Mémoire présenté le :

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

Par : Yacine Aboutaybi

Titre : Refonte de la Valeur Client Multi-risque Commerce - Étude de la résiliation

Confidentialité : ☐ NON ☒ (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. PIGEON
.....

A. YOU
.....

M. HUYGHUES-BEAUFOND
.....

Membres présents du jury de l'ISFA

P. RIBEREAU
.....

.....

.....

Entreprise : Generali France
Nom : Samia Nouar
Signature :

Directeur de mémoire en entreprise :
Nom : Samia Nouar
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

Résumé

Le marché de l'assurance est un marché mûr, les tarifs fixés ainsi que le service fourni et la marque de l'assureur détermine généralement le comportement du client après le devis. Ce climat hostile aux tarifs excessifs fait que l'assureur se doit de choisir convenablement des profils jugés rentables à long terme afin de pouvoir se montrer agressif sur le prix proposé, ce raisonnement est encore plus pertinent sur les produits obligatoires à l'exercice d'une activité ou d'un besoin nécessaire (Automobile, Multi-Risques Habitation, Multi-Risques Commerce, ...) où la concurrence se fait la plus rude.

En réponse à cela, Generali opte pour une étude de rentabilité à deux niveaux : par produit et par client. En effet, une approche de la rentabilité par contrat n'est pas suffisante compte-tenu de la corrélation qu'ont les contrats entre eux lors qu'ils sont détenus par le même client. C'est la seconde étude dont il sera question ici par la notion de Valeur Client. Une notion qui transcrit la rentabilité actuelle mais aussi prospective du client, en effet elle correspond la projection actualisée des résultats techniques et financiers des contrats du client sur une période probable de présence en portefeuille. Elle comporte également un volet valeur potentielle qui correspond à l'appétence de souscrire à de nouveaux contrats ayant eux mêmes une certaine rentabilité.

L'étude ici est de participer à la refonte de la Valeur Client du produit Multi-Risques Commerce (MRC) sur le portefeuille professionnel et petites entreprises représentant la grande majorité des clients de ce produit. L'apport sera multiple, ce mémoire permettra de créer des lois de chute grâce à notre modélisation de la résiliation (nécessaire à la projection des résultats techniques et financiers), de mieux expliquer les mouvements des clients et d'inclure cet indicateur à de plus nombreux processus.

Toute la question réside plus dans la sélection des profils de clients que sur le nombre de contrats souscrits. Ce mémoire donne des indicateurs indispensables à ce type de raisonnement.

Mots clés : valeur client, valeur potentielle, résiliation, lois de chute, modélisation, rentabilité, ratio technique, ratio combiné, Multi-Risque Commerce.

Abstract

The insurance market is mature, rates, the service provided and the brand of the insurer, generally determine the customer's behaviour after the quote. This climate of hostility to excessive premiums compels the insurer to choose carefully long-term profitable profiles in order to be aggressive on the price offered. This approach is even more relevant for mandatory products for the exercise of an activity or a necessary need (Motor, Home Insurance, Commercial insurance, ...) where competition is fiercest.

In response to this, Generali opts for a profitability study at two levels : by product and by client. Indeed, an approach to profitability per contract is not satisfactory given the correlation between contracts held by the same client. This is the second study that will be discussed here through the notion of Customer Value. A notion that combines the current but also prospective profitability of a client, in fact it refers to the discounted projection of the technical and financial results of the client's contracts over a probable period of presence in the portfolio. It also includes a potential value component which corresponds to the appetite for subscribing to new contracts which themselves have a certain profitability.

This study aims to rebuild the Commercial insurance (MRC) Customer Value for professionals and small businesses which represent the vast majority of the customers of this product. The objectives will be multiple, this thesis will allow us to determine more reliable lapse rates, to have a better understanding of client behaviour and to take this indicator into account in more processes. All the matter is more about the selection of the profiles than the number of subscribed contracts.

This thesis provides approaches that are essential for this type of analysis.

Key words : Custumer Value, Potential Value, Cancelling, Policy lapse rate, Modelling, Profitability, Technical ratio, Combined ratio, Commercial insurance.

Introduction générale

Contexte économique et émergence du concept de Valeur Client

L'entrée en vigueur de la loi Hamon 17 mars 2014 durcit une fois de plus la concurrence d'une activité déjà très compétitive, elle stipule que l'assuré, sous certaines conditions, peut résilier à tout moment son contrat après un an de couverture. Cet assouplissement des conditions de résiliation offre ainsi un moyen supplémentaire pour l'assuré de faire pression sur les différentes compagnies. Ceci accentue la compression des marges des assureurs et rend l'assuré plus versatile quant à son assureur qui se doit de lui proposer ce que la concurrence lui propose. Auparavant, l'inertie des comportements, forcée par les contraintes légales et administratives, dispensait l'assureur de se mettre nécessairement au diapason de ses rivaux en affaire.

Ce changement réglementaire est un signe supplémentaire que le contexte assurentiel est tendu et la concurrence âpre est exacerbé par un marché évoluant que très faiblement avec des produits pratiquement identiques d'un assureur à un autre.

Pour toutes ces raisons, l'assureur se doit de cibler sa clientèle afin de ressortir une marge décente d'un tarif agressif. La population doit être en adéquation avec l'objectif et la structuration du produit. Ce ciblage peut être opéré par la mise en place d'offres commerciales plus ou moins personnalisées attirant les clients à forte rentabilité future.

La question est de cerner ce type de client le plus rapidement possible et de les garder en portefeuille.

La notion de Valeur Client permet en partie de résoudre cette problématique en ayant une notion de rentabilité centrée sur le client et non sur le produit. Le calcul ainsi que l'idée de la Valeur seront expliqués plus en profondeur par la suite.

En quelques mots, il s'agit des projections actualisées des marges techniques et financières du client sur les contrats qu'il possède et qu'il pourrait potentiellement posséder.

Cet indicateur permet de segmenter le portefeuille afin d'établir des actions ciblées sur ceux-ci et également d'apporter des indicateurs de rentabilités transverses.
Malgré une émergence assez récente [1], plusieurs assureurs ont déjà implémentés cette

1. dans les années 2000

notion dans leur processus d'ajustement tarifaire ou d'approche du client. Generali est l'un d'entre eux.

Generali France

Generali France est une filiale du groupe Generali, l'un des plus importants groupes mondiaux d'assurance et de services financiers avec une riche expérience de près de 2 siècles et 61 millions de clients à travers le monde. La France est le troisième marché de Generali avec une forte position en Assurance Vie.

Naturellement, un groupe aussi important s'est intéressé à cette notion et lui a même dédié une équipe à part entière au sein du département "Technique Assurance".

Au sein de la direction "Données et approche client", l'équipe "Analytics Avancées" comporte deux pôles : le pôle "DataLab" et le pôle "Valeur Client" où a été réalisé cette étude.

Le service Valeur Client s'occupe essentiellement des missions suivantes :

↪ Le calcul et la sophistication de la Valeur Client.

↪ Les applications de la Valeur Client dans les processus d'ajustement tarifaire, de résiliations et d'offres commerciales.

↪ Création d'outils de reporting destinés aux réseaux de distribution, aux équipes produit et de tarification.

Les échanges entre l'équipe Valeur Client et les équipes de tarification sont nombreux ainsi qu'avec les équipes : Connaissance Client, Gestion de la donnée.

Sujet de ce mémoire

Le sujet de ce mémoire porte sur le produit MRC dont la Valeur est perfectible. Elle est revue afin d'être utilisée de manière plus directe dans les processus commerciaux. [2]

Cette refonte de la Valeur sera possible après un travail approfondi sur les indicateurs et données qui permettent son calcul.

2. Les raisons de l'étude seront détaillées dans la suite de ce mémoire.

Certains de ces indicateurs sont l'objet de ce mémoire : création de probabilité de résiliations sur un an selon le type de profil et élaboration de lois de chute ainsi que dans un second temps, l'élaboration de modèles d'appétence au multi-équipement pour servir le calcul de la valeur potentielle. Cette deuxième partie sera un enrichissement futur de ce mémoire et donc ne sera pas présentée ici.

Il s'agit donc d'un questionnement interne au service étudié par une approche académique dont les problématiques peuvent être explicités de la manière suivante :

Comment calculer prospectivement de manière fiable et explicable les probabilités de résiliation des contrats Multi-Risque Commerce ? Comment interpréter ces mouvements futurs ?

La résolution de ces interrogations servira à projeter les marges techniques et financières dans le futur en pondérant les marges futures par la probabilités de survie dans le portefeuille.

Pour tâcher d'apporter une réponse à cela, nous allons, dans un premier temps, expliquer la notion de Valeur Client puis décrire les pré-requis aux modélisations et enfin présenter la modélisation technique de la résiliation et ses résultats en traitant avec plusieurs types d'hypothèses et de modèles afin de répondre de la manière la plus complète à la problématique.

Il sera question de traiter la résiliation sur une courte période par le calcul de probabilités de résiliation à la maille contrat puis par leur agrégation d'établir des lois de résiliation sur une longue durée par groupe de contrats. Une note sera apportée en ce qui concerne le lissage des probabilités de résiliation par aggrégat.

Cette étude aura pour principe d'obtenir un compromis entre une modélisation performante mais opaque et une modélisation transparente, compréhensible mais moins performante. Ce travail a été fait en supprimant pas à pas certaines hypothèses de modélisation.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au bon déroulement et qui m'ont aidé lors de la rédaction de ce rapport.
Tout d'abord, je voudrais remercier ma tutrice de stage, Samia Nouar, responsable d'études d'actuariat au sein de l'équipe Valeur Client, de m'avoir encadré et s'être montrée toujours disponible pour répondre à mes interrogations. Elle a partagé ses connaissances de manière très pédagogique et a su m'accompagner tout en préservant mon autonomie et mon indépendance, le tout avec patience et gentillesse. Je remercie également le manager du service, Alexandre Cochard, et Hamza El Hassani, d'abord coordinateur de l'équipe puis manager de celle-ci, de m'avoir fait confiance pour ce poste d'alternant ainsi que dans les missions conférées durant ce stage. Je souhaite également remercier toute l'équipe Advanced Analytics, qui m'a accueilli avec beaucoup de gentillesse et de bienveillance.
La dualité Actuariat/Data science de l'équipe m'a permis d'avoir une vision plus globale du domaine dans lequel j'évolue.
Je souhaite ensuite adresser mes remerciements à l'Institut de Science Financière et d'Assurances dont la formation dispensée m'a permis d'effectuer mes tâches techniques et dont le Forum des Entreprises a rendu possible le premier contact avec Generali.

Table des matières

1 La Valeur Client : une mesure de rentabilité innovante

1.1 La modélisation de la valeur du client, ses composantes et ses applications

La Valeur Client est une notion assez récente dénotant d'une nouvelle vision de la rentabilité où le client est au coeur du sujet. En effet, elle peut permettre d'avoir une vision orthogonale à celle de la rentabilité par produit dans le cas des clients ayant plusieurs contrats (appelé ici multi-équipés). Il peut être alors possible d'envisager qu'une mauvaise rentabilité sur certains produits notamment des produits d'appel, peut à long terme être une situation positive si elle permet la conquête de profils à forte rentabilité sur d'autres produits. De ce fait, elle permet également de déterminer la limite supérieure des coûts d'acquisition d'un nouveau client compte tenu de son profil. Une comparaison peut être faite avec l'Embedded value qui se place du point de vue de l'actionnaire sur la branche Vie de l'assurance. Cet indicateur permet de se rendre compte des avantages d'une rétention du client à long terme et la nécessité de fidéliser les clients ayant une propension à se comporter ainsi. La valeur du client est en fait décomposée en deux versants : la valeur actuelle de l'assuré et la valeur potentielle. On peut remarquer qu'elle ne tient pas compte de la valeur passée du client mais seulement de la valeur que lui attribue ses contrats à l'instant du calcul (valeur actuelle) et la valeur que lui attribue son profil dans le futur (valeur potentielle). Concentrons-nous d'abord sur la valeur actuelle du client avec l'étude d'une maille plus fine : celle des contrats.

1.1.1 La valeur des contrats

La valeur d'un contrat est l'espérance de la somme des profits, techniques comme financiers, du contrat actualisés à la date du calcul.

La valeur du contrat d'assurance réside dans deux projections : celle de la marge technique et celle de la marge financière estimées sur un horizon pouvant aller jusqu'à 30 ans. Cet horizon diffère en fonction du produit et du profil client dont il est question. Pour le

produit MRC, il sera question d'une projection allant jusqu'à 27 ans.

1.1.1.1 Le résultat technique

Pour calculer le résultat technique, il sera question de projeter les primes, les frais et les sinistres et en découle naturellement, pour l'année i, le résultat technique :

$$\boxed{RT_i = Primes_i - Sinistres_i - Frais_i} \tag{1.1}$$

Pour projeter les primes, on applique la relation de récurrence suivante, pour i=1 :

$$\boxed{Primes_1 = Primes} \tag{1.2}$$

Pour i>1,

$$\boxed{Primes_i = Primes_{i-1} * (1 + Tx_i^{evolution}) * (1 - Tx_i^{chute})} \tag{1.3}$$

où Tx_i^{chute} représente la probabilité conditionnelle à la présence de l'assuré en i-1 de résilié son contrat en i et $Tx_{i-1}^{evolution}$ le taux prenant en compte les majorations probables de l'assuré.

Les indicateurs utilisés ici proviennent des Bureaux d'Etudes Techniques et sont révisés de manière périodique.

En ce qui concerne les sinistres, leur projection dépend également des lois de chute et le taux de majoration est remplacé par le taux d'inflation de chaque année (une sorte de majoration structurelle). Les primes, elles, sont fixées et le montant n'évolue qu'au gré des majorations et résiliations éventuelles.

De ce fait, pour ne pas être dépendant de l'année de projection et garder en robustesse, nous nous aidons des expériences passées et utilisons le S/P cible (indicateur de rentabilité Sinistres/Primes) de la branche que l'on multiplie à la prime, qui est elle connue pour la première année, on a :

$$\boxed{Sinistres_1 = S/P * Primes} \tag{1.4}$$

Pour i>1 :

$$\boxed{Sinistres_i = Sinistres_{i-1} * (1 + Tx_i^{inflation}) * (1 - Tx_i^{chute})} \tag{1.5}$$

Pour les frais, ils regroupent les frais de gestion et les commissions, dont le taux est noté Tx_{FGSCOM} et qui ne dépend pas de l'année d'évaluation. Des dispositifs le maintiennent relativement stable (pour le produit MRC, ce taux est d'environ 30 %) :

$$\boxed{Frais_i = Primes_i * Tx_{FGSCOM}} \tag{1.6}$$

Le calcul des trois projections a été présenté et repose sur plusieurs indicateurs économiques ainsi que sur les lois de chute que nous étudierons plus tard.

Une fois ce travail effectué, les résultats techniques des années de projections sont établis. Pour obtenir le résultat technique global il nous faut actualiser tout ces résultats à la date d'évaluation et prendre en compte l'imposition :

$$\boxed{RT = \sum_{i=1}^{max.projection} (RT_i * Deflateur_i * (1 + Tx_i^{impots}))} \tag{1.7}$$

La première partie du calcul de la valeur d'un contrat a été exposée. Passons donc à la seconde.

1.1.1.2 Le résultat financier

Le résultat financier se base sur la provision qui représente le montant provisionné et investi. On le calcule de la manière suivante :

$$\boxed{Provision_i = \sum_{j=1}^{i} Tx_k^{PSAP} * Sinistres_{i-k+1}} \tag{1.8}$$

où le Tx_k^{PSAP} est le taux de Provision pour Sinistres à Payer qui est le coût total estimé de tous les sinistres survenus jusqu'à la fin de l'exercice, déclarés ou non, déduction faite des sommes déjà payées au titre de ces sinistres.

L'assurance est par essence une activité particulière où le cycle de production est inversé et dont la santé financière est primordiale à l'équilibre économique et aux assurés. De ce

fait, la réglementation est très présente notamment sur la solvabilité de l'entreprise. Calculons alors cette marge de solvabilité qui exige de sécuriser la position financière de l'entreprise. Cette marge étant une exigence de capital, elle figurera au passif du résultat :

$$\boxed{Marge.Solva_i = Primes_i * Tx_{Marge.Solva} * Tx_{Cot.Capital}} \tag{1.9}$$

où le $Tx_{Cot.Capital}$ est le taux de rentabilité annuel attendu par les actionnaires et les créanciers, en retour de leur investissement, d'où le nom du coût du capital qui est ici de 6 % et le $Tx_{Marge.Solva}$ le taux réglementaire du taux de marge de solvabilité.

Il suffit donc d'actualiser et de prendre en compte l'imposition pour avoir le résultat financier :

$$\boxed{RF = \sum_{i=1}^{max.projection} (Provision_i * Tx_i^{Rendement} - Marge.solva_i) * Deflateur_i * (1 - Tx_i^{Chute})} \tag{1.10}$$

On peut faire l'analogie entre le résultat financier et le résultat technique en comparant les frais avec les capitaux réservés à la solvabilité et le taux de rendement du placement financier avec le S/P (un indicateur de rentabilité financier et un indicateur de rentabilité technique).

1.1.1.3 La valeur du contrat

La valeur du contrat est la somme du résultat technique et du résultat financier, il correspond aux résultats probables actualisés techniques et financiers. Cette dualité technique-financier dans l'étude de la rentabilité permet de capter des phénomènes difficilement compréhensibles lors de l'étude seule du résultat technique. Par exemple, un contrat ne dégageant que peu de rentabilité technique peut par sa longévité apporter un résultat financier intéressant.

Cette valeur est la première des briques du calcul de la valeur du client. Il est question de voir la valeur du client comme une agrégation de la valeur de ces contrats qu'il possède à l'instant t.

Il est légitime de ne pas s'arrêter à la vision du client comme possesseur de contrats à l'instant t mais d'intégrer à cela son appétence à en souscrire d'autres dans un futur plus

ou moins lointain.

1.1.2 La valeur potentielle

La valeur potentielle du client, comme dit précédemment, vient ajouter à l'étude du client une vision prospective de sa détention.

Son calcul s'appuie sur celui de la valeur contrat. En effet la valeur potentielle est le produit de la probabilité de souscrire à n produits et de la rentabilité de ces contrats. Un calcul que l'on peut apparenter à celui d'une espérance, ici, l'espérance de la rentabilité future ajoutée.

$$\boxed{Valeur.Potentielle = \sum_{\omega \in \Omega} p_\omega * VC_\omega} \tag{1.11}$$

où p_ω est la probabilité de souscrire à un produit ω, VC_ω la valeur contrat moyenne de ce même produit, Ω l'ensemble des contrats que le client est susceptible de souscrire.

Afin de faire ce travail, des scores d'appétence entre les produits souscrits par le client et les produits qu'il pourrait souscrire sont établis.

Cependant, la connaissance métier et la pertinence des scores font que seuls 5 produits sont envisagés comme étant des produits ayant un potentiel de souscription : l'Automobile, la MRC, la MRH, l'Epargne et la Retraite. Ce scoring a été établi par des modélisations de type gradient boosting [3]. Ces probabilités dépendent des produits détenus par le client et/ou du réseau de distribution.

Dans notre étude, il sera seulement question du produit MRC dont les liens avec les autres produits sont illustrés sur la figure suivante :

3. Méthode d'apprentissage automatisé produisant un modèle de prédiction à partir d'un ensemble de modèles de faible performance.

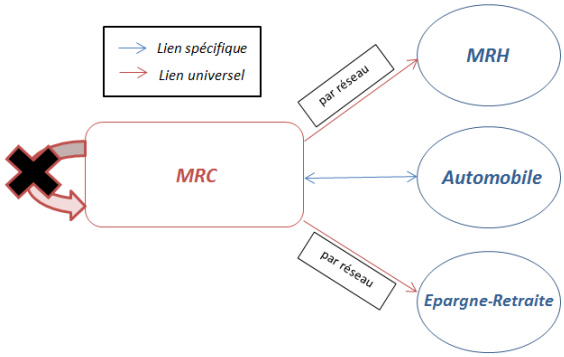


Schéma des liens d'appétence d'un client MRC entre produits

La mention "par réseau" indique une différentiation entre le réseau Salarié et le réseau Agent/Courtier, la mention "lien spécifique" dénote d'une relation d'un scoring d'un produit spécifique à un autre et "lien universel" de n'importe quel produit à un autre.

Il serait possible d'envisager d'autres produits à potentiel de souscription pour les détenteurs de contrat MRC et surtout d'inclure un scoring d'appétence de multi-équipement d'un produit vers lui-même. En effet, il est important de remarquer que les probabilités de souscription à un produit sont calculées pour un produit non souscrit par le client.

1.1.3 La valeur du client et de sa sphère

La valeur du client est composée de la valeur des contrats et de la valeur potentielle précédemment décrite.

La valeur du client permet d'avoir une vision plus étendue de la rentabilité en ne se focalisant pas sur un produit ou une branche isolée mais sur le panel de contrats d'un client. C'est en ce sens que par extension au client la notion de sphère a été créée. On parle alors de la valeur client de la sphère comme étant la somme des valeurs client et des valeurs potentielle de chacun des membres de cette dite sphère. La sphère d'un client est l'entourage proche du client, défini par trois liens :

> le lien entreprise avec la relation : dirigeant de/dirigé par
> le lien filial : est parent de/est enfant de
> le lien matrimonial : est conjoint(e) de

L'idée est que le comportement d'un des membres de la sphère a statistiquement une influence notable sur le comportement de chaque membre qui la compose. En ce sens, il

paraît alors naturel de considérer non seulement l'influence des contrats entre eux d'un client par son profil mais également le profil de la sphère par les profils la composant. Cette étape par la valeur du client est nécessaire pour connaître les clients qui avantage la sphère et ceux qui la desservent.

Le schéma suivant illustre cette idée et donne une version étendue de la notion de sphère :

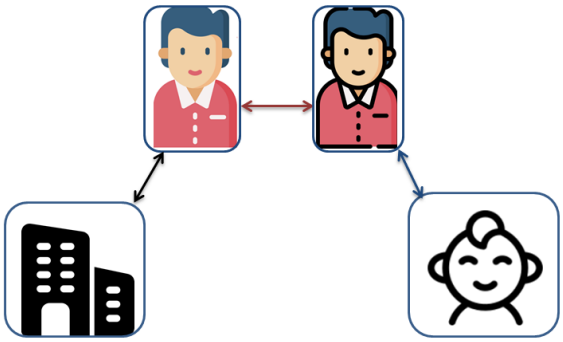


Schéma d'une sphère étendue (trois liens).

Suite à ces considérations, pour rendre ces métriques que sont la valeur client et la valeur potentielle opérationnelles, des clusters ont été faits. Ces derniers permettent de donner un sens plus général et compréhensible au comportement de la sphère. De plus, ils donnent un sens statistique plus fort à ces notions. Cinq clusters ont été faits et sont étiquetés par des étoiles allant de 1 à 5. Ils prennent en compte la valeur client de la sphère et la valeur potentielle de celle-ci. Voici les règles de répartition :

Nombre d'étoiles	Conditions d'attribution
5*	valeur actuelle > 750 avec potentiel de multi-équipement (valeur potentielle > 0)
4*	750 > valeur actuelle > 0 avec potentiel de multi-équipement (valeur potentielle > 0)
	valeur actuelle > 1600 sans potentiel de multi-équipement
3*	1600 > valeur actuelle > 0 sans potentiel de multi-équipement
2*	valeur actuelle < 0 avec potentiel de multi-équipement
1*	valeur actuelle < 0 sans potentiel de multi-équipement

Règles d'attribution des étoiles.

Cette partie théorique sur la valeur du client et de sa sphère étant terminée, intéressons-nous à l'application de ces indicateurs plus en détail.

1.1.4 Applications de la Valeur Client : outil d'ajustement tarifaire, de priorisation de la clientèle et de visualisation du portefeuille

1.1.4.1 Outil d'ajustement tarifaire

La valeur de la sphère et la valeur potentielle permettent de cibler les clients à forte rentabilité future afin de les garder en portefeuille. Ainsi avec cette vision client, il est possible de faire des offres commerciales quitte à dégrader fortement la rentabilité du produit impacté si cela permet d'augmenter la duration du client jusqu'à ce qu'il devienne rentable. Cet indicateur permet d'avoir une vision plus lointaine du parcours du client et d'avoir des outils quant à l'arbitrage à faire entre un manque à gagner à court terme avec une rentabilité améliorée à long terme, et de meilleurs résultats à court terme sans amélioration de sa rentabilité future.

