Institut des ACTUAIRES

ENSAE

IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 08/11/2022

Par : Staelle TSOTSOP DJIMENE

Titre : # Application des algorithmes de Machine Learning pour la réduction des écarts actuariels d'expérience

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

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

Membres présents du jury de la filière

Nom : Hillairet Caroline

Membres présents du jury de l'Institut des Actuaires

Entreprise : ALLIANZ VIE

Allianz Iard
Direction Actuariat
Tour Neptune
Case courrier 1818
20, Place de Seine
92086 PARIS LE DEFENSE cedex

Signature :

Directeur du mémoire en entreprise :

Nom : MPELI MPELI Ulrich
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

TABLE DES MATIÈRES

REMERCIEMENTS

Je remercie Guillaume METGE, l'Assistante exécutive du CFO, de m'avoir accepté dans son équipe Reporting MVBS/NBV, ainsi que mon encadrant d'alternance MPELI MPELI Ulrich, pour son professionnalisme, sa pédagogie et sa disponibilité. Leur encadrement a été remarquable.

Je remercie également Nicolas BOURE, directeur de l'actuariat vie d'Allianz France de m'avoir donné l'opportunité de travailler au sein de ses équipes. Je tiens à remercier les équipes Reporting MVBS/NBV, Risk Modeling LH, inventaire épargne de Allianz France avec laquelle j'ai passé ma première année d'expérience professionnelle dans les meilleures conditions. Je me souviens de leur bienveillance et leur bonne humeur.

Je souhaite exprimer mes remerciements à mon tuteur de mémoire ENSAE Caroline Hillairet, pour ses conseils avisés et sa prise de recul sur mon mémoire qui n'ont pas manqués de me challenger. Une mention spéciale à tout le corps administratif et enseignant de l'ENSAE-IP Paris pour l'enseignement de qualité que j'ai eu la chance de recevoir durant ces deux années.

Enfin, je remercie ma famille ainsi que mes ami(e)s pour leur soutien moral et physique durant la rédaction de ce mémoire.

RÉSUMÉ

La contrainte principale des compagnies d'assurance est de respecter leurs engagements vis-à-vis des assurés. Les nouveaux référentiels comptables prudentiels obligent les acteurs du secteur d'assurance à développer des modèles mathématiques dits modèles actuariels pour valoriser les risques futurs inhérents à leur activité. Toutefois, il découle de ces modèles, des écarts d'expérience entre les comportements modélisés et réels des assurés ; ce qui pourrait mettre en péril l'équilibre financier de l'assureur.

Afin de réduire les écarts d'expérience sur les provisions, une analyse est effectuée sur 288 produits entre 2017 et 2020. Il en découle que les rachats totaux et partiels contribuent majoritairement aux écarts observés sur les Provisions. Ainsi sur ce poste, des groupes homogènes de produits sont construits afin de calibrer des taux de chocs sur les lois de rachats totaux, de manière à augmenter les rachats si le modèle les sous-estime et vice-versa. Cette approche permet de réduire de manière significative les écarts d'expérience sur les rachats et par conséquent sur les provisions.

Afin de proposer une approche Machine Learning pour la calibration de nouvelles lois de rachats totaux, deux produits sur lesquels les écarts d'expérience sont les plus importants sont retenus. Sur ces produits, les lois de rachats en nombre et en montant sont calibrées par réseau de distribution en utilisant les caractéristiques des assurés et des contrats sur la période de 2015 à 2021. Nos modèles de régression pénalisés, SVM, KNN et les forêts aléatoires révèlent que l'ancienneté, l'encours sur le contrat et la CSP des assurés sont les caractéristiques les plus importantes dans la prédiction des taux de rachats. Sur 2021, les lois prédites par nos modèles approchent mieux la réalité que celles du modèle interne, ce qui est confirmé par la réduction des écarts d'expérience sur les rachats et les provisions. Par ailleurs, l'approche d'apprentissage supervisée marche moins bien lorsque les lois de rachats sont calibrées suivant le volume d'encours sur les contrats.

Malgré les résultats assez satisfaisants découlant de cette étude, elle présente toutefois des limites liées aux hypothèses de modélisation. Une perspective de ce travail consisterait à analyser la contribution de la politique de participation aux bénéfices aux écarts d'expérience observés sur les provisions.

Mots clés : Écarts d'expérience, Machine Learning, rachats

ABSTRACT

The main constraint for insurance companies is to respect their commitments to policyholders. The new prudential accounting standards oblige the players in the insurance sector to develop mathematical models, known as actuarial models, to value the future risks inherent in their activity. However, these models result in differences in experience between the modelled and actual behaviour of policyholders, which could jeopardise the insurer's financial equilibrium.

In order to reduce experience gaps on life insurance reserves, an analysis is carried out on 288 products between 2017 and 2020. It was found that total and partial lapses contribute the most to the gaps observed in reserves. Thus, for this item, homogeneous groups of products are constructed in order to calibrate shock rates on total lapses laws, so as to increase lapses if the model underestimates them and vice versa. This approach significantly reduces the experience gap on lapses and consequently on life insurance reserves.

In order to propose a Machine Learning approach for the calibration of new total lapses laws, two products with the largest experience gaps are selected. On these products, the total lapses laws in number and amount are calibrated by distribution network using the characteristics of the policyholders and the contracts over the period from 2015 to 2021. Our penalized regression, SVM, KNN and random forest models reveal that policyholder seniority, policyholder outstandings and CSP are the most important characteristics in predicting lapses rates. For the year 2021, the laws predicted by our models are closer to reality than those of the internal model, which is confirmed by the reduction of the experience gaps on Lapses and therefore reserves. Moreover, the supervised learning approach works less well when the lapses laws are calibrated depending on the volume of outstanding contracts.

Although the results of this study are quite satisfactory, it does have some limitations related to the modeling assumptions. One perspective of this work would be to analyze the contribution of the profit-sharing policy to the experience gaps observed on the reserves.

Key words : Experience gaps, Machine Learning, Lapses

INTRODUCTION GÉNÉRALE

Un contrat d'épargne en assurance vie est un contrat par lequel l'assureur garantit à l'assuré ou au bénéficiaire désigné par l'assuré, le versement d'une prestation, d'un capital ou d'une rente si l'événement garanti par le contrat survient. La contrainte principale des compagnies d'assurance est donc de respecter leurs engagements vis-à-vis des assurés. A cet effet, l'introduction des nouveaux référentiels comptables (IFRS 17) et prudentiels (Solvabilité II) oblige les acteurs du secteur d'assurance à réfléchir à l'évaluation de leurs engagements. Ces différentes normes ont pour objectif commun la quantification d'un risque futur en lui attribuant une juste valeur aujourd'hui.

La valorisation des risques futurs nécessite l'utilisation de modèles mathématiques dits modèles actuariels. Au cœur de la prise de décision dans une compagnie d'assurance, ces modèles ont pour vocation de projeter deux états comptables : le bilan et le compte de résultat à partir d'une image des actifs, des passifs et des fonds propres à la date d'évaluation, en utilisant des hypothèses adéquates. Ils interviennent de ce fait pour répondre aux besoins de tarification des produits, de calcul réglementaire de la solvabilité ou encore de pilotage de l'activité.

Cependant, l'utilisation de tels modèles n'est pas sans risque. George Box rappelle que « tous les modèles sont faux mais que certains sont utiles ». Le risque posé par l'utilisation de modèles se manifeste par l'écart entre la réalité et le phénomène modélisé : on parle d'écarts actuariels d'expérience. Les sources de cet écart proviennent d'une incertitude d'estimation, d'une utilisation inappropriée ou d'une déficience des données ou d'une mauvaise calibration des hypothèses utilisées. Le risque de surprovisionnement ou de sous-provisionnement émanant de ces écarts présente un véritable enjeu économique. Une situation de surprovisionnement ou de surcapitalisation signifie en effet une opportunité manquée. Le capital potentiellement libéré pourrait être utilisé pour développer davantage l'activité. De même, une situation de sous-provisionnement ou de sous-capitalisation peut mettre en péril l'équilibre financier et ainsi conduire à

la faillite de l'assureur. Par ailleurs, des écarts d'expérience sur la valeur actuelle des flux de trésorerie pourraient sous-estimés ou surestimés considérablement le résultat de l'assureur.

Dans ce contexte, différentes questions se posent : Comment mesurer la qualité des écarts d'expérience dans un modèle complexe ? Les algorithmes de Machine Learning peuvent-ils permettre de réduire les écarts actuariels d'expérience sur les provisions constituées par l'assureur dans le but de respecter ses engagements ?

Dans ce mémoire, nous allons tenter de répondre à ces différentes problématiques. Pour ce faire, nous ferons appel aux algorithmes de Machine Learning. Des méthodes de clustering permettront de définir des groupes de produits homogènes en terme d'écarts actuariels et d'un point de vue métier et ainsi calibrer des taux de chocs d'hypothèses de projection des flux. Par ailleurs, une approche alternative aux modèles actuariels complexes est testée dans ce mémoire afin de tenter de réduire les différents écarts observés sur les provisions et ainsi approcher au mieux la réalité. Cette approche consiste en effet à calibrer des lois basées sur les algorithmes de Machine Learning auxquelles sont comparées les hypothèses de base (d'un point de vue déterministe) du modèle interne.

Première partie

Cadre de l'étude et présentation du portefeuille

1 CADRE CONCEPTUEL ET THÉORIQUE DE L'ÉTUDE

1.1 Généralités sur l'assurance vie

Dans cette section, nous présentons une vision globale du fonctionnement et des types de contrats d'assurance vie.

1.1.1 Quelques concepts clés

Le concept d'assurance repose sur plusieurs principes de base décrits comme suit ((Ducos 2020)) :

- **Cycle de production inversé** : le secteur de l'assurance présente un cycle de production inversé car le coût du produit d'assurance est connu après sa vente. L'assureur perçoit une prime de la part de l'assuré qui est utilisée plus tard lorsqu'un éventuel sinistre survient. En assurance vie, la prestation en cas de sinistres est fixée à l'avance ;
- **Présence d'aléa** : l'aléa découle de la réalisation ou non du risque sur le lequel porte le contrat d'assurance.
- **Mutualisation des risques** : ce principe est au cœur de l'activité d'assurance. l'assureur regroupe les risques homogènes de sorte qu'ils se compensent entre eux. Ainsi, il consiste à répartir le coût de réalisation d'un sinistre entre les assurés soumis potentiellement au même risque ;
- **Protection** : Le métier d'assureur consiste à accepter, contre rémunération, le transfert de risque de l'assuré vers l'assureur. Ce qui garantit à l'assuré une prestation en cas de sinistres. La quantification et la gestion de ce risque représentent l'enjeu principal des assureurs ;

En assurance vie, un contrat d'assurance est un contrat par lequel l'assureur garantit à l'assuré ou au bénéficiaire désigné par l'assuré, le versement d'une prestation, d'un capital ou d'une rente si l'événement

garantit par le contrat survient ; l'événement défini est lié à la durée de la vie humaine. L'objectif pour l'assureur étant de constituer un capital ou de transmettre une épargne valorisée aux assurés ou aux bénéficiaires. L'assureur assure ce service moyennant une prime. Le contrat d'épargne peut prendre fin pour des raisons diverses : la demande de rachat total de l'encours par l'assuré, la transformation en rente dans le cas du départ en retraite par exemple, le décès de l'assuré ou encore l'arrivée à terme du contrat. Ainsi, on distingue :

- Les contrats d'assurance en cas de décès : le capital ou la rente est versé(e) au bénéficiaire si le souscripteur décède avant le terme du contrat. Ce dernier constitue donc une épargne au profit d'une tierce personne ;
- Les contrats d'assurance en cas de vie : l'assuré reçoit un capital ou une rente s'il est toujours en vie à la fin du contrat. Ce type de contrat sert un objectif d'épargne optimisée en termes de fiscalité de long terme ;
- Les contrats mixtes : à l'échéance de ce contrat, le versement d'un capital ou d'une rente est garanti, soit au souscripteur, s'il est en vie, soit à un bénéficiaire, si le souscripteur est décédé ;

Ces contrats sont des placements financiers considérés comme des contrats d'assurance vie aléatoires au sens du code des assurances ; en effet, l'échéance est aléatoire et dépend de la durée de vie de l'assuré.

1.1.2 Types et fonctionnement des contrats en assurance vie

Les contrats d'assurance vie peuvent également être catégorisés suivant leur nature comme suit :

- les contrats d'épargne : ce sont des contrats d'assurance en cas de vie comportant des garanties en cas de décès de l'assuré, généralement utilisés pour financer des projets futurs ou optimiser la transmission d'un patrimoine ;
- les contrats de prévoyance : ces contrats permettent de couvrir l'assuré, à titre privé ou professionnel, contre les aléas de la vie liés à la personne. Cette assurance permet donc de faire face à d'éventuelles difficultés liées à une hospitalisation, un accident, un décès ou encore une perte de revenus ;
- les contrats de retraite : l'épargne retraite est une épargne constituée pour créer un complément de retraite supplémentaire. L'épargnant se constitue en effet un capital qui est alimenté par le versement périodique de sommes, sans conditions de montant, jusqu'au départ à la retraite. Une fois sa retraite prise, l'épargnant perçoit son épargne soit sous forme de rente viagère, soit sous forme de capital (ou les deux).

