

Mémoire présenté le : 04 mai 2021

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires**

Par : Clara DESCOURVIERES

Titre : Apport de la Valeur Client dans la compréhension des portefeuilles d'assurance santé individuelle

Confidentialité : NON OUI (Durée : 1 an 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires

C. MEUNIER
R. CAILLET
.....
.....

Membres présents du jury de l'ISFA

D. CLOT
.....
.....
.....

Entreprise :

Nom : addactis France

Signature :

Directrice de mémoire en entreprise :

Nom : Eugénie POYET

Signature :



Invité :

Nom :

Signature :

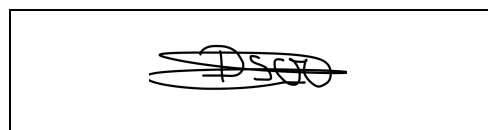
Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels

(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise



Signature du candidat





RESUME

Mots clés : Valeur Client – Assurance santé individuelle – Modèles linéaires généralisés (GLM) – Arbres CART – Forêts aléatoires – Comportement client - Segmentation portefeuille.

La relation entre l'assureur et ses assurés est en constante évolution depuis quelques années et change avec les différents contextes économique, réglementaire et technologique dans lesquels nous vivons. Dans un but de fidélisation de ses clients, la connaissance de leurs besoins, leur satisfaction et la maîtrise des offres proposées sont dorénavant des points cruciaux pour les assureurs, dans l'environnement de plus en plus compétitif qu'est le domaine de l'assurance santé individuelle.

La connaissance et l'optimisation de son portefeuille par l'assureur peuvent être atteints grâce à l'utilisation de la Valeur Client. Cet indicateur s'appuie entre autres sur la valeur passée et prospective des assurés. Un exemple de son implémentation est proposé dans cette étude, notamment en utilisant des méthodes de *datascience* afin de prédire le comportement des assurés, concernant leur résiliation et leurs consommations. L'utilité concrète de la Valeur Client est ensuite démontrée par son aide à la segmentation du portefeuille. Elle permet d'identifier les cibles de l'assureur et les actions potentielles à mener sur ces groupes clés. Enfin, les apports de la mise en place de ces actions à l'échelle du portefeuille global sont détaillés et analysés.



ABSTRACT

Key words: Lifetime Customer Value – Individual Health Insurance – Generalized Linear Models – Classification and Regression Trees – Random Forest – Customer Behavior – Portfolio segmentation.

The relation between an insurer and its customers has considerably evolved in the past few years influenced by the economical, reglementary and technological contexts. The insurer must know the needs of its clients in order to improve their loyalty, ensure their contentment and master his commercial offers, especially in the individual health insurance market, an environment becoming more and more competitive.

The insurer can have a good knowledge of its portfolio and reach its optimization through the Lifetime Customer Value metric, which relies mainly on the past and the future value of each policyholder. An example of its implementation is developed in this study, by applying *datascience* methods in order to predict the behavior of the customers, regarding their loyalty and consumptions. Special attention is placed on the Lifetime Customer Value because of its role in splitting the portfolio, which contributes to identifying the most important parts and the course of action of the insurer with those key segments. Finally, the consequences of these actions on the whole portfolio are described and analyzed.



REMERCIEMENTS

Je souhaite tout d'abord remercier ma directrice de mémoire, Eugénie POYET, pour son aide, ses conseils, son implication, sa confiance et sa patience tout au long de l'élaboration de ce mémoire. J'ai beaucoup appris sur le sujet, le métier et sur moi-même pendant cette période et ce, en grande partie grâce à toi. Alors encore merci, ce fut un plaisir de travailler avec toi.

J'adresse également mes remerciements au reste des consultants de l'équipe Pricing & Data d'addactis France pour leur accueil durant mes stages et année d'alternance, et plus particulièrement à Jean-Pascal HERMET et Cécile PARADIS pour leurs contributions et remarques précieuses lors de la finalisation de ce mémoire.

Je remercie ensuite l'ensemble de l'équipe pédagogique de l'ISFA pour les enseignements prodigués durant ces 3 années et notamment à Anne EYRAUD-LOISEL, pour ses conseils qui ont permis l'amélioration de ce mémoire.

Je tiens aussi à remercier mes (anciens) collègues et amis Aymeric, Sami, Fabien, Quentin, Mélodhy et Julie qui, par leur aide, écoute, conseils, encouragements ou divertissement, m'ont été d'une grande aide lors de la conception de ce mémoire.

Enfin, mes derniers remerciements sont adressés à mes parents Cécile et Gillou ainsi qu'à mon petit frère Martin, pour leur appui, soutien et encouragements tout au long de mes études. Je n'en serais pas là sans vous.



SOMMAIRE

RESUME	1
ABSTRACT	2
REMERCIEMENTS	3
INTRODUCTION	6
1. CONTEXTE ET OBJECTIFS DE L'ETUDE	7
1.1 Le marché de l'assurance santé : son cadre et ses enjeux	7
1.1.1 Les spécificités de l'assurance santé	7
1.1.2 Les différents réseaux de distribution	9
1.1.3 Le contexte économique actuel	10
1.1.4 Les évolutions réglementaires des dernières années	11
1.2 Définitions et intérêts de la Valeur Client	15
1.2.1 Présentation du concept de Valeur Client	15
1.2.2 L'indicateur de Valeur Client	17
1.3 Présentation de l'étude	20
1.3.1 Les objectifs de l'étude	20
1.3.2 Présentation des données utilisées	21
2. MISE EN PLACE D'UN INDICATEUR DE VALEUR CLIENT	30
2.1 Calcul et utilisation de la valeur historique	30
2.1.1 Calcul de la valeur historique	30
2.1.2 Démarche générale et hypothèses retenues	31
2.1.3 Résultats	31
2.1.4 Segmentation du portefeuille à partir du ratio S/P	35
2.2 Calcul de la Valeur Client prospective	39
2.2.1. Présentation du modèle de résiliation	39
2.2.2 Présentation du modèle de consommation	55
2.2.3 Projection des montants de primes	61
2.2.4 L'indicateur de valeur prospective	61



2.2.5	Limites de la modélisation	64
3.	APPLICATION : LA VALEUR CLIENT COMME OUTIL DE DIAGNOSTIC ET D'AIDE A LA DECISION	67
3.1	Segmentation du portefeuille grâce à la Valeur Client	67
3.1.1	Comparaison entre Valeur Client historique et marge prospective théorique	67
3.1.2	Comparaison entre probabilité de résiliation et marge prospective théorique	69
3.1.3	Comparaison entre probabilité de résiliation et valeur prospective	73
3.1.4	Conclusion et limites de la segmentation dans le cadre de l'étude	74
3.2	Impact de différents scénarios sur la Valeur Client du portefeuille	75
3.2.1	Calcul initial	75
3.2.2	Comparaison de deux scénarios	76
3.2.3	Présentation et comparaison des résultats	77
	CONCLUSION	80
	BIBLIOGRAPHIE	81
	ANNEXES	82



INTRODUCTION

Le secteur de l'assurance santé en France est un secteur dynamique qui comporte de nombreuses spécificités et qui est soumis à de nombreuses réglementations. La compétitivité du domaine est accrue et s'intensifie au fur et à mesure des avancées technologiques. Le comportement des clients se veut plus volatil car ces derniers, mieux informés sur les offres existantes, deviennent plus attentifs aux prix proposés. Ce marché est donc en constant renouvellement et le début du XXIe siècle a notamment été marqué par l'importance de la satisfaction des clients, critère important et influençant particulièrement leur fidélisation.

De nombreux outils commerciaux et marketings sont à disposition des assureurs afin d'analyser et gérer la relation avec leurs clients. Ces outils leur permettent d'optimiser leurs actions afin de maximiser à la fois la satisfaction de leurs clients et leurs propres profits. Parmi eux, l'indicateur de Valeur Client, mêlant, entre autres, valeur passée et valeur prédictive du résultat des contrats, présente un intérêt particulier puisque, bien exploité, son enjeu peut dépasser la simple recherche de rentabilité. Son implémentation et son utilisation permettent en effet d'optimiser les stratégies de gammes et le ciblage des actions, offres, services à proposer par l'assureur pour mieux attirer ses prospects, mieux fidéliser ses clients et in fine optimiser sa rentabilité. Les assurés peuvent également en tirer une plus grande satisfaction puisque les solutions qui leur sont proposées sont plus adaptées à leurs attentes et besoins, mieux identifiés par les diagnostics de portefeuille basés sur la Valeur Client.

La Valeur Client est une notion ouverte, dont la définition de calcul et la composition dépendent de l'utilisation cible souhaitée et des moyens disponibles à sa mise en place.

L'objectif de cette étude consiste en l'implémentation d'un exemple d'indicateur de Valeur Client en assurance santé individuelle puis d'en montrer l'utilité, l'intérêt et les apports directs de son utilisation.

Pour ce faire, la première partie de l'étude décrira le fonctionnement de l'assurance santé en France, le contexte économique actuel et les principales évolutions réglementaires des dernières années. Le concept de Valeur Client, ses composantes et son utilité (notamment dans le cadre de l'assurance santé) seront ensuite détaillées. Les données sur lesquelles cette étude est basée et les premiers traitements effectués seront ensuite présentés.

La deuxième partie de cette étude consistera en l'implémentation d'un exemple d'indicateur de Valeur Client. La valeur historique sera d'abord calculée puis analysée. Les méthodes utilisées pour la modélisation des différentes composantes de l'indicateur de la valeur prospective seront ensuite développées, les différents résultats obtenus seront étudiés puis discutés.

La troisième et dernière partie de ce mémoire développera les intérêts et applications possibles de l'indicateur de Valeur Client. Différentes segmentations du portefeuille seront effectuées et permettront de mettre en avant des actions à exécuter sur des segments cibles du portefeuille. Les apports de la mise en place de ces actions seront alors présentés.



1. CONTEXTE ET OBJECTIFS DE L'ETUDE

La première partie de cette étude consiste à en présenter le contexte en décrivant le marché de l'assurance santé et en proposant une définition de la notion de Valeur Client. Les objectifs de l'étude sont ensuite détaillés, suivis d'une présentation des données utilisées.

1.1 Le marché de l'assurance santé : son cadre et ses enjeux

Afin d'appréhender au mieux le contenu de cette étude, il est important d'avoir une vision globale du marché de l'assurance santé. Ainsi, ses spécificités, ses différents modes de distribution, le contexte économique actuel et les évolutions réglementaires des dernières années seront détaillés dans cette première partie.

1.1.1 Les spécificités de l'assurance santé

Le domaine de l'assurance santé présente de nombreuses spécificités qui sont détaillées dans cette partie.

a. Une assurance à deux niveaux

En France, l'assurance santé est composée de deux niveaux : d'une part, l'assurance maladie gérée par la Sécurité Sociale qui est obligatoire pour tous et d'autre part, la complémentaire santé gérée par les organismes assureurs.

L'assurance maladie

La Sécurité Sociale a été fondée en 1945 dans le but de proposer une couverture générale à la population française face aux conséquences financières de certains risques sociaux tels que les accidents du travail, la maladie, la vieillesse ou les charges familiales.

L'assurance maladie de la Sécurité Sociale est composée de plusieurs régimes qui correspondent chacun à un ensemble de droits et de devoirs réciproques des employés, des patrons et d'une caisse de la Sécurité Sociale. Un assuré est donc rattaché à un régime selon sa situation professionnelle. Il existe plusieurs régimes dont notamment le régime général qui concerne les travailleurs salariés (soit 80% de la population), le régime agricole qui couvre les exploitants et salariés agricoles, le régime des travailleurs non-salariés non agricoles, le régime Alsace-Moselle qui intervient en complément du régime général pour les personnes domiciliées en Moselle, Bas-Rhin et Haut-Rhin et d'autres régimes spéciaux.

Ce régime de base est obligatoire et permet de rembourser une partie des dépenses encourues pour les différentes prestations de santé d'un individu. Un code acte est assimilé à chaque prestation médicale et à chaque code acte correspond un taux de remboursement (TR) et une base de remboursement de la Sécurité Sociale (BRSS). Une prestation médicale n'ayant pas de code acte ne sera pas remboursée par le régime obligatoire. La différence entre la BRSS et le remboursement de la Sécurité Sociale est appelée Ticket Modérateur (TM). Il est également possible que le tarif appliqué à la prestation médicale soit supérieur au montant de la BRSS. Dans



ce cas, la différence entre le montant payé par l'assuré (autrement dit, ses Frais Réels) et la BRSS est appelée Dépassements d'honoraires. Ainsi, selon la prestation de santé et le tarif appliqué, le reste à charge pour l'assuré peut être plus ou moins important voire même total.

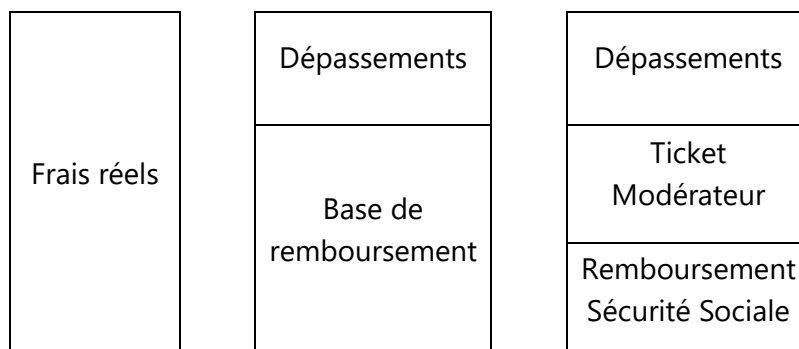


Figure 1.1 : Schéma du mécanisme de remboursement de la Sécurité Sociale

La complémentaire santé

Afin de diminuer le reste à charge de ses frais de santé, un individu peut souscrire à une assurance santé complémentaire auprès d'un organisme assureur de son choix (société d'assurance, institution de prévoyance ou mutuelle).

Il existe trois types d'assurance santé complémentaire :

- **La complémentaire santé collective** : Dans le secteur privé, une entreprise a pour obligation de proposer une assurance santé complémentaire collective à ses salariés depuis la mise en place de l'Accord National Interprofessionnel de 2013. L'adhésion est obligatoire pour l'employé (sauf cas de dispense particuliers) et l'employeur est tenu de prendre en charge une partie des cotisations. Le contrat peut également proposer des garanties optionnelles qui sont laissées au libre choix et à la charge du salarié.
- **La complémentaire santé individuelle** : La complémentaire santé peut également être souscrite à titre individuel. Ce type de contrat s'adresse aux étudiants (non couverts comme ayants droit de leurs parents), aux fonctionnaires, aux chômeurs, aux retraités, aux travailleurs indépendants mais aussi aux salariés du privé souhaitant compléter leur complémentaire santé collective par une surcomplémentaire individuelle.
- **La complémentaire santé solidaire** : Pour être éligible à cette protection complémentaire, l'assuré doit être affilié à un régime de la Sécurité Sociale et avoir des revenus inférieurs au plafond d'éligibilité (ressources annuelles pour une personne seule de 12 193€ en 2020). Les cotisations sont nulles ou très faibles (selon les revenus) et le contrat ne peut être résilié que par l'assuré et non pas par l'assureur. De plus, les garanties dont bénéficie l'assuré concerné par la CSS sont exactement les mêmes que celles des autres assurés du contrat collectif.

b. Une tarification sous contrainte

Le coût d'une prestation pour un organisme assureur dépend de l'objectif national des dépenses d'assurance maladie qui est fixé chaque année par la loi de financement de la Sécurité Sociale. Les remboursements de l'organisme assureur peuvent être exprimés en complément des



remboursements de la Sécurité Sociale (ex : 100% TM) ou en incluant ce remboursement (ex : 100% BR). Ainsi, la tarification des contrats est directement subordonnée aux modifications apportées par cette loi sur les bases de remboursements.

Il faut également tenir compte de certaines restrictions réglementaires et juridiques, tel que le fait que la loi interdit de différencier le tarif selon le sexe en assurance. De plus, la loi Evin de 1989 prohibe de tenir compte du passé de l'assuré pour le discriminer ou le favoriser dans la tarification de sa complémentaire santé.

Enfin, la grande diversité des risques couverts par l'assurance santé complexifie la tarification des contrats. En effet, certains actes, tels que les dépenses en pharmacie, présentent une fréquence de consommation élevée avec des coûts très dispersés. Les consultations sont également fréquentes mais présentent un coût faible et peu variable. A contrario, certains postes, comme les actes d'hospitalisation, ont une fréquence de consommation faible mais leur coût peut être très important. Les assureurs doivent donc tenir compte de chacune de ces spécificités dans les modèles de tarification de leurs contrats.

c. Une rentabilité incertaine

En assurance santé comme pour tout autre secteur d'assurance, les assureurs doivent faire face à l'inversion du cycle de production lors de la tarification. En effet, les tarifs des cotisations que doivent payer les assurés sont calculés en amont du versement des prestations. La rentabilité est donc incertaine pour l'organisme assureur et l'équilibre financier qui découle du contrat ne sera connu qu'à la fin de l'exercice de celui-ci.

Par ailleurs, la compétitivité s'est accrue ces dernières années sur ce marché, avec des marges techniques de plus en plus tirées et les récentes évolutions réglementaires, de même que le contexte sanitaire actuel, ajoutent encore de l'incertitude sur la rentabilité des gammes santé.

1.1.2 Les différents réseaux de distribution

Afin de maximiser la rentabilité de souscription d'un contrat, l'objectif pour l'assureur est de maîtriser et limiter les coûts engendrés par l'acquisition puis la gestion de ses clients. L'individu peut souscrire un contrat par différents réseaux de distribution, chacun ayant ses spécificités et donc des coûts différents. Parmi ces réseaux, il existe notamment :

- **L'agent général** : professionnel indépendant faisant l'intermédiaire entre l'assuré et l'entreprise d'assurance avec laquelle il a signé un mandat exclusif. L'agent reçoit un pourcentage de la prime pour chaque contrat vendu.
- **Le courtier** : possède le statut de commerçant et est mandataire de l'assuré. Il peut travailler pour plusieurs entreprises d'assurances car il est chargé de trouver le contrat le plus adapté pour son client. Il touche une commission pour chaque contrat signé et est le seul interlocuteur de l'assureur (l'assuré n'a pas de contact direct avec son assureur).
- **La vente directe** : concerne toutes les ventes qui s'effectuent sans contact physique. La vente par internet est en pleine croissance ces dernières années et permet aux clients d'effectuer un grand nombre d'opérations eux-mêmes et à distance, sans avoir recours à un conseiller.



Ce canal représente un moindre coût pour l'assureur puisque le client effectue une grande partie des recherches lui-même mais l'enjeu est grand puisqu'il augmente la compétitivité du fait de l'accessibilité à toutes les informations et une comparaison entre assureurs facilitée.

1.1.3 Le contexte économique actuel

Le marché de l'assurance est un marché dynamique, qui présente une évolution croissante d'année en année. Son modèle est en train de changer, au rythme du développement technologique, ce qui entraîne une réelle innovation du métier et des offres présentées. Cependant, l'année 2020, marquée par la crise de la COVID-19, a freiné voire stoppé cette croissance, d'autant plus que son impact risque de se faire ressentir sur les années à venir.

a. Un marché en pleine transformation

L'innovation numérique des dernières années a eu un impact important sur le marché de l'assurance notamment avec les nouveaux modes de distribution et communication qui ont été mis en place, l'apparition de nouveaux risques, la multiplication des services proposés et l'évolution de la relation entre les assureurs et leurs clients.

En effet, l'expansion numérique a amené les assureurs à se moderniser et à développer davantage leur communication digitale afin d'être en contact permanent avec leurs clients. La digitalisation présente pour l'assuré une simplification des démarches administratives, notamment avec la télétransmission entre l'assurance maladie de la Sécurité Sociale et sa complémentaire santé. Pour l'assureur, cela lui permet d'être plus proche de son client et de proposer des offres plus adaptées après étude des données collectées grâce à la data-science. Cependant, cela engendre de nouveaux risques à identifier et gérer tels que les cyber-risques.

Le secteur de l'assurance santé présente également de nouveaux enjeux liés à l'évolution des modes de vie qui doivent être pris en compte par les assureurs comme la prévention, les téléconsultations, l'aide au maintien à domicile ou l'assistance aux malades. Le numérique modifie le métier de l'assureur en l'encourageant à développer davantage son service d'accompagnement des assurés.

Enfin, le comportement des consommateurs est devenu plus volatile. Toute l'information se trouvant à portée de main sur internet, notamment grâce aux sites de comparateurs d'assurance, les clients sont devenus plus exigeants car mieux informés et plus attentifs aux prix proposés. La concurrence dans le secteur de l'assurance est donc en croissance constante.

b. Une année 2020 sous tension

Pour l'année 2020, le secteur de l'assurance avait prévu une croissance continue mais plus faible car marquée par de nombreux changements réglementaires et une baisse des taux pesant sur la rentabilité. Cependant, la crise de la Covid-19 est venue perturber l'évolution du marché. [1]

Le secteur de l'assurance a été lourdement critiqué durant la période du premier confinement au printemps 2020, par l'absence de prise en charge de la perte d'exploitation des commerces et



entreprises pénalisés par la crise. Pour pallier cela, nombre d'assureurs ont offert des solutions innovantes et variées pour leurs clients. Certains ont proposé des gestes commerciaux à leurs assurés (extensions de garanties, réductions tarifaires, allongement des délais de paiements), ont versé des indemnités aux secteurs les plus touchés ou ont suggéré de nouvelles offres et services sur mesure. Même si le secteur de l'assurance en ressort affaibli, les acteurs ont su montrer à nouveau leur capacité d'adaptation et d'innovation.

Cette crise a également accéléré la prise en charge par l'assurance santé des nouveaux enjeux mentionnés précédemment. En effet, la prévention a été mise en valeur, notamment par l'accompagnement psychologique lié à la pandémie, aux différents confinements et leurs conséquences. De plus, l'incitation à la distanciation sociale a entraîné une hausse importante du nombre de téléconsultations alors que cette méthode présentait un avis mitigé auparavant.

Enfin, il a été demandé aux organismes assureurs de prendre part à l'effort consenti par l'assurance maladie afin de soutenir le système de santé pendant cette crise sanitaire, avec une contribution exceptionnelle de 2 milliards d'euros sur deux ans. En effet, le gouvernement a estimé que les recettes des complémentaires santé étaient inchangées alors que leurs charges avaient largement diminué. Cependant, il est difficile de qualifier cette différence de « bénéfiques » puisqu'un report des soins a été observé avec une hausse de 30% en soins dentaires et 20% en optique à la suite du confinement du printemps 2020. De plus, la faillite de certaines entreprises à cause de la crise peut amener à une baisse des revenus des cotisations dans les mois à venir. L'évolution économique du secteur pour les années à venir est donc plus qu'incertaine.

1.1.4 Les évolutions réglementaires des dernières années

Le secteur de l'assurance santé a subi de nombreuses réformes ces dernières années. La première loi importante est celle de Douste-Blazy qui instaure en 2005 une participation forfaitaire de 1€ sur certains actes médicaux ainsi que la création du « parcours de soin »¹. Par la suite, plusieurs autres réglementations ont été mises en place dans le but, notamment, de réduire le déficit de l'assurance maladie.

Ces réformes représentent un enjeu non négligeable pour les organismes d'assurance puisqu'elles impliquent la tarification d'un risque futur à l'aide de données passées. En effet, les comportements des assurés sont amenés à être modifiés à la suite de ces changements de législation et les assureurs doivent en tenir compte afin de limiter leurs pertes.

¹ Le « parcours de soin » est un dispositif consistant à ce que l'assuré choisisse un médecin traitant qui le guidera lors de ses consultations de médecins spécialistes en échange d'un meilleur taux de remboursement de la Sécurité Sociale.



a. Mise en place du contrat responsable

A compter du 1^{er} janvier 2006, les assureurs doivent proposer des contrats d'assurance complémentaire santé collectifs et individuels respectant un certain cahier des charges initialement fixé par décret le 30 septembre 2005.

Un contrat est dit « responsable » s'il respecte les règles définies dans le code de la Sécurité Sociale. Le contrat responsable doit a minima rembourser le ticket modérateur des consultations, de certains médicaments et du forfait journalier hospitalier, respecter des plafonds et planchers de remboursement en optique et ne pas prendre en charge les dépassements d'honoraires des consultations hors parcours de soins. De plus, un contrat responsable ne peut pas prendre en charge les participations forfaitaires imposées par la loi Douste-Blazy. Enfin, il doit couvrir au moins deux actes de prévention par an tels que la réalisation de vaccins ou de dépistages.

Les règles imposées pour les contrats responsables ne présentent pas de caractère obligatoire. Cependant, depuis le 1^{er} avril 2015, une réforme incite les assureurs et employeurs à respecter ce cahier des charges en imposant une taxe de 14% aux contrats non-responsables contre 7% pour les contrats responsables, en appliquant des charges sociales sur la part patronale de la cotisation à un contrat non-responsable et en ne déduisant pas la cotisation du salarié à un contrat non-responsable de son revenu imposable.

b. Accord National Interprofessionnel (ANI)

L'Accord National Interprofessionnel de 2013 a permis la généralisation de la mise en place d'une complémentaire santé pour tous les salariés du secteur privé, à compter du 1^{er} janvier 2016.

Les employeurs du secteur privé ont désormais l'obligation de proposer une complémentaire santé collective à leurs employés et de prendre en charge au moins 50% de son financement. Les salariés ont obligation d'y adhérer, à l'exception des personnes bénéficiant de la CSS, des ayants droit d'un bénéficiaire d'une autre couverture complémentaire ou des personnes en contrat à durée déterminée de moins d'un an. La couverture des ayants droit n'est pas obligatoire mais l'employeur peut décider de les couvrir également. De plus, rien n'empêche l'employé de souscrire à une surcomplémentaire de manière individuelle afin d'être mieux protégé.

La complémentaire santé proposée par l'employeur doit respecter le panier de soins minimum défini par l'ANI. Il consiste en une prise en charge de l'intégralité du ticket modérateur pour les consultations, actes techniques et pharmacie, un remboursement d'au moins 125% BRSS pour les prothèses dentaires, la prise en charge du forfait journalier hospitalier et un remboursement forfaitaire d'au moins 100€ en optique par an. Ces garanties minimales à respecter peuvent être améliorées par accords de branche et conventions collectives. Les garanties proposées sont généralement plus avantageuses pour l'employé que par une assurance santé individuelle puisque le contrat est négocié collectivement.

Bien que seuls les actifs du secteur privé soient concernés par l'ANI, les assureurs ont vu une modification de la répartition du risque assurantiel entre le collectif et l'individuel. Un grand



nombre de salariés a résilié son assurance santé individuelle ce qui a entraîné une diminution du nombre d'assurés âgés de moins de 60 ans. Cependant, beaucoup d'employeurs ont opté pour des contrats collectifs ne répondant qu'aux exigences minimales imposées, ce qui a amené à une hausse du nombre de souscriptions à une surcomplémentaire.

c. La réforme 100% santé

Le cahier des charges des contrats responsables a été modifié afin de proposer des paniers de soins sans reste à charge pour l'optique, le dentaire et l'audiologie. Sa mise en application a été effectuée de manière progressive entre le 1^{er} janvier 2019 et le 1^{er} janvier 2021. L'objectif de cette réforme est de prévenir le renoncement aux soins pour raisons financières.

Trois postes de santé sont concernés : l'optique, le dentaire et l'audiologie. Pour chaque poste, des paniers de soins « 100% santé » permettant un reste à charge nul pour l'assuré ont été définis. Pour ce faire, les bases de remboursement des actes concernés ont été élargies, les professionnels de santé ont accepté des prix limites de vente et pour être responsables, les contrats d'assurance santé doivent prendre en charge l'intégralité du reste à charge des assurés sur ces équipements.

En optique, le panier de soins « 100% santé » doit contenir plusieurs modèles de montures et les verres doivent comporter certains traitements obligatoires. La prise en charge intégrale des équipements de ce panier est effective depuis le 1^{er} janvier 2020.

Sur le poste dentaire, les équipements sont découpés en trois paniers différents : un panier « 100% santé » sans aucun reste à charge, un panier aux honoraires maîtrisés qui peut engendrer un reste à charge mais limité par le plafond des honoraires de certains actes et un panier à honoraires libres pour lequel le reste à charge varie seulement en fonction du niveau de garantie de la complémentaire santé. Le plafonnement des tarifs des prothèses et l'évolution des bases de remboursement ont été effectués de manière progressive entre 2019 et 2021, selon le type de prothèses.

En audiologie, le panier de soins « 100% santé » contient plusieurs types d'équipements de qualités variées en termes esthétiques et techniques. La base de remboursement de la Sécurité Sociale pour ces équipements a progressivement augmenté depuis le 1^{er} janvier 2019 pour passer de 200€ à 400€. Les tarifs des professionnels de santé sont également limités afin de ne pas dépasser 950€ par oreille en 2021.

