qui ont répondu à nos interrogations au cours de notre cursus du CEA.

Merci à Fangrou Ji qui nous a beaucoup aidé au démarrage en nous conseillant et nous éclairant sur les méthodes de machine learning.

Et enfin, merci à nos familles qui nous ont poussé à avancer, ont participé à la relecture de ce mémoire.

Résumé

Alors que la Résiliation infra-annuelle des contrats santé facultatifs entre en application au 1er décembre 2020 et que la duration des contrats reste un élément clé de la rentabilité, la prédiction de la résiliation des clients semble être un avantage différenciant.

En prenant une photo du portefeuille en début d'année, serons-nous en mesure de déterminer quels assurés vont résilier dans l'année ?

L'étude du portefeuille de Santé Individuelle d'AXA France a permis d'expérimenter et comparer 2 générations de méthodes de prédiction. La régression logistique, méthode traditionnelle permet d'élaborer un score pour chaque individu représentant son risque de résiliation. La forêt aléatoire, elle, classifie les contrats résiliés ou non en réalisant une multitude d'arbres de décisions aléatoires sur des sous-échantillons qui permettent ensuite d'établir une prédiction.

Bien qu'enrichies par des données en open data, la régression logistique ne s'est pas révélée efficace.

En revanche, la forêt aléatoire a permis de créer une sous-population ayant un risque de résiliation plus de 5 fois supérieur à la population globale et d'identifier plus d'un tiers des résiliations.

Des mesures peu coûteuses, car concentrées sur les assurés « risqués », sont donc envisageables et pourraient s'avérer efficaces sur le taux de résiliation.

Mots clés : résiliations, contrats santés individuelles complémentaires, prédiction, régression logistique, forêt aléatoire

Abstract

When the infra-annual termination law for individual complementary health cover enters in applications starting from the 1st of December 2020, the duration of the Health insurance contracts become a key element for the portfolio profitability, such that the capacity of predicting the contracts termination seems to be a competitive advantage.

We focus on finding the efficient explanatory variables for predicting the termination of contracts.

Two methods, the logistic regression and random forest, are used to study the AXA France's portfolio.

Our study shows that even though our data was enhanced with the external open data, the logistic regression cannot identify the insureds who have the strong tendency to terminate contracts efficiently.

The random forest performs much better than the logistic regression and it identifies more than one third of the terminations.

In addition, the random forest allows us to establish a sub-set population with a termination risk five times more than that of the overall data set such that we may make the termination rates decrease a lot by taking preventive actions on the sub-set selected by the random forest.

Key words: termination, individual health insurance, prediction, logistic regression, random forest

Table des matières

Introduction

Alors que la Résiliation infra-annuelle des contrats santé facultatifs entre en application au 1er décembre 2020 et que la duration des contrats reste un élément clé de la rentabilité, la prédiction de la résiliation des clients semble être un avantage différenciant.

Certaines études ont d'ores et déjà été réalisées sur d'autres risques, telles que celle d'Antoine BRUN (2017) qui a étudié la résiliation en assurance emprunteur. Mais également sur des portefeuilles de Santé Individuelle, dont les travaux d'Arthur DUTEL (2017) qui a étudié l'impact de la stratégie tarifaire sur la résiliation des contrats, mais aussi ceux d'Anna GEHLER (2009) et son approche descriptive et la classification des profils de clients qui résilient et enfin le mémoire dont la problématique est la plus proche de notre objectif : celui de Matthias VALLA (2018).

Nous avons essayé, dans le cadre de cette étude, une nouvelle approche : celle d'identifier avec une bonne probabilité quels clients seraient sur le point de résilier leur contrat d'assurance santé individuelle afin de pouvoir mettre par la suite en place des mesures de rétentions et essayer de prévenir cette résiliation. Nous cherchons donc à attribuer un score de 0 (« non résilié dans l'année ») ou 1 (« résilié dans l'année ») à chacun de nos clients en portefeuille.

Pour ce faire, nous disposons de deux années d'expérience sur plus de 400 000 clients. Cela nous permettra de définir un modèle une année et de vérifier sa stabilité dans le temps l'année suivante.

Nous avons donc pour cela souhaité évaluer deux méthodes statistiques. La première étant une méthode statistique « traditionnelle », à savoir, la méthode de la régression logistique de la famille des modèles linéaires généralisés, grâce à laquelle nous avons souhaité mettre en place un score de résiliation.

La seconde méthode que nous avons souhaité tester est une méthode plus moderne puisqu'il s'agit de la méthode de forêt aléatoire. Grâce à celle-ci, nous avons pu classifier nos clients en deux catégories : ceux pour lesquels nous prédisons une résiliation dans l'année et les autres.

Après une présentation du contexte de l'assurance santé et de nos données, nous entrerons dans le sujet par une analyse des variables dont nous disposons.

Nous présenterons ensuite nos deux méthodes l'une après l'autre ainsi que notre cheminement et la façon dont nous avons personnalisé chaque méthode afin de parvenir aux meilleurs résultats.

Nous finirons par comparer les deux méthodes sur la base de plusieurs critères afin de définir quel modèle nous semble le plus pertinent et avec quelle efficacité.

Chapitre 1. Présentation du contexte, problématique et approche

1.1. *L'assurance Santé*

L'assurance santé en France propose une couverture sur principalement deux niveaux : un obligatoire et un complémentaire facultatif.

1.1.1. Le régime obligatoire ou Assurance Maladie :

Comme l'indique le site internet officiel de la Sécurité sociale, depuis les débuts, en 1928, des premiers régimes de Sécurité sociale couvrant les salariés de l'industrie et du commerce, le système de Sécurité sociale a beaucoup évolué. On notera évidemment la date de 1945, à laquelle les ordonnances d'octobre créent un modèle de Sécurité sociale « visant à assurer, à tous les citoyens, des moyens d'existence dans tous les cas où ils sont incapables de se les procurer par le travail » selon le programme du Conseil National de la Résistance. Ce nouveau régime ne couvre que les salariés. La gestion est allouée aux partenaires sociaux et son financement aux employeurs et aux salariés. Le régime est constitué comme un socle obligatoire, qui ne couvre pas l'intégralité du risque, mais permet à des opérateurs complémentaires de compléter la couverture, notamment en santé où la complémentaire peut prendre en charge le ticket modérateur destiné à laisser un reste à charge aux patients afin de limiter l'incitation au recours à ce nouveau régime. La mise en place permet également de maintenir les régimes spéciaux déjà en place.

En 1952, la Mutualité sociale agricole (MSA) est fondée pour assurer le risque de vieillesse des exploitants agricoles. 9 ans plus tard, le risque maladie aussi sera couvert par ce régime.

En 1966, les Travailleurs non-salariés non-agricoles se voient dotés d'un régime maladie-maternité. Et un an plus tard, la structure de la Sécurité sociale des salariés est donnée, à savoir le découpage en 3 branches distinctes : Santé, vieillesse et Famille. Elles sont maintenant 5 à ce jour, puisque l'AT-MP (accident du travail et maladie professionnelle) et la dépendance font désormais partie des branches de la Sécurité sociale.

En 1978, une loi ancêtre de PUMa (Protection Universelle Maladie) dont nous parlerons plus tard instaure un mécanisme d'assurance personnelle pour la « population résiduelle » qui ne relève pas d'un régime existant, notamment les ministres de cultes.

Depuis les années 1990, le rythme des réformes de la Sécurité sociale a beaucoup accéléré. En 1996, la catégorie des Lois de Financement de la Sécurité sociale est créée en parallèle des Lois de Finance. En effet, le budget de la Sécurité sociale est nettement supérieur au budget de l'Etat (en 2019 : la Sécurité sociale a eu 532 Mds € de recette, contre des recettes de l'Etat de 291,3 Mds € avant prélèvements de l'Union Européenne).

En 1999, la Couverture Maladie Universelle est créée pour couvrir l'ensemble des personnes résidant en France de façon régulière, y compris sans domicile fixe. Elle est gratuite pour les personnes à faible revenus et payante pour les autres. Elle peut être complétée par un dispositif de CMU-c (Couverture Maladie Universelle complémentaire) qui donne accès à une complémentaire santé gratuitement.

En 2006, les indépendants se regroupent au sein du nouveau Régime Social des Indépendants (RSI).

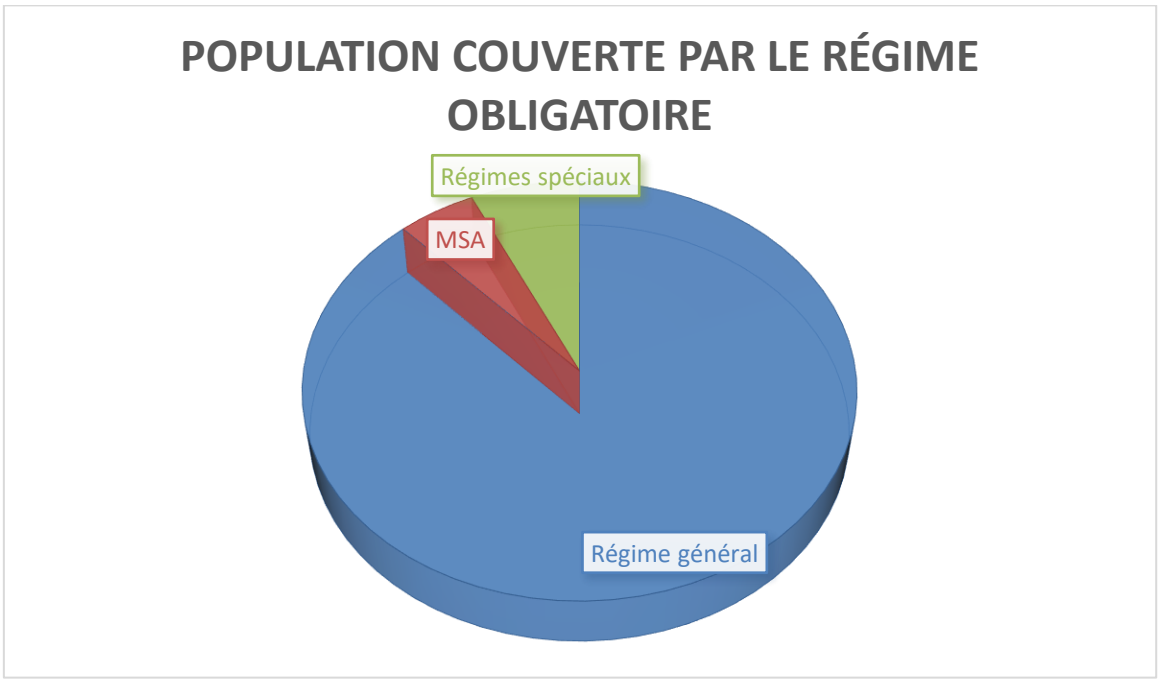
En 2015, le Tiers payant est généralisé à tous les assurés.

En 2018, le RSI est rattaché au Régime Général. Ils sont suivis de près par les régimes de Sécurité sociale étudiants (2019).

Aujourd'hui, le régime obligatoire couvre l'ensemble des personnes qui travaillent ou résident en France de manière stable et régulière (ameli.fr). C'est-à-dire, l'ensemble des régimes (général, agricole et spéciaux) au travers de différentes caisses : la Caisse Nationale d'Assurance Maladie (pour 88% de la population, y compris les indépendants depuis 2020) et la Caisse Centrale de la Mutualité Sociale Agricole (pour 5% de la population) qui se réunissent au sein de l'UNCAM et de 27 autres caisses pour les 7% de Français bénéficiant de régimes spéciaux.

La Direction de la Sécurité sociale est en charge de coordonner les différentes caisses en concevant les politiques publiques en matière d'accès aux soins et de régulation financière, afin de maintenir l'équilibre des régimes.

La répartition de la population couverte par le régime obligatoire est présentée ci-dessous dans le graphique 1 :



Graphique 1 : Population couverte par le régime obligatoire

Les cotisations à ce régime sont fonction des revenus.

Selon les postes de dépenses santé, le régime obligatoire couvre entre 0 et 100% des dépenses. Le reste des dépenses peut être pris en charge par des assurances complémentaires ou surcomplémentaires ou rester à la charge des patients. Le remboursement des régimes obligatoires et complémentaires ne peut dépasser les frais réellement engagés par le patient.

A noter : il existe un régime obligatoire complémentaire pour les salariés d'Alsace et de Moselle, qui représente un cas particulier sur lequel nous ne nous étendrons pas.

1.1.2. Les complémentaires santé

Le rôle des complémentaires santé est de diminuer le reste à charge des patients. L'ensemble des preneurs de risques se réunissent au sein de l'UNOCAM (Union Nationale des organismes complémentaires d'assurance maladie) qui regroupe les représentants de la FFA pour les assureurs, de la FNMF pour les mutuelles et du CTIP pour les Instituts de Prévoyance.

Les assureurs sont encadrés par le Code des Assurances ;
Les mutuelles par le Code de la Mutualité ;
Les IP par le Code de la Sécurité sociale.

La Direction de la Sécurité sociale prépare les lois de finance de la Sécurité sociale qui impactent ces 3 codes lors des réformes des régimes complémentaires.

Il s'agit de contrats qui sont généralement facultatifs pour les individus, mais obligatoires pour les entreprises dans le cadre de la couverture de leurs salariés. En 2019, en France, 95% de la population était couverte par une complémentaire santé individuelle ou collective d'après la DREES.

On le rappelle, les salariés ont été parmi les premiers à bénéficier de la Sécurité sociale. Ils sont également les premiers à bénéficier d'une couverture complémentaire obligatoire par leur employeur suite à l'Accord National Interprofessionnel de 2013, qui a été transcrit dans la Loi et entré en vigueur en 2016. Il instaure une couverture minimale obligatoire, financée à au moins 50% par l'employer. Quelques cas de dispenses sont toutefois prévus dans le cas de contrats courts, ou notamment lorsque la couverture est trop coûteuse pour un salarié. Le marché de la santé individuelle que nous étudierons s'est donc rétracté et se concentre principalement sur les indépendants, les étudiants, les retraités, les salariés dispensés, mais également par les salariés d'entreprises qui ne seraient pas encore en conformité avec la loi. Du fait du caractère durable de leur situation, les indépendants et les retraités sont qualifiés de cibles « pérennes », et sont donc des cibles de clientèle clients privilégiées en santé individuelle.

En collectif comme en individuelle, le choix de l'organisme assureur par le souscripteur du contrat est libre : le marché est concurrentiel. Un souscripteur peut résilier son contrat pour en souscrire un nouveau auprès d'un autre assureur. En revanche, un assureur a l'obligation de maintenir les garanties si le bénéficiaire en fait la demande (garanties santé viagères depuis la loi Evin du 31 décembre 1989).

Les cotisations sont réglées soit par le souscripteur, soit par l'adhérent ou le salarié. A la différence des régimes obligatoires, elles dépendent généralement de son niveau de risque ou du niveau de risque du groupe auquel il appartient pour des contrats collectifs.

Les prestations remboursent les frais des soins dont bénéficient les clients en complément d'un régime obligatoire. Certaines dépenses peuvent être parfois remboursées sans intervention du

Régime Obligatoire, mais les remboursements ne peuvent jamais dépasser les frais réels engagés ce sont des remboursements « en nature ».

Ils sont, pour de nombreux postes, encadrés par la règlementation, particulièrement pour 95% des contrats en France qui répondent à un cahier des charges dit des “contrats responsables”, que nous développerons en point *1.2 les contrats responsables et solidaires*.

1.1.3. L’assurance Santé chez AXA

AXA France fait partie du groupe AXA. L’entité française représente 25% du chiffre d’affaires du groupe. AXA France regroupe en réalité 4 entités : AXA France Vie et AXA France IARD qui sont des compagnies d’assurance et AXA Assurances Vie Mutuelle et AXA Assurances IARD Mutuelle qui sont des mutuelles d’assurance.

AXA France est le troisième plus grand assureur santé en France en 2019 derrière le groupe Vyv (regroupement des mutuelles Harmonie Mutuelle, MGEN et Istya) et le rapprochement Malakoff-Humanis avec un chiffre d’affaires de 2 339 millions d’euros (d’après le Top 30 de la Santé 2020 de l’Argus de l’Assurance). AXA France est donc le 1er assureur et suit 2 regroupements de mutuelles.

La branche Santé est portée par les compagnies Vie. Le chiffre d’affaires Santé est surtout porté par les assurances dites « Collectives », c’est-à-dire qui couvrent les salariés d’entreprises ou les adhérents d’associations. Environ 85% du chiffres d’affaires de la santé individuelle est chez AXA France Vie. 15% sont chez AXA Assurances Vie Mutuelle. La répartition se fait en fonction du distributeur de l’assurance (Agent général ou courtier par exemple), mais il est assez transparent pour les clients des deux entités, car le produit commercialisé et les process sont absolument identiques.

Concernant la Santé dite « individuelle », AXA France commercialise des produits au travers de nombreux canaux : réseau salarié, réseaux de plus de 4000 agents généraux et courtiers. La partie de portefeuille étudiée dans le cadre de ce mémoire est celle commercialisée par les réseaux dits « propriétaires », c’est-à-dire les réseaux salarié et d’agents généraux et une faible part du portefeuille issu du courtage.

Seule la partie du portefeuille constituée des deux dernières générations des produits AXA sont pris en compte. Ceux-ci se nomment « Référence » et « Ma Santé ». Il s’agit de générations de contrats dits « responsables et solidaires », qui ont été commercialisé depuis 2001.

1.2. *Les contrats responsables et solidaires*

Les contrats « responsables et solidaires » respectent certains critères définis par le législateur en contrepartie d'une fiscalité allégée.

Les contrats solidaires (article 1001 2bis du CGI) sont des contrats qui n’appliquent pas de sélection médicale avant la souscription. Ils ne fixent pas non plus les primes et cotisations en fonction de l’état de santé de l’assuré.

Les critères de “responsabilité” sont apparus en 2004 et ont évolué régulièrement. Aujourd’hui ils représentent un cahier des charges assez strict avec :

- Une obligation de communication des taux de frais aux adhérents ou aux souscripteurs.
- Une obligation de respect des pénalités imposées par le régime obligatoire : franchises, participations forfaitaires, plafonnement des honoraires, etc.
- Une obligation de respect des planchers et plafonds de remboursements en fonction du poste de santé :
  - Au minimum 100% du ticket modérateur sur la quasi-totalité des actes pris en charge par le régime obligatoire ;
  - Au minimum 100% des Prix limites de ventes et des honoraires limite de facturation des actes entrant dans les paniers de soins dits “100% Santé” en optique, dentaire et audition (en 2021) ;
  - Au minimum 100% du forfait journalier sans limite de durée et de la participation aux actes lourds ;
  - Planchers et plafonds en optique suivant la complexité des verres ;
  - Plafonnement des remboursements des honoraires de médecin n’ayant pas signé de Dispositif de Pratique Tarifaire Maîtrisé avec la Sécurité sociale.

1.3. *Les montages juridiques*

Les contrats de complémentaires santé peuvent suivre plusieurs montages juridiques : Collectifs obligatoires, collectifs facultatifs ou individuels.

Les contrats collectifs obligatoires sont typiquement des contrats d’entreprise : le souscripteur est l’entreprise et le contrat va couvrir de façon obligatoire les salariés (ceux-ci peuvent avoir des cas d’exonération).

Les contrats collectifs facultatifs sont par exemple des contrats dits “associatifs” : une personne morale souscrit un contrat et les personnes qui lui sont rattachées peuvent, si elles le souhaitent, s’affilier au contrat.

Enfin, dans le montage des contrats individuels, les personnes physiques souscrivent directement les contrats auprès des assureurs.

Nous nous intéresserons particulièrement aux contrats individuels dans ce mémoire. Sauf mention contraire, les éléments indiqués vaudront pour ce montage particulièrement.

1.4. *Les modalités de résiliation*

Les contrats sont généralement annuels avec une tacite reconduction à l’échéance principale.

Il existe toutefois un moyen de sortir de cette tacite reconduction, c'est ce qu'on appelle la résiliation, c'est-à-dire la fin du contrat de façon prématurée ou non par l'une ou l'autre des parties (compagnie ou assuré), dans un cadre défini par les lois et le contrat.

Pour résilier son contrat, la disposition principale est la résiliation à l'échéance par lettre recommandée : 2 mois avant l'arrivée de l'échéance principale, l'assuré peut demander le non-renouvellement de son contrat.

Il existe d'autres dispositions légales moins connues :

- Lorsque la compagnie n'informe pas 15 jours avant l'arrivée du délai préalable des 2 mois pour la résiliation, alors l'assuré a la faculté de résilier son contrat (Loi Chatel L. 113-15-1 du Code des assurances). La résiliation intervient alors le lendemain de la demande.
- Lorsque la situation de l'assuré change et que cela a un impact sur la prime (changement de régime obligatoire, éventuellement la situation matrimoniale, le domicile, la profession, le départ en retraite, la cessation d'activité, etc.). La résiliation intervient le mois suivant la demande de résiliation.
- Lorsque l'assuré se retrouve couvert par un contrat collectif obligatoire d'entreprise. Dans ce cas, la résiliation intervient à l'échéance suivante (des clauses contractuelles pouvant être plus souples pour le client).
- Lorsque la cotisation augmente au-delà des clauses d'indexation contractuelles. Dans ce cas, la résiliation intervient dans le délai prévu au contrat (généralement 15 jours ou le mois suivant).

Ces dernières dispositions sont moins connues et moins pratiquées. La plus courante est actuellement la clause de résiliation à l'échéance principale.

A noter que le taux de résiliation d'un portefeuille ayant un impact direct sur la rentabilité de celui-ci, il est particulièrement suivi. Nous définissons ainsi le taux de résiliation annuel :

$$Taux\ de\ résiliation_{année\ N} = \frac{nombre\ de\ résiliations\ reçues\ dans\ l'année\ N}{nombre\ de\ contrats\ effectifs\ au\ début\ de\ l'annéeN}.$$

1.5. *La résiliation infra-annuelle*

Ces démarches de résiliation, assez lourdes administrativement pour le client viennent à l'encontre de la libre concurrence : c'est une cause de rétention des clients.

C'est pourquoi, la loi du 14 juillet 2019 relative au droit de résiliation sans frais de contrats de complémentaire santé a été adoptée.

Cette loi comporte 4 volets :

- Volet de transparence : Les contrats responsable d'assurance Santé sont contraints depuis le 1er septembre 2020 d'indiquer avant la souscription et à chaque renouvellement le taux

de frais de l'assureur ainsi que la part de primes dédiée au paiement des sinistres Santé (appelé chez AXA France ratio sinistres sur cotisations ou primes soit « S/C » ou « S/P »).

- Volet de simplification de la résiliation : Les contrats santé et prévoyance peuvent, depuis le 1er décembre 2020 être résiliés par courrier simple ou tout support durable (voie électronique notamment). La lettre recommandée avec accusé de réception n'est donc plus obligatoire. L'assureur est également obligé d'accepter une résiliation effectuée par le même média que la souscription du contrat. Ainsi, un contrat souscrit par téléphone pourra être résilié par téléphone.
- Volet de résiliation à tout moment : après l'écoulement d'un délai d'un an à compter de la première souscription, un souscripteur peut résilier son contrat à tout moment. La résiliation prendra effet 1 mois après que l'assureur en a reçu notification par l'assuré. Elle sera sans frais ni pénalités. Une entreprise ou un assuré individuel pourront donc résilier leur assurance en 1 mois, sans attendre l'échéance anniversaire. En revanche, un salarié ne peut toujours pas résilier son contrat obligatoire d'entreprise. En complément la loi instaure l'obligation pour le nouvel assureur de s'assurer d'une continuité d'assurance pour l'assuré au moment de son changement.
- Un volet de suivi des droits en temps réel : La réforme a donné la mission à l'UNOCAM de mettre en place un système d'interrogation en direct des droits des assurés accessible aux professionnels de santé afin que ceux-ci puissent s'assurer (entre autres) que les cartes de tiers-payant présentées sont toujours valables.

L'ensemble de ces dispositions devrait faciliter le parcours client lors de la résiliation.

Ainsi, cette loi va grandement déstabiliser les portefeuilles. A l'heure où la duration des contrats est un élément clé de rentabilité, arriver à réduire la résiliation devient un enjeu majeur.

1.6. *Problématique*

Afin de pouvoir sécuriser les portefeuilles, avons-nous la possibilité de prédire la résiliation d'un client avant sa notification ?

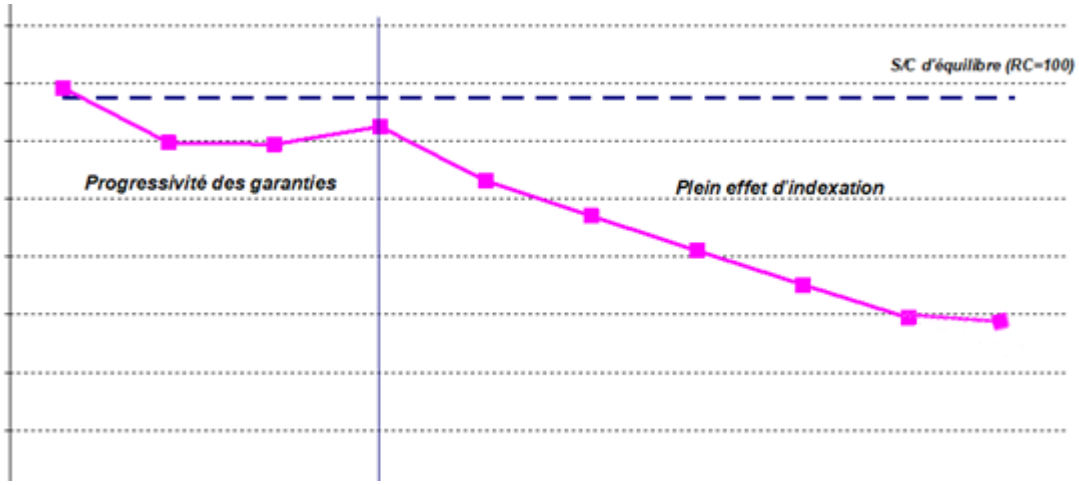
De façon plus précise :

En ayant une vision d'un portefeuille de contrats santé à une date, les méthodes de machine learning du XXIème siècle sont-elles plus efficaces que les méthodes GLM pour prédire les contrats qui seront résiliés dans l'année suivant celle-ci ?

La prédiction de ces résiliations sera un élément clé pour permettre des actions préventives auprès des clients les plus à-même de partir. En effet, la duration de ces contrats permettra une meilleure rentabilité (amortissement des frais assureurs, des commissions de 1ère année auprès du distributeur, amélioration de la rentabilité dans le temps).

En particulier, pour la gamme Ma Santé, les différentes formules intègrent une progressivité de garantie : par exemple, à partir de la 2ème année, nous donnons plus de garanties en optique, à partir de la 3ème année, les garanties augmentent en dentaire. Ainsi, malgré les majorations annuelles,

l'amortissement du contrat ne prend son plein effet qu'à compter de la 4ème année. L'enjeu est donc fort de conserver les contrats le plus longtemps possible. Un exemple d'évolution du S/C d'une génération de contrat est présenté dans le graphique 2.



Graphique 2 : Evolution du S/C d'une génération de contrat

1.7. *Approche*

Nous souhaitons donc détecter les clients ayant le plus de probabilité de partir afin de pouvoir envisager des mesures de rétention. Pour cela, nous allons comparer 2 méthodes statistiques. La première sera par le biais d'une régression GLM logistique afin de catégoriser les assurés. Le seconde, ayant la même finalité, reposera sur une méthode plus actuelle de machine learning.

Apprentissage et test des modèles :

Sur le portefeuille en vigueur au 31/12/2017, combien de contrats sont résiliés au cours de l'année 2018 : 70% des données sont utilisées comme données d'apprentissage et 30% des données sont utilisées pour le test de validation du modèle.

Ensuite, des prédictions seront faites sur le portefeuille en vigueur au 31/12/2018, avec les résiliations constatées au cours de l'année 2019. Si les résultats présentent des défauts, il y aura deux possibilités :

1. Le modèle de prédiction des résiliations n'est pas fiable sur l'échantillon de validation du portefeuille à fin 2017 ;
2. Le modèle est valide sur le portefeuille à fin 2017, mais non valide sur le portefeuille à fin 2018, donc il n'est pas stable dans le temps.