Le fonctionnement de ces contrats repose sur l'approche de capitalisation. Les primes versées par les assurés ne leur sont pas en effet reversées durant la vie du contrat ; celles-ci sont investis sous forme de placements financiers dont les revenus sont à chaque fois réinvestis et incorporés au montant épargné jusqu'à l'échéance du contrat. En assurance vie, on distingue les contrats d'épargne, les contrats de retraite pour lesquels le principe consiste schématiquement en la constitution d'un capital différé qui pourra être converti en rente et les contrats de prévoyance ayant pour objet d'une part la couverture des risques de décès selon que le décès résulte d'une maladie ou d'un accident et d'autre part des risques d'arrêt de travail et d'invalidité.

Pour ce qui est des contrats d'épargne, on distingue plusieurs supports d'investissement pour les contrats d'épargne :

- **Les contrats en euros**

Ce sont les produits d'épargne les « moins risqués » pour l'assuré car il ne présente pas de risque de perte en capital. En effet, la prime de l'assuré est majoritairement investie sur des titres de créance et d'obligations du secteur public ou privé et capitalisée chaque année à un taux de revalorisation qui dépend du taux minimum garanti (TMG) par l'assureur lors de la souscription et de sa production financière. Ces contrats sont donc peu volatiles mais présentent un rendement assez faible. Par ailleurs, le risque de perte est intégralement supporté par l'assureur. Sur ce support, une partie des résultats financiers et techniques de l'assureur sont versés aux assurés sous forme de participations aux bénéfices (PB). La PB représente pour les compagnies d'assurance, un levier de pilotage des risques.

Depuis plusieurs années, la diminution des taux de rendement des obligations et titres de créances ont entraîné une assez forte baisse des rendements des fonds euros. ainsi, dans le contexte actuel de taux bas voir négatif, Le TMG est un outil commercial dont le niveau devrait être fixé avec prudence par les assureurs.

- **Les contrats en Unités de Compte (UC)**

Pour ce type de contrat, la prime de l'assuré est investie sur des actifs financiers et valeurs mobilières de différentes natures, exprimés en nombre de parts de supports d'investissement et dont la valeur fluctue en fonction des mouvements observés sur les marchés financiers. Les parts d'UC sont des parts d'OPC : FCP, SICAV, parts de SCI ou de SCPI. Contrairement au fonds en euros, c'est l'assuré qui porte la totalité du risque lié aux investissements si aucune garantie n'est spécifiée. En effet, l'assureur garantit uniquement le nombre d'UC et non la valeur des parts.

Les UC ont l'avantage d'être divisible : ce qui signifie qu'un assuré qui souhaite investir une certaine

somme sur un fond UC donné va en acquérir la part correspondant à sa valeur liquidative et inversement lors d'une vente. Ainsi, Pour chaque support choisi, le nombre d'UC s'obtient en divisant la prime investie par la valeur liquidative du support au moment de l'investissement. Ce support présente également des rendements plus élevés que les fonds euros, mais plus volatiles. Les risques financiers sur ce support sont donc plus élevés.

- **Les contrats multi-supports**

Ce sont des contrats constitués de plusieurs fonds d'investissement. On distingue donc des contrats multi-supports UC et des contrats multi-supports UC/EURO ; la répartition de la prime sur les différents fonds suit un profil de gestion défini en fonction du niveau de risque de l'assuré. Plus la proportion d'UC est élevée, plus le risque et ainsi le rendement sont élevés pour le souscripteur. Ce type de contrat présente la particularité de permettre aux assurés d'effectuer des transferts de fonds d'un support à un autre : on parle d'acte d'arbitrage.

- **Les contrats dits croissance ou Eurocroissance**

Ces contrats sont à mi-chemin entre les contrats Euro et les contrats UC. Ils permettent une garantie du capital variant de 80% à 100% à une échéance choisie par l'adhérent et d'une durée minimum de 8 ans. Ces contrats ont principalement pour objectif d'encourager les investissements dans les PME notamment dans des sociétés contribuant au financement du logement social ou intermédiaire, les entreprises non-cotées.

1.1.3 Options et garanties sur les contrats multi-supports

Des mécanismes d'options et garanties financières permettent d'assurer aux épargnants un ensemble d'opérations sur tout ou une partie de leur investissement. Ces options et garanties entrent dans le calcul du Best Estimate et s'accompagnent d'un ensemble de risques supportés par l'assureur.

- **Garantie plancher**

Elle assure aux bénéficiaires du contrat un montant minimal de prestation appelé plancher. L'assureur supporte donc les risques de perte de capital moyennant des frais. Cette option ne concerne que les contrats d'assurance-vie en unité de compte et le montant n'est versé qu'en cas de décès de l'assuré. L'assureur est ainsi exposé au risque de mortalité s'il existe un écart entre l'évolution réelle de mortalité et la table de mortalité utilisée pour le calcul de la prime et des provisions. Avec le contexte actuel de taux durablement bas, cette option n'est plus offerte dans les contrats.

- **Taux minimum garanti (TMG)**

Cette garantie ne concerne que le support Euro. Le TMG se définit légalement comme le rendement minimum d'un contrat d'assurance vie. L'assureur garantit une revalorisation minimale de l'épargne du souscripteur (investie sur le fonds Euro) en fonction d'un taux fixé de manière contractuelle. Cette option garantit donc un rendement certain à l'assuré sur le capital investi. elle représente donc l'arme concurrentielle principale des assureurs.

Le calcul du TMG est réglementé par l'article 132 du code des assurances. Il répond aux exigences suivantes : Le taux proposé ne peut être supérieur à 85% de la performance des actifs de la société sur deux ans consécutifs, la promesse de revalorisation des contrats bénéficiant d'un TMG ne doit pas se faire au détriment de la rémunération d'autres contrats, le TMG est fixé sur une durée minimale de six mois et enfin le TMG ne peut dépasser un plafond dépendant de la stabilité des marchés obligataires. Il en découle donc un risque lié aux variations des taux de rendements sur le marché des obligations.

- **Sortie en rente/capital**

Cette option permet à l'assuré de récupérer son capital accumulé sous forme d'un versement unique à l'issu du contrat (sortie en capital) ou sous forme de versements périodiques d'une rente pendant une certaine durée (sortie en rente). Le montant de la rente (arrérage) dépend notamment de la valeur de rachat du contrat et de l'âge de l'assuré.

- **Participation aux bénéfices**

Le cycle de production propre à l'assurance étant inversé (prime versée avant la survenance d'un sinistre), la réglementation[1] exige aux assureurs de partager avec ses assurés ses bénéfices techniques et financiers probables sous forme de participation aux bénéfices (PB). On distingue :

— La PB réglementaire : son calcul se fait au niveau global de l'entreprise. Ainsi, la rémunération globale de la compagnie est au plus égal à 10% de son résultat technique et 15% du résultat financier.

- La PB contractuelle : le mécanisme de PB peut être défini par des clauses contractuelles au niveau du contrat ou d'un ensemble de contrats.
- La PB discrétionnaire : dans ce cas, l'assureur peut décider de servir en plus de la PB réglementaire ou contractuelle une PB discrétionnaire à certains assurés.

L'assureur dispose tout de même d'une certaine flexibilité vis-à-vis de la distribution de la participation aux bénéfices. L'attribution peut être directe à travers l'augmentation des provisions (épargne future) ou différée (sous une période maximale de 8 ans) par la mise en réserve au sein de la provision pour participation

1. Articles L. 331-3, A. 331-3 et A. 331-4 du Code des Assurances

aux excédents (PPE) afin de lisser dans le temps les rendements du contrat. Cette politique de distribution des profits et de gestion des provisions constitue la « crediting strategy ». Ce terme est crucial et stratégique dans la modélisation de la société d'assurance vie et a un impact très important sur les profits futurs de l'assureur. La PB ne s'applique qu'au support Euro.

- **Option d'arbitrage**

On ne parle d'arbitrage que sur des contrats multi-supports. Un acte d'arbitrage est l'opération qui consiste à réorienter tout ou une partie du capital constitué sur un ou plusieurs supports vers un ou plusieurs autres supports disponibles ; cette opération présente l'avantage de conserver l'antériorité fiscale. Les facteurs qui incitent les épargnants à arbitrer sont à la fois structurels (liés au contrat) et conjoncturels (liés aux conditions de marché ou aux mesures fiscales). Tout acte d'arbitrage donne lieu à une commission fixe ou variable en fonction du montant arbitré et du moment de l'arbitrage.

Cette option peut présenter un risque pour l'assuré non initié aux mécanismes liés à l'assurance vie ; le transfert de fonds sur un support en chute occasionnerait en effet des moins-values à la vente. Il en est de même pour un investissement sur un support en augmentation (à un prix élevé) sans garantie que sa nouvelle position ait la même dynamique dans le futur. Précisons en outre que la mauvaise calibration des lois d'arbitrage par les assurés est l'une des causes des écarts actuariels d'expérience observés sur les Best-Estimates et présentent un risque de moins-values.

- **Option de rachat**

Cette option donne la possibilité au souscripteur de récupérer une partie ou la totalité de son contrat avant le terme du contrat. En cas de rachat partiel, la partie restante continue à être investie. En revanche, le rachat total met fin au contrat. Elles donnent lieu à des pénalités généralement proportionnelles aux montants arbitrés. Cette option oblige l'assureur à constituer à chaque instant des provisions car l'assuré est susceptible d'effectuer un rachat à n'importe quel moment. Il s'agit donc d'un facteur de risque que l'assureur est amené à modéliser.

Les rachats sur un contrat peuvent être effectués pour des raisons structurels et conjoncturels dont les principales sont les suivantes.

- **Le cadre fiscal avantageux :** Ce cadre concerne essentiellement l'impôt sur le revenu ; les plus-values réalisées dans le cadre de l'épargne bénéficient d'un dispositif fiscal spécifique qui dépend de l'ancienneté du contrat comme le montre le tableau suivant :

  Précisons qu'après 8 ans d'ancienneté, un abattement annuel est effectué sur la part d'intérêts rachetée

Tableau 1.1 – Tableau simplifié du cadre fiscal sur le revenu

	Gains issus des versements avant le 27/09/2017	Gains issus des versements après le 27/09/2017
Entre 0 et 4 ans	52,20%	30%
Entre 4 et 8 ans	32,20%	30%
Après 8 ans	24,7%	24,7% pour les gains provenant des 150 000€ versés et 30% au-delà de 150 000€

de 4600€ pour une personne seule, 9200€ pour un couple.

— **La conjoncture économique et la concurrence :** Les rachats effectués pour cette raison sont des rachats dynamiques. Ils dépendent essentiellement du taux servi par l'assureur et le taux attendu par l'assuré (taux concurrentiel).

Les écarts actuariels d'expérience sur les options d'arbitrages et de rachats seront analysés dans le cadre de ce mémoire.

1.1.4 Mode de gestion des contrats multi-supports

Les modes de gestion au sein des contrats multi-supports dépendent principalement du profil de risque de l'assuré, des objectifs de rendement fixés par ce dernier et de l'horizon d'investissement. On distingue :

- **La gestion libre :** c'est le mode de gestion le plus souple et le plus risqué pour l'assuré. En effet, l'assuré gère son épargne et ses placements en totale autonomie, c'est-à-dire qu'il décide des supports sur lesquels son épargne est investie ainsi que des différents arbitrages ;
- **La gestion pilotée :** Contrairement à la gestion libre, les options d'arbitrage prédéfinies s'appliquent automatiquement notamment la sécurisation des plus-values, la répartition constante des fonds, la limitation des pertes ;
- **La gestion profilée :** Dans ce cas, l'assuré décide du profil de risque qu'il est prêt à prendre au moment de la souscription. Il peut s'agir d'un profil prudent, équilibré ou dynamique suivant le niveau d'aversion au risque de l'assuré. Ce sont les experts qui se chargent de la gestion et de l'arbitrage des investissements ;
- **La gestion sous mandat :** L'épargnant délègue la gestion de son épargne à l'assureur ou un

professionnel. Toutefois, les objectifs de rendements et le niveau de risque consenti par l'assuré sont définis au préalable avec le gestionnaire.

1.2 Réglementation en assurance vie : Solvabilité II et norme MCEV

Ce mémoire a pour but de proposer entre autres des algorithmes de Machine Learning permettant de réduire les écarts actuariels d'expérience observés sur les provisions Best Estimate calculés sous la directive Solvabilité II. Il est donc important de présenter les réglementations qui régissent le calcul de cet indicateur.

1.2.1 La directive Solvabilité II

La directive européenne Solvabilité II est en vigueur depuis le 1^{er} janvier 2016 et s'inscrit dans la lignée de la réforme Bâle 2. Elle a pour but d'inciter les organismes d'assurance à mieux connaître et gérer leurs risques et à mieux adapter leurs stratégies et leurs exigences réglementaires à leur profil de risque. Pour ce faire, de nouvelles méthodes d'évaluation du bilan économique des assureurs et réassureurs et de nouvelles exigences de capital de solvabilité sont mise en place sous cette directive. Elle comporte 3 piliers.