Cette réforme présente un risque important pour les assureurs et ses impacts sont complexes à modéliser. Si les coûts des différents dispositifs peuvent être déterminés (ou extrapolés), il est difficile d'anticiper les déformations futures des fréquences de consommation. Le moindre renoncement aux soins espéré, le déport des adhésions sur des contrats « entrée de gamme » devenus suffisamment couvrants ou les pratiques des professionnels de santé sont autant d'éléments incertains qui risquent de déformer les comportements de consommation des assurés. Ce point pourrait entraîner des coûts importants pour les assureurs.



d. La résiliation infra-annuelle

Faisant suite à la loi Hamon de 2014 visant à améliorer la protection des consommateurs et à rééquilibrer les pouvoirs entre clients et professionnels, à partir du 1^{er} décembre 2020, les contrats d'assurance complémentaire santé à tacite reconduction peuvent être résiliés à tout moment et de façon simplifiée.

Ce droit de résiliation s'applique aux contrats individuels souscrits dans n'importe lequel des trois types d'organisme assureur. La loi exclut cependant les contrats collectifs à adhésion obligatoire puisque seuls les employeurs peuvent bénéficier de la résiliation infra-annuelle et non pas leurs employés adhérents. Il faut de plus attendre un délai minimum de souscription d'un an avant de bénéficier de ce droit de résiliation.

La résiliation des contrats est également simplifiée puisque l'assuré est dorénavant en mesure de la demander via le même réseau de distribution que celui utilisé pour la souscription. Ainsi, un assuré ayant souscrit, par exemple, une assurance en ligne peut donc la résilier à tout moment sur internet.

Le comportement des assurés sera probablement modifié en assurance individuelle puisque l'assuré aura une volonté de rentabiliser sa cotisation, ce qui peut entraîner un risque d'antisélection et de surconsommation. La facilité de résiliation va également accroître la concurrence des organismes d'assurance santé qui seront poussés à proposer de nouvelles offres aux tarifs attractifs, en particulier aux personnes considérées comme « bons risques ». La rentabilité des produits ainsi que la mutualisation des risques sont donc menacées par cette réforme.

1.1 Synthèse

Le marché de l'assurance santé présente de nombreuses spécificités qui influencent la tarification et le paiement des prestations. La fréquence de consommation et le coût des prestations dépendent de ces particularités mais également de caractéristiques propres à l'assuré telles que son âge, son sexe ou sa catégorie socioprofessionnelle. La rentabilité des contrats découle de cette tarification mais également du canal de distribution par lequel il a été souscrit puisque les frais engendrés à l'acquisition diffèrent d'un canal à l'autre.

Le secteur est innovant et dynamique. Il a su s'adapter aux changements de la société et à la révolution technologique des dernières années. Cependant, la compétitivité du secteur de plus en plus importante, l'environnement réglementaire strict et la crise de la COVID-19, risquent de freiner la croissance de ce dernier.

Ainsi, de nouveaux enjeux sont apparus pour les organismes assureurs. La maîtrise et la bonne connaissance de leurs portefeuilles clients permettent d'augmenter leur rentabilité, de mieux comprendre les besoins de leurs assurés et donc d'y répondre au mieux. Il est également de leur devoir de tenir compte de la prise en charge de nouveaux risques émergents dus aux changements des modes de vie actuels.



Enfin, de nombreux changements de réglementations sont venus modifier le secteur de l'assurance complémentaire notamment en incitant les assureurs à proposer des garanties minimales, en rendant obligatoire l'adhésion à un contrat collectif pour les salariés ou en modifiant la structure des remboursements de la Sécurité Sociale. De plus, la résiliation du contrat de mutuelle a été récemment simplifiée et est dorénavant possible à tout moment, ce qui engendra de nouveaux défis de tarification et de compétitivité entre les différents acteurs du marché.

1.2 Définitions et intérêts de la Valeur Client

Objet de cette étude, la notion de Valeur Client est définie dans cette partie. Par la suite, son évolution et son intérêt pour le domaine de l'assurance sont présentés. [2] [3]

1.2.1 Présentation du concept de Valeur Client

a. Origine de la Valeur Client

Pour les secteurs de la vente de biens et services, l'importance de développer et d'entretenir une relation avec ses clients a largement évolué au cours de ces dernières années.

Les entreprises ont commencé à tenir compte de la satisfaction des clients en améliorant la qualité de leurs produits après s'être longtemps concentrées uniquement sur un objectif de vente de masse. L'évolution des modes de vie à la fin du XXe siècle a entraîné un développement de l'offre qui a accru la compétitivité des entreprises et a fait naître un objectif et une volonté de fidélisation des clients.

De ce fait, la stratégie de *Customer Relationship Management (CRM)*, connaît un grand succès dans les années 2000. Elle consiste à améliorer la rentabilité de l'entreprise en développant une relation durable avec les clients qui présentent un potentiel non négligeable. Le développement des nouvelles technologies a, comme expliqué précédemment, modifié le comportement des clients ce qui implique que leur satisfaction et la réponse à leurs besoins doivent dorénavant se trouver au cœur de toute stratégie marketing. L'évolution des outils technologiques depuis le début du XXIe siècle rend cela plus accessible et plus efficace. En effet, les méthodes de *data science* aident à l'étude des comportements et à la personnalisation des besoins, en mettant en lumière les préférences des clients plutôt que leurs simples intentions.

Dans le cadre de la stratégie de *CRM*, de nombreux indicateurs marketing dont la Valeur Client (ou *Customer Lifetime Value*) ont été développés afin de répondre à d'autres enjeux mis en avant par la volatilité des clients et la compétitivité croissante de certains secteurs, tout en utilisant au mieux les avantages du *big data*.

b. Définition de la Valeur Client



L'indicateur de Valeur Client n'a pas vraiment de définition de calcul fermée et précise : celle-ci varie selon l'usage qui doit en être fait et des données disponibles pour son implémentation.

La Valeur Client peut être basée sur la seule notion de rentabilité et être définie comme « la valeur de l'ensemble des achats d'un type de produit qu'un client fait en moyenne pendant toute sa vie ». [5] Elle apporte ainsi des informations sur le rôle joué par chaque client dans la création de valeur pour l'entreprise.

Une autre définition est donnée par « la contribution financière totale d'un client sur sa durée de vie au sein de la compagnie en s'appuyant sur la satisfaction à long terme du client plutôt que sur une optimisation des ventes à court terme ». [6] La notion de durée de la relation est ainsi prise en compte dans cette définition de l'indicateur.

La Valeur Client peut, utilisée dans un cadre d'analyse marketing, apporter des informations sur le bénéfice ou non de différentes actions menées par l'entreprise.

Globalement, la création d'un indicateur de Valeur Client vise donc à calculer la rentabilité d'un client, basée sur le profit dégagé depuis le début de sa relation avec l'entreprise et est assorti d'une mesure prédictive de son attitude future concernant la durée de leur relation, son coût et son apport pour l'entreprise.

c. Les apports possibles de la Valeur Client dans le secteur de l'assurance

Après s'être implanté dans le secteur des télécoms et de la grande distribution, l'indicateur de Valeur Client est désormais utilisé dans le secteur de l'assurance. [4]

En effet, dans un contexte de forte concurrence où les clients assurés sont sur-sollicités et ont un meilleur accès à l'information pour comparer les offres entre elles, disposer d'une vision claire et objective des besoins, attentes et comportements de ses clients est un véritable atout pour l'assureur dans la définition de sa stratégie.

Par ailleurs, identifier ses cibles rentables et/ou fidèles, trouver les facteurs impactant chacun de ces deux axes et anticiper les déformations futures de la rentabilité sont au cœur de l'activité des actuaires exerçant dans le secteur de l'assurance.

Que ce soit pour la tarification ou la modélisation de la rentabilité des produits d'assurance, les analyses sont toujours effectuées par une projection future des flux entrants et sortants. Le plus souvent, celles-ci sont effectuées de façon segmentée (par gamme ou selon les critères de segmentation tarifaire).

Dans ce contexte, la vision prospective et individualisée de la rentabilité d'un client, définition même de la Valeur Client, permet de façon évidente de nourrir ses intérêts.

Par une segmentation de son portefeuille selon cet axe, l'assureur en renforcera sa maîtrise puisqu'il pourra mener des actions adéquates afin d'en développer les portions en croissance tout en optimisant sa rentabilité globale.



Le client peut également en tirer un bénéfice puisque l'assureur a dorénavant la capacité d'ajuster les moyens et la stratégie qu'il lui consacre. Il peut ainsi mener la « meilleure action future » en lui présentant directement une nouvelle offre, un ajustement de celle en cours ou un geste commercial lui correspondant au mieux. L'objectif est d'augmenter également la satisfaction des clients afin de jouer un rôle dans leur fidélisation, qui représente un véritable enjeu dans un environnement compétitif et le sera d'autant plus avec la récente mise en place de la résiliation infra-annuelle présentée précédemment.

De plus, selon les critères sélectionnés pour la modélisation de l'indicateur, des défaillances et des points d'amélioration de la compagnie peuvent être mis en avant. Par exemple, un fort taux de résiliation sur un segment de portefeuille passant par un canal de distribution précis peut même être l'indication d'un problème et des améliorations à effectuer sur ce dernier.

1.2.2 L'indicateur de Valeur Client

a. Un indicateur ouvert

Comme expliqué précédemment, la notion de Valeur Client n'a pas de définition fermée. La façon de calculer cet indicateur et de l'analyser est donc à adapter en fonction du cadre d'usage souhaité et des données mises à disposition.

On peut néanmoins s'accorder à dire que le calcul de la Valeur Client doit être segmenté en fonction des caractéristiques de chaque assuré et de leur comportement spécifique. Cependant, dans le cadre de l'assurance complémentaire santé, son utilisation devra être adaptée à la réglementation car l'assureur n'a pas le droit de discriminer les assurés en fonction de leur état de santé. En particulier, la tarification proposée in fine ne pourra pas être individualisée, en étant indexée sur la Valeur Client calculée. L'article L112-1 du Code de la mutualité interdit en effet de prendre en considération le passé de l'assuré lors de la couverture de ses dépenses santé. Or, si un assuré présente une Valeur Client faible, cela peut s'expliquer par un coût de prestations important ou une fréquence de consommation non négligeable et donc probablement par un état de santé dégradé.

Les applications effectuées pourront cependant, comme pour la tarification, être définies sur des segments de portefeuille homogènes, construits et définis sans qu'il n'y ait de discrimination. De même qu'un tarif unisexe peut être construit en combinant des modélisations des prestations différenciées entre les Hommes et les Femmes, les aspects calculatoires et applicatifs de l'indicateur de Valeur Client peuvent être définis sur des axes non nécessairement identiques.

L'implémentation de l'indicateur de Valeur Client doit être effectué dans le respect du Règlement général sur la protection des données⁴. Ce dernier implique notamment que les

⁴ Règlement encadrant le traitement des données personnelles au sein de l'Union Européenne, entré en application en mai 2018.



données utilisées pour la mise en place de l'indicateur sont tenues d'être anonymisées et que la sécurité doit être maximisée lors de leur manipulation et stockage.

La Valeur Client est donc un indicateur ouvert dont les différentes composantes et hypothèses de calcul doivent tenir compte de la problématique visée et des données disponibles, et dont l'exploitation devra être réalisée dans le respect du cadre réglementaire.

b. Valeur passée et valeur prospective

La Valeur Client peut correspondre à une valeur passée, calculée sur les données disponibles du portefeuille et une valeur prospective, modélisée selon les données historiques. [7]

La valeur passée peut être calculée de différentes manières. Elle peut correspondre à une simple valeur comptable du client, représentée par la différence entre les primes perçues et les prestations payées. Dans ce cas, il faut tenir compte du temps passé par l'assuré au sein du portefeuille. Cependant, cette durée diffère pour chaque individu ce qui rend la comparaison des valeurs historiques impossible. Pour remédier à ce problème, une reconstitution des données manquantes peut être effectuée pour les assurés les plus récents. Il faut alors tenir compte de la fiabilité des données historiques comparativement aux données estimées afin de ne pas pénaliser ni les nouveaux assurés, ni les plus anciens. Il est également possible de pondérer les montants de primes perçues et prestations payées par la durée d'exposition au risque de chaque assuré. Cela permet de baser le calcul sur des données historiques uniquement tout en tenant compte de la durée de présence de chaque assuré dans le portefeuille. Enfin, si ces données sont mises à disposition, la valeur passée doit prendre en compte les frais d'acquisition et de gestion du client.

Plusieurs paramètres peuvent être considérés dans le calcul de **la valeur prospective** mais le principe reste cependant le même. Il consiste à utiliser les données historiques afin de prédire les comportements futurs, et ce, pour chaque paramètre pris en compte dans l'implémentation de l'indicateur. Cela peut être la probabilité de résiliation du contrat, la probabilité de multi-équipement (acquisition d'un nouveau contrat), la probabilité de montée en gamme (acquisition de meilleures garanties), la probabilité de sinistralité... Cette méthode peut être utilisée car les relations sont de nature contractuelle (donc durable) et le marché de l'assurance présente suffisamment de stabilité pour que le comportement futur puisse être à l'image du comportement passé. La durée sur laquelle est calculée la valeur future doit être fixée en amont et ne pas dépasser un certain nombre d'années afin que son calcul soit le plus pertinent possible. En effet, dans un marché en constante évolution où les offres proposées varient en fonction des caractéristiques de l'assuré qui changent tout au long de sa vie (ajout ou perte d'ayants droit, changement de régime social, de situation de famille...), le calcul de la Valeur Client prospective à l'ultime présente un intérêt plus théorique que réellement applicatif.

c. Autres composantes de l'indicateur

D'autres composantes de l'indicateur de Valeur Client existent et peuvent venir compléter la Valeur Client passée et prospective. Cependant, selon le contexte de l'étude et le marché concerné, leur implémentation présente un intérêt plus ou moins important.



L'*embedded value* correspond à la valeur actualisée des contrats détenus par un assuré à un instant fixé. Elle estime la contribution aux résultats futurs que peut avoir un contrat. Ainsi, plus l'assuré vieillit, plus l'*embedded value* diminue puisque l'assureur a de moins en moins de flux à recevoir de ce dernier. Le calcul de cette valeur dépend donc notamment de l'âge de l'assuré, de son ancienneté, de sa probabilité de résiliation etc. Cette composante est principalement utilisée en assurance vie.

La valeur maximale d'un client représente le nombre maximum d'achats qu'il peut effectuer ou, pour le marché de l'assurance, le nombre maximum de contrats auxquels il peut souscrire. Elle est calculée pour un assuré en se basant sur des données externes et marketing telles que des statistiques publiques ou des enquêtes de satisfaction client. Elle permet de ne pas surestimer le potentiel futur du client. Elle est liée à **la valeur compétitive** de la firme qui correspond à la capacité de l'entreprise à exploiter la valeur maximale du client et à préserver sa valeur future. Elle tient également compte de la part de client pouvant être reprise à la concurrence. Elle nécessite cependant une quantité d'information importante afin de répondre au mieux aux attentes du client et de pouvoir le persuader de modifier son comportement actuel. Elle doit également tenir compte des évolutions réglementaires et de l'arrivée potentielle de nouveaux acteurs.

Une fois les composantes de l'indicateur choisies, il convient de décider comment les lier entre elles afin de fournir l'objectif de l'étude. Un score total peut être calculé par exemple en sommant des flux actualisés ou en effectuant une moyenne entre valeur passée et valeur future. Une simple comparaison des différentes valeurs peut également être effectuée afin de bien cerner l'évolution et donc le potentiel de valeur des assurés.

1.2 Synthèse

La place de la relation client dans la vente de biens et services a évolué au cours du XXe siècle. La satisfaction client est désormais au cœur des préoccupations des entreprises, notamment dans les environnements compétitifs, où la fidélisation des clients est devenue un enjeu de taille.

L'évolution technologique du début du XXIe siècle a permis de faciliter l'étude du comportement des clients et de développer différents indicateurs de performance. Parmi eux, l'indicateur de Valeur Client est particulièrement utilisé dans le monde des assurances. Sa définition et ses composantes dépendent de l'utilisation qui en est faite, ainsi que des données disponibles lors de son implémentation.

Globalement, la Valeur Client sert à calculer la rentabilité d'un client sur la durée totale de sa relation avec l'entreprise, en s'appuyant sur les données historiques afin de modéliser le comportement de l'assuré et donc sa valeur prospective. Cet indicateur, une fois implémenté, permet à l'assureur de segmenter son portefeuille et ainsi détecter les groupes de clients à potentiel de valeur, repérer d'éventuelles défaillances de la compagnie et mettre en œuvre une stratégie plus proche des attentes du client afin de répondre au mieux à ses besoins.

Dans le cadre de l'assurance complémentaire santé, l'assureur doit faire attention à ne pas proposer une tarification individuelle basée sur la Valeur Client. En effet, l'usage de cet indicateur



doit se faire dans le respect des réglementations et la loi interdit notamment de pénaliser le tarif d'un contrat par l'état de santé de l'assuré.

L'indicateur de Valeur Client peut être défini par plusieurs valeurs telles que la valeur passée (calculée à partir des données historiques), la valeur prospective (calculée à partir de la modélisation des résiliations, prestations et primes à venir), l'*embedded value* (estimation de la contribution aux résultats futurs), la valeur maximale (le nombre maximum de contrats que peut avoir un client) ou la valeur compétitive (valeur du client pouvant être récupérée à la concurrence). La définition exacte de l'indicateur est choisie selon les données disponibles, l'étude réalisée et l'objectif souhaité.

1.3 Présentation de l'étude

Le cadre de l'étude ayant été présenté dans les deux précédentes parties, une description des objectifs de l'étude et des données utilisées à sa réalisation est effectuée dans la partie suivante.

1.3.1 Les objectifs de l'étude

L'objectif de cette étude consiste en l'implémentation d'un exemple d'indicateur de Valeur Client puis de son utilisation afin de réaliser une cartographie stratégique d'un portefeuille d'assurance santé individuelle.

En effet, cet indicateur est utilisé en assurance IARD depuis de nombreuses années et des études similaires ont déjà été menées pour ce secteur ([8] [9]) mais son implémentation reste à l'heure actuelle relativement peu courante en assurance de personnes.

La valeur historique de chaque assuré, calculée comme la différence entre les primes perçues par l'assureur et les remboursements des frais de santé, permettra d'avoir une photographie réelle du portefeuille et de sa composition sur les deux années constitutives de la base de données. Ces données serviront également à modéliser la valeur prospective des assurés qui, dans cette étude, considère uniquement les frais de santé remboursés et la probabilité de résiliation des assurés, ainsi que les primes touchées par l'assureur dans un futur proche.

A la différence d'autres études menées sur la Valeur Client en assurance santé ([10]), les intérêts de cet indicateur autres que la mesure de rentabilité du portefeuille seront présentés. Il sera en effet utilisé afin de proposer différentes segmentations du portefeuille et, selon l'objectif recherché par l'assureur, il lui sera donc possible d'en obtenir un diagnostic précis, d'identifier de nouvelles cibles de conquêtes et d'adapter sa stratégie future. Les bénéfices engendrés sur le portefeuille global par la mise en place des différents scénarios identifiés grâce à ces segmentations seront ensuite présentés.

Globalement, l'étude proposera donc un exemple d'implémentation de cet indicateur, puis démontrera son intérêt par une application directe sur le portefeuille d'assurance santé utilisé dans cette étude.



1.3.2 Présentation des données utilisées

a. Les données initiales

L'étude est initialement basée sur 4 bases de données d'assurance santé individuelle, qui contiennent des informations sur les prestations, les assurés, les contrats et les garanties correspondantes pour les années 2017 et 2018 :

- **La base de prestations** comporte les informations relatives à chaque opération effectuée telles que la date de survenance, l'acte concerné, le reste à charge de l'assuré, la part de l'assurance complémentaire ou encore la part prise en charge par la Sécurité Sociale.
- **La base des assurés** contient les informations tête par tête des bénéficiaires telles que leur sexe, date de naissance, état civil et les changements concernant leur régime social ou leur situation familiale.
- **La base des contrats** comporte les informations de différents contrats dont leur période d'effet, s'ils ont été résiliés et le motif de résiliation si elle a eu lieu.
- **La base des garanties** contient la gamme, le niveau de garantie et les dates de validité de chaque contrat.

b. Hypothèses retenues et traitements effectués

Dans un objectif de modélisation de la probabilité de résiliation d'un contrat et des consommations de santé des assurés, les bases de données ont été nettoyées, retraitées puis agrégées entre elles afin de comporter les informations nécessaires à la réalisation de cette étude.

Certaines hypothèses ont été prises en amont du traitement des données :

- Si un changement d'information a été enregistré concernant un assuré, un contrat ou une garantie, seule la plus récente a été conservée.
- Seules les informations comprises dans la période d'observation (du 1^{er} janvier 2017 au 31 décembre 2018) ont été conservées. Cette hypothèse a notamment entraîné la suppression de tout contrat ayant été résilié avant le début de cette période ou souscrit après la fin de cette période.
- Seules les observations qui ont trouvé correspondance avec chacune des bases ont été conservées.

Au final, cette étude s'appuie donc sur les informations d'environ 100 000 contrats pour un peu plus de 150 000 bénéficiaires et un total de 4 500 000 lignes de prestations par an pour 2017 et 2018.

c. Statistiques descriptives du portefeuille

Afin d'appréhender au mieux le portefeuille, cette partie présente une étude des effectifs, taux de résiliation et consommation moyenne selon différents critères.



Au global, l'âge moyen du portefeuille est de 44,3 ans et il est composé à 47% d'hommes et 53% de femmes.

De plus, 79% des contrats ont une situation de famille « Isolé » et seulement 14% couvrent des enfants (en moyenne 1,8 enfants par contrat).

Analyse par âge et par sexe

En santé, la consommation moyenne des assurés dépend généralement de l'âge : elle est conséquente à la naissance mais décroît rapidement puis augmente à nouveau avec l'âge. Ainsi, une proportion importante de très jeunes enfants et/ou personnes âgées au sein du portefeuille peut provoquer des coûts non négligeables pour l'assureur.

Au sein du portefeuille, les effectifs sont répartis de la façon suivante parmi les classes d'âges :

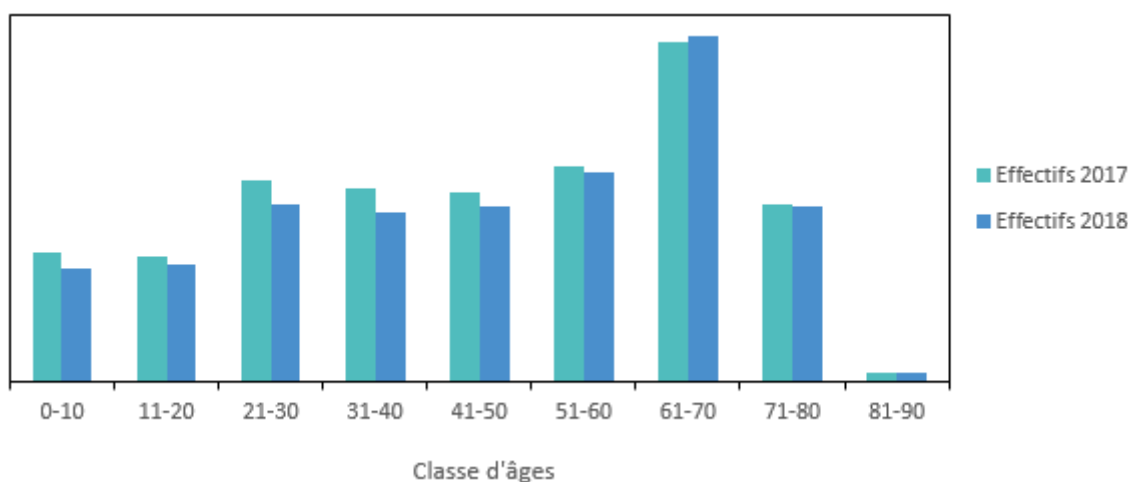


Figure 1.3.a : La classe des 61-70 ans est la plus représentée au sein du portefeuille.

L'histogramme montre donc que la classe des 61-70 ans est principalement représentée au sein du portefeuille. Le portefeuille comporte également un certain nombre de contrats couvrants plusieurs ayants droit puisque le nombre de bénéficiaires âgés de 0 à 20 ans est également considérable. Enfin, l'effectif de chaque classe d'âges décroît entre 2017 et 2018 à l'exception des plus de 60 ans. Le portefeuille a donc vieilli entre les deux années.

L'étude de la consommation moyenne par âge et par sexe du portefeuille montre que ce dernier suit les tendances attendues, décrites précédemment.

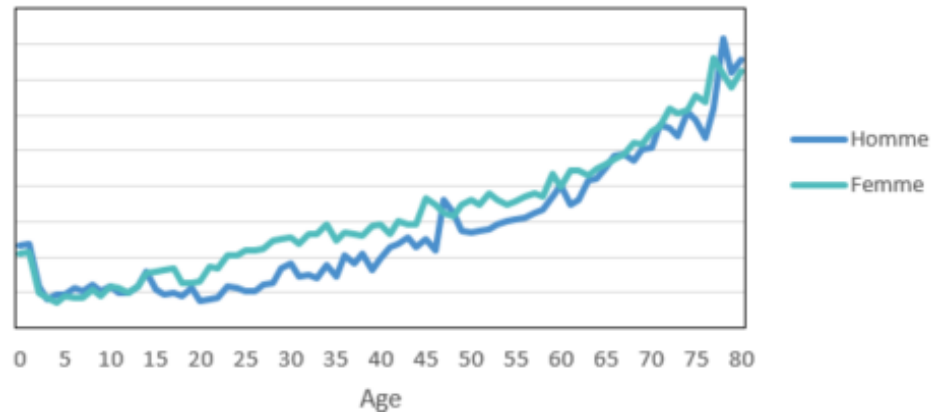


Figure 1.3.b : Le coût moyen des prestations croît avec l'âge et est plus important pour les femmes.

Le graphique représente bien un coût moyen plus conséquent pour les nouveaux nés et jeunes enfants qui croît ensuite avec l'âge. De plus, la consommation moyenne est plus importante chez les femmes que chez les hommes.

Enfin, l'âge peut avoir un impact sur le taux de résiliation moyen par année : il est de 26% en 2017 et 26% en 2018. Son évolution est représentée ci-dessous afin d'avoir une première idée de la tendance que peut suivre la résiliation au sein du portefeuille par classe d'âges.

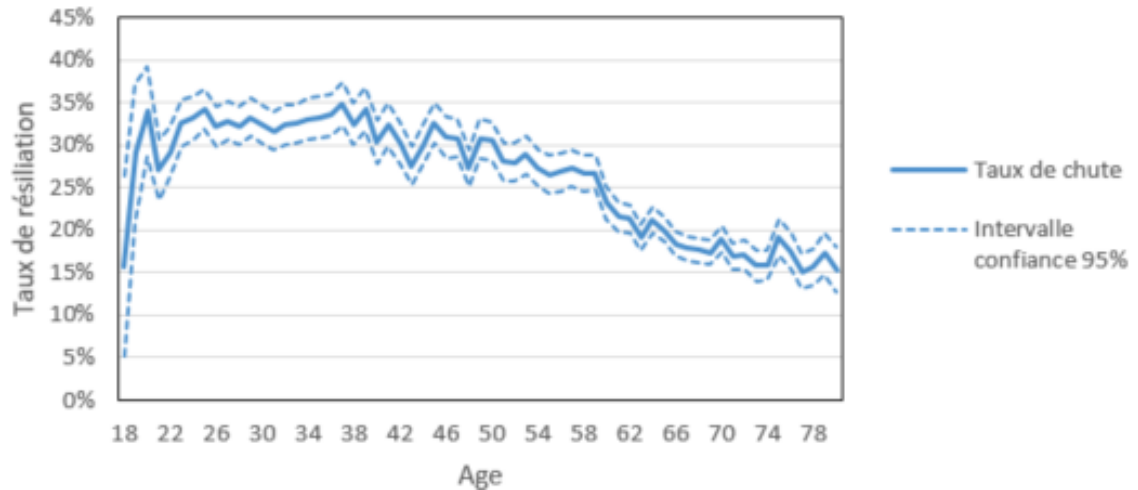


Figure 1.3.c : Le taux de résiliation moyen décroît avec l'âge pour l'année 2017.