Nous noterons que des travaux ont déjà été préalablement menés sur le portefeuille d'Axeria Prévoyance par Mathias VALLA (2018). Son approche première ayant été par les méthodes des séries temporelles, c'est-à-dire en utilisant le cycle des résiliations au cours des 3 premières années de lancement d'un produit, il a ensuite tenté de prédire les comportements individuels, mais la prédiction se concentrait uniquement sur les 5% du portefeuille les plus à risque. C'est pourquoi nous avons souhaité pousser plus loin cette approche en appliquant nos études sur la totalité du portefeuille avec une approche mais aussi des variables différentes.

Chapitre 2. Mise en forme effectuées – Préparation des données

Nos données ont dans un premier temps été traitées directement à partir de l'infocentre via l'outil WPS avant d'être analysées et traitées grâce au logiciel R.

La préparation des données, dans l'analyse, dans la modélisation et dans la projection, est primordiale, pour une raison simple : meilleures sont les données, meilleurs seront les modèles ! Autrement dit, si la base de données est propre, même un algorithme simple peut donner un bon aperçu des données. *A contrario*, des données non préparées peuvent consommer beaucoup de temps ou même fausser le résultat.

Voici les types de « problèmes » qui méritent un traitement :

- Données manquantes ;
- Données dupliquées / superflues ;
- Les valeurs extrêmes, qui peuvent être erronées ou utiles, et qu'il faudra traiter au cas par cas ;
- Données qualitatives avec trop de catégories ;
- Données incohérentes.

D'où viennent les « problèmes » :

- Erreurs de saisie humaines ;
- Fusion de données venant de différents systèmes d'information et donc hétérogènes ;
- Les besoins évoluant dans le temps sans que le système n'ait pu suivre ces évolutions.

Les méthodes utilisées sont les suivantes, en fonction du besoin, et la nature des données :

- Traitement des valeurs manquantes, nous ne pouvons pas tout simplement ignorer la présence de données manquantes car les algorithmes ne les acceptent pas ou les prennent mal en compte. Nous verrons en détail ultérieurement les méthodes utilisées.
- Enlever les observations inattendues, souvent les observations inattendues proviennent des données en duplicata ou la fusion des plusieurs bases de données.
- Traitement des valeurs aberrantes, telles que des majorations supérieures à 50% alors que la plupart des valeurs se trouvent entre 0% et 3%.
- Corriger les erreur structurelles, causées par des erreurs de typo ou incohérence de type majuscule-minuscule.
- Dans la modélisation régression logistique, le niveau associé à chaque modalité d'une variable quantitative est important, car le coefficient est calculé par rapport à la « modalité de référence », il n'y a pas de règle stricte dans le choix de « modalité de référence », mais les modalités peu présentées sont peu choisies dans la préparation des données pour la modélisation.

Nous avons donc pris du temps à bien collecter, nettoyer, regrouper, et enrichir nos données, pour nous assurer de la qualité des données et de leur bonne utilisation par la suite.

L'objectif est donc de garantir dans la mesure du possible l'exactitude et la cohérence des données utilisées.

2.1. *Présentation des données*

Nous avons fait le choix de sélectionner les contrats des deux dernières générations des produits de santé Individuelle AXA France que sont les produits « Référence » et « Ma Santé » (anciennement appelé « Modulango »), car elles représentent plus de 90% du portefeuille global. Ce sont des contrats commercialisés respectivement depuis 2001 et 2012, principalement par nos réseaux propriétaires (agents et réseau salarié), mais également dans une moindre mesure par des courtiers.

Nous avons 5 types de données :

- Celles se rapportant aux états des contrats à la date d'arrêté : dates d'effet, taux de commissionnement, etc. (bases « contrats ») ;
- Celles se rapportant aux états antérieurs des contrats : nombre de remplacements, cotisations des années passées (bases « ACN ») ;
- Celles se rapportant aux assurés : cotisations actuelles, réductions tarifaires, cible de clientèle, etc. (bases « personnes ») ;
- Celles se rapportant à la sinistralité : montant de sinistres réglés dans l'année (bases « sinistres ») ;
- Les données extérieures (bases de l'INSEE, de la DREES ou du fonds CMU notamment).

Ces ensembles de données sont dans des bases différentes qu'il a fallu traiter afin de permettre le rapprochement.

Après nettoyage de la base, récupération des données nécessaires et fusions des bases de données, nous avons notamment appliqué les traitements suivants :

- Suppression des contrats sans assurés dessus ;
- Suppression des contrats sans date d'effet ;
- Anonymisation de la base :
    - Hashage des numéros de contrats et de personne ;
    - Mise au 1er du mois des dates d'effet et de résiliation du contrat.
- Regrouper les niveaux de garantie similaires dans les deux dernières générations de produit santé individuelle, pour réduire le nombre de catégories à analyser associés à cette variable ;
- Identifier les cibles de clientèle dites « pérennes », en fonction de leur régime et leur âge, en « PROS », « SENIORS » et « Autres ».

In fine, nous disposons d'une base de 404 737 contrats en vigueur au 1er janvier 2018 et de 416 632 contrats au 1er janvier 2019.

Notre base possède 38 variables, que nous allons présenter en détail dans la partie *4.1. Analyse exploratoire* :

- 9 variables brutes extraites des différentes sources de données internes ;
- 20 recalculées à partir de variables brutes (les variables brutes n'ayant pas été conservées) ;
- 9 variables externes issues de l'open data.

2.2. *Agrégation / création de nouvelles variables*

Nous avons effectué des agrégations de données afin de créer des variables plus pertinentes. Il s'agit simplement de regrouper des données lorsque celles-ci sont trop inexploitables. Nous nous posons la question « comment peuvent-elles être agrégées de la manière la plus juste et efficace pour que l'algorithme puissent « apprendre » ? » L'objectif d'agrégation est de :

- Réduire les données : plus précisément de réduire le nombre de valeurs différentes ou attributs différentes. Cela réduit ainsi le temps de traitement du modèle. Par exemple avant de regrouper les niveaux de garanties pour les deux gammes « référence » et « ma santé », il existait 25 niveaux. Après regroupement des garanties similaires, nous descendons à 7 niveaux. Cela facilite par ailleurs l'interprétabilité et limite le surapprentissage. Nous pouvons également faire cela de façon arbitraire, lorsque nous estimons qu'une éventuelle séparation de 2 modalités est dû plus à du bruit qu'à une réelle distinction. Ainsi, nous avons regroupé les 2 produits Ma Santé et Modulango, car les produits sont absolument identiques et seul le nom a changé. Par conséquent, nous ne souhaitons pas que le modèle utilise cette distinction de nom comme variable explicative.
- Changement d'échelle : L'agrégation permet de changer l'échelle en donnant une vision de niveau supérieur des données. Par exemple, nous ne souhaitons pas avoir une ancienneté calculée au jour le jour. Par conséquent, nous avons souhaité la paramétrer en ancienneté annuelle.

2.3. *Traitement des données manquantes*

La donnée manquante est toujours une information. Le fait d'être « manquant » est en lui-même une information.

Deux façons courantes de traiter les données manquantes sont :

- Supprimer l'observation qui contient les valeurs manquantes ;
- Estimer la valeur manquante en fonction des valeurs sur les autres observations de la même variable, par exemple la moyenne.

La suppression est une méthode simple, mais qui n'est pas considérée comme une bonne méthode car on supprime ainsi l'information, ce qui n'est pas optimal pour le modèle. En outre, dans la base de validation, nous pouvons aussi avoir les valeurs manquantes, que nous devrons prédire également.

Pour ces raisons, nous créons une nouvelle catégorie « Autres » pour les manquants sur les variables catégorielles.

Concernant les variables quantitatives, nous avons choisi de l'estimer. Très souvent, s'il s'agit d'une variable continue, une moyenne peut être utilisée pour remplacer la valeur manquante.

2.4. *Enrichissement de données externes*

Nous sommes partis du postulat que le comportement du contrat du point de vue de l'assureur ne permet pas totalement d'expliquer les incitations qu'un client peut avoir à quitter son assureur santé.

Nous avons souhaité modéliser l'environnement de soin du client, mais également son environnement macro-économique. Nous avons souhaité ajouter des données de la DREES et de l'INSEE, notamment :

- Taux de chômage par localité (regroupement de communes) : nous avons pris l'hypothèse qu'une région où le chômage est plus important implique une tension budgétaire et donc une régulière mise en concurrence du contrat d'assurance.
- Taux d'ACS par localité : en effet, le bénéfice d'un contrat ACS (Aide à la Complémentaire Santé) est un motif de résiliation du contrat.
- Taux de CMUC par localité : en effet, le bénéfice d'un contrat CMU-C est un motif de résiliation du contrat.
- Nombre de médecins par localité : un client qui a un meilleur accès au soin aura plus intérêt à adapter son contrat à l'offre de soin plutôt que l'inverse.
    - Médecins ;
        - Généraliste ;
        - Spécialiste.
    - Chirurgien-dentiste ;
    - Infirmiers diplômés ;
    - Pharmaciens.

Les variables externes sont choisies sur la base d'une connaissance métier.

2.5. *Contrôle de cohérence*

Les contrôles de cohérence sont effectués pour s'assurer de la qualité des données, notamment sur :

- La fiabilité de la date de souscription : la gamme « référence bien être » est lancé en 2001, et la gamme « Ma Santé » (anciennement « Modulango ») est lancé en 2012, les dates de souscription du contrat doivent être cohérentes par rapport aux lancements des produits complémentaire santé.
- Les montants de primes : vérification si les montants de primes d'assurance sont dans une tranche prédéterminée.
- Nombre de bénéficiaires : vérification si les contrats existent bien dans la base personne, si ce n'est pas le cas, cela veut dire qu'il n'y a aucun assuré associé à ce contrat, il faut donc l'exclure de notre étude.
- Valeurs aberrantes d'indexation : l'indexation a été calculée sur le rapport de primes annuel sur année N et année N-1, nous pouvons donc trouver une indexation trop importante ou trop faible, par rapport à la tranche d'indexation « normal ». S'il s'agit d'une « indexation »

à +10%, cela est probablement expliqué par ajout d'un bénéficiaire du contrat, nous devons donc le « caper » pour ne pas le traiter comme une « indexation » classique. Idem pour une « indexation » à -10%, cela est probablement expliqué par un enfant qui devient majeur et qui part de ce contrat complémentaire santé du parent.

- Code de la commune : vérification de l'existence des codes de la commune.

2.6. *Règlement Général de Protection des Données*

Le respect de la règlementation en matière de données personnelles est un élément clé de l'ensemble des études actuarielles. Le Règlement Général de Protection des Données (RGPD) définit le cadre législatif permettant ou non d'utiliser les données personnelles des assurés.

Nous nous sommes assurés de la bonne information et du recueil du consentement des assurés. Celui-ci est effectué lors de la souscription du contrat.

Les données récoltées sont utilisées uniquement à des fins de gestion ou d'études statistiques. Notre clause type précise expressément l'utilisation des données pour « évaluer votre situation ou la prédire (scores d'appétence) ».

De manière générale, au sein d'AXA France, les piliers RGPD sont scrupuleusement respectés et de nombreux travaux sont en cours dans l'ensemble des lignes de métiers pour continuer à s'assurer de la conformité au règlement :

Consentement : le traitement des données est demandé à la souscription.

Information : les conditions générales ou notices d'informations précisent le contour de l'utilisation.

Finalité & limitation : nous utilisons les données uniquement dans le but pour lequel elles ont été fournies et comme indiqué dans les contrats. Nos données sont anonymisées tous les 3 ans.

Exactitude : Nous demandons à l'ensemble des collaborateurs et partenaires de maintenir à jour les informations dans le cadre des contrats de Santé Individuelle.

Sécurité : Nous sommes régulièrement formés à la sécurisation des données qui sont elles-mêmes sur des serveurs sécurisés avec des habilitations restreintes et mises à jour régulièrement.

2.7. *Anonymisation des données à la norme NPA5*

« La présente Norme de Pratique Actuarielle (ci-après « NPA 5») est une norme professionnelle de catégorie 3 adoptée par l'Institut des actuaires le 16 novembre 2017.

Elle vise à proposer un cadre d'utilisation des données applicables à tous les actuaires membres de l'Institut des actuaires (ci-après « l'actuaire ») face aux enjeux et aux risques liés à l'utilisation des données massives ou comportant des données personnelles ou des données de santé à caractère personnel (ci-après « les données »).

Elle concerne tout « actuaire » amené à utiliser « ces données » dans les cadres, non-limitatifs, d'analyses comportementales, de segmentation et de profilage, ainsi que certains types de tarification. »

INSTITUT DES ACTUAIRES (2017), *Norme de Pratique relative à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel - NPA 5*

Selon « NPA5 », il est préférable d'utiliser les données anonymisées et à défaut rendre impossible de réidentifier des individus. Nous avons ainsi rendu toutes nos données anonymes en supprimant certains indicateurs qui peuvent être considérés comme des données personnelles, par exemple l'adresse, les noms des clients, les dates de naissance, et nous avons également haché les numéros de contrats.

Afin d'éviter des risques de discrimination, des vérifications ont été effectuées sur l'ensemble des données utilisées, elles ne contiennent pas de variables discriminantes directes ou indirectes, par exemple sexe, nationalité, âge des bénéficiaires au contrat, etc. A noter que la modalité « seniors » de la variable « cible » fait référence au statut de retraité.

Aucune donnée de santé n'a été collectée au long de l'étude, conformément aux différents textes de lois, au respect du RGPD, et pour le but de respect du secret professionnel dans le cadre des données de santé à caractère personnel.

Cette approche est restrictive puisque nous aurions pu imaginer utiliser des données de santé pour avoir un modèle mieux adapté. Mais cela aurait apporté beaucoup de contraintes dans l'utilisation et la restitution des données.

Chapitre 3. Analyse exploratoire

Suite à la préparation des données, une première exploration a été nécessaire afin de s'assurer de la direction d'étude prise, de la qualité desdites données et afin d'avoir une première idée de l'impact des variables utilisées.

3.1. *Analyse exploratoire des variables brutes*

Nous présentons ici l'ensemble du portefeuille vu au 1er janvier 2018. C'est le portefeuille sur lequel nos modèles « apprendront » avant d'être testés sur le portefeuille au 1er janvier 2019.

Les variables ont été analysées dans l'ordre d'apparition dans nos bases de données. Cela n'a aucun rapport avec un quelconque classement.

- TCION :

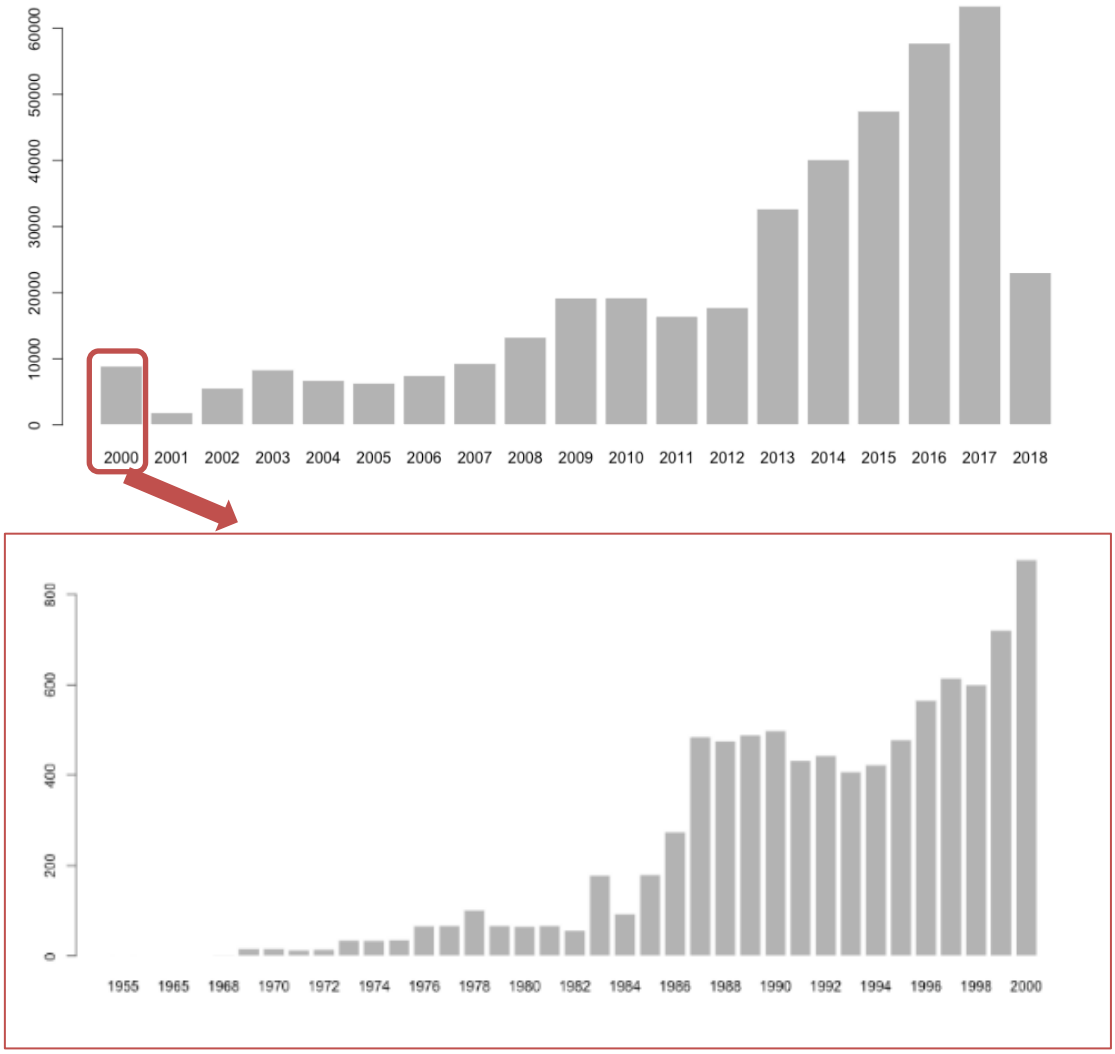
Taux de commissionnement du contrat pour le distributeur. Cette variable pourrait être intéressante si on estime qu'un distributeur mieux rémunéré conserve mieux ses contrats. Ces taux de commissions sont fixés par les mandats des distributeurs et ne varient donc pas d'un client à l'autre, mais en fonction des agents et des types de distributeurs.

	base_2018$RESEAU		
base_2018$TCION	1	2	3
0	0.90	0.43	99.99
2	0.00	0.00	0.00
5	0.00	0.01	0.00
7	0.00	0.03	0.00
8	0.42	0.00	0.00
9	0.12	0.00	0.00
10	35.05	27.20	0.00
11	1.45	0.00	0.00
12	59.19	0.15	0.00
13	0.59	0.00	0.01
14	1.75	0.00	0.00
15	0.51	71.87	0.00
16	0.00	0.32	0.00
20	0.01	0.00	0.00
Sum	100.00	100.00	100.00

Graphique 3 : Répartition des taux de commissionnement par réseaux de distribution

Le graphique 3 ci-dessus montre la répartition des taux de commissionnement par réseau de distribution, environ 60% de contrats souscrits avec les agents généraux (réseau 1) ont un taux de commission à 12%, 35% ont un taux de commissionnement à 10% ; quant aux courtiers (réseau 2) étudiés ici, 72% des courtiers commissionnent à 15%, et le reste à 27% ; le réseau de distribution salariés AXA (réseau 3) a, quant à lui, un système de commissionnement décorrélé des contrats qui n'apparaît donc pas dans les bases de données.

- DTEFFAN : date d'effet des affaires en portefeuille (graphique par année)



Graphique 4 : Nombre d'affaires en portefeuille par année de souscription

Dans le fonctionnement d'un contrat d'assurance, plusieurs dates sont importantes : la date de souscription ou date d'émission, les dates de prise et de fin d'effet c'est-à-dire de couverture, et la date d'émission de la résiliation.

En effet, un assuré qui souhaite se couvrir va signer son contrat chez son distributeur à la date de souscription, mais il va décider de la date à laquelle il souhaite commencer à être couvert, qui peut être le jour même ou un jour futur. Dans de rares cas, la date peut être rétroactive, mais cela reste de l'ordre de l'exception, car l'assureur refuse souvent au vu des risques que cela comporte en termes d'aléa moral. De la même façon, lorsqu'il souhaitera mettre fin à sa couverture, il demandera la résiliation à la date d'émission de la résiliation mais celle-ci prendra effet souvent à une autre date. Par exemple, dans le cas de la résiliation à l'échéance principale, l'émission de la résiliation doit avoir lieu au moins 2 mois avant l'échéance du contrat.

La date d'effet des affaires nouvelles nous permet de connaître l'ancienneté du contrat. La plupart des contrats étudiés sont concentrés sur les 5 dernières années. Le graphique ci-dessus représente la répartition des contrats en portefeuille sur les 2 dernières gammes de produits commercialisées « Référence » et « Ma Santé » qui sont respectivement commercialisées depuis 2001 et 2012.

Néanmoins, nous avons constaté qu'il y a des contrats avec des dates de début antérieures à 2001, c'est parce qu'AXA permet de modifier le contrat, et de basculer d'une gamme à l'autre, tout en gardant la date d'origine de souscription. Nous avons donc des assurés en portefeuille qui étaient sur une gamme antérieure à « Référence », mais qui ont basculé sur « Référence » ou « Ma Santé » en cours de contrat.

Sur les produits AXA, c'est la date d'effet qui détermine la date d'échéance du contrat. Celle-ci n'est pas systématiquement fixée au 1er janvier, mais au 1er du mois de la souscription. Cela implique une saisonnalité moins marquée.

- DTEFFRES :

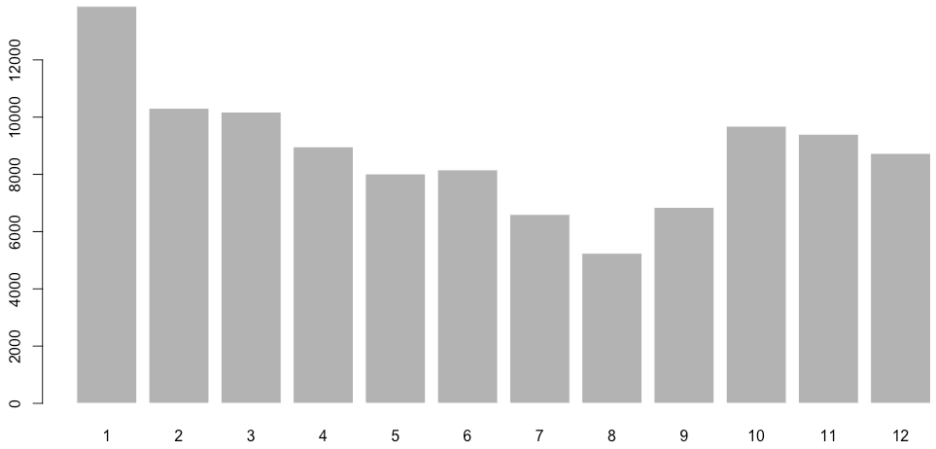
La date d'effet des résiliations indique la date de fin de couverture et donc si le contrat est toujours en vigueur à la date souhaitée. Elle nous sert donc pour regarder le portefeuille effectif au 1er janvier.

- DTEMIRES :

La date d'émission des résiliations nous permet de définir s'il y a eu ou non résiliation dans l'année étudiée et sa saisonnalité.

	N	%
Non-Résiliation	355246	87,8 %
Résiliation	49491	12,2 %

Le taux de résiliations annuel tel que défini en 1.4 est à 12.2% sur le portefeuille présent au 1er janvier 2018.

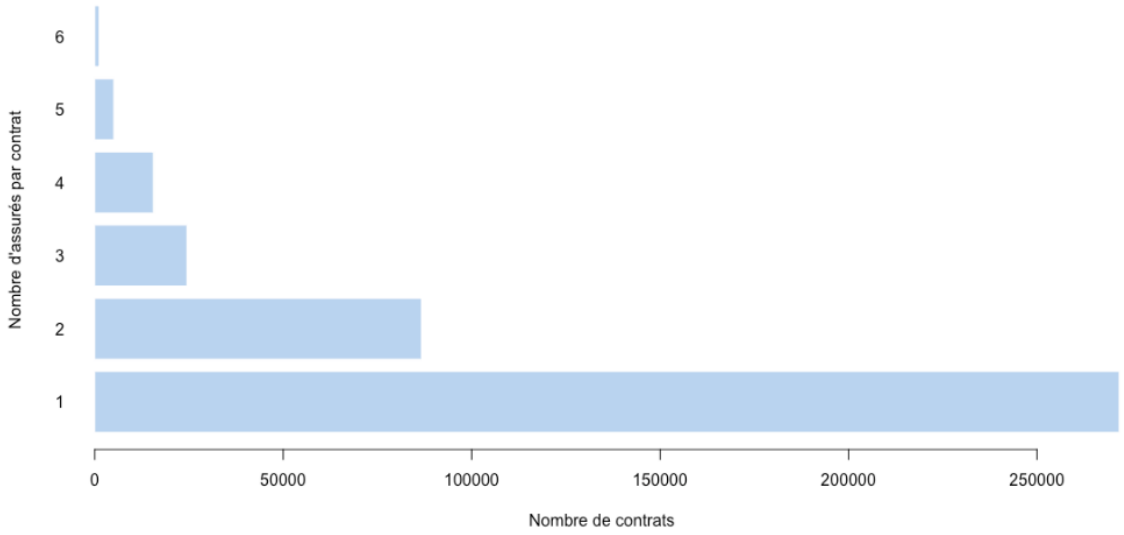


Graphique 5 : Nombre de résiliations par mois

Il y a une saisonnalité des résiliations qui surviennent plutôt au 1er et au 4e trimestre de l'année. C'est cohérent avec la saisonnalité de la souscription et des échéances principales des affaires.

A noter que sur le portefeuille étudié, les dates d'échéances sont réparties sur toute l'année. Sur d'autres portefeuilles qui seraient en échéance au 1er janvier uniquement, les résiliations seraient présentes quasi-uniquement au mois d'octobre pour la résiliation à l'échéance et au mois de janvier dans le cadre de l'application de la loi Chatel.

- NBASSUR : nombre d'assurés attaché à ce contrat



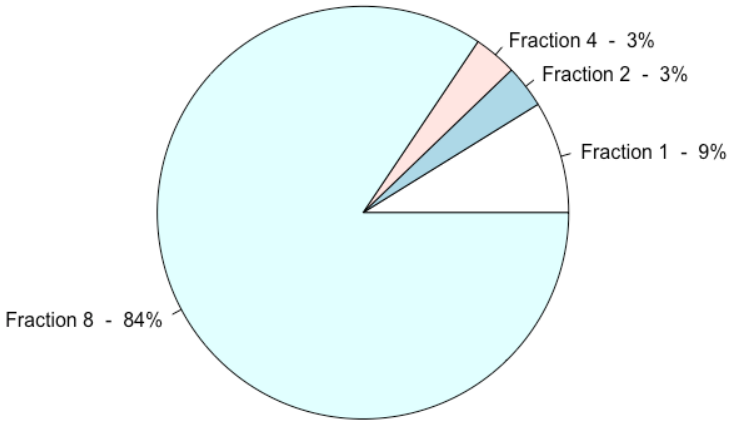
Graphique 6 : Nombre d'ayants-droit par contrat

L'ensemble des membres d'une même famille, appelé ayants-droits, sont acceptés sur les contrats. L'ayant-droit dans presque tous les cas fait partie de la famille de l'assuré. Il peut être le conjoint ou les descendants (jusqu'à 21 ans) ou ascendants qui vivent sous le même toit de l'assuré. La majorité des contrats couvre un seul assuré, mais environ d'un quart des contrats assure des ayants-droits. A noter que pour des contraintes informatiques, le maximum d'assurés associés à un contrat complémentaire santé individuelle est de six dans la base de données étudiée. Lorsque la famille est plus élargie, un second contrat est dressé.