1.2.1.1 Pilier I : exigences quantitatives

L'objectif du pilier 1 de la réforme Solvabilité II consiste à calculer les exigences quantitatives de la compagnie d'assurance. Contrairement à Solvabilité I, le bilan de la compagnie n'est plus évalué sous une vision comptable mais plutôt économique. Ce qui permet ainsi de refléter la richesse réelle de l'entreprise. Les actifs sont en valeur de marché et les passifs sont évalués sous une vision « Best Estimate ».

Les provisions techniques de ce bilan correspondent à la somme de la marge pour risque et du Best Estimate (BE). La réglementation Solvabilité 2 définit les cash-flows Best Estimate comme « la moyenne pondérée par leur probabilité des flux de trésorerie futurs compte tenu de la valeur temporelle de l'argent estimée sur la base de la courbe des taux sans risque pertinente, soit la valeur actuelle attendue des flux de trésorerie futurs »[2]. Il s'agit donc de l'espérance des flux futurs de règlements actualisés au taux sans risque. Pour ce qui est de la marge pour risque (ou RM pour Risk Margin), elle représente le coût lié aux risques non couvrables entraînant un aléa sur le Best Estimate. Elle « est calculée de manière à garantir que la valeur des provisions techniques est équivalente au montant que les entreprises d'assurance et de

2. Article R351-2 du Code des Assurances, transposition en droit français de l'article 77 de la Directive Solvabilité

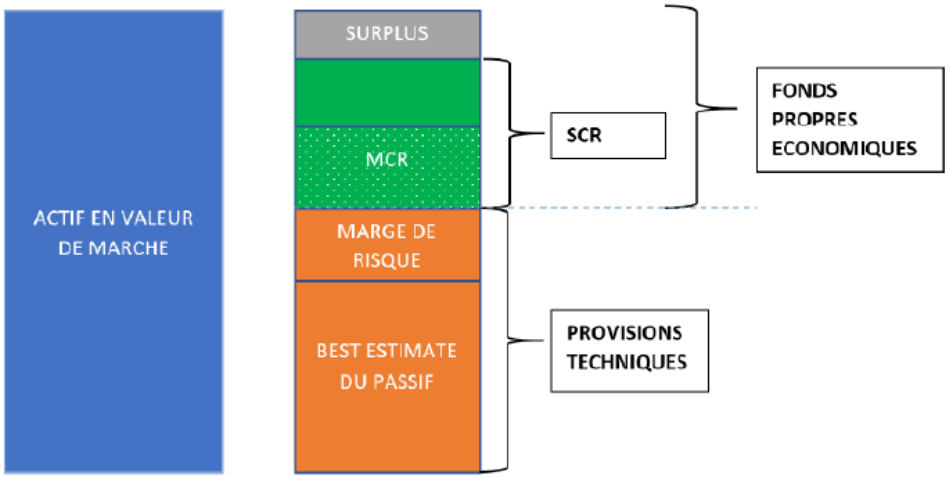


FIGURE 1.1 – Bilan économique sous Solvabilité II

réassurance demanderaient pour reprendre et honorer les engagements d'assurance et de réassurance ».

Mathématiquement, le BE se calcule comme suit :

$$BE = \sum_{t=1}^{\infty} E_{\mathbb{Q}} \left[\prod_{s \leqslant t} \frac{1}{(1+r_s)} (CF_t^{out} - CF_t^{in}) \right]$$

$\mathbb{Q}$ représente la probabilité risque-neutre, CF_t les cash-flows entrants et sortants à la date de projection t et r_s le taux sans risque forward à la date s.

Les fonds propres économiques représentent sous Solvabilité II la richesse réelle de la compagnie. Il s'agit de la différence entre l'actif en valeur de marché et les provisions techniques. Les exigences en capital se définissent par :

- Le MCR (Minimum Capital Requirement) : représente le niveau de fonds propres en dessous de laquelle l'intervention de l'autorité de contrôle est automatique. Le MCR se situe entre 25% et 45% du SCR.
- Le SCR (Solvency Capital Requirement) : capital dont doit disposer l'assureur pour faire face en ruine à horizon 1 an avec une probabilité de 99.5%.

1.2.1.2 Pilier II : les exigences qualitatives

Le pilier 2 définit les exigences en matière de supervision. Les objectifs sont de renforcer la gouvernance et la gestion du risque des assureurs. Solvabilité II introduit la démarche ERM (Enterprise Risk Management) qui définit la gestion du risque comme étant l'identification, l'évaluation et la priorisation des risques suivies par la coordination et l'application de ressources pour minimiser, piloter et contrôler la probabilité et/ou

l'impact d'évènements non voulus et maximiser la réalisation d'opportunités. Cela passe notamment par l'instauration de quatre fonctions clés définies dans les articles 44,46 à 48 de la directive : l'audit interne, la fonction de conformité, la fonction de gestion des risques et la fonction actuarielle.

1.2.1.3 Pilier III : les exigences de reporting

Ces exigences sont relatives au reporting à destination du public et du régulateur. Ce pilier a pour but d'améliorer, d'une part la transparence du secteur de l'assurance vis-à-vis du public en assurant un niveau d'informations suffisant aux assurés et aux acteurs financiers et d'autre part, de garantir un niveau de reporting suffisamment détaillé afin de permettre l'évaluation de la bonne gestion des risques de la part des compagnies d'assurance par le régulateur. La figure 4.30 en annexe A résume les trois piliers présentés.

1.2.2 Calcul des provisions "Best Estimate" sous solvabilité 2

Selon la directive Solvabilité 2, le "Best Estimate" ou meilleure estimation « correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque pertinents. Le calcul de la meilleure estimation est fondé sur des informations actualisées et crédibles et des hypothèses réalistes et il fait appel à des méthodes actuarielles et statistiques adéquates, applicables et pertinentes ». Les provisions "Best estimate" représentent la dette probable de l'assureur vis à vis de ses assurés. A la souscription, il existe par construction un équilibre entre les engagements respectifs de l'assureur et de l'assuré. Dès que l'assuré a payé la première prime, un déséquilibre s'instaure : l'engagement de l'assuré devient généralement inférieur à celui de l'assureur. Après la souscription, l'assureur a donc une dette probable vis-à-vis de l'assuré supérieure à sa créance de primes.

Les provisions sont inscrites donc au passif de son bilan comme une somme représentative de sa «dette nette».

Seule la méthode de calcul des provisions par la méthode prospective est reconnue par la réglementation (Article R. 343-3 du Code des Assurances). Selon l'article R343-3, les Provisions sont calculées comme la différence entre la valeur actuelle probable des engagements de l'assureur (paiement des prestations futures et frais associés) et la valeur actuelle probable des engagements de l'assuré (paiement des primes futures) : $Provisions_t = VAP(A,t) - VAP(a,t)$.

Sur la base des hypothèses retenues pour le provisionnement, les ressources prévisionnelles sont égales aux dépenses prévisionnelles à partir d'une formule de récurrence :

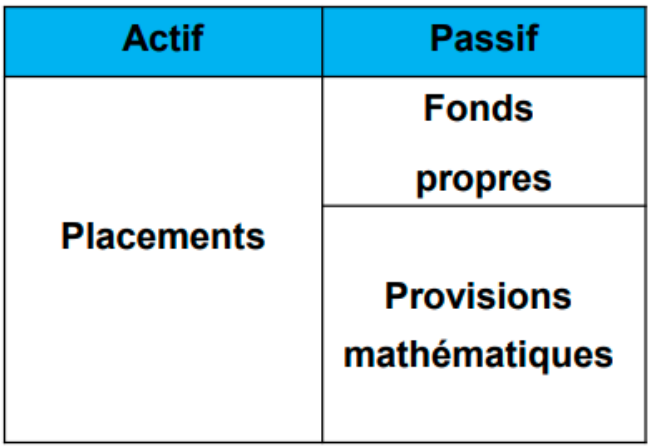


FIGURE 1.2 – Bilan simplifié d'un assureur

Provisions début de période + Primes versées au cours de la période + Produits financiers (sur la base du taux technique) = Prestations probables (sur la base de la table de mortalité ou de la loi de maintien utilisée dans le provisionnent) + Frais probables (pris égaux aux chargements escomptés dans le calcul des provisions) + provisions probables de fin de période (sur la base de la table de mortalité ou de la loi de maintien utilisée dans le provisionnement).

Les modèles de projection des prestations et frais probables des différents contrats d'assurance-vie nécessitent :

- Des données comptables : provisions, frais, etc.
- Des données de gestion : nombre de contrats, sexe, etc.
- Les dispositions contractuelles : taux de chargements, taux de commissions, taux garantis, revalorisation minimale/indexation ;
- Des sources extérieures (réglementaires ou non) : tables de mortalité (données de la population nationale, etc.), d'incapacité/invalidité ;
- Des modèles et statistiques annexes : lois de rachat, lois d'arbitrage, rendement des actifs (ESG), tables de mortalité (données internes) et autres lois biométriques.
- De règles contractuelles relatives à différentes provisions ou de règles de gestion : politique d'investissement et de participation aux bénéfices, revalorisation discrétionnaire dès lors qu'elle a été constatée plusieurs fois par le passé, dotation/reprise à la provision pour égalisation, traitement des fonds de revalorisation, etc.

1.3 Notion d'écarts actuariels d'expérience

En assurance vie, les engagements futurs ainsi que les indicateurs de rentabilité attendus par l'assureur sont calculés à partir des modèles de projection définis à cet effet. Ces modèles permettent à l'entreprise d'avoir une idée de l'évolution de son passif et de son actif au fil des années. Cependant ces projections sont calculées à partir des hypothèses définies l'année d'évaluation. Un an après l'évaluation, les engagements de l'entreprise sont recalculés et certaines des hypothèses remises à jour. Ainsi, pour un flux ou un indicateur donné, l'écart constaté entre la valeur réelle (calculée l'année N) et la valeur attendue (calculée l'année N-1) est appelé écart actuariel d'expérience. Par exemple, si l'on s'attend à avoir 500 000 € de sinistres à payer sur une année, et que réellement on ne réalise que 150 000 € de sinistres à payer, alors il existe un écart d'expérience.

Les écarts actuariels incluent[3] :

- **Les ajustements liés à l'expérience** : ce sont les effets des différences entre les hypothèses actuarielles antérieures et ce qui s'est effectivement produit (non réalisation en N d'hypothèses posées en N-1) ;
- **Les effets des changements d'hypothèses actuarielles** : il s'agit du changement d'hypothèses jugées obsolètes ;

En particulier dans le cas des provisions, il s'agit donc des montants résultant des écarts entre les anticipations (hypothèses) et la réalité économique. Dans la pratique, les sources d'écarts actuariels sont diverses ; elles découlent des écarts d'hypothèses suivantes définies sur le passif et sur l'actif :

— les hypothèses démographiques et biométriques : elles regroupent les tables de mortalité, les lois de rachats et d'arbitrages ;

- les hypothèses financières : elles regroupent les chroniques de rendements financiers, les taux d'actualisation et les taux d'inflation ;
- les hypothèses de revalorisation des contrats : il s'agit des taux minimum garanti, des taux de participation aux bénéfices (PB) et des taux de clauses de PB ;
- les hypothèses de coûts : elles se composent des coûts de gestion sur encours, des coûts d'acquisition et des coûts de structure ;

3. Paragraphe 7 de la norme IAS 19

- les hypothèses de chargements et de commissions : elles regroupent les commissions et les chargements d'acquisition et sur encours ;
- les hypothèses de taxes.

Des écarts actuariels significatifs impliquent deux situations : une situation de surprovisionnement ou de surcapitalisation qui signifie une opportunité manquée car le capital potentiellement libéré pourrait être utilisé pour développer davantage l'activité et une situation de sous-provisionnement et/ou de sous-capitalisation qui peut mettre en péril l'équilibre financier. La réduction des écarts actuariels présente donc à la fois des enjeux économiques et réglementaires.

L'ampleur et la récurrence des écarts d'expérience prouvent ainsi que les hypothèses établies ne sont plus représentatives de la réalité, et par conséquent la nécessité de les réajuster.

Dans le cadre de nos travaux, nous nous sommes intéressés aux écarts actuariels d'expérience sur la provisions de clôture. Les différents postes de calcul de ces provisions sont les suivants :

$$Provision\_cloture = Provision\_ouverture + primes + arbitrages\_nets - sinistres(dontlesrachats) - frais\_et\_PS + bénéfices + resultat\_technique.$$

Illustrons la notion d'écarts d'expérience cette fois sur les provisions. Soit un contrat d'assurance vie multi-supports pour lequel la provision d'ouverture sur le support Euro est de 100€. Le tableau 1.2 ci-contre présente les écarts d'expérience sur les postes susmentionnés :

TABLEAU 1.2 – Exemple d'écarts d'expérience sur les provisions de clôture (en €)

Poste	valeur réelle	valeur modélisée	ecart d'experience
PM d'ouverture	100	100	0
Primes	50	50	0
Arbitrages nets	16	-25	41
Sinistres	-40	-10	-30
Autres	11	5.5	5.5
PM de clôture	137	120.5	16.5

Les écarts d'expérience sur un poste correspond donc à la différence entre les données réelles et les données modélisées. Dans cet exemple, les sinistres sur le support Euro de ce contrat sont sous-estimés par le modèle car les sinistres modélisés sont 4 fois inférieurs aux sinistres réels. Il en est de même pour

les arbitrages entrants (arbitrages nets = arbitrages entrants - arbitrages sortants). Finalement, les écarts d'expérience sur la provisions de clôture sont de 16.5€. Elles sont donc sous-estimées par le modèle.