Le cluster dans lequel se trouve le client lui permet d'accéder à différents niveaux de rabais, cela permet de faire de la rétention de "bons" profils et d'inciter au multi-équipement.

En effet, le rabais CVM (Client Value Manager) permet d'adapter l'offre au profil du client. Avant cela, la majorité des intermédiaires (Generali France est entièrement intermédié) accordait 12 % de rabais à tous les nouveaux contrats, le maximum étant de 12 % en moyenne sur le portefeuille client de l'intermédiaire. La distribution des rabais était uniforme ou sans réel arbitrage consistant.

Le CVM permet d'avoir un arbitrage quantitatif et défendable sur le rabais à accorder. L'intermédiaire continue d'avoir une certaine souplesse pendant la négociation du contrat, avec une enveloppe appelée latitude commerciale.

Mais plus que cela, la valeur client permet également de mettre en place des boucliers de revalorisation (système permettant de protéger les profils à fort potentiel et/ou à forte rentabilité de trop importantes revalorisations), de proposer de fortes revalorisations aux clients jugés déficitaires et de geler la défense de contrats des clients dits indéfendables [4].

La valeur client permet d'ajuster le tarif d'un contrat en fonction du profil du client, comme une couche supplémentaire d'étude tarifaire avec un argument prospectif.

4. Clients fortement destructeurs de valeur.

1.1.4.2 Outil de sélection de profils client

La valeur client est aussi précieuse dans la restructuration du portefeuille. En effet les rabais et autres offres commerciales "discriminent" les clients et de ce fait, à long terme, restructure le portefeuille augmentant les profils jugés rentables ou à forte rentabilité future.

Ce type de raisonnement est essentiel compte tenu du fonctionnement économique d'une assurance où l'on sait à l'avance les recettes mais pas les dépenses. Une meilleure mutualisation se fera si l'on peut influer sur les données du risque sans altérer les garanties ou la couverture.

1.1.4.3 Outil de visualisation de portefeuille

On peut également voir la valeur client comme un outil de visualisation du risque et de la rentabilité du portefeuille à un instant fixé ou de manière prospective. Durant des projets de reportings ou de bilans, elle peut être pertinente pour avoir une vision sous la maille client d'indicateurs clés. Par exemple, il est possible de prendre comme indicateur de rentabilité technique la proportion de 4-5 étoiles (clients à forte valeur) du portefeuille ou la diminution de profils destructeurs de valeurs (à valeur négative) pendant une certaine durée.

1.2 Utilité de l'étude

1.2.1 Projet de refonte de la Valeur MRC : explications et objectifs

Le calcul de la Valeur Client va être étudié à nouveau en profondeur : cette étude est fortement corrélée à de nombreux travaux faits dans l'équipe. On pourra citer le projet de transparence de la Valeur pour que cette dernière soit mieux interprétable pour les autres équipes comme le marketing, la distribution (surtout les agents) qui ont à disposition en plus en plus d'outils quantitatifs. Mais également pour les équipes de tarification et les équipes produits dont les résiliations ou les fortes revalorisations se basent au moins en partie sur la valeur client.

De plus, de nouvelles applications de la valeur concernant d'autres offres commerciales sont étudiées comme la priorisation de clients vis-à-vis de l'indemnisation ou de certains services comme le dépannage pour l'Auto.

Plus spécifiquement, le calcul de la valeur de la MRC est perfectible et est souvent mis en cause, mais en cette période de nouvelle tarification du produit MRC avec la mise en place de nouveaux rabais liés à ce nouveau tarif, la valeur client MRC doit être repensée en profondeur.

Un objectif sur le long terme consiste à implémenter le calcul de la valeur en temps réel, en effet pour l'instant ce calcul est fait à chaque mensualité. Elle est implémentée sur SAS Enterprise Guide et se base sur des tables SAS dans lesquelles sont descendus toutes les variables et indicateurs nécessaires à ce calcul qui sont mis à disposition mensuellement.

Ce travail serait utile pour ajouter une actualisation plus rapide de la valeur des clients, pour mieux comprendre et influer sur le multi-équipement simultané (deux contrats signés à la souscription par exemple). En cas de la concrétisation de deux contrats dans l'intervalle d'un mois, le deuxième contrat sera tarifé sur la valeur du client avant le premier contrat. Cette étude sera effectuée dans l'esprit du calcul en temps réel de la valeur c'est-à-dire avec des données disponibles presque instantanément et à n'importe quel moment.

Nous pouvons également discuter des "packs" de produits mis en place, l'Auto-MRH et l'Auto-MRH-GAV afin de favoriser le multi-équipement. Ces packs complétés accordent un rabais conséquent sur les produits le composant. Il pourra être utile d'avoir à disposition une valeur client revue récemment, à jour et plus précise dans le cas de l'étude de pack comportant de la MRC.

1.2.2 Le Multi-Risque Commerce : un produit clé

Le contrat MRC est destiné aux professionnels et entreprises qui souhaitent assurer leur commerce : leur bien, la perte de leur exploitation ainsi qu'une garantie responsabilité civile. Les clients susceptibles de souscrire un tel contrat sont de tous secteurs d'activités avec des tarifs différenciés. Il est distribué par deux réseaux de distribution que sont les agents et les courtiers.
Les chiffres et graphiques présentés ont tous une vision datant de fin 2019 avec des données de l'année 2019.

Les produits proposés sont les suivants, avec la répartition du nombre de contrats et de montant de prime :

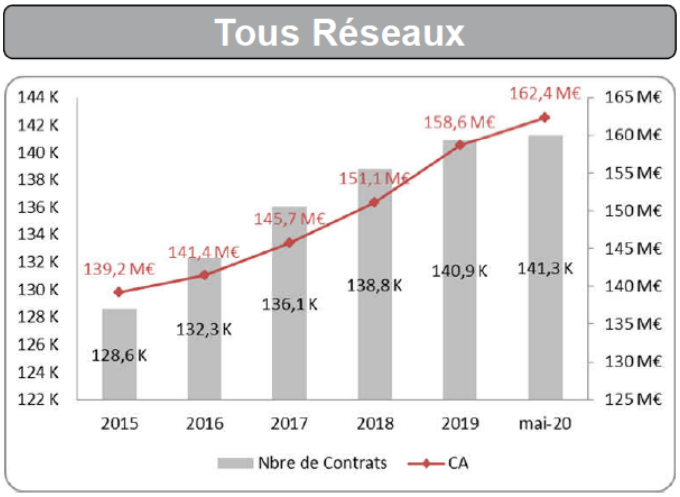
Répartition du nombre de contrats et de montant de prime

En M€

Chiffre d'affaires	170,9	**Réassurance**	-11,5
Courant	171,1	Primes acquises cédées	-21,0
Antérieurs	-0,2	Sinistres cédés	5,3
Acceptations	0,0	Commissions de cession	4,2
Primes acquises brutes	170,1	**Primes acquises Nettes**	149,1
Sinistralité courante brute - hors EE	-103,0	**Sinistralité courante nette - hors EE**	-91,4
Sinistralité antérieure brut - hors EE	8,9	**Sinistralité antérieure nette - hors EE**	2,6
Commissions brutes	-40,1	**Commissions nettes**	-35,9
Frais généraux bruts	-20,2	**Frais généraux nets**	-20,2
		CoR Courant Hors EE	1,5
		CoR Net Hors EE	4,1
		EE M€	-8,6
		CoR Net yc EE	-4,5

Chiffres clés du produit MRC vision fin 2019 où EE signifie Évènements Exceptionnels

Malgré son relatif faible nombre de contrats (seulement 140 000 avec 132 000 clients contre 770000 contrats environ en Auto), le produit MRC représente une important partie de l'assiette des primes en Non-Vie avec des primes annuelles moyennes d'environ 1100 euros. Le produit représente près de 10% du chiffre d'affaire de Generali France et est composé en très grande majorité de professionnels et de petites entreprises.

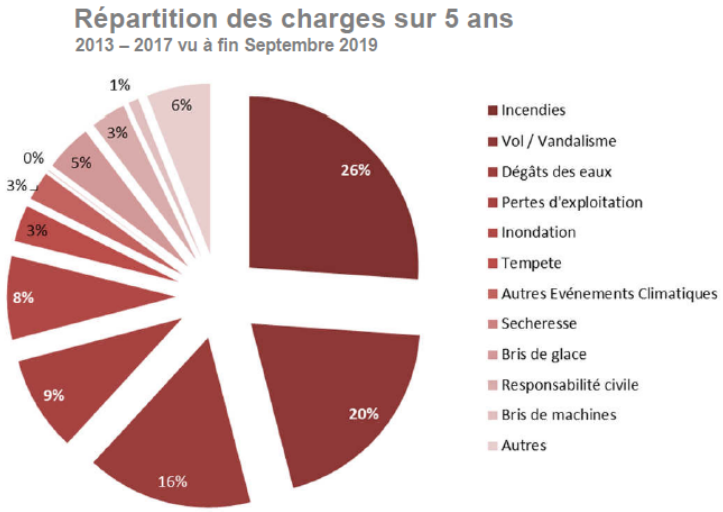


Évolution du chiffre d'affaire et du nombre de contrats

M€	T4 2019	P 2020 v19	F2 2020 Hors COVID	F2 2020 yc COVID
Mono véhicule	396,7	402,8	403,4	387,2
MRH	320,8	316,5	322,1	315,6
Particuliers	717,5	719,3	725,5	702,8
MRC	170,9	171,2	174,7	142,9
MRI	111,9	115,2	119,7	116,4
Risque Agricole	37,2	37,3	41,6	40,9
RC Générale	102,0	102,2	104,1	94,9
Flottes	99,6	101,0	102,2	98,6
RI	125,5	126,7	133,4	128,9
Construction	59,9	68,1	66,4	58,9
Pro de l'auto	34,1	33,9	33,7	31,3
RT	21,3	22,3	23,6	22,2
Transport	43,1	44,1	47,1	43,8
Pro PE & Ent.	805,4	822,1	846,3	778,7
Total IARD	1 522,9	1 541,4	1 571,8	1 481,5

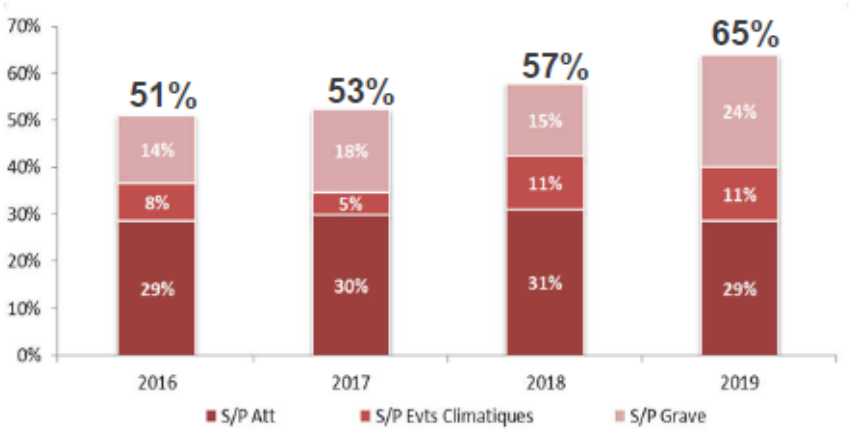
Répartition du chiffre d'affaire des produits IARD

Diamétralement opposé à cela, il sera présenté quelques informations sur la sinistralité et le rentabilité du produit MRC :



Répartition du montant des charges par garantie

On appelle ratio technique le ratio suivant Sinistres/Primes noté S/P, il faut encore ajouter les frais et commissions pour obtenir le ratio combiné ou encore le COR qui rend véritablement compte de la perte ou du gain du produit dans la comparaison à 1.



Résultats du ratio technique sur les 4 dernières années

On remarque que les sinistres dits Graves dégrade beaucoup la rentabilité.

%	T4 2019	P 2020 v19	F2 2020 Hors COVID	F2 2020 yc COVID
Mono véhicule	99,4%	99,7%	97,0%	91,2%
MRH	98,2%	97,3%	98,3%	95,8%
Particuliers	98,9%	98,7%	97,6%	93,2%
MRC	104,7%	99,7%	99,5%	136,0%
MRI	100,1%	93,3%	88,7%	88,0%
Risque Agricole	111,7%	94,9%	91,0%	93,9%
RC Générale	88,3%	97,3%	90,8%	95,6%
Flottes	102,1%	102,2%	102,3%	96,9%
RI	108,0%	98,2%	120,2%	119,2%
Construction	113,0%	115,2%	109,7%	123,5%
Pro de l'auto	103,9%	103,8%	97,3%	99,8%
RT	82,3%	73,9%	78,5%	84,3%
Transport	96,7%	102,6%	103,7%	105,1%
Pro PE & Ent.	101,7%	99,1%	100,0%	107,6%
Total IARD	100,4%	98,9%	98,9%	100,6%

Résultats du COR net courant comprenant les évènements exceptionnels depuis 2019

Il est à noter que la crise sanitaire puis économique a fortement dégradé le ratio combiné du produit MRC à cause notamment du non paiement des primes de nombre de clients et d'un plan de sauvegarde des clients a été mise en place avec une enveloppe conséquente pour diminuer les primes des assurés pendant ces temps d'instabilité économique.

Ces chiffres sur la rentabilité sont à prendre avec des précautions puisque l'année 2020 est une année exceptionnelle et que ce plan de sauvegarde permet de garder en portefeuille les clients.

Concernant la rentabilité avant la crise sanitaire, elle est à relativiser par le fait que le produit MRC est un **produit d'appel** pour les professionnels et petites entreprises qui ont une forte propension à se multi-équiper. De ce fait, sa rentabilité se voit améliorée par une plus importante duration (plus de contrats par client implique une duration plus importante) et par le gain porté par les contrats souscrits sur d'autres produits que la MRC a permis de concrétiser.

Étudions ce caractère de produit d'appel et de multi-équipement.

Produit d'entrée	Répartition
MRC	75%
AUTO	11%
RC	8%
MRH	2%

Répartition du produit d'entrée des clients MRC

On remarque que les clients MRC entrent dans le portefeuille avec un contrat MRC, il s'agit donc d'un produit d'appel compte tenu du fait que le nombre moyen de contrats par client est de 2,2 et que 43 % des clients sont multi-équipés.

La proportion de résiliation à un an quant à elle de 8,2 % sur les contrats de plus d'un an et sur les contrats de moins d'un an, elle est de 22%.

En résumé, l'importance de la MRC réside dans son apport significatif dans le chiffre d'affaire et dans son potentiel à multi-équiper ses clients, elle est la porte d'entrée des professionnels et petites entreprises de Generali France.

1.2.3 L'apport de la nouvelle modélisation de la Valeur

La modélisation des résiliations permettra de mettre à jour les lois de chute du contrat MRC avec un approche nouvelle et d'en changer la maille. Il est primordial d'actualiser ces probabilités compte tenu des changements liés au portefeuille et au contexte concurrentiel. Cette modélisation se fera de manière indépendante de l'ancienne. Les probabilités de résiliations annuelles interviennent dans le calcul de la valeur des contrats et le calcul de la duration du contrat.

En fusionnant avec les études récentes effectuées dans le service sur l'élasticité au prix, l'étude permettra de mieux comprendre la rétention de son portefeuille et donc de mieux adapter les rabais et offres commerciales des nouveaux clients. Il pourra également être le support de réflexion quant aux moyens de restructuration du portefeuille par des profils plus appropriés au produit MRC de Generali France. En effet mieux comprendre les intérêts du prospect permettra de mieux sélectionner les profils par le biais de actions commerciales.

2 Pré-requis aux modélisations : structuration de bases de données viables

2.1 Périmètre de l'étude

L'étude se limite aux clients professionnels et aux petites entreprises détenant au moins un contrat MRC où le calcul de la valeur client est perfectible.

Bien entendu, il y a plusieurs produits dans la branche MRC. On peut dès à présent séparer ces produits en deux catégories : ceux qui ne sont plus commercialisés (en "run-off") c'est-à-dire que l'on laisse le stock de ces contrats s'écouler en résiliation sans faire de conquête dessus et ceux qui sont commercialisables. Ces derniers représentent environ 75 % du stock des contrats MRC. Notre étude se désintéressera des produits désuets dont la résiliation est parfois trompeuse (ils sont invités à résilier leur contrats pour prendre un produit plus récent) ce qui ajouterait un biais trop important à notre travail.

Les produits étudiés seront les suivants : *"100 % PRO ACPS","100% PRO ARTISANS-COMMERCANTS", "100% PRO ASSOCIATION", "100% PRO FABRICATION", "100% PRO SERVICE","ANCIENNE GENERATION".*

On remarque que les produits sont différenciés selon le secteur d'activité de l'assuré qui est généralement la variable la plus discriminante quant au risque lié au contrat, de plus la différentiation par le secteur d'activité permet de proposer des garanties plus spécifiques à l'assuré compte tenu de son activité. Ce raisonnement poussé à l'extrême pourrait conduire à une hyper-segmentation des clients donc à dégrader la mutualisation des sinistres par produit.

2.2 Structuration de la base résiliation

La base nécessaire au travail développé ici a été construite à partir des contrats existants dans le portefeuille au 31/12/2016 afin de modéliser l'état du contrat à la fin de l'année 2017. L'année 2017 a été choisie puisque des imprécisions datant de cette année ont été soulevées. Pour ne pas se baser uniquement sur cette année qui peut être singulière vis-à-vis de la concurrence, du contexte économique, une base 2018 a été faite de la même

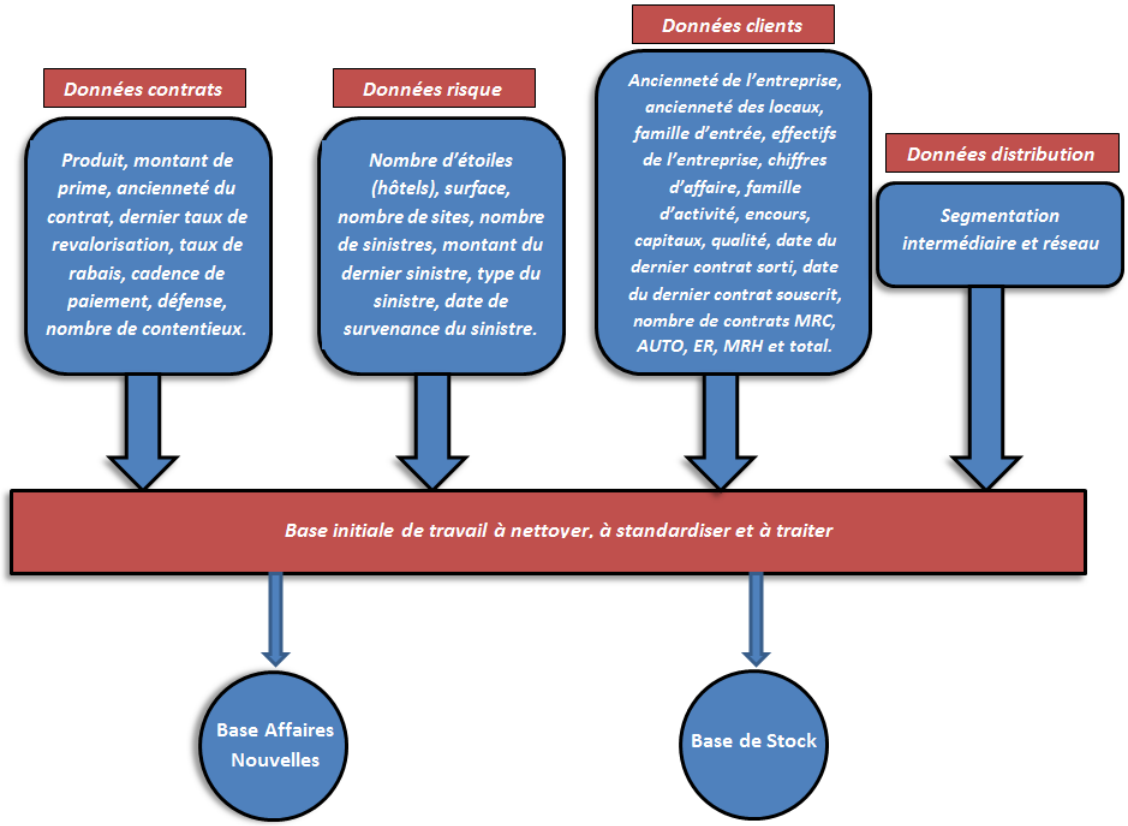
manière et pourra être utilisée pour prouver la non-singularité de cette année.

En effet, les lois de chute sont des n-uplets de probabilités de résiliation à un an conditionnées à leur présence dans le portefeuille au dernier jour de l'année écoulée.

De ce fait, la base a été construite pour avoir des probabilités de résiliations à la fin de cette année 2017 et le conditionnement à l'année de présence sera contenu dans la variable "ancienneté du contrat".

La base a été conçue en essayant avant toute modélisation de considérer le plus de variables explicatives possibles et donc des jointures entre des données contrats, clients, risques, distributions ainsi que des données externes ont été faites.

Par souci de clarté, il sera illustré ci-dessus l'ensemble des variables considérées.



La base est donc composée de variables qualitatives et quantitatives et est comporte des Affaires Nouvelles et des contrats en stock dans le portefeuille.

Elle a donc été subdivisée pour séparer les nouveaux contrats des plus anciens puisque l'on ne dispose pas des mêmes informations pour ces contrats. En effet, les variables historiques n'existent pas pour les nouveaux contrats. De plus, leur comportement est sensiblement

différent [5].

La prudence s'impose dans la construction de bases à des fins de modélisation prédictive dans le sens où toutes les variables se doivent de dater de la fin de l'année 2016, sinon des leakages peuvent être observés.

Un premier nettoyage des données, une fois ces données jointes et fiabilisées, est fait. Les valeurs manquantes de variables a priori significatives (de manière statistique) ont été enlevées de la base si elles étaient très minoritaires et sinon elles ont été remplacées par la moyenne ou la modalité la plus représentée quand la variance de la variable était raisonnable. Si la variance était trop importante, les valeurs manquantes ont été remplacées par les réalisations d'une variable aléatoire distribuée comme la variable censurée des valeurs manquantes.

Cette dernière méthode permet d'être plus précis que d'assigner la modalité la plus représentée à toutes les variables manquantes.

Les outliers ont été également enlevés de la base, sauf si la variable a été segmentée.

Ce nettoyage assez radical quant à la suppression de contrats est possible du fait du peu de défauts de la base créée. En effet, lors de la sélection des variables par connaissance métier, a été pris en compte le taux de remplissage de la variable. Une variable étant fournies à 30% ne sera pas retenue car inexploitable puisqu'elle dépendra trop fortement de son imputation pour les valeurs manquantes.

Une chose est également à retenir : la disponibilité des variables de la base de données créée est un critère important. En effet une variable, même supposée très explicative se doit d'être disponible facilement pour rendre la modélisation opérationnelle.

Puis, vient ensuite la standardisation des variables c'est-à-dire respecter les formats pour les modélisations futures, la fiabiliser par la comparaison d'études descriptives de certaines variables de la base par rapport aux bases originelles disponibles sur SAS Guide Enterprise.

Les données sur le contrat impactent a priori la résiliation. Par exemple le dernier taux de majoration (s'il est important), le nombre de contentieux ou la cadence de paiement encouragent a priori la résiliation du client selon les statistiques descriptives effectuées. A

5. Une plus forte réalisation est observée à la fin de la première année de contrats. C'est le premier moment où les clients peuvent résilier facilemment (loi Hamon 2014)

contrario, la défense du contrat, le taux de rabais et l'ancienneté du contrat l'inhibent. Les variables comme le produit et la plupart des données risques permettent d'estimer au mieux la sinistralité, et font partie des variables tarifantes du produit.

Les variables clients permettent d'essayer de comprendre l'environnement à la fois contractuel du client (tendance d'équipement ou de dés-équipement), économique : le secteur d'activité, son encours, sa qualité (Propriétaire/Locataire),...

Enfin, les données sur le réseau de distribution permettent d'ajouter un aspect commercial et gestion de contrat de manière concrète à la base de données. En effet le type de réseau influe fortement sur la concrétisation et sur la résiliation, puisque la façon d'appréhender le client est différente.