Le nombre moyen d'assurés est de 1,5 bénéficiaires/contrat.

- FRAC :

Fractionnement de la prime, annuel, semestriel ou trimestriel :



Graphique 7 : Répartition du portefeuille par fractionnement de la prime

Il y a 4 modalités possibles associées à cette variable, 1 signifie prime annuelle, 2 signifie prime semi-annuelle, 4 signifie prime trimestrielle, et 8 signifie prime mensuelle. La fraction mensuelle des primes représente la majorité dans notre base de données étudiée.

Les fractionnements annuels, semestriels et trimestriels peuvent être par prélèvement ou par paiement manuel. En revanche, les contrats mensuels sont uniquement par prélèvement.

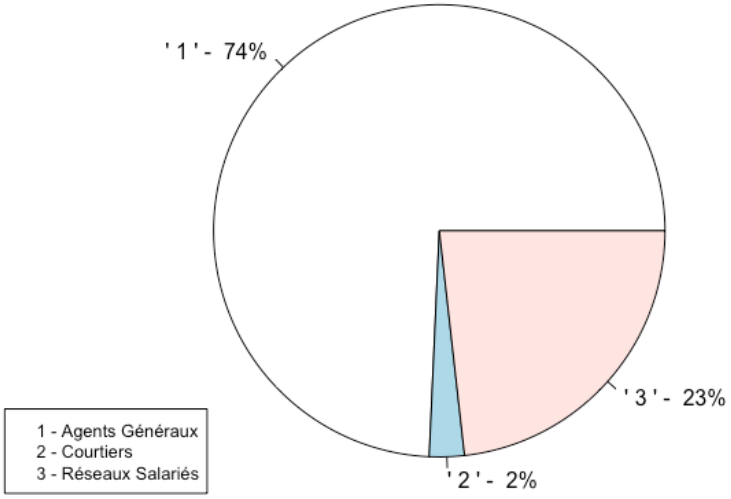
Une hypothèse que le modèle pourrait vérifier : les contrats qui sont en prélèvement mensuel résilieraient moins, car le prélèvement est plus « indolore » que le paiement manuel.

- MTPCPTE :

Ce montant représente, le précompte, c'est-à-dire, la commission de la 1ère année, lorsqu'elle est supérieure au commissionnement linéaire annuel. En effet, les agents généraux peuvent opter pour une méthode de sur-commissionnement la première année avec une baisse du commissionnement récurrent. Cette méthode est un incitatif à produire. Inversement, cette variable pourrait avoir tout son sens puisque cela peut inciter à faire de la « mauvaise production » et moins encourager le distributeur à conserver son portefeuille.

- RESEAU :

Réseau de distribution : agents généraux, courtiers et réseaux salariés



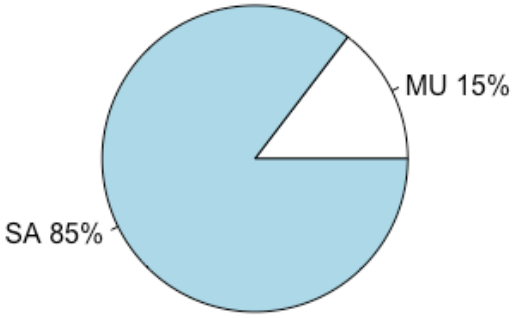
Graphique 8 : Répartition du portefeuille par réseaux de distribution

Une grande partie du portefeuille étudiée dans le cadre de ce mémoire est celle commercialisée par les réseaux dits « propriétaires », c'est-à-dire les réseaux salarié et d'agents généraux et une faible part du portefeuille est issue du courtage. Le réseau du courtage est incité à vendre une autre gamme de produit AXA France qui n'est pas étudiée. Néanmoins, le réseau de plus de 4000 agents généraux a une forte contribution sur l'ensemble du portefeuille avec près des ¾ des contrats en cours.

Réseau de distribution	Taux de résiliation 2018
Tous réseaux	12,2%
Agents Généraux AXA	12,4%
Courtiers	15,5%
Réseau salarié AXA	11,3%

Nous avons constaté que les comportements des résiliations varient d'un réseau à l'autre : le réseau salariés AXA est relativement faible en résiliation par rapport aux 2 autres réseaux, 11,3% versus un taux de résiliation global à 12,2% ; le réseau courtage a un taux de résiliation à 15,5%, qui est le plus élevé. Nous étudions par la suite si le réseau de distribution est une variable significative dans la modélisation.

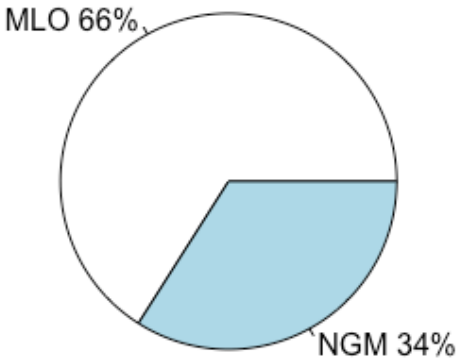
- Société : SA (AXA France Vie) et MU (AXA Assurances Vie Mutuelle)



Graphique 9 : Répartition du portefeuille par les sociétés

La majorité des contrats complémentaire santé individuelle sont souscrits chez AXA France Vie. On rappelle ici, que les assurés ne choisissent pas s'ils souhaitent être assurés par AXA France Vie ou AXA Assurances Vie Mutuelles. La répartition se fait en fonction des portefeuilles des agents, présentés dans la partie *1.1.3 L'assurance Santé chez AXA*.

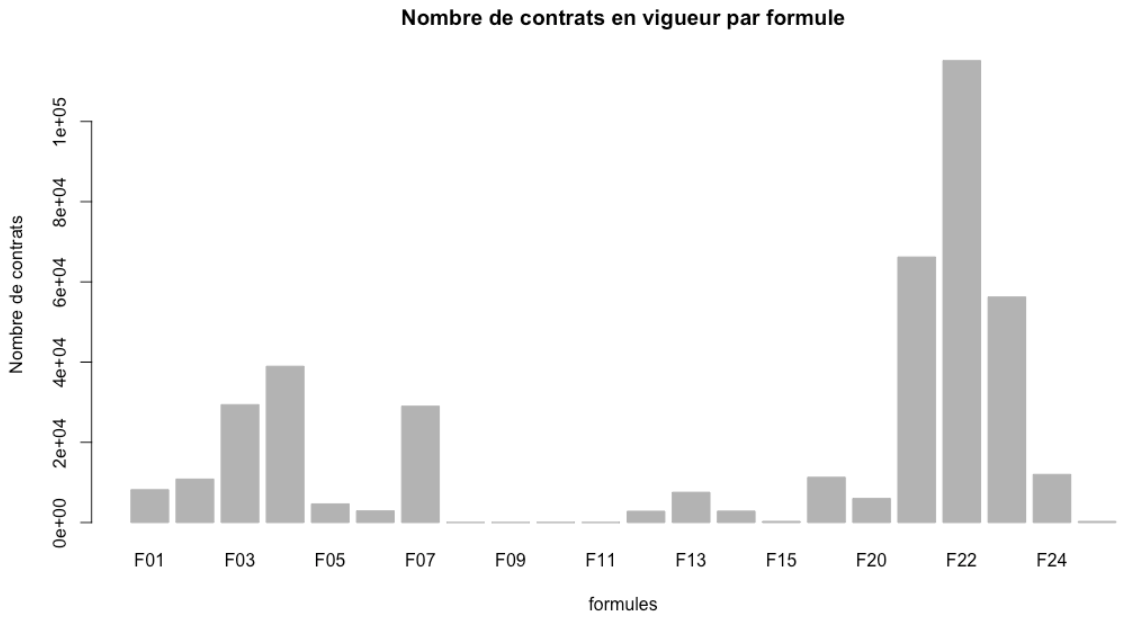
- Caractéristiques sur les garanties :
  - Gamme



Graphique 10 : Répartition du portefeuille par les gammes des produits

La gamme « Ma Santé » (MLO) représente deux tiers de l'ensemble des contrats étudiés. Elle va continuer à prendre plus de poids avec le temps car la gamme « Référence » (NGM) n'est plus commercialisée.

- Formule



Graphique 11 : Nombre de contrats en vigueur par formules

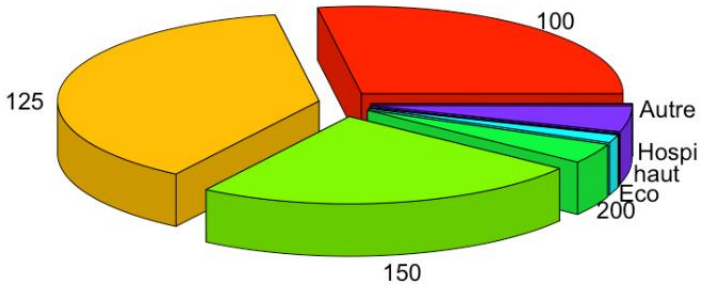
La variable Formule correspond à 25 formules de garanties couvertes par le contrat santé individuelle commercialisées. Les 15 premières correspondent à la gamme « Référence », et les formules 16 à 25 sont les formules de la gamme « Ma Santé ».

Les 25 formules ne sont pas évidentes à exploiter, car certaines sont similaires entre les deux gammes et d'autres non.

Nous avons donc décidé de les retraiter en regroupant les niveaux de garanties similaires, nous avons ainsi obtenu 7 niveaux de garanties pour faciliter la modélisation.

Par ailleurs, il y a 142 valeurs manquantes sur cette variable, nous avons créé une valeur « Autre » pour remplir ces valeurs manquantes afin d'éviter de perdre ces observations.

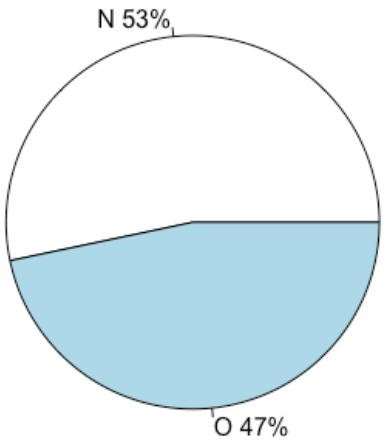
- Niveau de garantie



Graphique 12 : Répartition du portefeuille par niveaux de garantie

Les niveaux de garanties souscrits sont plutôt concentrés sur les garanties de « cœur de gamme » : 100, 125 et 150. Les niveaux d'entrée de gamme « Hospi » et « Eco » et les niveaux hauts de gamme « 200 » et « haut » sont marginaux.

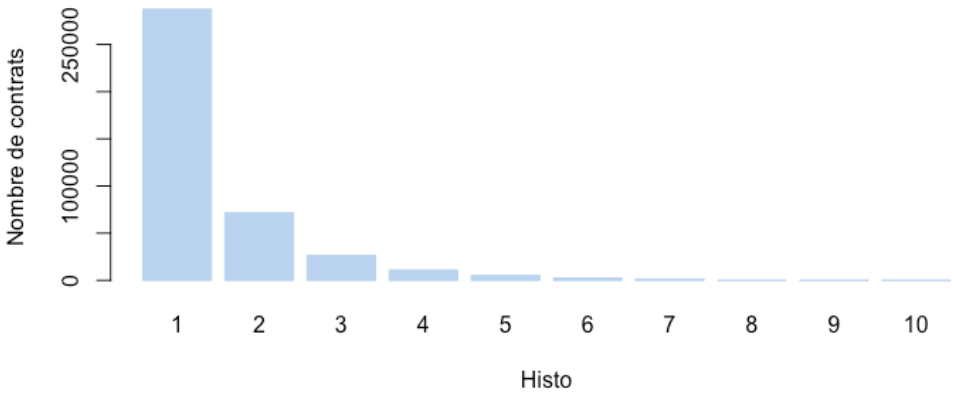
- Renfort



Graphique 13 : Répartition du portefeuille - avec ou sans renfort

La gamme Référence propose un seul renfort de garanties facultatif. Ma Santé en propose trois. Nous avons souhaité utiliser la variable sous la forme d'un top « Le contrat a au moins un renfort » afin de décorréler la réponse de la gamme. Parmi les 400 milliers de contrats étudiés, presque la moitié des clients ont adopté une option renfort.

- Nombre de changements totaux



Graphique 14 : Nombre de contrats en vigueur par changements totaux

Les changements consistent à identifier toutes modifications d'un même contrat, sans résiliation. Dans le jargon, nous appelons ces changements des « histos » car la situation précédente est historisée. Par exemple :

- ajout ou enlèvement d'un assuré ;
- revue à la hausse ou à la baisse du niveau de couverture complémentaire santé ;
- changement de l'adresse ;

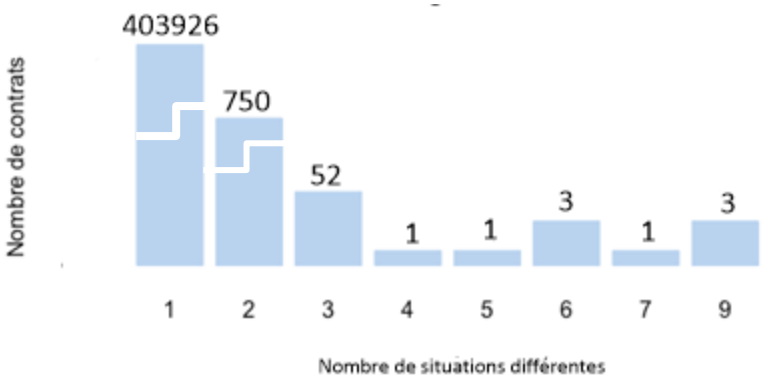
- etc.

Nos suppositions sur cette variable vont dans les 2 sens :

- Nous supposons d'un côté qu'un contrat qui « bouge » beaucoup est plus à même de partir.
- D'un autre côté, s'il bouge, c'est que nous avons répondu à son besoin, donc peut-être qu'il n'a plus ce besoin de se tourner vers la concurrence.

Nos modèles pourront éventuellement vérifier ces hypothèses.

▪ Nombre de changements au cours des 12 mois précédents



Graphique 15 : Nombre de contrats en vigueur par changements dans l'année

Le nombre de modifications d'un contrat au cours de 12 derniers mois a été collecté pour voir sa significativité. On constate que c'est très rare d'effectuer plus d'une modification dans l'année.

▪ AN3 :

Cet indicateur identifie les contrats qui ont été issue d'une résiliation – affaire nouvelle au lieu d'une modification du contrat initial.

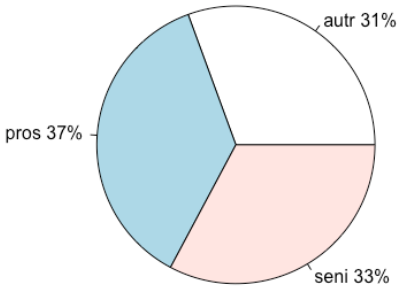
Cette situation est effectuée lorsque le remplacement n'est pas possible et que le distributeur est obligé de recréer un contrat. Lorsque le distributeur saisit l'affaire, il indique comme « origine de l'affaire nouvelle » la 3ème modalité proposée « Affaire issue du portefeuille » d'où le nom de la clause AN3.

▪ Localité

Cette variable représente le code INSEE de la commune de l'assuré. Elle a été utilisée afin de pouvoir lier des variables externes, mais nous ne l'utilisons pas comme tel dans nos modèles.

3.2. *Variables recalculées*

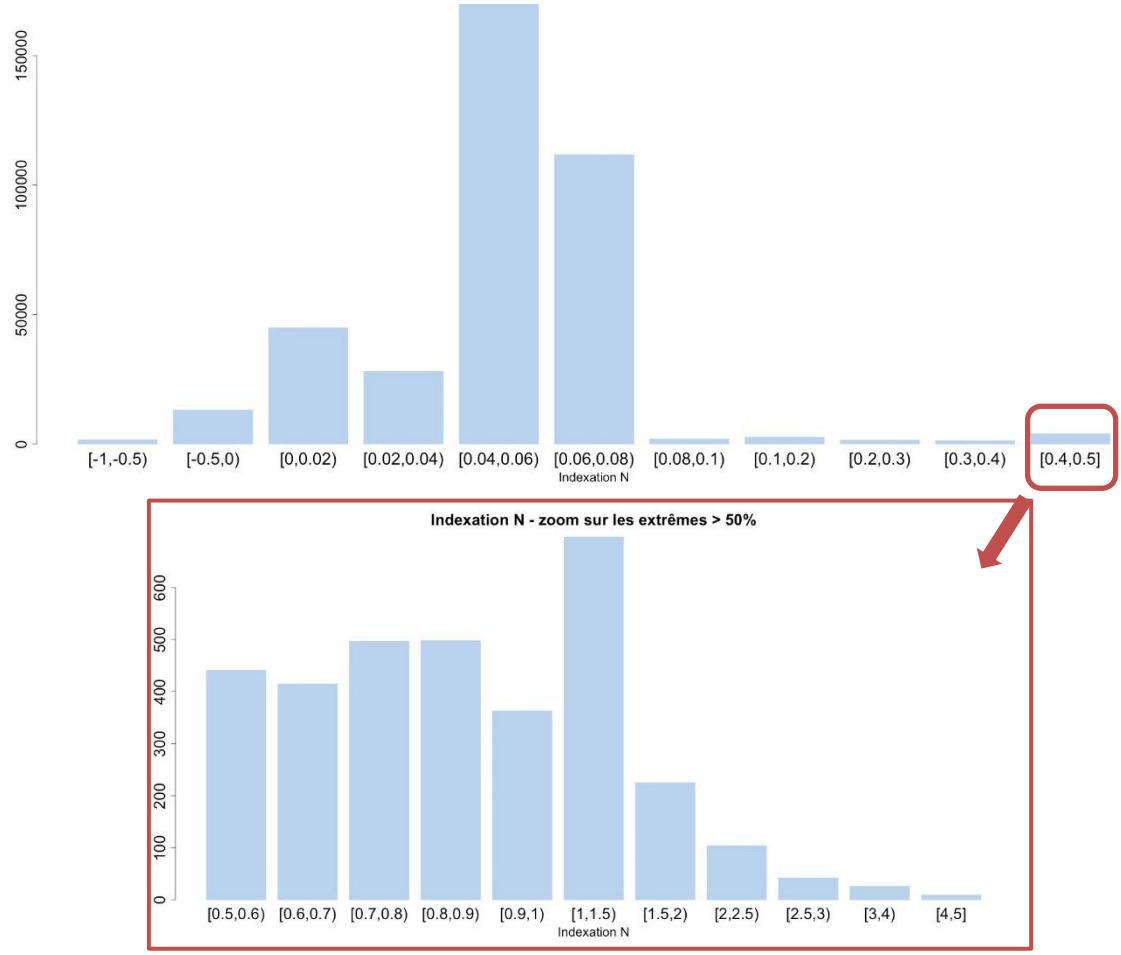
▪ Cible : cibles de clientèle



Graphique 16 : Répartition du portefeuille par cibles de clientèle

Nous avons classé les cibles de clientèle dites « pérennes » (au sens décrit dans le paragraphe *1.1.2. Les complémentaires santé*), en fonction de leur régime et leur âge, en « PROS » - Travailleurs non-salariés agricoles ou non, « SENIORS » - retraités et « Autres » - qui se compose principalement des affiliés au régime général.

- Niveau d'indexation
  - Index_N :

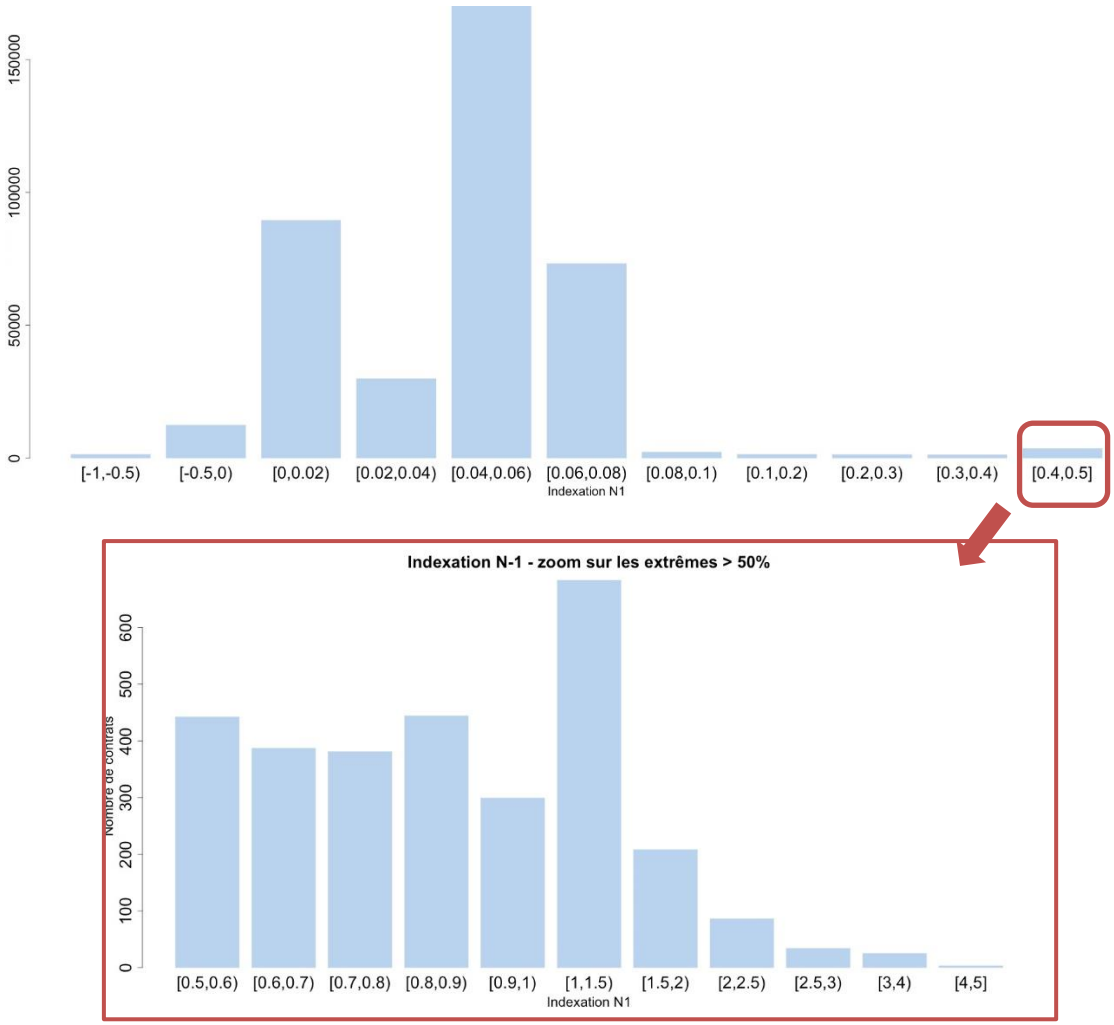


Graphique 17 : Répartition des contrats par niveaux d'indexation N

L'indexation N est calculée sur le rapport de primes annuel d'année N et année N-1. Pour calculer cet indicateur, nous avons souhaité aborder cette indexation du point de vue du client et non d'un point de vue « technique », c'est-à-dire que nous regardons la prime TTC, comprenant l'ensemble

des éventuelles réductions tarifaires à laquelle il a le droit d'une année sur l'autre, les éventuels mois gratuits, etc.

- Index_N-1 :



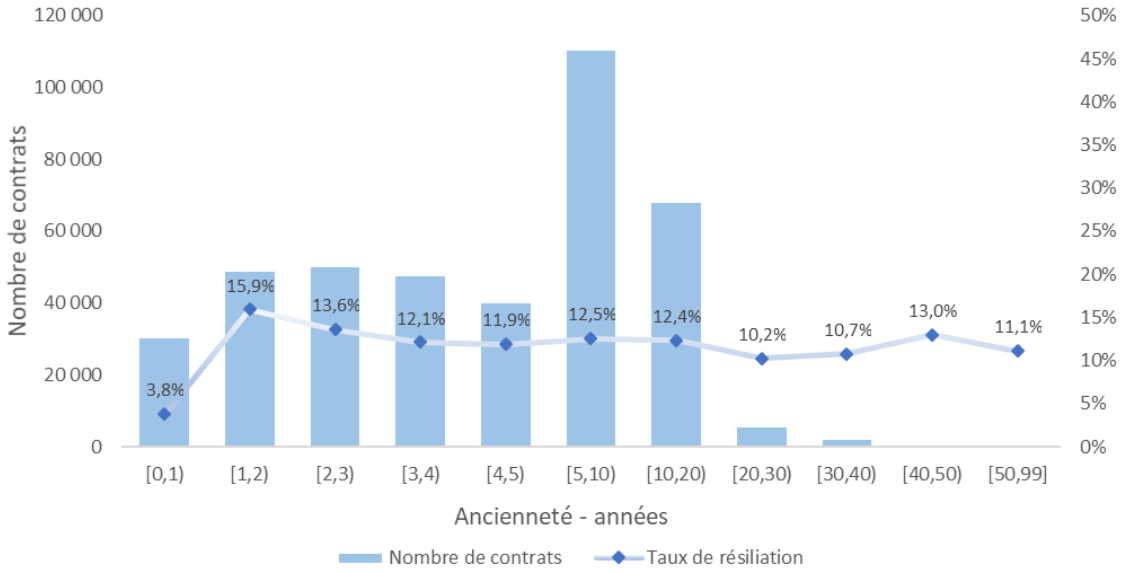
Graphique 18 : Répartition des contrats par niveaux d'indexation N-1

L'indexation N-1 est calculée sur le rapport de primes annuel sur année N-1 et année N-2.

Nous pouvons donc calculer une indexation trop importante ou trop faible, par rapport à la tranche d'indexation « normal ». S'il s'agit d'une « indexation » supérieure à +50%, cela est probablement expliqué par ajout d'un bénéficiaire du contrat, nous devons donc le « caper » pour ne pas le traiter comme une « indexation » classique. Idem pour une « indexation » à -10%, cela est probablement expliqué par un enfant qui devient majeur et qui quitte le contrat complémentaire santé du parent, ou par la radiation d'un bénéficiaire. Nous avons ainsi décidé de « caper » le maximum et le minimum pour obtenir les « vraies » indexations en excluant les « évolutions sur les primes annuelles anormales ».

Dans le cadre ou la valeur est manquant, la seule possibilité est que l'indicateur primes annuelle d'année précédente est manquant, donc c'est un contrat qui est nouvellement souscrit dans l'année. C'est donc logique d'imposer la valeur manquant à 0.