2 ANALYSE EXPLORATOIRE DU PORTEFEUILLE D'ÉTUDE

Dans cette partie, il est question d'analyser dans un premier temps les écarts d'expérience obtenus sur différents produits du périmètre Épargne-retraite. Par la suite, sont présentés les clusters de produits constitués dans le but de définir des chocs moyens à appliquer sur les hypothèses de rachats et d'arbitrages afin d'en réduire les écarts actuariels.

2.1 Présentation du portefeuille

Cette étude porte sur les contrats mono et multi-supports d'assurance vie individuelle et collective du périmètre épargne et retraite d'ALLIANZ France. Les données nécessaires au calcul des écarts d'expérience sur ces produits proviennent des bases suivantes fournies par le système Inventaire d'ALLIANZ :

- **La base VIPR** : elle regroupe l'ensemble des données liées aux flux de trésorerie réels observés sur l'ensemble des produits. On y retrouve entre autres les montants réels de rachats, d'arbitrages et de provisions par produit ;
- **La base R4** : elle contient l'ensemble des flux de trésorerie modélisés par groupe de produits. Il s'agit notamment des prestations, des provisions, coûts et commissions issus des modèles actuariels ;

2.1.1 Description des produits du portefeuille

Le portefeuille de cette étude est constitué de 474 produits. Ces produits sont en grande majorité de la compagnie Allianz VIE ("FR0002-AZVIE"), seulement 19% des produits appartiennent à la compagnie Allianz Retraite ("FR1764-AZRET"). Par ailleurs, respectivement 35% et 38% des produits appartiennent aux entités partenariat vie "PVSF" et d'épargne vie individuelle "VIEP". Les produits de l'entité "Retraite" sont minoritaires. Près de 4 produits sur 5 sont des produits d'épargne individuelle ; les produits de retraite

collective représentent 19% des produits étudiés dans cette étude. Une minorité (3%) de produits appartient aux produits de retraite individuelle. Près de 70% des produits de cette étude ont au moins un support d'investissement en Euro. Aussi, les produits sont présents sur 33 cantons ; respectivement 29% et 12% des produits étudiés appartiennent aux cantons d'investissement "21_PG_VIE" et "AT_PG_RET".

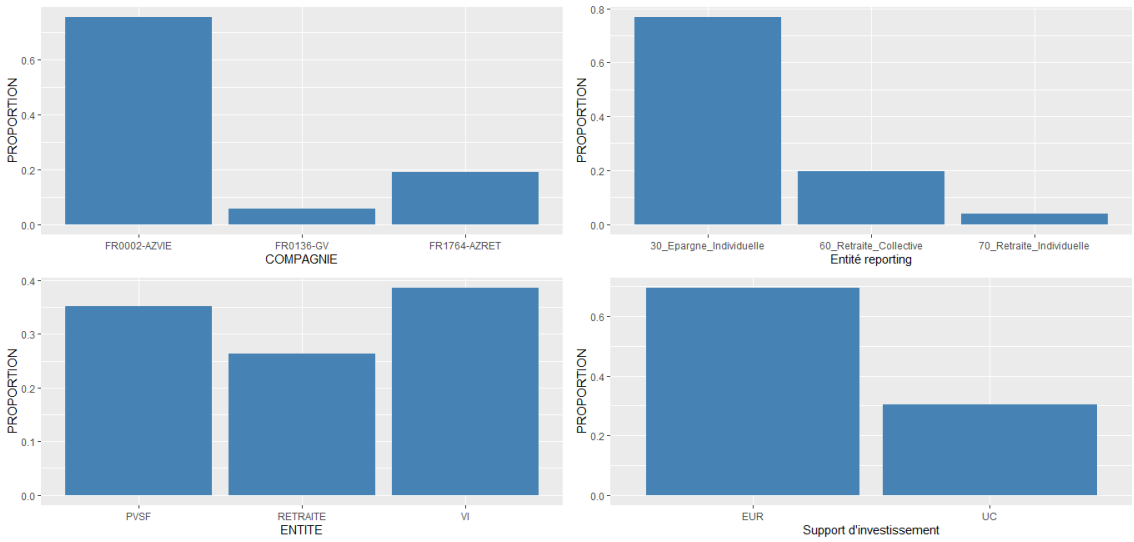


FIGURE 2.1 – Description des produits de l'étude

2.1.2 Calcul des écarts d'expérience sur le portefeuille

Pour un flux donné, l'écart actuariel d'expérience se calcule comme la différence entre le montant modélisé de la base R4 et le montant réel de la base VIPR. La principale difficulté dans le calcul de ces écarts réside dans l'établissement d'une correspondance fidèle entre la maille de produits utilisée dans la base R4 et celle de la base VIPR. Pour ce faire, une maquette a été développée sous Excel ; elle intègre toutes les corrections de produits et correspondances de mailles nécessaires au calcul des écarts d'expérience (Tableau 4.15 en annexe A). Finalement, la formule de calcul des écarts d'expérience sur un cash flow donné est la suivante : $Ecart\_cash\_flow = cash\_flow\_VIPR - cash\_flow\_R4$. Tous les chiffres fournis dans ce mémoire ne sont pas ceux du portefeuille réelle d'Allianz ; ils ont été modifiés(par application d'un "scaling factor") par souci de confidentialité.

2.2 Analyse des écarts d'expérience sur les produits

2.2.1 Écarts d'expérience sur les provisions de clôture

Cette analyse est faite uniquement sur les produits pour lesquels les provisions d'ouverture réelles et modélisées sont relativement proches. En effet, les provisions d'ouverture correspondent aux primes versées par l'assuré ; par conséquent, elles devraient être identiques dans les bases R4 et VIPR. Sur le graphique 2.2 ci-contre, on observe des écarts relatifs de provisions d'ouverture assez importants sur certains produits ; ceci est lié notamment à des ajustements dans le modèle actuariel stochastique. Seuls les produits pour lesquels les écarts relatifs sur provisions d'ouverture, inférieurs à 5% sont retenus dans cette étude ; ils représentent 58% du portefeuille soit 288 produits.

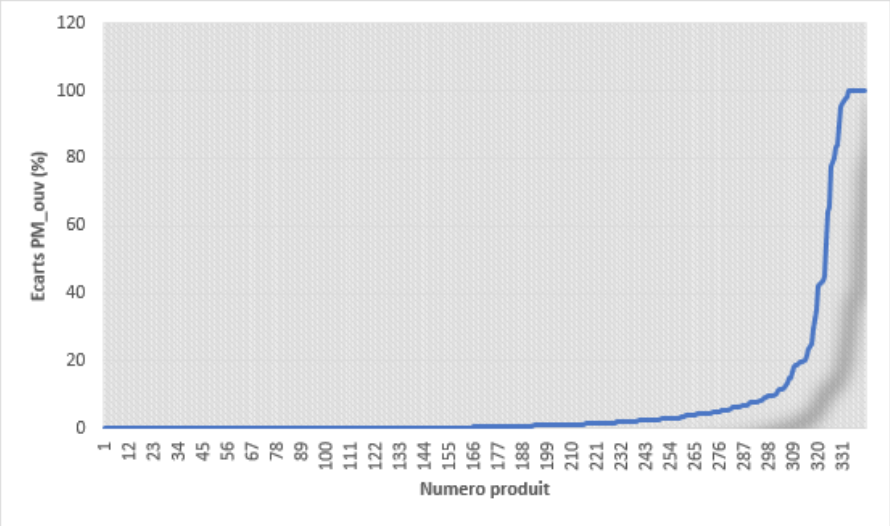


FIGURE 2.2 – Écarts d'expérience (en %) sur les provisions d'ouverture

2.2.2 Description des écarts sur les provisions de clôture

Globalement sur l'ensemble des produits étudiés, la somme des provisions de clôture réelles issues de la VIPR est de l'ordre de 993 M€ ; quant aux provisions modélisées issus du R4, elles sont de l'ordre de 972 M€. Il est découle donc des écarts d'expérience de l'ordre de 21 M€ ; ce qui signifie que le modèle interne de projection de cash-flows surestime les provisions de clôture sur les produits modélisés. On observe par ailleurs sur le graphique 4.31 en annexe A que de nombreux produits se démarquent avec des écarts assez importants notamment le produit "prod_eur" [1] (produit monosupport Euro) ; l'écart important observé sur

1. Il s'agit d'un nom fictif de produits par souci de confidentialité

les provisions de ce produit est du à l'annulation de la mise en place d'actions commerciales visant à réduire la duration du produit afin de compenser la baisse des taux.

Afin de comparer les différents écarts d'un produit à l'autre, des écarts relatifs sont calculés de la manière suivante : $écart\_relatif\_PM = \frac{Ecart\_PM\_clo}{PM\_clo\_R4}$. En moyenne, les écarts sur les produits du portefeuille de cette étude représentent 2.8% des provisions modélisées.

TABLEAU 2.1 – Distribution des écarts sur provisions de clôture

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Ecarts (en M €)	-0.09	-0.01	0.00	0.09	0.06	0.31
ecarts relatifs (%)	-50.763	-1.554	1.116	2.807	6.615	65.581

Les produits sur lesquels les écarts sur les provisions de clôture sont les plus importants sont majoritairement les produits de la compagnie Allianz-Vie "FR0002-AZVIE", de l'entité "VI" et du canton "21_PG_VIE".

- **Caractéristiques des produits suivant les écarts sur provisions de clôture**

La répartition de ces écarts suivant les caractéristiques des produits (graphique 2.3 ci-dessous) révèle que les produits de la compagnie Allianz GV ("FR0136-GV") et d'Allianz Vie ("FR0002-AZVIE") présentent en moyenne des écarts d'expérience plus importants contrairement aux produits d'allianz Retraite (FR1764-AZRET). Les écarts sur les produits d'Allianz Vie ("FR0002-AZVIE") sont quant-à-eux assez disparates. Par ailleurs, l'entité des produits d'épargne Vie Individuelle "VI" est celle sur laquelle les écarts sont les plus importants sur les provisions. Sur les entités partenariat vie et retraite ("PVSF" et "RETRAITE"), les écarts d'expérience sont relativement faibles. Comparés à la retraite individuelle, les produits de retraite collective présentent également des écarts assez importants. En particulier, un produit sur deux de la retraite individuelle présente des écarts deux fois plus importants. Aussi, d'un support à l'autre, les écarts d'expérience sont très disparates ; en effet, les écarts sont nettement plus importants sur le support UC. Il est donc plus difficile de calibrer les hypothèses de projection des provisions sur le support UC.

Ainsi, les hypothèses de projection des cash-flows les moins bien calibrées sont celles des produits appartiennent à la compagnie Allianz GV, l'entité "VI", l'entité reporting "retraite individuelle" ou le support "UC". La réduction des écarts d'expérience ainsi mis en évidence sur les provisions de clôture passe par une analyse de la contribution de chaque poste de calcul de cette provision aux écarts observés.

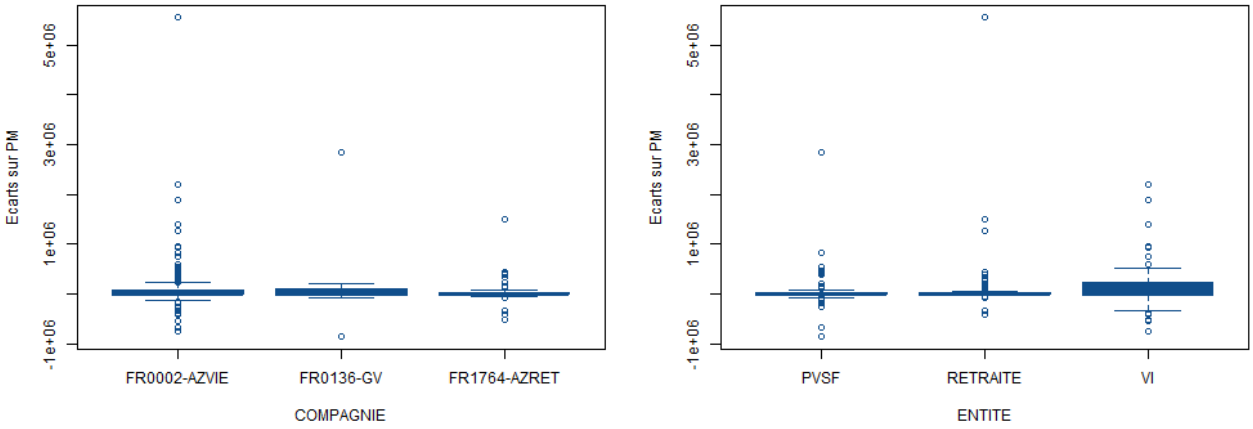


FIGURE 2.3 – Répartition des écarts relatifs sur les provisions de cloture en fonction des caractéristiques des contrats

2.2.3 Choix des postes d'analyse : enroulé de provisions

L'enroulé de provisions permet de mettre en évidence les postes sur lesquels les écarts d'expérience sont les plus importants. Ainsi, les différents postes de calcul de la provision de clôture sont les suivants :

PM_clôture=PM_ouverture + primes + arbitrages nets - sinistres (dont les rachats) - frais et prélèvements sociaux + bénéfices + résultat technique.