Generali est une entreprise dont les produits sont totalement distribués par des intermédiaires et ne fait pas de vente directe, donc le type de relation entre le client et l'intermédiaire est à prendre en considération.

La segmentation de l'agent est, elle, basée sur l'efficacité et la performance de l'intermédiaire prenant en compte sa concrétisation, la duration de ses contrats et la rentabilité marginale qu'il dégage. Cette variable a trois modalités.

Par exemple, une information coûteuse comme le "credit score" n'apparaîtra pas dans cette base.

Plus d'informations sur la santé financière des clients aurait été bénéfique pour le risque d'impayé ou sur la résiliation pour disparition du risque. Pour essayer d'induire ce genre de considération, l'ancienneté de l'entreprise, ses locaux, son effectif, son chiffre d'affaire et ses capitaux (ou encore le nombre d'étoiles pour les hôtels) mis en jeu dans le fond de commerce ont été représentés dans la base.

Le grand nombre de contrats dans les deux bases créées (Affaires Nouvelles 22 000 contrats et Stock 108 000 contrats) me permet d'avoir une certaine souplesse vis-à-vis de la modélisation, comme par exemple envisager le Machine Learning.

3 Modélisation de la résiliation

3.1 Objectifs de l'étude

L'objectif de cette partie est de modéliser la probabilité qu'a un client de résilier son contrat Multi-Risque Commerce durant l'année. Cette probabilité sera calculée pour tous les clients professionnels présents en 2017. L'étude a pour résultat l'élaboration de lois de chute des contrats MRC en fonction de certaines variables explicatives afin de réduire la variance des probabilités de résiliations calculées. En effet, le but de la modélisation est de donner une valeur moins agrégée que celle qu'un taux de résiliation. Avec un ensemble de probabilité très important, les probabilités sont différenciées en fonction des profils et la modélisation permet aussi de connaître les variables impactant la résiliation ou non d'un contrat.

Pour rappel, une loi de chute sera considérée ici comme un n-uplet de probabilité conditionnelle de survie du contrat avec n étant la durée de la projection en année. Plus concrètement, il s'agira d'une table renseignant la probabilité conditionnelle de résiliation sous un an par maille et par année d'ancienneté.

Un travail concernant l'ajustement de ces probabilités par une loi paramétrique sera réalisé pour obtenir des lois de chute lissées, continues et facilement projetables dans un futur plus lointain tout en restant cohérent avec la réalité de la durée maximale d'un contrat.

Ces lois de chute, comme déjà mentionné auparavant, influeront sur la valeur des contrats et donc sur la valeur potentielle et actuelle du client et également sur les possibles actions à entreprendre avant une possible résiliation.

Durant l'élaboration de ce mémoire, une possible application a été envisagée et est intimement liée à une valeur en temps réel. Elle serait de créer une alerte pour les clients à forte valeur lorsque leur probabilité de résiliation augmente à cause d'un changement de situation, de caractéristique ou d'une nouvelle action. Cette alerte serait être accompagnée d'une proposition d'action permettant de diminuer cette probabilité comme un rabais, un incitation à un nouveau produit ou encore un changement sur son contrat (une garantie supplémentaire, un contrat plus adapté à sa nouvelle situation).

3.2 Hypothèses des modèles

L'objectif est donc de résoudre un problème de classification binaire : la résiliation ou non du contrat sur une année. La modélisation de cet état nous présentera des probabilités de résiliation qui ,une fois croisées avec l'ancienneté du contrat, permettront de créer les lois de chute. En effet, les probabilités sont fortement impactées par l'ancienneté du contrat et donc seront discrétisées par cette variable.

Une autre étude aurait pu être faite en étudiant les mêmes contrats suivis sur plusieurs années avec un modèle de durée où le temps d'arrêt serait la résiliation du contrat. Cependant, prendre des contrats assez anciens afin de les suivre durant de nombreuses années pour avoir le recul nécessaire à l'élaboration de lois de chute, ne serait pas forcément pertinent à cause du contexte économique changeant et de l'évolution des tarifs du produit étudié. La question de l'accessibilité et de la qualité des données anciennes pose également problème.

Afin de rester le plus actuel possible, l'option de prendre en compte de l'ancienneté du contrat comme une variable explicative dans un modèle sur une année a été choisie.

De plus, par souci de fiabilité, la modélisation a été déployée sur des données de l'année 2018 également afin de vérifier que les résultats étaient similaires. Ce travail permet d'écarter une possible singularité de l'année. Avec un exemple extrême, prendre l'année 2020 comme année d'étude de la résiliation pour prédire le futur de la résiliation serait très malvenu considérant le caractère exceptionnel du climat économique de cette année.

3.3 La régression logistique : Théorie et Applications

3.3.0.1 Réflexion sur le modèle

Il sera discuté des modèles linéaires généralisés (GLM) et en particulier de la régression logistique.

Le premier argument de ce choix est l'explicabilité de ce modèle. En effet, le GLM n'est pas souvent le modèle le plus performant, surtout dans des problèmes de classification binaire comme celui-ci, mais il nous donne des résultats tout à fait explicables quand le travail sur les variables explicatives est bien mené. Il nous donne l'impact et l'importance

de chaque variable et/ou de chaque modalité des variables catégorielles dans l'estimation de la variable à expliquer. Ce fait permet de mieux expliquer la variable réponse et ses corrélations ; de plus cela permet de déployer rapidement et facilement le modèle sur des données où la variable réponse doit être prédite. Ces raisons font du GLM une solution perfectible mais aisément opérationnelle dans l'étude que nous menons. En effet, ceci est en lien avec le projet d'actualisation de la valeur en temps réel et de l'adaptation de la probabilité de résiliation en instantanée.

D'autres types de modèles, malgré une meilleure prédiction, ne donnent pas autant d'interprétabilité sur leur estimation et donnent des scores de sortie où l'utilisateur se doit de faire aveuglément confiance en son modèle avec généralement une adaptation au temps réel moins aisée. Le vrai gain se situe également dans les estimations d'incertitudes que les modèles statistiques peuvent donner et que les modèles type Machine Learning ne peuvent donner qu'après un travail conséquent.

3.3.0.2 Le modèle linéaire classique

Les modalités de la variable réponse (0 ou 1) nous conduisent logiquement vers une approche par régression logistique dans un premier temps.

Les familles exponentielles apparaissent de façon naturelle dans la recherche de distributions lors d'applications statistiques ; en effet, elle comprend de nombreuses distributions usuelles : normale, exponentielle, gamma, chi-carré, bêta, Bernoulli, Poisson, etc.

Cette famille est "exponentielle" lorsque la fonction de densité prend une forme algébrique particulière entre la variable aléatoire et les paramètres : la séparation des facteurs.

Cette famille se présente sous la forme suivante :

$$\boxed{f(y|\theta,\phi) = exp(\frac{y*\theta - b(\theta)}{a(\phi)} + c(y,\phi))} \tag{3.1}$$

où $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont des fonctions, et θ est appelé paramètre naturel. Le paramètre θ est le paramètre dit d'intérêt et ϕ est considéré comme un paramètre de nuisance.

Nous retrouvons facilement les expressions suivantes pour l'espérance et la variance (pour

b deux fois dérivable) :

$$\boxed{E(Y) = b'(\theta)} \qquad \boxed{V(Y) = b''(\theta) * \phi} \tag{3.2}$$

Une distribution de cette famille est la loi de Bernoulli pour laquelle la variable aléatoire est réduite à deux valeurs (0 ou 1) avec comme paramètre p, la probabilité de la résiliation de l'évènement (1) ici la résiliation du contrat dont la fonction de masse est la suivante :

$$\boxed{f(y|\theta,\phi) = p^y * (1-p)^{1-y} = exp(y * ln(\frac{p}{1-p}) - ln(\frac{1}{1-p}))} \tag{3.3}$$

Par identification, il vient : a(ϕ)=1, θ = ln($\frac{p}{1-p}$) et c(θ,ϕ)=0.

On appelle fonction canonique g d'une distribution de la famille exponentielle, la fonction qui vérifie la relation suivante : g(E(Y))=θ ici g(p)=ln($\frac{p}{1-p}$). La fonction canonique de la loi Bernoulli est la fonction de [0,1] dans **R** suivante :

$$\boxed{g(x) = ln(\frac{x}{1-x})} \tag{3.4}$$

On appelle cette fonction la fonction logit.

Ce lien entre l'espérance de Y et le paramètre de la famille exponentielle permet de considérer comme fonction "lien" par la suite l'adoption de la fonction logit dont on va expliciter dans la partie suivante la définition ainsi que l'utilité.

En effet, par la suite, il sera question de supposer un lien entre certaines variables dites explicatives et une variable à expliquer. La fonction lien sera présente pour faire le lien entre variable explicative et variable réponse.

3.3.1 Le Modèle Linéaire Généralisé : régression logistique

Avant d'expliquer en quoi consiste le modèle linéaire généralisé, il serait pertinent d'expliciter une version particulière : le modèle linéaire.

Pour cette description, notons Y la variable à expliquer, X=$(Xi)_{i\in[\![1;n]\!]}$ les variables

explicatives.

Ce dernier s'appuie sur des hypothèses assez fortes sur les variables explicatives ainsi que réponse.

On fait l'hypothèse de l'exactitude des réalisations des X_i avec Var($X_{i,j}$)=0, $\forall i \in [\![1;n]\!]$ et $\forall j$ avec j>i c'est-à-dire plus d'observations que de variables explicatives. Naturellement, X doit être de rang plein (hypothèse de non-coliénarité des X_i).

Le modèle est bien spécifié donc E(ϵ)=0.

L'hypothèse d'homoscédasticité de la variable dépendante Y qui est définie par $Var(Y) = \sigma^2$=constante. Également, on admet que les Y_i ne sont pas corrélés, les réalisations de la variable dépendante n'influent pas sur les suivantes.

Les paramètres à estimer sont le fond de cette modélisation : $\beta = (\beta_i)_{i \in [\![1;n]\!]}$.

$$\boxed{Y = X * \beta + \epsilon} \tag{3.5}$$

où ϵ est le vecteur de erreurs, n réalisations indépendantes,centrés et de variance constante.

Les paramètres sont estimés par la méthode des moindres carrés visant à minimiser la somme des carrés des résidus c'est-à dire que l'on minimise leur dispersion.
De manière plus précise : $\widehat{\beta} = argmin_\beta(||Y - X\beta||_2)$.

Ceci revient à résoudre n+1 équations : $\frac{\partial(\Sigma\epsilon^2)}{\partial\beta i}$=0. Avec un peu de travail, on conclut que $\widehat{\beta} = (X^T X)^{-1}(X^T Y)$ où X^T est la transposée de X.

Nous pouvons conclure que cet estimateur des moindres carrés ordinaires sous les hypothèses citées est le meilleur estimateur linéaire sans biais.

Les problématiques resteront les mêmes pour la partie suivante, notons entre autres celles-ci : les variables explicatives à considérer (sera étudié par la suite), l'estimation des β, leur précision, quantifier le pouvoir explicatif du modèle, quantifier l'influence de chaque variable la prédiction et son importance et également déceler les observations atypiques du jeu de données.

3.3.1.1 Théorie sur le modèle linéaire généralisé

Cette introduction ainsi faite nous permet de discuter du modèle linéaire généralisé. Ce modèle peut être utilisé avec n'importe quelle loi de la famille exponentielle précédemment présentée en supposant une relation entre Y|X et les variables explicatives X.

Le modèle se fonde sur trois piliers :

* La variable à expliquer, Y, de famille exponentielle.
* Le vecteur des variables explicatives X formant une famille libre.
* La fonction lien notée g.

Cette relation fait intervenir une fonction appelée fonction lien telle que :

$$\boxed{g(E(Y|X)) = X * \beta = \beta_0 + X_1 * \beta_1 + ... + X_n * \beta_n} \tag{3.6}$$

Avec g étant une fonction bijective de [0,1] dans R (la bijectivité est nécessaire pour assurer l'existence et l'unicité et pour pouvoir retrouver $E(Y/X)=g^{-1}(X\beta)$*).*

Rappelons également une hypothèse récurrente ici aussi : l'indépendance conditionnelle des résidus.

Reste maintenant à choisir une fonction lien. Ici la forme linéaire des paramètres bêta n'est pas supposée égale à la variable réponse mais à sa transformation par g, en cela le choix de cette dernière doit être pensé.

Dans la modélisation nous concernant, la variable à expliquer est binaire et sa loi de type exponentielle est la loi de Bernoulli.

Compte tenu ce qui a été expliqué précédemment, la fonction canonique logit semble être un candidat sérieux pour g. Le choix a été fait de prendre cette fonction comme étant la fonction lien de notre GLM (Generalized Linear Model). On parle alors de régression logistique lorsque le GLM repose sur une loi de Bernoulli et une fonction logit comme lien.

Nous pouvons réécrire la relation principale du modèle de la sorte :

$$\boxed{ln(\frac{p}{1-p}) = X * \beta = \beta_0 + X_1 * \beta_1 + ... + X_n * \beta_n} \tag{3.7}$$

Avec p le paramètre de la loi de Bernoulli, p ∈]0,1[.

Estimation des coefficients β : Dans le modèle linéaire classique, cette estimation était faite par la minimisation de la somme des carrés des résidus. Ici, la méthode est celle du maximum de vraisemblance.

En inférence bayésienne, la vraisemblance peut être vue comme la densité de probabilité des données conditionnellement à une valeur des paramètres (ici une variable aléatoire).On peut également l'interpréter comme état un indicateur quantitatif de l'information apportée par les données sur la valeur des paramètres Sa maximisation est donc logique.

Elle se présente sous la forme suivante pour une distribution de famille exponentielle :

$$\boxed{L(\theta, \phi|Y) = \prod_1^n f(y_i|\theta_i, \phi) = exp(\sum_{i=1}^n \frac{y_i * \theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi))} \tag{3.8}$$

On étudie généralement la log-vraisemblance par souci de simplification car un maximum de la vraisemblance est aussi un maximum de la log-vraisemblance et inversement (la fonction ln étant bijective).

Il faut alors résoudre les équations suivantes :

$$\boxed{\sum_i \frac{\partial log(L_i)}{\partial \beta_j} = 0} \tag{3.9}$$

Pour la i^{me} réalisation et le j^{me} paramètre β_j

Il vient :

$$\boxed{\forall j, \sum_i \frac{\partial \mu_i}{\partial \eta_j} * \frac{y_i - \mu_i}{V(Y_i)} * X_{i,j} = 0} \tag{3.10}$$

$$Avec\ \mu_i = E(Y_i),\ \eta_i = X_i\beta.$$

Ces équations pourront être numériquement résolues avec une méthode de descente de gradient.

3.3.1.2 Sélection de variables par différentes méthodes

La première étape à l'élaboration de modèle se situe dans la standardisation des données au modèle étudié (nettoyage, formatage,...). Cette étape a été traité précédemment. Après cela et avoir présenté le modèle, il nous faut disposer des variables explicatives mentionnées plus haut et les sélectionner pour garder l'explicabilité du modèle et même améliorer sa performance. C'est le critère de parcimonie des variables explicatives. Ces trois types de sélection de variables ont été étudiés.

Méthode Stepwise : la régression pas à pas

Il s'agit d'une procédure de sélection de variables qui est une amélioration de la méthode ascendante, qui examine un modèle avec une seule variable explicative puis ajoute une à une d'autres variables explicatives.

La procédure Stepwise est l'une des plus utilisée tant elle est un séduisant compromis entre la procédure exhaustive (la meilleure théoriquement) et la méthode ascendante (moins performante).

Cette dernière consiste à chaque étape à réexaminer les variables introduites et de ce fait, une variable significative peut ne plus l'être dans l'étape suivante. Cette versatilité s'explique par les corrélations existant après coup avec les autres variables ajoutées.

Dans les faits, la procédure stepwise introduit une nouvelle variable et évalue une nouvelle fois les tests de Student des variables déjà introduites et sélectionne la variable la moins significative parmi celles qui ne le sont plus, et la retire. La procédure se termine une fois qu'aucune variable n'est plus ni ajoutée ni enlevée.

A chaque itération, l'indicateur de performance est évalué ; ici il s'agit de l'AIC [6] que l'on doit minimiser.

Il est à noter que les résultats peuvent s'avérer trompeurs lorsqu'il existe une colinéarité forte entre variables explicatives car l'une peut masquer le pouvoir explicatif de l'autre.

Cette procédure ne donne pas le meilleur modèle possible dans l'absolu mais permet d'obtenir un résultat acceptable avec une notion de parcimonie. Cependant, l'utilisation même du GLM,ici, n'apporte pas la meilleure modélisation et préfère composer avec l'explicabilité, la transparence et le souplesse.

Sélection LASSO [7]

La régression LASSO introduit un terme de pénalisation dans la fonction objectif (la fonction à minimiser ou maximiser) dépendant des paramètres β du modèle dans un soucis de parcimonie du modèle. Ceci ajoute un biais dans le modèle mais permet la réduction de la variance de celui-ci.

LASSO attribut un paramètre $\widehat{\beta_i}$ nul à la variable à éliminer afin de ne garder que les variables plus explicatives à la variable réponse.

La fonction objectif est donc la somme des carrés des résidus pour la régression linéaire et pour une régression logistique la vraisemblance du modèle. Dans ces deux cas lors de la méthode LASSO, sera ajoutée une pénalisation de norme l_1 [8].

L'expression des paramètres β est la suivante :

$$\boxed{\widehat{\beta_{MV}^{Lasso}} = argmax_\beta(ln(L) - \lambda \sum_{i=1}^{n} |\beta_i|)} \tag{3.11}$$

où $\lambda > 0$ contrôle la puissance de la régularisation.

Il faut choisir le paramètre λ entre 0 qui correspond à une régression non pénalisée et ∞

6. Akaike's Information Criteria, égal à 2k -2ln(L) où L est la vraisemblance du modèle et k le nombre de paramètres. Cet indicateur sera décrit et utilisé par la suite.

7. Least Absolute Shrinkage and Selection Operator

8. Il s'agit d'un terme positif pour un modèle linéaire et négatif pour une régression logistique puisqu'il faut respectivement minimiser et maximiser la fonction objectif dans ces deux cas

provoquant la nullité de tous les coefficients. Plus λ est élevé, plus la pénalisation est forte et donc plus le modèle est parcimonieux.

Nous comprenons ainsi que la valeur attribuée au paramètre λ est très importante. Ainsi nous expliquons brièvement la méthode utilisée dans l'étude pour la déterminer.

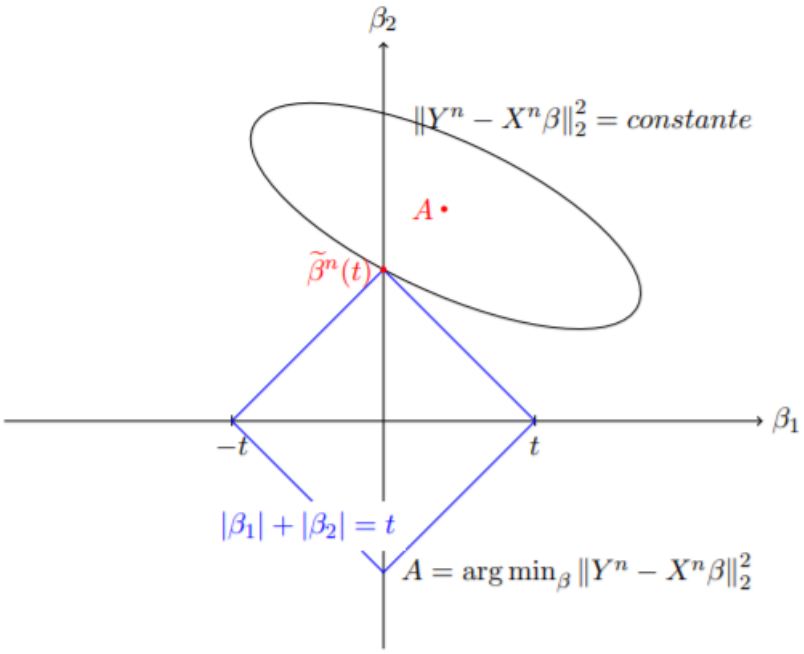
L'objectif étant de prendre une valeur minimisant l'erreur de validation croisée, la déviance, entre parcourant des valeurs comprises entre λ_{max} et λ_{max} sur une échelle logarithmique. On appelle ici λ_{max} la valeur minimale pour que tous les coefficients du modèle soient réduits à 0 et λ_{min} généralement un millième de λ_{max}.

La valeur de λ sera celle du λ minimisant l'erreur de validation croisée.

Le résultat de cette régression LASSO réside uniquement dans l'information des variables à conserver dans le modèle celles dont $\widehat{\beta_{MV}^{Lasso}} \neq 0$. La valeurs des paramètres ou les prédictions de la variable à expliquer ne sont pas à conserver puisque le modèle a été volontairement biaisé par la pénalisation.

Il est à noter que dans le cas de fortes corrélations entre les variables explicatives, le LASSO sera moins efficace puisqu'il choisit une variable parmi les variables fortement corrélées au détriment des autres.

Illustrons ce type de pénalisation de la manière suivante [9] :



9. Ici avec une régression linéaire avec méthode des moindres carrés

Ici le coefficient β_2 est annulé

Sélection RIDGE

La sélection RIDGE ressemble dans la théorie à celle du LASSO tant il s'agit dans les deux cas d'une régression pénalisée. Cependant, la sélection pratique une pénalisation avec une norme l_2.

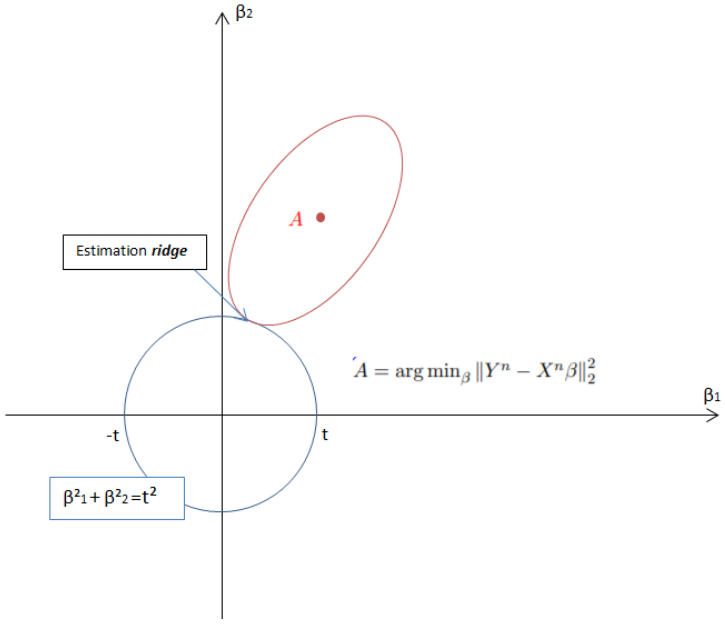
$$\boxed{\widehat{\beta_{MV}^{Ridge}} = argmax_\beta(ln(L) - \lambda||\beta||_2^2)} \tag{3.12}$$

où $\lambda > 0$ contrôle la puissance de la régularisation avec $||.||_2^2$ est la norme 2.

La régression *ridge* conserve toutes les variables mais sa pénalisation empêche les valeurs trop élevées des coefficients et limite la variance du modèle.

Elle permet de garder les variables explicatives pour interpréter le modèle et améliore les propriétés numériques et la variance des estimations.

Illustrons ce type de pénalisation de la manière suivante [10] :



Le point de concurrence des ellipses donne les coefficients ridge

10. Ici avec une régression linéaire avec méthode des moindres carrés

Sélection Elasticnet

La méthode Elasticnet est une extension du LASSO dans le sens où il permet dans notre cas de pouvoir améliorer une limitation du LASSO, celle mentionnée précédemment sur les variables fortement corrélées. Il sera question de rajouter une pénalité *ridge* à la pénalité LASSO afin d'allier les bénéfices de chacune des méthodes.

La fonction objectif sera pénalisée par le terme : $\lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$.