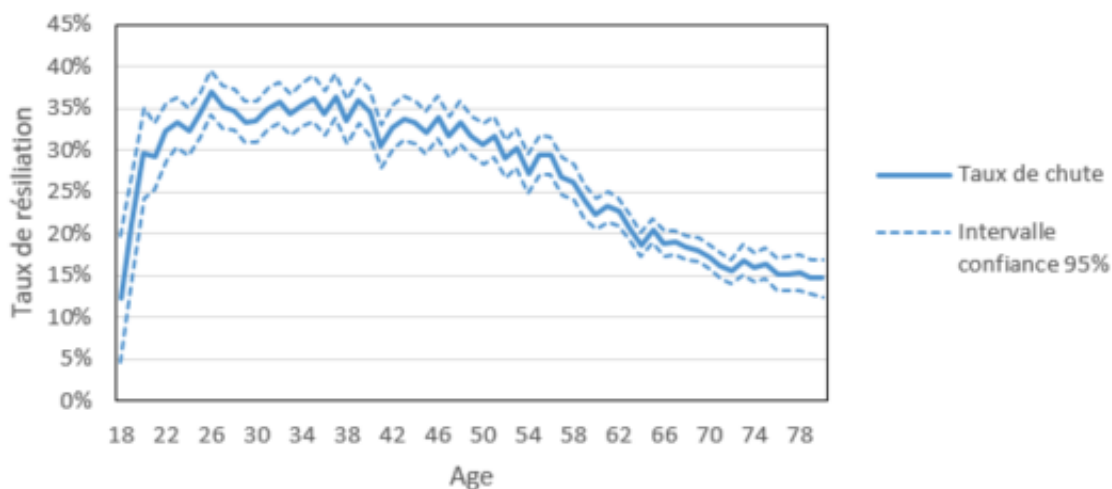


Figure 1.3.d : Le taux de résiliation moyen décroît avec l'âge pour l'année 2018.

A première vue, pour les deux années, le taux de résiliation est plutôt constant et élevé pour la classe d'âges de 20 à 50 ans puis décroît à partir de cet âge.

Analyse par gamme et niveau de garantie

Les assurés du portefeuille ont le choix parmi trois gammes de garanties et cinq niveaux de couvertures différents. L'impact de la résiliation de ces contrats ainsi que la consommation des assurés seront plus ou moins importants selon le choix effectué. Il peut donc être intéressant d'avoir une idée de la répartition des effectifs par gamme et niveau de garantie.

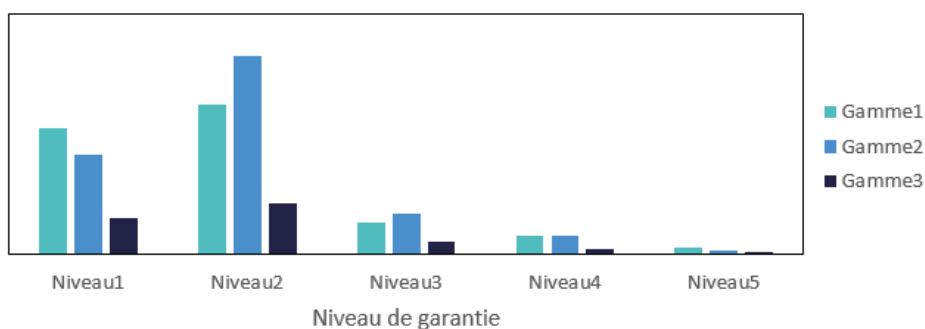


Figure 1.3.e : Les niveaux 1 et 2 sont les plus souscrits, particulièrement pour les gammes 1 et 2.

L'histogramme montre que la gamme 2 est la gamme la plus souscrite au sein du portefeuille parmi les différents contrats. La gamme 3 est relativement peu présente comparativement aux gammes 1 et 2. De plus, toutes gammes confondues, les niveaux de garantie les plus souscrits sont les niveaux 1 et 2.

L'analyse bivariée de l'âge de l'adhérent principal et de la gamme de garantie de chaque contrat peut donner des indications sur la consommation attendue de chaque gamme selon l'âge de l'assuré principal.

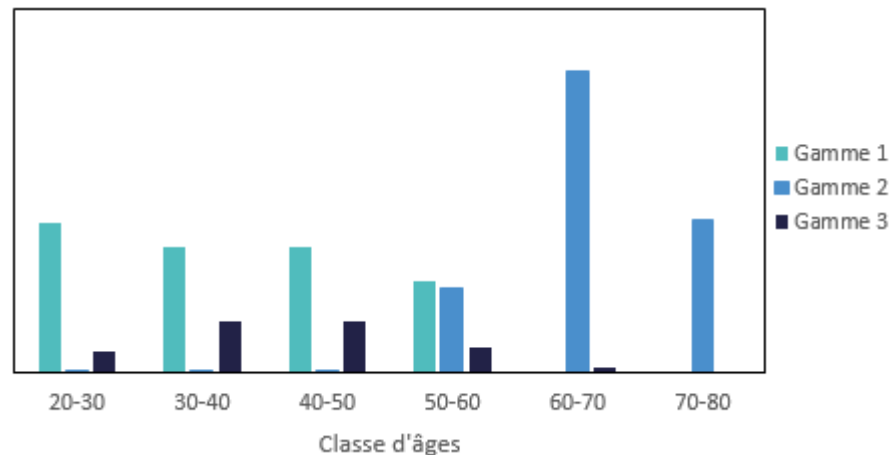


Figure 1.3.f : Les gammes 1 et 3 sont souscrites par les moins de 60 ans et la gamme 2 par les plus de 60 ans.

L'histogramme ci-dessus indique que la gamme 2 est uniquement souscrite par les personnes les plus âgées du portefeuille. A contrario, les gammes 1 et 3 ont été choisies par les classes d'âges plus jeune. La consommation moyenne par gamme peut donc être corrélée à l'âge, notamment pour la gamme 2 puisqu'il a été vu précédemment que la consommation était croissante avec l'âge.

Le nombre moyen de consommations selon la gamme de garantie est le suivant :

Gamme 1	Gamme 2	Gamme 3
7	15	5

Tableau 1.3.a: Nombre moyen de consommations par gamme de garantie.

Il apparait donc clairement que le nombre moyen de consommations est supérieur pour les assurés de la gamme 2.

Enfin, il peut être intéressant d'identifier les gammes et niveaux de garanties qui présentent un fort taux de résiliation.

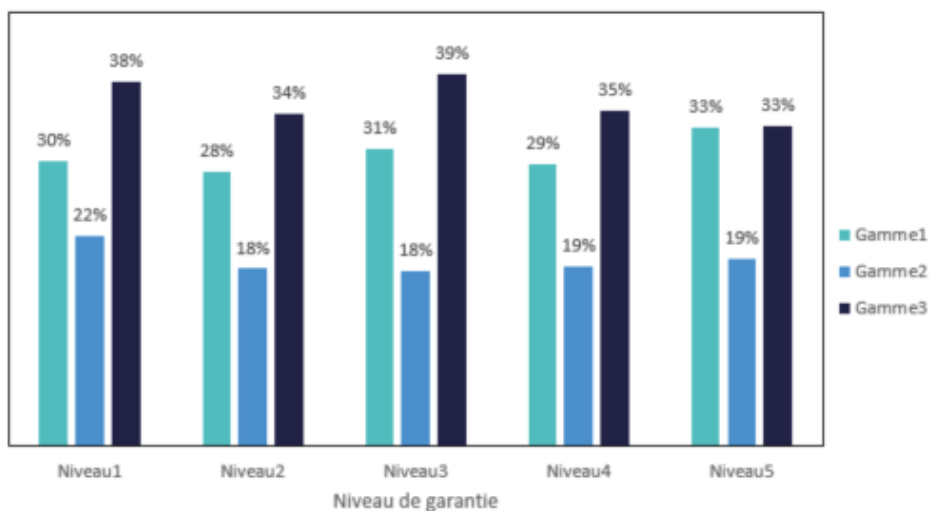


Figure 1.3.g : La gamme 3 a le taux de résiliation le plus important en 2017.

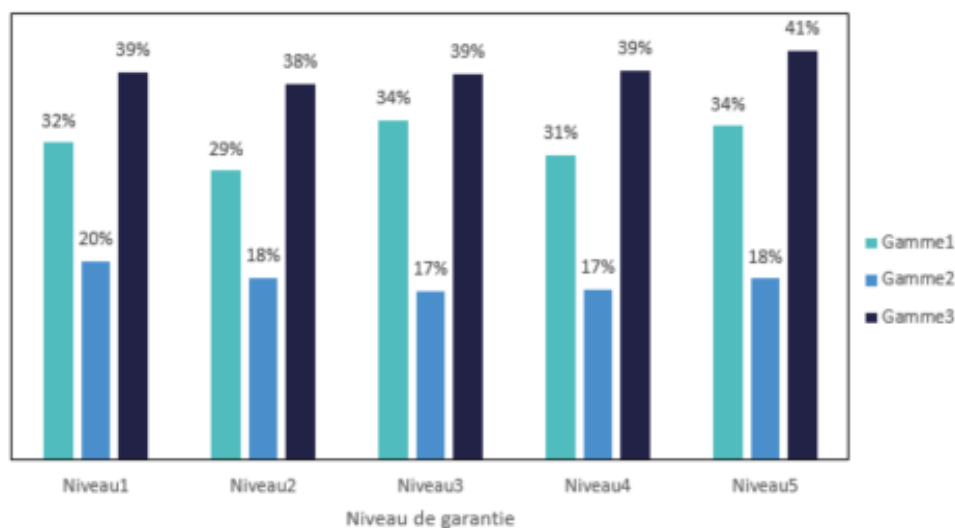


Figure 1.3.h : La gamme 3 a le taux de résiliation le plus important en 2018.

Le taux de résiliation a augmenté pour les gammes 1 et 3 sur tous les niveaux de garantie entre 2017 et 2018 mais a contrario, a diminué pour la gamme 2. La gamme 3 présente chaque année un fort taux de résiliation (36% en 2017 et 39% en 2018) alors que la gamme 2 présente un taux de résiliation relativement faible (19% en 2017 et 18% en 2018). Encore une fois, cela peut être expliqué par l'analyse précédente indiquant un taux de chute décroissant avec l'âge et un âge moyen de 67 ans pour les assurés de la gamme 2.

De plus, le taux de résiliation ne semble pas être dépendant du niveau de garantie souscrit.

Analyse par région

Le portefeuille est composé d'assurés de toutes les régions de France Métropolitaine et Outre-Mer. Le coût des prestations dépend également du lieu où le soin a été effectué. En effet, les praticiens de certaines régions sont réputés pour appliquer d'importants dépassements



d'honoraires, la consommation de médicaments est différente en milieu rural et en milieu urbain, le régime de Sécurité Sociale d'Alsace-Moselle propose des remboursements plus avantageux etc.

Le tableau ci-dessous va donc présenter une répartition de l'effectif du portefeuille par région, ainsi que le taux de résiliation moyen pour l'année 2018 et le coût moyen d'une prestation de soins pour l'assureur.

Région	Effectif	Conso. Moyenne	Taux de chute moyen
Auvergne-Rhône-Alpes	9%	11€	27%
Bourgogne-Franche-Comte	3%	10€	26%
Bretagne	3%	10€	25%
Centre-Val de Loire	5%	11€	25%
Corse	1%	11€	23%
Grand Est	7%	9€	24%
Hauts-de-France	8%	10€	26%
Ile-de-France	27%	12€	27%
Normandie	3%	11€	26%
Nouvelle-Aquitaine	7%	10€	25%
Occitanie	10%	11€	26%
Outre-Mer	7%	12€	26%
Pays de la Loire	3%	10€	25%
Provence-Alpes-Côte d'Azur	7%	11€	26%

Tableau 1.3.b : Effectif, coût moyen de prestation de soins et taux de chute moyen par région.

Le portefeuille est composé principalement d'assurés domiciliés en Ile-de-France. Cette région ainsi que les régions d'Outre-Mer, sont les régions où les remboursements complémentaires sont les plus élevés. De plus, la région Grand-Est (dont les habitants sont couverts par le régime Alsace-Moselle) est la région présentant le coût moyen le plus faible. Enfin, concernant les résiliations, la région ne semble pas être un impact discriminant (à l'exception de la Corse qui a un taux de résiliation relativement plus faible mais qui représente seulement 1% du portefeuille).

Analyse par situation de famille

La situation de famille de chaque contrat a été définie selon le nombre et le type de bénéficiaires. Les impacts de la résiliation d'un contrat ne seront pas les mêmes pour l'assureur en fonction du nombre d'ayants droit couverts.

Le tableau ci-dessous va donc présenter une répartition de l'effectif du portefeuille ainsi que le taux de résiliation moyen pour l'année 2018 par situation de famille.



Situation de famille	Effectif	Taux de chute moyen
Isolé	72%	24%
Duo	12%	23%
Couple + Enfants	6%	40%
Monoparentale	10%	37%

Tableau 1.3.c : Effectif et taux de chute moyen par situation de famille.

La majorité des contrats a une situation de famille « Isolé », l'impact de la résiliation sera donc moindre pour l'assureur. De plus, le taux de chute est inférieur au taux de chute moyen du portefeuille pour ces contrats.

Enfin, même si les contrats de type « famille » ne sont pas majoritaires dans le portefeuille, leur taux de chute est important. La résiliation de l'un de ces contrats entraîne cependant la perte simultanée de plusieurs assurés et donc une diminution de cotisation plus importante pour l'assureur.

Analyse par Régime de Sécurité Sociale

Selon le régime social auquel appartient un assuré, la base de remboursement de la Sécurité Sociale diffère, et par conséquent, il en sera de même concernant le remboursement de la complémentaire santé.

Le tableau ci-dessous va donc présenter une répartition de l'effectif du portefeuille ainsi que le taux de résiliation moyen pour l'année 2018 et le coût moyen d'une prestation de soins par régime social.

Régime social	Effectif	Conso. Moyenne	Taux de chute moyen
Régime général	89%	11	26%
MSA	1%	11	24%
Alsace Moselle	2%	5	23%
TNS	2%	12	24%
Autres régimes	5%	11	24%

Tableau 1.3.d : Effectif et taux de chute moyen par régime social.

La grande majorité du portefeuille appartient au régime général (89% du portefeuille), ce qui est légèrement supérieur à la moyenne nationale (80% de la population). De plus, il apparaît clairement que le coût moyen d'une prestation de soins est plus faible pour le régime Alsace-Moselle que pour les autres régimes puisque le taux de remboursement est plus élevé.



1.3 Synthèse

L'objectif de cette étude consiste en l'implémentation et l'utilisation d'un indicateur de Valeur Client sur un portefeuille d'assurance santé individuelle.

Cet indicateur sera implémenté en se basant sur des données concernant les contrats, les assurés, les garanties et les prestations de santé à la suite du nettoyage et tri des observations à l'aide de différentes hypothèses.

Après avoir présenté le portefeuille par une analyse statistique traditionnelle, la Valeur Client sera utilisée afin de détailler cette étude. Elle permettra notamment de segmenter le portefeuille en différents groupes homogènes, d'identifier les segments pouvant présenter ou non un intérêt pour l'assureur en termes de marge et de mettre en parallèle la probabilité de résiliation et les marges théoriques prospectives générées.

Enfin, la finalité de cette étude consistera à étudier l'impact de différents scénarios sur la Valeur Client du portefeuille global afin que l'assureur puisse mettre en place une stratégie lui permettant d'anticiper, gérer et maîtriser les risques encourus.



2. MISE EN PLACE D'UN INDICATEUR DE VALEUR CLIENT

Dans le cadre de cette étude, l'indicateur de Valeur Client est implémenté en deux étapes avec, d'une part, le calcul de la valeur historique et, d'autre part, celui de la valeur prospective. La valeur historique est obtenue à partir des bases de données présentées dans la partie précédente. La valeur prospective est calculée ici, après projection des primes et modélisation du comportement attendu de l'assuré concernant sa résiliation et ses consommations de santé.

2.1 Calcul et utilisation de la valeur historique

La première étape de la mise en place de l'indicateur de Valeur Client implémenté dans cette étude a pour but d'obtenir une photographie de l'état actuel du portefeuille. Pour cela, la valeur historique des assurés peut être calculée grâce aux données disponibles et permettra d'obtenir un premier diagnostic du portefeuille.

2.1.1 Calcul de la valeur historique

A l'aide des primes 2018 connues et du taux d'indexation appliqué par l'assureur au 1^{er} janvier 2018 (2,5% uniformément), les primes 2017 ont été recalculées. La prime utilisée est la prime commerciale nette de taxes qui permet ainsi de tenir compte de la stratégie de tarification d'origine de l'assureur. Elle est calculée selon la formule suivante :

$$Prime_{HT2017}(\text{âge}) = \frac{Prime_{HT2018}(\text{âge})}{1,025}$$

La prime perçue pour chaque assuré et pour chaque année peut ainsi être calculée selon son âge, sa gamme de garantie, son niveau de garantie et son temps de présence au sein du portefeuille sur l'année⁵. Pour l'année n et un assuré i , on a donc :

$$Prime_{Perçue_{i,n}} = Prime_n * Exposition_{i,n}$$

Le montant total de consommation de chaque assuré par année est également calculé. Pour l'année n et un assuré i , ayant consommé N prestations par année de montants X , on a donc :

$$Consommation_{i,n} = \sum_{j=1}^{N_{i,n}} X_j$$

La Valeur Client historique de l'assuré i pour l'année n est alors obtenue de la manière suivante :

$$Valeur_{Historique_{i,n}} = Prime_{Perçue_{i,n}} - Consommation_{i,n}$$

⁵ Cette durée sera appelée *exposition* dans la suite de l'étude. Elle prend une valeur entre 0 et 1.



La Valeur Client historique du portefeuille global pour l'année n peut être obtenue en sommant les valeurs historiques des N assurés couverts dans l'année :

$$ValHistTotale_n = \sum_{i=1}^N ValeurHistorique_{i,n}$$

2.1.2 Démarche générale et hypothèses retenues

Seule la valeur historique des clients n'ayant pas résilié leur contrat à la fin de l'année 2018 a été calculée. En effet, la suite de l'étude estime la valeur prospective d'un client notamment à partir de sa valeur historique et l'intérêt n'est donc pas porté sur les clients qui ont quitté le portefeuille en 2018 puisque leur valeur prospective sera assurément nulle.

a. Etude de la valeur historique 2018

La valeur historique des assurés présents sur l'année 2018 a été calculée dans un but de comparaison avec leur valeur prospective et ainsi pouvoir déterminer quels segments du portefeuille présentent un enjeu pour l'assureur. Ce calcul est basé sur une base d'environ 90 000 assurés.

Les valeurs historiques de chaque assuré ont été normalisées afin de simplifier la comparaison entre les différentes valeurs qui seront calculées dans cette étude. La valeur historique normalisée de l'assuré i est calculée de la façon suivante :

$$VHnorm_i = \frac{ValeurHistorique_i - \min(ValeurHistorique)}{\max(ValeurHistorique) - \min(ValeurHistorique)}$$

La Valeur Client historique a ensuite été découpée en 5 classes (mauvaise, faible, moyenne, bonne et forte), déterminées selon les quantiles de valeur historique normalisée du portefeuille pour l'année 2018. Ainsi, par exemple, la classe de valeur « forte » contient les 20% meilleurs assurés du portefeuille au sens de la valeur historique.

b. Comparaison entre 2017 et 2018

Une comparaison entre l'année 2017 et l'année 2018 a été effectuée afin de repérer si les assurés ont plutôt une valeur historique stable dans le temps ou au contraire, volatile.

Cette comparaison a été effectuée sur les assurés présents au sein du portefeuille sur les 2 années (ceux ayant résilié leur contrat en 2017 ou qui ont souscrit en 2018 ne sont donc pas concernés). Le calcul de cette valeur historique est donc basé sur un portefeuille d'environ 70 000 assurés.

Comme pour l'année 2018, la valeur historique normalisée de chaque assuré a été calculée puis des classes de valeurs historiques ont été déterminées selon les quantiles de cette valeur.

2.1.3 Résultats

Les résultats des deux études présentées précédemment sont exposés dans la partie suivante.



a. Etude de la valeur historique 2018

La valeur historique de 2018 a été calculée pour les assurés présents au sein du portefeuille en 2018 et n'ayant pas résilié leur contrat au cours de l'année.

Une étude des caractéristiques discriminantes de chaque classe de valeur a été menée afin d'identifier les segments du portefeuille principalement constitutifs de chaque classe.

Pour ce faire, les répartitions des effectifs de chaque critère au sein des classes de valeur sont comparées à la répartition des effectifs du portefeuille global selon le même critère. Cette analyse permet alors d'identifier directement quels critères sont principalement constitutifs de chaque classe de valeur. Toutes les analyses de cette étude suivent cette même logique.

Etude de l'âge

La répartition des assurés par âges au sein de chaque classe de valeur historique normalisée et du portefeuille global est présentée dans le tableau ci-dessous :

Classe âges	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
0-10	9%	16%	8%	0%	0%	5%
11-20	7%	14%	10%	1%	0%	6%
21-30	13%	20%	16%	3%	0%	11%
31-40	11%	12%	16%	10%	1%	12%
41-50	11%	10%	15%	13%	4%	12%
51-60	12%	9%	12%	18%	16%	14%
61-70	21%	12%	16%	34%	42%	27%
71-80	14%	5%	7%	19%	33%	12%
81-90	1%	0%	0%	1%	4%	1%
Total	100%	100%	100%	100%	100%	100%

Tableau 2.1.a : Répartition des assurés par âge au sein des différentes classes de valeur historique normalisée pour l'année 2018.

Les valeurs « bonne » et « forte » sont principalement composées d'assurés de la tranche d'âges 60-80 ans (53% pour la classe bonne valeur et 75% pour la classe forte valeur) alors qu'ils ne représentent que 39% du portefeuille global.

Il est également notable que les moins de 30 ans sont surreprésentés pour la classe de valeur « faible » puisqu'ils la composent à 50% alors qu'ils ne représentent que 22% du portefeuille.

De plus, l'âge moyen de la classe de valeur « forte » est de 66,8 ans contre 34,3 ans pour la classe de valeur « faible ».

L'étude de ce tableau permet de conclure que l'assureur aurait intérêt à cibler les plus de 60 ans afin d'augmenter sa rentabilité et qu'un travail ciblé peut être effectué sur les moins de 30 ans particulièrement représentés pour la classe de valeur « faible ». On peut en effet s'attendre à ce



que, parmi eux, certains présentent tout de même un potentiel financier pour l'assureur et qu'il peut être intéressant pour lui de déterminer précisément sur quel segment devrait-il concentrer ses efforts.

Etude de la gamme et du niveau de garantie

La répartition des assurés par gamme et niveau de garantie au sein de chaque classe de valeur historique normalisée et du portefeuille global est présentée dans le tableau ci-dessous :

Gamme	Niveau	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Gamme1	1	12%	19%	22%	4%	0%	7%
	2	16%	15%	16%	19%	6%	12%
	3	3%	2%	2%	4%	2%	3%
	4	2%	1%	1%	2%	2%	2%
	5	1%	0%	0%	0%	1%	1%
Gamme2	1	11%	8%	11%	26%	4%	13%
	2	23%	11%	13%	27%	56%	30%
	3	4%	2%	2%	4%	14%	4%
	4	2%	1%	1%	1%	8%	2%
	5	0%	0%	0%	0%	2%	1%
Gamme3	1	10%	21%	6%	1%	0%	8%
	2	12%	14%	20%	8%	3%	12%
	3	3%	3%	4%	2%	1%	3%
	4	1%	1%	1%	2%	1%	1%
	5	0%	0%	0%	0%	0%	1%
Total		100%	100%	100%	100%	100%	100%

Tableau 2.1.b : Répartition des assurés par gamme et niveau de garantie au sein des différentes classes de valeur historique normalisée pour l'année 2018.

Les valeurs « bonne » et « forte » sont principalement composées d'assurés de la gamme 2 (58% pour la classe valeur « bonne » et 84% pour la classe valeur « forte ») bien qu'ils ne représentent que 50% du portefeuille. Ces deux classes sont très peu représentées par la gamme 3 (5% pour la valeur « forte ») alors qu'elle concerne un quart des assurés du portefeuille.

Les classes de valeurs moindres sont donc principalement constituées des assurés des gammes 1 et 3. Plus particulièrement, le niveau 1 représente 48% de la classe de valeur « faible » alors qu'il ne concerne que 28% des contrats du portefeuille.

L'assureur peut donc miser sur la gamme 2 (et notamment le niveau 2 de garantie) afin d'augmenter sa rentabilité. A contrario, il peut être intéressant de cibler le niveau 1 des gammes 1 et 3 dans un but d'amélioration de leur valeur, par exemple par une revalorisation des primes de ce niveau de garantie.



Etude de la situation de famille

La répartition des assurés par situation de famille au sein de chaque classe de valeur historique normalisée et du portefeuille global est présentée dans le tableau ci-dessous :

Situation famille	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Isole	61%	50%	59%	73%	76%	68%
Duo	10%	6%	8%	14%	20%	10%
Couple+Enf	10%	15%	12%	6%	2%	7%
MonoParent	19%	29%	22%	7%	2%	15%
Total	100%	100%	100%	100%	100%	100%

Tableau 2.1.c : Répartition des assurés par situation de famille au sein des différentes classes de valeur historique normalisée pour l'année 2018.

Les valeurs « bonne » et « forte » comportent une plus grosse part de contrats « Isolé » ou « Duo » que le portefeuille global et il y a très peu de contrats avec enfants pour ces classes. A contrario, la classe de valeur « faible » contient une plus grande proportion de contrats avec enfants que le portefeuille global.

On retrouve donc les résultats obtenus avec l'étude de l'âge, indiquant que les jeunes (soit ici, les contrats avec enfants) ont tendance à avoir une valeur historique plutôt faible.

Conclusion

Cette étude a permis d'identifier les critères impactant la valeur historique du portefeuille global : **les assurés de plus de 60 ans, de la gamme 2 et sans enfant semblent particulièrement présents au sein de la classe de valeur « forte »**. Ces 3 critères sont cependant corrélés puisque la gamme 2 est principalement souscrite par des personnes âgées et qui, de surcroît, n'ont pas d'enfants à charge. A contrario, **les moins de 30 ans et assurés des contrats de type « famille », semblent plutôt avoir une valeur historique faible**, tout comme les assurés des niveaux 1 de chaque garantie. Cela peut s'expliquer notamment par le montant de prime plus faible pour les jeunes et bas niveaux de garantie.

Une étude ciblant les assurés de moins de 30 ans a été menée afin de repérer quels critères sont impactant sur la répartition entre les différentes classes de valeur historique pour ces assurés. Il apparaît alors que les classes de valeurs historiques « forte » ou « bonne » sont réparties sur les plus hauts niveaux de garantie. A l'inverse, les niveaux 1 et 2 de garantie se partagent la quasi-totalité des classes de valeur « mauvaise » et « faible ». Ainsi, parmi les moins de 30 ans, l'assureur pourrait tout de même s'intéresser à ceux ayant souscrit aux niveaux 3 ou supérieur puisque la valeur historique de ce segment d'individus n'est pas négligeable et qu'ils représentent tout de même 4% du portefeuille global.



b. Comparaison entre 2017 et 2018

La même étude a été réalisée sur les valeurs historiques de l'année 2017 afin de repérer si au global, les assurés ont une valeur historique stable dans le temps ou au contraire, volatile.

Le détail de cette étude est reporté en [annexe](#) puisque la démarche est similaire à celle de la partie précédente.

Les mêmes segments de portefeuille identifiés pour chaque classe de valeur ont été obtenus pour l'année 2017 que pour l'année 2018. A l'échelle du portefeuille global, il semble donc que les assurés ont une valeur historique stable.

Il est cependant possible qu'à l'échelle micro, des changements de classe de valeur historique se produisent d'une année à l'autre. Une analyse plus précise a été effectuée afin de déterminer si les assurés appartenant à une certaine classe de valeur en 2017 sont les mêmes qu'en 2018 ou si la stabilité de la valeur historique normalisée n'est valable qu'à l'échelle du portefeuille.

		Valeur hist. 2018				
		Mauvaise	Faible	Moyenne	Bonne	Forte
Valeur hist. 2017	Mauvaise	9%	4%	3%	2%	2%
	Faible	4%	9%	3%	2%	1%
	Moyenne	3%	4%	9%	3%	1%
	Bonne	2%	2%	3%	9%	3%
	Forte	2%	1%	1%	3%	13%

Tableau 2.1.d : Répartition des changements de classe de valeur des assurés entre 2017 et 2018.

Le tableau de répartition des effectifs ci-dessus montre que 49% des assurés du portefeuille total ne changent pas de classe de valeur entre les deux années. Dans le cas contraire, les assurés restent principalement dans une classe de valeur historique proche. Ainsi, à échelle micro, les assurés ne changent en général pas drastiquement de valeur aux yeux de l'assureur entre les deux années.

Après calcul de la valeur historique numérique totale du portefeuille pour 2017 et 2018, on remarque qu'elle augmente d'environ 5% entre les deux années.