- Ancienneté du contrat :



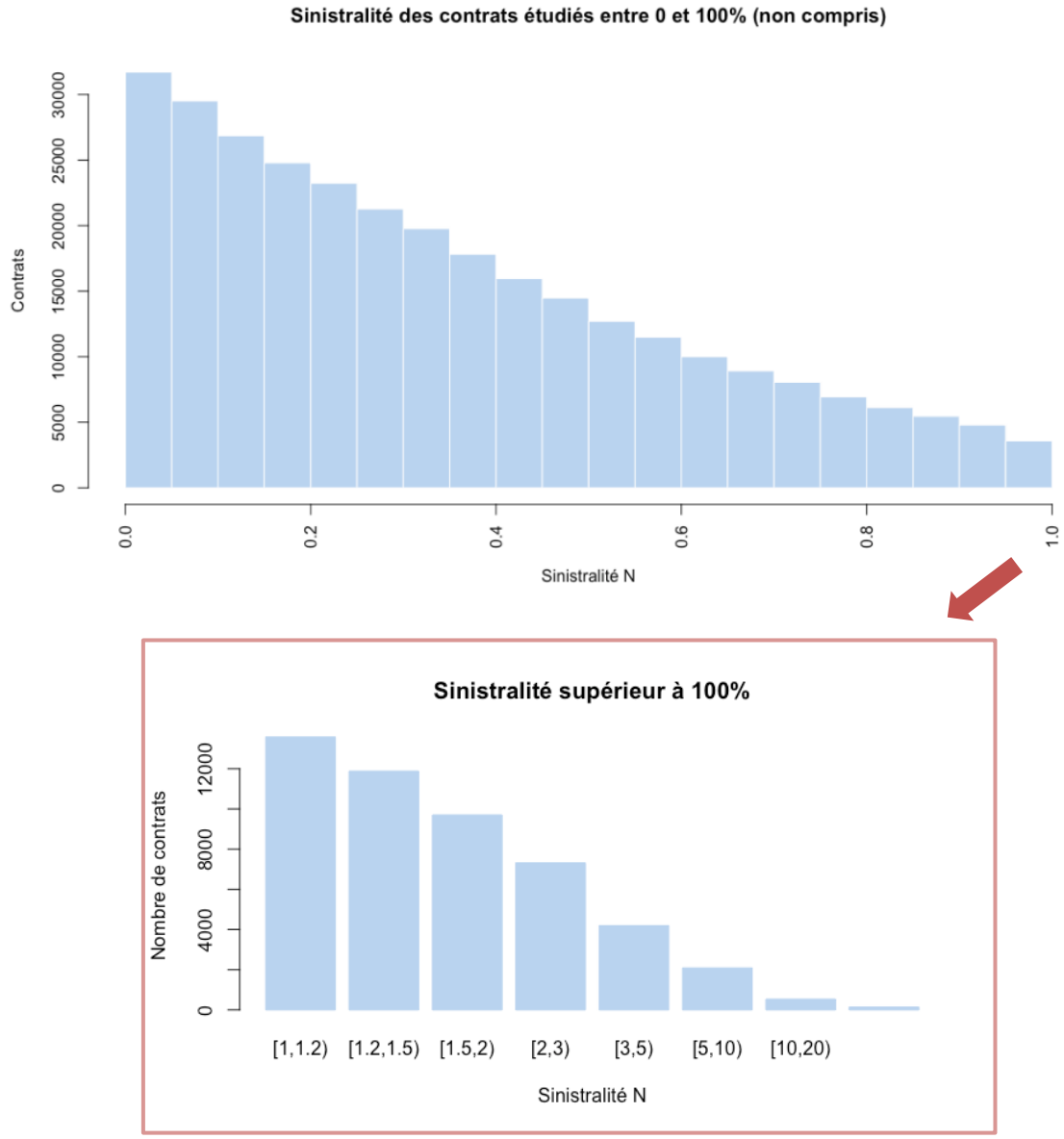
Graphique 19 : Nombre de contrats par ancienneté

La plupart des contrats ont une ancienneté inférieure à 5 ans et une grande partie la quasi-totalité a une ancienneté inférieure à 20 ans. Or, selon l'indication donnée en *2.1. Problématique*, plus longtemps le contact reste, plus il est rentable. Nous avons donc intérêt à conserver le contrat dans portefeuille le plus longtemps possible.

L'ancienneté du contrat n'est pas une information dans notre système de gestion. Elle a été calculée comme la différence entre la date d'arrêtée de la base de données et la date du début des contrats, en nombre des jours, nous avons par la suite regroupé en années d'ancienneté.

- Rentabilité :

- SCN_N : S/C de l'année N



Graphique 20 : S/C N des contrats

Le S/C représente le ratio de montant de sinistres sur les cotisations perçues.

$$S/C = \frac{Sinistres\ réglés\ par\ la\ compagnie\ sur\ la\ période}{Cotisations\ perçues\ au\ titre\ de\ la\ période}.$$

Il représente une vision technique. Pour avoir un aperçu de la rentabilité, du contrat, il faut ajouter les frais de la compagnie au numérateur pour avoir un ratio combiné. Si le ratio combiné est inférieur à 100%, le contrat (ou le portefeuille) est techniquement rentable. S'il est supérieur, alors le contrat est techniquement déficitaire.

A noter que la sinistralité contrat par contrat est utilisée ici dans le cadre d'un scoring, toutefois la Loi Evin de 1989 indique bien dans l'article 6 que nous ne pouvons utiliser cette information pour majorer un contrat spécifique :

> *« L'organisme ne peut ultérieurement augmenter le tarif d'un assuré ou d'un adhérent en se fondant sur l'évolution de l'état de santé de celui-ci.*
>
> *Si l'organisme veut majorer les tarifs d'un type de garantie ou de contrat, la hausse doit être uniforme pour l'ensemble des assurés ou adhérents souscrivant ce type de garantie ou de contrat. »*

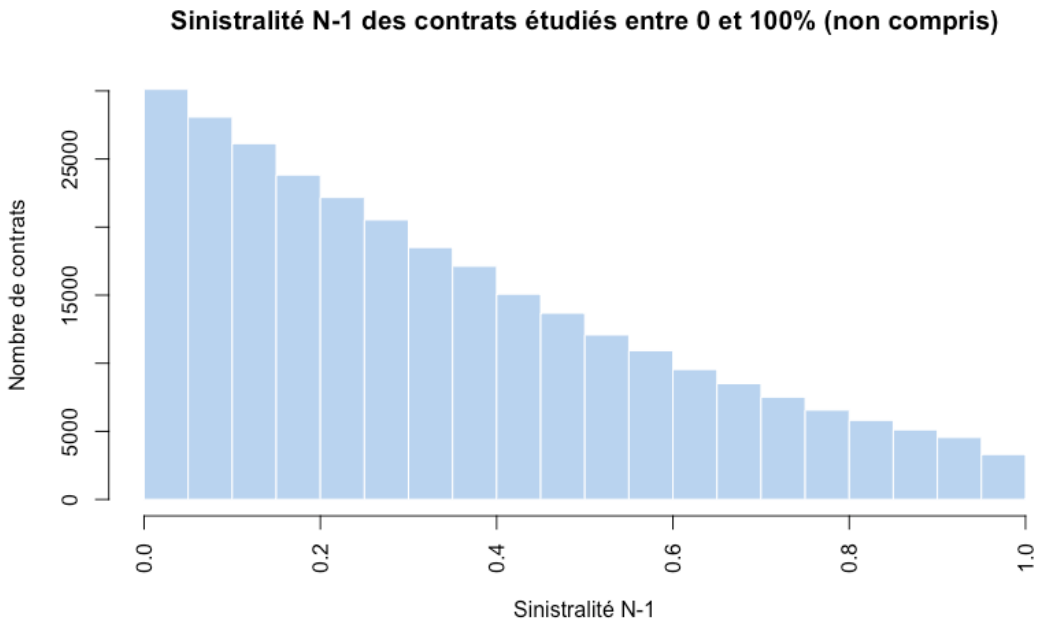
Par ailleurs, ce même article interdit à l'assureur de résilier le contrat tant que l'assuré paie sa prime. Ainsi, à la différence des contrats IARD, l'assurance santé est contrainte de conserver un client « risqué » et en cas de portefeuille déficitaire, l'effort doit être porté par l'ensemble des membres du groupe homogène.

Dans le portefeuille étudié, 51 854 contrats ont une sinistralité nulle, 49 418 contrats ont une sinistralité supérieure à 100%, un zoom a donc été fait sur ces contrats.

Quand la valeur est manquante, la moyenne sur les autres observations est calculée pour remplacer la valeur manquante. La majorité des contrats dans la base étudiée ont une sinistralité faible.

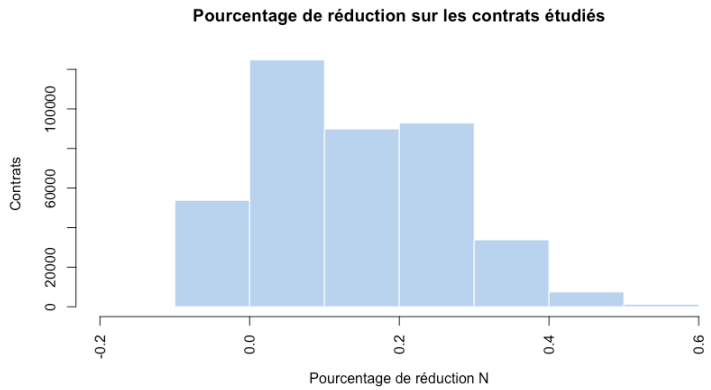
- SCN_N1 : S/C N-1

La S/C de l'année précédente est également calculée sur le même principe.



Graphique 21 : S/C N-1 des contrats

- Niveau de réduction :



Graphique 22 : Pourcentage de réduction sur les contrats étudiés

Le niveau de réduction s'exprime en pourcentage par rapport à la prime, il prend en compte les réductions liées aux contrats s'il en a bénéficié lors de la mise en place par le biais d'une offre, mais également, les éventuelles réductions liées aux bénéficiaires (conjoints assurés, 3ème enfant assuré au contrat gratuitement, etc.). Nous supposons qu'un contrat qui dispose d'avantages tarifaires est moins prédisposé à résilier.

- w_dept

Nous avons choisi de conserver la localisation du contrat, mais uniquement sur la base du département et non du code commun.

- Motif de la résiliation :

Cette valeur existe dans la base de données, mais elle n'est pas renseignée partout, en outre les informations renseignées ne sont pas fiables. Nous n'avons donc pas exploité cette variable. Des chantiers sont en cours afin de fiabiliser cette donnée, notamment en réduisant le nombre de motifs et en rendant obligatoire le remplissage lors de la saisie de résiliation.

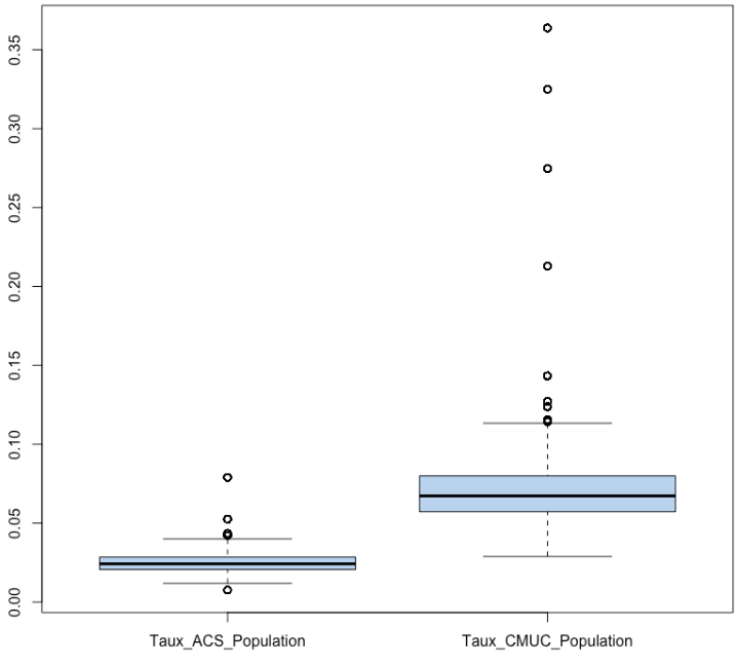
3.3. *Variables externes*

Nous avons pour hypothèse qu'une part de la décision de la résiliation provient de données non directement liées au contrat. C'est pourquoi nous avons souhaité insérer dans notre modèle des données venant de sources extérieures à nos bases initiales. Notamment en utilisant les données ouvertes (ou « open data »).

- Nombre de chômeurs par localité

Les informations sont obtenues sur le site INSEE.fr. Nous supposons ainsi que les contrats dans des zones économiquement plus tendues peuvent avoir un comportement différent que ceux des zones économiquement plus dynamiques. Nous avons également un plus fort risque que les assurés résilient pour bénéficier de l'ACS ou de la CMU-C.

- Nombre d'ACS et CMUC :



Graphique 23 : Taux de ACS et CMUC

Les dispositifs ACS et CMU-C permettaient jusqu'en novembre 2020 de fournir une complémentaire santé gratuite (CMU-C) ou à faible coût (ACS) aux foyers à faibles ressources. En cas de bénéfice de ce dispositif, un assuré est en droit de résilier son contrat individuel, d'où l'impact envisagé de cette variable externe. Cela peut également être un cas de dispense au contrat collectif obligatoire, mais nous ne nous étendrons pas sur le sujet.

En France en 2018, en moyenne il y 2,5% de ACS, et 7,1% de CMUC.

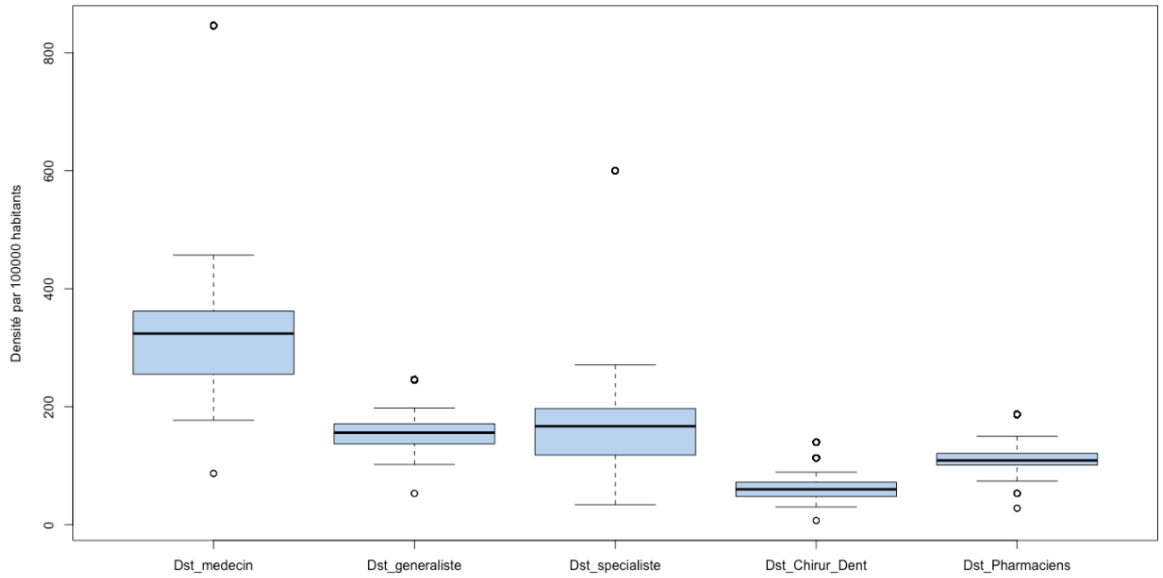
- Nombre de médecins par localité :

Nous estimons que les assurés des zones ayant un meilleur recours au soin peuvent avoir un comportement différent des autres. Par exemple, parce qu'ils pourraient souhaiter optimiser leur contrat en fonction de cette offre de soins.

Nous avons donc récupéré les densités par localité des professionnels de santé suivants :

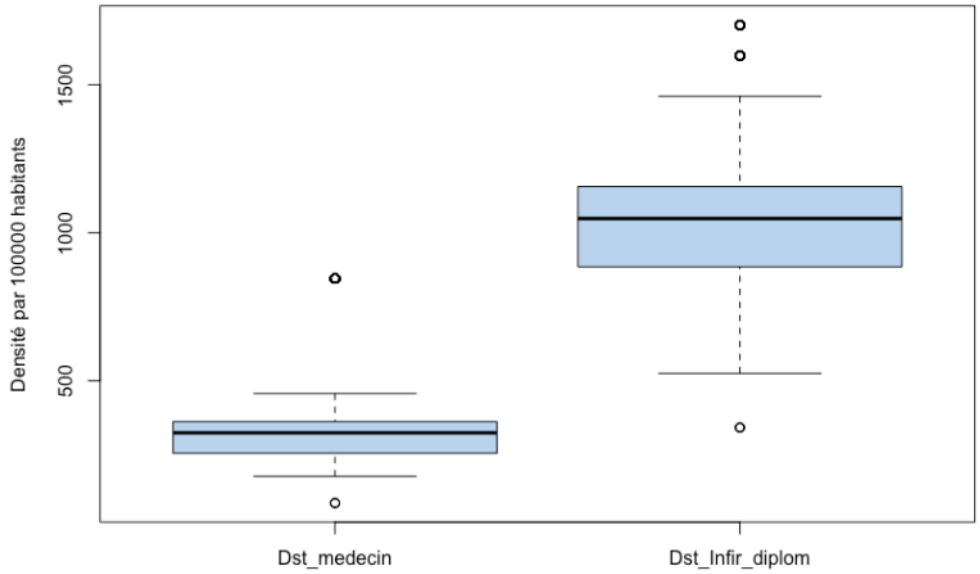
- Médecins :
  - Généraliste ;
  - Spécialiste.
- Chirurgien-dentiste ;

- Infirmiers diplômés ;
- Pharmaciens.



Graphique 24 : Densité des médecins (généraliste et spécialiste) par 100 000 habitants

En 2018 en France, pour tous les 100 milles habitants, en moyenne il y a 154 médecins généralistes, 174 médecins spécialistes, un peu moins de chirurgiens-dentistes à 62, 112 pharmaciens, et 1 036 infirmières diplômés.



Graphique 25 : Densité des médecins versus infirmières par 100 000 habitants

3.4. *Test sur la corrélation des variables*

Le test de corrélation est une mesure de la liaison entre deux variables. La valeur du coefficient de corrélation varie entre -1 et +1. Le 1 indique que les deux variables sont parfaitement liées, et la ligne + signifie une liaison positive et - signifie une liaison négative ; quand la valeur s'approche de 0, la relation de liaison diminue, et 0 indique une absence totale de liaison. Normalement si l'absolu

du coefficient de corrélation est entre [0.1, 0.3), on considère que la liaison est petite, s'il est entre [0.3, 0.5), cela peut être une liaison moyenne, au-delà de 0.5 signifie une liaison forte.

La table de corrélation (cf. annexe 3) montre que le taux de commissionnement (*TCION*) est très fortement corrélé au Réseau Agent (*RESEAU1*) et au Réseau Salariés (*RESEAU3*), nous avons déjà parlé de ce phénomène précédemment quand on présentait les variables dans la partie *4.1.1. Analyse des variables brutes.* La variable *gamme* est fortement corrélée avec la variable Option Renfort (*w_renfort*) et la variable ancienneté(*r_ancien_m*). Les autres variables n'ont pas de liaisons fortes entre elles.

Ce test permet de mieux connaitre les relations entre les variables explicatives et nous n'avons pas constaté de points bloquants pour la modélisation par la suite.

Chapitre 4. Modélisation prédictive

Nous avons souhaité comparer deux méthodes de modélisation. Nous rappelons ici que nous avons 2 objectifs à remplir : le premier étant d'avoir une bonne capacité de prédiction des résiliations et la seconde étant la recherche d'un modèle stable dans le temps.

4.1. *Modèle Classique : GLM - Régression logistique :*

4.1.1. Mise en place de la régression

4.1.1.1. Théorie de la régression logistique

Nous cherchons à prédire si un contrat sera résilié ou pas, donc une variable Y catégorielle à 2 modalités : 1 (représente un contrat résilié durant l'année) ou 0 (représente un contrat en vigueur en fin d'année) est attendue ; de plus les variables explicatives sont catégorielles ou quantitatives, c'est l'objet des méthodes de régression logistique, présenté notamment par Ricco RAKOTOMALA (2017), dont la formule s'écrit comme suit :

$$\boldsymbol{logit(P(Y=1|X)) = ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.}$$

$$\begin{aligned} &\textit{avec} \quad \pi(X) = P(Y=1|X). \\ &\textit{et} \quad 1-\pi(X) = P(Y=0|X). \end{aligned}$$

La formule peut également s'écrire comme suit :

$$\pi(X) = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}.$$

Et $\frac{\pi(X)}{1-\pi(X)} = \frac{P(Y=1|X)}{P(Y=0|X)}$ représente une cote (« Odds » en anglais). La notion de cote s'introduit naturellement quand on voit la formule $\frac{p}{1-p}$. La cote s'interprète comme un rapport rapportant l'événement réalisé (avec la probabilité p) sur l'événement qui ne se réalise pas (avec la probabilité $1-p$). Par exemple, si la cote est égale à 4 (⇔ P=0,8), l'individu a 4 fois plus de chance de gagner que de perdre.

L'application du logarithme sur les produits des cotes permet de faire apparaître une somme, mais surtout il permet de recentrer $ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$ autour de 0.

	S1 (central)	S2 (+)	S3 (-)
P	0,5	0,6	0,4
$1-P$	0,5	0,4	0,6
Cotes $\frac{P}{1-P}$	1,0	1,50	0,67
Log(cotes) $ln\left(\frac{P}{1-P}\right)$	0,0	0,18	-0,18

Tableau 1 : Logarithme des cotes

Et en ce qui concerne le rapport des cotes, ou en anglais l'« Odds ratio », c'est une comparaison des cotes sur deux phénomènes. Le rapport des cotes montre s'il y a une différence forte ou faible entre 2 groupes. Exemple, si un autre groupe a une cote égale à 2, alors le rapport des cotes est égal à 2 (4/2), ainsi le premier groupe a une plus grande probabilité de se réaliser.

Nous pouvons donc avoir très facilement les propriétés ci-dessous :

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \in (-\infty; +\infty).$$

$0 < \pi(X) < 1$ (Puisque c'est une probabilité)

Et les relations équivalentes :

$$\frac{\pi(X)}{1-\pi(X)} > 1 \leftrightarrow \frac{P(Y=1|X)}{P(Y=0|X)} > 1 \rightarrow \hat{Y} = 1.$$

$$\pi(X) > 0.5 \rightarrow \hat{Y} = 1.$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k > 0 \rightarrow \hat{Y} = 1.$$

La fonction *glm* de R permet de dresser directement le modèle.

4.1.1.2. Application sur nos données

Un premier résultat du modèle ci-dessous avec l'ensemble de variables explicatives est présenté par un tableau de coefficients.

$$\begin{aligned} r_{Resil} \sim\ & RESEAU{:}TCION + NBASSUR + FRAC + MTCPTE + RESEAU + societe + gamme \\ & + an3_{tech} + w_{renfort} + w\_dept + n\_histo\_tot + n\_histo\_annee + w\_cible \\ & + w\_index\_N + w\_index\_N1 + w\_chomage\_N + Taux\_ACS\_Population \\ & + Taux\_CMUC\_Population + Dst\_medecin + Dst\_generaliste \\ & + Dst\_specialiste + Dst\_Chirur\_Dent + Dst\_Infir\_diplom \\ & + Dst\_Pharmaciens + SCN\_N + SCN\_N1 + pctreductot + niveau \\ & + r\_ancien\_m. \end{aligned}$$

Les coefficients sont donnés par ce tableau ci-dessous :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.770e+02	5.978e+05	-0.001	0.999497	
NBASSUR	1.833e-02	6.805e-03	2.694	0.007068	**
FRAC2	-4.498e-01	3.962e-02	-11.353	< 2e-16	***
FRAC4	-2.063e-01	3.531e-02	-5.843	5.12e-09	***
FRAC8	-4.865e-01	1.976e-02	-24.624	< 2e-16	***
MTPCPTE	-6.908e-04	6.059e-05	-11.402	< 2e-16	***
RESEAU2	7.520e-01	1.902e-01	3.954	7.68e-05	***
RESEAU3	-1.204e-01	6.104e-02	-1.973	0.048531	*
RESEAUAutre	-1.953e+01	1.371e+04	-0.001	0.998864	
societeSA	2.897e-02	1.786e-02	1.622	0.104848	
gammeNGM	1.657e-01	2.263e-02	7.323	2.42e-13	***
an3_techO	-4.206e-01	1.692e-02	-24.864	< 2e-16	***
w_renfortO	-1.199e-01	1.494e-02	-8.026	1.01e-15	***
w_dept01	1.658e+02	2.415e+05	0.001	0.999452	

w_dept02	5.325e+01	8.125e+04	0.001	0.999477	
w_dept03	-2.833e+01	4.700e+04	-0.001	0.999519	
w_dept04	4.764e+01	6.945e+04	0.001	0.999453	
w_dept05	1.607e+02	2.364e+05	0.001	0.999458	
[…]					
w_dept99	9.220e-01	1.331e+00	0.693	0.488575	
n_histo_tot	2.637e-02	6.808e-03	3.874	0.000107	***
n_histo_annee	-2.326e+01	7.523e+03	-0.003	0.997533	
w_ciblepros	-6.798e-01	1.562e-02	-43.508	< 2e-16	***
w_cibleseni	-6.315e-01	1.576e-02	-40.074	< 2e-16	***
w_index_N	-1.315e+00	6.262e-02	-21.005	< 2e-16	***
w_index_N1	-8.153e-02	7.147e-02	-1.141	0.253991	
w_chomage_N	4.166e-03	6.254e-03	0.666	0.505353	
Taux_ACS_Population	-4.870e+03	8.627e+06	-0.001	0.999550	
Taux_CMUC_Population	7.297e+03	1.066e+07	0.001	0.999454	
Dst_medecin	NA	NA	NA	NA	
Dst_generaliste	NA	NA	NA	NA	
Dst_specialiste	NA	NA	NA	NA	
Dst_Chirur_Dent	NA	NA	NA	NA	
Dst_Infir_diplom	NA	NA	NA	NA	
Dst_Pharmaciens	NA	NA	NA	NA	
SCN_N	1.688e-01	5.122e-03	32.949	< 2e-16	***
SCN_N1	-1.604e-02	6.651e-03	-2.411	0.015890	*
pctreductot	-1.438e+00	6.579e-02	-21.856	< 2e-16	***
niveau125	-1.585e-01	1.449e-02	-10.944	< 2e-16	***
niveau150	-1.931e-01	1.716e-02	-11.251	< 2e-16	***
niveau200	-1.381e-01	3.630e-02	-3.805	0.000142	***
niveauEco	1.297e-01	4.213e-02	3.079	0.002076	**
niveauhaut	-5.109e-01	1.933e-01	-2.643	0.008216	**
niveauHospi	-1.282e-01	2.886e-02	-4.441	8.93e-06	***
niveauAutre	-5.954e-01	2.675e-01	-2.226	0.026041	*
r_ancien_m	-3.369e-04	5.736e-05	-5.873	4.28e-09	***
RESEAU1:TCION	-9.079e-03	5.133e-03	-1.769	0.076957	.
RESEAU2:TCION	-5.384e-02	1.351e-02	-3.985	6.76e-05	***
RESEAU3:TCION	-1.824e+00	9.606e+03	0.000	0.999848	
RESEAUAutre:TCION	NA	NA	NA	NA	

Sortie 1 : Sortie R de la régression logistique sur l'ensemble des variables

Une première analyse des coefficients est possible. Le modèle logistique donne l'équation linéaire pour le log des rapports de cotes, nous pouvons donc conclure beaucoup de choses de ce *summary* de la régression logistique.

La fractionnement (*FRAC*) de la prime est une variable significative pour la résiliation, de plus il a un signe négatif, ce qui veut dire que plus la valeur est importante (par exemple la valeur = 8), moins il est probable que le contrat soit résilié. C'est en lien avec nos hypothèses de départ.

Le pourcentage de réduction (*pctreductot*) est une variable explicative significative pour la résiliation, de même avec son signe négatif, quand le pourcentage de réduction augmente, la chance de résiliation diminue. C'est également en lien avec nos hypothèses de départ.

La sinistralité de l'année en cours (*SCN_N*) est une variable significative, et avec son coefficient à 0.1688, on peut conclure que la hausse de sinistralité induit une hausse de la probabilité de résiliation.

Le nombre de modifications totales (*n_histo_tot*) est une variable significative. Son signe positif indique que pour les contrats sur lesquels il y a eu plus de modifications, les risques de résiliation sont plus importants. Cela permettrait de conclure qu'un assuré qui fait évoluer régulièrement son contrat risque de quitter son assureur.

Cependant le nombre de modification dans l'année (*n_histo_annee*) n'est pas une variable significative, cela est compréhensible puisque la variable *n_histo_annee* pour la majorité des contrats prend la valeur 1.

La dernière indexation (*w_index_N*) est aussi significative et, contrairement à notre intuition, la hausse d'indexation réduit la probabilité de résiliation avec son coefficient négatif à -1,315.

Le modèle ici montre également que la probabilité de résiliation est plus faible lorsque les contrats sont commercialisés par le réseau salarié (*RESEAU3*), plus fort lorsqu'ils sont commercialisés par le réseau courtage (*RESEAU2*). La valeur 0.752 est le coefficient associé à la modalité 2 de la variable RESEAU, et s'interprète par rapport à la modalité 1 (réseau des agents).