3.3.1.3 Deux modèles : Affaire Nouvelles et Affaires en Stock

Une fois que nous avons choisi notre modèle et avant de sélectionner les variables explicatives, nous pouvons nous demander s'il est pertinent de déployer ce modèle sur l'ensemble du portefeuille d'étude. En effet, après un travail nécessaire de synthèse des données, des tendances de résiliations ont été observées, ce qui est utile à la vérification et à la compréhension de nos profils mais également des résultats de nos modèles. Cependant, cette étude a révélé que naturellement les "Affaires Nouvelles" (les contrats ayant moins d'un an d'ancienneté) n'observent pas les mêmes proportions de résiliation. La résiliation est beaucoup plus forte sur ce segment, à tel point qu'il parut judicieux de les traiter à part et de leur réserver un modèle exclusif.

Le comportement des clients possédant ce genre de contrats différait totalement des autres, et ceci s'explique par plusieurs choses. L'assuré adoptant récemment un contrat MRC est généralement un nouveau client, car la MRC est un produit d'appel pour les professionnels et le nombre de contrats MRC détenus un même client est rarement plus grand que 1. De ce fait, il est plus susceptible de résilier qu'un client ayant plusieurs contrats et ayant une plus grande ancienneté dans le portefeuille, on peut également ajouter que la disparition du risque (faillite, rachat,...) est plus fréquente.

Cependant, ceci aurait pu être pris partiellement en compte par la variable ancienneté de contrat présente dans la base construite, mais une contrainte opérationnelle ne permettait pas également de penser à un modèle performant regroupant l'ensemble du portefeuille, celle des données. En effet, les nouvelles affaires ne sont pas pourvues des mêmes données, que ce soit sur le client ou sur le contrat, par exemple le taux de revalorisation ou encore le nombre de défense effectuée.

Ainsi pour des raisons opérationnelles mais également après une réflexion métier sur la question, la séparation de ces deux parties du portefeuille semble justifiée. Il s'agit de deux types de comportements différents capturés donc par deux modèles différents.

Il sera effectué un modèle sur le "stock" des contrats (hors Affaires Nouvelles) et les "Affaires Nouvelles" prenant effet en 2016.

3.3.1.4 Le Modèle Linéaire Généralisé : régression logistique

La base de donnée utilisée a été présentée précédemment, la base d'entraînement du modèle représente 80% de la base initiale et celle de test 20%.

Plusieurs variables n'ont pas été sélectionnées à cause de leur colinéarité avec d'autres variables explicatives, malgré le fait que la presque totalité des variables soient corrélées. Par exemple, le nombre de contrats Vie et le nombre de contrats Non-Vie rendent le nombre de contrats total inutile puisqu'il est la somme de ces deux variables.

La réelle sélection de variables se fait par la pénalisation LASSO présentée plus haut qui donnait de meilleurs résultats que les autres méthodes explicitées.

Afin de choisir les variables explicatives du modèle, il est convenable de faire varier la pénalisation λ de la méthode. Pour chaque valeur de λ, plusieurs modèles sont entraînés via une validation croisée à partir de l'échantillon d'entraînement.

Pour améliorer les performances de la régression logistique et diminuer le poids de l'hypothèse linéaire, il convient de segmenter les variables explicatives numériques ayant une grande variance. En effet, une variable explicative avec des valeurs extrêmes aurait l'effet de diminuer le "Bêta" de la variable et donc de possiblement sous-estimer l'effet marginal de la variable ou encore de la rendre non significative. Cette segmentation a été faite en étudiant la distribution de la variable par rapport à la variable réponse ou encore avec le graphique de splines cubiques où les "pics" seront les modalités de la variable.

Devant le grand nombre de variables présentes, nous garderons empiriquement les variables semblant apporter un pouvoir explicatif significatif. Sont écartées les caractéristiques trop spécifiques de l'activité du commerce (comme les étoiles pour les hôtels), ou redondantes, et les variables présentant beaucoup de valeurs manquantes.

3.3.1.5 Modèle des contrats en stock

Dans un premier temps, nous allons présenter le modèle concernant les données du stock comportant 94606 observations.

Deux modèles ont été établis : la différence entre les deux réside dans le nombre de variables prises en compte.

En effet, les contraintes opérationnelles de l'entreprise liées à une éventuelle mise en production font que le nombre de variables doit être restreint ; il sera question d'un modèle dit allégé sans forte perte de performance. Les variables les moins significatives ont été écartées, ainsi que celles dont l'accès est plus difficile et ne peut être obtenu à tout moment.

Le modèle allégé comporte 22 variables explicatives à la fois quantitatives et catégorielles. Le modèle ayant plus de degrés de liberté comporte lui 41 variables.

Présentons le modèle allégé et ses variables explicatives :

Variables explicatives du modèle allégé sur le stock	*Type de variable*	*Commentaires*
Famille d'entrée (segmentée)	4 modalités	AUTO/MRC/MRH et Autres
Paiement mensuel	1/0	Paiement mensuel ou non
Nombre de sinistres	Quantitative	Nombre de sinistres sous 36mois
Effectif de l'entreprise	4 modalités	0/1 à 9/ +10 salariés et sans employeur; variable retravaillée
Reseau	2 modalités	Agents/Courtiers
Année de création entreprise	Quantitative	Variable retravaillée
Taux de revalorisation (sans défense)	Quantitative	
Nombre de contrats IARD	Quantitative	
Nombre de contrats VIE	Quantitative	
Nombre de contrats sortis	Quantitative	
Produit	6 modalités	Les produits cités précédemment
Année de survenance du dernier sinistre	Quantitative	
Année du dernier contrat souscrit	Quantitative	
Année du dernier contrat sorti	Quantitative	
Anciennete du client (en année)	Quantitative	
Surface à assurer (segmentée)	3 modalités	
Capitaux à assurer (segmentée)	4 modalités	
Nombre de sites assurés	Quantitative	
Propriétaire	1/0	Variable retravaillée
Contentieux	1/0	
Famille d'activité (segmentée)	2 modalités	Segmentation entre famille cible ou non cible
Anciennete du contrat	Quantitative	

*Présentation des variables explicatives du modèle allégé sur le stock avec la mention "**variable retravaillée**" pour les valeurs manquantes recréées*[11]

On notera que la prime n'a pas été prise en compte, elle est très corrélée avec de

11. Voir paragraphe traitement et nettoyage des données.

nombreuses variables explicatives qui sont aussi tarifaires et de ce fait elle n'améliorait pas la performance. De plus, le tarif est amené à changer considérablement durant les prochains mois.

Afin d'analyser les résultats de la régression logistique, il sera présenté la valeur des coefficients de chaque variable ainsi que l'erreur et la p-value du test de Student réalisé. Il sera considéré comme significatif une p-value inférieure à 5%.

Variables du modèle allégé	*Coefficients*	*Erreur*	*Statistique*	*P-value*	*Significativité*
Famille d'entrée Autres	0,158	5,78E-02	2,73	6,42E-03	**
Famille d'entrée MRC	-0,101	4,77E-02	-2,11	3,49E-02	*
Famille d'entrée MRH	-0,279	1,25E-01	-2,24	2,53E-02	*
Paiement mensuel	0,135	2,99E-02	4,51	6,43E-06	***
Nombre de sinistres	-0,090	3,70E-02	-2,43	1,51E-02	*
1-9salariés	-1,595	3,97E-02	-40,18	<2e-16	***
Plus de 10salariés	-1,418	5,97E-02	-23,74	<2e-16	***
SSEMPL	-1,325	5,18E-02	-25,59	<2e-16	***
ReseauCourtier	-0,074	2,88E-02	-2,56	1,04E-02	*
Année de création entreprise	-0,004	9,79E-04	-3,85	1,18E-04	***
Taux de revalorisation (sans défense)	8,382	1,03E+00	8,18	2,89E-16	***
Nombre de contrats IARD	-0,019	6,20E-03	-3,01	2,65E-03	**
Nombre de contrats VIE	-0,167	5,18E-02	-3,22	1,27E-03	**
Nombre de contrats sortis	0,031	3,21E-03	9,60	<2e-16	***
PRODUIT100%PROARTISANS-COMMERCANTS	0,019	3,49E-02	0,53	5,94E-01	
PRODUIT100%PROASSOCIATION	-0,298	1,06E-01	-2,81	4,92E-03	**
PRODUIT100%PROFABRICATION	0,137	5,65E-02	2,42	1,57E-02	*
PRODUIT100%PROSERVICE	-0,167	4,18E-02	-4,00	6,31E-05	***
PRODUITANCIENNEGENERATION	0,464	8,37E-02	5,54	2,97E-08	***
Année de survenance du dernier sinistre	0,000	1,54E-05	1,37	1,70E-01	
Année du dernier contrat souscrit	-0,022	4,79E-03	-4,54	5,53E-06	***
Année du dernier contrat sorti	0,000	1,73E-05	15,53	<2e-16	***
Ancienneté du client (en année)	-0,150	5,37E-03	-27,92	<2e-16	***
Surface à assurer (moyenne)	-0,109	3,84E-02	-2,85	4,43E-03	**
Surface à assurer (grande)	-0,118	5,26E-02	-2,24	2,50E-02	*
Capitaux à assurer (moyen)	-0,162	3,32E-02	-4,86	1,16E-06	***
Capitaux à assurer (important)	-0,208	3,98E-02	-5,24	1,62E-07	***
Nombre de sites assurés	-0,112	2,74E-02	-4,08	4,54E-05	***
Propriétaire	-0,213	4,26E-02	-5,01	5,58E-07	***
Contentieux	0,853	2,80E-02	30,51	<2e-16	***
Famille d'activité segment2	-0,224	4,90E-02	-4,56	5,06E-06	***
Ancienneté du contrat	0,106	6,51E-03	16,31	<2e-16	***
Echelle de significativité	*0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1*			**AIC : 43017**	

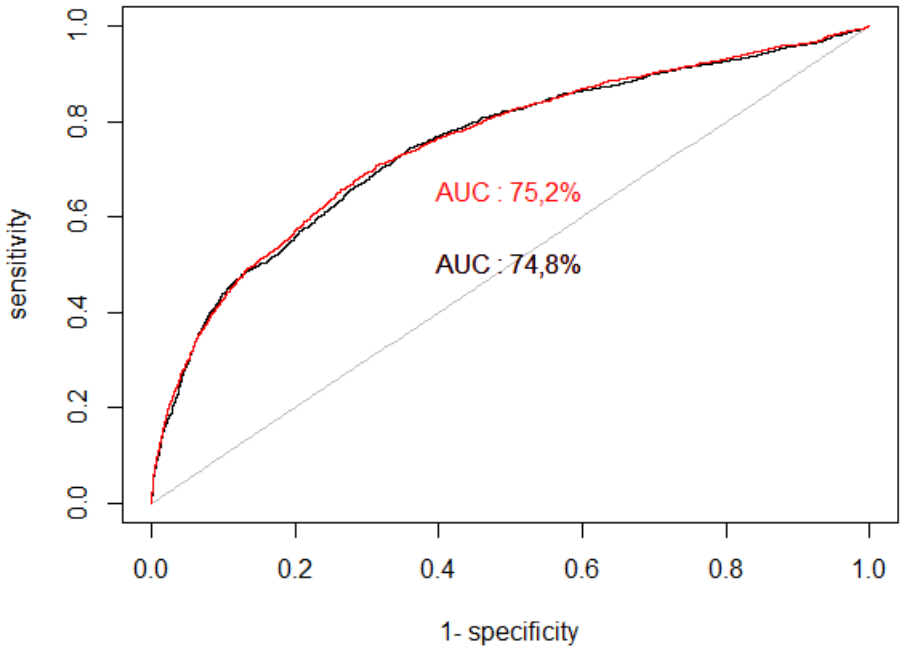
Présentation des résultats du GLM avec une déviance de 43010, les variables sur fond blanc ne sont pas significatives.

Après avoir retiré l'année de survenance du dernier sinistre, on note aucun gain AIC. Rappelons rapidement que l'AIC est le critère d'information d'Akaike qui est une mesure de la qualité d'un modèle statistique. Son expression est la suivante : $AIC = 2k - 2ln(L)$

avec k le nombre de paramètres du modèle et L le maximum de la vraisemblance. Cet indicateur doit être minimisé et il sera l'un des principaux indicateurs pour choisir notre modèle.

Le modèle plus complet comprend en plus comme variables explicatives : le taux de rabais, le nombre de contrats en Auto/MRC/MRH/Epargne-retraite, la défense ou non du contrat, le chiffre d'affaire, l'année de création des locaux, l'encours, la prime, le nombre de contentieux et même le nombre d'étoiles pour les hôtels. Ces variables ont été sélectionnées par LASSO également mais sans considération par rapport à l'accès au données et avec une moins forte parcimonie.

Afin de comparer nos deux modèles, l'un allégé et l'autre plus complet, affichons leur courbe ROC ainsi que leur AUC [12].



Courbes ROC des modèles sur leur échantillon Test (en rouge : le modèle complet, en noir : le modèle allégé).

Compte tenu de la faible différence entre le modèle allégé et le modèle complet (AIC = 42758 contre 43017), nous adopterons le modèle allégé.

12. Area Under Curve, plus d'informations dans "A new look at the statistical model identification" par Hirotugu Akaike.

Intéressons-nous aux effets marginaux de chaque variable.

Variables du modèle allégé	Effet sur la résiliation	Variables du modèle allégé	Effet sur la résiliation
Famille d'entrée Autres	1,17	PRODUIT100%PROFABRICATION	1,15
Famille d'entrée MRC	0,90	PRODUIT100%PROSERVICE	0,85
Famille d'entrée MRH	0,76	PRODUITANCIENNEGENERATION	1,59
Paiement mensuel	1,14	Année de survenance du dernier sinistre	1,00
Nombre de sinistres	0,91	Année du dernier contrat souscrit	0,98
1-9 salariés	0,20	Année du dernier contrat sorti	1,00
Plus de 10 salariés	0,24	Ancienneté du client (en année)	0,86
SSEMPL	0,27	Surface à assurer (moyenne)	0,90
ReseauCourtier	0,93	Surface à assurer (grande)	0,89
Année de création entreprise	1,00	Capitaux à assurer (moyen)	0,85
Taux de revalorisation (sans défense) en %	2,31	Capitaux à assurer (important)	0,81
Nombre de contrats IARD	0,98	Nombre de sites assurés	0,89
Nombre de contrats VIE	0,85	Propriétaire	0,81
Nombre de contrats sortis	1,03	Contentieux	2,35
PRODUIT100%PROARTISANS-COMMERCANTS	1,02	Famille d'activité segment2	0,80
PRODUIT100%PROASSOCIATION	0,74	Ancienneté du contrat (en année)	1,11

Présentation de l'effet marginal de chaque variable sur la probabilité de résiliation.

Avant de rentrer dans le détail des commentaires sur les chiffres présentés, notons que l'effet de la variable "Ancienneté du contrat" sur laquelle repose l'étude est trompeuse. En effet, elle est fortement corrélée à la variable "Ancienneté du client" : ainsi il faut prendre des précautions et étudier l'effet marginal de l'un avec l'autre. Plus concrètement, l'augmentation d'une unité de l'ancienneté du contrat multiplie la probabilité de résiliation par $\beta_{anc.client} + \beta_{anc.contrat}$ soit 0,97 *toutes choses égales par ailleurs* car l'augmentation de l'ancienneté du contrat induit la même augmentation de l'ancienneté du client par définition. Ce lien est même causal.

Il est également à noter que le taux de revalorisation impacte la prime et peut donc pousser le client à résilier le contrat, mais cette variable a une faible valeur moyenne et une faible variance, donc son effet marginal est à relativiser. L'augmentation d'un point de revalorisation est une action forte quand la moyenne de cette variable est de 2,8%.

Après avoir relativisé les chiffres, relevons les effets les plus forts sur la résiliation :

► Les caractéristiques augmentant le plus fortement la probabilité de résiliation :

↪ Le taux de revalorisation à considérer avec les précautions décrites précédemment. Un fort taux de revalorisation est parfois utilisé comme levier à la résiliation d'un contrat.

↪ L'indicatrice CONTENTIEUX qui renseigne le fait que le client soit

déjà entré en contentieux avec l'entreprise. Cela représente 27% du portefeuille étudié, signe qu'un travail doit être fait sur le tarif, l'ajustement tarifaire et la gestion du client.

↪ Le produit "ANCIENNE GENERATION" est en run-off et donc n'est plus commercialisé pour les nouveaux clients. Les clients ayant encore ces produits sont "invités" à basculer sur un produit commercialisé. Cela représente seulement 2.6% du portefeuille.

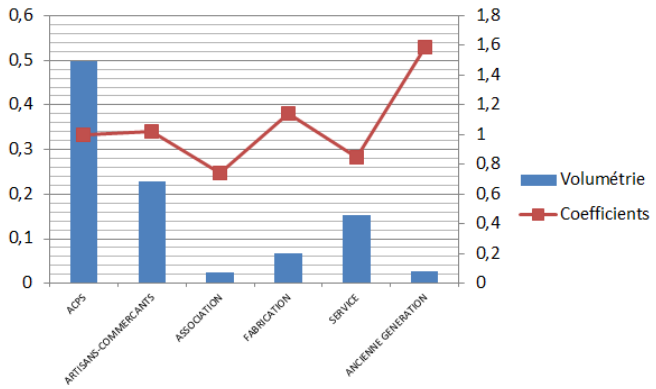
▶ Les caractéristiques diminuant le plus fortement la probabilité de résiliation :

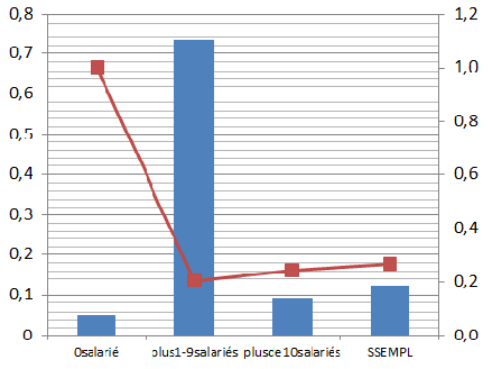
↪ Un effectif de l'entreprise plus grand que 0 est la modalité qui diminue le plus la probabilité de résiliation. En effet, cela donne des indications sur la santé financière de l'entreprise et si l'on combine cette variable avec celle des capitaux à assurer, on peut avoir une vision correct des finances de l'assuré.

↪ Le produit PRO-ASSOCIATION est selon le modèle un produit dont la probabilité de résiliation *toutes choses égales par ailleurs* est la plus faible.

↪ L'entrée dans le portefeuille par un contrat Multi-Risque Habitation est également un indicateur de faible résiliation *toutes choses égales par ailleurs*.

La volumétrie ainsi que les coefficients associés aux modalités seront représentés pour les variables "Produit" et "Effectifs" :





Nous pouvons penser que les lois de chute seront créées avec une pente de 0.97 qui est l'effet marginal de l'ancienneté du contrat, cependant, beaucoup de variables évoluent en même temps que l'ancienneté du contrat. En effet, figer les caractéristiques de l'assuré en faisant vieillir son contrat n'est pas raisonnable tant ses caractéristiques évoluent

également avec le temps, notamment son nombre total de contrats, les capitaux à assurer ou encore dans le cas d'une expansion, sa surface ou son nombre de salariés. De plus, l'effet ici est seulement linéaire et ne tient donc pas compte de l'ancienneté du contrat à vieillir : le coefficient lié à l'ancienneté du contrat reste le même.

L'étude menée ici, tentera de prendre en compte cela en segmentant encore plus le portefeuille en le clusterisant puis en étudiant l'effet de l'ancienneté du contrat pour chaque année et la probabilité de résiliation conditionnée à l'ancienneté actuelle du contrat.

L'AIC nous donne des indications quant au modèle à adopter, c'est un indicateur de performance statistique relatif tandis que l'AUC est plus universel même si les deux dépendent fortement de la qualité et de la volumétrie des données.

Regardons d'une autre manière l'adéquation de notre régression à la réalité en comparant les résultats de la prédiction sur les contrats réellement résiliés et non résiliés à travers un diagramme quantile-quantile.

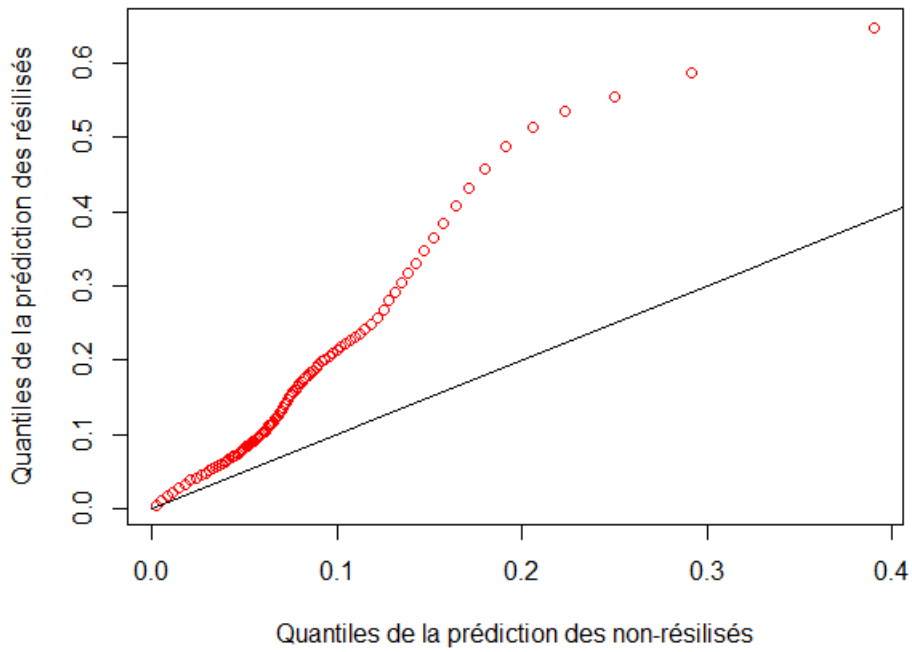


Diagramme quantile-quantile des prédictions selon la résiliation ou non du contrat vu à posteriori avec comme repère la droite noire

Nous remarquons que pour les quantiles élevés les points s'écartent de la 1ère bissectrice ce qui traduit une bonne différenciation entre les prédictions des futurs résiliés et les futurs contrats en stock fin 2017. En effet, plus les points s'éloignent de la droite plus la différence entre ces deux types de contrats (différence que l'on veut faire) est grande. Une courbe proche de la bissectrice dénote d'une uniformité des probabilités de résiliation et n'a pas de pouvoir de prédiction quant à la résiliation ou non du contrat.

Un problème peut être souligné pour les faibles valeurs de probabilités qui n'ont que peu de pouvoir de prédiction compte tenu de la proximité des distributions de probabilités entre ces deux groupes.

3.3.1.6 Modèle des contrats Affaires Nouvelles

Afin de ne pas répéter les propos tenus précédemment, nous allons seulement discuter des performances de la régression logistique réalisée sur ce périmètre.

Le modèle a été construit de la même façon ainsi que les graphiques, la seule différence notable est la proportion de résiliation et les variables n'existant plus dans ce modèle , à savoir : taux de revalorisation, contentieux (évènement trop rare sur une si courte période de couverture) et la modalité Ancienne Génération de produit naturellement.

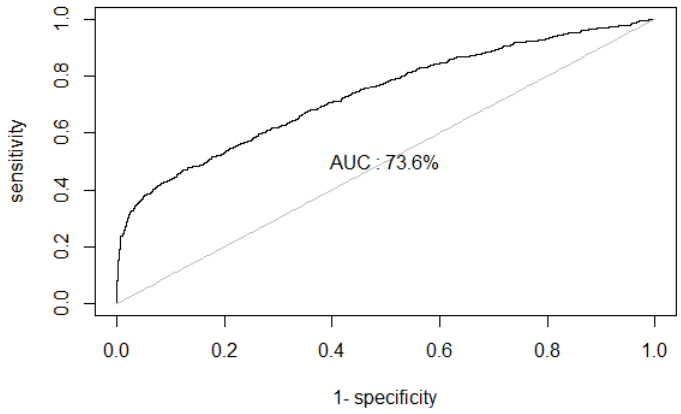
A noter que d'autres segmentations pour les variables quantitatives avec une forte variance ont été faites sur le même schéma que dans la partie précédente (utilisations de splines cubiques et de clustering).

Les variables les plus influentes sont les variables économiques : le chiffre d'affaires en premier lieu (plus le montant est élevé, moins la probabilité de résiliation est importante), la segmentation du secteur d'activité, l'ancienneté du contrat (même si il est d'un an au maximum) et les capitaux assurés avec la même logique que le chiffre d' affaire.