L'analyse des écarts d'expérience sur chaque poste (tableau 2.5 ci-contre) montre que les actes de rachats (total et partiel) contribuent à hauteur de 55% aux écarts observés sur les provisions de clôture. La contribution des arbitrages nets n'est que de 3.39%.

TABLEAU 2.2 – Enroulé/décomposition de la provision de clôture en millions d'euros

Postes d'analyse	VIPR	R4	Écarts d'expérience
PM ouverture	1 000	1000	0
Primes	47.27	47.27	0
Arbitrages nets	-0.71	0.03	-0.73
Rachats	-42.87	-54.29	11.42
Autres sinistres	-17.90	-24.41	6.51
Autres postes	7.28	3.85	3.43
PM clôture	993.11	972.45	20.65

Ainsi au regard de la forte contribution des rachats aux écarts d'expérience sur les provisions de clôture, on s'attellera dans la suite de cette étude à réduire les écarts d'expérience uniquement sur les actes de rachats. Par ailleurs, il existe dans la littérature une étude antérieure à celle-ci (mémoire (Miralles 2021)) qui traitent des arbitrages.

2.2.4 Analyse des écarts d'expérience sur les rachats

Ces rachats concernent à la fois les rachats partiels et totaux réels et modélisés sur les années 2020, 2019 et 2017. Ces 3 années sont considérées d'exploiter toute l'information disponible ; et ainsi obtenir un choc moyen le plus représentatif possible des écarts actuariels. L'année 2018 n'est pas utilisée dans cette étude car la base R4 des sorties de modèle sur cette année n'était pas disponible.

Globalement, les écarts sur les rachats sont les plus faibles en 2020 ; sur les années 2019 et 2017, les écarts sur l'ensemble des produits sont relativement assez proches. Une analyse produit par produit fait ressortir les produits sur lesquels les écarts sur les rachats sont les plus importants ; ils sont majoritairement les produits de la compagnie Allianz-Vie "FR0002-AZVIE", de l'entité assurance Vie Individuelle "VI". Par ailleurs, les rachats modélisés sont surestimés sur 80% de ces produits. En particulier, sur les produits phares d'épargne individuelle "Prod_E1" et "prod_E2"[2], les écarts sont causés par la non prise en compte des transferts liés au changement du type support des contrats (du type support "mono-support" vers "multi-supports") dans la calibration des lois de rachats. Ces transferts sont considérés en pratique comme des rachats totaux.

2.2.4.1 Analyse des écarts relatifs (taux de chocs) sur les rachats

Au cours des années étudiées, les écarts relatifs d'expérience sur les rachats sont assez disparates comme le montre la figure 2.4. Ils varient assez fortement d'une année à l'autre. L'année 2020 est celle qui présente les taux écarts les plus importants en moyenne. Pour la moitié des produits étudiés, les écarts relatifs sont supérieurs à -16.9% ; tandis que sur les autres années, les écarts médians sont nuls. Au regard de cette forte disparité, seuls les écarts observés en 2020 sont utilisés dans la suite de l'étude.

2.2.4.2 Analyse des écarts sur les produits

Les produits de la compagnie Allianz Retraite ("FR1764-AZRET") sont les produits sur lesquels les écarts sur les rachats sont les plus faibles en moyenne contrairement aux produits de la compagnie Allianz

2. Ce sont des noms fictifs de produits par souci de confidentialité

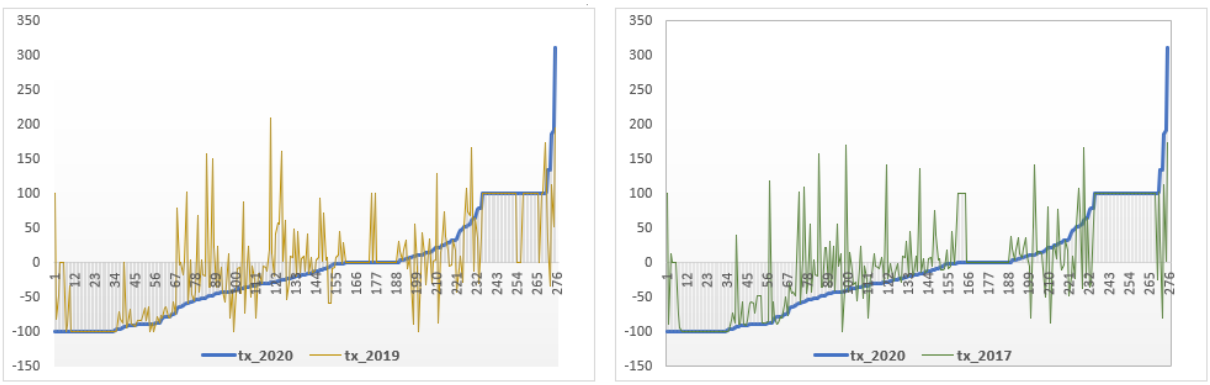


FIGURE 2.4 – Écart d'expérience (en %) sur les rachats en 2020, 2019 et 2017

VIE ("FR0002-AZVIE") et Génération Vie ("FR0136-GV") ; Par conséquent, les hypothèses de rachats sur les produits d'Allianz Retraite seraient mieux calibrées sur certains produits comparées aux autres entités. Sur les produits d'Allianz VIE et GV, un produit sur deux présente des écarts relatifs supérieurs respectivement à -34,65% et -24,09% ; les rachats issus de la modélisation de ces produits sont donc globalement surestimés. Afin de corriger ces écarts, les hypothèses sur les rachats seront modifiées à la baisse.

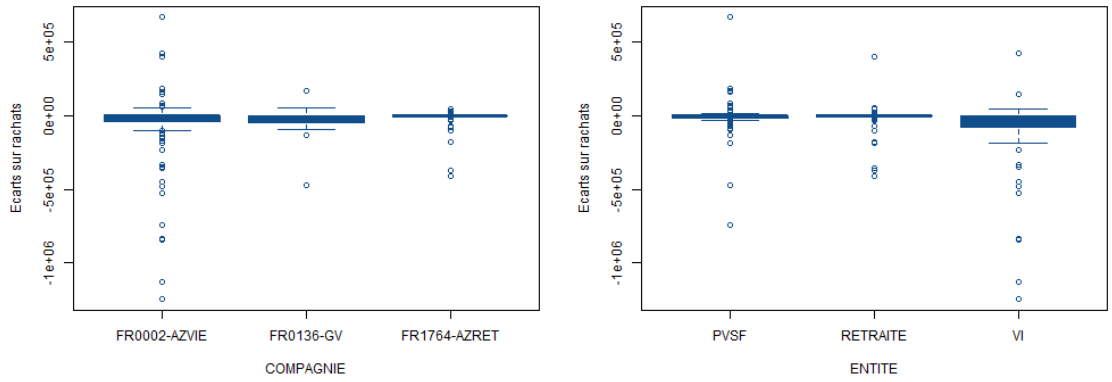


FIGURE 2.5 – Répartition des écarts d'expérience sur les rachats suivant la compagnie et l'entité

Contrairement aux produits des entités Partenariat Vie et Services Financiers[3] et retraite "PVSF" et "RETRAITE" sur lesquels les écarts d'expérience sur les rachats relativement faibles, ceux de l'entité Vie individuelle "VI" présentent les écarts assez importants. Par conséquent, les rachats modélisés sont surestimés sur ce périmètre ; En revanche, les lois de rachats sont donc mieux calibrées sur la retraite individuelle. Pour ce qui est des supports d'investissement, un produit sur deux du support UC présente des écarts 4 fois supérieurs au support Euro.

Ces importants écarts d'expérience ainsi mis en évidence sur les rachats sont dus aux erreurs d'esti-

3. L'entité "Partenariat vie regroupe les produits vie distribués à travers des partenaires haut de gamme d'Allianz

mation des lois de rachats, à la non modélisation des rachats sur certains produits ou encore aux biais de modélisation connus. En effet, sur les 288 produits de cette étude, 34 produits ont des rachats non modélisés ; il est donc impossible d'évaluer ni de corriger d'éventuels écarts d'expérience ; ces produits sont donc écartés de l'étude. Ils appartiennent tous à la compagnie Allianz Retraite ; en effet, ce sont des produits de rente sur lesquels il n'existe pas d'option de rachats ; par conséquent, aucune loi de rachats n'est calibrée sur ces produits. Toutefois, il existe des situations exceptionnelles (définies contractuellement) qui autorisent des rachats.

Afin de réduire ces écarts sur les rachats, des chocs sont appliqués aux hypothèses de rachats associées aux produits modélisés, de manière à augmenter les rachats du modèle interne si ils sont sous-estimés et vice-versa.

2.3 Détermination des chocs à appliquer aux lois de rachat

Les chocs à appliquer aux lois de rachats sont déterminés en adoptant la méthodologie suivante :

- **Calcul des taux de chocs sur les hypothèses produit par produit** : ce calcul permet d'obtenir les taux de chocs à appliquer aux hypothèses de rachats. Pour chaque produit considéré, le taux de choc est calculé comme suit : $taux\_choc = \frac{rachats\_VIPR - rachats\_R4}{rachats\_R4} = \frac{Ecart\_rachats}{rachats\_R4}$. Ainsi, des taux de -100% sont attribués aux produits pour lesquels les rachats réels sont nuls et ceux modélisés non nuls. Les produits non modélisés (produits pour lesquels les rachats du R4 sont nuls) sont analysés séparément car aucune hypothèse sur les rachats n'est faite dans le modèle ;
- **Construction des clusters de produits** :

  Appliquer un choc par produit rendrait la modélisation lourde. Les méthodes de clustering permettent de créer des classes de produits homogènes ou similaires d'un point de vue métier, sur lesquelles un choc moyen par classe est appliqué. Ces classes sont constituées en fonction de l'intensité des écarts d'expérience sur les produits et de leurs caractéristiques ;
- **Application des chocs moyens** : Pour tous les produits d'un cluster, un unique taux de chocs moyen est appliqué. En particulier, pour les produits "VIEP", les lois de rachats totaux sont directement modifiées. En revanche, pour les produits "PVSF" et "Retraite", les chocs sont appliqués aux coefficients de correction de rachats totaux car il n'existe pas de correspondance directe entre les codes produits associés aux chocs et ceux associés aux lois de rachats totaux. Les lois de rachats partiels ne sont pas modifiées pour la même raison. Finalement, pour les produits "PVSF" et "Retraite", les

chocs sont appliqués de la manière suivante :

$$\begin{cases} coef\_cor\_ractot\_final = coef\_cor\_ractot\_initial * (1 + taux\_choc\_moyen) \\ taux\_rachat\_total\_final = taux\_rachat\_total\_initial * coef\_cor\_ractot\_final \end{cases}$$

Un $coef\_cor\_ractot$ est associé à un groupe rachats qui regroupe un ensemble de produits sur lesquels les comportements clients en terme de rachats sont similaires. Précisons que ce coefficient prend la valeur 100 si aucune correction n'est appliquée aux taux de rachats.

Pour les produits "VIEP", le taux de rachats totaux après chocs est calculé comme suit :

$$taux\_rachat\_total\_final = taux\_rachat\_total\_initial * (1 + taux\_choc\_moyen)$$

2.3.1 Présentation des méthodes de clustering utilisées

Cette section s'attelle à présenter l'ensemble des méthodes d'apprentissage statistique non-supervisé utilisées dans ce mémoire pour le partitionnement des produits.

L'apprentissage non-supervisé consiste à extraire d'un jeu de données des groupes d'individus présentant des caractéristiques communes. Les algorithmes développés visent ainsi à découvrir et mettre en évidence des structures sous-jacentes aux données non étiquetées. Il permet, entre autres, le partitionnement des données et la réduction de dimensions. Dans ce mémoire, les techniques d'apprentissage non-supervisé K-médoïdes et DBSCAN sont utilisées afin de créer des groupes homogènes de produits suivant leurs caractéristiques et les écarts actuariels observés.

2.3.1.1 K-médoïdes : algorithme PAM

Il s'agit d'un algorithme de partitionnement directement dérivé des K-Means et idéal pour le partitionnement des variables mixtes (numériques et catégorielles). La méthode des K-médoïdes se base sur des médoïdes pour créer des partitions (clusters). Un médoïde est en effet l'élément d'un ensemble qui minimise la somme des distances entre lui et chacun des autres éléments de cet ensemble. Comme le K-Means, le K-Médoïdes va minimiser l'erreur quadratique moyenne qui est la distance entre les points du cluster et le médoïde. La différence avec le K-Means est que les centroïdes du K-Medoïdes sont choisis parmi les points du jeu de données.

Parmi les algorithmes de K-Medoïdes, l'algorithme PAM (Partionning Around Medoid) est celle utilisée dans ce mémoire. Elle se décompose en deux étapes :

- **Etape 1 (Initialisation)** : Le premier médoïde est le point du jeu de données minimisant la fonction objectif $C = \sum_i d(x_i, m_1)$. Le second médoïde est un point autre que m_1, minimisant la même fonction

objectif. On détermine ainsi les k premiers médoïdes. Chaque point est ensuite affecté au groupe dont le médoïde est le plus proche ;

- **Etape 2 (recherche des meilleurs médoïdes)** : Pour tout couple de points du jeu de données (i,j) tel que i soit un des k médoïdes et j un point autre qu'un médoïde, l'algorithme permute i et j si la fonction objective baisse. Cette étape est répétée jusqu'à stabilisation de la partition.