2.1.4 Segmentation du portefeuille à partir du ratio S/P

Après avoir calculé la consommation et les primes de chaque assuré, une analyse du portefeuille selon le ratio sinistres/primes peut également être proposée. De la même manière que précédemment, l'étude de chaque critère est effectuée sans tenir compte de leurs potentielles corrélations. Il est ainsi possible d'identifier si les segments rentables selon le ratio S/P pour l'année 2018 sont les mêmes que ceux obtenus grâce à l'indicateur de valeur historique.



Un ratio inférieur à 100% est significatif de la rentabilité du portefeuille : plus le ratio est faible, plus la rentabilité est importante. A l'inverse, un ratio supérieur à 100% indique que le portefeuille est déficitaire.

Sur la totalité du portefeuille, le S/P est de 60% pour l'année 2018 : le portefeuille est donc globalement rentable. Les S/P des tableaux suivants seront exprimés en pourcentage du ratio du portefeuille global.

a. Par classe d'âges

Le tableau ci-dessous représente la valeur du ratio S/P par classe d'âges par rapport au S/P du portefeuille ainsi que la répartition des assurés au sein du portefeuille selon le même critère :

Classe âges	S/P	Répartition globale
0-10	90%	7%
11-20	82%	6%
21-30	95%	11%
31-40	98%	12%
41-50	102%	12%
51-60	95%	14%
61-70	100%	27%
71-80	110%	11%
81-90	105%	1%
Portefeuille	100%	100%

Tableau 2.1.e : S/P par tranche d'âges et répartition des effectifs au sein du portefeuille.

La rentabilité du portefeuille varie globalement peu d'une tranche d'âges à l'autre. Cependant, les assurés de 11 à 20 ans, sont quand même significativement plus rentables que le reste du portefeuille avec un S/P de 82% de celui du portefeuille mais n'en constituent que 6%. A contrario, les 71 à 80 ans, dont le S/P avoisine 110% de celui du portefeuille, sont les moins rentables.

b. Par gamme et niveau de garantie

Le tableau ci-dessous représente la valeur du ratio S/P selon la gamme et le niveau de garantie par rapport au S/P du portefeuille ainsi que la répartition des assurés au sein du portefeuille selon les mêmes critères :



Gamme	Niveau	S/P	Répartition Globale
Gamme1	1	97%	10%
	2	97%	13%
	3	118%	3%
	4	105%	2%
	5	108%	1%
Gamme2	1	108%	9%
	2	105%	27%
	3	98%	4%
	4	95%	2%
	5	73%	1%
Gamme3	1	92%	9%
	2	85%	13%
	3	90%	3%
	4	88%	1%
	5	75%	1%
Portefeuille		100%	100%

Tableau 2.1.f : S/P par gamme et niveau de garantie et répartition des effectifs au sein du portefeuille.

Les gammes 2 et 3 sont les plus rentables et leur rentabilité croît globalement avec le niveau de garantie puisque le ratio S/P décroît.

La gamme 1 est la moins rentable, notamment pour le niveau 3 de cette gamme puisque c'est pour ce niveau que le ratio S/P est le plus important. Cependant, il ne concerne que 3% des assurés.

Le niveau 5 de garantie est particulièrement rentable pour les gammes 2 et 3 mais ne concerne qu'une très faible proportion du portefeuille. Comparativement au ratio global, le niveau 2 de la gamme 3 est particulièrement rentable et concerne 13% du portefeuille ce qui n'est pas négligeable pour l'assureur.

c. Par situation de famille

Le tableau ci-dessous représente la valeur du ratio S/P selon la situation de famille par rapport au S/P du portefeuille ainsi que la répartition des assurés au sein du portefeuille selon le même critère :



Situation famille	S/P	Répartition globale
Isole	103%	64%
Duo	95%	12%
Couple+Enf	85%	7%
MonoParent	93%	17%
Portefeuille	100%	100%

Tableau 2.1.g : S/P par situation de famille et répartition des effectifs au sein du portefeuille.

La situation de famille n'est pas particulièrement explicative du S/P mais il est tout de même notable que le S/P est plus faible pour les situations de famille « Couple + Enfants » qui sont donc plus rentables.

d. Conclusion

Le portefeuille est globalement rentable. Les gammes 2 et 3 sont les plus profitables selon le calcul du S/P et leur rentabilité croit avec le niveau de garantie.

Par comparaison avec l'étude menée sur la valeur historique et les critères caractéristiques de chaque classe de valeur, on remarque que le premier indicateur donne plus d'indications sur la composition du portefeuille.

En effet, le S/P ne permet pas de séparer de façon distincte la rentabilité selon la situation de famille ou les classes d'âges alors que l'utilisation de la valeur historique a identifié clairement les plus de 60 ans comme cible pour l'assureur et les moins de 30 ans comme peu profitables.

De plus, le S/P ne permet pas de différencier sur laquelle des gammes 2 et 3 l'assureur devrait porter son attention alors que cette distinction était claire lors de la première étude.

Cette comparaison montre donc l'intérêt d'utiliser l'indicateur de Valeur Client puisqu'il apporte une vision plus détaillée du portefeuille que les indicateurs plus couramment utilisés.

2.1 Synthèse

La valeur historique de chaque assuré du portefeuille a été calculée à l'aide des primes versées à l'assureur et des remboursements effectués par ce dernier.

Une première étude a été effectuée sur les contrats existants en 2018 dans un but de comparaison avec la valeur prospective calculée par la suite. Les assurés de la gamme 2 sont apparus comme présentant un intérêt particulier en termes de marge pour l'assureur. A contrario, les moins de 30 ans présentaient une valeur assez faible : un ciblage de cette population peut donc être réalisé par l'assureur afin d'identifier si certains de ces assurés présentent tout de même une valeur « bonne » ou « forte ».



Une comparaison a été effectuée avec l'année 2017 afin de vérifier la stabilité de ces critères au cours du temps, à l'échelle globale du portefeuille puis en analysant les changements ou non de classe de valeur de chaque assuré entre les deux années d'observation.

Enfin, une analyse du portefeuille basée sur le ratio sinistres/primes a également été réalisée afin de comparer les résultats obtenus à ceux de la première partie. L'intérêt de l'indicateur de Valeur Client a été mis en avant puisqu'il permet un découpage du portefeuille selon certains critères tel que l'âge qui ne sont pas visibles avec l'utilisation du S/P.

2.2 Calcul de la Valeur Client prospective

La seconde étape de la mise en place de l'indicateur de Valeur Client implémenté dans cette étude consiste à obtenir une idée de la valeur future attendue de chaque assuré. Pour cela, la valeur prospective peut être calculée après modélisation du comportement des assurés, concernant leur durée de vie au sein du portefeuille et leur consommation en santé.

2.2.1. Présentation du modèle de résiliation

La première partie du calcul de la valeur prospective de chaque assuré consiste à modéliser sa probabilité de résiliation.

a. Utilisation d'outils de *datascience*

Différentes méthodes de *datascience* ont été utilisées lors de la création de l'indicateur de Valeur Client. Une présentation de ces méthodes est effectuée ci-dessous. [8] [11] [12] [13]

Les modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés (GLM) sont utilisés afin de prédire le comportement d'une variable aléatoire Y à partir de variables explicatives X_i et d'une fonction g exprimant le lien entre ces deux types de variables. La relation est exprimée ainsi :

$$g[E(Y|X_1 \dots X_n)] = \sum_{i=1}^n a_i * X_i$$

La construction des GLM suit les étapes suivantes :

- **Choix du modèle** : la loi de probabilité que suit la variable Y doit être déterminée parmi les lois de la famille exponentielle⁶ et la fonction lien g doit être choisie selon la loi suivie par Y .

⁶ Pour appartenir à la famille exponentielle, la variable Y doit avoir une densité de la forme :

$f_{\theta, \varphi}(y) = \exp(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi))$ avec $a()$, $b()$ et $c()$ des fonctions, $\theta \in \mathbb{R}$ le paramètre de la moyenne et $\varphi \in \mathbb{R}_+^*$ la paramètre de dispersion.



- **Estimation des coefficients de régression** : les coefficients de régression a_i sont calculés afin de déterminer l'impact (positif ou négatif) de chaque variable explicative X_i sur la variable Y , relativement à la valeur de référence pour chaque variable qualitative (une modalité est désignée ou choisie aléatoirement comme référence à la mise en place du modèle) ou proportionnellement à la valeur de chaque variable quantitative.
- **Sélection des variables explicatives** : il s'agit ici de choisir la meilleure combinaison de variables X_i d'un point de vue qualitatif pour le modèle. Afin de n'obtenir que des variables significativement explicatives, certaines variables peuvent être ôtées du modèle et des regroupements de modalités peuvent être effectués pour les variables qualitatives.
- **Validation du modèle** : plusieurs indicateurs peuvent être calculés afin de comparer la qualité de différents modèles entre eux, notamment le Critère d'Information Akaike (AIC). Ce dernier propose une estimation de la perte d'information en fonction du nombre de variables explicatives utilisées. Le meilleur modèle au sens de l'AIC sera celui dont ce critère est le plus faible. L'étude des résidus (partiels, de Pearson ou de déviance) peut également être effectuée afin d'obtenir des informations sur la qualité d'ajustement du modèle et sur les éventuelles tendances non linéaires entre les variables X_i et Y . Les observations aberrantes altérant la qualité du modèle peuvent être mises en avant en calculant leur influence sur le modèle, par exemple grâce à la distance de Cook.
- **Prédictions du modèle** : Il est possible d'obtenir une idée de la performance prédictive du modèle en l'appliquant sur un échantillon de validation et en comparant les prédictions obtenues aux observations réelles. La méthode de validation diffère selon la loi que suit la variable Y et sera détaillée par la suite.
- **Interprétation et utilisation du modèle** : une fois le modèle sélectionné et les coefficients de régression calculés, l'impact de chaque variable explicative peut être déterminé et interprété.

Les arbres CART⁷

Les arbres CART désignent les arbres de régression si la variable à expliquer est quantitative (comme l'âge, la distance...) ou les arbres de classification si la variable à expliquer est qualitative (comme le sexe, une couleur...). Ils consistent à séparer la population initiale en différentes classes homogènes. A chaque étape de classification, la population est séparée en deux selon la réponse à un test binaire, concernant une variable explicative à la fois.

Les arbres CART sont donc composés :

- D'une **racine** qui regroupe l'ensemble de la population étudiée
- De **branches** qui correspondent aux règles de segmentation de la population initiale
- De **feuilles** qui contiennent les sous-groupes de la population initiale après classification

⁷ Classification And Regression Trees



La construction d'un arbre CART doit suivre les étapes suivantes :

- **Construction de l'arbre maximal** : la classification de la population est effectuée initialement sans contrainte ni règle d'arrêt. L'arbre maximal présente cependant un problème d'*overfitting* : il a le défaut d'être extrêmement fidèle aux données d'apprentissage ce qui rendra l'utilisation du modèle difficile sur une autre population et son interprétation trop complexe.
- **Elagage** : l'arbre maximal est ensuite élagué à l'aide de règles d'arrêt définies, par exemple, sur le nombre maximal de segmentations, le nombre minimal d'observations devant être contenues dans chaque feuille etc. Plusieurs sous-arbres de l'arbre maximal qui respectent les règles passées en paramètres sont ainsi obtenus.
- **Sélection finale de l'arbre optimal** : parmi tous les arbres élagués à disposition, il s'agira de choisir celui qui optimise la classification en minimisant l'erreur relative.
- **Prédictions du modèle** : Il est possible d'obtenir une idée de la performance prédictive du modèle en l'appliquant sur un échantillon de validation et en comparant les prédictions obtenues aux observations réelles. La méthode utilisée pour l'étude de performance, qui diffère selon l'utilisation d'un arbre de classification ou de régression, sera détaillée par la suite.
- **Interprétation des résultats** : Les variables utilisées comme premiers critères de segmentation sont celles qui ont le plus d'impact sur la variable à expliquer. De plus, la probabilité de succès (dans le cadre d'une classification) ou la valeur moyenne (dans le cadre d'une régression) ainsi que la proportion de la population initiale que représente le sous-groupe sont obtenus pour chaque feuille.

Les forêts aléatoires

L'algorithme des forêts aléatoires est basé sur l'utilisation des arbres CART décrits ci-dessus. Il consiste en l'implémentation de plusieurs arbres CART, dont les variables explicatives et les individus sélectionnés pour l'apprentissage sont choisis de façon aléatoire parmi les données initiales. Plusieurs arbres sont ainsi obtenus et la prédiction finale pour un individu est choisie selon la modalité la plus fréquente qui apparaît sur l'ensemble des arbres modélisés.

La construction d'une forêt aléatoire doit suivre les étapes suivantes :

- **Construction du modèle initial** : la forêt aléatoire est initialement construite avec la variable à expliquer et toutes les variables explicatives, ainsi qu'un nombre important d'arbres la composant.
- **Choix des paramètres** : le nombre d'arbres composant la forêt aléatoire ainsi que le nombre minimal de variables utilisées peuvent être paramétrés et sélectionnés afin de minimiser l'erreur de prévision appelée erreur *out-of-bag* (OOB).
- **Prédictions du modèle** : Il est possible d'obtenir une idée de la performance prédictive du modèle en l'appliquant sur un échantillon de validation et en comparant les prédictions obtenues aux observations réelles. La méthode utilisée pour l'étude de performance, qui diffère selon le type de la variable à expliquer, sera détaillée par la suite.



- **Interprétation des résultats** : les forêts aléatoires ont le défaut d'avoir un effet « boîte noire » qui rend l'interprétation des résultats plus difficile que les deux méthodes précédentes. L'importance des variables peut être déterminée mais il est impossible d'obtenir les règles de classification retenues pour le modèle.

Dans le cadre de cette étude, l'effet « boîte noire » pourra être contourné en étudiant les caractéristiques des individus pour lesquels la prédiction de résiliation présente une forte ou faible probabilité afin d'obtenir une idée de l'impact de chaque caractéristique sur le résultat.

Etude de la prédiction des modèles

Une fois le modèle implémenté à l'aide de l'une des méthodes décrites auparavant, une étude de sa capacité de prédiction doit être effectuée. Elle consiste à appliquer le modèle sur un « échantillon de validation » et à comparer les résultats prédits à ceux réellement observés.

Dans le cas où la variable à expliquer est qualitative, il est possible d'obtenir une matrice de confusion afin de comparer les prédictions du modèle aux observations. Par exemple, si la prédiction n'a que deux issues possibles (comme c'est le cas pour le modèle de résiliation), la matrice de confusion obtenue sera de la forme suivante :

		Prédictions	
		Echec	Succès
Observations	Echec	Vrais négatifs	Faux positifs
	Succès	Faux négatifs	Vrais positifs

Tableau 2.2.a : Matrice de confusion

Initialement, le seuil de décision qui permet de classer une prédiction comme échec ou succès est fixé à 0,5. Grâce à cette matrice, plusieurs métriques indiquant la qualité du modèle peuvent être calculées :

- Le **recall ou taux de vrais positifs** représente la performance du modèle à prédire un succès lorsqu'il est observé comme tel. Il est donné par la formule suivante :

$$\text{recall} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

- Le **taux de faux positifs** correspond au taux de succès prédits mais qui n'ont pas été observés comme tel. Il est utilisé afin de tracer la courbe ROC du modèle (décrite ci-après) et est obtenu grâce à la formule suivante :

$$\text{TFP} = \frac{\text{faux positifs}}{\text{faux positifs} + \text{vrais négatifs}}$$

- La **precision** désigne le taux de succès prédits qui sont en réalité observés et est calculée par la formule :

$$\text{precision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$



La **courbe ROC**⁸ indique la performance du modèle pour la bonne prédiction des échecs et des succès puisqu'elle représente le taux de vrais positifs en fonction du taux de faux positifs, pour différentes valeurs du seuil de décision. La courbe ainsi obtenue peut être comparée à la bissectrice $y = x$ afin de déterminer la performance du modèle. En effet, plus la précision du modèle est bonne, plus la courbe ROC se rapprochera de la droite $y = 1$. A l'inverse, si elle se rapproche de la bissectrice, cela signifie que le modèle n'apporte pas plus d'informations qu'une prédiction effectuée au hasard. Le **score AUC ROC**⁹, qui correspond à l'aire sous la courbe ROC, peut ensuite être calculé. Plus il sera proche de 1, plus le modèle sera performant pour prédire les succès.

La **courbe PR**¹⁰ représente la valeur de la *precision* en fonction du *recall*, pour différentes valeurs du seuil de décision. Cet indicateur ne prend pas en compte la bonne prédiction des échecs et représente donc seulement la bonne prédiction des succès. De la même manière que pour l'AUC ROC, l'**AUC PR** peut ensuite être calculé.

L'abscisse de l'intersection des courbes ROC et PR est généralement choisie comme seuil de décision optimal pour la classification entre échec et succès.

Dans le cas où la variable à expliquer est quantitative, la prédiction du modèle peut être calculée à l'aide de plusieurs indicateurs dont :

- Le **Root mean square error (RMSE)**, qui correspond à l'erreur moyenne entre les valeurs prédites et observées. Il est donné par la formule suivante :

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

avec n le nombre de données, y_i la consommation observée et \hat{y}_i la consommation prédite.

- Le **coefficient prédictif Q_2** peut ensuite en être déduit par la formule :

$$Q_2 = 1 - \frac{RMSE^2}{Var(y)}$$

Plus sa valeur est proche de 1, plus la prédiction du modèle est satisfaisante.

Enfin, une **validation croisée** sera effectuée afin de vérifier la robustesse des modèles. Pour cela, la base initiale est découpée en K segments. Le modèle utilise, à tour de rôle, $K-1$ segments de l'échantillon comme base d'apprentissage et effectue la validation sur le K -ème segment restant, comme le montre le graphique suivant :

⁸ Receiving Operating Characteristics

⁹ Area under curve

¹⁰ Precision - Recall

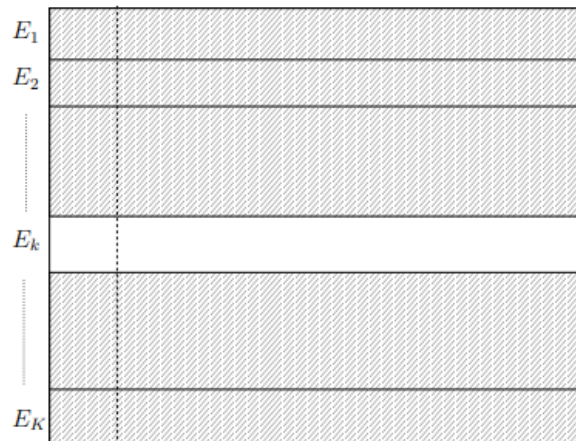


Figure 2.2.a : Découpage de l'échantillon en K segments pour la validation croisée.

Une prévision est ainsi obtenue pour chaque individu. Elles sont ensuite comparées aux valeurs observées afin d'obtenir un RMSE et un Q_2 pour chaque échantillon. Le modèle retenu sera celui qui obtient la meilleure estimation moyenne pour les K validations effectuées.

b. Hypothèses retenues pour la mise en place de la modélisation

Cette étude a été effectuée sur les données des contrats de l'année 2018 uniquement, soit environ 90 000 contrats. Les données sont alors séparées en un échantillon d'apprentissage (80% des données initiales) et un échantillon de validation du modèle (les 20% restant). Une fois la validation effectuée, le modèle pourra être utilisé afin de prédire la probabilité de résiliation d'un assuré sur les années à suivre.

Des variables ont été initialement sélectionnées afin d'expliquer la probabilité de résiliation d'un contrat. Parmi elles, on retrouve les caractéristiques propres à chaque contrat, des informations sur l'assuré principal et d'autres variables calculées à l'aide de la base de prestations afin de tenir compte des consommations de chaque contrat (seuls les postes optique et hospitalisation ont été retenus car la consommation ou non de ces deux postes est connue dans le domaine de l'assurance santé pour avoir un impact sur le comportement des assurés). L'étude se fait donc à l'aide des variables suivantes :

- L'ancienneté du contrat
- La gamme de garantie du contrat
- Le niveau de garantie du contrat
- Le nombre de bénéficiaires du contrat
- Le nombre d'enfants bénéficiaires du contrat
- La situation de famille du contrat
- Le régime social de l'assuré principal
- Le sexe de l'assuré principal
- L'âge de l'assuré principal
- La région où sont domiciliés les bénéficiaires



- Le nombre de prestations des bénéficiaires du contrat dans l'année ou la classe du nombre de prestations dans l'année
- Un indicateur de consommation de l'un des bénéficiaires en hospitalisation dans l'année
- Un indicateur de consommation de l'un des bénéficiaires en optique dans l'année

Un total de 13 variables est utilisé au départ afin de prédire la probabilité de résiliation d'un contrat. Cependant, l'étude des corrélations entre les variables peut amener à réduire ce nombre dans la suite de l'étude.

Les classes du nombre de prestations de l'année ont été créées à l'aide des quantiles sur les données et suivent donc le découpage suivant pour les modèles linéaires généralisés :

Classe de volume	Nombre de prestations
NUL	= 0
FAIBLE]0 ; 5]
MOYEN]5 ; 10]
FORT]10 ; 20]
IMPORTANT	> 20

Tableau 2.2.b : Définition des classes de volume de prestations pour le GLM

Pour les modèles CART et de forêt aléatoire, la consommation a été découpée en 10 classes différentes. En effet, ces méthodes de *datascience* utilisent la classification selon les différentes modalités qui peut alors être plus précise que pour les GLM, qui calculent un coefficient pour chaque modalité et dont la différenciation peut être moindre. Elles sont définies de la façon suivante :

Classe de volume	Nombre de prestations
Classe 0	= 0
Classe 1	= 1
Classe 2]1 ; 4]
Classe 3]4 ; 6]
Classe 4]6 ; 9]
Classe 5]9 ; 12]
Classe 6]12 ; 16]
Classe 7]16 ; 20]
Classe 8]20 ; 30]
Classe 9	> 30

Tableau 2.2.c : Définition des classes de volume de prestations pour les modèles CART et forêts aléatoires.



c. Etude réalisée et résultats obtenus

Etude des corrélations

Les variables explicatives d'un modèle linéaire généralisé doivent suivre une hypothèse d'indépendance.

Une première sélection des variables qualitatives peut être effectuée grâce à l'étude des corrélations par un test de Chi-deux. Ce dernier est utilisé pour tester l'hypothèse qu'il n'y a pas de relation entre deux variables qualitatives. Les résultats obtenus après avoir effectué le test sur les variables deux à deux ne permettent pas de repérer une dépendance entre deux variables. Elles seront donc toutes conservées.

Pour les variables quantitatives, le calcul du coefficient de corrélation de Pearson permet de déterminer la nature de la relation entre deux variables. Il est calculé à l'aide de la formule

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Une valeur proche de 0 indique l'absence de corrélation linéaire alors qu'une valeur proche de -1 ou 1 indique une relation linéaire forte.

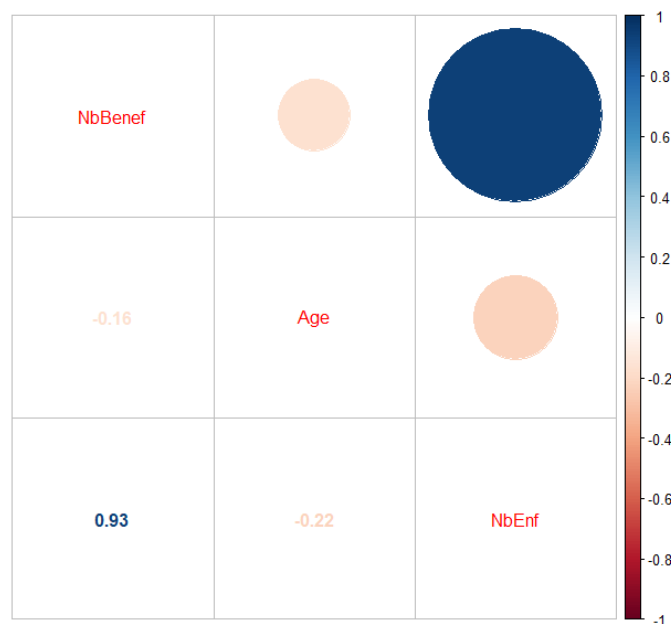


Figure 2.2.b : Les variables NbBenef et NbEnf sont fortement corrélées

Le graphique ci-dessus indique que le coefficient de Pearson obtenu pour le nombre de bénéficiaires du contrat et le nombre d'enfants couverts est de 0,93. Ces deux variables sont fortement corrélées et seule la variable indiquant le nombre de bénéficiaires sera conservée puisque c'est elle qui présente la corrélation la plus faible avec la variable Age.

Ainsi, l'étude des corrélations a permis de supprimer la variable indiquant le nombre d'enfants bénéficiaires du contrat.



Modèle linéaire généralisé

Un modèle linéaire généralisé a été implémenté afin d'expliquer les résiliations de l'année 2018 à l'aide des variables décrites ci-dessus et de l'échantillon d'apprentissage.

La loi de probabilité associée à la variable aléatoire de résiliation est la loi de Bernoulli puisque cette variable ne peut avoir que deux valeurs possibles : 0 si l'assuré n'a pas résilié son contrat, 1 le cas échéant.

La fonction lien utilisée dans ce modèle est la fonction *logit* $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ avec $\mu = E[Y]$. C'est la fonction lien généralement utilisée avec la loi de Bernoulli.

Les résultats obtenus pour ce premier modèle indiquent que la variable « Région » n'est globalement pas significative. De plus, certaines modalités des variables « Ancienneté », « Régime Social » et « Niveau de garantie » ne sont pas significatives. Les modifications suivantes ont été effectuées :

- Les niveaux de garantie 3, 4 et 5 ont été regroupés en un niveau 3+.
- Les anciennetés de 4, 5, 6, et 7 ans ont été regroupées en une ancienneté 4+.
- Les régimes spéciaux ont été regroupés afin d'avoir pour cette variable, uniquement les modalités « Régime général » et « Autres régimes ».
- La variable « Région » n'a pas été conservée.
- Le modèle a été implémenté à l'aide des classes de nombre de consommations et non pas du nombre exact de consommations.

L'étude du critère AIC est ensuite réalisée afin de sélectionner le meilleur modèle et de supprimer les variables augmentant l'erreur du modèle. Un algorithme *stepwise* de direction *backward* est appliqué au modèle : le critère AIC est calculé initialement avec toutes les variables explicatives du modèle puis la variable la moins significative est retirée à chaque étape de l'étude. Si le critère AIC diminue après retrait de la variable, elle ne sera pas retenue dans le modèle optimal. Ici, le modèle initial est le plus performant et toutes les variables sont conservées.

L'étude des résidus ou de la distance de Cook n'a pas vraiment de sens dans le cadre d'un modèle logistique du fait que la variable à expliquer soit binaire. Afin de valider la qualité du modèle, sa qualité de prédiction va être étudiée grâce à l'échantillon de validation.

Les prédictions des résiliations obtenues à l'aide du modèle implémenté donnent la matrice de confusion suivante avec un seuil $\alpha = 0,5$:

	Prédiction 0	Prédiction 1
Observation 0	13655	370
Observation 1	3822	649

Tableau 2.2.d : Matrice de confusion des prédictions du GLM.

Le seuil n'est pas adapté puisque peu de résultats sont prédits à 1, donc peu de contrats sont prédits comme résiliés. Ici, le *recall* est de 14,5% et la *precision* est de 63,7%.



En effet, un déséquilibre apparaît dans les données puisque la probabilité de résiliation observée en réalité n'est pas de 0,5.

Lorsque les courbes ROC et PR sont tracées, leur intersection permet de déterminer le seuil α optimal afin d'obtenir un équilibre entre le *recall* et la *precision*. Le graphique suivant est alors tracé :

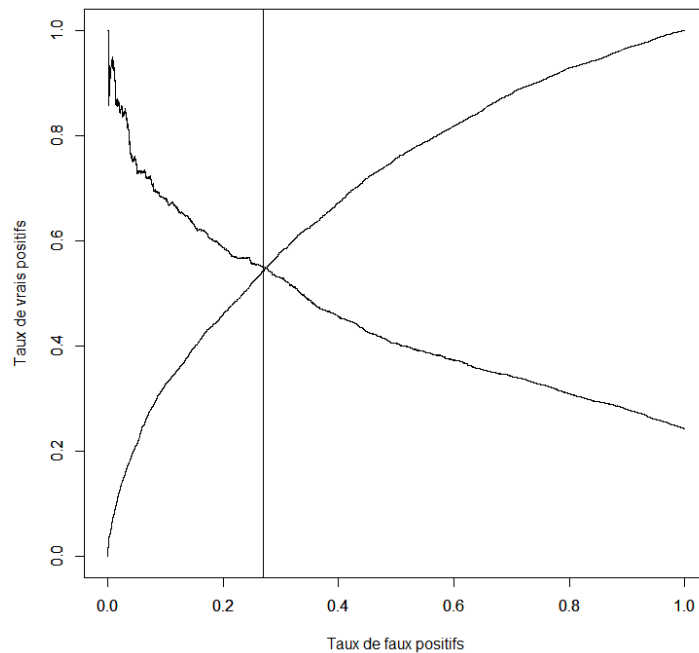


Figure 2.2.c : Les courbes ROC et PR des prédictions du GLM s'intersectent en un point d'abscisse 0,27.