Parmi les variables favorisant la résiliation se trouvent : les produits PRO ARTISAN et PRO FABRICATION ainsi que le réseau Courtier (contrairement au stock).

Les détails de ce modèle seront mis en annexe.

Malgré ces absences, le modèle donne les résultats suivants pour les 20202 contrats concernés : un AIC de 14708 et un AUC de 73,6%.



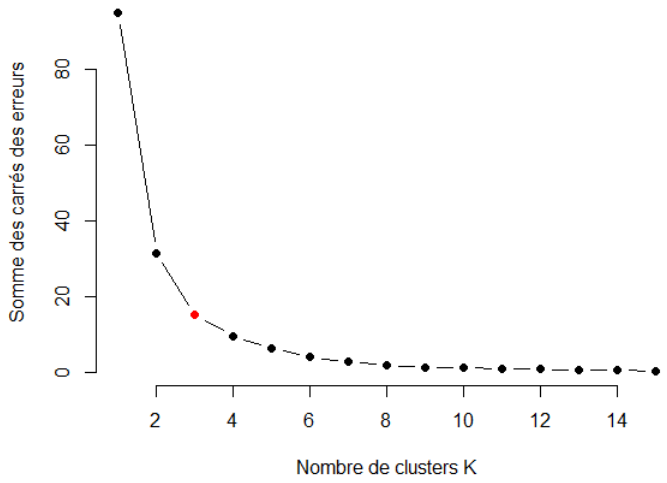
Les modèles qui vont suivre seront également élaborés pour les contrats Affaires Nouvelles, cependant les modèles d'apprentissage automatisé n'apporteront pas de réelle plus-value compte tenu de la faible volumétrie de ce périmètre.

3.3.2 Création d'un zonier à partir des résidus du modèle

Pour les deux modèles étudiés, n'a été prise en compte aucune variable géographique. En effet, le département de l'assuré n'est pas significatif compte tenu du grand nombre de ses modalités et du faible volume par modalité. Il faut donc créer une variable géographique clusterisée avec peu de modalités pour qu'une variable géographique devienne significative.

Le but serait d'analyser la résiliation par zone géographique, de faire un zonier simplifié comme cela existe dans les processus de tarification avec comme métrique la sinistralité par exemple.

De ce fait, une étude des résidus du GLM par département a été faite. Des clusters de départements ont été faits sur la mesure de la déviance par département du modèle finalisé du GLM. Le regroupement de département est choisi puisqu'il est suffisamment macroscopique pour être significatif et est suffisamment fin pour ne pas composer avec une trop importante variance intra-groupe de la déviance. Le regroupement a été fait par la méthode de k-means et pour déterminer le nombre de clusters à créer, la méthode graphique du coude a été choisie pour sa simplicité.



Méthode du coude avec comme point rouge le "coude" pour K=3

L'idée sous-jacente est de faire l'hypothèse que la composante non-expliquée par le modèle actuel est explicable en partie par une variable géographique segmentée par département et que cette variable n'est pas corrélée trop fortement avec les autres variables explicatives.

L'ajout de cette variable créée n'apporte pas de réel gain de pouvoir prédictif dans la modélisation de la résiliation.

La variable géographique apporte cependant des indications sur la répartition géographique de l'assuré et permet de faire un zonier sur la performance du modèle par zone géographique.

Cette variable ne se trouvera donc pas dans les prochaines modélisations.

3.4 Mélange de régressions logistiques

Il s'agit d'une extension de la simple régression logistique avec la même théorie, les mêmes données et les mêmes variables explicatives que dans la partie précédente. L'avantage de ce mélange est de prendre en compte l'hétérogénéité non-observable des données. Il s'agit d'une généralisation bien connue de la régression où la loi sous-jacente est une loi mélange. Brièvement, la densité du mélange s'écrit de la manière suivante :

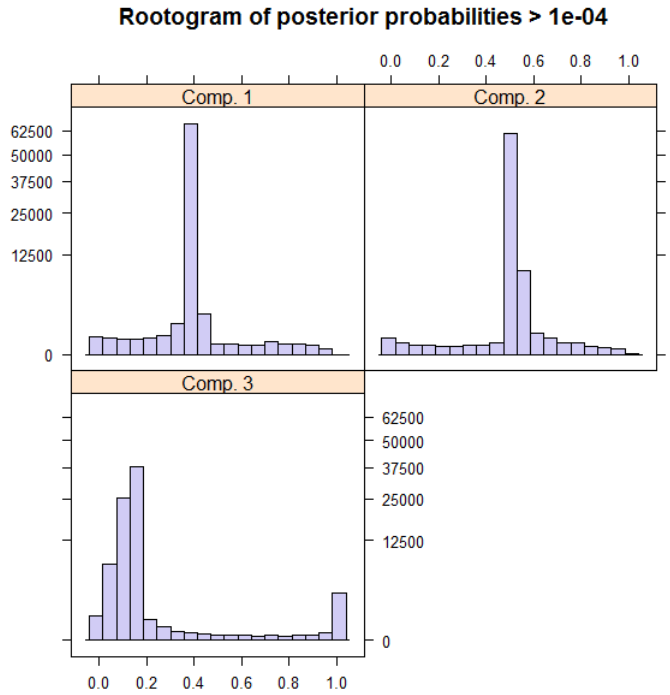
$$\boxed{f(y_i) = \sum_{j=1}^{k} \pi_j f_j(x_j)} \tag{3.13}$$

où f est la densité de Y, les f_j les densités des composantes, k le nombre de composantes et $\sum_{j=1}^{k} \pi_j = 1$

Ici, il s'agira uniquement de mélanges de distributions logistiques. Chaque observation appartient à un groupe du mélange, le nombre de groupes est fini et est déterminé inférentiellement. Les probabilités d'appartenance à tel ou tel groupe doivent être estimées dans le même temps.

Le package utilisé [13] nous permet de choisir le nombre de clusters en fonction de l'AIC du modèle et fournit les probabilités de résiliation des n composantes avec le cluster d'appartenance de chaque observation.

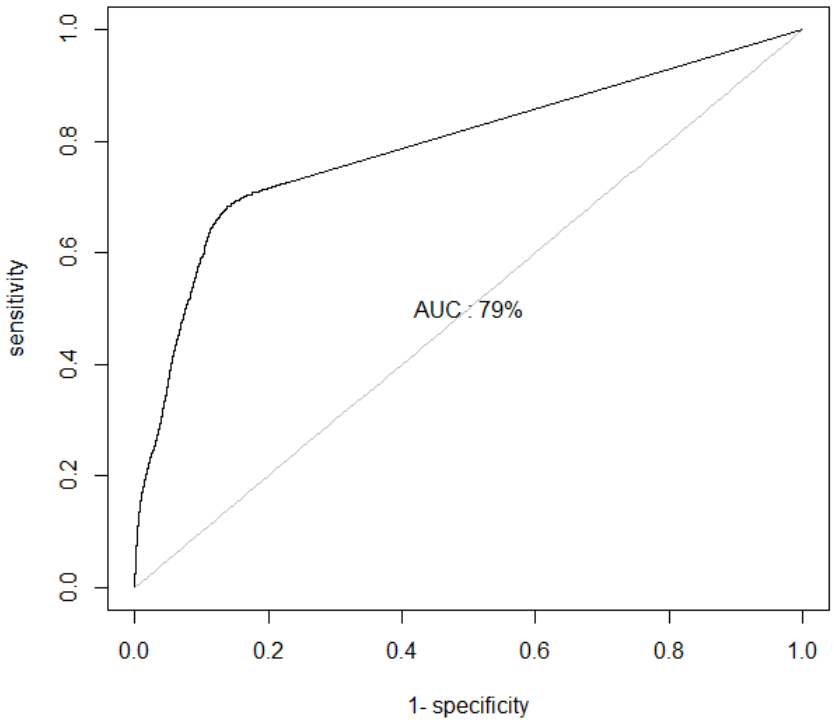
L'AIC est maximisé pour 3 clusters. Le portefeuille sera réparti suivant les 3 composantes avec une distribution de probabilité pour chaque composante suivante :



Où la taille des groupes sont respectivement de : 1091, 71452 et 3141 avec des probabilités moyenne de 0,361, 0,482 et 0,157.

13. Package Flexmix, R.

Le modèle choisi nous donne une performance de : 27776 pour l'AIC et un AUC de 79%.



On remarque un point d'inflexion significatif de la courbe ROC

Ce modèle apporte de meilleurs résultats mais gêne par son manque de flexibilité dans le sens où les clusters créés sont difficilement déployables sur des données indépendantes et différentes de l'échantillon d'entraînement.

De plus, la clusterisation du portefeuille induit une complexification de l'interprétabilité de la modélisation. Il faut associer en même temps que le calcul des probabilités, un cluster à chaque observation.

Pour toutes ces raisons, il ne sera pas retenu.
Bien que sa mise en pratique apporte un regard nouveau sur le portefeuille (clusterisation par homogénéité de la variable réponse) et donc peut permettre de subdiviser le portefeuille pour réaliser plusieurs modèles, son manque de flexibilité et le fait qu'il ne peut être calculé instantanément (dans le cadre de la valeur en temps réel) à cause de l'association des observations à un cluster le rendent inadapté à l'étude ici développée.

Une autre façon d'étendre la régression logistique va être explicitée dans la partie suivante.

3.4.1 Le Modèle Additif Généralisé

3.4.1.1 Justification d'un tel modèle et théorie

Après avoir déployé le GLM et expliqué sa théorie, on a mentionné le caractère linéaire entre la transformation de la variable réponse par la fonction lien g et les variables explicatives quantitatives. Les variables catégorielles n'ont pas cette condition de linéarité. Une méthode serait de segmenter toutes les variables numériques pour n'avoir que des variables catégorielles mais une fois encore on perdrait beaucoup en terme de clarté, et on diminuerait possiblement la performance du modèle à cause des erreurs de segmentation.

Le Modèle Additif Généralisé se soustrait à cette forme linéaire : $X\beta$.
Sa structure reste, cependant, la même que celle des GLM de ce fait, nous n'allons pas revenir trop longuement sur son développement théorique. Il s'agit grossièrement d'une fusion du GLM et du modèle additif.

Au lieu de d'estimer les β, on estime des fonctions paramétriques ou non des variables explicatives ce qui rend l'adaptabilité du modèle assez importante et le cadre assez vaste.

Grâce à cela, la contribution marginale de chaque variable ne sera pas seulement de $\beta_i X_i$ mais de $f_i(X_i)$ la variable explicative est transformée pour estimer au mieux la variable dépendante.

Sa forme est la suivante :

$$\boxed{g(E(Y|X)) = \beta_0 + \sum_{i=1}^{n} f_i(x_i)} \tag{3.14}$$

où g est la fonction lien, les fonctions f_i peuvent être cantonnées à un certain domaine de fonctions.

Rappelons tout de même qu'ici nous préférons préserver l'explicabilité et l'interprétabilité du modèle autant que faire se peut, ainsi nous choisissons pour les fonctions f_i des splines cubiques.
Elle permet à l'observateur de comprendre les données disponibles et de se forger un premier avis de la résiliation étudiée.

Ainsi, il est possible de garder une souplesse et une certaine transparence dans la modélisation et dans le même temps, se soustraire à la linéarité hypothétique de la relation $g(E(Y)) = X\beta$.

Un écueil peut être observé pour le modèle additif généralisé, c'est celui du sur-apprentissage.

Le nombre de paramètres permettant ce lissage peut être spécifié et se doit d'être suffisamment faible pour ne pas sur-apprendre : très inférieur au nombre de degrés de liberté des données. De manière assez standard et utilisé dans les prochaines modélisations, la validation croisée pourra servir à détecter et réduire les problèmes de sur-apprentissage éventuels.

L'arbitrage sera de préférer un modèle GLM si un modèle additif généralisé ne permet pas d'accroître sensiblement les performances de la prédiction.

Ce modèle peut s'avérer également utile quand, pour une GLM ou un autre modèle, il est nécessaire de segmenter des variables explicatives. En effet, le modèle par sa modélisation, ici en splines ($s(X_i)$) donne des indications sur le nombre de segments et les intervalles de la variable X_i.

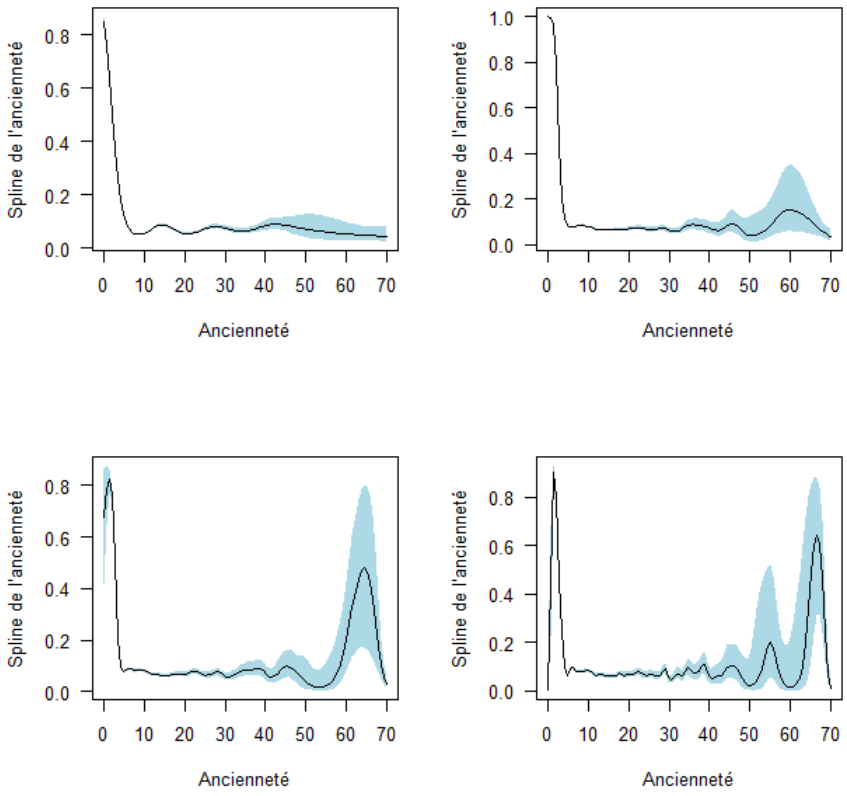
3.4.1.2 Le Modèle Additif Généralisé

Nous avons gardé les mêmes variables que précédemment avec l'idée de ne pas segmenter les variables quantitatives comme la surface, les capitaux, etc. Les variables numériques non binaires ont été transformées en splines cubiques.

Les paramètres de lissage vont également être estimés [14].

Présentons d'abord l'ajustement des splines pour l'ancienneté du client.

14. noté λ il s'agit d'un paramètre positif qui conditionne la courbure de la courbe, un λ élevé sera plus précis mais plus complexe à estimer.



L'effet de l'ancienneté sur la résiliation avec l'ancienneté comme seule variable explicative avec un k le nombre de dimensions du terme de lissage de 9,20,30,50. [15]

Il est aisé de comprendre que l'effet est mieux pris en compte lorsque le nombre de dimensions est grand avec comme désavantage une modélisation plus coûteuse.

Ici, pour les variables avec peu de variance et d'importance, le nombre de dimensions est réduit. Pour les variables ancienneté client et contrat, k a été choisi assez élevé pour rendre la variable significative et assez explicative.

La méthode choisie a été celle de REML, maximum de vraisemblance restreint, qui se base sur la maximisation de la vraisemblance sur les données ainsi transformées. Elle permet de déterminer des paramètres sans intervention du statisticien.

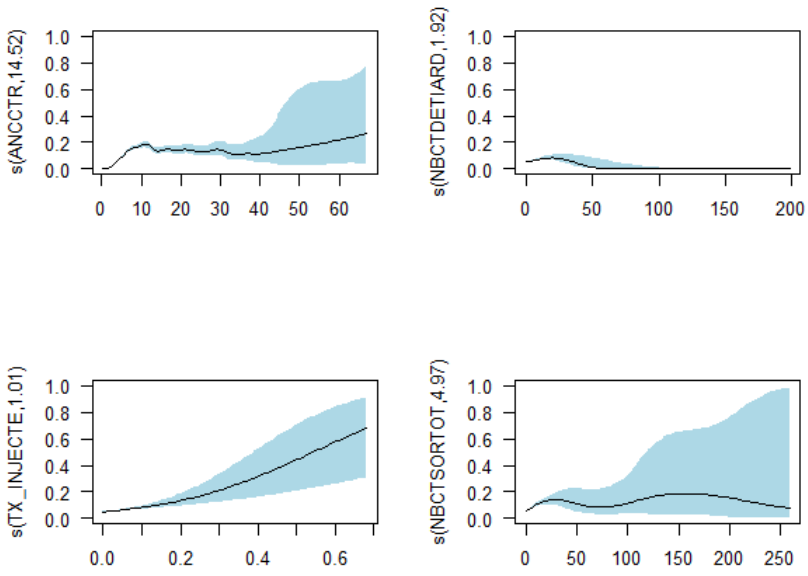
Nous allons présenter également la significativité des variables, leurs coefficients et leurs effets marginaux :

15. La courbe correspond à l'inverse de la fonction logistique de la sortie du modèle GAM.

Variables explicatives	Coefficients	Erreur	Effets	P-value	Significativité
Famille d'entrée Autres	0,134	0,060	1,14	2,55E-02	*
Famille d'entrée MRC	-0,116	0,052	0,89	2,41E-02	*
Famille d'entrée MRH	-0,208	0,125	0,81	9,67E-02	.
Nombre de contrats VIE	-0,029	0,050	0,97	5,56E-01	
Paiement mensuel	0,046	0,032	1,05	1,51E-01	
PRODUIT100%PROARTISANS-COMMERCANTS	-0,187	0,046	0,83	4,37E-05	***
PRODUIT100%PROASSOCIATION	-0,330	0,116	0,72	4,56E-03	**
PRODUIT100%PROFABRICATION	0,137	0,060	1,15	2,18E-02	*
PRODUIT100%PROSERVICE	-0,193	0,049	0,82	8,22E-05	***
PRODUITANCIENNEGENERATION	0,367	0,088	1,44	2,80E-05	***
1-9salariés	-1,592	0,045	0,20	<2e-16	***
Plus de 10salariés	-1,286	0,064	0,28	<2e-16	***
SSEMPL	-1,235	0,055	0,29	<2e-16	***
ReseauCourtier	-0,141	0,032	0,87	9,78E-06	***
Propriétaire	-0,211	0,045	0,81	2,39E-06	***
Contentieux	0,633	0,030	1,88	<2e-16	***
Famille d'activité segment2	-0,108	0,053	0,90	4,07E-02	*
Echelle de significativité	'***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' '			AIC : 37408	

On notera que les variables augmentant le plus fortement la probabilité de résiliation reste toujours : "Contentieux" et "Ancienne génération" pour le produit.

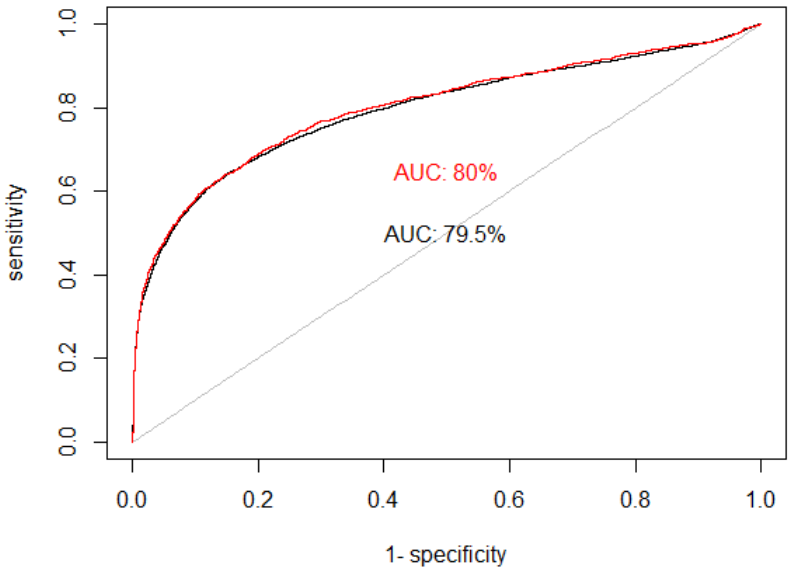
Deux variables deviennent non significatives, le nombre de contrats en VIE qui est généralement nul et la cadence mensuelle de paiement seront enlevés du modèle. Les autres variables ont été transformées par des splines cubiques et elles sont toutes très fortement significatives (à l'exception de la surface qui elle est moyennement significative) [16].



Les effets marginaux de 4 variables ont été présentés ici où "$TX\_INJECTE$" est le taux de revalorisation.

16. Très fortement significative correspond à '***' et moyennement à '*' selon l'échelle présentée plus haut.

Enfin, il sera question de la performance du modèle et son pouvoir prédictif.



En rouge : la courbe ROC sur l'échantillon Test et en noir : sur l'échantillon d'entraînement.

L'AIC également est meilleur que les anciens modèles avec une valeur de 37408.
On note la faible différence de performance entre l'échantillon Test et celui d'entraînement qui montre l'absence de sur-apprentissage du modèle.

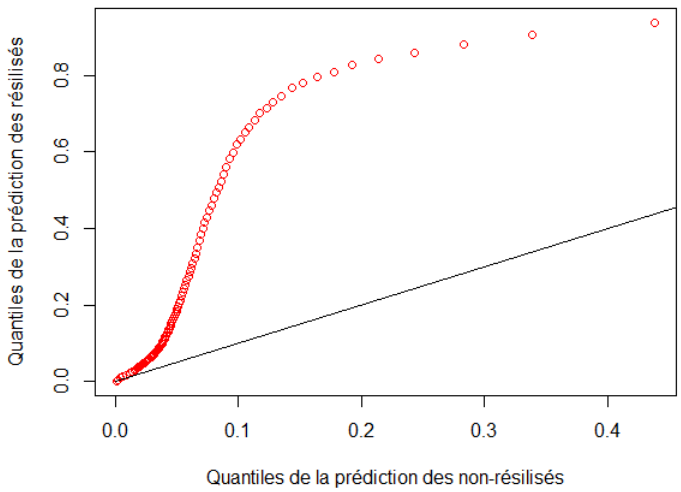


Diagramme quantile-quantile des prédictions selon l'état du contrat a posteriori.

Le modèle différencie bien, par les probabilités qui en émergent, les contrats en ayant une valeur plus forte pour les futurs résiliés.

L'interrogation se portera quant au choix du modèle pour les contrats en stock au 1er Janvier 2017. Les résultats pour les modèles statistiques présentés ici donnent raison au modèle additif généralisé qui, tout en restant aisément interprétable, fournit des performances meilleures.

Ce modèle est plus explicite que celui du mélange de régression logistique, et beaucoup plus performant que la régression linéaire simple.

Le fait de faire l'économie du caractère linéaire de la relation des variables avec la variable réponse transformée (par la fonction lien) est payant et dénote d'une situation où la linéarité est une hypothèse (trop) forte qui ne doit pas être sous-estimée.

Cependant, ces modèles ne prennent pas en compte les interactions entre les variables et gardent une hypothèse mathématique sur le lien entre les variables explicatives et la variable réponse. Ces raisons amènent à se poser la question de l'apprentissage automatisé.

3.5 Modèles Machine Learning : Théorie et Applications

Avant de mentionner les arbres CART puis les forêts aléatoires, les modèles MARS (Multivariate Adaptive Regression Spline) représentent une transition entre GAM et CART dans notre cheminement de modélisation. Les MARS s'affranchissent également de l'hypothèse d'additivité avec une utilisation de splines. Plus précisément, un MARS est combine des splines linéaires pour former un modèle de prédiction non linéaire. L'utilisation de splines est un choix fort de fonctions prédictives et permet de mieux lutter contre le sur-apprentissage. L'interaction entre variables est pris en compte ici contrairement aux modèles vus précédemment. Autrement, l'automatisation de la procédure marque la distinction entre les modèles statistiques et l'apprentissage automatisé.