Pour juger de la qualité des clusters, l'algorithme définit une métrique nommée silhouette. En effet, pour chaque point, est calculé le coefficient $S_i = \frac{b_i - a_i}{max(a_i, b_i)}$ avec a_i la dissimilarité moyenne x_i au sein de son groupe et b_i la dissimilarité moyenne de x_i par rapport au groupe le plus voisin de x_i (dissimilarité moyenne la plus faible entre i et les objets de l'autre groupe). Ce qui permet de représenter dans chaque classe, la silhouette des points et ainsi visualiser leur bonne affectation. Plus la silhouette est grande, plus l'algorithme est performant.

La méthode des K-médoïdes présente le grand avantage d'avoir une grande robustesse vis à vis des données aberrantes. Elle est généralement la plus efficace car elle converge vers la classification la plus évidente contrairement au K-Means qui converge vers un minimum local.

2.3.1.2 DBSCAN : density-based spatial clustering of applications with noise

DBSCAN est un algorithme de partitionnement qui intègre la notion de clusters basée sur la densité, permet découvrir des clusters de forme arbitraire (Courjault-rade 2018). Son principe repose sur la notion de ϵ-voisinage d'un point défini comme l'ensemble des points appartenant à la boule de rayon ϵ centrée en ce point. En plus du rayon ϵ, le paramètre $MinPts$ est considéré : il s'agit du nombre minimum de points à prendre en compte dans la boule de rayon ϵ. Cet algorithme introduit entre autres les définitions suivantes (Courjault-rade 2018) :

- **Point coeur** : Un point x_i est un coeur si son ϵ-voisinage contient au moins $MinPts$ points en l'incluant. Un point du ϵ-voisinage de x_i est dit densément atteignable depuis x_i ;
- **Points de densité atteignable** :Deux points sont dits de densité atteignable s'il existe une chaine de points deux à deux directement densité-atteignables qui lie ces deux points ;
- **Points non atteignables** : Tous les points non atteignables de tout autre point, donc isolés, sont considérés comme atypiques (outliers).
- **Distance-Cœur** : la distance-coeur d'un point du jeu de données est la plus grande des distances à ses $MinPts$ plus proches voisins. Elle est notée $d_{core}(x_p)$,

- **Mutual Reachability Distance** : La "Mutual reachability distance" entre deux points x_p et x_q est définie par la formule $d_{mreach}(x_p, x_q) = max(d_{core}(x_p), d_{core}(x_q), d(x_p, x_q))$.

L'algorithme DBSCAN nécessite de fixer les valeurs des paramètres ϵ et $MinPts$. Pour trouver un cluster, DBSCAN commence par fixer un point arbitraire p et recherche tous les points de densités accessibles à partir de p (en calculant les distances "mutual reachability" pour toutes les paires de points formées par les éléments du jeu de données et le point fixé). Si le ϵ-voisinage de ce point contient suffisamment de points alors une classe se crée ; sinon, ce point est temporairement considéré comme atypique. Si un point est un coeur, tous les points de son ϵ-voisinage sont affectés à sa classe. Par ailleurs, si l'un de ces points est un coeur, son ϵ-voisinage vient compléter la classe précédente. Le processus continue jusqu'à ce que tous les points atteignables complètent la classe. Puis, l'algorithme fixe un autre point pas encore classé et renouvelle la procédure jusqu'à ce que tous les points du jeu de données soient étiquetés.

Grâce à l'utilisation des paramètres ϵ et $MinPts$, DBSCAN peut fusionner 2 clusters dans le cas où ils sont proches l'un de l'autre (avec des densités différentes). Cet algorithme est mis en défaut si les classes présentent des densités locales hétérogènes ou si les groupes de points ne sont pas suffisamment distincts les uns des autres.

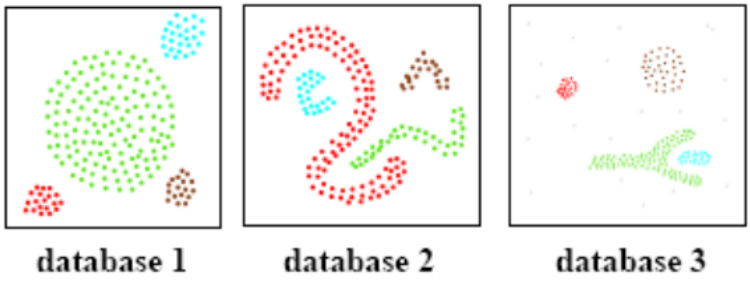


FIGURE 2.6 – Clusters trouvés par DBSCAN sur 3 jeux de données

2.3.1.3 Metrique de Gower : clustering sur variables mixtes

Une partie importante de la recherche de clusters sur un ensemble de données avec des variables de types mixtes est de trouver une métrique de distance capable de traiter différents types de variables (les variables catégorielles et numériques simultanément). La métrique de distance de Gower est capable de le faire en calculant les composantes de la distance entre deux instances X_i et X_j différemment pour chaque variable. pour deux instances X_i et X_j avec deux variables, désignées par X_{ik} et X_{jk} pour k $\in 1, 2$, supposons que la première variable soit catégorique et que la seconde soit numérique. Pour la première variable catégorielle,

la différence entre les valeurs de X_{ik} et X_{jk} est définie comme une fonction indicatrice (Gower 1971) :

$$S_{ijk} = \begin{cases} 0 & \text{si } X_{ik} = X_{jk} \\ 1 & \text{sinon.} \end{cases}$$

Pour une variable numérique, la mesure de similitude entre les valeurs de X_{ik} et X_{jk} est définie comme suit :

$$S_{ij} = \frac{\sum_{k=1}^{n} w_{ijk} S_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

w_{ijk} Représente le poids de la variable k entre les observations X_i et X_j. Dans ce mémoire, $w_{ijk} = w_k$ ce qui revient à attribuer un poids par variable.

2.3.2 Application à la construction des clusters

Les méthodes d'apprentissage non-supervisé suivantes sont comparées afin de choisir les meilleurs clusters : K-médoïdes et DBSCAN. La métrique de similarité/dissimilarité utilisée dans cette étude est la métrique de Gower car elle est adaptée pour les variables mixtes et permet d'affecter des poids différents en fonction de l'importance des variables.

Dans cette étude, quatre variables sont utilisées pour la construction des clusters ; il s'agit des taux de chocs sur rachats et des caractérisques des produits notamment la compagnie, l'entité et l'entité reporting. Un poids de 1 est alloué aux taux de chocs tandis qu'un unique poids de 0,4 est alloué aux autres caractérisques ; Il s'agit des poids qui optimisent le partitionnement. La variable "taux de chocs sur rachats" est donc la plus importante dans le clustering.

Les produits ayant des taux de chocs sur rachats de 100% et -100% sont regroupés en deux clusters. Le taux de chocs de -100% est considéré pour les produits modélisés et dont les rachats réels sont nuls afin d'annuler les rachats modélisés.

Les autres clusters sont construits sur 222 produits. Les produits les plus similaires suivant la distance de Gower sont les suivants :

Tableau 2.3 – Les produits les plus similaires

Produits	Entité	Entité reporting	Compagnie	taux_chocs
MODARIS_35_EUR	VI	Epargne Individuelle	FR0002-AZVIE	-77.3407
MODARIS_35_UC	VI	Epargne Individuelle	FR0002-AZVIE	-77.3464

Les produits les plus dissimilaires sont les suivants :

TABLEAU 2.4 – Les produits les plus dissimilaires

Produits	Entité	Entité reporting	Compagnie	taux_chocs
AV_MS_C_287_338_UC	PVSF	Epargne Individuelle	FR0002-AZVIE	347.09
BENEF_OPTI_325_35	VI	Retraite Individuelle	FR1764-AZRET	-88.77

La métrique de Gower est donc une bonne mesure de dissimilarité sur nos données. La matrice de dissimilitudes obtenue est utilisée comme "input" dans les algorithmes K-médoïdes et DBSCAN. Rappelons que l'indicateur "silhouette moyenne" (qui combine à la fois les variances intra-classes et inter-classes) permet d'évaluer chacun de ces modèles. Le nombre de clusters optimal est donc celui qui maximise la "silhouette moyenne".

- **Partitionnement par les K-médoïdes**

Sur cet algorithme, les nombres k de clusters testés vont de 2 à 30. Pour choisir le nombre k optimal, le graphique des silhouettes moyennes est tracé et k est choisi tel que la silhouette moyenne soit maximisée. Il est alors retenu 20 clusters pour lesquels la silhouette moyenne est de 0.66. Les taux de chocs par clusters semblent homogènes.

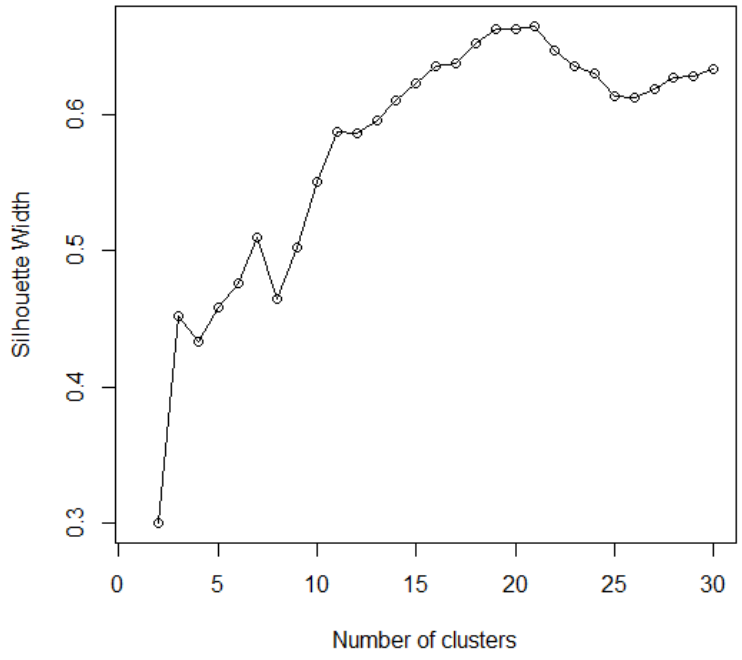


FIGURE 2.7 – Silhouette plot des K-médoïdes

- **Partitionnement par l'algorithme DBSCAN**

Pour cet algorithme, il est nécessaire de spécifier les valeurs "eps"(epsilon) et "$MinPts$" optimales ; ce

qui constitue une limitation du DBSCAN car il devient sensible au choix de ces paramètres, en particulier si les clusters ont des densités différentes. La méthode de détermination de la valeur "*eps*" optimale consiste à calculer la moyenne des distances de chaque point à ses k plus proches voisins. Ensuite, ces k-distances sont tracées dans un ordre croissant. Le paramètre "*eps*" optimal correspond donc à un seuil où un changement brusque se produit le long de la courbe des k-distances. Comme le montre le graphique 2.8, paramètre optimal "eps" est de 0.06. Finalement, 11 clusters sont construits par cet algorithme avec une silhouette

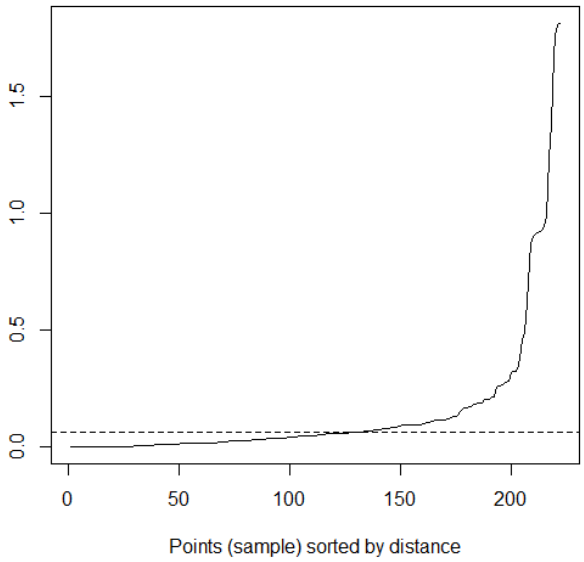


FIGURE 2.8 – Choix du paramètre "eps"

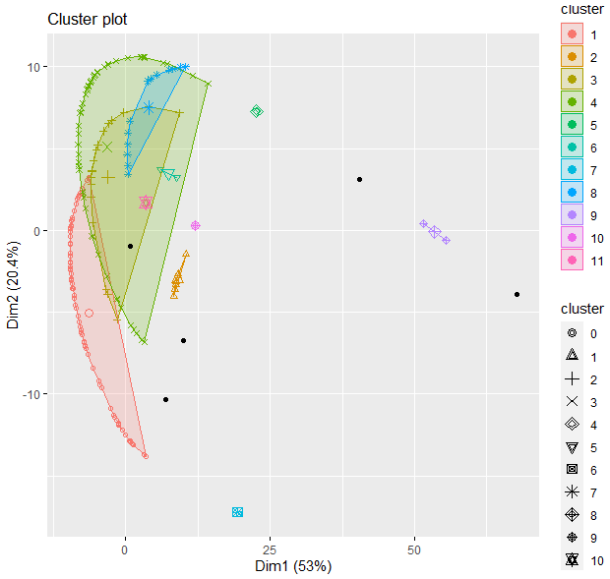


FIGURE 2.9 – Clusters obtenus par DBSCAN

moyenne de 0.51. Toutefois, les taux de chocs par clusters sont assez disparates.