La courbe ROC est supérieure à la bissectrice ce qui indique que le modèle est plus informatif que le hasard. De plus, l'AUC ROC du modèle est de 0,7 et l'AUC PR de 0,451. La prédiction du modèle est donc acceptable.

Le seuil α est modifié afin d'être fixé à 0,27, valeur qui correspond à l'intersection des deux courbes sur le graphique ci-dessus. La matrice de confusion suivante est ainsi obtenue :

	Prédiction 0	Prédiction 1
Observation 0	2692	4530
Observation 1	1779	2846

Tableau 2.2.e : Matrice de confusion des prédictions du GLM après modification du seuil.

Le *recall* est ici de 60,2% mais la *precision* diminue à 37,3%.

Arbre CART

La méthode des arbres CART a ensuite été utilisée afin d'implémenter un arbre de classification à l'aide des données d'apprentissage. L'arbre maximal a été construit en utilisant les 12 variables explicatives non corrélées utilisées initialement lors du modèle linéaire généralisé (c'est la variable des classes de prestations qui a été conservée pour ce modèle également). L'arbre



ainsi obtenu indique que l'âge, la région, la classe du nombre de prestations et le niveau de garantie sont les variables les plus importantes. Cependant, l'arbre contient environ 20 000 feuilles ce qui est un nombre trop important pour être exploité.

L'arbre maximal va être élagué et pour cela, des règles d'arrêt vont être définies afin de minimiser l'erreur. Ici, l'objectif est d'obtenir un arbre élagué dont le paramètre de complexité¹¹ cp minimise l'erreur. Le graphique ci-dessous représente l'erreur relative de l'arbre selon la valeur de cp choisie :

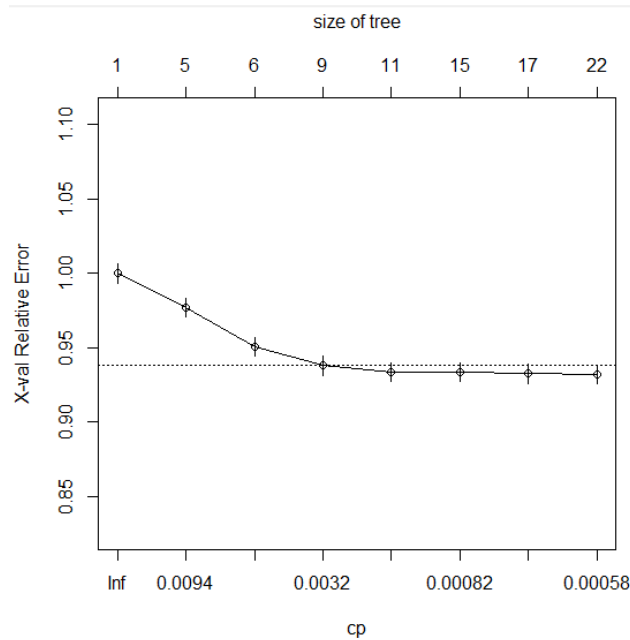


Figure 2.2.d : Le cp minimisant l'erreur donne un arbre CART de 22 feuilles.

Le cp minimisant l'erreur donne un arbre composé de 22 feuilles. Un problème d'*overfitting* apparaît alors : l'arbre obtenu est trop précis et découpe la population initiale en un nombre de classes trop élevé.

Un cp de 0,0023 est finalement choisi puisque l'erreur relative avec cette valeur est relativement proche de celle obtenue avec le cp minimal. L'arbre obtenu sera plus facile à interpréter puisqu'il ne contiendra que 11 classes différentes.

L'arbre élagué suivant est donc obtenu :

¹¹ Ce paramètre indique que si la séparation n'améliore pas l'adéquation d'au moins la valeur cp , l'arbre ne segmente pas.

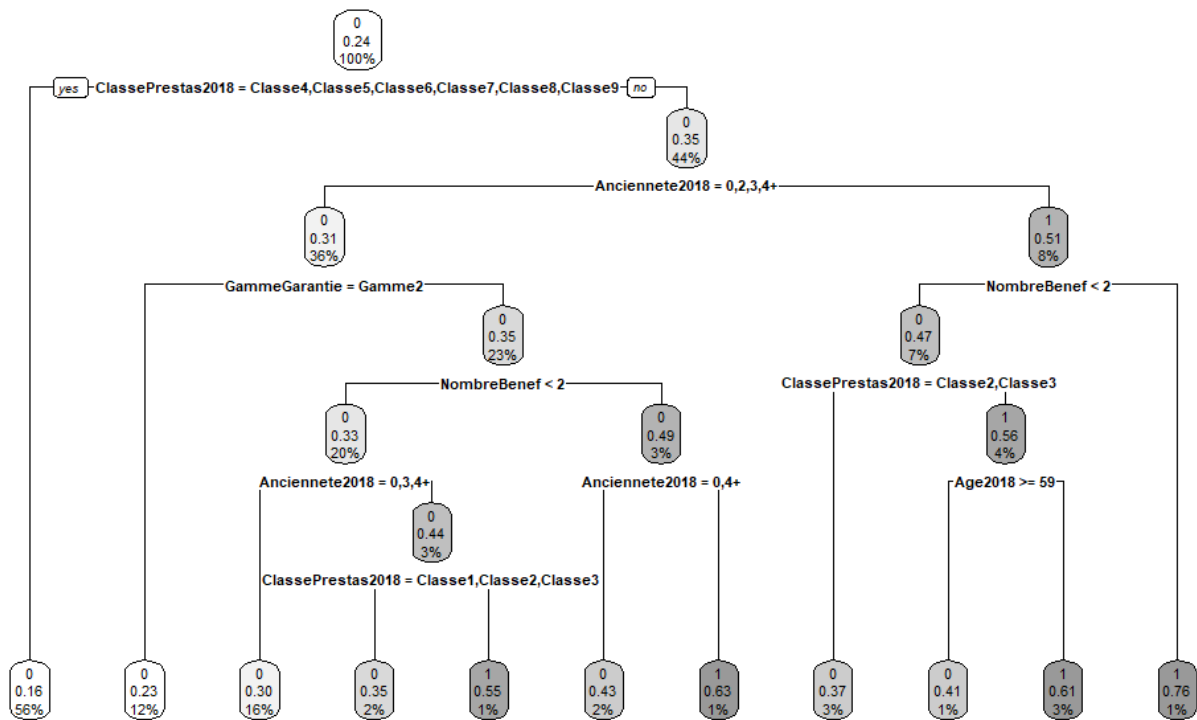


Figure 2.2.e : La classe du nombre de prestations dans l'année est la variable la plus discriminante de la probabilité de résiliation.

Les variables discriminantes concernant la résiliation sont donc le nombre de consommations dans l'année, l'ancienneté du contrat, la gamme de garantie et le nombre de bénéficiaires.

Les prédictions des résiliations obtenues à l'aide du modèle implémenté donnent la matrice de confusion suivante avec un seuil $\alpha = 0,5$:

	Prédiction 0	Prédiction 1
Observation 0	13525	500
Observation 1	3686	785

Tableau 2.2.f : Matrice de confusion des prédictions du CART.

Le seuil n'est cependant pas adapté puisque peu de résultats sont prédits à 1, donc peu de contrats sont prédits comme résiliés. Ici, le *recall* est de 17,6% et la *precision* est de 61,1%.

En effet, un déséquilibre apparaît dans les données puisque la probabilité de résiliation observée en réalité n'est pas de 0,5. Les courbes ROC et PR ont été tracées afin de déterminer un seuil de précision plus adapté. Le seuil α est modifié afin d'être fixé à 0,27 qui correspond à l'intersection des deux courbes.

De plus, l'AUC ROC du modèle est de 0,65 et l'AUC PR de 0,42. La prédiction du modèle est donc plutôt faible.

Après modification du seuil, la matrice de confusion suivante est ainsi obtenue :



	Prédiction 0	Prédiction 1
Observation 0	10316	3709
Observation 1	2204	2267

Tableau 2.2.g : Matrice de confusion des prédictions du CART après modification du seuil.

Le *recall* est ici de 50,7% mais la *precision* diminue à 37,9%.

Forêt aléatoire

Un modèle de forêt aléatoire a été implémenté sur les 12 variables présentées auparavant (c'est la variable des classes de prestations qui a été conservée pour ce modèle également) [14].

Le modèle initial est composé de 1000 arbres et sans indication concernant le nombre minimal de variables utilisées pour chaque arbre (appelé *mtry*). Il est fixé par défaut à la racine carré du nombre de variables explicatives (ici $mtry_{default} = \sqrt{14} \approx 3$). Le taux d'erreur OOB ainsi obtenu est de 22,63%.

Le taux d'erreur en fonction du nombre d'arbres composant la forêt aléatoire est tracé et le taux d'erreur se stabilise à partir de 800 arbres. Le paramètre du nombre d'arbres est donc défini à 800. De plus, la valeur de *mtry* est généralement choisie parmi $mtry_{default}$, $mtry_{default} * 2 = 6$ et $\frac{mtry_{default}}{2} \approx 2$. La valeur de ce paramètre retenue sera celle qui minimise le taux d'erreur OOB soit $mtry = 3$.

Ce modèle indique que l'âge, le nombre de prestations, la région, l'ancienneté et le niveau de garantie sont les variables qui ont le plus d'impact sur la résiliation.

Afin de valider la qualité du modèle, la matrice de confusion, les courbe ROC et PR et l'AUC ROC vont de nouveau être utilisées.

Les prédictions des résiliations obtenues à l'aide du modèle implémenté donnent la matrice de confusion suivante avec un seuil $\alpha = 0,5$:

	Prédiction 0	Prédiction 1
Observation 0	13131	894
Observation 1	3374	1097

Tableau 2.2.h : Matrice de confusion des prédictions du modèle de forêt aléatoire.

Le seuil n'est cependant pas adapté puisque peu de résultats sont prédits à 1, donc peu de contrats sont prédits comme résiliés. Ici, le *recall* est de 24,5% et la *precision* est de 55,1%.

En effet, un déséquilibre apparaît dans les données puisque la probabilité de résiliation observée en réalité n'est pas de 0,5. Les courbes ROC et PR ont été tracées afin de déterminer un seuil de précision plus adapté. Le seuil α est modifié afin d'être fixé à 0,27 qui correspond à l'intersection des deux courbes.

De plus, l'AUC ROC du modèle est de 0,68 et l'AUC PR de 0,45.



Après modification du seuil, la matrice de confusion suivante est ainsi obtenue :

	Prédiction 0	Prédiction 1
Observation 0	9089	4936
Observation 1	1678	2793

Tableau 2.2.i : Matrice de confusion des prédictions du modèle de forêt aléatoire après modification du seuil.

Le *recall* est ici de 62,5% mais la *precision* diminue à 36,1%.

Choix du modèle

Trois modèles de *datascience* ont été utilisés afin de modéliser les résiliations des assurés. Les critères AUC ROC et AUC PR des trois modèles sont rappelés ci-dessous afin de sélectionner le plus performant.

	GLM	Arbre CART	Forêt aléatoire
AUC ROC	0,7	0,65	0,68
AUC PR	0,45	0,42	0,45

Tableau 2.2.j : Récapitulatif des critères AUC ROC et AUC PR des différents modèles

Le modèle GLM sera donc finalement retenu afin d'identifier les segments du portefeuille étant les plus susceptibles de résilier leur contrat puisqu'il présente l'AUC ROC et l'AUC PR les plus importants.

La probabilité de résiliation des assurés dans les années à venir peut être déterminée grâce à ce modèle selon les caractéristiques des contrats, de l'assuré principal et de consommation qui ont été retenues comme variables explicatives. Pour rappel, la fonction *logit* a été utilisée pour ce modèle. Si on note a_i les coefficients obtenus lors de la modélisation pour chaque variable X_i et p la probabilité de résiliation, on a donc en théorie :

$$\sum_{i=1}^n a_i * X_i = \ln \left(\frac{p}{1-p} \right)$$

Cependant, dans la pratique, la relation n'est pas totalement linéaire et une variable appelée « intercept » a été ajoutée lors de la modélisation. Elle correspond à la valeur de cette relation pour l'individu de référence. Elle est notée β , et la formule de la relation est donc la suivante :

$$\beta + \sum_{i=1}^n a_i * X_i = \ln \left(\frac{p}{1-p} \right)$$

La probabilité de résiliation d'un contrat peut donc être obtenue de la façon suivante :

$$p = \frac{1}{1 + \exp(-(\beta + \sum_{i=1}^n a_i * X_i))}$$



d. Analyse des résultats

Grâce aux coefficients obtenus avec le modèle GLM (dont les résultats détaillés sont reportés en [annexe](#)) et à la formule ci-dessus, il est alors possible de calculer la probabilité de résiliation assimilée à chaque contrat.

Afin d'obtenir la probabilité de résiliation pour les années à venir, il convient de tenir compte des changements de valeur de certaines variables d'année en année. Certaines variables varient de façon certaine (comme l'âge et l'ancienneté), d'autres sont considérées par hypothèse comme invariables dans le cadre de cette étude (comme la situation de famille, la gamme ou le niveau de garantie) et d'autres devront être modélisées (comme la consommation en hospitalisation ou optique).

Une étude détaillée sur la probabilité de résiliation ainsi modélisée a été menée afin d'identifier les profils des contrats fidèles et volatiles obtenus, à partir de la base de données de l'année 2018 (pour l'âge de l'assuré principal, l'ancienneté du contrat et les variables indicatrices de la consommation). A l'aide de la fonction de répartition de la probabilité de résiliation, il a été décidé de concentrer l'étude sur les assurés ayant une probabilité de résiliation inférieure ou égale à 10% (considérée comme faible) et ceux ayant une probabilité de résiliation supérieure ou égale à 50% (considérée comme forte). En effet, c'est principalement entre ces deux valeurs que la fonction de répartition croît : une probabilité de résiliation hors de ces bornes est donc notable et peut faire l'objet d'une étude plus approfondie.

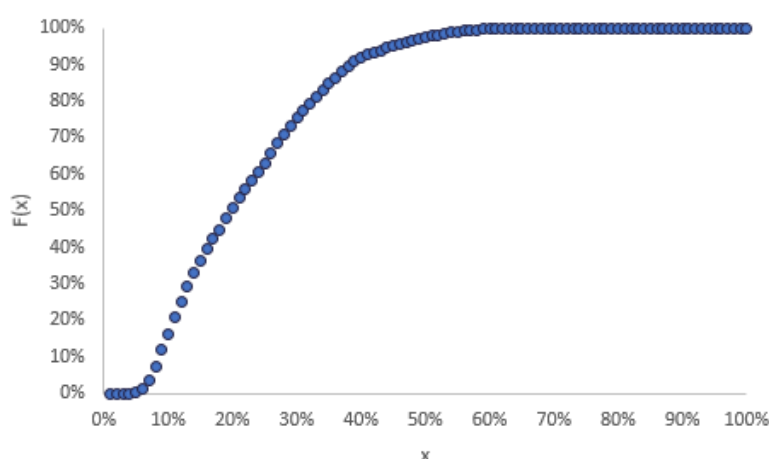


Figure 2.2.f : La courbe de la fonction de répartition croît principalement entre une probabilité de 10% et 50%.

Le tableau suivant compare la répartition des contrats selon différents critères, pour le portefeuille global, pour les contrats ayant une probabilité de résiliation faible ($\leq 10\%$) et ceux ayant une forte probabilité de résiliation ($\geq 50\%$). Seuls les critères pour lesquels un résultat notable a été remarqué sont reportés dans ce tableau.



	Portefeuille global	Faible proba. résil.	Forte proba. résil.
Proportion portefeuille	100%	19%	3%
Age moyen assuré principal	55,26	67,85	37,96
Nb moyen de bénéficiaires	1,29	1,05	2,14
Gamme Garantie 1	37%	9%	52%
Gamme Garantie 2	53%	91%	1%
Gamme Garantie 3	10%	0%	47%
Ancienneté : 0	22%	25%	17%
Ancienneté : 1	15%	0%	68%
Ancienneté : 2	15%	7%	8%
Ancienneté : 3	16%	16%	4%
Ancienneté : 4+	31%	51%	4%
Classe Prestas : NUL	16%	0%	56%
Classe Prestas : FAIBLE	19%	0%	34%
Classe Prestas : MOYEN	15%	1%	6%
Classe Prestas : FORT	23%	34%	2%
Classe Prestas : IMPORTANT	27%	65%	1%

Tableau 2.2.k : Répartition des effectifs selon plusieurs critères parmi les contrats à faible et forte probabilité de résiliation

Les assurés principaux des contrats à **faible probabilité de résiliation** ont un âge moyen supérieur à celui du portefeuille global. En effet, ils concernent à 91% la **gamme de garantie 2** qui est souscrite majoritairement par les séniors. De plus, l'ancienneté influence fortement la probabilité de résiliation puisque 51% de ces adhérents fidèles ont une **ancienneté supérieure ou égale à 4 ans**. Enfin, la totalité des contrats à faible probabilité de résiliation concerne les personnes ayant une **consommation forte ou importante**.

Les contrats à **forte probabilité de résiliation** ont un âge moyen inférieur à celui du portefeuille global. Il s'agit également des contrats couvrants **plusieurs bénéficiaires** (2,14 en moyenne) et principalement au bout **d'un an d'ancienneté** (68%). Enfin, la quasi-totalité des contrats à forte probabilité de résiliation (90%) concerne les personnes ayant une **consommation nulle ou faible**. Ces contrats ne représentent que 3% du portefeuille et pourraient donc être négligés sans que leur résiliation n'ait un réel impact sur le portefeuille. Cependant, on peut s'attendre à ce que leur valeur prospective soit intéressante pour l'assureur puisque les assurés de ces contrats ne consomment pas ou peu.

La probabilité de résiliation modélisée est donc directement dépendante de la consommation ou non des assurés. Elle est importante au bout d'un an d'ancienneté, ce qui correspond à la durée minimale légale afin de pouvoir résilier un contrat d'assurance santé individuelle. Au contraire, si l'assureur arrive à fidéliser les contrats pendant au moins 4 ans, la probabilité de résiliation de ces



derniers chute drastiquement. Enfin, les contrats les plus susceptibles d'être résiliés couvrent en moyenne 2 bénéficiaires ce qui entraîne une perte de prime plus importante pour l'assureur.

2.2.2 Présentation du modèle de consommation

La probabilité de résiliation ayant été modélisée pour chaque contrat, les consommations des assurés vont maintenant être modélisées afin d'obtenir leur valeur prospective.

a. Montant de consommation moyen

Le montant de consommation moyen a été calculé à l'aide de la base de données de prestations sur 3 postes différents. En effet, le modèle de résiliation implémenté précédemment nécessite la différenciation de consommation des postes Hospitalisation, Optique et autres postes. Le même découpage a donc été effectué pour le coût de ces prestations. Pour l'Hospitalisation et l'Optique, le coût total par personne a été considéré alors que pour les autres postes, c'est un coût moyen par prestation qui a été calculé. Ces coûts ont été différenciés selon les caractéristiques suivantes de chaque assuré :

- L'âge
- Le sexe
- La gamme de garantie
- Le niveau de garantie

L'âge et le sexe ont été sélectionnés puisque ces critères sont apparus comme influents sur le montant de consommation lors de l'étude descriptive des données, détaillée en première partie. De plus, il est connu qu'il existe des biais de consommation selon les garanties proposées pour certains postes comme l'optique ou le dentaire. Les critères de la gamme et du niveau de garantie ont donc également été retenus.

L'âge est la seule variable étant amenée à changer dans cette étude d'une année à l'autre et donc à différencier le montant de consommation moyen de chaque assuré selon les années. Ainsi, un lissage par moyenne mobile a été effectué sur ce critère afin d'harmoniser les montants de consommation des individus. Par sexe, gamme et niveau de garantie, le montant de consommation pour l'âge n est donc calculé de la façon suivante :

$$Conso_{lissée}[n] = \frac{Conso[n-2] + Conso[n-1] + Conso[n] + Conso[n+1] + Conso[n+2]}{5}$$

b. Modélisation des données explicatives de la résiliation

Il s'est avéré que certaines données de consommation étaient explicatives de la probabilité de résiliation d'un contrat. Dans un but d'estimation de la résiliation sur les années à venir, il faut donc réussir à prédire ces données. Le comportement des assurés quant à leur consommation ou non en hospitalisation et optique et le nombre moyen de consommations par an doivent donc être déterminés afin que toutes les variables nécessaires à la prédiction de la probabilité de résiliation soient disponibles.



Etude de la consommation en hospitalisation et optique

La consommation en hospitalisation et optique a été modélisée à l'aide de la base de données de prestations. La probabilité de consommation pour chacun des postes a été calculée selon les caractéristiques suivantes de chaque assuré :

- Son âge
- Son sexe
- Sa gamme de garantie
- Son niveau de garantie

Ces critères de différenciation sélectionnés sont les mêmes que pour ceux du coût moyen afin de garder une certaine cohérence dans la modélisation des consommations.

L'âge étant la seule variable amenée à varier dans cette étude d'une année à l'autre et donc à différencier la probabilité de consommation de chaque assuré, un lissage par moyenne mobile a été effectué sur ce critère. Par sexe, gamme et niveau de garantie, la probabilité de consommation pour l'âge n est donc calculée de la façon suivante :

$$Proba_{\text{lissée}}[n] = \frac{Proba[n - 2] + Proba[n - 1] + Proba[n] + Proba[n + 1] + Proba[n + 2]}{5}$$

La consommation ou non en Optique et Hospitalisation des assurés suivent donc des lois de Bernoulli de probabilité ainsi calculée. Pour chaque individu, 1000 simulations de Bernoulli ont ensuite été effectuées afin d'obtenir leur consommation en Optique et Hospitalisation de chaque année.

En regroupant le nombre de consommations de chaque assuré par contrat en optique et en hospitalisation, les indicateurs de consommation de ces postes nécessaires à la modélisation de la résiliation ont ainsi pu être obtenus.

Modélisation du nombre de consommations

Afin de modéliser le nombre de consommations par assuré, les mêmes méthodes de *datascience* vont être utilisées que pour la prédiction de la probabilité de résiliation.

Choix des variables explicatives

Cette étude a été effectuée sur les données des assurés pour l'année 2018 uniquement, qui n'ont pas résilié ni en 2017 ni en 2018. De plus, seules les données attritionnelles sont conservées afin que les consommations exceptionnelles d'un assuré n'influencent pas la qualité du modèle. De ce fait, les assurés ayant consommé plus que le quantile à 95% du nombre de consommations de la base de données ne sont pas considérés au sein de l'étude. Les données sont alors séparées en un échantillon d'apprentissage (80% des données initiales) et un échantillon de validation du modèle (les 20% restant).

Des variables ont été initialement sélectionnées afin d'expliquer le nombre de consommations d'un assuré. Parmi elles, on retrouve les caractéristiques propres à chaque individu, des informations sur son contrat et d'autres variables indicatrices de sa consommation



calculées à l'aide de la base de prestations de l'année précédente. En effet, une personne ayant fortement consommé lors de l'année n peut présenter des problèmes de santé et donc consommer fortement également en année $n+1$. De plus, les mêmes postes de soins que pour la modélisation de la résiliation ont été retenus. L'étude se fait donc à l'aide des variables suivantes :

- L'ancienneté du contrat de l'assuré
- La gamme de garantie du contrat de l'assuré
- Le niveau de garantie du contrat de l'assuré
- La situation de famille du contrat de l'assuré
- Le régime social de l'assuré principal du contrat
- Le sexe de l'assuré
- L'âge de l'assuré
- La région de l'assuré principal du contrat
- Le nombre de consommations de l'assuré dans l'année précédente
- Un indicateur de consommation en hospitalisation de l'assuré dans l'année précédente
- Un indicateur de consommation en optique de l'assuré dans l'année précédente
- Un indicateur de consommation en hospitalisation de l'assuré dans l'année en cours
- Un indicateur de consommation en optique de l'assuré dans l'année en cours

Un total de 13 variables est donc sélectionné afin de prédire le nombre de consommations d'un assuré. Cependant, l'étude des corrélations peut amener à supprimer certaines variables.

Le test du Chi-deux sur les variables qualitatives a fait ressortir une corrélation entre le régime social de l'assuré et la consommation en optique dans l'année précédente et actuelle. Le régime social de l'assuré n'a donc pas été conservé comme variable explicative.

Le coefficient de corrélation de Pearson entre l'âge et le nombre de consommations de l'année précédente est de 0,27. Les deux variables quantitatives sont conservées pour la suite de l'étude.

Modèle linéaire généralisé

Un modèle linéaire généralisé a été implémenté afin d'expliquer le nombre de consommations d'un assuré à l'aide des 12 variables listées ci-dessus et de l'échantillon d'apprentissage.

La loi de probabilité associée à la variable aléatoire de comptage correspond généralement à une loi de Poisson, avec une fonction lien *log*. Pour cela, les conditions suivantes doivent être respectées :

- Indépendance des données
- Distribution des données selon une loi de Poisson de paramètre λ
- Absence de sur-dispersion (soit $E[Y] = \text{Var}[Y] = \lambda$)

Les données sont supposées indépendantes puisque la consommation d'un assuré n'est pas conditionnée à la consommation d'un autre. La fonction de répartition des données observées est



ensuite comparée à la fonction de répartition théorique d'une loi de Poisson de paramètre $\lambda = E[Y]$.

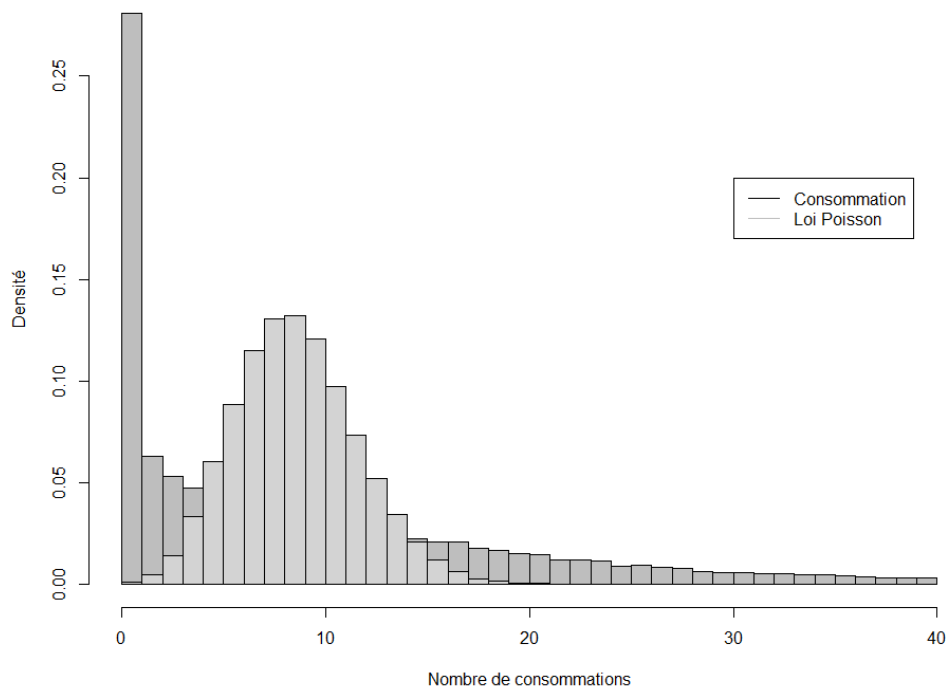


Figure 2.2.g : La loi de Poisson ne semble pas adéquate pour modéliser le nombre de consommations.

Il apparaît sur le graphique que la loi de Poisson n'est pas vraiment adaptée pour cette variable. De plus, la sur-représentation des valeurs nulles risque d'entraîner de la sur-dispersion dans les données, soit une variance empirique supérieure à la moyenne. En effet, l'espérance prend pour valeur $E[Y] = 9,06$ et la variance $Var[Y] = 90,89$ donc $E[Y] < Var[Y]$.