Les arbres de décision sont des modèles d'apprentissage non linéaires dans une continuité d'apprentissage automatisé initiés par les MARS, notamment par l'utilisation de splines et de prise en considération d'interactions entre variables. Il s'agit d'établir une classification

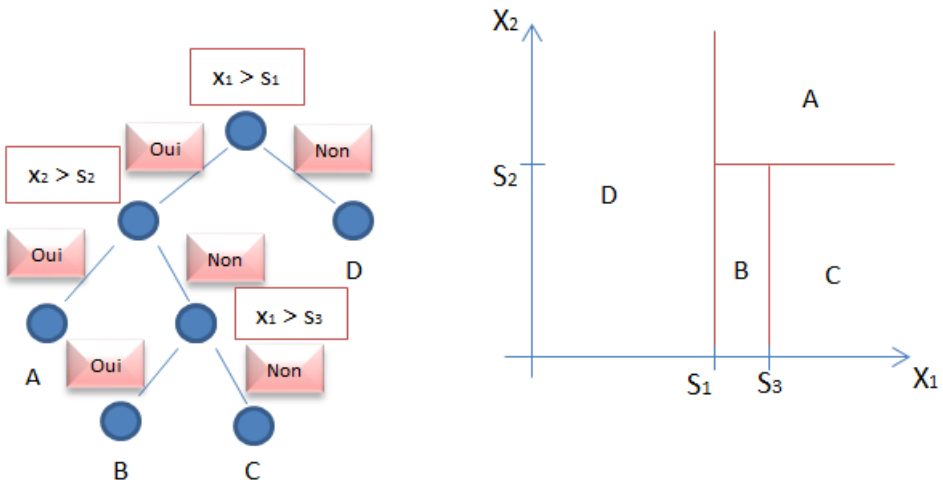
hiérarchique descendante des données en des groupes homogènes par rapport à la variable à expliquer. Chacun de ces groupes est associé à la valeur moyenne des résultats qu'il contient.

De manière plus explicite, l'arbre est constitué de noeuds qui séparent en deux la population du sous-ensemble. Chaque partition est basée sur une seule variable où un seuil est considéré afin de rendre le partitionnement binaire. On appelle noeud terminal le noeud après lequel aucun découpage du sous-ensemble n'est effectué. La méthode de croissance des arbres est dite sans regret, une fois la partition choisie, l'algorithme traite les sous-espaces générés de façon indépendante.

La croissance de l'arbre réside dans un partitionnement récursif des contrats étudiés selon les profils $(X_i)_i$ en minimisant à chaque étape l'erreur de prédiction du modèle. Nous discuterons, pour introduire la forêt aléatoire, de l'algorithme CART [17].

À chaque nœud de l'arbre, les indicatrices $1_{X_i<s_j}$, pour toutes les variables quantitatives X_i avec s_j comme seuil, sont utilisées afin de séparer en deux sous-ensembles de l'ensemble précédent.

Pour les variables qualitatives, toutes les combinaisons de modalités sont évaluées.



A gauche, se trouve une visualisation de l'arbre avec 2 variables quantitatives explicatives et à droite, l'espace des données partitionné.

L'objectif est de maximiser l'homogénéité des sous-ensembles créés et d'imposer à l'algorithme des conditions d'arrêt, autrement ce dernier ne composera qu'avec une

17. Classification And Regression Tree

observation par feuille qui sera l'homogénéité extrême. On appelle arbre saturé, un arbre dont une nouvelle division de sous-ensemble produit un noeud vide. Avant d'expliciter les indicateurs d'homogénéité à maximiser, notons tout de suite les conditions d'arrêts de l'algorithme les plus utilisées :

— Le profondeur maximale de l'arbre
— L'erreur augmente
— Le nombre d'observations minimum d'une feuille
— Le nombre d'observations du noeud terminal est de 1
— L'homogénéité du sous-ensemble est suffisamment élevée (avec un seuil à fixer)

Ces conditions sont très importantes et certaines doivent être arbitrées par le statisticien. L'un des risques du manque de maîtrise de ces restrictions est le sur-apprentissage. En effet, si le nombre de feuilles est trop important et donc que la population des feuilles est faible, un apprentissage trop fort des données rend le modèle inflexible à une autre base de données Test indépendante. Il s'agit d'un sur-ajustement des données qui prive le modèle d'un possible généralisation de son pouvoir prédictif.
En effet, une fois déployé sur une base test indépendante, il ne pourra pas prendre suffisamment de recul à cause de son ajustement extrême aux données d'entraînement. Ceci peut être causé par une trop importante profondeur et une trop grande complexité de l'arbre. Inversement à cela, un arbre peu profond ou complexe ne donnera pas des résultats assez performants et admettra un biais de construction.

Heureusement, pour pallier ce problème, il est important *d'élaguer* son arbre afin de le rendre plus robuste à la diversité de tests auquel il pourra être soumis. Cet élagage peut être fait de plusieurs manières : par l'observation des erreurs de prédiction des sous-arbres ou par des méthodes de pénalisation sur le nombre de feuilles [18].

Revenons alors à la métrique d'homogénéité et son expression. Il faut adapter l'indicateur d'homogénéité en fonction de la nature de la variable : pour une variable quantitative, l'indicateur naturel est la variance dans le sous-ensemble, pour une variable qualitative avec n modalités se présentent comme les plus courantes l'indice de diversité de Gini [19] et l'entropie de Shannon.

L'indice de Gini s'exprime de la manière suivante :

18. Avec une idée similaire aux sélections de variables vues précédemment
19. Ce critère est utilisé pour les arbres CART

$$\boxed{I_{Gini}(f) = \sum_{i=1}^{n}(1 - f_i) * f_i} \tag{3.15}$$

où f est la distribution des n modalités de la variable réponse

Cet indice rend compte de la probabilité d'erreur de classification d'une observation dont l'étiquette serait choisie aléatoirement selon la distribution des étiquettes dans le sous-ensemble.

Le gain d'information repose sur l'entropie de Shannon :

$$\boxed{I_{entropie} = -\sum_{i=1}^{n} f_i * log_2(f_i)} \tag{3.16}$$

où f est la distribution des n modalités de la variable réponse

Résumons de manière plus grossière les arbres CART en rappelant qu'ils présentent un modélisation automatisée non linéaire prenant en compte les interactions entre variables contrairement aux GLM ou GAM qui restent à l'initiative du statisticien.
C'est une modélisation interprétable quand elle n'est pas agrégée comme dans les forêts aléatoires [20].

Cependant, l'arbre CART par nature est facilement sujet au sur-apprentissage et est relativement instable quand il s'agit de modifier ses données d'apprentissage.

Pour se soustraire à ces lacunes, nous allons discuter des forêts aléatoires dont la base a été présentée dans cette partie.

L'idée est d'agréger les arbres CART pour que le modèle et ses prédictions soient plus robustes. Afin d'avoir avant l'agrégation des arbres décorrélés et diversifiés, un aléa est introduit dans le processus de croissance des arbres. De ce fait, le modèle agrégeant ces arbres sera plus généralisables et réduira le sur-apprentissage.

20. étudiées dans la prochaine partie

3.5.1 Modèle par forêts aléatoires : théorie

Le *random forest*[21] est une méthode d'agrégation de ces arbres. Elle est parmi les algorithmes les plus utilisés en Machine Learning.

Discutons donc de ses principes fondateurs.

La singularité réside dans les deux types d'aléas mis en jeu dans l'algorithme, l'un impactant les observations et l'autre les variables.

> Le premier est le bagging : contraction de bootstrap (en français, rééchantillonage aléatoire) et aggreging . Il influe sur l'échantillon qui est à la base des arbres mentionnés. A chaque étape, un échantillon avec remise est considéré et est servi de base de construction à un arbre, cet échantillon est de même taille que celui de la base initiale. Le nombre d'arbres peut être choisi par le statisticien. Enfin, les arbres sont agrégés, pour une variable réponse quantitative en prenant la valeur moyenne des arbres et pour une variable catégorielle la modalité la plus représentée, ceci finit de constituer la notion de bagging pour ce bref résumé.

> Le deuxième aléa concerne les variables : à chaque itération, durant la création des critères de partionnement du sous-ensemble, une variable ainsi qu'une division a lieu et cette variable sera cette fois sélectionnée dans un sous-ensemble de variables créé aléatoirement. La raison invoquée de cette complexification du processus est d'empêcher une trop forte corrélation entre les arbres. En effet si une ou plusieurs variables ont un pouvoir explicatif démesuré par rapport aux autres alors elles seront présentes dans la quasi-totalité des arbres et donc créerai par construction une corrélation entre des derniers.

Le sous-ensemble de variables candidates à chaque split a une taille qui peut être déterminé par validation croisée.

Une brève explication de la validation croisée s'impose : la validation croisée est une technique d'étude de performance de prédiction d'un modèle par sa généralisation à des bases de test indépendantes.
Une manière de faire de la validation croisée est par "k-fold" où la base donnée est divisée en k parties distinctes et dont chacune de ces parties sera la base test du modèle tandis que les autres en seront la base d'apprentissage au cours des itérations de cet algorithme.

21. traduit par forêt aléatoire

A noter que le partionnement de cette base doit permettre de garantir une indépendance entre les parties créées.

Illustration de la k-fold cross validation. ***Source :*** *stackoverflow.com*

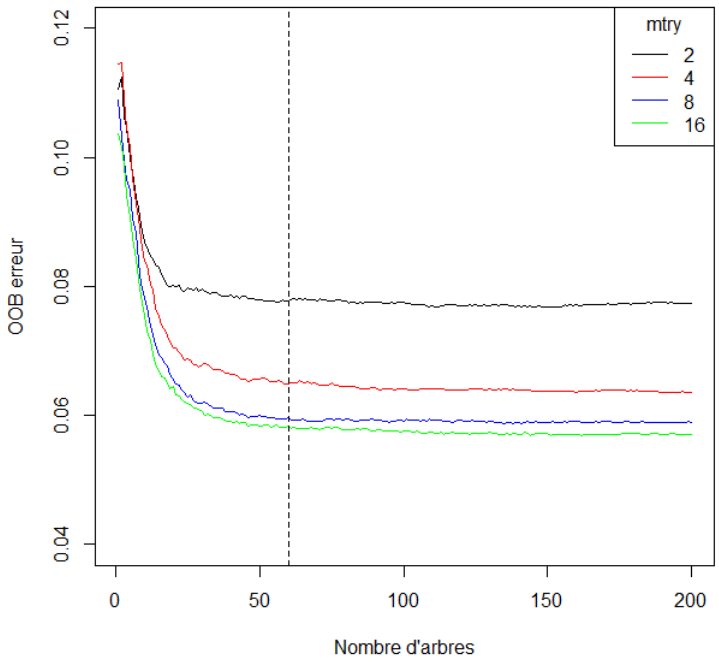
3.5.2 Modèle par forêts aléatoires : application

Une validation croisée en k-fold (avec k=5) comme décrite auparavant a été réalisée pour déterminer les hyper-paramètres optimaux au sens de l'erreur Out-Of-Bag [22]. La validation croisée permet d'éviter de biaiser la valeur de ces paramètres.

Les deux hyper-paramètres à optimiser,ici, sont le nombre d'arbres et le nombre de variables "candidates" à chaque *split* (noté mtry) (discutée dans la partie théorique du Random Forest).

De ce fait, le nombre d'arbres va être variable avec un mtry constant tout en restant raisonnable quant au temps de calcul de l'algorithme et vigilant quant au sur-apprentissage éventuel.

22. Out-Of-Bag correspond aux observations non pris en compte dans l'échantillon bootstrap



L'erreur Out-of-Bag en fonction du nombre d'arbres et du mtry.

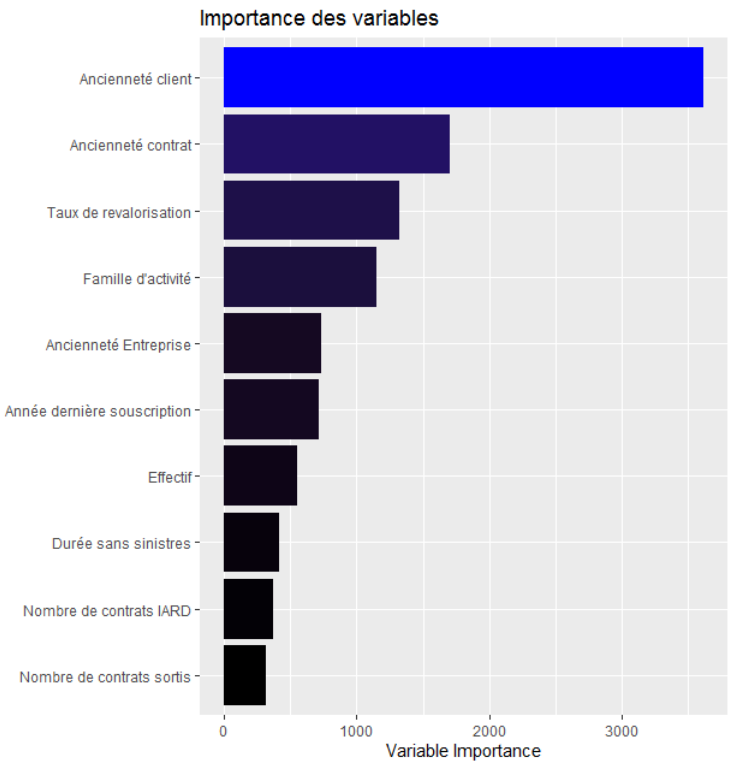
Les hyper-paramètres choisis permet d'éviter tout sur-apprentissage en effet avoir un fort mtry fait que le nombre de variables candidates est important et donc que les mêmes variables (à fort pouvoir explicatif) se retrouvent dans de nombreux arbres, ceci mettrait à mal l'hypothèse d'arbres non corrélés ainsi le sur-apprentissage sera plus fort.

Les paramètres sont les suivants : 60 arbres et 8 variables candidates par split.

Compte tenu de la faible proportion de la modalité résilié du contrat, une méthode de *oversampling* a été implémentée.
Elle consiste à créer des doublons pour la modalité sous-représentée (résilié) pour permettre de mieux apprendre sur cette modalité mais elle augmente le risque d'overfitting.

La modalité "résilié" a été forcée à représenter 30% de la base d'étude. Cependant, après validation croisée, ceci n'apporte qu'un gain très faible d'AUC et donc ne sera pas gardé pour la suite.

Une fois notre modèle établi, compte tenu de la faible explicabilité de la méthode seulement l'importance des variables par l'index de Gini va être affichée.



Le graphique ci-dessus montre l'importance des 10 variables les plus explicatives du modèle sur la probabilité de résiliation. On notera que ici la variable "Famille d'activité" n'a pas été segmenté puisque la forêt aléatoire n'a pas besoin de segmentation et l'explicabilité étant réduite cette segmentation n'a pas lieu d'être.

Cependant, d'autres segmentations ont été gardées sans perte notable de performance.

Les variables les plus importantes sont les variables d'ancienneté et le taux de revalorisation déjà importantes dans les précédents modèles. La variable de la famille d'activité est devenue ici plus influente que précédemment.

Il est important de voir que le produit n'est plus aussi explicatif notamment la modalité "Ancienne Génération". Ceci peut s'expliquer par le fait que ce produit n'est plus commercialisé et donc les clients ayant ce produit ont une ancienneté élevée.

De ce fait, le pouvoir prédictif de cette modalité a été pris par les variables d'ancienneté. Ceci n'a pas été visible dans les modèles statistiques car l'ancienneté (client et contrat) a un effet contraire à la résiliation et donc les produits en run-off ont été traités séparément en conférant à la modalité Produit une significativité que l'on ne retrouve pas ici.

En effet, ici il n'est pas question de linéarité ou même de propension à résilier ou non mais

seulement d'influence ce qui peut expliquer les changements relevés. A ajouter également, que les contrats en run-off ne représente qu'une infime partie du portefeuille.

Regardons maintenant les résultats du modèle :

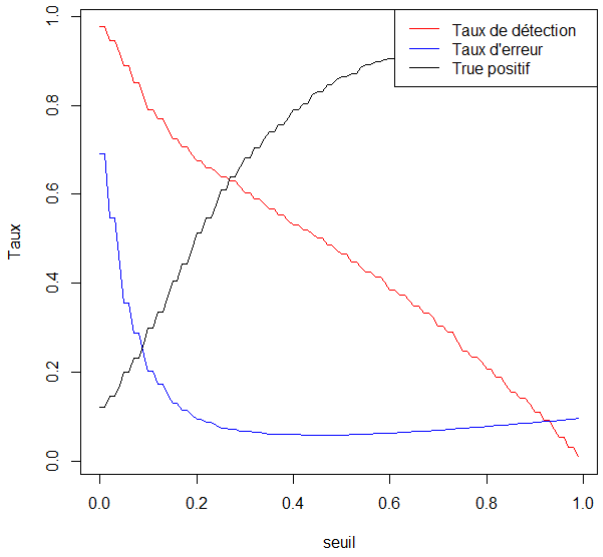
		Prédictions	
		0	1
Etats	0	16929	140
	1	986	867

Matrice de confusion des prédictions de l'échantillon Test avec un seuil de probabilité sans arbitrage.

Cette matrice de confusion n'est qu'indicative car en effet elle dépend du seuil de probabilité à partir duquel on assigne la valeur 1 ou 0 à la prédiction.

En augmentant le seuil, plus de contrats se verront attribuer la modalité "résilié" mais le taux de vrai positif diminuera en même temps. Une pondération peut être faite entre se tromper en attribuant un caractère de résiliation à un futur non-résilié et celui de ne pas détecter une résiliation future.

Voici un graphique permettant de se rendre compte de l'arbitrage à faire.



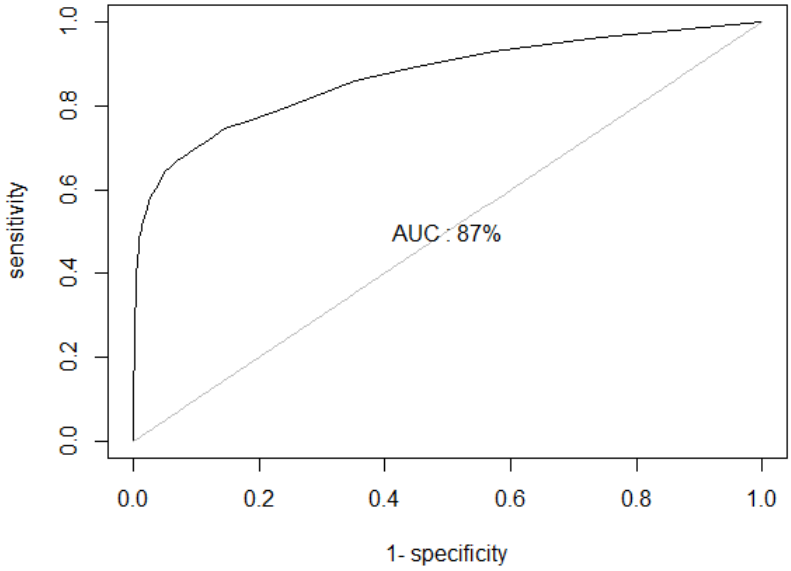
Graphique montrant l'évolution des taux de détection, d'erreur et de vrai positif des prédictions en fonction du seuil de probabilité.

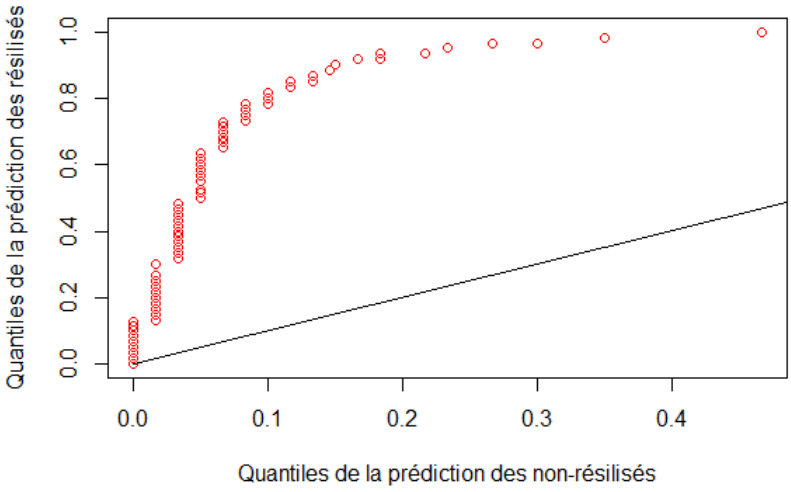
On appelle,ici le taux de détection : le rapport du nombre de contrats prédits comme

résiliés et du nombre de contrats résiliés, le taux d'erreur étant le nombre de contrats mal prédits divisé par le nombre de contrats total.

Cependant, les probabilités nous suffisent pour cette étude qui sera utile pour effectuer des calculs d'espérance de rentabilité et de score de fragilité.

Intéressons-nous pour pouvoir réellement comparer les modèles à la courbe ROC et à son indicateur associé l'AUC.





L'AUC dénote du fort pouvoir prédictif de ce modèle mais le coût que cela engendre est le peu d'explications qu'il nous fournit : absence ou difficile estimation de coefficients

marginaux des variables explicatives et l'incertitude qui les accompagne.

Notons ici qu'un autre modèle d'apprentissage automatisé a été implémenté : le gradient boosting qui fournissait des résultats similaires à la forêt aléatoire.

3.6 Choix du modèle le plus adapté

Le modèle le plus performant est sans nul doute l'apprentissage automatisé cependant, compte tenu des attentes d'explicabilité que requiert l'application opérationnelle de cette étude, le modèle additif généralisée semble être un compromis acceptable.

En effet, ce dernier est plus performant qu'une régression logistique et moins opaque que les forêts aléatoires puisqu'il apporte des indications quantitatives quant à l'effet marginal de chaque variable.

Dans un second temps, où la maîtrise de la résiliation sur ce produit sera plus grande, les modèles les plus performants pourront être utilisés en toute confiance dans le calcul de la Valeur.

3.7 Création d'une table de lois de chute

L'objectif est d'agréger les probabilités résultantes du modèle choisi selon quelques variables afin de créer une table de lois de chute.

Ces lois de chute permettent de projeter les probabilités de résiliations sur plusieurs années.

Pour faire cela, la solution préconisée et expliquée au début de l'étude est d'agréger ces probabilités de résiliation sous un an afin que, dans les groupes, la variable d'ancienneté du contrat sert à créer les probabilités de résiliation à la maille du groupe sur de nombreuses années comme des probabilités conditionnelles de résiliation à la présence du contrat à l'année n. L'agrégation sera donc par maille et par ancienneté de contrat (discrétisée en année entière) la moyenne des probabilités de résiliation sous un an.

Ces probabilités pour chaque maille serviront à calculer les marges techniques et financières futures pondérées par la probabilité de rétention.

Ce regroupement des contrats doit être assez fin pour minimiser la variance intra-groupe de

la probabilité sinon l'élaboration du modèle sous-jacent serait moins utile. Or, un maillage trop fin ne permettra pas d'avoir des valeurs d'ancienneté de contrat assez étendues pour projeter sur le long terme.

Un compromis doit également être fait sur les variables du maillage pour minimiser la variance des clusters créés.

Pour ce faire, le maillage va être réalisé avec un arbre de classification selon les variables du modèle du stock sans intégrer bien entendu la variable d'ancienneté du contrat.

L'arbre ainsi que le reste de l'élaboration des lois de chute agrégées n'inclue pas les Affaires Nouvelles car en effet, ces contrats ne peuvent être regroupés par les variables les plus discriminantes du modèle. En effet, des variables comme le contentieux ou le taux de revalorisation ne concernent pas les Affaires Nouvelles ainsi inclure ces contrats et donc enlever les variables citées dans l'agrégation des probabilités augmentera trop fortement la variance intra-groupe.

Ceci n'est pas un problème car la première année du contrats étant très particulière sera étudiée non pas de manière agrégée mais de manière personnalisée.

Rappelons que l'objectif central de ce regroupement est de visualiser le portefeuille de manière macroscopique mais surtout de projeter la propension des assurés à résilier avec des courbes lissées qui feront l'objet de la partie suivante.