Toutes ces interprétations reposent sur la vision que donne le modèle calculé. Elles ne présagent rien sur la qualité de celui-ci et peuvent même s'avérer erronées si celui-ci est mauvais, ce que nous allons vérifier.

4.1.1.3. Evaluation du modèle brut

A la différence du modèle linéaire de régression simple, nous n'avons pas d'indicateur tel que le R^2 ajusté qui permet de mesurer la qualité du modèle rapidement. Ce que nous faisons c'est calculer un pseudo-R^2 de McFadden présenté par Daniel MCFADDEN (1974) :

$$R^2_{MF} = 1 - \frac{LL(Modèle\ proposé)}{LL\ (Modèle\ trivial)}.$$

Pour pouvoir interpréter ce pseudo-R^2 de McFadden, nous parlons d'abord du modèle trivial, modèle proposé, ainsi que la vraisemblance ($\boldsymbol{L}$) et la log-vraisemblance ($\boldsymbol{LL}$) :

Le modèle trivial est le modèle avec 1 seul paramètre (la constante) tandis que le modèle proposé correspond au modèle avec p paramètre + 1 constante

La vraisemblance s'écrivant : $\boldsymbol{L} = \prod \pi^Y \times (1-\pi)^{1-Y}$; $avec\ Y = 1\ ou\ 0$.
On en déduit la log-vraisemblance :$\boldsymbol{LL} = \sum Y \times \ln \pi + (1-Y) \times \ln(1-\pi)$; $avec\ \ Y = 1\ ou\ 0$.

La méthode du maximum de vraisemblance consiste à estimer les paramètres $\beta_0,\ \beta_1,\ \beta_2,\ \cdots,\ \beta_k$ de la régression logistique, qui maximisent $\boldsymbol{L}$. Comme max de $\boldsymbol{L}$ est également le max de $\boldsymbol{LL}$ grâce à la monotonie de la fonction LOG, on privilégie donc $\boldsymbol{LL}$ pour faciliter la manipulation.

Le calcul sous R de R^2_{MF} nous donne une valeur de 0,0411064.

Ce R^2_{MF} peut prendre des valeurs entre 0 et 1.

$$Min(R^2_{MF}) = 0, \qquad quand\ LL(Modèle\ proposé) = LL\ (Modèle\ trivial).$$

$$Max(R^2_{MF}) = 1, \qquad quand\ LL(Modèle\ proposé) = 0.$$

R^2_{MF} est considéré comme correct que quand il est supérieur à 1. Ce modèle n'est donc pas efficace selon ce premier résultat de pseudo-R^2, cela sera confirmé par le résultat du modèle tel que nous le verrons plus tard.

4.1.2. Sélection de variables explicatives

4.1.2.1. Sélection des variables par AIC

Nous allons par la suite utiliser la fonction *step* pour sélectionner les variables explicatives significatives un par un et voir si le résultat est cohérent avec ce test de significativité global.

On utilise la fonction *step* pour sélectionner le meilleur modèle, en minimisant AIC, un indicateur de la qualité d'un modèle, présenté par Hirotugu AKAIKE (1998). Plus il est faible, meilleur sera le modèle.

$$AIC\ (Akaike\ Information\ Criterion) = 2k - 2\ln \hat{L}.$$

k est le nombre de paramètres du modèle et $\hat{L}$ le maximum de vraisemblance du modèle. AIC est un estimateur fondé sur la théorie des informations, quand on s'attend à ce qu'un modèle représente la réalité, il ne peut jamais la faire à 100% : il « perd » des informations. AIC estime cette « perte » des informations lié à un modèle. Plus AIC est petit, plus l'information « perdue » est faible, la qualité de ce modèle est donc bonne. Le niveau d'AIC seul n'a pas de sens, sauf pour comparer un modèle à un autre. nous pouvons choisir le « meilleur » modèle qui contient le plus d'informations tout en gardant le moins de variables possible, c'est-à-dire le plus optimal.

Avec la fonction *step*, le logiciel R peut sélectionner les variables explicatives significatives ainsi : il rajoute ou supprime à chaque fois une variable, et calcule l'AIC associé à chacun des modèles obtenus avec ou sans cette variable. Il sélectionne ensuite le meilleur et réitère l'opération. Les variables supprimées (ou rajoutées) sont dans l'ordre suivant :

- Dst_Pharmaciens ;
- Dst_Infir_diplom ;
- Dst_Chirur_Dent ;
- Dst_specialiste;
- Dst_medecin;
- Dst_generaliste;
- w_chomage_N;
- Taux_ACS_Population ;
- w_index_N1.

Les étapes intermédiaires de R sont présentées en annexe 2.

Nous pouvons voir que les valeurs externes ne semblent pas appropriées. En effet, elles sont éliminées en premier, par exemple la densité des pharmaciens de la commune en premier, suivie de la densité des infirmières, dentistes, spécialistes, médecins, et généralistes. Par la suite, le taux de chômage et le taux ACS sont également éliminés et considérés comme non significatifs. Jusque-

là, la seule variable significative conservée est le taux de CMUC par la fonction *step*. Nous pouvons voir dans le dernier « step » de l'annexe 2 que, la suppression du taux CMU n'a que très peu d'impact au niveau d'AIC (202300 versus 202299 retenu par le modèle).

Par rapport aux variables internes, l'indexation de l'année N-1 n'est pas significative non plus, mais dans le même temps l'indexation de l'année N est conservée, ce qui est intuitivement parlant : nous avons tendance à résilier un contrat d'assurance si la dernière majoration est trop importante, et pas spécialement l'avant-dernière.

Le modèle optimisé donné par la fonction *step* comprend donc les variables suivantes :

- NBASSUR
- SCN_N
- FRAC
- r_ancien_m
- MTPCPTE
- Pctreductot
- RESEAU
- Societe
- Gamme
- an3_tech
- w_renfort
- w_dept
- n_histo_tot
- n_histo_annee
- w_cible
- w_index_N
- Taux_CMUC_Population
- SCN_N1
- Niveau
- RESEAU:TCION

Ce modèle donné par la fonction « *step* », avec les variables explicatives selon leur AIC associé, est-il optimal ? Le résultat de la dernière étape de la fonction « *step* » montre que l'ajout ou suppression de certaines variables ne change que marginalement AIC, comme présenté ci-dessous :

Le modèle sélectionné est donc celui qui présente un AIC de 202298,7, à savoir :

$$r\_Resil \sim NBASSUR + SCN\_N + FRAC + r\_ancien\_m + MTPCPTE + pctreductot + RESEAU + societe + gamme + an3\_tech + w\_renfort + w\_dept + n\_histo\_tot + n\_histo\_annee + w\_cible + w\_index\_N + Taux\_CMUC\_Population + SCN\_1 + niveau + RESEAU{:}TCION$$

	Df	Deviance	AIC
<none>		202023	202299
- societe	1	202025	202299
+ w _ inde_N1	1	202021	202299
- Taux_CMUC_Population	1	202026	202300
- SCN_N1	1	202029	202303
- NBASSUR	1	202030	202304
- RESEAU:TCION	3	202041	202311
- n_histo_tot	1	202038	202312
- r_ancien_m	1	202059	202333
- gamme	1	202076	202350
- w_renfort	1	202087	202361
- w_dept	105	202353	202419
- MTPCPTE	1	202156	202430
- n_histo_annee	1	202185	202459
- niveau	7	202230	202492
- w_index_N	1	202442	202716
- pctreductot	1	202510	202784
- FRAC	3	202639	202909
- an3_tech	1	202669	202943
- SCN_N	1	203246	203520

- w_cible	2	204428	204700

Sortie 2 : Sélection de variables par critère AIC

Nous avons donc observé la sortie du modèle obtenu par le critère AIC :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.547e+01	2.204e+03	-0.034	0.972691	
NBASSUR	1.771e-02	6.780e-03	2.612	0.008989	**
SCN_N	1.688e-01	5.123e-03	32.953	< 2e-16	***
FRAC2	-4.495e-01	3.962e-02	-11.346	< 2e-16	***
FRAC4	-2.059e-01	3.531e-02	-5.832	5.47e-09	***
FRAC8	-4.866e-01	1.976e-02	-24.627	< 2e-16	***
r_ancien_m	-3.391e-04	5.735e-05	-5.912	3.38e-09	***
MTPCPTE	-6.887e-04	6.058e-05	-11.369	< 2e-16	***
pctreductot	-1.436e+00	6.576e-02	-21.837	< 2e-16	***
RESEAU2	7.490e-01	1.902e-01	3.938	8.21e-05	***
RESEAU3	-1.182e-01	6.103e-02	-1.937	0.052695	.
RESEAUAutre	-1.257e+01	4.222e+02	-0.030	0.976255	
societeSA	2.914e-02	1.786e-02	1.631	0.102798	
gammeNGM	1.644e-01	2.260e-02	7.276	3.44e-13	***
an3_techO	-4.205e-01	1.692e-02	-24.857	< 2e-16	***
w_renfortO	-1.204e-01	1.494e-02	-8.058	7.73e-16	***
w_dept01	2.817e+01	7.039e+02	0.040	0.968083	
w_dept02	1.392e+01	3.311e+02	0.042	0.966465	
w_dept03	-7.157e+00	2.033e+02	-0.035	0.971919	
w_dept04	6.453e+00	1.416e+02	0.046	0.963664	
w_dept05	3.261e+01	8.116e+02	0.040	0.967955	
w_dept06	1.075e+01	2.518e+02	0.043	0.965944	
w_dept07	1.650e+01	4.004e+02	0.041	0.967138	
[…]					
w_dept978	**2.391e+00**	**1.148e+00**	**2.082**	**0.037349**	*
w_dept98	-1.227e+01	4.403e+02	-0.028	0.977775	
w_dept99	9.226e-01	1.331e+00	0.693	0.488301	
n_histo_tot	2.682e-02	6.794e-03	3.948	7.88e-05	***
n_histo_annee	-1.231e+01	3.149e+01	-0.391	0.695860	
w_ciblepros	-6.796e-01	1.562e-02	-43.516	< 2e-16	***
w_cibleseni	-6.313e-01	1.576e-02	-40.067	< 2e-16	***
w_index_N	-1.314e+00	6.261e-02	-20.992	< 2e-16	***
Taux_CMUC_Population	1.209e+03	3.087e+04	0.039	0.968770	
SCN_N1	-1.620e-02	6.655e-03	-2.434	0.014929	*
niveau125	-1.589e-01	1.448e-02	-10.975	< 2e-16	***
niveau150	-1.936e-01	1.716e-02	-11.282	< 2e-16	***
niveau200	-1.393e-01	3.628e-02	-3.839	0.000124	***
niveauEco	1.301e-01	4.212e-02	3.087	0.002020	**
niveauhaut	-5.125e-01	1.933e-01	-2.651	0.008023	**
niveauHospi	-1.277e-01	2.886e-02	-4.424	9.69e-06	***
niveauAutre	-5.934e-01	2.675e-01	-2.218	0.026548	*
RESEAU1:TCION	-8.896e-03	5.133e-03	-1.733	0.083033	.
RESEAU2:TCION	-5.348e-02	1.351e-02	-3.959	7.54e-05	***
RESEAU3:TCION	-9.610e-01	4.002e+01	-0.024	0.980845	
RESEAUAutre:TCION	NA	NA	NA	NA	

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					
Number of Fisher Scoring iterations: 13					

Sortie 3 : Sortie R de la régression logistique après sélection de variables par critère AIC

Plusieurs variables restent non significatives : société, département, nombre de changement (histo) dans l'année, Taux de CMUC, sinistralité de N-1. Nous n'allons pas les garder. Nous obtenons ainsi le modèle retenu sur notre ensemble de données.

$$r_{Resil} \sim NBASSUR + SCN\_N + FRAC + r\_ancien\_m + MTPCPTE + pctreductot + RESEAU + gamme + an3\_tech + w\_renfort + n\_histo\_tot + w\_cible + w\_index\_N + niveau.$$

4.1.2.2. Mesure de la colinéarité

Un modèle linéaire généralisé (GLM) exige la non-existence de la multicolinéarité parfaite, qui est traduit par : une variable explicative est la combinaison linéaire de(s) autre(s) variable(s) explicative(s).

Par exemple si $X_3 = 2X_1 + 3X_2$, alors X_1, X_2 et X_3 seront multicolinéaires.

La mesure de colinéarité permet de vérifier que différentes variables explicatives ne mesurent pas le même phénomène. Multicolinéarité et corrélation ne sont pas à confondre ; deux variables corrélées ne sont pas forcément colinéaires.

Si deux variables sont colinéaires, elles sont fortement corrélées, mais si deux variables sont corrélées, elles ne sont pas forcément colinéaires. Si certaines variables mesurent le même phénomène, on peut dire qu'elles sont colinéaires, c'est le cas par exemple si nous avions une variable présentant le S/C du contrat et une seconde qui représente le ratio combiné. Si les taux de frais sont identiques, alors les 2 variables sont absolument colinéaires. Il faut donc en éliminer une des 2 pour ne pas fausser le modèle.

Dans une régression, la multicolinéarité peut augmenter la variance des coefficients de régression, elle induit ainsi que le coefficient associé à chaque variable explicative n'est plus valide, et il ne peut plus donc être interprété de manière fiable.

L'indicateur VIF (variance inflation factor) capte l'augmentation de la variance à cause d'un coefficient qui est linéaire avec d'autre variables explicatives. Donc un VIF de 1 signifie un manque de corrélation avec les autres variables.

> vif(final_2018)			
	GVIF	Df	GVIF^(1/(2*Df))
NBASSUR	1.137475	1	1.066525
SCN_N	1.009321	1	1.004650
FRAC	1.167353	3	1.026125
r_ancien_m	2.570203	1	1.603185
MTPCPTE	1.550037	1	1.245005
pctreductot	1.557564	1	1.248024
RESEAU	17.583565	3	1.612567
gamme	3.368019	1	1.835216
an3_tech	1.206238	1	1.098289
w_renfort	1.568636	1	1.252452
n_histo_tot	1.261109	1	1.122991
w_cible	1.395054	2	1.086795
w_index_N	1.020939	1	1.010415

niveau	1.323932	7	1.020245
TCION	16.656423	1	4.081228

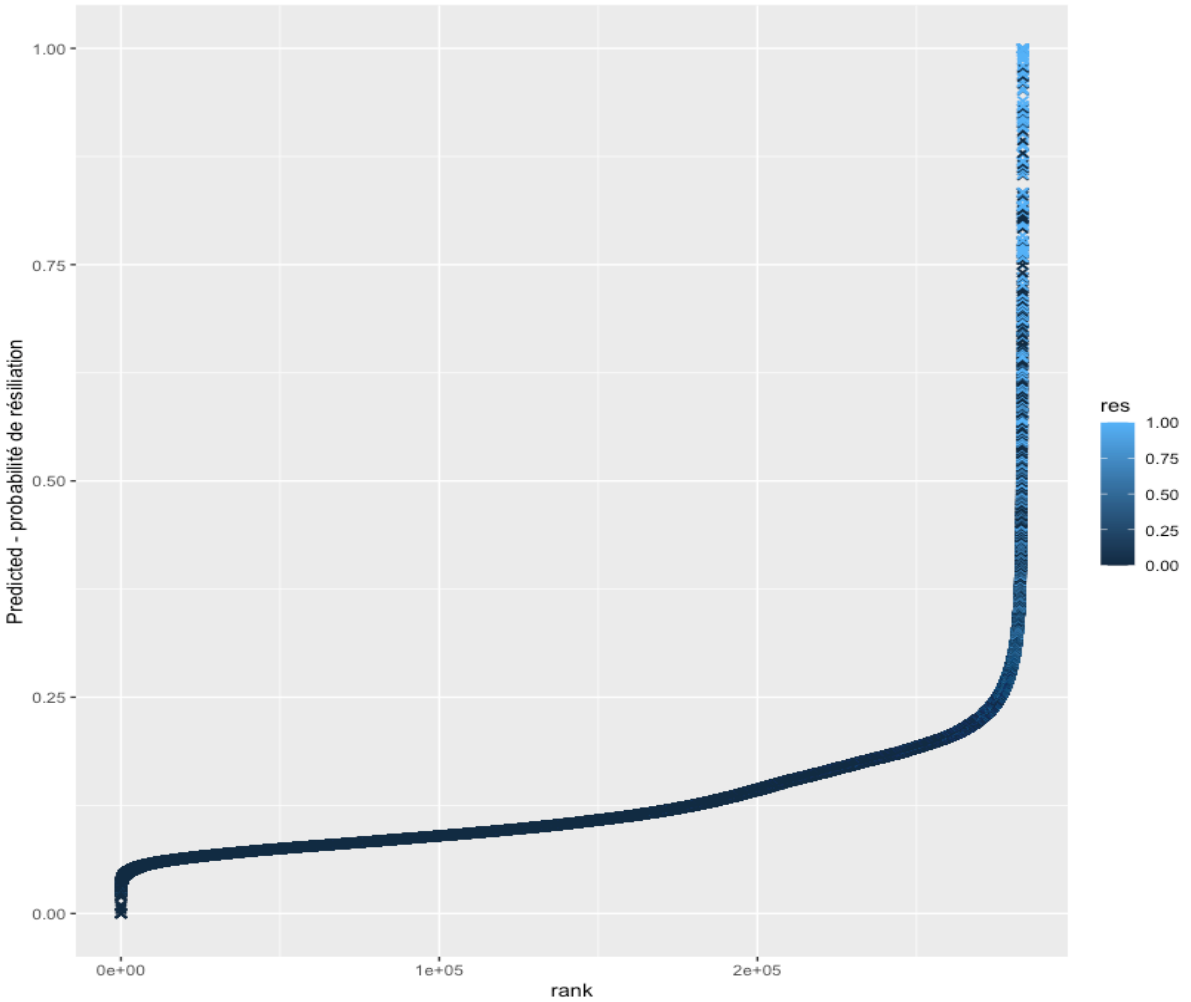
Sortie 4 : Calcul de la Variance VIF pour tester la colintéarité

Pas de surprise que la forte corrélation existe sur la variable *TCION*, *RESEAU*, c'est pour cela que nous allons créer la nouvelle variable *RESEAU:TCION* pour éliminer cet effet de forte liaison ; cependant de la corrélation est également observée entre variables *gamme* et *r_ancien_m*, mais il ne sont pas parfaitement multicolinéaires, donc ce n'est pas gênant de les laisser dans le modèle. Nous avons déjà vu ces relations dans la partie test de corrélation.

4.1.2.3. Présentation graphique du modèle

Comme tous les modèles logistique, la présentation graphique n'est pas ce qu'on a l'habitude de voir, dans le modèle linéaire par exemple.

On peut présenter cependant un graphique avec la valeur de prédiction (résilié ou non) et la réalité (résilié ou non) :



Graphique 26 : Présentation graphique du modèle

Les points bleu clair, qui se trouvent en majorité en haut du graphique, sont les vraies résiliations, qui sont également prédits par le modèle avec une forte probabilité de résiliation, et les points bleu foncé, qui se trouvent plutôt en bas du graphique, sont les contrats non résiliés, qui sont prédits par

le modèle avec une faible probabilité de résiliation. Avec un modèle parfait, l'ensemble des contrats résiliés (bleus clairs) auraient une probabilité calculée supérieure à 0,5, et les contrats non résiliés auraient une probabilité calculée inférieure à 0,5. Ici nous observons que plusieurs contrats non résiliés sont mal prédits. Le graphique est trop dense pour voir si les contrats résiliés sont mal prédits ou pas.

4.1.3. Application du modèle

Le modèle choisi est appliqué d'abord sur la base de données de validation 2018, qui correspond au tiers des observations choisies au hasard (qui n'ont pas été utilisées pour l'apprentissage du modèle), pour voir si la modélisation est valide. Ensuite il sera appliqué sur la base de données 2019 pour voir s'il est valide d'une année à l'autre. Le même traitement de données a été fait dans les 2 bases d'applications du modèle, par exemple calculer l'ancienneté du contrat en nombre du mois, limiter l'indexation à 50% pour exclure une éventuelle hausse de prime liée à l'ajout d'un ayant-droit, etc.

4.1.4. Evaluation et validation du modèle

Dans un premier temps, après avoir sélectionné les variables explicatives de la régression logistique, nous pouvons déjà évaluer la qualité du modèle en comparant la prédiction (valeur entre 0 et 1) et la réalité (résilié ou non).

Si la valeur calculée est supérieure à 0,5, le contrat est prédit comme résilié dans l'année. S'il est réellement résilié, nous le comptabilisons comme un « réussi » ; si la valeur calculée est inférieure à 0,5, il est prédit comme étant non résilié, et s'il est en réalité non résilié, c'est aussi compté comme un « réussi ». Le sous-total des « réussis » rapporté à l'ensemble de la base d'apprentissage donne un taux de réussite, à 87,8%, cela ne parait pas trop mal pour le premier aperçu. Nous nous apercevrons plus tard qu'il ne faut pas s'arrêter à la première impression.

Dans le cadre d'un modèle logistique, les rapports de cotes (valeurs exponentielles des coefficients) du modèle sont présentés :

> exp(coef(final_2018))				
(Intercept)	**NBASSUR**	**SCN_N**	**FRAC2**	**FRAC4**
0.5293195862	1.0132026836	1.1794711996	0.6383105605	0.8093186458
FRAC8	**r_ancien_m**	**MTPCPTE**	**pctreductot**	**RESEAU2**
0.6162201920	0.9996750447	0.9993404790	0.2495434050	2.1030854692
RESEAU3	**RESEAUAutre**	**gammeNGM**	**an3_techO**	**w_renfortO**
0.9079612074	0.0005843055	1.1670168556	0.6558017228	0.8681800223
n_histo_tot	**w_ciblepros**	**w_cibleseni**	**w_index_N**	**niveau125**
1.0257446406	0.5222887577	0.5372571013	0.2724997410	0.8456720153
niveau150	**niveau200**	**niveauEco**	**niveauhaut**	**niveauHospi**
0.7903720339	0.8189513402	1.1388767385	0.5429088704	0.8810212742
niveauAutre	**RESEAU1:TCION**	**RESEAU2:TCION**	**RESEAU3:TCION**	**RESEAUAutre:TCION**
0.5387203524	0.9917558762	0.9475726480	0.5679778006	NA

Sortie 5 : Rapport des cotes du modèle après sélection de variables par critère AIC.

Concernant le nombre d'assurés associés à un contrat (*NBASSUR*), le rapport de cotes est calculé à 1,013, c'est-à-dire pour chaque assuré en plus, il n'y a pas de différence significative dans le taux de résiliations (1,3% de chances de plus de résilier par assuré supplémentaire).

En revanche, pour le réseau de distribution courtage (*REASEAU2*), le rapport de cotes est de 2,103, par rapport au réseau de distribution agents généraux (*REASEAU1*), les risques de résiliation du réseau courtage semble 2 fois plus importante que ceux du réseau agents toutes choses égales par ailleurs.

4.1.4.1. Matrice de confusion

La matrice de confusion permet de confronter les réels et les prédictions, avec sa répartition de 2 types d'erreur pour facilement voir la qualité de la prédiction. Elle se présente généralement comme suit :

		Prédictions	
		Résilié	**Non-Résilié**
Réels	**Résilié**	Vrai Positif (Vp)	Faux Négatif (Fn)
	Non-Résilié	Faux positif (Fp)	Vrai Négatif (Vn)

Tableau 2 : Matrice de confusion : définition

A partir de ce tableau, nous pouvons voir parmi les résiliations prédites par le modèle, combien sont réellement résiliés avant la fin d'année (Vp), combien ne sont pas résilié (Fp) ; et les parmi les prédits non-résiliés par le modèle, combien sont réellement en vigueur en fin d'année (Vn), combien sont cependant résiliés (Fn). Nous pouvons ainsi déduire le taux d'erreur ($\frac{Fp+Fn}{Vp+Fp+Fn+Vn}$), le taux de succès ($\frac{Vp+Vn}{Vp+Fp+Fn+Vn}$).

Pour analyser la qualité d'un modèle, le taux d'erreur ne suffit pas à lui seul : la sensibilité (ou le taux de vrais positifs $\frac{Fp}{Vp+Fn}$), la précision ($\frac{Fp}{Vp+Fp}$), la spécificité ($\frac{Vn}{Fp+Vn}$) ou encore le taux de faux positifs ($1-\frac{Vn}{Fp+Vn}$) sont des éléments à prendre en compte. Mais nous verrons cela en détail ultérieurement.

4.1.4.2. Validation sur la base 2018

Pour tester la qualité de ce modèle, obtenu à partir de la base d'apprentissage (70% de la base 2018), nous allons appliquer ce modèle dans la base de validation de la même période (30% de la base 2018) et la base de validation de l'année suivante (base 2019), que nous supposons toutes les deux indépendantes de la base d'apprentissage. Une matrice de confusion est calculée pour chaque validation – la méthode *predict* avec l'argument *type="response"* - nous appliquons notre modèle logistique à ces deux tableaux de données et il renvoie pour chaque individu la probabilité qu'il ait vécu le phénomène étudié.

Nous devons transformer nos probabilités prédites en une variable du type binaire, donc « oui / non ».

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	81	14 724	14 805
	Non-Résilié	91	106 360	106 451
	Total	172	121 084	121 256

Tableau 3 : Matrice de confusion : base de validation 2018, modèle initial

Nous avons donc 14 815 (14 724+91) prédictions incorrectes sur un total de 121 256, soit un taux de mauvais classement de 12,22 %. Ceci ne parait pas très mauvais comme taux d'erreur dans un premier temps, mais si on regarde plus en détail, dans la base validation qui contient 121 256 contrats, il y a réellement 106 451 contrats en vigueur en fin d'année, et le modèle a réussi à classer 106 360 contrats dans la bonne catégorie, et il s'est trompé sur 91 contrats ; les faux positifs ne sont pas importants, à 0,085%. Cependant, dans les vrais résiliés 14 805 contrats, seule 81 contrats sont correctement classés, c'est-à-dire que le modèle n'a pas pu identifier 14 724 contrats qui sont réellement résiliés, soit un taux de faux négatifs de 99,5%, cette modélisation n'est pas efficace, dans le sens où nous cherchons à identifier les contrats qui sont susceptibles de résilier, mais seuls 0,5% peuvent être identifiés, donc la majorité des risques de rachat se sont échappés. Ce modèle a classé la quasi-totalité des contrats dans la catégorie Non-Résilié.

Les faux positifs engendrent un coût important pour la rétention des contrats, par exemple pour « éviter » les « éventuels » résiliés, on pourrait investir sur eux avec des mesures les incitant à rester, mais dans ce cas, l'argent investi sur les faux positifs serait inefficient.

Pour voir si ce problème survient à cause de la suppression de certaines variables explicatives, nous avons repris le modèle avec l'ensemble de variables explicatives initiales.

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	86	14 719	14 805
	Non-Résilié	87	106 364	106 451
	Total	173	121 083	121 256

Tableau 4 : Matrice de confusion : base de validation 2018, modèle complet

Le résultat n'est pas plus satisfaisant. En fait, nous constatons qu'avant et après avoir sélectionné les variables, le modèle donne les résultats très similaires. Le mécanisme de sélection des variables a donc bien fonctionné, mais la régression logistique telle qu'appliquée ici n'est pas efficace.

4.1.4.3. Modèle avec données rééquilibrées

Nous avons choisi de modéliser une nouvelle sensibilité afin de traiter le déséquilibre des données (en anglais : Imbalanced Data) entre nos deux classes à prédire, « résilié » ou « non résilié », qui représentent respectivement 12% et 88% de notre base. « non-résilié » est surreprésenté.

De fait, un modèle brut qui classerait l'ensemble des observations dans la catégorie « non résiliée », obtiendrait un taux d'erreur global de 12%, qui pourrait sembler satisfaisant.

La plupart des algorithmes fonctionne mieux quand les nombres d'observations dans chacune des classes sont équivalents, car ces algorithmes cherchent maximiser les réussis et minimiser les erreurs. Nous utilisons donc la technique de rééchantillonnage (en anglais : resampling techniques) – undersample. Il s'agit de diminuer la quantité de la classe majoritaire, de façon aléatoire, pour obtenir un échantillon de même taille que la classe minoritaire, afin de « rééquilibrer » les deux classes étudiées.