Au regard des performances des différents modèles, le partitionnement par les K-médoïdes est donc le meilleur algorithme ; les caractéristiques des clusters construits sont les suivantes :

Près de 45% des clusters présentent des taux de chocs positifs ; les hypothèses de rachats sur les produits de ces clusters seront donc modifiées à la hausse comme détaillée dans la méthodologie. En revanche, pour les taux moyens négatifs, les hypothèses sont modifiées à la baisse. Le cluster 10 regroupe les produits sur lesquels les écarts d'expérience sur les rachats sont nuls ; toutefois, ces produits sont non modélisés. Les clusters 11 et 12 regroupent les produits les mieux modélisés en terme de rachats. En revanche, les clusters pour lesquels les taux de chocs sont les plus importants regroupent les produits d'épargne et de retraite individuelle.

Les différents clusters ainsi obtenus regroupent des produits appartenant à des groupes rachats différents (Tableau 4.17 en annexe A). Ce qui soulève des limites dans la méthodologie de construction de ces groupes

TABLEAU 2.5 – Taux de chocs moyens par cluster

N° clusters	nombre produits	taux_moyen(%)	N° clusters	nombre produits	taux_moyen(%)
1	4	286.26	11	14	-0.659
2	3	138.28	12	9	-1.75
3	3	87.44	13	6	-6.58
4	9	70.56	14	21	-18.26
5	7	41.92	15	15	-23.19
6	5	32.51	16	18	-41.43
7	24	18.52	17	4	-59.62
8	10	16.23	18	11	-79.29
9	7	5.70	19	21	-81.12
10	29	0.00	20	2	-90.31

rachats.

2.3.3 Application des chocs et analyse critique des lois de rachats actuels

Comme présentée dans la méthodologie, seules les hypothèses sur les rachats totaux sont modifiées, en particulier les hypothèses de correction des rachats totaux. Les chocs sont appliqués de la manière suivante :

$$coef\_cor\_ractot\_final = coef\_cor\_ractot\_initial * (1 + taux\_choc\_moyen)$$

Les résultats obtenus après les différents chocs sont les suivants :

TABLEAU 2.6 – Enroulé/décomposition de la provision de clôture en millions d'euros avant et après chocs

Postes d'analyse	VIPR	R4 initial	R4 après choc
PM ouverture	1000	1000	1000
Primes	47.27	47.27	47.27
Arbitrages nets	-0.67	0.03	-0.20
Rachats	-42.87	-54.29	-44.56
Autres sinistres	-17.90	-24.41	14.85
Autres postes	7.28	3.63	5.92
PM cloture	993.10	972.45	981.97

L'enroulé de provisions effectué (Tableau 2.6 ci-dessus) permet d'évaluer l'impact des chocs d'hypothèses de rachats sur différents postes d'analyse. Sur les rachats (partiels et totaux), les taux de chocs estimés permettent de réduire de 85% les écarts d'expérience, ce qui constitue un gain très significatif. Par ailleurs,

Tableau 2.7 – Comparaison des flux modélisés avant et après les chocs

Ecarts d'expérience au global (en millions d'€)			
flux	Ecart initial	Ecart après choc	Montants réduits
PM cloture	20.65	11.14	9.51
Rachats	11.42	1.69	9.73
Arbitrages nets	-0.73	-0.97	0.24

ces chocs induisent une réduction des écarts d'expérience sur les arbitrages nets de l'ordre de 32%. Il existe donc une assez forte corrélation entre les mouvements de rachats et d'arbitrages. Finalement, les écarts d'expérience sur les provisions de clôture sont réduits de près de 50% ; l'écart non réduit est causé par les interactions (effets croisés) qui existent entre les différents mouvements (rachats, arbitrages, participation aux bénéfices,...) d'une part et les écarts déjà existants sur les autres postes d'autre part.

Finalement, les écarts d'expérience assez importants constatés sur les rachats remettent en question la méthodologie de calibration des lois de rachats. En effet, les lois de rachats (totaux et partiels) des produits de la vie individuelle, du partenariat vie, de certains contrats d'assurance collective sont calibrées à des mailles "groupe rachats" différentes. Le groupe rachats pour les contrats d'épargne vie individuelle et la retraite collective est la combinaison du code du produit (famille homogène de produits), du réseau de distribution et du type de support. Pour les produits de type "Partenariats", la maille est la combinaison du réseau statistique et du type de support. Le tableau ci-dessous présente une synthèse de la méthodologie :

	Groupe rachats	méthode
Vie individuelle/retraite collective	CodeProd x réseau x type support	Kaplan-Meier/calcul direct
Partenariats	réseau stat. x type support	calcul direct

Par ailleurs, l'assez forte hétérogénéité des groupes rachats au sein des clusters construits remet également en question la calibration des lois de rachats à une maille agrégée. Une calibration à une maille plus fine (maille produit) est faite dans la suite de ce mémoire afin de réduire les écarts d'expérience observés. Cette approche est testée et évaluée sur les produits sur lesquels les écarts sont les importants sur les rachats.

2.4 Sélection des produits à écarts importants

Il s'agit dans cette partie de sélectionner les produits sur lesquels les écarts d'expérience sur les rachats sont les plus importants. Cette sélection se fait en définissant un seuil au delà duquel un écart sur les rachats est considéré comme important. Pour ce faire, deux méthodes sont utilisées :

- La calibration d'une loi statistique sur la distribution des écarts d'expérience ;
- La discrétisation des écarts d'expérience.

2.4.1 Méthodologie de sélection des produits

2.4.1.1 Estimation paramétrique d'une loi statistique

Une estimation est dite paramétrique lorsque l'on suppose que les données obéissent à un modèle (probabiliste ou non) défini par un nombre fini de paramètres à estimer. L'objectif est de déterminer les paramètres avec la meilleure précision possible. Soit $(\Omega, \mathcal{A}, \mathcal{P})$ un espace probabilisé et X une variable aléatoire $(\Omega, \mathcal{A})$ dans $(E, \mathcal{E})$. Soit un modèle statistique $\{P_\theta, \theta \in \Theta\}$, une famille de probabilités sur $(E, \mathcal{E})$. Si la loi de X est paramétrique, alors elle appartient à la famille $\{P_\theta, \theta \in \Theta\}$. Soit g une fonction définie de $\Theta \longrightarrow \mathcal{R}^k$, on appelle estimateur de $g(\theta)$ toute application T définie de $\Omega \longrightarrow \mathcal{R}^k$ de la forme $T = h(X)$ où h $(\mathcal{E} \longrightarrow \mathcal{R}^k)$ est mesurable. T est un estimateur sans biais de $g(\theta)$ si $\forall \theta \in \Theta$, $\mathbb{E}[T] = g(\theta)$. Le risque quadratique de T est défini par :

$$R(T, g(\theta)) = \mathbb{E}_\theta[(T - g(\theta))^2]$$

- **Estimation de θ par la méthode des moments**

Soit X le vecteur formé par un n échantillons $(X_1, ..., X_n)$. Les X_i sont à valeurs dans un ensemble $\mathcal{X}$. Soit $f = (f_1, ..., f_k)$ une application de $\mathcal{X}$ dans $\mathbb{R}^k$ telle que l'application suivante définie de $\Theta \to \mathbb{R}^k$ soit injective : $\Phi : \theta \mapsto \mathbb{E}_\theta[f(X_1)]$.

On définit l'estimateur $\hat{\theta}_n$ par la méthode des moments comme la solution dans Θ de l'équation :

$$\Phi(\theta) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Lorsque $\mathcal{X} \subset \mathbb{R}$, alors $f(X_i) = x^i$ et Φ correspond au $i^{\text{ème}}$ moment de la variable X_1 sous P_θ.

- **Estimation de θ par maximum de vraisemblance**

Soit $\{E, \mathcal{E}, \{P_\theta, \theta \in \Theta\}\}$ un modèle statistique, où $\Theta \subset \mathbb{R}^k$. On suppose qu'il existe une mesure σ-finie μ qui domine le modèle, c'est-à-dire que $\forall \theta \in \Theta$, P_θ admet une densité $p(\theta, .)$ par rapport à μ. On appelle vraisemblance de X, l'application :

$$\begin{cases} \Theta \to \mathbb{R}_+ \\ \theta \mapsto p(\theta, X) \end{cases}$$

Si, les X_i forment un n-échantillon de loi Q_{θ_0} ($\theta_0 \in \Theta \subset \mathbb{R}^k$) et Q_θ absolument continue par rapport à une mesure ν sur $\mathcal{X}$, en notant : $q(\theta, x) = \frac{dQ_\theta}{d\nu}(x)$, on a la vraisemblance suivante :

$$p(\theta, X) = \prod q(\theta, X_i)$$

L'estimateur du maximum de vraisemblance de θ est :

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} log[q(\theta, X_i)]$$

Cet estimateur n'existe pas toujours et n'est pas toujours unique.

- **Test d'adéquation : Test Kolmogorov-Smirnov**

Le test Kolmogorov-Smirnov est un test d'ajustement à une loi continue. Il s'agit d'un test d'hypothèse utilisé pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue ou encore si deux échantillons suivent la même loi. En pratique, ce test cherche à obtenir une estimation de la fonction de répartition à partir de l'échantillon observé afin de la comparer ensuite à la fonction de répartition de la loi théorique. Soit le modèle suivant : un échantillon $(X_1, ..., X_n)$ d'une loi inconnue P. L'hypothèse nulle est : $\mathcal{H}_0$: la loi P a pour fonction de répartition F_0, où F_0 est la fonction de répartition d'une loi continue donnée. Si l'hypothèse $\mathcal{H}_0$ est correcte, alors la fonction de répartition empirique $\hat{F}$ de l'échantillon doit être proche de F_0. La fonction de répartition empirique est la fonction de $\mathbb{R}$ dans [0,1], qui vaut :

$$\hat{F}(x) = \begin{cases} 0 & \text{pour x} < X_{(1)} \\ \frac{i}{n} & \text{pour} X_i < x < X_{(i+1)} \\ 1 & \text{pour x } \geq X_{(n)} \end{cases}$$

où les $X_{(i)}$ sont les statistiques d'ordre de l'échantillon (valeurs de l'échantillon rangées par ordre croissant). En d'autres termes, $\hat{F}(x)$ est la proportion d'éléments de l'échantillon qui sont inférieurs ou égaux à x. On mesure l'adéquation de la fonction de répartition empirique à la fonction F_0 par la distance de Kolmogorov-Smirnov, qui est la distance de la norme uniforme entre fonctions de répartition :

$$D_{KS}(F_0, \hat{F}) = max_{(i=1,...n)} \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Sous l'hypothèse $\mathcal{H}_0$, la loi de statistique $D_{KS}(F_0, \hat{F})$ ne dépend pas de F. On compare la valeur obtenue à une valeur critique $D_\alpha(n)$ fournie par les tables de Kolmogorov-Smirnov.

2.4.1.2 Méthodes de discrétisation

La discrétisation consiste à transformer une variable initialement numérique en une variable ordinale. Il s'agit en effet de trouver une subdivision optimale de l'intervalle de variation de la variable numérique initiale, et d'identifier cette variable avec un caractère dont les différentes modalités correspondraient aux sous-intervalles ainsi mis en évidence. Les méthodes de discrétisation à amplitudes et fréquences égales présentent l'inconvénient d'être sensible aux valeurs aberrantes. Des méthodes plus sophistiquées impliquent l'utilisation d'algorithmes non supervisés pour définir les catégories optimales. Dans ce mémoire, sont utilisées les méthodes de discrétisation par arbre de décision et k-means.

- **Discretisation par arbre de décision (CART)**

Il s'agit d'une méthode de discrétisation non supervisée qui n'utilise aucune information autre que la distribution des variables continues pour créer les groupes contigus dans lesquels les valeurs seront affectées. de façon spécifique, l'objectif est de découper la variable continue Y en classes $R_1, ..., R_J$ qui minimisent : $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_j)^2$ avec $\hat{y}_j = \frac{1}{n} \sum_{i \in R_j} y_i$, n_j étant le nombre d'observations dans la feuille R_j. Les seuils s de discrétisation sont ceux qui minimisent la variance suivante :

$variance_s = \sum_{i \in R_s^-} (y_i - \hat{y}_{R_s^-})^2 + \sum_{i \in R_s^+} (y_i - \hat{y}_{R_s^+})^2$ avec $R_s^- = \{Y < s\}$ et $R_s^+ = \{Y \geqslant s\}$

Cet algorithme fonctionne de façon récursive : une fois le premier seuil optimal s choisi, des découpages similaires sont effectués sur les régions R_s^- et R_s^+ jusqu'à ce que les critères d'arrêts soient respectés (nombre minimum d'observations, profondeur de l'arbre, ...).