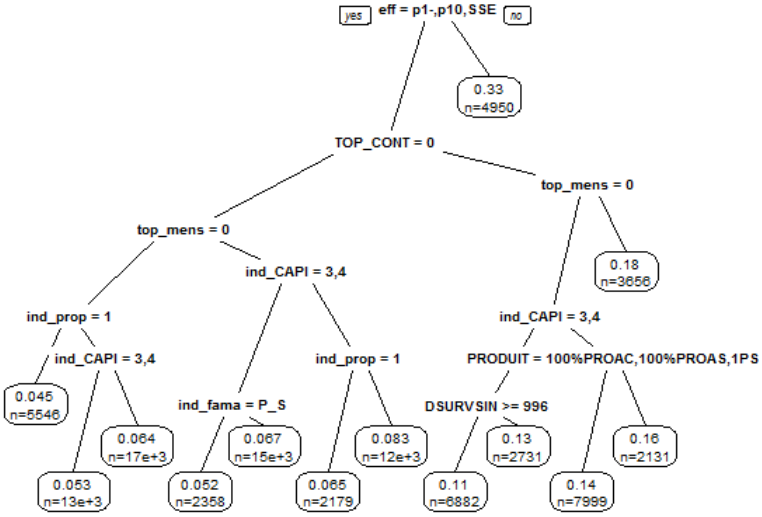
La résiliation des Affaires Nouvelles sera étudiée sans lissage et avec une vision unique à la fin de l'année.

L'arbre ici sera paramétré par le nombre d'observations par feuille (nombre d'observations minimal par cluster) et la profondeur maximale (nombre de critères maximal pour former chaque maille).

Arbitrairement, la profondeur maximale sera fixée à 5 et le nombre d'observations minimales par groupe à 2000 afin d'avoir tout de même un nombre suffisant d'année d'ancienneté.

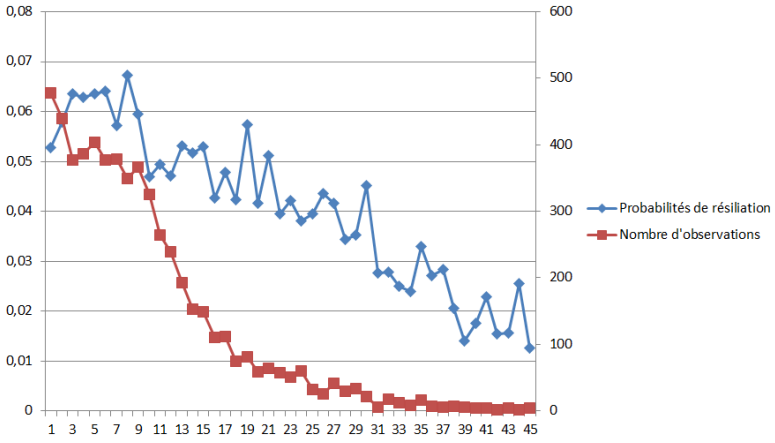
A titre d'exemple, il a été choisi les résultats du modèle de régression logistique afin de créer le maillage suivant.

Présentons d'abord l'arbre de classification où l'on retrouve naturellement les variables les plus impactantes du modèle.



où TOP_CONT est l'indicatrice de contentieux, eff est l'effectif de l'entreprise avec "p1-,p10" les modalités : 1 à 9 salariés et plus de 10 salariés, DSURVSIN l'année du dernier sinistre,ind_prop est l'indicatrice de propriété de l'assuré,ind_CAPI le cluster des capitaux assuré (3-4 sont les deux plus élevés) et ind_fama est la variable sur le secteur d'activité.

Les feuilles renseignent le nombre d'observations présentes et la moyenne des prédictions.



Probabilités de résiliation sous un an conditionnées à l'ancienneté du contrat en fonction de cette même ancienneté sur une maille du portefeuille.

Le nombre de clusters ainsi créé est de 13 dont la répartition est contenue dans l'arbre de régression. Ceci permettra de remplir une table de lois de chute par maille sur laquelle peut se baser la projection des marges dans le cadre du calcul de la valeur du contrat MRC.

3.8 Ajustement de lois paramétriques

Il serait intéressant de se demander si la granularité annuelle de ces probabilités est souhaitable. Il nous est possible d'être plus fin dans l'agrégation des probabilités en agrégeant par mensualité plutôt que par année. Mais une idée serait d'extrapoler ces résultats en ajustant une distribution de probabilité usuelle sur ces probabilités modélisées. Ceci permettrait de pouvoir estimer les mouvements de manière continue donc d'avoir des probabilités de résiliation pour tout horizon.

Cette méthode a un coût, c'est celle de l'erreur d'ajustement de la distribution aux données. Ceci constituera la suite de cette partie.

Rappelons tout de même qu'un ajustement correct peut être d'une grande utilité tant l'on possède d'information sur ce genre de distribution. Une distribution usuelle telle que la loi gaussienne, gamma, logistique parmi tant d'autres sont continues, certaines ont des moments facilemment étudiables. En outre, cela permettrait de prévoir de manière encore plus fine le comportement de l'assuré.

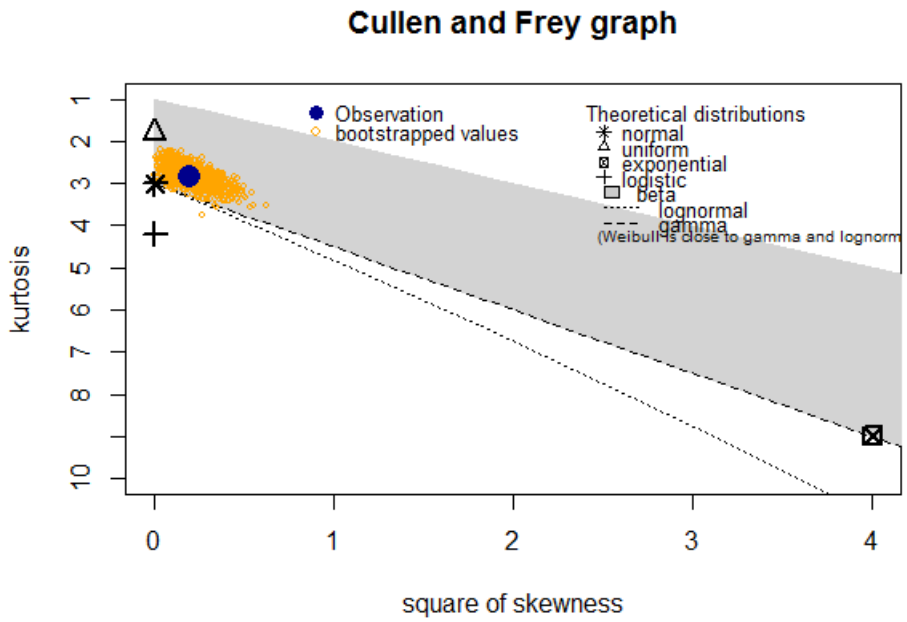
Ceci étant rappelé, l'agrégation se fera mensuellement par rapport à l'ancienneté du contrat puisqu'un ajustement sur seulement une vingtaine de points (une vingtaine d'années) peut être critique. Le versant d'une agrégation plus fine est l'augmentation de la volatilité des probabilités agrégées et somme toute le manque de sens du résultat. Cependant, cette plus fine agrégation est rendue possible par les contraintes de volumétrie imposées à l'arbre de classification.

De ce fait, nous nous retrouvons avec plusieurs centaines de points par maille (les points n'ayant que peu d'observations ont été enlevés) et des distributions à ajuster.

Il s'agit ici de ne pas entrer dans un test de toutes les lois d'ajustement possibles et imaginables mais seulement celles qui nous apporterait des informations faciles d'accès sur la résiliation (distributions continues, à moments définis, ...).

Pour ce faire, pour chaque maille [23], sera étudié le graphique de Cullen et Frey :

23. Seulement 13 mailles ont été créées.



Graphique de Cullen et Frey présentant le coefficient d'aplatissement ***kurtosis*** *en fonction du carré du coefficient d'asymétrie* ***skewness*** *d'une feuille du portefeuille.*

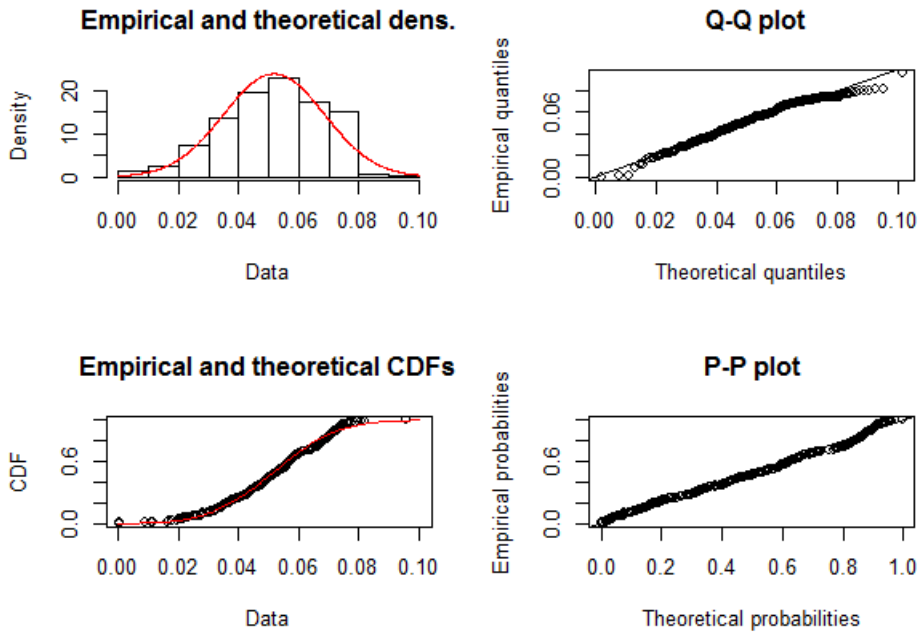
Ce graphique permet de réduire le nombre de distributions candidates pour s'ajuster aux données.

On notera que les observations ont été rééchantillonnées pour assurer la robustesse du graphique.

Après avoir effectué cette sélection, nous devons prendre un indicateur permettant de choisir la distribution adéquate. L'AIC est choisi.

Restons sur l'exemple de cette maille et après le test de la distribution bêta, logistique et même gamma la distribution gaussienne obtient le meilleur AIC et est donc choisie. L'ajustement et la recherche de paramètres ont été faits par maximum de vraisemblance.

Présentons alors les résultats de cet ajustement afin de savoir s'il est pertinent dans l'absolu :



Graphiques présentant l'ajustement de la loi normale aux probabilités de la maille considérée.

On rappelle seulement que le premier graphique représente la juxtaposition des densités théorique et empirique, le deuxième est un diagramme quantile-quantile des des données théoriques et empirique dont l'idéal est l'alignement sur la première bissectrice. Les deux suivants sont la juxtaposition des fonctions de répartition des résultats empiriques et théoriques et un digramme probabilités empiriques - probabilités théoriques.

L'ajustement est très bon et la p-value du test de Kolmogorov-Smirnov indique une valeur de 10.5% ce qui n'infirme pas l'hypothèse nulle qui est que les données suivent une loi normale de paramètres : environ 0,0517 pour l'espérance et de 0,0168 pour l'écart-type.

Ce même travail a été fait pour toutes les mailles du portefeuille et les lois d'ajustement suivantes ont été les plus courantes : la loi logistique (dans la quasi totalité des cas) et la loi normale.

Ceci peut s'expliquer par le fait que le modèle sous-jacent est une régression logistique, cependant rappelons qu'un même travail peut être fait avec un autre des modèles présentés précédemment.

Ainsi avec les résultats des modèles statistiques des Affaires Nouvelles et les lois de chute ajustées par maille des contrats en stock, nous avons les outils nécessaires pour répondre

à la problématique initiale sur l'étude de la résiliation.

3.9 Applications possibles

Rappelons que plusieurs types de modèles ont été implémentés selon le degré de transparence et de performance souhaité et de manière différenciée entre les Affaires Nouvelles et les contrats en stock.

Ensuite des clusters ont été créés permettant ainsi que créer non plus des probabilités de résiliation sous un an mais de projeter des probabilités moyennes sur une vingtaine d'années.

Pour finir, des distributions ont été ajustées sur chacune des mailles créées.

Ce travail dans sa globalité apporte différents niveaux de compréhension de la résiliation et permet de fournir des informations quantitatives utiles à l'entreprise.

Ceci pourra servir comme indicateur aux intermédiaires afin d'adapter leur politique commerciale notamment par rapport aux clients multi-équipés et à forte valeur afin de les garder dans leur portefeuille. Un score de fragilité facilement actualisable pourra ainsi être créé, ce dernier permettra d'établir des actions à réaliser afin de consolider la présence du client dans le portefeuille ou de l'inciter à résilier si celui-ci n'est pas rentable.

La plus importante utilisation est celle de la projections des marges techniques et financières du contrat MRC où les lois de chute jouent un rôle prépondérant [24].

24. Voir partie consacrée

4 Applications réalisées

4.1 Vision de la rentabilité revue

Une fois le travail réalisé, la question était de se questionner sur la vision de la rentabilité choisie avant sa projection par les lois de chute créées. En effet, bien que les probabilités de résiliation ont un fort impact dans la valeur, la rentabilité du contrat (ici MRC) est la base de l'indicateur.

La problématique centrale est la suivante : prendre un S/P modélisé ou "historique". Si l'on choisit des S/P empiriques, la volatilité de la valeur contrat sera forte par le fait qu'il dépendra essentiellement du caractère aléatoire du risque.

De manière plus concrète, un client subissant un sinistre "grave" peu de temps après la date d'effet de son contrat aura naturellement une valeur contrat fortement négative même si ce dernier a une importante duration, les majorations successives ne permettront pas de lui accorder une valeur décente. Ce dernier est peut être un client rentable dans le temps mais son premier sinistre biaise trop fortement la modélisation pour le savoir.

Le contraire est également possible, surtout pour un produit dont les sinistres sont relativement rares : beaucoup de clients se retrouveront avec un S/P à 0 n'ayant pas assez d'ancienneté pour être réellement exposés au risque. Pour un produit, comme La Santé par exemple, la question serait moins ardue car les sinistres sont assez fréquents et donc permettent en peu de temps de se figurer du profil du client et donc de diminuer la volatilité de la sinistralité.

En ce qui concerne le premier exemple, une façon de faire, récurrente en tarification, serait d'enlever les "graves" de la sinistralité mais deux questions surviennent la première étant de savoir si ce type de sinistre est inhérent au client et à son comportement et la deuxième est la non-résolution du problème soulevé par le second exemple.

Si l'on choisit alors la version modélisée par les équipes de tarification du S/P, alors l'on considérera seulement quelques S/P modélisés selon quelques variables. Un S/P modélisé concernera un sous-groupe du portefeuille, le risque étant qu'il ne dénotera que peu du comportement particulier du contrat, l'on perdrait de la variance dans les valeurs

du contrat : tous les contrats du sous-groupe aurait sensiblement la même valeur et la segmentation des clients n'en serait que moins pertinente.

Un autre problème se pose : la valeur a un but d'aide à la décision et est exposée à la distribution de ce fait, ces derniers remontent souvent le fait que la rentabilité affichée dans leur outil ne correspond pas au parcours du client dont ils gèrent le contrat. Ceci crée de la défiance vis-à-vis des travaux entrepris et diminue les impacts opérationnels de notre indicateur.

Avant ces travaux, la vision prise était celle d'un S/P moyen pour le produit dans sa globalité.

Le problème se résume de la façon suivante : travailler avec la sinistralité empirique du contrat ajoute des problèmes avec les "nouveaux" contrats ainsi qu'un manque de robustesse dans l'évaluation de sa valeur et travailler avec des S/P modélisés sur des sous-groupes trop importants du portefeuille améliore en effet la robustesse mais néglige la singularité de chaque client.

Ce type de réflexion est contenu dans la théorie de la crédibilité où il est question de juger pertinent ou non de l'expérience individuelle par rapport à l'expérience collective. Des études sur ce sujet sont réalisées au sein des bureaux d'études techniques.

Une manière simplifiée de résoudre ce problème appelant à de futures améliorations a été mise en place. En effet, la rentabilité individuelle est jugée pertinente qu'après un certain laps de temps qui lui-même dépend des caractéristiques du contrat. Avant cela, il s'agira d'un S/P d'achat autrement dit modélisé.

La rentabilité est ici un pré-requis au travail exposé et est fournie par les équipes de tarification mais par son importance, il était important d'en discuter la vision.

Cette discussion a été menée et la réception des S/P a été validée avec les équipes concernées.

La question des frais et commissions a également été traitée et ces derniers ont été inclus dans le calcul de la valeur alors même que ces derniers n'étaient pas entièrement pris en compte auparavant.

Dans l'optique de prendre en compte la duration du contrat de manière plus importante

et au vu de la politique actuelle de revalorisation qui sera faite après cette période de crise, des hypothèses de revalorisation minimum ont été ajoutées au taux d'inflation des primes.

De ce fait, un contrat avec un ratio combiné légèrement supérieur à 1 pourra avoir une valeur positive si sa duration est importante. L'ensemble de ses contrats sera appelé "zone grise" en raison de la non-adéquation de la rentabilité du contrat avec les bureaux d'études techniques.

4.2 Application opérationnelle des lois de chute

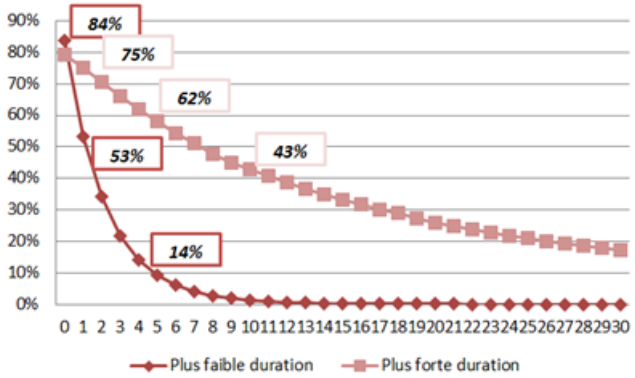
4.2.1 Résultats et discutions

L'implémentation des lois de chute dans le calcul de la valeur a été présentée à toutes les équipes concernés et réalisée dans les chaînes de calcul SAS. Elle impacte la revalorisation [25], la visualisation du portefeuille et influence la politique de rabais à la souscription.

Les précédentes visions de la rentabilité étaient anciennes et les lois de chute devenaient désuètes c'est pour cette raison que la valeur contrat MRC a sensiblement changé.

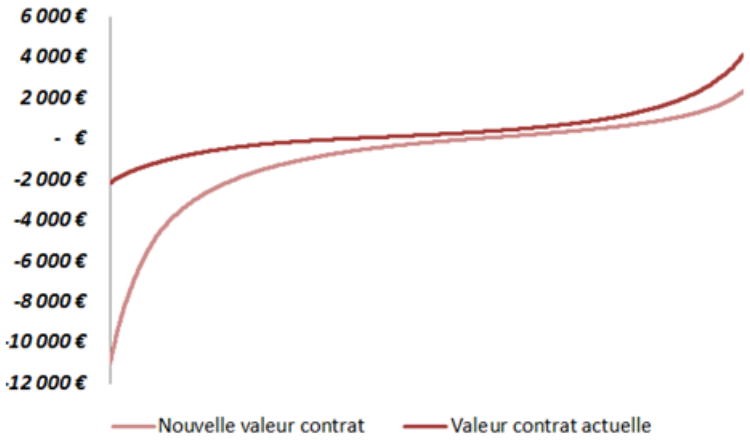
Présentons alors les résultats de la nouvelle valeur en comparaison avec l'ancienne.

Affichons d'abord les deux plus extrêmes lois de chute calculées, pour améliorer la lisibilité du propos ces dernières seront des fonctions de survie en portefeuille [26] :



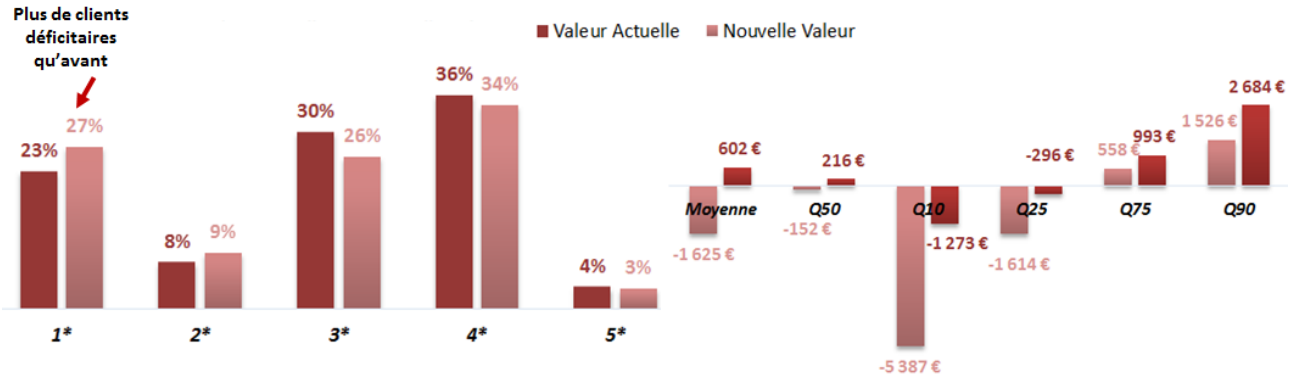
Présentons maintenant les distributions de valeurs ainsi obtenues :

25. Par le biais de "bouclier" de revalorisation pour le segment des clients à forte valeur, de revalorisations plus importantes pour les clients destructeurs de valeur.
26. La version cumulative des probabilités de rétention sous un an, cette dernière comme une fonction de répartition caractérise entièrement la loi.



On notera une baisse significative de la valeur avec un poids non négligeable des clients à valeur fortement négative dans le portefeuille (ces derniers ont déjà été ciblés par les équipes de tarification et feront l'objet d'une forte revalorisation).

L'impact sur la segmentation des clients selon le système d'étoiles présenté précédemment est la suivante :



Le changement de valeur impacte notamment les clients destructeurs de valeur dont le nombre augmente de 4 points au détriment des anciens 3 et 4 étoiles, cela va permettre d'entamer une politique de revalorisation ciblée non pas au niveau du contrat, chose que les bureaux d'études techniques font déjà, mais au niveau du client. Des clients avec une rentabilité dégradée mais ayant de nombreux contrats à valeur seront donc épargnés afin de maximiser les marges techniques et financières dans le futur.

Les discussions avec les différentes équipes concernées ont eu pour résultante un accord de revalorisation supposée a priori dans les projections de 3% et a confirmé les hypothèses

économiques et financières effectuées.

Le sous-groupe de clients ayant une valeur contradictoire avec celle des bureaux d'études techniques représente environ 10% du portefeuille. Cela est dû à la duration supposée des clients dont les contrats s'améliore avec le temps dans le portefeuille. Une divergence de cet ordre a toujours été constatée cependant maintenant, elle est totalement assumée et maîtrisée. En effet, elle ne concerne que les contrats dont le ratio combiné se trouve entre 1 et 1,07 avec une moyenne de 1,03 donc relativement proche de la limite de rentabilité de 1.

4.2.2 Migration vers Hadoop initiée

Le service fonctionnant essentiellement sur des chaînes SAS avec une actualisation des données mensuelles, des projets ont été lancés afin de transcrire dans un premier temps le calcul à proprement parler avec des diverses agrégations et le rendu aux différentes équipes sur Hadoop. [27] De plus les bases de données sur Hadoop sont pour la plupart actualisées quotidiennement ce qui permet, surtout pour le cas de clients réalisant des mouvements durant une courte période de temps, d'obtenir un indicateur plus pertinent et plus réactif.

Cette revue a été pensée et codée de façon à être facilement exportable sur ce type de support.

La migration de la valeur sera l'un de mes projets cette année et la MRC pourra être l'un des produits les plus rapidement opérationnels sur cette plateforme.

27. Hadoop,framework Open Source, est conçu pour stocker, analyser et manipuler des volumes de données importants, il est un élément clé actuellement du traitement Big Data.

5 Conclusion

Ce mémoire nous donne plusieurs niveaux de réponse à la problématique sous-jacente ce travail et propose des outils concrets pour une mise en production. Des applications de cette étude ont également été citées.

Tout au long de ce mémoire, les hypothèses de modélisation ont été débattues afin d'obtenir le modèle le plus approprié possible.

Pas à pas, des hypothèses ont été supprimées rendant la modélisation de plus en plus performante mais de moins en moins explicable.

La première modélisation par régression logistique a laissée place à un mélange de régressions logistiques afin de généraliser son propos et d'améliorer ses performances.

Puis l'hypothèse de linéarité du modèle linéaire généralisé a été remplacée par une hypothèse moins forte d'additivité malgré tout de même un choix particulier de fonctions qui sont les splines cubiques. Et enfin, afin de prendre en compte davantage cette non-linéarité et d'intégrer les interactions entre les variables, une modélisation basée sur l'apprentissage automatique a été étudiée avec le modèle de forêt aléatoire.