Deux nouveaux modèles ont alors été mis en place afin de pallier ce problème :

- **Un modèle « quasi-poisson »** : il est ici considéré que $V[Y] = \alpha * E[Y]$ et la variance n'est donc plus considérée égale à l'espérance mais est définie comme une fonction linéaire de l'espérance. Les coefficients obtenus par ce modèle sont les mêmes que pour une régression de Poisson mais l'erreur standard des paramètres obtenus est multipliée par $\sqrt{\alpha}$.
- **Un modèle « Zero Inflated Poisson »** : il est ici pris en compte le fait que la valeur nulle est sur-représentée. En effet, dans cette étude, un assuré peut avoir une consommation nulle car il ne consomme pas en règle générale ou parce qu'il n'a pas eu l'opportunité de consommer sur la période étudiée. Ce modèle permet alors de combiner une loi Binomiale (survenance ou non de $Y=0$) et une loi de Poisson afin de compter le nombre de consommations.

La prédiction des modèles a été confrontée aux valeurs observées pour l'échantillon de validation. Les RMSE et Q_2 des deux modèles précédents ont été calculés afin de pouvoir comparer



le taux d'erreur et la performance prédictive de ces deux modèles. Les résultats sont reportés dans le tableau suivant :

	Quasi-Poisson	Zero Inflated Poisson
RMSE	7,48	7,04
Q ₂	38,76%	45,75%

Tableau 2.2.1 : Comparaison des résultats des deux modèles.

Les résultats obtenus sont convenables, notamment pour le modèle ZIP. Cependant, la mauvaise adéquation de la loi de Poisson aux données incite à essayer d'améliorer ces prédictions à l'aide des autres méthodes de *datascience* présentées précédemment.

Arbre CART

La méthode des arbres CART a ensuite été utilisée afin d'obtenir un arbre de régression à l'aide des données d'apprentissage. Cela permet de ne pas avoir à modéliser la loi de probabilité suivie par les données, qui n'était pas optimale lors de l'utilisation du GLM.

L'arbre maximal obtenu ne peut pas être utilisé puisqu'il contient environ 40 000 feuilles. L'arbre optimal obtenu après élagage selon le paramètre de complexité est un arbre d'une centaine de feuilles. Un paramètre de complexité permettant de réduire ce nombre de feuilles et avoir ainsi un arbre interprétable est donc utilisé. L'arbre suivant sera retenu :

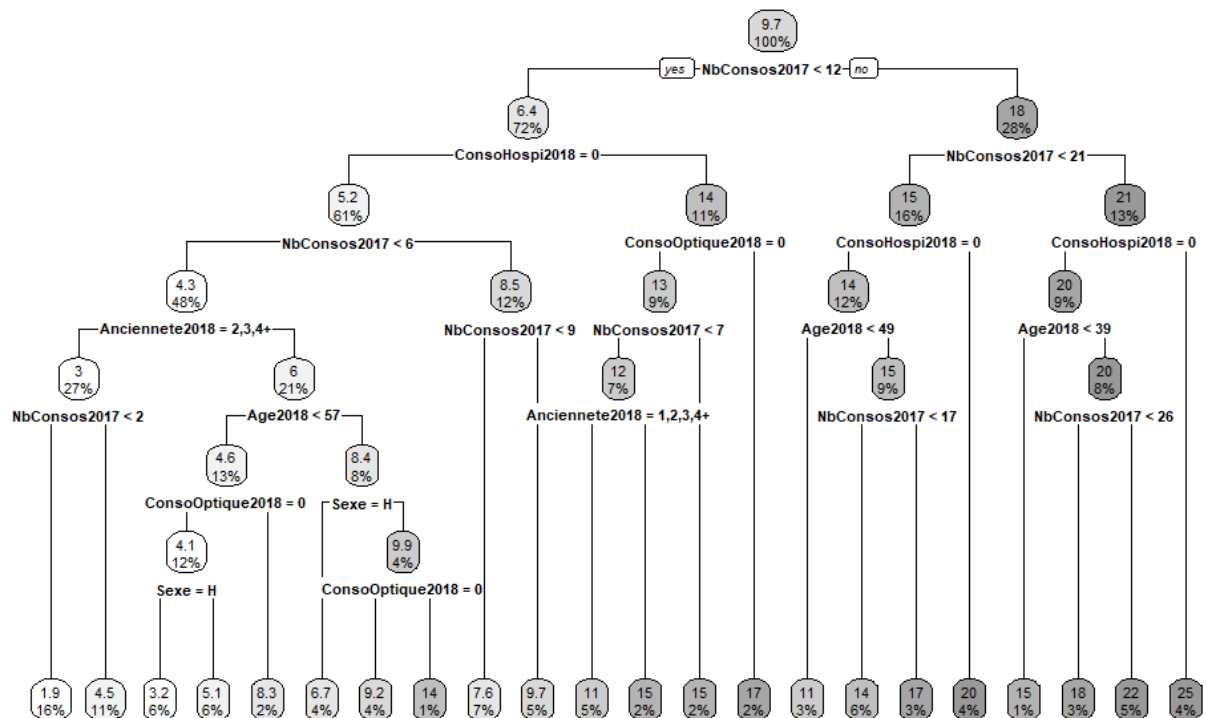


Figure 2.2.h : Le nombre de consommations de l'année précédente est la variable la plus influente sur le nombre de consommations.



Les variables discriminantes sur le nombre de consommations sont donc le nombre de consommations de l'année précédente, la consommation en hospitalisation et en optique, l'ancienneté et l'âge.

La prédiction du modèle a été comparée aux valeurs observées pour l'échantillon de validation. Le RMSE est alors de 6,96 et le coefficient prédictif de 46,94%.

Forêt aléatoire

Un modèle de forêt aléatoire a été implémenté sur les 12 variables présentées initialement.

Le paramètre du nombre d'arbres a été fixé à 50 afin de limiter le temps de calcul. Il a cependant été vérifié que le nombre d'arbres choisi était suffisant pour stabiliser le taux d'erreur. En faisant varier la valeur de $mtry$, on retient que la valeur de ce paramètre qui minimise le RMSE et maximise le Q_2 avec des prédictions des données de l'échantillon de validation est $mtry = 4$.

Le RMSE est alors, dans ce cas, de 6,77 et le Q_2 de 49,87%.

Ce modèle indique que le nombre de consommations en 2017, l'âge, la consommation en hospitalisation et la région sont les variables qui ont le plus d'impact sur la consommation de 2018.

Conclusion

Quatre modèles de *datascience* ont été utilisés sur le même échantillon de validation afin de prédire les consommations des assurés. Les critères RMSE et Q_2 des quatre modèles sont rappelés ci-dessous afin de sélectionner le plus performant.

	GLM Quasipoisson	GLM ZIP	Arbre CART	Forêt aléatoire
RMSE	7,48	7,04	6,96	6,77
Q_2	38,76%	45,75%	46,94%	49,87%

Tableau 2.2.m : Comparaison des résultats des différents modèles.

Le modèle de forêt aléatoire sera donc retenu pour modéliser le nombre de consommations de chaque assuré puisqu'il a le RMSE le plus petit et un coefficient prédictif de 49,87%.

Grâce au modèle de forêt aléatoire, le nombre de consommations de chaque assuré a pu être modélisé pour les années à venir. Les indicatrices de consommation en optique et hospitalisation ont également été prises en compte afin qu'un assuré n'ait pas un nombre de consommations total inférieur à la somme de ces deux indicatrices. En regroupant le nombre de consommations de chaque assuré par contrat, les classes de consommation nécessaires à la modélisation de la résiliation ont ainsi pu être obtenues.

c. Calcul du montant de consommation total

Les différentes modélisations décrites précédemment ont permis d'obtenir les informations suivantes pour chaque assuré :

- Le montant de consommation moyen en optique, noté Mtn_{opt}



- Le montant de consommation moyen en hospitalisation, noté Mtn_{hospi}
- Le coût moyen des autres prestations de santé, noté Mtn_{autres}
- Le nombre de prestations attendues dans l'année, noté $NbConsos_{tot}$
- Un indicateur de consommation en optique, noté Ind_{opt}
- Un indicateur de consommation en hospitalisation, noté Ind_{hospi}

Une estimation du montant de consommation de chaque assuré est alors obtenue de la façon suivante :

$$Mtn_{tot} = (NbConsos_{tot} - Ind_{opt} - Ind_{hospi}) * Mtn_{autres} + Ind_{opt} * Mtn_{opt} + Ind_{hospi} * Mtn_{hospi}$$

Le même calcul sera effectué pour chaque année en tenant compte des changements de caractéristiques des assurés d'une année à l'autre.

Le montant de consommation de chaque assuré pour chaque année a donc été modélisé.

2.2.3 Projection des montants de primes

Comme décrit précédemment, seuls les montants des primes 2018 et le taux d'indexation appliqué par l'assureur au 1^{er} janvier 2018 (2,5% uniformément) se trouvent à notre disposition.

Le montant de prime payé par l'assuré dépend de l'âge, de la gamme et du niveau de garantie et est donc calculé de la façon suivante pour l'année $n + 1$:

$$Prime_{n+1}(\text{âge}) = Prime_n(\text{âge}) * 1,025$$

De cette manière, les primes payées par chaque assuré pour les années à venir sont obtenues.

2.2.4 L'indicateur de valeur prospective

a. Implémentation de l'indicateur

Les modélisations des montants de consommation et le calcul des primes à percevoir pour les années à venir permettent de définir la marge prospective théorique de l'assuré i pour l'année n de la manière suivante :

$$MargeProspTheo_{i,n} = Prime_{i,n} - MtnTot_{i,n}$$

Afin d'obtenir la valeur prospective d'un assuré, il faut également rajouter une dimension de probabilité. Dans le cadre de cette étude, la valeur prospective sera donc obtenue grâce à la probabilité de résiliation modélisée précédemment. Ainsi, l'assuré i ayant une probabilité de résiliation p pour l'année n a une valeur prospective calculée de la manière suivante :

$$ValeurProspective_{i,n} = MargeProspTheo_{i,n} * (1 - p_{i,n})$$

La valeur prospective totale de l'assuré i pour les N années à venir est obtenue de la façon suivante :

$$ValeurProspectiveTot_i = \sum_{k=1}^N ValeurProspective_{i,k}$$



De la même manière que pour l'étude de la valeur historique, des classes ont été créées à l'aide des quantiles de valeur prospective afin de simplifier la comparaison de cet indicateur entre les différentes années ou entre les groupes d'assurés selon leur valeur prospective totale.

La valeur prospective du portefeuille global peut être obtenue en sommant les valeurs prospectives de tous les assurés, par année ou pour les N années à venir.

b. Analyse des résultats

Afin d'obtenir la valeur prospective de chaque année à venir, il convient de tenir compte des changements de la valeur de certaines variables d'année en année. Certaines d'entre elles varient de façon certaines (comme l'âge et l'ancienneté), d'autres sont considérées par hypothèse comme invariables dans le cadre de cette étude (comme la situation de famille, la gamme ou le niveau de garantie) et d'autres devront être modélisées (comme la consommation en hospitalisation ou optique de chaque année ainsi que le nombre de consommations de l'année précédente à partir de l'année 2020).

Une étude détaillée de la marge prospective théorique a été réalisée à partir de la base de données de 2018 (pour l'âge de l'assuré principal, l'ancienneté du contrat et les variables indicatrices de la consommation) afin d'identifier les assurés présentant un potentiel de valeur ou non pour l'assureur. Les critères étudiés sont les mêmes que lors de l'analyse effectuée sur les classes de valeur historique.

Etude de la répartition selon la gamme et le niveau de garantie

La répartition des assurés par gamme et niveau de garantie au sein de chaque classe de valeur prospective et du portefeuille global est présentée dans le tableau suivant :

Gamme	Niveau	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Gamme1	1	17%	12%	15%	13%	0%	7%
	2	19%	15%	15%	15%	8%	12%
	3+	7%	4%	4%	5%	6%	6%
Gamme2	1	17%	12%	1%	18%	11%	13%
	2	24%	19%	19%	20%	47%	30%
	3+	4%	5%	5%	7%	21%	7%
Gamme3	1	3%	14%	16%	5%	0%	8%
	2	6%	16%	19%	13%	4%	12%
	3+	2%	4%	5%	5%	3%	5%
Total		100%	100%	100%	100%	100%	100%

Tableau 2.4.a : Répartition des assurés par gamme et niveau de garantie au sein des différentes classes de valeur prospective.



Le tableau ci-dessus met en évidence le fait que les assurés ayant une valeur prospective mauvaise sont majoritairement des assurés de la gamme 1, niveau 1 ou 2 puisque 36% des assurés de cette classe de valeur ont souscrit à ces niveaux de garantie contre 19% des assurés du portefeuille global.

Les valeurs « faible » et « moyenne » contiennent une plus grande proportion d'assurés de la gamme 3, niveau 1 ou 2 que le portefeuille global.

Les assurés ayant une valeur prospective « forte » sont principalement assurés de la gamme 2, notamment du niveau 2 ou supérieur (environ 80% des individus de cette classe alors qu'ils ne représentent que 37% du portefeuille).

Etude des répartitions selon la situation de famille du contrat

La répartition des assurés par situation de famille au sein de chaque classe de valeur prospective et du portefeuille global est présentée dans le tableau ci-dessous :

Situation famille	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Isole	65%	55%	48%	66%	75%	68%
Duo	11%	10%	8%	11%	18%	10%
Couple+Enf	5%	12%	15%	8%	3%	7%
MonoParent	20%	23%	29%	14%	4%	15%
Total	100%	100%	100%	100%	100%	100%

Tableau 2.4.b : Répartition des assurés par situation de famille au sein des différentes classes de valeur prospective.

Une part plus importante de contrats avec enfants dans les classes de valeur « faible » ou « mauvaise » que dans le portefeuille global (35% des assurés ayant une valeur prospective faible contre 22% au total) est mise en avant par ce tableau.

A l'inverse, les assurés de la classe de valeur « forte » sont principalement assurés sans ayant droit ou sans enfant.

Etude des répartitions selon l'âge

La répartition des assurés par classe d'âges au sein de chaque classe de valeur prospective et du portefeuille global est présentée dans le tableau suivant :



Classe d'âges	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
< 20 ans	8%	21%	28%	9%	0%	11%
20-30 ans	9%	16%	18%	9%	1%	11%
30-40 ans	10%	13%	14%	12%	3%	12%
40-50 ans	16%	8%	9%	13%	8%	12%
50-60 ans	16%	11%	8%	19%	13%	14%
60-70 ans	25%	20%	14%	30%	37%	27%
> 70 ans	15%	13%	9%	9%	38%	13%
Total	100%	100%	100%	100%	100%	100%

Tableau 2.4.c : Répartition des assurés par tranche d'âges au sein des différentes classes de valeur prospective.

Les assurés de la classe de valeur « faible » ont pour la plupart moins de 30 ans (environ 37% alors qu'ils ne représentent que 22% du portefeuille).

La majorité de la classe de valeur « forte » est âgée de plus de 60 ans (75% des assurés ayant une valeur « forte » contre 40% du portefeuille). La classe de valeur « bonne » présente également une proportion non négligeable de 60-70 ans (30% des assurés de cette classe).

Conclusion

L'étude par critère a permis d'identifier que les **assurés de plus de 60 ans, de la gamme 2 et sans enfant** semblent particulièrement présents au sein des **classes de valeurs « bonne » et « forte »**. Ces 3 critères sont cependant corrélés puisque la gamme 2 est principalement souscrite par des personnes âgées et qui, de surcroît, n'ont pas d'enfant à charge. A contrario, **les moins de 30 ans assurés des niveaux 1 et 2 des gammes 1 et 3** semblent plutôt avoir une **valeur prospective « mauvaise » ou « faible »**.

Comme pour l'étude de la valeur historique, les moins de 30 ans ont fait l'objet d'une analyse plus ciblée afin de déterminer lesquels, parmi ces individus, peuvent tout de même présenter un potentiel de valeur pour l'assureur. Il apparaît alors que la classe de valeur prospective « forte » est répartie uniquement entre les niveaux 3 ou supérieurs alors qu'ils ne représentent que 16% des souscriptions pour les moins de 30 ans. La classe de valeur « bonne » comporte une majorité d'assurés de niveau 2 ou plus de la gamme 3. A contrario, le niveau 1 de la gamme 3 est particulièrement représenté pour les classes de valeurs « faible » et « moyenne ».

2.2.5 Limites de la modélisation

L'implémentation de l'indicateur de valeur prospective effectuée dans cette étude présente un certain nombre de limites qu'il est important de relever et d'expliquer afin d'avoir une idée claire et précise de la fiabilité des résultats obtenus.

Tout d'abord, les bases de données sur lesquelles l'étude est effectuée ne présentent pas un nombre suffisant de variables pouvant être utilisées afin d'expliquer les modèles. En effet, lors de l'implémentation de modèles de *datascience*, un grand nombre de variables explicatives est



nécessaire afin de prédire au mieux le résultat recherché. Dans cette étude, les modèles ne sont basés que sur une douzaine de variables explicatives ce qui est relativement peu. De plus, un certain nombre d'entre elles sont des variables devant être modélisées également, ce qui augmente l'erreur et l'incertitude des résultats de ces modèles.

De fortes hypothèses ont été retenues lors du traitement de base de données. Par exemple, la stabilité de la situation de famille des adhérents ou de la garantie des contrats ont été retenues comme hypothèses. En réalité, ce sont des variables étant généralement amenées à varier (avec la naissance d'un ayant droit ou avec le changement de situation matrimoniale des assurés pour la situation de famille, ou par la volonté d'un changement de couverture pour la gamme de garantie). D'autres hypothèses ont ensuite été considérées afin de mener à bien la modélisation et le calcul de la valeur prospective, telle que l'absence d'interaction entre les différents actes de santé. Une étude plus détaillée des bases de prestations aurait pu mener à considérer ce biais.

En pratique, l'assureur peut tenir compte de toute autre information qu'il juge pertinente à la mise en place de l'indicateur de Valeur Client. Dans cette étude, l'indicateur a été implémenté à partir des informations disponibles dans les bases de données mises à disposition par l'assureur et a principalement pour but de servir d'exemple afin d'illustrer l'utilisation possible et l'intérêt que présente la Valeur Client dans le cadre de l'assurance santé individuelle. En pratique, la robustesse et l'interprétabilité des résultats pourraient sans doute être améliorées en utilisant d'autres variables ou hypothèses. En particulier, certaines données sur les dimensions gestion ou marketing seraient pertinentes à introduire. Les incidents de paiements, le multi-équipement, des appels, courriers ou connexions sur le site assurés sont en effet autant de variables généralement connues pour leur explicativité dans les modèles de résiliation. Par ailleurs, l'absence de profondeur d'historique n'a pas permis de capturer les effets d'évolution dans le temps des caractéristiques et de la structure de la consommation de soins. Ces points pourraient permettre de disposer d'une vision complète lors de prochaines études.

2.2 Synthèse

L'utilisation de différentes méthodes de *datascience* a permis les modélisations de la probabilité de résiliation et de la consommation des assurés.

La résiliation des assurés pour les années à venir a été modélisée grâce à un modèle linéaire généralisé ainsi que différentes variables explicatives. Les seniors ayant une forte consommation sont les personnes les moins susceptibles de résilier. A contrario, les contrats qui consomment peu ont une forte probabilité de résiliation, notamment au bout d'un an d'ancienneté.

La consommation des assurés a été modélisée en utilisant un modèle de forêt aléatoire pour le nombre de consommations par an et en utilisant le coût moyen des prestations selon l'âge des assurés, leur sexe et leur garantie. De plus, des simulations ont été effectuées afin de modéliser, séparément, la consommation en optique et en hospitalisation, puisque ces postes sont apparus comme particulièrement discriminants lors de la modélisation de la probabilité de résiliation. Les



coûts moyens de ces postes par an selon l'âge, le sexe et la garantie des assurés ont été calculés de façon distincte. Le coût total de consommations par assuré a ainsi pu être modélisé.

Enfin, la marge prospective théorique de chaque assuré peut être obtenue à partir de la projection des primes à percevoir par l'assureur et du montant de consommations des assurés pour les années à venir. Les assurés de plus de 60 ans, de la gamme 2 et sans enfant semblent avoir une marge prospective théorique importante alors que les moins de 30 ans assurés des niveaux 1 et 2 des gammes 1 et 3 semblent plutôt avoir une marge prospective théorique « mauvaise » ou « faible ».

La somme des marges prospectives théoriques pondérées par la probabilité de présence pour chaque année donne alors la valeur prospective de chaque assuré.

Les limites de la modélisation et les axes d'amélioration possibles ont ensuite été présentés.



3. APPLICATION : LA VALEUR CLIENT COMME OUTIL DE DIAGNOSTIC ET D'AIDE A LA DECISION

Une fois la valeur historique calculée et la valeur prospective modélisée, différentes applications peuvent alors être proposées. Cette partie consiste à présenter l'intérêt de la création d'un tel indicateur au travers de différentes segmentations, dans un but de diagnostic du portefeuille puis d'aide à la décision pour l'assureur concernant les assurés à cibler et les actions à réaliser. Les apports directs de la mise en place de ces opérations sont ensuite présentés afin de conclure sur l'utilité de l'utilisation de la Valeur Client.

3.1 Segmentation du portefeuille grâce à la Valeur Client

Différentes analyses de la Valeur Client peuvent être effectuées. L'approche présentée dans cette étude consiste en une segmentation du portefeuille à l'aide des différents éléments calculés précédemment. Selon les critères de segmentation étudiés et les informations dont dispose l'assureur, le portefeuille peut être découpé en un nombre plus ou moins important de segments.

3.1.1 Comparaison entre Valeur Client historique et marge prospective théorique

Pour commencer, il peut être intéressant de croiser la Valeur Client historique à la marge prospective théorique de chaque assuré. On ne s'intéresse ici qu'à l'évolution de la valeur de chaque assuré, sans tenir compte de leur probabilité de résiliation. Cette segmentation peut servir à définir la stratégie de fidélisation de l'assureur (pour les segments à forte valeur) ou de changement de tarification (pour les segments à faible valeur).

La marge prospective théorique de chaque assuré a été normalisée de la même manière que pour la valeur historique et des classes de valeur (mauvaise, faible, moyenne, bonne et forte) ont été définies à partir des quantiles de marge prospective théorique normalisée pour les quatre années.

Une segmentation du portefeuille peut donc être effectuée selon le changement de classe de valeur de chaque assuré. En particulier, cinq groupes différents d'assurés ont été identifiés :

- Les assurés ayant une **valeur historique et prospective « faible » ou « mauvaise »** pour chaque année : cette cible ne présente pas de potentiel de valeur pour l'assureur et n'en présentera pas dans les années à venir. Il peut alors mettre en place une stratégie visant à limiter leur coût mais n'a pas d'intérêt a priori à mener des actions commerciales visant à les fidéliser.
- Les assurés ayant une **valeur historique « faible » ou « mauvaise » mais une marge prospective théorique « forte » ou « bonne »** pour chaque année : ces assurés sont, à l'heure actuelle, coûteux mais seront, dans le futur, profitables pour l'assureur. Il doit donc miser sur leur fidélisation et identifier les éventuelles opportunités qu'ils présentent afin



que les coûts importants engendrés par le passé aient une chance d'être amortis par la suite.

- Les assurés ayant une **valeur historique « forte » ou « bonne » mais une marge prospective théorique « faible » ou « mauvaise »** pour chaque année : ces assurés étaient profitables par le passé mais vont potentiellement présenter des coûts importants dans les années futures. L'assureur doit donc élaborer une stratégie afin de limiter leur coût futur mais peut également tenter d'identifier le facteur explicatif de ce changement de valeur. Si un tel facteur dépend directement de ses services, il peut alors tenter de résoudre ce dysfonctionnement et ainsi changer les prévisions de rentabilité de ces assurés.
- Les assurés ayant une **valeur historique et prospective « forte » ou « bonne »** pour chaque année : ces assurés sont à l'heure actuelle intéressants pour la rentabilité du portefeuille et le seront aussi par la suite. La fidélisation et la conquête de ces cibles représentent donc un enjeu important pour l'assureur qui doit s'investir dans cette relation.
- Les assurés ayant une **valeur historique et prospective « moyenne »** : ces assurés ne sont pas fortement rentables à un moment précis de leur relation client mais le sont de manière plus modérée et constante. L'assureur peut tenter d'améliorer leur valeur après identification des opportunités qu'ils présentent mais n'a globalement pas grand intérêt à consacrer beaucoup de ressources sur ce segment.

Les efforts fournis par l'assureur pour chaque catégorie d'assurés doivent également dépendre de la proportion qu'ils représentent au sein de son portefeuille. Il n'est en effet pas forcément pertinent pour l'assureur de concentrer des ressources (en temps, en budget, en employés...) sur un groupe d'assurés si celui-ci ne présente qu'une proportion minimale du portefeuille.

Une étude détaillée des caractéristiques de chaque groupe d'assurés a été réalisée en utilisant l'année 2018 comme année de référence (pour l'âge et l'ancienneté) puisque les valeurs historiques de cette année sont comparées avec les résultats prospectifs des autres années.

Les résultats sont détaillés en [annexe](#) et sont reportés ci-dessous sous forme d'un graphique :

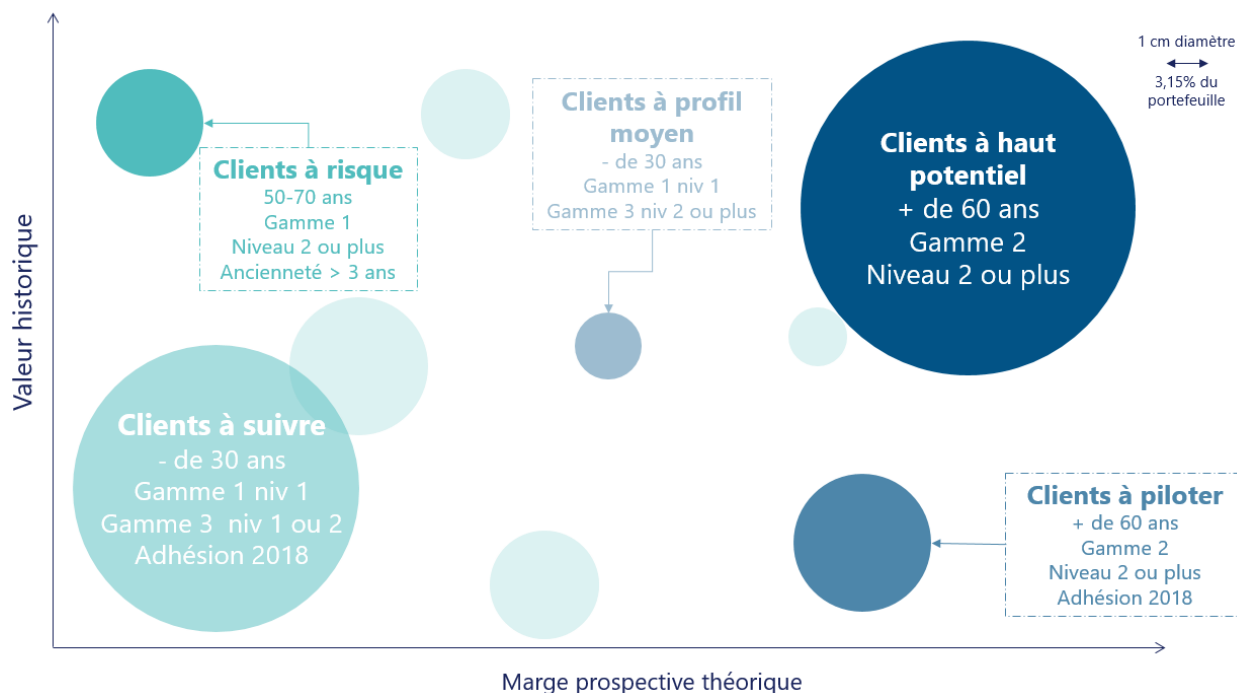


Figure 3.1.a : Les moins de 30 ans assurés du niveau 1 des gamme 1 et 3 sont à suivre, les assurés des niveaux 2 ou plus de la gamme 2 sont à haut potentiel.

À la suite de cette analyse, deux objectifs principaux sont mis en avant pour l'assureur.

Il doit d'abord tenter **d'améliorer la marge des moins de 30 ans**. Pour ce faire, il peut envisager une augmentation de prime afin d'amortir les consommations de ces assurés. Ces profils sont également proches de ceux à profil moyen : il peut tenter de promouvoir les niveaux 2 ou supérieurs de la gamme 3 afin d'inciter à une montée en gamme de la part de ses assurés et ainsi améliorer leur marge prospective théorique.

Il doit également s'assurer de **conserver ses assurés de la gamme 2 âgés de plus de 60 ans** puisque ces clients à haut potentiel représentent un quart de son portefeuille et que les clients à piloter ont des profils similaires, à l'exception de l'ancienneté. Leur fidélisation est donc un objectif important pour la rentabilité future du portefeuille.

L'assureur peut également essayer d'investiguer et comprendre la perte de valeur de ses clients à risque même si la faible proportion de ce segment ne justifie pas nécessairement un effort important de sa part.