Nous avons donc recalculé un nouveau GLM sur cette nouvelle base de données, en suivant les mêmes étapes que pour le modèle initial :

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	8 365	6 420	14 785
	Non-Résilié	5 295	9 751	15 046
	Total	13 660	16 171	29 831

Tableau 5 : Matrice de confusion : base de validation 2018, modèle "balanced data"

Le modèle semble bien meilleur au premier abord : en effet, contrairement aux 2 modèles précédents, il est bien capable de prédire des résiliations en nombre significatif.

Ainsi, ce modèle permet d'avoir 62% de chances que les contrats prédits comme résiliés le soient vraiment. Et avec un nombre de vrais positifs important (8 365), contrairement au modèle retenu.

En revanche, nous avons 11 715 (6 420+5 295) prédictions incorrectes sur un total de 29 831, soit un taux de mauvais classement très important de 39,27 %. Si on ne s'intéresse qu'aux 14 785 contrats vraiment résiliés, 8 365 contrats sont correctement classés, le taux de faux négatifs est de 43,4%, c'est-à-dire que le modèle peut classer correctement un peu plus que la moitié des résiliations réelles et deux tiers des non-résiliations réelles. Au niveau des faux négatifs ce dernier modèle est plus performant, mais les faux positifs semblent importants et cela peut engendrer un coût important pour les actions préventives.

Ce modèle nous semble globalement meilleur que les deux précédents. C'est pourquoi c'est celui-ci que nous allons comparer à la méthode de machine learning de random forest.

Voici les sorties concernant le modèle :

$$r\_Resil \sim FRAC + MTPCPTE + RESEAU + gamme + an3\_tech + w\_renfort + n\_histo\_tot + w\_cible + w\_index\_N + SCN\_N + pctreductot + niveau$$

> summary(final_2018_BD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.0641	-1.0844	-0.5946	1.1204	2.0286

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.179e+00	3.588e-02	32.856	< 2e-16	***
FRAC2	-3.891e-01	5.153e-02	-7.551	4.31e-14	***
FRAC4	-2.073e-01	4.875e-02	-4.253	2.11e-05	***
FRAC8	-4.254e-01	2.717e-02	-15.658	< 2e-16	***
MTPCPTE	-4.755e-04	7.019e-05	-6.774	1.25e-11	***

RESEAU2	1.895e-01	4.831e-02	3.922	8.79e-05	***
RESEAU3	2.046e-02	2.179e-02	0.939	0.34767	
RESEAUAutre	-7.821e+00	4.395e+01	-0.178	0.85878	
gammeNGM	1.274e-01	2.348e-02	5.426	5.78e-08	***
an3_techO	-4.187e-01	2.172e-02	-19.281	< 2e-16	***
w_renfortO	-1.476e-01	1.992e-02	-7.409	1.27e-13	***
n_histo_tot	2.741e-02	8.822e-03	3.107	0.00189	**
w_ciblepros	-6.757e-01	2.029e-02	-33.305	< 2e-16	***
w_cibleseni	-6.548e-01	2.082e-02	-31.447	< 2e-16	***
w_index_N	-2.145e+00	1.192e-01	-17.999	< 2e-16	***
SCN_N	1.712e-01	8.182e-03	20.929	< 2e-16	***
pctreductot	-1.326e+00	8.244e-02	-16.087	< 2e-16	***
niveau125	-1.681e-01	1.957e-02	-8.589	< 2e-16	***
niveau150	-2.321e-01	2.229e-02	-10.415	< 2e-16	***
niveau200	-1.531e-01	4.693e-02	-3.263	0.00110	**
niveauEco	8.003e-02	6.134e-02	1.305	0.19199	
niveauhaut	-4.962e-01	2.241e-01	-2.214	0.02683	*
niveauHospi	-9.190e-02	3.906e-02	-2.353	0.01863	*
niveauAutre	-4.226e-04	4.001e-01	-0.001	0.99916	

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 96191 on 69386 degrees of freedom					
Residual deviance: 91689 on 69363 degrees of freedom					
AIC: 91737					
Number of Fisher Scoring iterations: 7					

Sortie 6 : Résumé de la régression logistique sur la base de données rééquilibrée

Le R^2_{MF} de ce nouveau modèle est à 0,047, légèrement plus élevé que celui du modèle initial à 0,041.

4.1.4.4. Validation sur la base 2019

Bien que nous sachions que notre modèle de régression logistique n'est pas suffisamment performant, nous l'avons testé sur la base 2019 pour vérifier sa stabilité.

La règle est la suivante : si l'estimateur issu de la régression logistique est supérieur à 50%, alors nous estimons que le contrat est prédit comme résilié. Cela nous donne la matrice de confusion suivante :

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	305	47 064	47 369
	Non-Résilié	202	369 060	369 262
	Total	507	416 124	416 631

Tableau 6 : Matrice de confusion : base de validation 2019, modèle initial

Nous avons donc 47 266 (47 064+202) prédictions incorrectes sur un total de 416 631 observations, soit un taux de mauvais classement de 11,34 %. Cependant, parmi 369 262 contrats non résiliés,

seul 202 sont mal classés, soit un taux d'erreur de faux positifs à 0,055% (202 / 369 262). En ce qui concerne les faux négatifs, notre modèle se trompe de 99,36% (47 064/47 369).

On a donc un modèle qui, bien qu'ayant seulement 11% d'erreurs, est incapable d'identifier les contrats résiliés. Il n'apporte donc pas beaucoup d'informations.

La sensibilité aux variables sélectionnées a été mesurée avec le modèle qui contient l'ensemble de variables explicatives :

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	30 062	17 307	47 369
	Non-Résilié	236 354	132 906	369 260
	Total	266 416	150 213	416 629

Tableau 7 : Matrice de confusion : base de validation 2019, modèle complet

Le modèle complet (avec toutes les variables) a un taux de mauvais classement de 60,88% ; le modèle avec la sélection des variables est visiblement mieux et plus stable dans le temps. Cependant, il parvient mieux à identifier des vrais positifs que le modèle initial.

Malheureusement, le nombre de faux positifs est beaucoup trop important : la probabilité d'avoir un contrat effectivement résilié parmi les prédits comme tel est seulement de 11,3%, ce qui est plus faible que le hasard dans la population totale. Ce modèle est donc contre-productif.

Le modèle obtenu après avoir traité les données « non équilibrées » (Inbalanced Data) a été également appliqué sur la base 2019 pour valider le modèle :

		Prédictions		
		Résilié	Non-Résilié	Total
Réels	Résilié	24 949	22 420	47 369
	Non-Résilié	119 247	250 015	369 262
	Total	144 196	272 435	416 631

Tableau 8 : Matrice de confusion : base de validation 2019, modèle "balanced data"

Ce modèle a plusieurs avantages : contrairement au modèle initial, il parvient bien à identifier des vrais positifs en quantité significative. En outre, il parvient à identifier plus de la moitié des contrats résiliés, sans avoir autant de faux positifs que le modèle complet.

En revanche, ce taux de faux positifs reste très important, en effet, la probabilité d'avoir un contrat effectivement résilié parmi les prédits comme tel sont seulement de 17,3%, ce qui est à peine plus élevé que le hasard.

L'efficacité de l'ensemble des modèles est donc plutôt décevante sur 2019.

4.2. *Modèles Random forest*

Le machine learning utilise des méthodes statistiques connues de longue date, mais qui ont été rendues plus efficaces et possibles depuis quelques années par l'apparition d'ordinateurs plus puissants, plus rapides et avec des capacités de stockage de données plus adaptées.

C'est pourquoi ces méthodes sont en pleine expansion au XXI^{ème} siècle.

La méthode de machine learning que nous prendrons est le "Random Forest", méthode présentée au début des années 2000 par Leo BREIMAN (1984, 2001) réexpliqué par GENUER et POGGI (2017), Fabien MOUTARDE (2017) et Sébastien VEREL (2016).

On suppose que les variables aléatoires {(X1,Y1), … (Xn,Yn)} sont indépendantes et identiquement distribuées (i.i.d.) de même loi que (X,Y).
On a n = le nombre de contrats en portefeuille au 1er janvier de l'année.
$X \in R^p$ et $Y \in \{0 ;1\}$, avec

0 ⇔ « le contrat n'a pas été résilié dans l'année »
1 ⇔ « le contrat a été résilié dans l'année »

Nous souhaitons construire un prédicateur $\hat{a} : R^p \rightarrow \{0 ;1\}$.

4.2.1. Théorie de Random Forest

Le random forest est une méthode qui se compose d'une multitude d'arbres aléatoires qui auront chacun une vision parcellaire du problème et dont l'analyse d'ensemble permet de tirer un résultat fiable. Chaque arbre pris seul apporte une fraction de la solution, mais l'ensemble donne une robustesse et une précision particulièrement reconnues. Concrètement, pour chaque entrée, chaque arbre va "voter" et ainsi l'entrée sera classifiée selon le résultat donné par la majorité des arbres dans le cas d'une classification. Dans le cas d'une régression, la valeur estimée sera la moyenne des valeurs données par l'ensemble des arbres.

La méthode repose sur 3 concepts : l'arbre de décision CART, le "tree baging" et le "features sampling".

Nous noterons que le modèle random forest est non paramétrique et ne nécessite donc pas d'hypothèses fortes sur la structure des données (par exemple, une éventuelle normalité des résidus). Ainsi, il sait parfaitement faire abstraction des éventuelles corrélations de variables.

4.2.1.1. L'arbre CART

L'acronyme CART signifie "Classification And Regression Trees". Pour simplifier la présentation de l'approche, nous ne prendrons le cas que de la classification sachant que la méthode fonctionne également dans le cas d'une régression.

La racine de l'arbre correspond à l'ensemble des n observations et des p variables de la base à classifier. L'arbre CART se compose de nœuds successifs auxquels on applique des tests qui vont partitionner les données. L'arbre CART étant un arbre binaire, chaque nœud va effectuer un découpage en 2 partitions. La partition se fait en fonction d'un critère de coût qu'on essaie de

minimiser. Dans le cas de la classification, on va minimiser l'impureté (l'indice de Gini) des partitions obtenues à chaque nœud et essayer donc de les homogénéiser. Un nœud étant parfaitement homogène ou « pur » s'il ne contient que des observations de la même classe.

L'indice de Gini ou Coefficient de Gini est un indicateur de l'hétérogénéité d'un groupe. Il varie de 0 à 1, 0 signifiant que le groupe est pur ; 1 indiquant une hétérogénéité totale.

Il se calcule ainsi :

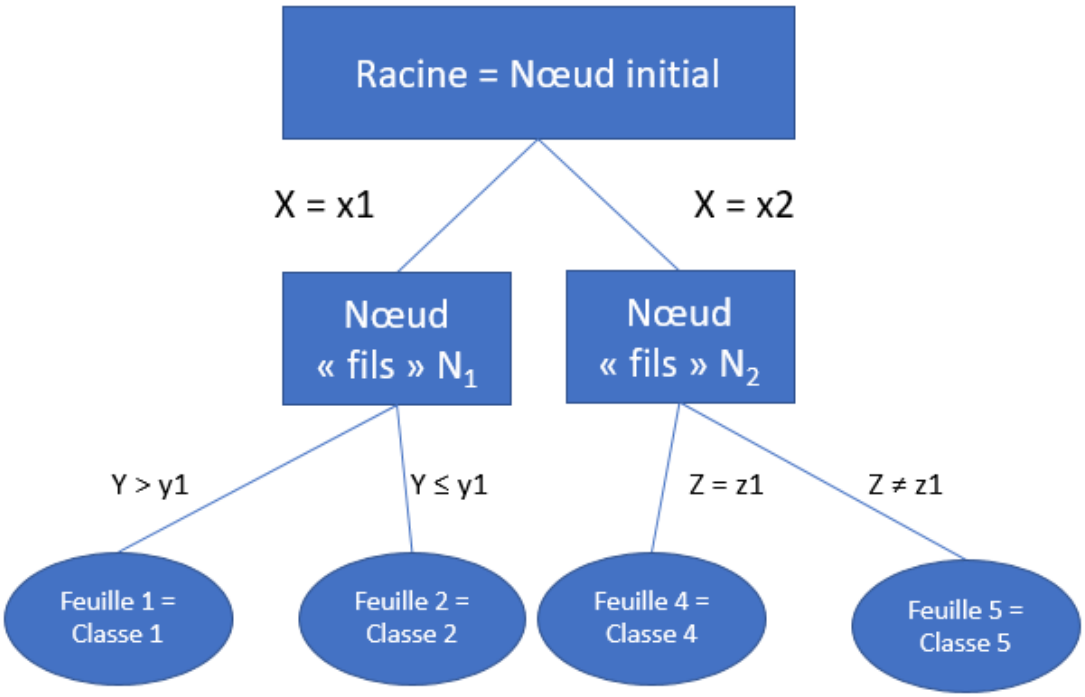
$$Gini = 1 - \sum_k p^2(w_k).$$

Avec $p(w_k)$, la probabilité de la classe w_k (estimée par la proportion N_k/N)

Concrètement, à chaque nœud N nous allons tester les différentes alternatives possibles de variables. Nous sélectionnerons la variable qui apportera le meilleur gain tel que pour chaque nœud « fils » N_j on a :

$$Gain = Gini(N) - \sum_j p(N_j) Gini(N_j).$$

Une fois la racine découpée, nous réitérons ce processus sur chacun des nœuds "fils". Nous n'appliquons pas le process aux nœuds ne contenant plus suffisamment d'observations ou lorsqu'il est "pur". In fine, l'arbre va donner ainsi une partition (les "feuilles") de la racine.



Graphique 27 : Exemple d'arbre CART

Ainsi l'algorithme d'arbre se découpe en 3 étapes :

1- Définir un critère qui permet de décider quelle est la meilleure division. Dans la stratégie CART, on va utiliser une stratégie dite « gloutonne », c'est-à-dire qu'on va utiliser à chaque nœud le test qui permet de faire décroître au maximum l'indice de Gini (voir ci-dessous).

2- Mettre en place une règle indiquant si un nœud est terminal (c'est-à-dire qu'il ne sera plus partitionné et devient donc une « feuille »). La feuille peut être « pure » ou non. Si la règle est trop stricte, on risque de rentrer dans du surapprentissage ou « overfitting », c'est-à-dire que le modèle est excellent sur les données en présence, mais non généralisable.
3- Affecter chaque « feuille » à une classe. En théorie, on l'affecte à la classe majoritaire. En cas d'égalité, on peut définir une règle de choix au hasard ou on définit une classe par défaut.

L'algorithme CART prévoit une 2ème étape appelée "élagage" (ou « pruning ») qui consiste à optimiser le nombre de nœuds, afin d'obtenir un bon équilibre entre la variance et le biais : en effet, la racine (arbre sans nœud) a un biais nul, mais une forte variance, et l'arbre le plus développé a une variance faible, mais un biais élevé. L'idée de l'élagage est donc de trouver un arbre "réduit" qui apporte un bon équilibre entre ces deux éléments en pénalisant le nombre de feuilles (on veut conserver une erreur raisonnable en diminuant la complexité de l'arbre). Nous enlèverons donc mécaniquement les nœuds qui apportent le moins à l'arbre (i.e. ceux qui ont des erreurs trop importante). Nous noterons, que pour effectuer cet élagage un critère différent du critère de Gini est nécessaire. Nous utilisons donc le critère de coût suivant :

$$Err\ (A) + a \times L(A).$$

Avec :
- $Err\ (A)$ la fraction des observations de validation mal classées par l'arbre A ;
- $L(A)$ le nombre de feuilles dans l'arbre A ;
- a, le coût de pénalité par nœud, variant de 0 (on prend l'arbre le plus complet) à l'infini (on ne prend qu'un nœud), choisi par l'utilisateur.

Dans l'algorithme de Random Forest, les arbres ne sont pas élagués.

4.2.1.2. Tree Baging

Nous allons présenter le concept du "tree baging", méthode de Leo Breiman datant de 1996 :

En théorie, pour former les sous-échantillons pour dresser les arbres, l'idéal est de sélectionner des observations sans remise. Dans la pratique, nous n'avons généralement pas assez d'observations pour nous permettre d'avoir une forêt d'arbres sans remise. C'est pourquoi, nous allons créer les arbres de décisions sur des sous-échantillons sélectionnés par "baging", c'est à dire par tirage de n' (avec n' < n) observations avec remise.

Exemple, pour un échantillon composé des lettres de l'alphabet, un premier arbre serait appris sur le sous-échantillon de 4 lettres [A, F, H, X], un second [E, F, M, P], un troisième [B, G, M, X], etc.

Cette méthode de baging permet de réduire la variance et donc les erreurs de prévisions. Cette propriété étant vraie malgré la non-indépendance des sous-échantillons et donc des estimateurs.

4.2.1.3. Feature Sampling

Pour chaque sous-échantillon sélectionné, l'algorithme de forêt aléatoire va dresser un arbre proche de l'algorithme CART (méthode de Leo Breiman, 1984) : en effet, l'arbre ne va pas utiliser l'ensemble

des variables disponibles, mais à chaque nœud, il va tirer un échantillon sans remise de m variables parmi les p variables disponibles.

L'objectif est de rendre plus indépendants les arbres en ajoutant du hasard dans le choix des variables qui interviennent dans les modèles. En effet, sans cela, dans l'éventualité où une variable aurait énormément de poids, dans la décision, alors la majorité des arbres donneraient la même décision, ce qui ne convient pas avec l'idée de la random forest d'un choix « démocratique » dans la classification.

Finalement, on remarque que plusieurs paramètres pourront être "réglés" afin d'optimiser la classification. Les 2 principaux étant : le nombre d'arbres (de sous-échantillons) de la régression et le nombre de variables sélectionnées à chaque nœud des arbres. Ce réglage s'appelle l'hyperparamétrage. Nous allons parvenir par tâtonnement à régler ces paramètres en utilisant le critère de l'erreur Out-of-Bag (OOB).

4.2.1.4. Le vote des arbres

Pour rappel, nous cherchons un prédicateur â : $R^p \rightarrow \{0 ;1\}$.

En créant une forêt de q arbres, chaque arbre l nous donnera un prédicateur noté $\hat{a}(X, l)$. Ces prédicateurs d'arbres sont iid.

On fait ensuite voter les prédicateurs d'arbres pour trouver la valeur de prédiction (0 ou 1) qui remporte le plus de voix : $\hat{a}(X) = \underset{0 \leq c \leq 1}{\operatorname{argmax}} \sum_{l=1}^{q} \hat{a}(X, l) = c$.

4.2.1.5. Erreur OOB

L'erreur Out-of-bag est une estimation du taux d'erreur de la classification. Il indique la probabilité que l'estimation soit différente de la réalisation. Elle repose sur le fait que les estimateurs sont réalisés sur des sous-échantillons "baggés" et donc la partie de la base de données non utilisée pour l'estimation sert d'échantillon de validation.

Plus concrètement, l'algorithme de l'Erreur OOB, va pour chaque observation (X_i, Y_i), regarder le vote de l'ensemble des arbres de la forêt aléatoire qui n'utilisent pas cette observation. On réitère l'opération pour l'ensemble des observations 1 ≤ i ≤ n. Et enfin, l'erreur OOB est estimée comme la fréquence des erreurs sur l'ensemble des observations : $ErrOOB = \frac{1}{n} \times \sum_{i=1}^{n} Y_i \neq \hat{Y}_i$.

L'erreur OOB est donc représentative de la robustesse du modèle sur l'échantillon d'apprentissage.

A noter cependant qu'elle n'est pas suffisante pour qualifier un modèle de "bon modèle" : en effet, dans le cas d'un surapprentissage, le modèle produira une erreur OOB très faible, mais le modèle ne sera pas applicable à un autre échantillon de données. C'est pourquoi dans le cadre de de cette étude, nous veillerons à utiliser un échantillon de validation en complément de l'échantillon d'apprentissage.

4.2.2. Analyse d'un modèle Random Forest

La difficulté reconnue de nombreux modèles de machine learning est le côté "boîte noire", c'est à dire, la difficulté à analyser en profondeur les résultats fournis par les algorithmes sous-jacents. La méthode Random Forest n'échappe pas à ces travers.

Contrairement aux méthodes traditionnelles de régression telles que la régression logistique, avec le random forest, nous n'avons pas la possibilité de connaître la participation de chaque variable explicative au résultat. Par exemple, dans notre cas, nous ne pourrons pas savoir si l'ancienneté est un critère favorable ou non à la résiliation.

Toutefois, un critère peut être étudié pour pouvoir avoir une idée du poids de chaque variable : l'indice de la diminution moyenne de Gini.

Comme nous l'avons vu précédemment, l'indice de Gini représente l'hétérogénéité d'une classe. On va regarder donc sur l'ensemble des arbres, quel paramètre influe le mieux, c'est-à-dire permet de baisser au maximum cet indice. L'inconvénient de cette méthode est qu'on peut ainsi savoir quelle variable permet d'aider au mieux à la classification des données, cependant on ne peut pas savoir dans quel sens elle influe, et ne permet pas une bonne interprétation. C'est ici que la forêt aléatoire révèle son côté « boîte noire ».

4.2.3. Application du Random Forest sur nos données

4.2.3.1. Modèle standard

Nous recherchons un modèle qui soit fiable sur la base de 2018, et sur 2019 afin de vérifier une stabilité dans le temps. En effet, un bon modèle qui n'aurait pas cette propriété ne serait pas utilisable dans le cadre d'une prévision.

Nous allons donc apprendre le modèle sur un échantillon d'apprentissage représentant 70% de la base de 2018. Nous le testerons ensuite sur un échantillon de validation contenant les 30% des observations restantes.

Enfin, nous validerons la stabilité dans le temps du modèle en contrôlant son bon fonctionnement et son bon résulat sur la base 2019.

Nous allons commencer par appliquer l'algorithme R du package « random forest » avec les paramètres par défaut et en utilisant l'ensemble des variables utilisables.

La sortie de R nous donne la matrice de confusion suivante :

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 5

OOB estimate of error rate: 8.8%
Confusion matrix :

	Non-résiliés dans l'année	Résiliés dans l'année	class.error
Prediction de non-résiliation	246 585	2 008	0.00807746
Prédiction de résiliation	22 896	11 658	0.66261504

Le paramétrage sélectionné par défaut est celui-ci :

- 500 arbres forment la forêt ;
- A chaque nœud, l'algorithme sélectionne uniquement 5 variables à choisir. Ce paramètre est par défaut √(q), avec q : le nombre de variables du modèle dans le cas d'une classification. Ici, nous avons q=29 variables explicatives.

L'erreur OOB du modèle est 8,8% ce qui est faible. Notre modèle semble plutôt bien prédire nos résiliations ou non-résiliations. Nous observons que nous arrivons très bien à prédire les non-résiliation avec un taux d'échec de moins de 1%. Cependant, nous avons du mal à prédire les résiliations avec un taux d'échec de 66%.

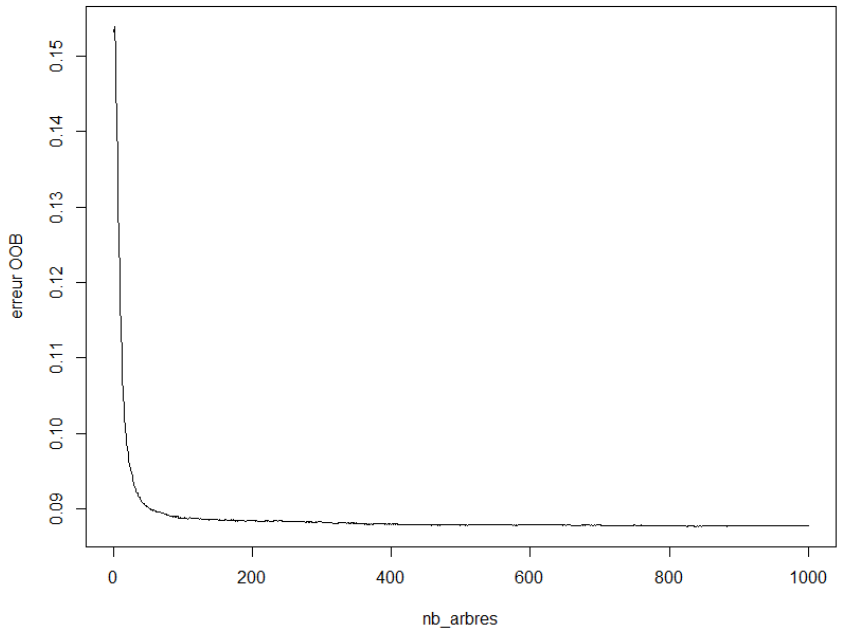
Comme pour la régression logistique, le nombre de résiliation étant faible sur la masse d'observations, le modèle a intérêt à privilégier la non-résiliation.

Dans la pratique, nous parvenons tout de même à identifier 34% des résiliations, ce qui permet de cibler 3 fois mieux qu'en effectuant un tirage aléatoire.

4.2.3.2. Hyperparamétrage

Nous allons tester plusieurs paramétrages par tâtonnement afin d'obtenir un rapport complexité/erreur suffisant.

Nous allons pour cela regarder le taux d'erreur OOB en fonction du nombre d'arbres dans la forêt. Pour cela, nous allons tester une forêt avec un grand nombre d'arbres (1000) et allons regarder à partir de quel nombre d'arbres, le gain d'erreur est devenu quasi-nul.



Graphique 28 : Erreur OOB en fonction du nombre d'arbre

On remarque ainsi que l'erreur OOB commence à se stabiliser aux environs de 150 arbres et continue de progresser jusqu'aux 500 arbres puis le gain devient négligeable. Nous allons donc conserver ce réglage de 500 arbres qui est celui par défaut.

Nous allons ensuite pouvoir régler le paramètre « mtry » qui concerne donc le nombre de variables présélectionnées à chaque nœud. Nous allons tâtonner en multipliant ou divisant par 2 ce paramètre.

Nous obtenons ainsi pour 2 variables :

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2
OOB estimate of error rate: 10.28%

	Non-résiliés dans l'année	Résiliés dans l'année
Prediction de non-résiliation	248062	28572
Prédiction de résiliation	531	5982
Erreur de classe	0.002136022	0.826879667

Et pour 10 variables :

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 10
OOB estimate of error rate: 8.58%

	Non-résiliés dans l'année	Résiliés dans l'année
Prediction de non-résiliation	246012	21699
Prédiction de résiliation	2581	12855
Erreur de classe	0.01038243	0.62797361

En prenant 2 variables à chaque nœud, nous remarquons que le taux d'erreur est plus important et surtout, elle est faite au niveau de la prédiction des résiliations, ce qui est l'élément que nous avons particulièrement du mal à isoler et qui est pourtant notre objectif principal. Nous allons donc écarter cette possibilité.

Le taux d'erreur OOB est légèrement plus faible en prenant un mtry de 10 variables, Ce qui est intéressant, c'est que même si on augmente sensiblement le taux de mauvaise prédiction des non-résiliation (qui est estimée avec plus de 1% d'erreur dans ce modèle), nous avons gagné 3,5 points de bonne prédiction de la résiliation qui est maintenant détectée dans plus de 37% des cas.

Etant donné que la mise en production de notre étude ne nécessitera pas une réponse « rapide » (nous n'avons pas de contraintes de l'ordre de la minute), nous pouvons nous permettre de rechercher le modèle avec le plus faible taux d'erreur.