Les données des contrats en stock ont été les plus traitées et représentent la grande majorité des contrats. Ces observations se prêtaient bien à ce type de travail puisque la performance des modèles augmente en même temps que l'abandon progressif des hypothèses initiales. Leur caractère hautement non linéaire donne du sens à ce raisonnement.

Une fois cette modélisation réalisée, une agrégation des résultats a permis la création de lois de chute et leur ajustement à des distributions usuelles pour encore une fois apporter une réponse à la problématique plus profonde et même plus opérationnelle.

De ce fait, ce mémoire permet de choisir la modélisation offrant le meilleur compromis entre performance et complexité par l'analyse des avantages et inconvénients de chacun des modèles. Sans être exhaustif sur les modélisations à réaliser pour ce type de travail, il a tout de même permis de présenter et de comparer les modélisations les plus classiques.

Les applications de cette étude ont été citées mais rappelons seulement que les probabilités

de résiliation sous un an ou les lois de chute peuvent faire l'objet d'applications concrètes.

Les principales étant leur intervention dans le calcul de la valeur des contrats Multi-Risques Commerce comme pondérateur des marges techniques et financières futures mais également dans la création de score de fragilité pouvant être l'activateur d'offres commerciales pour les clients à forte valeur ou plus simplement être un outil de négociation pour les intermédiaires.

La valeur potentielle, bien que peu étudiée dans ce mémoire, n'est pas pour autant omise. Il est question de l'intégrer pleinement dans les applications mentionnées. Cependant, une réflexion sérieuse de ce sujet nécessiterait un travail à part entière. Nous pouvons tout de même en dire quelques mots. Le calcul de cette valeur additionnelle a été balayé précédemment, il repose sur une modélisation de l'appétence des clients à souscrire un certain type de contrat compte-tenu de leur détention actuelle. Une étude est menée sur ce sujet par l'équipe dans laquelle j'évolue en étroite collaboration avec le Data Lab, équipe sous la même direction que la nôtre. La question est d'améliorer les modèles déjà existants pour leur donner une plus grande visibilité et accroître leur champ d'application, en comblant certaines lacunes qu'ils comportent.

L'une d'entre elles étant l'ordre de grandeur des probabilités calculées, en effet la valeur potentielle prend généralement des valeurs trop faibles par rapport à la valeur actuelle des clients. Les explications sont multiples : les modèles (XGBoost) actuels n'incluent de très peu de règles "métier" ce qui tend à calculer des probabilités d'appétence à un produit précis à des clients ne pouvant par nature y accéder et donc diminue fortement l'espérance des probabilités. Une autre explication est que la modélisation ne comprend pas de seuils d'appétence mais évolue de manière continue, en d'autres termes, une possible amélioration de ces modélisations est d'envisager un client comme étant appétant quand la probabilité de multi-équipement à un produit est supérieure à un certain seuil. Un seuil qui serait évalué en minimisant l'erreur de prédiction à partir de données antérieures où le multi-équipement serait observé sur une période de temps beaucoup plus longue.

Un autre désavantage actuel de celle-ci est le manque d'explicabilité et la difficulté de maintenance en effet les modèles existants prennent en entrée un très grand nombre de variables explicatives qui ne sont pas forcément bien renseignées ou à jour.

Pourtant, la valeur potentielle revêt une importance notable dans la segmentation du client et peut être un indicateur très puissant dans nos process si l'on acquière suffisamment de maîtrise sur ce sujet. Elle peut être par exemple à l'initiative de nouveaux "packs" de produits comme il en existe déjà, avec une plus grande précision quant au résultat attendu ou encore nous faire susciter de l'intérêt pour un sous-groupe précis du portefeuille de clients qui hésite à se multi-équiper et qui aurait besoin d'une action de notre part pour le faire.

Ces deux indicateurs complémentaires coordonnés avec pour chacun des hypothèses raisonnables en accord avec tous les bureaux d'études techniques, un calcul sérieux, un support novateur (Hadoop), des données actualisées quotidiennement et un code suffisamment souple pour évoluer en même temps que nos idées apporteraient énormément à une compagnie d'assurance comme Generali tant cela permet d'avoir une maîtrise exacerbée de son portefeuille et de ses mouvements, d'initier des actions précises et pertinentes en outre de mettre le client au coeur de sa stratégie économique de manière sereine et réfléchie.

Ce mémoire tente de s'inscrire dans cet optique avec l'exemple d'un des produits clés de Generali France IARD qu'est le produit Multi-Risque Commerce.

6 Bibliographie

↪ **CANS C., LAVERGNE C. [1995]**, "De la régression logistique vers un modèle additif généralisé : un exemple d'application", *Revue de statistique appliquée*, tome 43, no 2, 77-90.

↪ **Périclès Actuarial**, "Machine Learning :Du GLM à l'arbre CART en passant par le Random Forest", article.

↪ **NAKACHE S. [2016]**, "Aperçu des méthodes de sélection de variables", article [en ligne] disponible sur "$https://lilbigdataboy.wordpress.com/2016/01/04/apercu-des-methodes-de-selection-de-variables-avec-r/$".

↪ **KHANEBOUBI M. [2016]**, "Introduction à Random Forest avec R", exposé [en ligne] disponible sur "$http://mehdikhaneboubi.free.fr/random\_forest\_r.html$".

↪ **DENIL M. et *al.* [2014]**, "Narrowing the Gap : Random Forests In Theory and In Practice", article.

↪ **BERTRAND F. [2010]**, "Choix du modèle", cours de l'Université de Strasbourg.

↪ **MAINDONALD J. [2010]**, "Smoothing Terms in GAM Models", article.

↪ **MILHAUD X. [2012]**, "Mélanges de GLMs et nombre de composantes : application au risque de rachat en Assurance Vie", thèse, 69-85.

↪ **GEIGLE T., VIGNAL G. [2016]**, "Modélisation de la valeur du client entreprise et utilisations opérationnelles", Mémoire, 2016.

↪ **BALLINA R. [2014]**, "Méthodes d'apprentissage appliquées à la tarification non-vie", mémoire.

↪ **NETO J. [2015]**, "Fitting Distributions", notebook [en ligne] disponible sur "$http://www.di.fc.ul.pt/\ jpn/r/distributions/fitting.html$".

↪ **GUILLOT A. [2015]**, "Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?", mémoire d'actuaire.

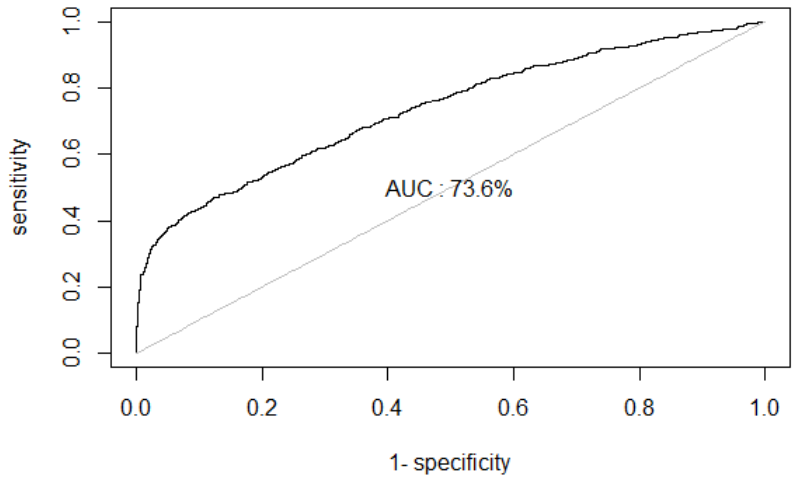
↪ **BEL L. et *al.* [2016]**, "Le Modèle Linéaire et ses Extensions", mémoire.

7 Annexes

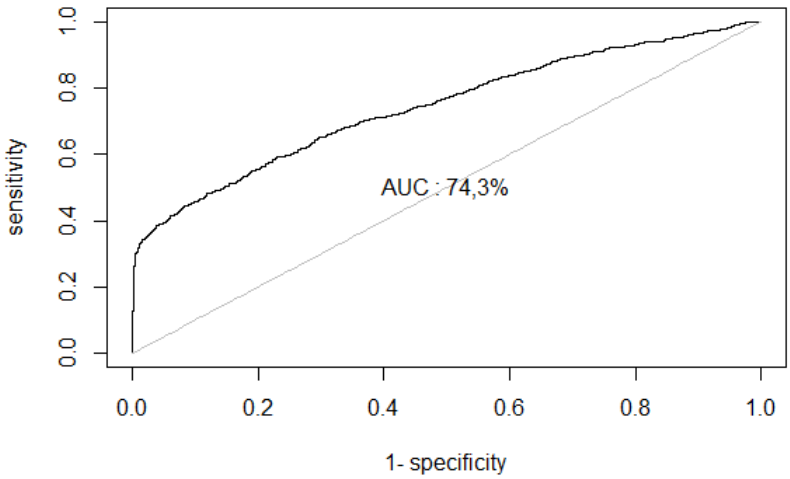
Modèle statistique pour les Affaires Nouvelles

Variables du modèle allégé	Coefficients	Erreur	Effets	P-value	Significativité
Constante	-0,433	6,66E-01	0,65	5,16E-01	
Chiffres d'affaire (moyen)	-0,248	6,01E-02	0,78	3,76E-05	***
Chiffres d'affaire (important)	-0,324	8,03E-02	0,72	5,60E-05	***
Chiffres d'affaire (très important)	-0,492	7,22E-02	0,61	9,59E-12	***
2-60 salariés	-0,726	3,69E-01	0,48	4,94E-02	*
60-100 salariés	-0,715	3,73E-01	0,49	5,48E-02	.
Plus de 100 salariés	-0,795	3,73E-01	0,45	3,30E-02	*
Nombre de contrats IARD	-0,145	1,51E-02	0,87	<2e-16	***
Nombre de contrats VIE	-0,227	1,14E-01	0,80	4,68E-02	*
Nombre de contrats sortis	0,043	4,71E-03	1,04	<2e-16	***
PRODUIT100%PROARTISANS-COMMERCANTS	0,603	5,46E-01	1,83	2,69E-01	
PRODUIT100%PROFABRICATION	0,367	5,70E-01	1,44	5,20E-01	
PRODUIT100%PROSERVICE	0,077	5,47E-01	1,08	8,89E-01	
Année du dernier contrat sorti	0,001	2,66E-05	1,00	<2e-16	***
ReseauCourtier	0,323	4,38E-02	1,38	1,75E-13	***
Ancienneté du client (en année)	-0,105	7,21E-03	0,90	<2e-16	***
Surface à assurer (moyenne)	-0,178	5,81E-02	0,84	2,26E-03	**
Surface à assurer (grande)	-0,212	6,47E-02	0,81	1,08E-03	**
Surface à assurer (très grande)	-0,151	7,70E-02	0,86	5,01E-02	.
Capitaux à assurer (moyen)	-0,163	5,91E-02	0,85	5,84E-03	**
Capitaux à assurer (important)	-0,445	7,22E-02	0,64	7,46E-10	***
Nombre de sites assurés	-0,032	1,62E-02	0,97	5,11E-02	.
Propriétaire	-0,130	4,34E-02	0,88	2,69E-03	**
Famille d'activité segment2	-0,217	6,54E-02	0,81	9,12E-04	***
Famille d'activité segment3	-0,593	9,13E-02	0,55	8,42E-11	***
Ancienneté du contrat	-0,443	7,43E-02	0,64	2,60E-09	***
Echelle de significativité	0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1			AIC : 14708	

Résultats de la régression logistique sur les contrats Affaires Nouvelles.



Courbe ROC de la régression logistique sur les contrats Affaires Nouvelles.



Courbe ROC du modèle GAM sur les contrats Affaires Nouvelles.

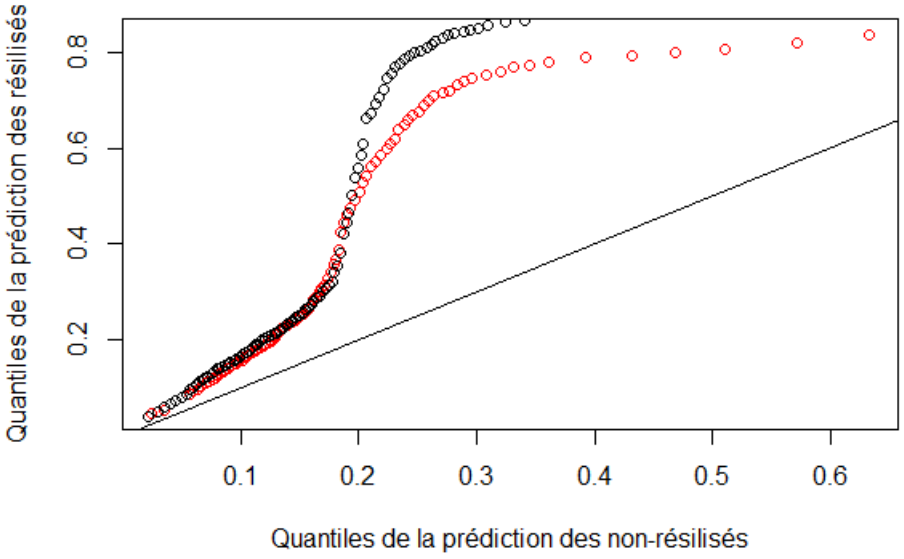


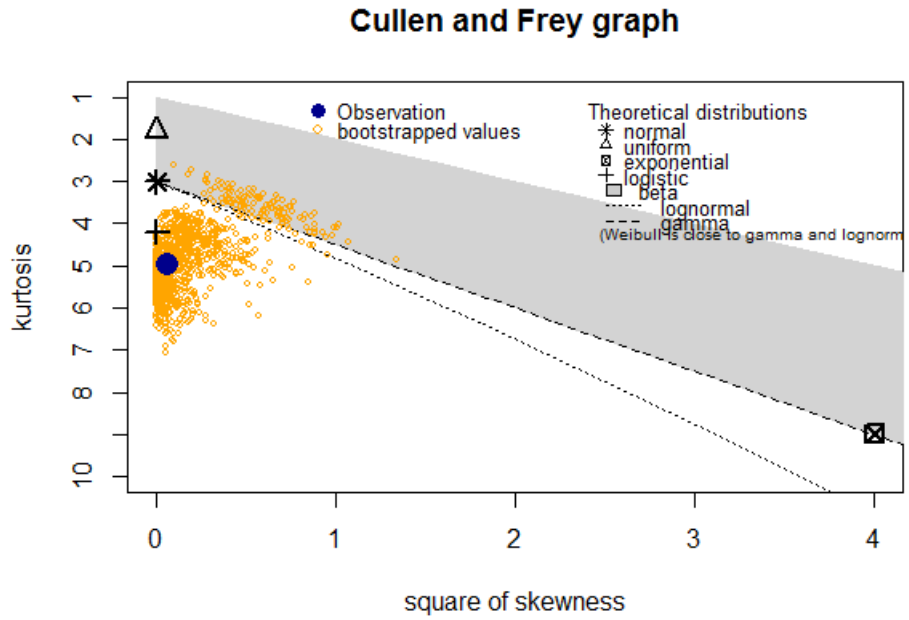
Diagramme quantile-quantile des prédictions selon l'état du contrat a posteriori. En rouge : prédiction de la régression logistique et en noir : prédiction de modèle additif généralisé

L'AIC du modèle GLM est de 14708 et celui du modèle GAM de 14166.

Compte tenu de la faible différence de performances entre ces deux modèles statistiques, la régression logistique a été préférée en ce qui concerne les contrats en Affaires Nouvelles.

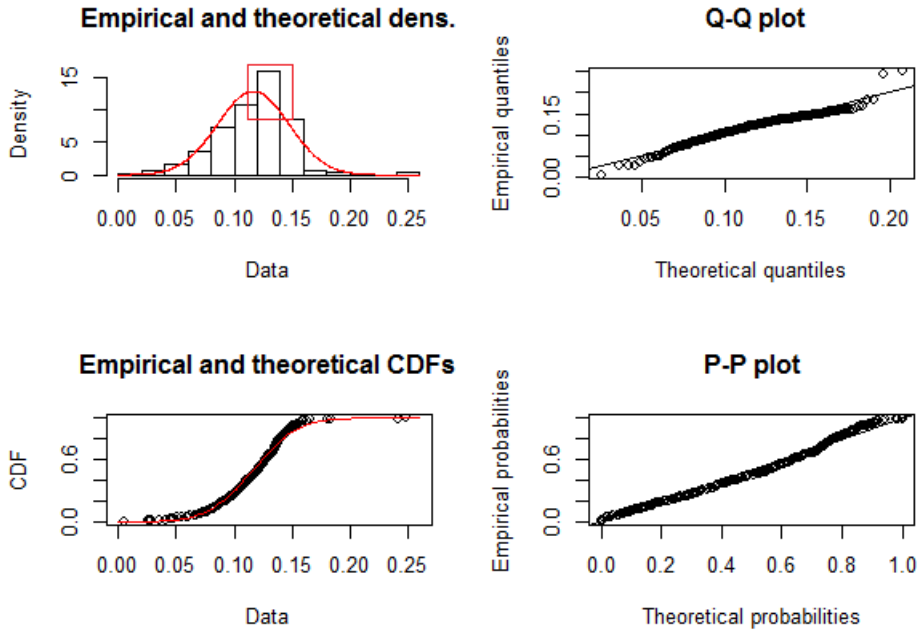
Présentation de l'ajustement d'une autre maille

Ajustement de la feuille 3 :

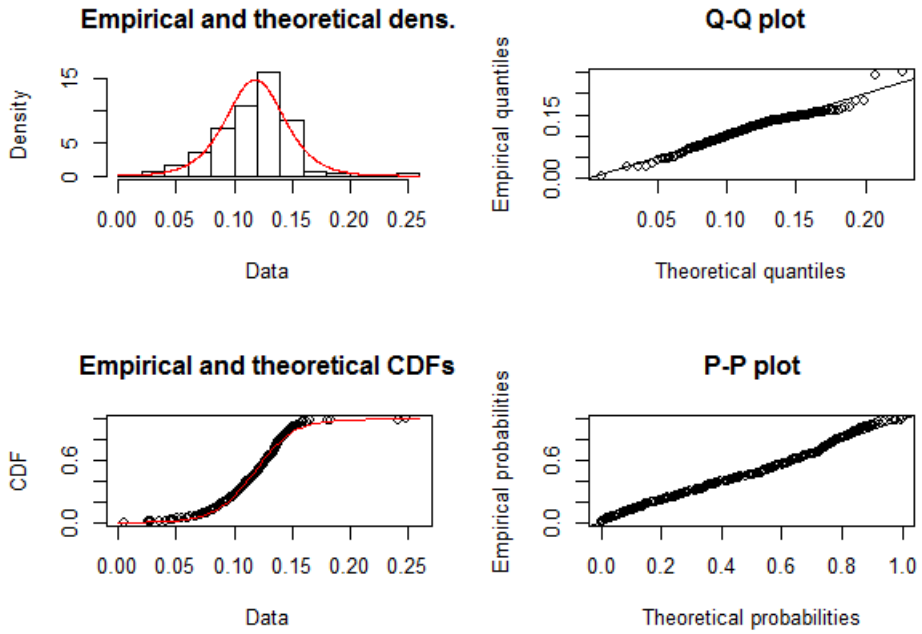


Graphique présentant le coefficient d'aplatissement ***kurtosis*** *en fonction du coefficient d'asymétrie* ***skewness****.*

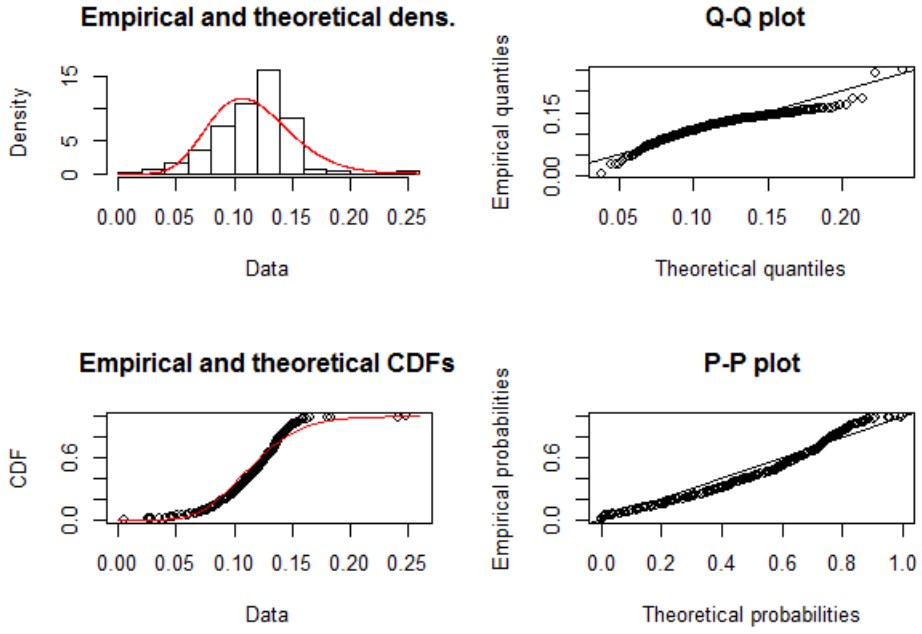
Les distributions envisagées sont la loi normale, logistique, beta (à cause de certaines valeurs de bootstrap) et la loi gamma.



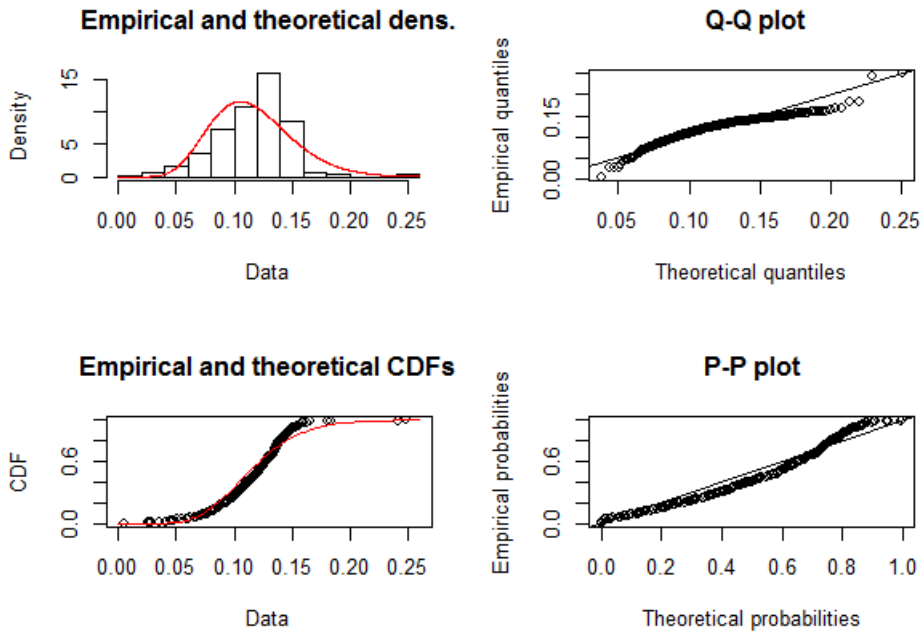
Graphiques présentant l'adéquation de la distribution gaussienne aux données de la feuille



Graphiques présentant l'adéquation de la distribution logistique aux données de la feuille



Graphiques présentant l'adéquation de la distribution bêta aux données de la feuille



Graphiques présentant l'adéquation de la distribution gamma aux données de la feuille

Les deux dernières lois s'ajustent sensiblement moins bien aux données pour les plus grandes valeurs de probabilités.

Il semble également que la distribution logistique s'adapte mieux au "pic" encadré en rouge sur le graphique des densités.

Regardons alors leur AIC et leur p-value du test de Kolmogorov-Smirnov.

	Normale	Logistique	Bêta	Gamma
AIC	-1276	-1291	-1214	-1204
P-value	12,46%	21,18%	0,40%	0,18%

Comparaison de l'AIC et de la p-value du test de Kolmogorov-Smirnov

Les distributions gaussienne et logistique sont possibles (p-value supérieure à 5%) mais nous garderons la distribution logistique comme loi d'ajustement.