3.1.2 Comparaison entre probabilité de résiliation et marge prospective théorique

Afin de déterminer un plan de pilotage stratégique et identifier les segments présentant un enjeu financier important sur le long terme, l'assureur peut segmenter son portefeuille en fonction de la probabilité de résiliation et de la marge prospective théorique de chaque assuré. **Cette segmentation peut servir à définir la stratégie de fidélisation de l'assureur** (pour les segments



à forte valeur), **de changement de tarification** (pour les segments à faible probabilité de résiliation) et à **identifier ses cibles de conquête**.

La segmentation peut s'effectuer sur chaque année dont la marge prospective théorique et la probabilité de résiliation ont été calculées afin d'obtenir une analyse plus précise et une stratégie année par année pour l'assureur. Dans le cadre de cette étude, une seule segmentation sera effectuée à l'aide de la moyenne des probabilités de résiliation et des différentes classes de la marge prospective théorique totale (dont les bornes ont été définies dans la segmentation précédente) afin d'avoir une vision globale sur le long terme. En particulier, neuf groupes différents d'assurés ont été identifiés :

- Les assurés ayant une **marge prospective théorique « faible » ou « mauvaise » et une probabilité de résiliation faible** (inférieure à 0,15) : l'assureur doit minimiser cette proportion d'assurés au sein de son portefeuille. Il a pour objectif d'améliorer leur marge prospective théorique puisque la probabilité que ces assurés soient fidèles est importante. Il est en effet, à première vue, plus intéressant d'un point de vue financier de risquer de perdre des assurés à faible valeur en essayant d'améliorer leur marge prospective théorique plutôt que d'être certain de conserver des clients coûteux pour lui.
- Les assurés ayant une **marge prospective théorique « faible » ou « mauvaise » et une probabilité de résiliation moyenne** (entre 0,15 et 0,35) : l'assureur doit se comporter avec ces assurés de la même manière que pour le segment précédent mais dans une priorité moindre puisqu'il y a une plus grande probabilité que ces clients quittent le portefeuille d'eux même.
- Les assurés ayant une **marge prospective théorique « faible » ou « mauvaise » et une probabilité de résiliation importante** (supérieure à 0,35) : ces assurés ne présentent pas un grand intérêt d'un point de vue financier pour l'assureur. Il peut cependant essayer de comprendre quelle est la cause de leur résiliation. S'il le peut, il doit alors travailler sur cette cause dans un but d'amélioration de ses services.
- Les assurés ayant une **marge prospective théorique « forte » ou « bonne » et une probabilité de résiliation faible** (inférieure à 0,15) : ces assurés sont à très fort potentiel de valeur pour l'assureur. Il doit donc s'assurer de leur satisfaction et peut également tenter de déterminer quels sont les facteurs garantissant leur fidélisation afin de le développer et améliorer celle des autres segments.
- Les assurés ayant une **marge prospective théorique « forte » ou « bonne » et une probabilité de résiliation moyenne** (entre 0,15 et 0,35) : l'assureur doit déterminer quels sont les facteurs influençant leur probabilité de résiliation et, si possible, travailler en ce sens afin d'améliorer leur satisfaction et donc leur fidélisation. Ces assurés présentent un potentiel de rentabilité mais ne sont cependant pas une priorité pour l'assureur puisqu'il y a tout de même une chance non négligeable qu'ils ne résilient pas.
- Les assurés ayant une **marge prospective théorique « forte » ou « bonne » et une probabilité de résiliation importante** (supérieure à 0,35) : comme pour le segment



précédent, l'assureur doit améliorer la fidélisation de ce segment. A l'inverse, il constitue une priorité pour lui puisqu'il y a de fortes chances que ces assurés quittent le portefeuille dans les années à venir.

- Les assurés ayant une **marge prospective théorique « moyenne » et une probabilité de résiliation faible** (inférieure à 0,15) : ces assurés ont une marge dans la moyenne du portefeuille et sont donc profitables pour l'assureur. L'assureur ne doit pas consacrer beaucoup de ressources sur ce segment mais doit continuer de s'assurer de leur satisfaction puisqu'ils sont tout de même profitables.
- Les assurés ayant une **marge prospective théorique « moyenne » et une probabilité de résiliation moyenne** (entre 0,15 et 0,35) : l'assureur doit améliorer leur fidélisation. Cependant, la marge de chaque assuré étant relativement faible, l'utilisation d'un nombre important de ressources sur ce segment risquerait d'être plus coûteux que l'apport potentiel de valeur des assurés. De plus, leur marge prospective théorique est moins intéressante que celle d'autres segments, ce qui n'en constitue pas une priorité pour l'assureur. Il doit donc trouver un équilibre entre amélioration de leur apport potentiel et de leur fidélisation tout en limitant les moyens déployés pour ces assurés.
- Les assurés ayant une **marge prospective théorique « moyenne » et une probabilité de résiliation forte** (supérieure à 0,35) : ces assurés ont une marge dans la moyenne du portefeuille mais vont certainement quitter le portefeuille. L'assureur peut tenter d'améliorer la fidélisation de ces assurés mais ce segment ne présente qu'un intérêt limité pour lui.

Une étude détaillée des caractéristiques de chaque groupe d'assurés a été réalisée en utilisant l'année 2019 comme année de référence (pour l'âge et l'ancienneté) puisque cette segmentation se base sur les résultats prospectifs. Cette étude ne tient donc pas compte des futurs assurés (nouveaux adhérents ou nouveaux nés affiliés comme ayants droit) mais seulement de ceux déjà présents au sein du portefeuille en 2018. De plus, la probabilité de résiliation étant calculée par contrat et la valeur prospective par assuré, la probabilité de résiliation attribuée à chaque assuré correspond à celle de son contrat. Ainsi, tous les bénéficiaires d'un même contrat auront la même probabilité de résiliation.

Les résultats de la segmentation ainsi effectuée sont détaillés en [annexe](#) et seul un résumé est ici rapporté sous forme de graphique :

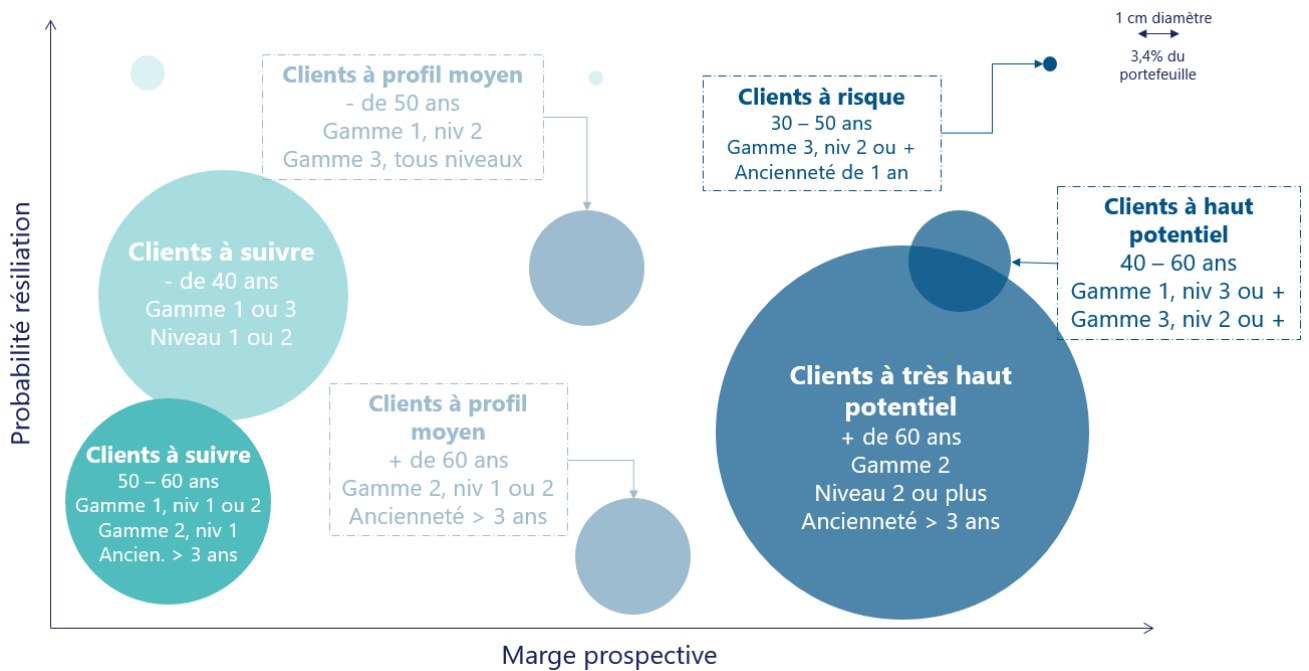


Figure 3.1.b : Les assurés du niveau 1 ou 2 de la gamme 1 sont à suivre, les plus de 60 ans assurés de la gamme 2 sont à très haut potentiel.

À la suite de cette analyse, plusieurs objectifs sont mis en avant pour l'assureur.

Il doit d'abord tenter **d'améliorer la marge des niveaux inférieurs de la gamme 1**. Pour ce faire, il peut envisager une augmentation de prime afin d'amortir les consommations de ces assurés. Ces segments sont également proches de ceux à profil moyen : il peut tenter de promouvoir la gamme 2 pour les cinquantenaires ou les niveaux supérieurs de la gamme 3 afin d'inciter à leur montée en niveau de gamme et ainsi améliorer leur marge prospective théorique.

Il doit également **s'assurer de conserver ses assurés de la gamme 2 âgés de plus de 60 ans** puisque ces clients à très haut potentiel représentent un tiers de son portefeuille. De plus, il doit travailler sur une montée en niveau de gamme des soixanténaires assurés au niveau 1 afin de les faire passer de profil moyen à très haut potentiel.

Enfin, il apparaît clairement que **la probabilité de résiliation est importante pour les assurés ayant souscrits depuis un an ou les assurés de la gamme 3**. Un problème existe donc avec la gamme 3 et l'assureur doit tenter d'améliorer la satisfaction de ces assurés. L'insatisfaction ou un comportement plus volatil est également notable lors de la première année de souscription et l'assureur doit essayer de remédier à cela puisque les coûts d'acquisition des clients ne seront pas amortis sur la durée. Les clients à risque ne représentant qu'une partie minimale de son portefeuille, il est alors plus judicieux pour l'assureur de se concentrer en priorité sur les profils à probabilité de résiliation moyenne.



Les cibles de conquête identifiées sont donc en premier lieu les seniors de plus de 60 ans pour la gamme 2 et, en second lieu, les 40-60 ans pour les gammes 1 et 3, à partir du niveau 2 de garantie, s'il arrive à améliorer leur fidélisation.

3.1.3 Comparaison entre probabilité de résiliation et valeur prospective

Une dernière segmentation entre la probabilité de résiliation et la valeur prospective est détaillée ci-dessous.

Même si semblable à la segmentation précédente, il a tout de même été jugé intéressant d'effectuer cette comparaison afin d'observer l'influence de la probabilité de résiliation sur la valeur prospective des clients. En effet, un assuré ayant une marge prospective théorique forte et une probabilité de résiliation forte, pourrait en réalité avoir une valeur prospective moyenne tandis qu'un assuré ayant une marge prospective théorique moyenne mais une probabilité de résiliation faible, pourrait avoir une valeur prospective forte. **L'assureur pourra ainsi définir une stratégie de ciblage et de fidélisation sur le long terme en s'appuyant sur toutes les informations modélisées au sein de l'étude.**

Les mêmes groupes d'assurés que lors de la segmentation précédente ont été créés et la même analyse a été effectuée. Les résultats sont détaillés en [annexe](#) et seul un résumé est ici rapporté sous forme de graphique :

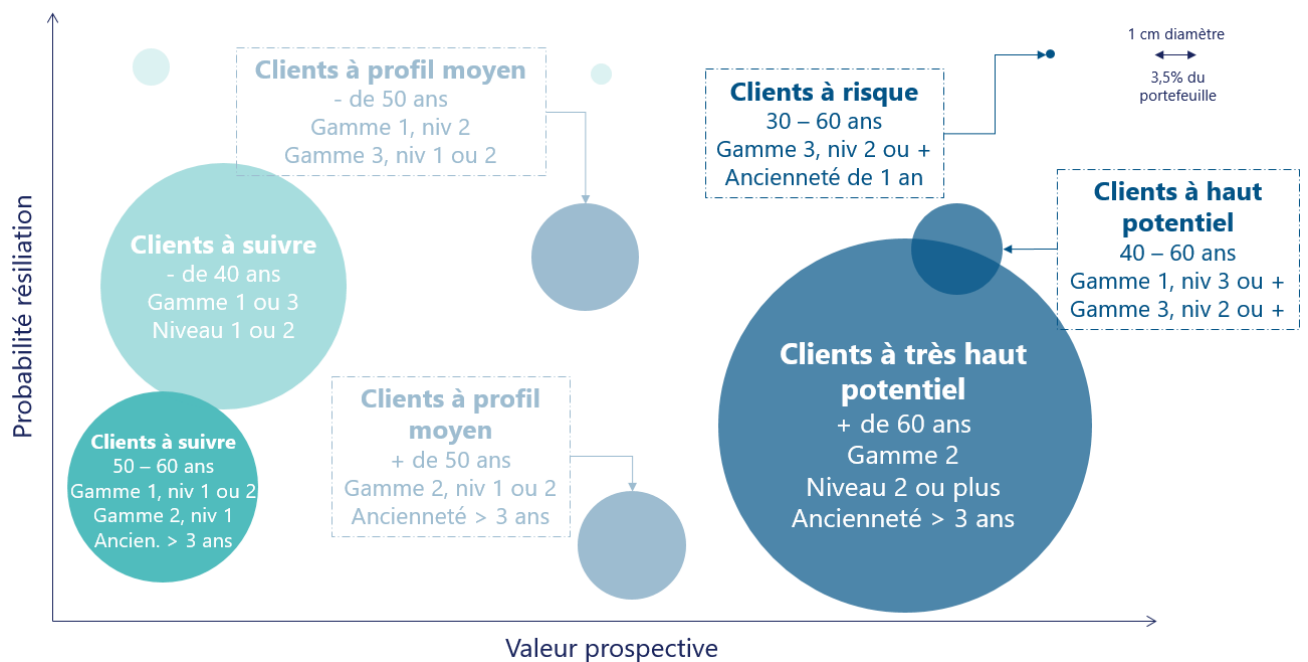


Figure 3.1.c : Les assurés du niveau 1 ou 2 de la gamme 1 sont à suivre, les plus de 60 ans assurés de la gamme 2 sont à très haut potentiel.

Les différents groupes obtenus sont très similaires à ceux de la segmentation précédente, à l'exception de changement dans la proportion des segments, de l'inclusion des 50-60 ans assurés de la gamme 3 depuis 1 an dans les profils à risque et de l'exclusion des moins de 50 ans assurés des niveaux supérieurs de la gamme 3 des profils moyens.



Globalement, cette segmentation ne permet donc pas de définir de nouveaux objectifs que ceux déjà connus par l'assureur à l'aide de la comparaison précédente.

3.1.4 Conclusion et limites de la segmentation dans le cadre de l'étude

La segmentation du portefeuille à partir des valeurs historique et prospective a permis d'identifier deux objectifs pour l'assureur : il doit s'assurer de conserver ses assurés de la gamme 2, âgés de plus de 60 ans et doit tenter d'améliorer la Valeur Client des moins de 30 ans assurés des gammes 1 et 3.

La segmentation du portefeuille selon la probabilité de résiliation et la marge prospective théorique a confirmé **les plus de 60 ans assurés de la gamme 2 comme cible de conquête** puisque leur valeur prospective est importante et que leur probabilité de résiliation diminue avec l'ancienneté. Le portefeuille contient déjà une grande proportion de ce segment.

Il est également notable que, parmi **les moins de 30 ans**, identifiés comme segment prioritaire dans la première segmentation, l'assureur doit particulièrement s'intéresser à l'amélioration de la valeur prospective **des assurés de la gamme 1** puisque leur probabilité de résiliation est faible (ceux de la gamme 3 ont une plus forte probabilité de résiliation).

Enfin, **les 30-60 ans assurés des niveaux 2 ou supérieurs de la gamme 3** présentent une marge prospective théorique importante mais également une forte probabilité de résiliation. L'assureur doit donc concentrer ses efforts de fidélisation sur ce segment.

L'interprétation des résultats de la mise en application de l'indicateur implémenté dans cette étude présente cependant quelques limites. En effet, il est difficile pour l'assureur d'agir directement sur les variables identifiées comme facteurs de segmentation (âge, ancienneté...). Des segments cibles de portefeuille sont ici déterminés mais, au vu des variables disponibles, il est complexe d'imaginer des changements de stratégies directs autres qu'une modification de la tarification des contrats. En pratique, l'assureur pourra se servir de données supplémentaires et ainsi identifier d'autres facteurs de segmentation. Par exemple, l'utilisation des données marketing peut permettre d'identifier l'impact de certaines offres commerciales sur la Valeur Client, les données d'autres contrats d'assurance permettent de tenir compte du multi-équipement...

3.1 Synthèse

Trois segmentations du portefeuille ont été effectuées à l'aide des éléments modélisés précédemment.

La segmentation du portefeuille à partir de la valeur historique et de la marge prospective théorique permet d'identifier si les assurés profitables dans le passé risquent de perdre de la valeur dans le futur ou si une partie du portefeuille présentant une faible valeur à l'acquisition peut s'améliorer avec le temps. Deux objectifs sont ressortis de cette segmentation : l'assureur doit conserver ses assurés de la gamme 2 et doit améliorer la marge prospective théorique des moins de 30 ans assurés des gammes 1 et 3.



La segmentation du portefeuille à partir de la marge prospective théorique et de la probabilité de résiliation permet à l'assureur de définir sa stratégie de fidélisation et de tarification sur les années à venir, ainsi que d'identifier des cibles de conquêtes. La segmentation met en avant le fait que l'assureur doit améliorer la marge des niveaux inférieurs de la gamme 1 et améliorer la fidélisation des assurés de la gamme 3 ainsi que celle des nouveaux adhérents. Les assurés de la gamme 2 âgés de plus de 60 ans sont clairement identifiés comme cible de clients à haut potentiel.

La segmentation du portefeuille à partir de la valeur prospective et de la probabilité de résiliation ne donne globalement pas beaucoup plus d'informations que la segmentation précédente.

Les limites de ces segmentations dans le cadre de l'étude ont ensuite été présentées.

3.2 Impact de différents scénarios sur la Valeur Client du portefeuille

Grâce aux segmentations présentées précédemment, plusieurs axes stratégiques ont été mis en avant pour l'assureur. Cette partie consistera en la simulation de l'application des différentes stratégies et de leur impact sur le portefeuille global. L'intérêt de l'indicateur de Valeur Client sera alors présenté de manière concrète.

3.2.1 Calcul initial

Dans un but de modification de son portefeuille, un assureur dispose de différents axes sur lesquels s'appuyer. Il peut en effet travailler sur :

- Le **nombre d'assurés** contenus dans son portefeuille en incitant à la souscription de nouveaux adhérents ou à la résiliation de certains autres,
- La **somme totale des primes perçues** grâce à sa politique de tarification,
- La **valeur client prospective** de son portefeuille, grâce à sa politique de tarification, sa stratégie de fidélisation-rétention, sa stratégie de conquête, le multi-équipement ou la montée en gamme de ces clients et de tout autre stratégie identifiée lors de l'implémentation de cet indicateur.

Afin d'étudier l'impact des différentes stratégies mises en exergue dans la partie précédente, il peut être intéressant d'observer l'évolution des 3 axes décrits ci-dessus.

Au vu des hypothèses retenues dans le cadre de cette étude, le nombre d'assurés peut sembler difficile à mesurer puisque seuls les assurés présents à la fin de l'année de référence sont considérés, sans tenir compte des informations sur la souscription de nouveaux assurés.

La somme totale des primes perçues et la valeur client prospective totale du portefeuille vont donc être calculées. Il est important de noter que le premier axe ne représente que la valeur monétaire du client alors que le deuxième tient également compte, dans cette étude, de sa probabilité de résiliation.



La somme totale des primes perçues a été calculée en sommant les primes perçues de chaque assuré du portefeuille, sans tenir compte de sa probabilité de résiliation. La valeur prospective de chaque assuré a été calculée à l'aide des primes projetées, de sa probabilité de résiliation et de son montant de consommation. La somme des valeurs prospectives de chaque assuré a ensuite été calculée afin d'obtenir la valeur prospective originale du portefeuille.

3.2.2 Comparaison de deux scénarios

La somme des primes et de valeur prospective de deux scénarios différents vont être comparées :

- **Le premier scénario consiste en une augmentation des primes sur l'ensemble du portefeuille** sans tenir compte des différents segments identifiés précédemment. Le taux d'augmentation des primes considéré a été fixé précédemment par hypothèse à 2,5% uniformément pour chaque année. **La probabilité de résiliation des assurés n'est pas modifiée** et celle obtenue par modélisation a été conservée.
- **Le deuxième scénario consiste en l'application de différentes stratégies selon les segments identifiés précédemment :**
 - o **Les « clients à suivre »** (soient les moins de 60 ans, assurés des niveaux 1 et 2 de la gamme 1) subiront une **augmentation de prime de 10% et on s'attend à observer 25% de résiliations en plus que dans le scénario central sur ce sous-segment** : économiquement parlant, il est pertinent pour l'assureur d'améliorer la marge dégagée par ces assurés et leur fidélisation peut ne pas être perçue comme prioritaire.
 - o **Les « assurés à profils moyens »** (soient les moins de 50 ans, assurés des niveaux 1 et 2 de la gamme 3) subiront une **hausse de 5% des tarifs, ce qui induira un accroissement des résiliations supposé de 20%** comparativement au scénario central. De la même manière que pour le segment précédent, l'assureur peut tenter d'améliorer la marge de ces assurés mais de façon plus limitée afin de réduire la résiliation de ces assurés qui sont déjà profitables.
 - o Le bénéfice dégagé par la hausse des primes de ces segments permettra de disposer d'une enveloppe pour proposer **des réductions tarifaires de 10% aux « clients à risque »** (soient les 30-60 ans assurés des niveaux 3 ou supérieurs de la gamme 3) **ainsi qu'au moins de 30 ans assurés des niveaux 3 ou supérieurs**¹². Le bénéfice escompté de cette action en termes de fidélisation est une **diminution des résiliations de 25% par rapport au scénario central sur ces segments**.
 - o Les primes des autres assurés seront augmentées de la même manière que dans le premier scénario et leur probabilité de résiliation ne sera pas modifiée.

¹²Pour rappel, une étude complémentaire a été menée sur cette classe d'âge afin d'identifier quels assurés de ce segment à faible valeur prospective présentent tout de même un potentiel de valeur pour l'assureur.



Ces pourcentages sont appliqués à chaque année de 2019 à 2022. Une comparaison par année puis au global sera effectuée entre les deux scénarios concernant la somme totale des primes perçues et la Valeur Prospective totale du portefeuille.

3.2.3 Présentation et comparaison des résultats

Au global, 44% des assurés du portefeuille sont concernés par les changements de prime et de probabilité de résiliation.

Pour chaque année, une augmentation du montant total de prime apparaît mais il est important de noter que l'augmentation de la valeur prospective est encore plus importante. Le tableau ci-dessous présente la différence des primes perçues et de la valeur prospective totale, entre le scénario central et le second scénario, par année puis en vision globale sur les quatre années.

	Primes perçues	Valeur prospective
Pour l'année 2019	+1,74%	+4,94%
Pour l'année 2020	+1,72%	+4,95%
Pour l'année 2021	+1,68%	+4,66%
Pour l'année 2022	+1,64%	+4,97%
Au total	+1,70%	+4,88%

Tableau 3.2.a : Comparaison des primes perçues et de la valeur prospective du portefeuille à l'issue du deuxième scénario par rapport au scénario initial.

La mise en place du deuxième scénario entraîne une augmentation des primes perçues de 1,70% sur les 4 années par rapport à l'augmentation des primes de manière uniforme. La valeur prospective du portefeuille a, quant à elle, augmenté de 4,88%.

La segmentation du portefeuille selon la valeur prospective et la probabilité de résiliation a été effectuée de nouveau sur le portefeuille modifié par le scénario précédent. Il est alors possible de comparer la proportion et la composition des différents segments par rapport au portefeuille initial.

La composition des différents segments n'a pas changé considérablement et les groupes d'assurés ont globalement gardé les mêmes caractéristiques. Cependant, les différences de proportion de chaque segment sont notables comme le montre le tableau ci-contre :



	Portefeuille scénario 1	Portefeuille scénario 2
« Clients à suivre »	36,8%	33,7%
« Assurés à profil moyen »	18,2%	19,5%
« Clients à risque »	0,8%	0,5%
« Clients à haut potentiel »	7,7%	10,6%
« Clients à très haut potentiel »	31,5%	30,2%
Autres segments	5,0%	5,7%

Tableau 3.2.b : Répartition des effectifs par segment pour chaque scénario

Le deuxième scénario a réduit la proportion des « clients à suivre » (ie ceux ayant une valeur prospective faible et une probabilité de résiliation faible ou moyenne) tandis que la proportion d'autres segments (entre autres les assurés à faible valeur prospective et forte probabilité de résiliation) a bien augmenté.

La part de « clients à risque » a diminué au profit des « clients à haut potentiel » et la proportion de « clients à très haut potentiel » n'a que très peu diminué. Cela signifie que la réduction proposée aux assurés de ce segment a légèrement réduit la marge théorique mais que l'objectif de fidélisation est bien atteint ici.

Différentes actions ont été appliquées à des groupes d'assurés variés, identifiés dans le respect de la législation (pas de différenciation par le sexe, ni par l'historique de consommation en santé) et ce, grâce à l'indicateur de Valeur Client implémenté dans cette étude. Ces scénarios, bien qu'hypothétiques, permettent de mettre en avant l'intérêt de la Valeur Client pour l'assureur en augmentant la valeur prospective du portefeuille. Les facteurs à modifier sont limités dans cette étude mais l'assureur pourra en réalité s'appuyer sur d'autres valeurs afin d'atteindre les objectifs qu'il aura identifié par suite de la segmentation de son portefeuille.

3.2 Synthèse

Plusieurs actions ont été appliquées sur différents segments du portefeuille, identifiés à l'aide des segmentations réalisées précédemment.

Les primes des assurés à faible ou moyenne valeur prospective ont subi une hausse tarifaire au dépend d'une augmentation de leur probabilité de résiliation. A l'inverse, une réduction tarifaire a été appliquée pour les assurés à forte valeur dans un but de rétention de ces personnes au sein du portefeuille.

Les impacts de ces différentes actions sur la valeur prospective et le montant de primes total perçu par l'assureur ont ensuite été analysés et comparés à la valeur prospective initiale.



Cette partie applicative sert à prouver de manière concrète l'intérêt de l'implémentation de l'indicateur de Valeur Client en assurance santé individuelle. L'assureur peut en effet augmenter la valeur prospective de son portefeuille tout en améliorant la satisfaction client (ici, par une réduction tarifaire), et ce grâce à des stratégies identifiées à la suite du diagnostic de son portefeuille à l'aide de l'indicateur de Valeur Client.



CONCLUSION

L'apport de la Valeur Client dans la compréhension des portefeuilles santé individuelle a été démontré dans cette étude.

Après avoir présenté le secteur de l'assurance santé, ses spécificités et enjeux dans le contexte actuel, la notion de Valeur Client a été introduite. Cet indicateur, dont la définition et l'implémentation dépendent de l'objectif recherché et des moyens disponibles à sa création, mêle à la fois la rentabilité monétaire des assurés, leur satisfaction et la longévité de leur relation avec l'assureur. Utilisé dans de nombreux domaines et notamment en assurance IARD depuis quelques années, son implémentation n'est encore, pour l'instant, que très peu commune en assurance de personnes.

Une première analyse du portefeuille a été effectuée en calculant la valeur historique de chaque assuré. Une photographie du portefeuille pour l'année de référence a ainsi été présentée. La suite de l'étude a consisté en l'utilisation d'outils de *datascience* afin de modéliser le comportement futur des assurés, concernant leur probabilité de résiliation et leur consommation en santé. La valeur prospective de chaque assuré a ainsi été calculée sur plusieurs années.

Une fois l'indicateur de Valeur Client implémenté, son intérêt comme outil de diagnostic et d'aide à la décision a été présenté. Pour ce faire, différentes segmentations du portefeuille ont été réalisées afin d'identifier les groupes d'assurés rentables ou non au cours du temps et d'observer les segments cibles, à risque et à potentiel du portefeuille. Certaines stratégies, déterminées à l'aide de ces segmentations, ont enfin été appliquées au portefeuille afin de montrer de manière concrète les apports de l'utilisation d'un indicateur de Valeur Client

L'aspect monétaire n'est pas le seul point sur lequel l'assureur peut tenter d'influer afin d'améliorer la valeur globale de son portefeuille et l'indicateur de Valeur Client permet de mettre en avant ces autres axes. L'assureur peut augmenter la valeur de son portefeuille tout en améliorant la satisfaction client, et ce grâce à des stratégies identifiées à la suite du diagnostic de ce dernier à l'aide de l'indicateur de Valeur Client.