Number of trees: 500
No. of variables tried at each split: 9
OOB estimate of error rate: 8.58%

	Non-résiliés dans l'année	Résiliés dans l'année
Prediction de non-résiliation	246070	21770
Prédiction de résiliation	2523	12784
Erreur de classe	0.01014912	0.63002836

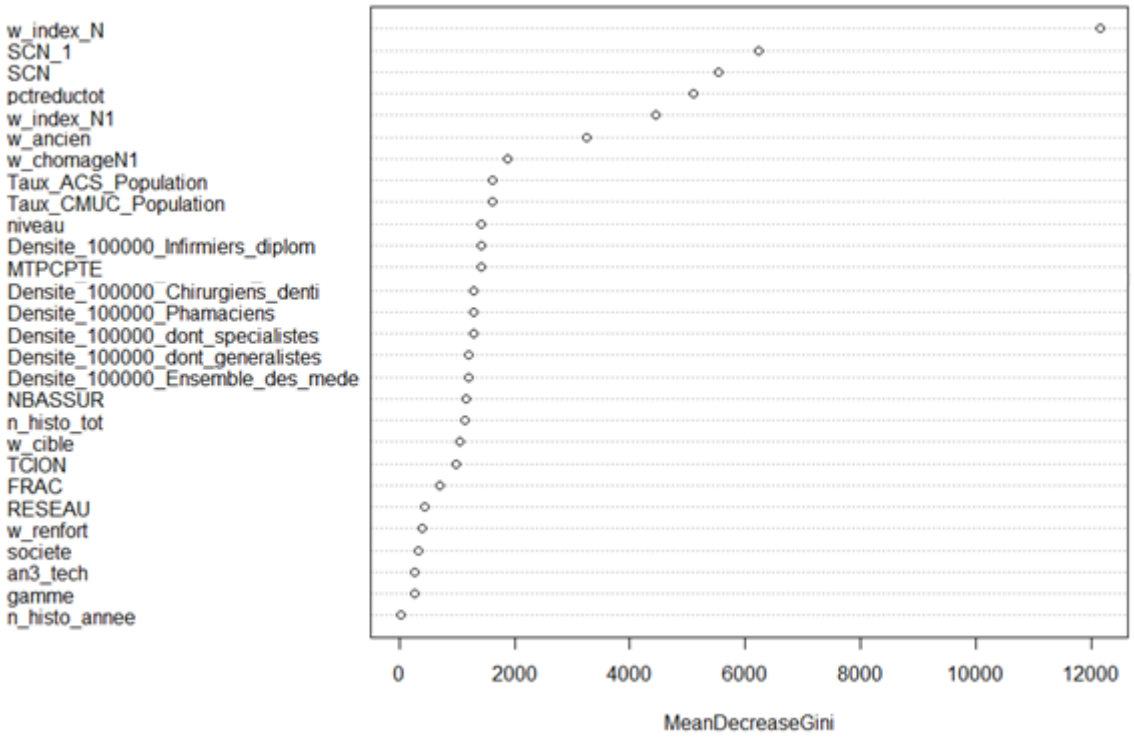
Number of trees: 500
No. of variables tried at each split: 8
OOB estimate of error rate: 8.6%

	Non-résiliés dans l'année	Résiliés dans l'année
Prediction de non-résiliation	246172	21922
Prédiction de résiliation	2421	12632
Erreur de classe	0.00973881	0.63442727

La baisse du nombre de variables diminue légèrement la qualité de la classification. Nous conserverons donc le modèle avec 10 variables par nœud. Nous rappelons que le risque d'un grand nombre de variables sélectionnées à chaque nœud est une trop grande dépendance entre les arbres et donc un surapprentissage. Nous contrôlerons cet aspect sur l'échantillon de validation.

4.2.3.3. Premières interprétations

Afin d'avoir une première interprétation du modèle, nous allons regarder l'impact moyen de chaque variable sur la diminution de l'indice de Gini.

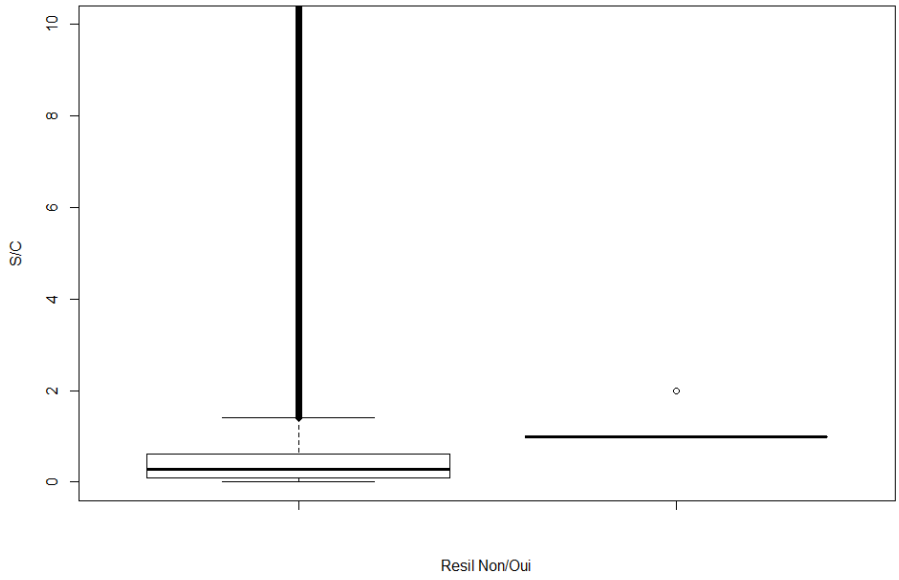


Graphique 29 : Importance des variables dans le modèle suivant le critère de la décroissance de Gini

Nous remarquons ainsi que le critère principal qui permet de classifier de la façon la plus discriminante nos observations est la majoration appliquée l'année précédente. Le modèle ne nous permet pas d'interpréter le sens de l'impact de cette variable, mais nous pouvons supposer que plus un contrat a été majoré l'année précédente, plus l'assuré a de chances de résilier.

Par ailleurs, certains impacts de variables ne semblent pas clairs : nous pouvons interpréter l'impact des S/C ainsi : « lorsqu'un assuré a un S/C faible, il a sûrement l'impression de payer cher pour ce qu'il dépense et souhaitera changer. » ou bien nous pouvons estimer au contraire : « les assurés ayant un S/C fort sont des opportunistes qui 'switchent' entre les différents assureurs pour optimiser leurs remboursements et donc résilient plus ».

Pour essayer de vérifier ces impacts, nous allons faire une analyse empirique, toutes choses égales par ailleurs :



Graphique 30 : sinistralité des contrats résiliés et non résiliés

Il semble qu'en moyenne, le S/C des contrats résiliés est supérieur à celui des contrats non résiliés.

Résiliation	Non-résilié	Résilié
S/C année précédente	49,04 %	80,11 %

Ce serait donc notre seconde hypothèse qui semble tenir la corde, à savoir : les opportunistes qui changent régulièrement de contrat.

A noter qu'on remarque dans la boxplot qu'il y a dans les non-résiliés une quantité non-négligeable d'assurés ayant des forts S/C. Ceux-ci représentent sûrement les assurés ayant été hospitalisés ou ayant des traitements lourds et coûteux. Ceux-ci n'ont donc pas le profil des « opportunistes » de notre hypothèse.

On pourrait supposer par exemple que le modèle classifie selon le S/C en séparant 3 classes :

- Les faibles consommateurs qui ne cherchent pas l'optimisation du contrat ;
- Les opportunistes qui changent de contrat pour profiter de réductions tarifaires ;
- Les forts consommateurs qui ne résilient pas car ayant de forts besoins.

Cette hypothèse ne pourra toutefois pas être vérifiée et on touche ici aux inconvénients quant à la lisibilité du modèle random forest. Dans tous les cas, nous savons que certains arbres de la forêt aléatoire pourraient effectuer la classification ainsi et d'autres non en fonction des variables explicatives retenues à chaque nœud.

Nous notons que 6 variables semblent être particulièrement discriminantes au vu de la décroissance moyenne de Gini :

- Les majorations des 2 dernières années ;
- Les S/C des 2 dernières années ;

- La réduction tarifaire appliquée au contrat ;
- L'ancienneté.

L'ensemble des données extérieures semblent être légèrement en retrait en termes de significativité.

A l'opposé, des variables semblent faiblement intéressantes :

- Le nombre de remplacements dans l'année : cette variable que nous avons créée semble avoir peu d'intérêts ce qui ne semble pas évident au premier abord.
- La gamme : cette variable est fortement corrélée à l'ancienneté, c'est donc rassurant que le modèle ne l'utilise que peu.
- La compagnie d'assurance AXA France Vie ou AXA Assurances Vie Mutuelle : là encore, sa faible prise en compte est rassurante, car l'immense majorité des clients ignorent s'ils sont assurés par la société anonyme ou la mutuelle, car les pièces contractuelles sont similaires. Les conditions tarifaires sont également identiques par conséquent, un fort impact de cette variable aurait révélé un modèle défaillant.

Etant donné que 6 variables semblent particulièrement importantes, nous allons tester un modèle réduit n'utilisant que celles-ci et allons tester s'il est aussi performant que les autres.

Il s'agit des variables :

- Ancienneté ;
- indexation des 2 années précédentes ;
- S/C des 2 années précédentes ;
- Pourcentage de réduction sur le contrat.

Number of trees: 300
No. of variables tried at each split: 2
OOB estimate of error rate: 8.69%

Résultat	Résilié dans l'année	Non-Résilié dans l'année
Prédiction de résiliation	12 958	3 013
Prédiction de non-résiliation	21 596	245 584

Tableau 9 : Matrice de confusion : base d'apprentissage, modèle réduit à 6 variables

Avant toute chose, on remarque que pour ce modèle à 6 variables, l'erreur OOB se stabilise aux alentours de 300 arbres aléatoires dans la forêt. Cela ajouté au plus faible nombre de variables, c'est un réel gain de temps de calcul.

Ainsi, bien que nous ayons un taux d'erreur légèrement plus élevé de 0,09% environ nous pourrions donc utiliser ce modèle afin de gagner du temps.

4.2.3.4. Validation du modèle et stabilité dans le temps

Etant donné que nous n'avons pas de contraintes de temps de calcul, notre modèle étant destiné à être joué une fois par an, nous allons conserver le modèle le plus fiable, c'est-à-dire celui sélectionnant à chaque nœud une variable parmi 10 et modélisant 500 arbres.

Nous allons dans un premier temps nous assurer que le modèle est cohérent sur l'échantillon de validation et n'a pas fait notamment de surapprentissage :

Résultat	Résilié dans l'année	Non-Résilié dans l'année
Prédiction de résiliation	5 531	1 055
Prédiction de non-résiliation	9 406	105 592

Tableau 10 : Matrice de confusion : base de validation 2018, modèle complet

Les résultats sont très bons puisque nous avons une erreur de seulement 8,6% (ce qui est cohérent avec l'erreur OOB du modèle), mais l'erreur de prédiction des résiliations effectives est également de 37%. Nous sommes donc totalement en phase avec les erreurs réalisées sur l'échantillon de test. Le modèle est donc consistant et nous allons pouvoir le tester sur la base de données 2019 afin de contrôler sa stabilité dans le temps.

Cette propriété est essentielle pour nous, car notre modèle serait inutilisable s'il ne durait pas dans le temps : nous ne pourrions pas prédire les résiliations, mais uniquement les analyser.

Résultat	Résilié dans l'année	Non-Résilié dans l'année
Prédiction de résiliation	16 282	9 278
Prédiction de non-résiliation	31 087	359 982

Tableau 11 : Matrice de confusion : base de validation 2019, modèle complet

Nous avons donc un taux d'erreur global de 9,7% et une erreur dans la prédiction des résiliations de 34,4%. Nous avons donc une bonne stabilité des résultats entre les années 2018 et 2019.

La spécificité est égale à

$$Spécificité = \frac{non - résiliés\ prédits\ et\ réalisés}{non - résiliés} = \frac{359\ 982}{(359\ 982 + 9\ 278)} = 97{,}5\%.$$

Notre modèle est donc très spécifique. Il y a peu de faux positifs : la probabilité d'être classé comme résilié dans l'année alors que le contrat ne le sera pas est faible. C'est un avantage pour éviter de disperser les ressources et les actions de rétention possibles auprès d'une population de contrats qui ne souhaitent pas résilier.

La sensibilité est, elle de :

$$Sensibilité = \frac{résiliés\ prédits\ et\ réalisés}{résiliés} = \frac{16\ 282}{(16\ 282 + 31\ 087)} = 34{,}4\%.$$

La sensibilité est moyenne. C'est-à-dire que le modèle ne parvient pas à identifier tous les contrats qui seront résiliés. Nous savons donc qu'en utilisant le modèle pour effectuer des actions de

rétention, nous ne pouvons espérer retenir qu'un tiers des résiliations (et surtout moins si nous considérons un taux d'échec desdites actions).

A noter que la valeur prédictive positive du modèle, c'est-à-dire la quantité de résiliés réellement parmi les prédits comme tel, est de :

$$Valeur\ prédictive\ positive = \frac{résiliés\ prédits\ et\ réalisés}{prédits\ résiliés} = \frac{16\ 282}{(16\ 282 + 9\ 278)} = 63{,}7\%.$$

La valeur prédictive est 5 fois supérieure au taux global de résiliation. Ainsi, en prenant au hasard un contrat, nous avions 12,8% de chances de « piocher » un contrat qui serait résilié dans l'année. Grâce au modèle, ce taux se porte maintenant à 63,7%.

Nous estimons donc que le modèle peut être utilisé pour mieux prédire les futures résiliations.

Chapitre 5. Comparaison des modèles et actions opérationnelles possibles

Nous avons pu dresser un modèle avec chacune des méthodes et allons devoir comparer ceux-ci pour choisir le meilleur, mais également voir les actions concrètes que ce dernier permettra d'aborder afin de répondre aux problématiques opérationnelles que représentent les résiliations des assurés.

5.1. *Comparaison des modèles*

Afin d'identifier la meilleure méthode de prédiction de résiliation, nous allons observer plusieurs critères :

- Taux d'erreur ;
- Efficacité à prédire des résiliations ;
- Stabilité dans le temps ;
- Interprétabilité des résultats.

5.1.1. Taux d'erreur

Les 2 méthodes ont permis de dresser une matrice de confusion, que nous rappelons ici :

		Prédictions	
		Résilié	**Non-Résilié**
Réels	**Résilié**	Vrai Positif (Vp)	Faux Négatif (Fn)
	Non-Résilié	Faux positif (Fp)	Vrai Négatif (Vn)

Tableau 12 : Matrice de confusion : définition

Nous pouvons ainsi définir le taux d'erreur qui nous donnera ainsi la précision « Accuracy » du modèle :

$$\text{Précision} = 1 - Erreur = 1 - \frac{Fp + Fn}{Vp + Fn + Fp + Vn}.$$

La régression logistique, avec le modèle « balanced data » a permis d'obtenir un taux d'erreur de 34,0 %.

La random forest a, elle, permis d'atteindre un taux d'erreur de 8,6% sur les prédictions de 2019.

Ainsi le modèle random forest pourrait sembler meilleur. Cependant, l'accuracy n'est pas un bon moyen de déterminer l'efficacité d'un modèle. En effet, ici, le taux de résiliation étant d'environ 11,4%, un modèle qui prédirait dans tous les cas une non-résiliation aurait une accuracy de 88,6%. Et pourtant il ne prédirait aucune résiliation.

5.1.2. Efficacité à prédire une résiliation

L'accuracy ne nous permet pas de choisir un modèle réellement efficace. En effet, nous avons besoin de prédire des résiliations et pas uniquement de prédire au mieux des non-résiliations.

Afin de prendre en compte ce besoin, nous pouvons par exemple utiliser le F-Score qui se présente ainsi :

$$F - Score = \frac{Vp}{Vp + \frac{1}{2(Fp + Fn)}}.$$

Le F-score ne prend pas en compte les Vrai négatifs (les contrats correctement prédit comme non-résiliés). Il est donc plus adapté à déterminer le bon modèle.

On a ainsi :

F-score(Logistique_BalancedData) = 0,260.

F-score(RandomForest) = 0,447.

Le random Forest est donc bien plus efficace à prédire la résiliation effective. En effet, la régression logistique a beaucoup de mal à prédire une résiliation en limitant les erreurs.

5.1.3. Stabilité dans le temps

Comme vu au chapitre 4.2.3.4, le modèle de random forest reste bien stable dans le temps avec des taux d'erreurs et de prédiction quasi-identiques entre les prédictions sur les échantillons d'apprentissage et de validation de l'année 2018 et sur celles de l'année 2019. Le modèle remplit donc correctement cette propriété.

Le modèle de régression logistique sur les « balanced data » est moins stable, car les résultats sont assez différents entre 2018 et 2019, mais il semble être en mesure de conserver une capacité de prédiction sur 2019.

5.1.4. Interprétabilité des résultats

La régression logistique est parfaitement interprétable : les sorties du logiciel nous permettent d'interpréter et même de recalculer les prédictions. Nous pouvons précisément savoir le poids de chaque variable.

Le modèle de Random Forest ne nous permet pas de connaître précisément les contributions des variables au modèle. Le critère de Gini nous permet de connaître leur importance, mais pas le sens de la contribution.

Nous pourrions imaginer fusionner les 2 modèles et utiliser l'interprétation de la régression logistique pour connaître les critères qui amènent à la résiliation des clients. Cependant cette hypothèse est à exclure : en effet, la régression logistique ayant un fonctionnement totalement différent de la random forest, il ne serait pas juste d'étendre les interprétations de l'un des modèles au second.

5.1.5. Choix du modèle

Le modèle de régression logistique initial, qui a été complètement « embrouillé » par le nombre bien supérieur de non-résiliation, a pu être partiellement redressé par le modèle « balanced data ». Malgré cela, ce dernier prédit les résiliations avec un très grand nombre de faux positifs sur les données que nous avons en stock.

La méthode s'avèrerait donc pas assez précise et trop coûteuse pour une utilisation ultérieure et se trouve donc disqualifiée.

De son côté, la random forest sans être brillante, a permis d'identifier plus d'un tiers des résiliations. Ce niveau est bien supérieur à l'aléatoire. La faible interprétabilité du modèle restant son seul défaut, celle-ci peut être compensée par les critères de décroissance de Gini couplée à quelques vérifications.

C'est donc le modèle que nous retiendrons dans le cadre de notre étude.

5.2. *Actions opérationnelles possibles*

Nous sommes parvenus à obtenir un modèle permettant d'isoler avec environ 2 chances sur 3 les contrats qui devraient être résiliés dans l'année. Sans être un résultat totalement satisfaisant, nous sommes en mesure de toucher de façon plus efficace nos clients « à risque ».

En prévoyant des actions sur la totalité de ces assurés, nous toucherions, environ 1/3 des contrats qui seraient résiliés dans l'année. En supposant que nos actions fonctionnent dans 50% des cas, c'est 1/6 des résiliations qui pourraient être évitées.

Nous proposons une série d'actions basées sur les variables ayant le plus d'impacts au sens de la décroissance moyenne de Gini dans le modèle Random Forest.

L'indexation de l'année précédente étant la variable avec le plus d'impact, nous suggérons de proposer un avantage en nature aux assurés sélectionnés par le modèle. Nous pourrions imaginer :

- Leur fournir par exemple un « chèque » pour une séance médicale chez un ostéopathe ou bien un supplément de garantie pour l'année qui vient. Cela permettrait de signifier à l'assuré que malgré la forte majoration, nous souhaitons rester à ses côtés pour l'assurer.
- Leur donner un engagement de non-majoration l'année suivante dès le début de l'année afin de les assurer que la majoration est nécessaire, mais ne sera pas réitérée chaque année.

Si ces mesures sont trop coûteuses nous pourrions imaginer créer des groupes d'assurés et ne cibler que les groupes qui ont un S/C faible. En effet, nous avons vu que les assurés opportunistes étaient plus à-même de « switcher », cependant, étant donné leur faible rentabilité, nous aurions intérêt à mettre des barrières à l'entrée de ceux-ci dans nos portefeuilles.

Cela pourrait s'effectuer de plusieurs façons :

- Ne pas permettre à un assuré qui nous a quitté de revenir dans l'année qui suit.

- Mettre des frais à la souscription qui seraient déduits de la cotisation annuelle lissée dans le temps : bien qu'il y ait des risques de perte de souscription, cela permettrait de garantir une souscription de qualité.
- Augmenter la progressivité des garanties : c'est une forme de délai d'attente caché, mais cela s'avère efficace pour conserver les clients. Nous pourrions élargir le fait de donner des meilleures garanties les 2èmes, 3èmes et 4èmes années à plus de postes que simplement en optique et en dentaire comme cela se fait actuellement.
- Permettre aux clients d'accumuler une cagnotte qui pourrait être dépensé en cas de gros reste à charge lors de la prise en charge partielle d'un soin coûteux. Cette cagnotte pourrait être rechargée chaque année d'un certain montant. Cela inciterait les assurés à rester dans le contrat.

Une mesure intéressante, bien que plus coûteuse en analyse, serait d'étudier profondément les profils de sinistralité et de remboursement des assurés lors de l'année écoulée afin de voir si la formule qu'ils ont choisie est adaptée ou si une autre formule leur aurait fait gagner de l'argent car mieux adaptée en rapport coût/remboursement. Cela permettrait de fournir un meilleur conseil au client et donc de se positionner comme conseiller avec une vraie valeur ajoutée.

Dans tous les cas, il serait intéressant de fournir les données des clients identifiés à leurs distributeurs afin de leur transmettre l'information que leur client risque de changer d'assurance et donc de les quitter. Cela permettrait aux conseillers de contacter lesdits clients afin de pouvoir éventuellement leur proposer une nouvelle offre plus adaptée à ce qu'ils recherchent ou de leur proposer une réduction tarifaire issue de leur enveloppe commerciale.

Conclusion

Bien que notre modèle final nous permette d'identifier une part significative (environ 1/3) des futures résiliations avec une bonne probabilité, on remarque que notre modèle ne semble pas suffisamment spécifique pour permettre d'isoler de façon plus précise les contrats qui sont résiliés dans l'année.

Nous avons envisagé plusieurs pistes d'amélioration de notre modèle pour de futurs travaux.

Dans un premier temps, nous pourrions envisager de récupérer des données par exemple issues des fichiers clients qui tiendraient compte des interactions entre les clients et AXA. Par exemple, le nombre d'appels répondus ou non, le nombre de réclamations, etc. Nous pourrions également récupérer des données d'éventuelles connexions de l'assuré sur l'espace de devis d'axa.fr ou d'autres sites d'assurance ou des comparateurs.

Par la suite, nous pourrions envisager d'effectuer une ACP avant l'application de la méthode random forest sur nos données. Cela permettrait de maximiser la variance entre les données. Et d'éliminer les corrélations. La contrepartie étant de perdre l'intégralité de la possibilité d'analyser les sorties du modèle.

Par une autre approche, nous pouvons utiliser de nouveaux modèles tels que les réseaux de neurones ou les SVM, mais aussi redonner sa chance à la régression logistique. Par exemple en modifiant la fonction afin de mieux prédire les résiliations en utilisant par exemple un algorithme qui prendrait en compte un critère issu du F-score (ou autre critère similaire) au lieu de l'AIC.

Enfin, tout notre modèle pourrait être à revoir. En effet, comme énoncé dans la partie *1.5 La résiliation infra-annuelle*, la réforme (qui prend effet au 1er décembre 2020) risque de fortement bouleverser le champ des résiliations, et particulièrement en santé individuelle, où le démarchage par la concurrence pourra s'avérer bien plus efficace : en effet, la concurrence pourra être mandatée pour résilier le contrat d'un client, non plus à la date d'échéance, mais dans un délai d'un mois.
L'avantage de notre modèle, est qu'il ne prédit pas les résiliations dans un délai court, mais prédit les résiliations à venir dans l'année. Il semble donc mieux armé pour résister aux modifications de comportement : on peut ainsi supposer que les souscripteurs identifiés ou non identifiés auront le même comportement et que seule la date de leur résiliation pourrait être modifiée dans l'année.
Si cette supposition s'avère exacte, alors le modèle restera utilisable en 2021. En revanche, il semble évident qu'une révision du modèle sera nécessaire en 2022, pour revalider le modèle sur la base des résiliations survenues en 2021. Un recalibrage sera peut-être nécessaire, voire une refonte totale du modèle si nos suppositions s'avèrent fausses.

Bibliographie

[1] AKAIKE, H., (1998), *Selected Papers of Hirotugu Akaike,* Springer, New York, NY

[2] BREIMAN, L., FRIEDMAN, J., OLSHEN, R., and STONE, C., (1984), *Classification and Regression Trees*. Chapman & Hall, New York.

[3] BREIMAN, L., (2001) *Random Forests*, Machine Learning, vol. 45, no 1

[4] GENUER, R., POGGI, J.-M., (2017) *Arbres CART et Forêts aléatoires, Importance et sélection de variables*, hal-01387654v2.

[5] INSTITUT DES ACTUAIRES, (2017), *Norme de Pratique relative à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel - NPA 5.*

[6] MCFADDEN D., (1974), *Conditional Logit Analysis of Qualitative Choice Behaviour*, Institute of Urban and Regional Development, University of California.

Mémoires d'actuariat

[7] BRUN, A (2017), *Modélisation et étude d'impacts du phénomène de résiliation en assurance emprunteur*, Mémoire d'actuariat, Université Louis Pasteur de Strasbourg.

[8] DUTEL, A., (2017), *Analyse des données pour la prise de décision opérationnelle : exemple de la dynamique de résiliation sur un contrat d'assurance complémentaire santé individuelle,* Mémoire d'actuariat, Centre d'études actuarielles.

[9] GEHLER, A., (2009), *Etude des profils de résiliation sur un portefeuille santé individuelle,* Mémoire d'actuariat, Université Louis Pasteur de Strasbourg.

[10] VALLA, M., (2018) *Modélisations du risque de résiliation dans un portefeuille de santé individuelle,* Mémoire d'actuariat, ISFA.

Document de travail

[11] MOUTARDE F., (2017), *Arbres de décision et Forêts aléatoires*, PSL

[12] RAKOTOMALALA, R., (2017), *Pratique de la régression logistique*, Université Lyon 2

[13] VEREL, S., (2016), *Outils avancés du data scientist*, LISIC

Webographie

[14] Site de la Sécurité sociale française, https://www.securite-sociale.fr

[15] WIKISTAT, *Agrégation de modèles*, http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf

Langages

[16] R CORE TEAM (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Version 4.0.3.

[17] WPS Workbench Version: 3.3.1.0.21515 © Copyright World Programming Limited 2002-2017. All rights reserved.

Annexes

Annexe 1 : Résumé de la base de données

```
> summary(base_2018)
     TCION              DTEFFAN            DTEFFRES           DTEMIRES          NBASSUR       FRAC          MTPCPTE           RESEAU       societe     gamme
 Min.   : 0.000   2018-01-01: 23037   2019-01-01: 10728   2018-11-20:  1129   Min.   :1.000   1: 35630   Min.   : -13.56   1    :300665   MU: 59154   MLO:267141
 1st Qu.:10.000   2017-01-01: 22283   2020-01-01:  8474   2019-11-19:   756   1st Qu.:1.000   2: 13591   1st Qu.:   0.00   2    : 10108   SA:345583   NGM:137596
 Median :10.000   2016-01-01: 19008   2018-04-01:  2462   2018-10-23:   646   Median :1.000   4: 13746   Median :   0.00   3    : 93958
 Mean   : 8.668   2015-01-01: 12243   2018-03-01:  2355   2018-04-17:   640   Mean   :1.515   8:341770   Mean   :  64.93   Autre:     6
 3rd Qu.:12.000   2014-01-01: 10736   2018-10-01:  2255   2018-03-20:   629   3rd Qu.:2.000              3rd Qu.:  66.42
 Max.   :20.000   2010-01-01:  7668   (Other)   : 79753   (Other)   :102227   Max.   :6.000              Max.   :4756.10
                  (Other)   :309762   NA's      :298710   NA's      :298710
 an3_tech      formule       w_renfort      w_dept        n_histo_tot      n_histo_annee    w_cible        w_index_N          w_index_N1          w_ancien
 N:316973   F22    :115167   N:215713   33     : 11848   Min.   : 1.000   Min.   :1.000   autr:123785   Min.   :-0.90500   Min.   :-0.84300   Min.   :    1
 O: 87764   F21    : 66202   O:189024   76     : 10292   1st Qu.: 1.000   1st Qu.:1.000   pros:148248   1st Qu.: 0.04300   1st Qu.: 0.01500   1st Qu.:  732
            F23    : 56229              59     :  9835   Median : 1.000   Median :1.000   seni:132704   Median : 0.05400   Median : 0.04900   Median : 1462
            F04    : 38963              35     :  9787   Mean   : 1.485   Mean   :1.002                 Mean   : 0.04344   Mean   : 0.03724   Mean   : 2083
            F03    : 29393              06     :  9673   3rd Qu.: 2.000   3rd Qu.:1.000                 3rd Qu.: 0.06000   3rd Qu.: 0.05800   3rd Qu.: 2982
            F07    : 28999              74     :  9090   Max.   :10.000   Max.   :9.000                 Max.   : 0.50000   Max.   : 0.50000   Max.   :22917
            (Other): 69784              (Other):344212
  w_chomage_N     Taux_ACS_Population Taux_CMUC_Population  Dst_medecin    Dst_generaliste Dst_specialiste Dst_Chirur_Dent  Dst_Infir_diplom Dst_Pharmaciens
 Min.   : 5.150   Min.   :0.007607    Min.   :0.02887      Min.   : 87.0   Min.   : 53.0   Min.   : 34.0   Min.   :  7.00   Min.   : 342     Min.   : 28.0
 1st Qu.: 8.700   1st Qu.:0.020645    1st Qu.:0.05720      1st Qu.:255.0   1st Qu.:137.0   1st Qu.:118.0   1st Qu.: 48.00   1st Qu.: 885     1st Qu.:101.0
 Median : 9.083   Median :0.024241    Median :0.06727      Median :324.0   Median :156.0   Median :167.0   Median : 60.00   Median :1048     Median :109.0
 Mean   : 9.083   Mean   :0.025148    Mean   :0.07138      Mean   :328.1   Mean   :153.9   Mean   :174.2   Mean   : 62.25   Mean   :1036     Mean   :112.1
 3rd Qu.: 9.083   3rd Qu.:0.028470    3rd Qu.:0.07992      3rd Qu.:362.0   3rd Qu.:171.0   3rd Qu.:197.0   3rd Qu.: 72.00   3rd Qu.:1156     3rd Qu.:121.0
 Max.   :16.550   Max.   :0.079008    Max.   :0.36379      Max.   :846.0   Max.   :246.0   Max.   :600.0   Max.   :140.00   Max.   :1702     Max.   :187.0

     SCN_N1            SCN_N           pctreductot          niveau        r_DateRes            r_Resil         r_ancien_m       r_ancien_a
 Min.   : 0.00000   Min.   :   0.0000   Min.   :-1.5200   125    :154143   Min.   :2018-01-02   Min.   :0.0000   Min.   :   0.0   Min.   : 0.000
 1st Qu.: 0.05369   1st Qu.:   0.0841   1st Qu.: 0.0669   100    :112093   1st Qu.:2018-07-31   1st Qu.:0.0000   1st Qu.:  61.0   1st Qu.: 2.000
 Median : 0.25540   Median :   0.2884   Median : 0.1490   150    : 97450   Median :2019-01-16   Median :0.0000   Median : 122.0   Median : 4.000
 Mean   : 0.46014   Mean   :   0.5268   Mean   : 0.1517   Hospi  : 19467   Mean   :2019-02-21   Mean   :0.1223   Mean   : 173.5   Mean   : 5.716
 3rd Qu.: 0.57370   3rd Qu.:   0.6149   3rd Qu.: 0.2238   200    : 14905   3rd Qu.:2019-10-07   3rd Qu.:0.0000   3rd Qu.: 248.0   3rd Qu.: 8.000
 Max.   :64.64179   Max.   :1609.3890   Max.   : 1.2097   Eco    :  5993   Max.   :2020-07-03   Max.   :1.0000   Max.   :1910.0   Max.   :63.000
                                                          (Other):   686   NA's   :298710
```

Annexe 2 : Sélectionner les variables

```
> fitAlltBo<-step(fitAllt_2018,direction="both")
Start:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_medecin + Dst_generaliste + Dst_specialiste + Dst_Chirur_Dent +
    Dst_Infir_diplom + Dst_Pharmaciens + SCN_N1 + niveau + RESEAU:TCION


Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_medecin + Dst_generaliste + Dst_specialiste + Dst_Chirur_Dent +
    Dst_Infir_diplom + SCN_N1 + niveau + RESEAU:TCION


Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_medecin + Dst_generaliste + Dst_specialiste + Dst_Chirur_Dent +
    SCN_N1 + niveau + RESEAU:TCION


Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_medecin + Dst_generaliste + Dst_specialiste + SCN_N1 +
    niveau + RESEAU:TCION


Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_medecin + Dst_generaliste + SCN_N1 + niveau + RESEAU:TCION
```

```
Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    Dst_generaliste + SCN_N1 + niveau + RESEAU:TCION


Step:  AIC=202302.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    w_chomage_N + Taux_ACS_Population + Taux_CMUC_Population +
    SCN_N1 + niveau + RESEAU:TCION

                        Df Deviance    AIC
- w_chomage_N            1   202021 202301
- Taux_ACS_Population    1   202021 202301
- w_index_N1             1   202022 202302
<none>                       202020 202302
- societe                1   202023 202303
- Taux_CMUC_Population   1   202024 202304
- SCN_N1                 1   202026 202306
- NBASSUR                1   202028 202308
- RESEAU:TCION           3   202038 202314
- n_histo_tot            1   202035 202315
- r_ancien_m             1   202056 202336
- gamme                  1   202074 202354
- w_renfort              1   202085 202365
- w_dept               105   202345 202417
- MTPCPTE                1   202155 202435
- n_histo_annee          1   202183 202463
- niveau                 7   202227 202495
- w_index_N              1   202440 202720
- pctreductot            1   202508 202788
- FRAC                   3   202637 202913
- an3_tech               1   202667 202947
- SCN_N                  1   203243 203523
- w_cible                2   204425 204703
```

```
Step:  AIC=202300.8
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    Taux_ACS_Population + Taux_CMUC_Population + SCN_N1 + niveau +
    RESEAU:TCION

                         Df Deviance    AIC
- Taux_ACS_Population     1   202021 202299
- w_index_N1              1   202022 202300
<none>                        202021 202301
- societe                 1   202023 202301
+ w_chomage_N             1   202020 202302
- Taux_CMUC_Population    1   202024 202302
- SCN_N1                  1   202027 202305
- NBASSUR                 1   202028 202306
- RESEAU:TCION            3   202039 202313
- n_histo_tot             1   202036 202314
- r_ancien_m              1   202056 202334
- gamme                   1   202075 202353
- w_renfort               1   202085 202363
- w_dept                105   202346 202416
- MTPCPTE                 1   202155 202433
- n_histo_annee           1   202183 202461
- niveau                  7   202227 202493
- w_index_N               1   202440 202718
- pctreductot             1   202509 202787
- FRAC                    3   202637 202911
- an3_tech                1   202667 202945
- SCN_N                   1   203244 203522
- w_cible                 2   204427 204703

Step:  AIC=202299.4
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 +
    Taux_CMUC_Population + SCN_N1 + niveau + RESEAU:TCION

                         Df Deviance    AIC
- w_index_N1              1   202023 202299
<none>                        202021 202299
- societe                 1   202024 202300
- Taux_CMUC_Population    1   202024 202300
- SCN_N1                  1   202027 202303
- NBASSUR                 1   202029 202305
- RESEAU:TCION            3   202040 202312
```

```
- n_histo_tot            1   202036 202312
- r_ancien_m             1   202057 202333
- gamme                  1   202075 202351
- w_renfort              1   202086 202362
- w_dept               105   202352 202420
- MTPCPTE                1   202156 202432
- n_histo_annee          1   202184 202460
- niveau                 7   202228 202492
- w_index_N              1   202441 202717
- pctreductot            1   202510 202786
- FRAC                   3   202638 202910
- an3_tech               1   202667 202943
- SCN_N                  1   203244 203520
- w_cible                2   204428 204702

Step:  AIC=202298.7
r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE + pctreductot +
    RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
    n_histo_tot + n_histo_annee + w_cible + w_index_N + Taux_CMUC_Population +
    SCN_N1 + niveau + RESEAU:TCION

                       Df Deviance    AIC
<none>                      202023 202299
- societe               1   202025 202299
+ w_index_N             1   202021 202299
- Taux_CMUC_Population  1   202026 202300
- SCN_N1                1   202029 202303
- NBASSUR               1   202030 202304
- RESEAU:TCION          3   202041 202311
- n_histo_tot           1   202038 202312
- r_ancien_m            1   202059 202333
- gamme                 1   202076 202350
- w_renfort             1   202087 202361
- w_dept              105   202353 202419
- MTPCPTE               1   202156 202430
- n_histo_annee         1   202185 202459
- niveau                7   202230 202492
- w_index_N             1   202442 202716
- pctreductot           1   202510 202784
- FRAC                  3   202639 202909
- an3_tech              1   202669 202943
- SCN_N                 1   203246 203520
- w_cible               2   204428 204700
```

Annexe 3 : Codes R du modèle GLM

```
fitAllt_2018<-glm(r_Resil ~ RESEAU:TCION+NBASSUR+FRAC+MTPCPTE+RESEAU+
        societe+gamme+an3_tech+w_renfort+w_dept+n_histo_tot+n_histo_annee+
        w_cible+w_index_N+w_index_N1+w_chomage_N+Taux_ACS_Population+
        Taux_CMUC_Population+Dst_medecin+Dst_generaliste+Dst_specialiste+
        Dst_Chirur_Dent+Dst_Infir_diplom+Dst_Pharmaciens+SCN_N+SCN_N1+
        pctreductot+niveau+r_ancien_m,
        family=binomial(link='logit'), data=baseT_2018)
```

Modèle initial complet

```
modelV.null <- fitAllt_2018$null.deviance/-2
modelV.pred <- fitAllt_2018$deviance/-2
(modelV.null-modelV.pred)/modelV.null
[1] 0.0411064
```

Calcul de R^2_{MF}

```
fitAlltBo<-step(fitAllt_2018,direction="both")
```

Optimisation par le critère AIC

```
fitAlltR_2018<-glm(r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE +
            pctreductot + RESEAU + societe + gamme + an3_tech + w_renfort + w_dept +
            n_histo_tot + n_histo_annee +w_cible + w_index_N +
            Taux_CMUC_Population + SCN_N1 + niveau + RESEAU:TCION,
        family=binomial(link='logit'), data=baseT_2018)
```

Modèle optimisé par AIC

```
final_2018<-glm(r_Resil ~ NBASSUR + SCN_N + FRAC + r_ancien_m + MTPCPTE +
            pctreductot + RESEAU + gamme + an3_tech + w_renfort +
            n_histo_tot + w_cible + w_index_N + niveau,
        family=binomial(link='logit'), data=baseT_2018)
```

Modèle retenu sur l'ensemble des données

```
y=baseT_2018$r_Resil-trunc(2*final_2018$fitted)
reussi=sum(y==0)
tauxreussi=reussi/length(y)
tauxreussi
[1] 0.8777202
```

Calcul de la précision du modèle

```
baseV_2018.pred<- predict(final_2018, type="response", newdata=baseV_2018)
```

Validation du modèle sur l'échantillon de test 2018

```
fitAllt_2018_BD <-glm(r_Resil ~ NBASSUR+FRAC+MTPCPTE+RESEAU+societe+
        gamme+an3_tech+w_renfort+w_dept+n_histo_tot+n_histo_annee+
        w_cible+w_index_N+w_index_N1+w_chomage_N+Taux_ACS_Population+
        Taux_CMUC_Population+Dst_medecin+Dst_generaliste+Dst_specialiste+
        Dst_Chirur_Dent+Dst_Infir_diplom+Dst_Pharmaciens+SCN_N+SCN_N1+
        pctreductot+niveau+r_ancien_m,
        family=binomial(link='logit'), data=baseT _2018_BD)
fitAlltBo_BD <-step(fitAllt_2018_BD, direction="both")
final_2018_BD <-glm(r_Resil ~ FRAC+MTPCPTE+RESEAU+ gamme+an3_tech+
        w_renfort+n_histo_tot+w_cible+w_index_N+SCN_N+pctreductot+niveau,
        family=binomial(link='logit'), data=baseT _2018_BD)
```

Modélisation sur la base de données rééquilibrée

Annexe 4 : Codes R random forest

```
rand_forest<-randomForest(Resil~TCION + w_ancien + NBASSUR + FRAC + MTPCPTE + RESEAU + societe + gamme + an3_tech + w_renfort + n_histo_tot + n_histo_annee + w_cible + w_index_N1 + w_index_N + w_chomageN1 + Taux_ACS_Population + Taux_CMUC_Population + Densite_100000_Ensemble_des_mede + Densite_100000_dont_generalistes + Densite_100000_dont_specialistes + Densite_100000_Chirurgiens_denti + Densite_100000_Infirmiers_diplom + Densite_100000_Phamaciens + SCN + SCN_1 + pctreductot + niveau ,data=tdata, na.action=na.omit)
```

Application du modèle random forest par défaut avec l'ensemble des variables

```
rand_forest1<-randomForest(Resil~ TCION + w_ancien + NBASSUR + FRAC + MTPCPTE + RESEAU + societe + gamme + an3_tech + w_renfort + n_histo_tot + n_histo_annee + w_cible + w_index_N + w_index_N1 + w_chomageN1 + Taux_ACS_Population + Taux_CMUC_Population + Densite_100000_Ensemble_des_mede + Densite_100000_dont_generalistes + Densite_100000_dont_specialistes + Densite_100000_Chirurgiens_denti + Densite_100000_Infirmiers_diplom + Densite_100000_Phamaciens + SCN + SCN_1 + pctreductot + niveau,

data=tdata, na.action=na.omit, ntree=1000)

plot(rand_forest1$err.rate[,1], type="l", xlab = "nb_arbres", ylab = "erreur OOB")
```

Hyperparamétrage du nombre d'arbres optimal

```
> rand_forest5<-randomForest(Resil~ w_ancien + w_index_N1 + w_index_N + SCN + SCN_1 + pctreductot,data=tdata,na.action=na.omit, ntree=1000)
> plot(rand_forest5$err.rate[,1], type="l", xlab = "nb_arbres", ylab = "erreur OOB")
> rand_forest6<-randomForest(Resil~ w_ancien + w_index_N1 + w_index_N + SCN + SCN_1 + pctreductot,data=tdata,na.action=na.omit, ntree=300)
> print(rand_forest6)
```

Application d'un modèle réduit de random forest

```
> rand_forest4<-randomForest(Resil~TCION + w_ancien + NBASSUR + FRAC + MTPCPTE + RESEAU + societe + gamme + an3_tech + w_renfort + n_histo_tot + n_histo_annee + w_cible + w_index_N1 + w_index_N + w_chomageN1 + Taux_ACS_Population + Taux_CMUC_Population + Densite_100000_Ensemble_des_mede + Densite_100000_dont_generalistes +Densite_100000_dont_specialistes + Densite_100000_Chirurgiens_denti + Densite_100000_Infirmiers_ Densite_100000_Phamaciens + SCN + SCN_1 + pctreductot + niveau, data=tdata,na.action=na.omit, ntree=500,mtry = 10)
```

Modèle final retenu

Annexe 5 : Matrice de coefficient de corrélation

```
> cor(data.indicator)
```

	r_Resil	TCION	FRAC.FRAC1	FRAC.FRAC2	FRAC.FRAC4	FRAC.FRAC8	MTPCPTE	RESEAU.RESEAU1	RESEAU.RESEAU2	RESEAU.RESEAU3	RESEAU.RESEAUAutre	gamme.gammeML0	gamme.gammeNGM	an3_tech.an3_techN
r_Resil	1.0000000000	0.020434130	0.032318107	-0.0134762861	0.0097490606	-0.0234383454	-0.034463250	0.009729934	0.0161422563	-0.016025335	-1.437113e-03	-0.007910265	0.007910265	0.043185518
TCION	0.0204341297	1.000000000	0.101807712	0.0808875644	0.0497033831	-0.1446283700	0.129779509	0.869820431	0.1579179087	-0.958725861	-6.713974e-03	-0.228160961	0.228160961	-0.230150442
FRAC.FRAC1	0.0323181075	0.101807712	1.000000000	-0.0579145958	-0.0582554498	-0.7238394326	-0.043152964	0.092131399	-0.0101609118	-0.091626789	1.068545e-03	-0.045304572	0.045304572	-0.112126096
FRAC.FRAC2	-0.0134762861	0.080887564	-0.057914596	1.0000000000	-0.0349511161	-0.4342768986	-0.028074912	0.067052815	-0.0015315165	-0.068839104	-7.177094e-04	-0.081202182	0.081202182	-0.136286139
FRAC.FRAC4	0.0097490606	0.049703383	-0.058255450	-0.0349511161	1.0000000000	-0.4368328173	-0.006533056	0.029948314	0.0321419833	-0.042875041	-7.219335e-04	-0.107087171	0.107087171	-0.051414349
FRAC.FRAC8	-0.0234383454	-0.144628370	-0.723839433	-0.4342768986	-0.4368328173	1.0000000000	0.050952844	-0.120315439	-0.0073584318	0.127268845	-1.178242e-04	0.129291391	-0.129291391	0.181083738
MTPCPTE	-0.0344632499	0.129779509	-0.043152964	-0.0280749120	-0.0065330562	0.0509528440	1.000000000	0.250178960	0.0057831394	-0.261100924	-1.828352e-03	0.032903906	-0.032903906	0.248226538
RESEAU.RESEAU1	0.0097299343	0.869820431	0.092131399	0.0670528149	0.0299483139	-0.1203154390	0.250178960	1.000000000	-0.2720273717	-0.934577964	-6.544355e-03	-0.219453798	0.219453798	-0.185471778
RESEAU.RESEAU2	0.0161422563	0.157917909	-0.010160912	-0.0015315165	0.0321419833	-0.0073584318	0.005783139	-0.272027372	1.0000000000	-0.087999290	-6.162125e-04	-0.010416734	0.010416734	0.041181881
RESEAU.RESEAU3	-0.0160253346	-0.958725861	-0.091626789	-0.0688391043	-0.0428750414	0.1272688452	-0.261100924	-0.934577964	-0.0879992899	1.000000000	-2.117061e-03	0.230998925	-0.230998925	0.176757202
RESEAU.RESEAUAutre	-0.0014371130	-0.006713974	0.001068545	-0.0007177094	-0.0007219335	-0.0001178242	-0.001828352	-0.006544355	-0.0006162125	-0.002117061	1.000000e+00	0.002763281	-0.002763281	0.002026000
gamme.gammeML0	-0.0079102653	-0.228160961	-0.045304572	-0.0812021824	-0.1070871711	0.1292913906	0.032903906	-0.219453798	-0.0104167336	0.230998925	2.763281e-03	1.000000000	-1.000000000	-0.021068804
gamme.gammeNGM	0.0079102653	0.228160961	0.045304572	0.0812021824	0.1070871711	-0.1292913906	-0.032903906	0.219453798	0.0104167336	-0.230998925	-2.763281e-03	-1.000000000	1.000000000	0.021068804
an3_tech.an3_techN	0.0431855180	-0.230150442	-0.112126096	-0.1362861387	-0.0514143493	0.1810837385	0.248226538	-0.185471778	0.0411818810	0.176757202	2.026000e-03	-0.021068804	0.021068804	1.000000000
an3_tech.an3_tech0	-0.0431855180	0.230150442	0.112126096	0.1362861387	0.0514143493	-0.1810837385	-0.248226538	0.185471778	-0.0411818810	-0.176757202	-2.026000e-03	0.021068804	-0.021068804	-1.000000000
w_renfort.w_renfortN	0.0365022286	0.195484558	0.048156325	0.0459472948	0.0604436850	-0.0906885860	0.060981537	0.174013837	0.0290614434	-0.190839222	-4.113127e-03	0.570139600	-0.570139600	0.041010826
w_renfort.w_renfort0	-0.0365022286	-0.195484558	-0.048156325	-0.0459472948	-0.0604436850	0.0906885860	0.060981537	-0.174013837	-0.0290614434	0.190839222	4.113127e-03	0.570139600	-0.570139600	-0.041010826
w_cible.w_cibleautr	0.1228358681	0.116384696	-0.079846999	-0.0900340992	-0.0346096528	0.1244638500	-0.168509746	0.085938533	0.0324583790	-0.100947888	-1.162982e-03	0.006740440	-0.006740440	0.070445805
w_cible.w_ciblepros	-0.0660691258	-0.043735129	0.153288910	0.0850852213	0.0083261301	-0.1662813717	0.014272309	-0.024526623	-0.0356696968	0.038575453	-2.633167e-04	0.043358200	-0.043358200	-0.064867802
w_cible.w_cibleseni	-0.0527583356	-0.069347877	-0.078949924	0.0010465894	0.0254247793	0.0484919961	0.150747243	-0.059178254	0.0047495398	0.059491803	1.411728e-03	-0.051114540	0.051114540	-0.002569587
w_index_N	-0.0395312802	-0.022107160	-0.007136608	-0.0149038896	-0.0117122657	0.0188395843	-0.021801947	-0.019548229	0.0044334834	0.018539328	6.357170e-03	0.023038337	-0.023038337	0.018852752
w_index_N1	0.0006782824	0.031640094	0.003107009	-0.0031964714	-0.0041807831	0.0012491624	-0.030436144	0.028860629	0.0123659075	-0.034446075	-3.539673e-05	-0.062892186	0.062892186	-0.016299028
SCN_N	0.0338024362	-0.005122057	-0.006302072	-0.0014322755	-0.0007522450	0.0060143581	0.005511867	-0.005433747	0.0019858980	0.004890414	5.453400e-05	0.016958418	-0.016958418	0.004975050
SCN_N1	0.0110863258	0.014056738	-0.003985157	-0.0026718546	0.0029076219	0.0029902342	0.016020784	0.009861216	0.0166333151	-0.016382046	2.873983e-03	0.030384368	-0.030384368	-0.002158894
pctreductot	-0.0567768490	-0.186620995	-0.025110930	-0.0218159377	0.0053034623	0.0278226654	0.107282616	-0.132568931	-0.0186658817	0.144089633	4.600983e-03	-0.344049342	0.344049342	0.207847217
r_ancien_m	0.0021208512	0.293534393	0.141286657	0.1464990421	0.1024952065	-0.2344829957	-0.135746632	0.260713217	-0.0065776787	-0.267423671	-3.310518e-03	-0.695303214	0.695303214	-0.142796895

	an3_tech.an3_tech0	w_renfort.w_renfortN	w_renfort.w_renfort0	w_cible.w_cibleautr	w_cible.w_ciblepros	w_cible.w_cibleseni	w_index_N	w_index_N1	SCN_N	SCN_N1	pctreductot	r_ancien_m
r_Resil	-0.043185518	0.036502229	-0.036502229	0.122835868	-0.0660691258	-0.052758336	-0.039531280	6.782824e-04	0.033802436	0.011086326	-0.056776849	0.002120851
TCION	0.230150442	0.195484558	-0.195484558	0.116384696	-0.0437351293	-0.069347877	-0.022107160	3.164009e-02	-0.005122057	0.014056738	-0.186620995	0.293534393
FRAC.FRAC1	0.112126096	0.048156325	-0.048156325	-0.079846999	0.1532889105	-0.078949924	-0.007136608	3.107009e-03	-0.006302072	-0.003985157	-0.025110930	0.141286657
FRAC.FRAC2	0.136286139	0.045947295	-0.045947295	-0.090034099	0.0850852213	0.001046589	-0.014903890	-3.196471e-03	-0.001432276	-0.002671855	-0.021815938	0.146499042
FRAC.FRAC4	0.051414349	0.060443685	-0.060443685	-0.034609653	0.0083261301	0.025424779	-0.011712266	-4.180783e-03	-0.000752245	0.002907622	0.005303462	0.102495207
FRAC.FRAC8	-0.181083738	-0.090688586	0.090688586	0.124463850	-0.1662813717	0.048491996	0.018839584	1.249162e-03	0.006014358	0.002990234	0.027822665	-0.234482996
MTPCPTE	-0.248226538	-0.060981537	0.060981537	-0.168509746	0.0142723087	0.150747243	-0.021801947	-3.043614e-02	0.005511867	0.016020784	0.107282616	-0.135746632
RESEAU.RESEAU1	0.185471778	0.174013837	-0.174013837	0.085938533	-0.0245266231	-0.059178254	-0.019548229	2.886063e-02	-0.005433747	0.009861216	-0.132568931	0.260713217
RESEAU.RESEAU2	-0.041181881	0.029061443	-0.029061443	0.032458379	-0.0356696968	0.004749540	0.004433483	1.236591e-02	0.001985898	0.016633315	-0.018665882	-0.006577679
RESEAU.RESEAU3	-0.176757202	-0.190839222	0.190839222	-0.100947888	0.0385754533	0.059491803	0.018539328	-3.444607e-02	0.004890414	-0.016382046	0.144089633	-0.267423671
RESEAU.RESEAUAutre	-0.002026000	-0.004113127	0.004113127	-0.001162982	-0.0002633167	0.001411728	0.006357170	-3.539673e-05	0.000054534	0.002873983	0.004600983	-0.003310518
gamme.gammeML0	0.021068804	-0.570139600	0.570139600	0.006740440	0.0433582001	-0.051114540	0.023038337	-6.289219e-02	0.016958418	0.030384368	-0.344049342	-0.695303214
gamme.gammeNGM	-0.021068804	0.570139600	-0.570139600	-0.006740440	-0.0433582001	0.051114540	-0.023038337	6.289219e-02	-0.016958418	-0.030384368	0.344049342	0.695303214
an3_tech.an3_techN	-1.000000000	0.041010826	-0.041010826	0.070445805	-0.0648678016	-0.002569587	0.018852752	-1.629903e-02	0.004975050	0.002158894	0.207847217	-0.142796895
an3_tech.an3_tech0	1.000000000	-0.041010826	0.041010826	-0.070445805	0.0648678016	0.002569587	-0.018852752	1.629903e-02	-0.004975050	0.002158894	-0.207847217	0.142796895
w_renfort.w_renfortN	-0.041010826	1.000000000	-1.000000000	0.048271367	-0.0830624526	0.037868215	-0.013923615	1.650529e-02	-0.015351107	-0.035894165	0.105565011	0.414797748
w_renfort.w_renfort0	0.041010826	-1.000000000	1.000000000	-0.048271367	0.0830624526	-0.037868215	0.013923615	-1.650529e-02	0.015351107	0.035894165	-0.105565011	-0.414797748
w_cible.w_cibleautr	-0.070445805	0.048271367	-0.048271367	1.000000000	-0.5046357288	-0.463606117	0.010047813	3.367254e-02	-0.008538868	0.002661156	-0.215124597	-0.061829725
w_cible.w_ciblepros	0.064867802	-0.083062453	0.083062453	-0.504635729	1.0000000000	-0.530996014	0.023774385	2.761905e-03	0.003028452	0.009361141	0.082729015	0.040733935
w_cible.w_cibleseni	0.002569587	0.037868215	-0.037868215	-0.463606117	-0.5309960136	1.000000000	-0.034261832	-3.588468e-02	0.005272921	-0.012219342	0.126243067	0.018881450
w_index_N	-0.018852752	-0.013923615	0.013923615	0.010047813	0.0237743846	-0.034261832	1.000000000	-3.239902e-03	-0.003660334	0.017378892	-0.011249016	-0.015121293
w_index_N1	0.016299028	0.016505287	-0.016505287	0.033672540	0.0027619049	-0.035884682	-0.003239902	1.000000e+00	-0.001749211	0.011511981	-0.020069552	0.065889680
SCN_N	-0.004975050	-0.015351107	0.015351107	-0.008538868	0.0030284523	0.005272921	-0.003660334	-1.749211e-03	1.000000000	0.095448258	0.012912672	-0.015502508
SCN_N1	0.002158894	-0.035894165	0.035894165	0.002661156	0.0093611406	-0.012219342	0.017378892	1.151198e-02	0.095448258	1.000000000	0.025121398	-0.013904357
pctreductot	-0.207847217	0.105565011	-0.105565011	-0.215124597	0.0827290153	0.126243067	-0.011249016	-2.006955e-02	0.012912672	0.025121398	1.000000000	0.057185070
r_ancien_m	0.142796895	0.414797748	-0.414797748	-0.061829725	0.0407339355	0.018881450	-0.015121293	6.588968e-02	-0.015502508	-0.013904357	0.057185070	1.000000000