Cette étude présente tout de même quelques limites, notamment sur les variables caractérisant les assurés qui ont pu être testées, ainsi que sur la qualité des modèles. Si la démarche et l'intérêt de cette approche ont bien pu être validées, il pourrait être intéressant pour aller plus loin d'enrichir les modélisations par d'autres données (données marketing, données de gestion, données économiques...) afin d'identifier d'autres axes explicatifs. L'indicateur de Valeur Client n'en serait que plus segmentant et informatif. Ainsi, avec une vision encore plus précise du portefeuille, les stratégies de ciblage de certains groupes d'assurés pourraient être plus pertinentes et, a fortiori efficaces, en termes d'amélioration de la rentabilité, d'optimisation des services proposés par l'assureur et plus globalement de la satisfaction client.



BIBLIOGRAPHIE

- [1] D. Le Nail et E. Poyet, «La créativité de l'assurance à l'épreuve du Covid-19,» *Blog addactis*, 2020.
- [2] A. Dillard, *Est-il encore pertinent pour les assureurs d'investir dans la valeur client ?*, Thèse de doctorat, ENASS, 2014.
- [3] A. Barral et E. Poyet, «Valeur client : cibler les segments pour maximiser la rentabilité des assureurs en santé individuelle,» *La tribune de l'assurance*, 2020.
- [4] A. Severin, «Valeur Client : radiographier ses clients pour jauger leur potentiel,» *www.argusdelassurance.com*, 2004.
- [5] J. Lendrévie et J. Lévy, *Mercator, Théories et nouvelles pratiques du marketing*, 2013.
- [6] V. Kumar, *Customer Lifetime Value - The path to Profitability*, 2007.
- [7] B. Donker, P. Verhoef et M. de Jong, «Modeling CLV: A test of competing models in the insurance industry,» 2006.
- [8] D. Hennom, *Création d'un indicateur de valeur client en assurance non-vie*, Mémoire IA, 2016.
- [9] S. Couaillac, *Création d'un indicateur de Valeur Client en Assurance Non Vie*, Mémoire IA, 2015.
- [10] M. Piffault, *La valeur client comme mesure de la rentabilité d'un portefeuille d'assurance santé individuelle*, Mémoire IA, 2015.
- [11] E. Masiello, *Cours Modèles linéaires généralisés*, ISFA, 2019.
- [12] X. Milhaud, *Cours de Tarification et Datascience*, ISFA, 2020.
- [13] R. Genuer et J.-M. Poggi, *Arbres CART et Forêts aléatoires, importance et sélection de variables*, 2017.
- [14] D. Bhalla, «A complete guide to Random Forest in R,» 2014. [En ligne]. Available: <https://www.listendata.com/2014/11/random-forest-with-r.html>.



ANNEXES

SOMMAIRE

Etude de la valeur historique pour l'année 2017	2
Etude de l'âge	2
Etude de la gamme et du niveau de garantie	3
Etude de la situation de famille	3
Résultats du GLM de la probabilité de résiliation des contrats	5
Segmentation du portefeuille : comparaison entre valeur historique et prospective	6
Segmentation du portefeuille : comparaison entre probabilité de résiliation et marge prospective	8
Segmentation du portefeuille : comparaison entre probabilité de résiliation et valeur prospective	10



ÉTUDE DE LA VALEUR HISTORIQUE POUR L'ANNEE 2017

La même étude a été réalisée que pour l'année 2018 sur les classes de valeur historique de l'année 2017.

Etude de l'âge

La répartition des assurés par âge au sein de chaque classe de valeur historique et du portefeuille global est présentée dans le tableau ci-dessous :

Classe âges	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
0-10	8%	14%	7%	0%	0%	5%
11-20	8%	12%	9%	1%	0%	5%
21-30	13%	23%	15%	2%	0%	11%
31-40	11%	12%	15%	10%	1%	13%
41-50	11%	10%	15%	14%	3%	11%
51-60	13%	10%	14%	17%	16%	12%
61-70	23%	14%	18%	36%	45%	28%
71-80	13%	5%	7%	19%	31%	13%
81-90	1%	0%	0%	1%	3%	1%
Total	100%	100%	100%	100%	100%	100%

Tableau annexe 1.a : Répartition des assurés par âge au sein des différentes classes de valeur historique pour l'année 2017.

Comme pour 2018, les valeurs « bonne » et « forte » sont principalement composées d'assurés de la tranche d'âges 60-80 ans (55% pour la classe bonne valeur et 76% pour la classe forte valeur) alors qu'ils ne représentent que 41% du portefeuille global.

Il apparait également que les moins de 30 ans sont surreprésentés pour la classe de valeur « faible » puisqu'ils la composent à 49% alors qu'ils ne représentent que 21% du portefeuille.

De plus, l'âge moyen de la classe de valeur « forte » est de 66,5 ans contre 35,5 ans pour la classe de valeur « faible ».

Les moins de 30 ans n'étant pas ou très peu représentés pour les classes de valeur « bonne » ou « forte », l'étude de ce tableau permet de conclure que l'assureur aurait intérêt à cibler les plus de 60 ans afin d'augmenter sa rentabilité et qu'un travail peut être effectué sur les moins de 30 ans afin d'améliorer leur valeur. Il est cependant possible que, parmi eux, certains présentent tout de même un potentiel financier pour l'assureur et qu'il peut être intéressant pour lui de déterminer précisément sur quel segment de ces assurés devrait-il concentrer ses efforts.



Etude de la gamme et du niveau de garantie

La répartition des assurés par gamme et niveau de garantie au sein de chaque classe de valeur historique et du portefeuille global est présentée dans le tableau ci-dessous :

Gamme	Niveau	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Gamme1	1	13%	25%	19%	2%	0%	10%
	2	16%	12%	19%	19%	6%	13%
	3	3%	2%	2%	3%	2%	2%
	4	2%	1%	1%	2%	3%	2%
	5	1%	0%	0%	0%	1%	1%
Gamme2	1	11%	8%	13%	26%	3%	10%
	2	25%	12%	14%	29%	59%	32%
	3	4%	2%	2%	4%	13%	5%
	4	2%	1%	1%	1%	8%	2%
	5	0%	0%	0%	0%	2%	0%
Gamme3	1	8%	19%	5%	1%	0%	9%
	2	12%	14%	19%	8%	2%	12%
	3	3%	2%	3%	1%	1%	1%
	4	1%	1%	1%	1%	1%	1%
	5	0%	0%	0%	0%	0%	0%
Total		100%	100%	100%	100%	100%	100%

Tableau annexe 1.b : Répartition des assurés par gamme et niveau de garantie au sein des différentes classes de valeur historique pour l'année 2017.

Les valeurs « bonne » et « forte » sont principalement composées d'assurés de la gamme 2 (60% pour la classe bonne valeur et 85% pour la classe valeur « forte ») alors qu'ils ne représentent que 49% du portefeuille. Ces deux classes sont très peu représentées par la gamme 3 (4% pour la valeur « forte ») alors qu'elle compose à 23% le portefeuille.

Les classes de valeurs moindres sont donc principalement constituées des assurés des gammes 1 et 3. Le niveau 1 représente notamment 52% de la classe de valeur « faible » alors qu'il ne concerne que 29% des contrats du portefeuille.

L'assureur peut donc miser sur la gamme 2 (et notamment le niveau 2 de garantie) afin d'augmenter sa rentabilité. A contrario, il peut être intéressant de cibler le niveau 1 des garanties et les gammes 1 et 3 afin d'améliorer la valeur de son portefeuille.

Etude de la situation de famille

La répartition des assurés par situation de famille au sein de chaque classe de valeur historique et du portefeuille global est présentée dans le tableau ci-contre :



Situation famille	Valeur mauvaise	Valeur faible	Valeur moyenne	Valeur bonne	Valeur forte	Portefeuille Global
Isole	61%	53%	61%	72%	75%	65%
Duo	11%	7%	9%	16%	21%	13%
Couple+Enf	9%	14%	11%	5%	2%	8%
MonoParent	19%	26%	19%	6%	2%	15%
Total	100%	100%	100%	100%	100%	100%

Tableau annexe 1.c : Répartition des assurés par situation de famille au sein des différentes classes de valeur historique pour l'année 2017.

Les valeurs « bonne » et « forte » comportent une plus grosse part de contrats « Isolé » ou « Duo » que le portefeuille global et il y a très peu de contrats avec enfants pour ces classes. A contrario, la classe de valeur « faible » contient une plus grande proportion de contrats avec enfants que le portefeuille global.

On retrouve donc ici les résultats obtenus avec l'étude de l'âge, indiquant que les jeunes (soit ici, les contrats avec enfants) ont tendance à avoir une valeur historique plutôt faible.



RESULTATS DU GLM DE LA PROBABILITE DE RESILIATION DES CONTRATS

Le GLM implémenté afin de modéliser la probabilité de résiliation des contrats donne les résultats suivants :

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.8122914	0.0528856	-15.359	< 2e-16	***
NombreBenef	0.1161497	0.0216189	5.373	7.76e-08	***
SituationFamilleCouple+Enf	0.5939187	0.0905361	6.560	5.38e-11	***
SituationFamilleDuo	0.4190549	0.0445715	9.402	< 2e-16	***
SituationFamilleMonoParent	0.3857145	0.0635736	6.067	1.30e-09	***
Anciennete20181	0.7658915	0.0277384	27.611	< 2e-16	***
Anciennete20182	0.3597696	0.0302034	11.912	< 2e-16	***
Anciennete20183	0.1291255	0.0311096	4.151	3.32e-05	***
Anciennete20184+	-0.1315910	0.0276186	-4.765	1.89e-06	***
GammeGarantieGamme2	-0.3265548	0.0351912	-9.279	< 2e-16	***
GammeGarantieGamme3	0.0346306	0.0570870	0.607	0.5441	
NiveauGarantie2	0.1083249	0.0207773	5.214	1.85e-07	***
NiveauGarantie3+	0.1343816	0.0277141	4.849	1.24e-06	***
RegimeSocialeAutres régimes spéciaux	-0.1834798	0.0336750	-5.449	5.08e-08	***
SexeH	-0.0591522	0.0192363	-3.075	0.0021	**
Age2018	-0.0044252	0.0009911	-4.465	8.01e-06	***
ClassePrestasGLMFORT	-0.9550539	0.0295267	-32.345	< 2e-16	***
ClassePrestasGLMIMPORTANT	-1.2858212	0.0326777	-39.349	< 2e-16	***
ClassePrestasGLMMOYEN	-0.5393636	0.0296672	-18.180	< 2e-16	***
ClassePrestasGLMNUL	0.2511184	0.0254040	9.885	< 2e-16	***
ConsoHospita20181	0.2369944	0.0256705	9.232	< 2e-16	***
ConsoOptique20181	-0.2630756	0.0299834	-8.774	< 2e-16	***

Figure annexe 2 : Résultats du GLM de modélisation de la probabilité de résiliation.

Après avoir regroupé les modalités de certaines variables, elles apparaissent comme significatives. Seules les modalités de la gamme de garantie n'ont pas été regroupées ici car la distinction apparaissait comme importante pour la suite de l'étude.

Le contrat de référence du modèle n'a pas de bénéficiaire, a une situation de famille « Isolé », est d'ancienneté 0 et a souscrit au niveau 1 de la gamme 1 de garantie. L'assuré principal est une femme d'âge 0, appartenant au régime général, ayant consommé un faible nombre de prestations et hors optique et hospitalisation.

La probabilité de résiliation augmente avec le nombre de bénéficiaires et pour toute autre situation de famille que celle de l'individu de référence. Elle est croissante en fonction de l'ancienneté puis décroît à partir de 4 ans. Les contrats de la gamme 2 ont une plus faible probabilité de résiliation, contrairement aux assurés de la gamme 3 et la probabilité de résiliation croît avec le niveau de garantie. Elle décroît en fonction de l'âge et de la consommation. Enfin, les contrats ayant consommé en hospitalisation ont une plus forte probabilité de consommation, contrairement à ceux ayant consommé en optique.



SEGMENTATION DU PORTEFEUILLE : COMPARAISON ENTRE VALEUR HISTORIQUE ET MARGE PROSPECTIVE THEORIQUE

Le tableau ci-dessous compare la répartition des assurés selon différents critères, pour le portefeuille global, les assurés ayant une valeur historique et prospective faible, les assurés ayant une valeur historique et prospective importante, les assurés ayant une valeur historique faible mais une marge prospective théorique importante, les assurés ayant une marge historique importante et une marge prospective théorique faible et les assurés ayant une valeur historique et prospective moyenne.

	Portefeuille global	Faible valeur	Forte valeur	Faible puis forte valeur	Forte puis faible valeur	Valeur moyenne
Prop. portefeuille	100%	21,5%	25,2%	10,3%	8,1%	5,1%
Gamme 1 – Niv 1	7%	24%	1%	2%	5%	18%
Gamme 1 – Niv 2	12%	18%	6%	8%	27%	15%
Gamme 1 – Niv 3+	6%	5%	4%	5%	11%	2%
Gamme 2 – Niv 1	13%	7%	10%	12%	24%	13%
Gamme 2 – Niv 2	30%	5%	49%	44%	23%	9%
Gamme 2 – Niv 3+	7%	0%	22%	16%	2%	1%
Gamme 3 – Niv 1	8%	23%	1%	1%	1%	9%
Gamme 3 – Niv 2	12%	14%	4%	7%	6%	25%
Gamme 3 – Niv 3+	5%	4%	2%	5%	2%	6%
Age < 20 ans	11%	35%	0%	2%	1%	21%
Age : 20-30 ans	11%	24%	1%	2%	5%	15%
Age : 30-40 ans	12%	14%	3%	6%	13%	16%
Age : 40-50 ans	12%	10%	5%	9%	16%	15%
Age : 50-60 ans	14%	8%	14%	15%	22%	14%
Age : 60-70 ans	27%	7%	41%	37%	30%	15%
Age > 70 ans	13%	2%	37%	30%	13%	4%
Ancienneté : 0	22%	35%	13%	33%	9%	18%
Ancienneté : 1	16%	15%	16%	12%	18%	18%
Ancienneté : 2	15%	12%	17%	14%	18%	14%
Ancienneté : 3	16%	13%	19%	14%	18%	16%
Ancienneté : 4+	31%	24%	35%	26%	38%	34%

Tableau annexe 3 : Répartition des effectifs selon plusieurs critères parmi les différents segments du portefeuille identifiés grâce aux valeurs historiques et marges prospectives.

Les assurés ayant une Valeur Client faible représentent 21,5% du portefeuille. Ce sont principalement les plus jeunes assurés puisque 59% des assurés de cette classe ont moins de 30 ans (alors qu'ils ne représentent que 36% du portefeuille global). Ces assurés sont principalement



assurés de la gamme 1, niveau 1 ou de la gamme 3 pour les niveaux 1 et 2. De plus, ce segment contient une part importante de nouveaux assurés (en 2018) puisque 36% des assurés ont moins d'un an d'ancienneté.

Les assurés ayant une Valeur Client forte représentent 25,2% du portefeuille. Ce sont principalement les plus vieux assurés puisque 78% des assurés de cette classe ont plus de 60 ans. Ces assurés sont principalement assurés de la gamme 2, niveaux 2 ou supérieurs. L'ancienneté ne semble pas avoir d'influence même s'il y a peu de nouveaux assurés (de 2018).

Les assurés ayant une valeur historique faible et une marge prospective théorique forte représentent 10,3% du portefeuille. Ce sont principalement les plus vieux assurés (plus de 60 ans) puisque leur proportion au sein de cette classe est supérieure à celle au sein du portefeuille global (ils représentent 67% de cette classe contre 19% du portefeuille). Ces assurés sont notamment assurés de la gamme 2, niveaux 2 ou supérieurs et ont pour la plupart, une ancienneté (en 2018) inférieure à 1 an.

Les assurés ayant une valeur historique forte et une marge prospective théorique faible représentent 8,1% du portefeuille. Les 50-70 ans sont particulièrement plus présents au sein de cette classe que dans le portefeuille global. Ce sont principalement les assurés de la gamme 1 qui ont une ancienneté (en 2018) d'au moins 4 ans.

Les assurés ayant une Valeur Client moyenne sur toute la durée de leur relation représentent 5,1% du portefeuille. Ce sont principalement les assurés de moins de 40 ans, du niveau 1 de la gamme 1 ou des niveaux 2 et supérieurs de la gamme 3. L'ancienneté ne semble pas avoir de réel impact sur ce segment.



SEGMENTATION DU PORTEFEUILLE : COMPARAISON ENTRE PROBABILITE DE RESILIATION ET MARGE PROSPECTIVE THEORIQUE

Le tableau ci-dessous compare la répartition des assurés selon différents critères, pour le portefeuille global, les assurés ayant une marge prospective théorique faible et une probabilité de résiliation faible (<15%) et moyenne (entre 15% et 35%), les assurés ayant une marge prospective théorique importante et une probabilité de résiliation faible (<15%), moyenne (entre 15% et 35%) et importante (> 35%), les assurés ayant une marge prospective théorique dans la moyenne du portefeuille et une probabilité de résiliation faible (<15%) et moyenne (entre 15% et 35%). Les segments de marge prospective théorique faible et moyenne ayant une forte probabilité de résiliation n'ont pas été détaillés ici puisque leur proportion est faible au sein du portefeuille et qu'ils ne présentent qu'un intérêt limité pour l'assureur.

	Ptf. global	Apport faible, proba faible	Apport faible, proba moy.	Apport fort, proba faible	Apport fort, proba moy.	Apport fort, proba forte	Apport moy. Proba faible	Apport moy. Proba moy
Prop. portefeuille	100%	16,8%	20,4%	30,5%	8,3%	0,7%	9,5%	9,5%
Gamme 1 – Niv 1	7%	26%	21%	1%	3%	3%	8%	13%
Gamme 1 – Niv 2	12%	20%	23%	4%	18%	19%	11%	23%
Gamme 1 – Niv 3+	6%	5%	7%	2%	12%	8%	3%	6%
Gamme 2 – Niv 1	13%	25%	1%	13%	4%	0%	31%	2%
Gamme 2 – Niv 2	30%	19%	1%	56%	17%	0%	40%	3%
Gamme 2 – Niv 3+	7%	2%	0%	22%	9%	0%	5%	1%
Gamme 3 – Niv 1	8%	2%	21%	0%	3%	4%	1%	14%
Gamme 3 – Niv 2	12%	1%	22%	0%	20%	29%	1%	30%
Gamme 3 – Niv 3+	5%	0%	5%	0%	13%	37%	0%	8%
Age < 20 ans	11%	2%	34%	0%	4%	11%	0%	23%
Age : 20-30 ans	11%	11%	25%	0%	4%	10%	2%	17%
Age : 30-40 ans	12%	14%	19%	0%	12%	28%	3%	22%
Age : 40-50 ans	11%	15%	13%	1%	21%	32%	7%	19%
Age : 50-60 ans	15%	19%	6%	8%	31%	13%	19%	14%
Age : 60-70 ans	26%	27%	1%	43%	22%	1%	45%	4%
Age > 70 ans	14%	12%	0%	47%	7%	1%	24%	0%
Ancienneté : 1	22%	8%	29%	11%	41%	85%	12%	30%
Ancienneté : 2	16%	14%	20%	15%	14%	9%	15%	17%
Ancienneté : 3	15%	18%	13%	18%	10%	3%	18%	12%
Ancienneté : 4+	47%	60%	38%	56%	34%	3%	55%	41%

Tableau annexe 4 : Répartition des effectifs selon plusieurs critères parmi les différents segments du portefeuille identifiés grâce aux probabilités de résiliations et marges prospectives.



Remarques :

- La répartition des assurés par âge et ancienneté au sein du portefeuille global diffère légèrement de l'annexe précédente puisque 2019 est considérée comme année de référence alors que celle de l'étude précédente était 2018.
- Il n'y a pas de nouveaux assurés dans cette segmentation. L'ancienneté minimale est donc d'un an.

Les assurés ayant une marge prospective théorique faible et une probabilité de résiliation faible représentent 16,8% du portefeuille. 46% de ces personnes sont des assurés de la gamme 1, niveau 1 ou 2 et 25% sont assurés de la gamme 2, niveau 1. De plus, ce segment du portefeuille contient une part importante d'anciens assurés puisque 60% des assurés de ce segment ont au moins 4 ans d'ancienneté (en 2019).

Les assurés ayant une marge prospective théorique faible et une probabilité de résiliation moyenne représentent 20,4% du portefeuille. Ce sont principalement des assurés de moins de 40 ans qui ont choisi les gammes 1 et 3, de niveau 1 ou 2 qui apparaissent dans ce groupe d'assurés. L'ancienneté ne semble pas avoir d'impact particulier.

Les assurés ayant une marge prospective théorique forte et une probabilité de résiliation faible représentent 30,5% du portefeuille. 90% des assurés ont plus de 60 ans et que la quasi-totalité de ces personnes sont des assurés de la gamme 2 (92% d'entre elles). De plus, l'ancienneté ne semble pas avoir d'impact significatif même s'il y a une légère prépondérance d'assurés depuis au moins 4 ans.

Les assurés ayant une marge prospective théorique forte et une probabilité de résiliation moyenne représentent 8,3% du portefeuille. Ils sont généralement assurés de la gamme 1, niveau 3 ou plus ou de la gamme 3, niveau 2 ou plus. La majorité de ce groupe d'assurés est âgé entre 30 et 50 ans et l'ancienneté du contrat ne semble pas avoir d'influence particulière.

Les assurés ayant une marge prospective théorique forte et une probabilité de résiliation forte représentent 0,7% du portefeuille. Ce sont principalement les assurés de 30 à 50 ans qui sont présents dans ce segment. 70% de ces personnes sont des assurés de la gamme 3. De plus, une part importante de nouveaux assurés est contenue dans ce segment du portefeuille puisque 85% des assurés ont 1 an d'ancienneté en 2019.

Les assurés ayant une marge prospective théorique moyenne et une probabilité de résiliation faible représentent 9,5% du portefeuille. Ce sont principalement les assurés de plus de 60 ans, assurés des niveaux inférieurs de la gamme 2 qui sont constitutifs de ce segment. De plus, l'ancienneté ne semble pas avoir d'impact significatif même s'il y a une légère prépondérance d'assurés depuis au moins 4 ans en 2019.

Les assurés ayant une marge prospective théorique moyenne et une probabilité de résiliation moyenne représentent 9,5% du portefeuille. Les assurés sont âgés de moins de 50 ans et sont assurés de la gamme 3, tous niveaux confondus, ou de la gamme 1, niveau 2. L'ancienneté du contrat ne semble pas avoir d'impact.



SEGMENTATION DU PORTEFEUILLE : COMPARAISON ENTRE PROBABILITE DE RESILIATION ET VALEUR PROSPECTIVE

Le tableau ci-dessous compare la répartition des assurés selon différents critères, pour le portefeuille global, les assurés ayant une valeur prospective faible et une probabilité de résiliation faible (<15%) et moyenne (entre 15% et 35%), les assurés ayant une valeur prospective importante et une probabilité de résiliation faible (<15%), moyenne (entre 15% et 35%) et importante (> 35%), les assurés ayant une valeur prospective dans la moyenne du portefeuille et une probabilité de résiliation faible (<15%) et moyenne (entre 15% et 35%). Les segments de valeur prospective faible et moyenne ayant une forte probabilité de résiliation n'ont pas été détaillés ici puisque leur proportion est faible au sein du portefeuille et qu'ils ne présentent qu'un intérêt limité pour l'assureur.

	Ptf. global	Apport faible, proba faible	Apport faible, proba moy.	Apport fort, proba faible	Apport fort, proba moy.	Apport fort, proba forte	Apport moy. Proba faible	Apport moy. Proba moy
Prop. portefeuille	100%	16,1%	20,7%	31,5%	7,7%	0,8%	9,1%	9,1%
Gamme 1 – Niv 1	7%	27%	21%	1%	3%	2%	9%	13%
Gamme 1 – Niv 2	12%	21%	23%	4%	18%	16%	12%	23%
Gamme 1 – Niv 3+	6%	5%	7%	2%	13%	11%	3%	6%
Gamme 2 – Niv 1	13%	24%	1%	14%	4%	0%	31%	2%
Gamme 2 – Niv 2	30%	18%	1%	56%	18%	0%	38%	3%
Gamme 2 – Niv 3+	7%	2%	0%	22%	10%	0%	5%	1%
Gamme 3 – Niv 1	8%	2%	21%	0%	3%	4%	1%	14%
Gamme 3 – Niv 2	12%	1%	22%	0%	19%	29%	1%	29%
Gamme 3 – Niv 3+	5%	0%	5%	0%	14%	39%	0%	8%
Age < 20 ans	11%	2%	34%	0%	2%	8%	0%	21%
Age : 20-30 ans	11%	12%	25%	0%	4%	7%	2%	16%
Age : 30-40 ans	12%	14%	19%	0%	11%	25%	4%	22%
Age : 40-50 ans	11%	15%	14%	1%	20%	37%	7%	20%
Age : 50-60 ans	15%	19%	6%	8%	31%	20%	20%	15%
Age : 60-70 ans	26%	26%	2%	43%	23%	2%	45%	4%
Age > 70 ans	14%	12%	0%	47%	8%	1%	22%	1%
Ancienneté : 1	22%	8%	30%	11%	42%	81%	13%	26%
Ancienneté : 2	16%	14%	20%	15%	14%	11%	15%	18%
Ancienneté : 3	15%	18%	13%	18%	10%	3%	18%	13%
Ancienneté : 4+	47%	60%	38%	56%	34%	5%	55%	43%

Tableau annexe 5 : Répartition des effectifs selon plusieurs critères parmi les différents segments du portefeuille identifiés grâce aux probabilités de résiliations et valeurs prospectives.



Remarque : Il n'y a pas de nouveaux assurés dans cette segmentation. L'ancienneté minimale est donc d'un an.

Les assurés ayant une valeur prospective faible et une probabilité de résiliation faible représentent 16,1% du portefeuille. 48% de ces personnes sont des assurés de la gamme 1, niveau 1 ou 2 et 24% sont assurés de la gamme 2, niveau 1. De plus, ce segment du portefeuille contient une part importante d'anciens assurés puisque 60% des assurés ont au moins 4 ans d'ancienneté (en 2019).

Les assurés ayant une valeur prospective faible et une probabilité de résiliation moyenne représentent 20,7% du portefeuille. Ce sont principalement des assurés de moins de 40 ans qui ont choisi les gammes 1 et 3, de niveau 1 ou 2 qui sont présents dans ce segment. L'ancienneté ne semble pas avoir d'impact particulier.

Les assurés ayant une valeur prospective forte et une probabilité de résiliation faible représentent 31,5% du portefeuille. 90% des assurés ont plus de 60 ans et la quasi-totalité de ces personnes sont des assurés de la gamme 2 (92% d'entre elles). De plus, l'ancienneté ne semble pas avoir d'impact significatif même s'il apparait une légère prépondérance d'individus assurés depuis au moins 4 ans.

Les assurés ayant une marge prospective théorique forte et une probabilité de résiliation moyenne représentent 7,7% du portefeuille. Ils sont généralement assurés de la gamme 1, niveau 2 ou plus ou de la gamme 3, niveau 2 ou plus. La majorité de ce groupe d'assurés est âgé entre 40 et 60 ans et l'ancienneté du contrat ne semble pas avoir d'influence particulière.

Les assurés ayant une valeur prospective forte et une probabilité de résiliation forte représentent 0,8% du portefeuille. Ce sont principalement les assurés de 30 à 60 ans qui sont constitutifs de ce segment et 68% de ces personnes sont des assurés de la gamme 3. De plus, ce segment du portefeuille contient une part importante de nouveaux assurés puisque 81% d'entre eux ont 1 an d'ancienneté en 2019.

Les assurés ayant une valeur prospective moyenne et une probabilité de résiliation faible représentent 9,1% du portefeuille. Ce sont principalement les assurés de plus de 60 ans, assurés des niveaux inférieurs de la gamme 2 qui apparaissent ici. De plus, l'ancienneté ne semble pas avoir d'impact significatif même s'il y a une légère prépondérance d'assurés depuis au moins 4 ans en 2019.

Les assurés ayant une valeur prospective moyenne et une probabilité de résiliation moyenne représentent 9,1% du portefeuille. Les assurés sont âgés de moins de 50 ans et sont assurés de la gamme 3, niveau 2 ou plus, ou de la gamme 1, niveau 2. L'ancienneté du contrat ne semble pas avoir d'impact.


THE RISKTECH FOR INSURANCE