- **Discretisation par K-means**

Le K-means est un algorithme de minimisation alternée qui, étant donné un entier K, va chercher à séparer un ensemble de points en K clusters. Cet algorithme minimise le critère d'erreur (distorsion) suivant par rapport aux centres des classes $(\mu_1^{(0)}, ..., \mu_K^{(0)})$ et les classes $(_1, ..., _n)$:

$$J(\mu_1, ..., \mu_K, z) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \left\| x_i - \mu_k \right\|^2$$

qui correspond à la distance euclidienne totale entre chaque données x_i et le centre μ_{z_i} dont elle est la plus proche. Dans l'expression du critère J , z_{ik} est une variable binaire qui vaut 1 si la classe de x_i est k et 0 sinon. L'algorithme K-means est composée des trois étapes suivantes (Chamroukhi 2016) :

— **Initialisation** : On initialise les centres des classes $(\mu_1^{(0)}, ..., \mu_K^{(0)})$ (au choix) pour donner le pas de départ de l'algorithme en choisissant par exemple aléatoirement des centres "virtuels", ou K données parmi les données à traiter ;

— **Étape d'affectation (classification)** : Chaque donnée est assignée à la classe du centre dont elle est la plus proche. $\forall i \in (1, ..., n)$,

$$z_{ik}^t = \begin{cases} 1 & \text{si } k = argmin_{z \in \{1,..,K\}} \|x_i - \mu_z\|^2 \\ 0 & \text{sinon.} \end{cases}$$

— **Étape de recalage des centres** : le centre μ de chaque classe k est recalculé comme étant la moyenne arithmétique de toutes les données appartenant à la classe obtenue à l'étape précédente.

$$\mu_k^{t+1} = \frac{\sum_{i=1}^n z_{ik}^{(t)} x_i}{\sum_{i=1}^n z_{ik}^{(t)}}$$

— **retour à l'étape 1 jusqu'à convergence.**

Cet algorithme converge en un nombre fini d'opérations. Cependant la convergence est locale, ce qui pose le problème de l'initialisation. Aussi, cet algorithme ne fonctionne que sur des variables numériques. Pour juger de la qualité des clusters, l'algorithme définit les inerties intra-classe et inter-classe comme suit :

$$I_{inter} = \sum_{k=1}^{K} p_k d^2(g_k, g), I_{intra} = \sum_{k=1}^{K} \sum_{i \in C_k} w_i d^2(x_i, g_k)$$

p_k et w_i représentent respectivement le poids de de la classe C_k et de l'individu x_i ; g_k représente le centre de gravité de la classe C_k. Une faible valeur de l'inertie intra-classe indique une homogénéité des classes tandis qu'une grande valeur d'inertie interclasse indique une bonne séparation des classes. Par conséquent, maximiser l'inertie interclasse c'est minimiser l'inertie intra-classe. Une stratégie simple permet d'identifier le nombre de classes : elle consiste à faire varier K et surveiller l'évolution de l'inertie intra-classes W. L'idée est de visualiser un « coude » comme le montre le figure suivante :

Les k-means sont également des algorithmes de discrétisation uni-variée non supervisée qui consiste à appliquer le clustering classique à une variable continue unidimensionnelle. Les seuils de discrétisation T_i sont définis comme : $T_i = \frac{K_i + K_{i+1}}{2}$.

2.4.2 Résultats de la sélection de produits

2.4.2.1 Calibration des paramètres de la loi des écarts d'expérience

La calibration est faite sur les écarts d'expérience (pris en valeur absolue) non nuls observés en 2020. La figure 2.11 montre une distribution des écarts très asymétriques, fortement étalées vers la droite. Un

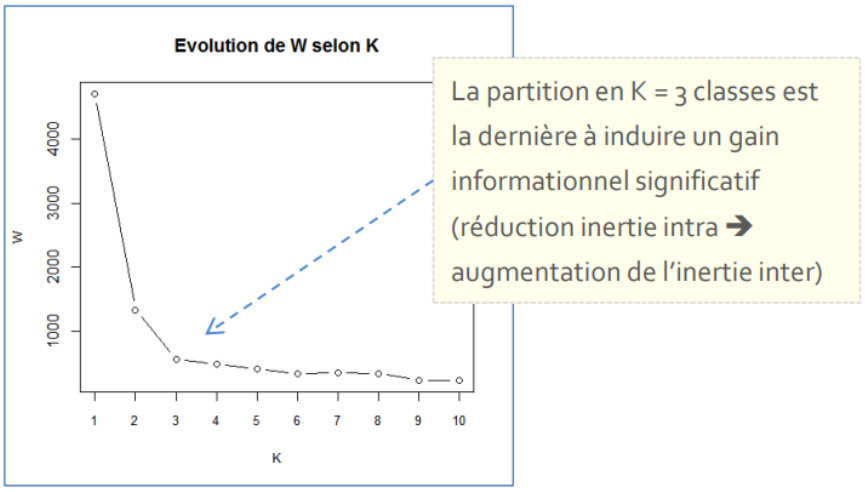


FIGURE 2.10 – Principe du choix du nombre de classes K

ensemble de lois statistiques définies sur R^+ est testé grâce au package "Fitter" de python ; il permet en effet de tester plus de 80 distributions et fournit celles qui s'ajustent le mieux à nos données ainsi que les paramètres associés. Comme le montre la figure 2.11, la loi de Galton-Gibrat plus connue sous le nom de loi log-normale, est celle qui s'ajuste le mieux à nos données.

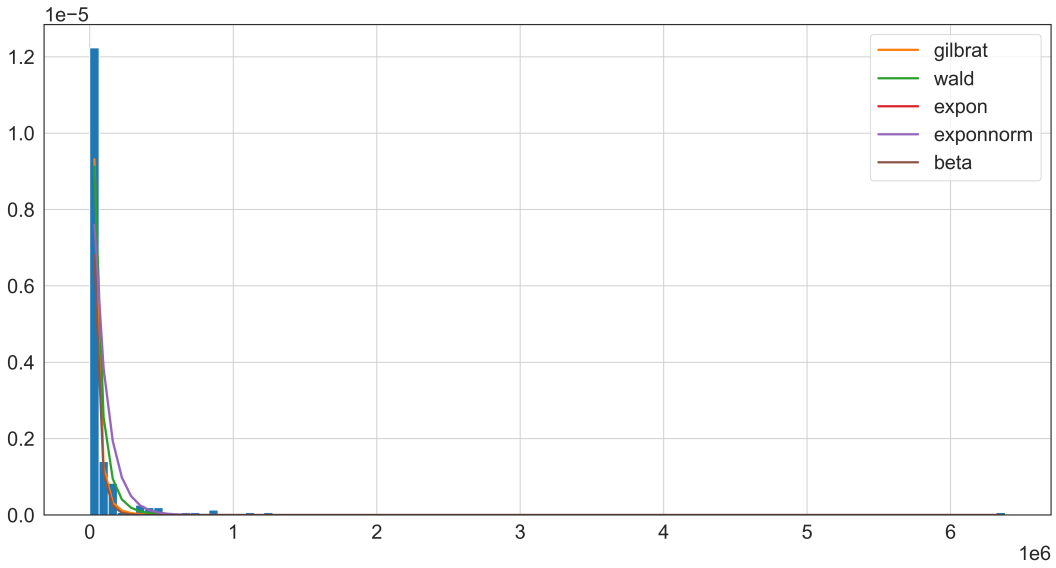


FIGURE 2.11 – Densité des lois qui s'ajustent le mieux aux écarts

Par ailleurs, c'est celle qui minimise l'erreur d'ajustement et les critères d'information ; le test d'ajustement à une loi de Kolmogorov-Smirnov confirme la qualité de l'ajustement (ks_pvalue = 0.57) comme le montre le tableau ci-après.

Tableau 2.8 – Qualité d'ajustement des distributions

Loi	sumsquare_error	AIC	BIC	ks_pvalue
gilbrat	8.8e-12	5387.70	7602.80	0.57
wald	1.1e-11	7690.77	7543.62	0.21
expon	2.9e-11	9116.76	7305.77	0.17
exponnorm	2.9e-11	9118.88	7300.00	0.09
beta	3.0e-11	16571.54	7292.18	0.03

Les paramètres de la loi de Gilbrat (log-normale) sont estimés par la méthode du maximum de vraisemblance sur 50 échantillons bootstrappés :

Tableau 2.9 – Paramètres estimés : loi log-normale

paramètres	*Shape*	*Scale*
Valeur Médiane	0.43	2.11e06

Finalement, le seuil retenu correspond à un centile supérieur de la distribution log-normale dont les paramètres figurent plus haut. Sur 50 échantillons bootstrappés, on obtient les centiles suivants :

Tableau 2.10 – Centiles de la loi de log-normale (en millions d'€)

	centile médian	95% IC	nb produits
95^e centile	0.44	[0.30, 0.57]	11
98^e centile	0.79	[0.58 , 1.05]	5

Avec un seuil 0.44 M€, 11 produits présentent des écarts d'expérience extrêmes sur les rachats. Toutefois, il est difficile de toutes les modéliser dans le cadre d'un mémoire à cause des problèmes opérationnels. Finalement, le 98^e centile est retenu correspondant au seuil de 0.79 M€ et pour lequel sont sélectionnés les produits suivants : "prod_A", "prod_B", "prod_eur", "Prod_E1" et "prod_E2"[4].

2.4.2.2 Discrétisation des écarts d'expérience

Cette deuxième approche est comparée à la précédente afin de sélectionner le seuil qui sépare le mieux les écarts importants des autres. Les méthodes de discrétisation par arbre de décision et K-means permettent d'obtenir les 3 classes suivantes (montants en millions d'€) :

4. Ce sont des noms fictifs de produits par souci de confidentialité

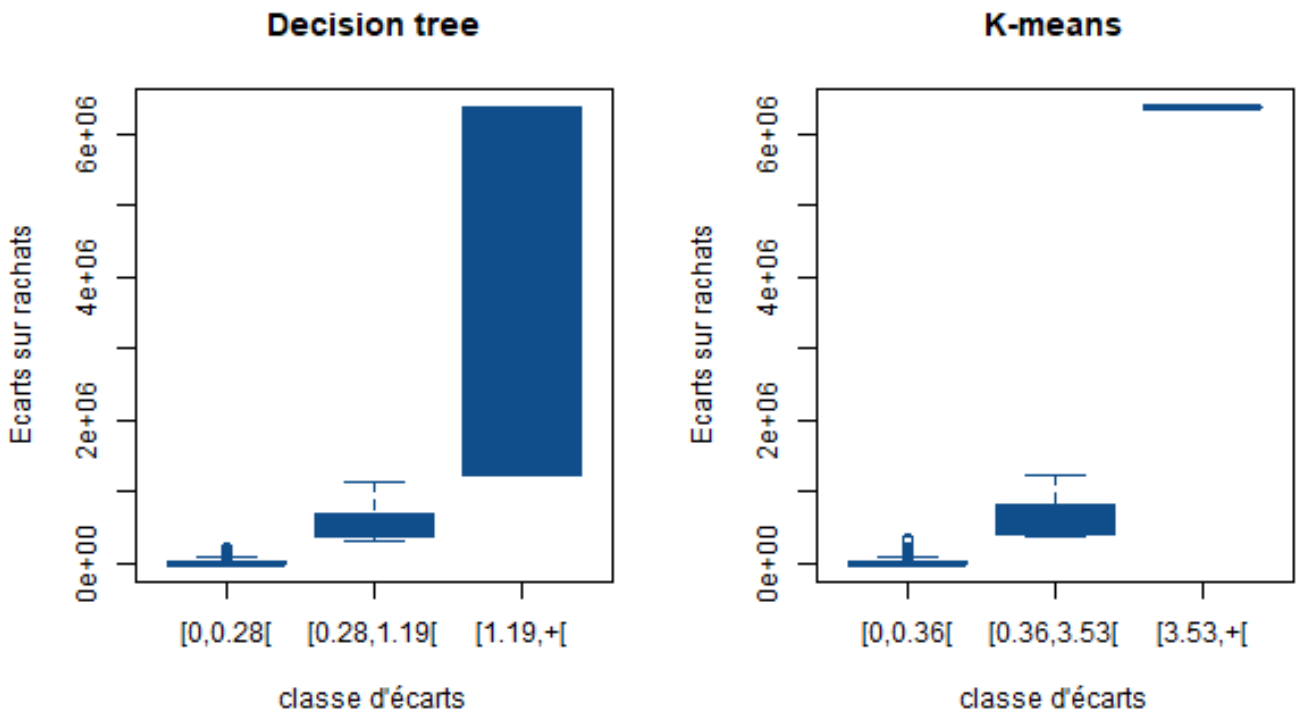


FIGURE 2.12 – Discrétisation des écarts sur les rachats totaux

Avec la méthode de discrétisation par les k-means, le seuil est de 3.53 M€ et permet de sélectionner un seul produit (produit "prod_eur"). La méthode par arbre de décision donne un seuil de 1.19 M€ qui permet de bien séparer les écarts importants des autres. Ce seuil permet de sélectionner deux produits : il s'agit des produits "prod_eur" et "prod_E1". Ces produits font partie de ceux retenus par la méthode précédente.

Afin de challenger les méthodes internes de calibration des lois de rachat sur ces produits, les algorithmes de Machine Learning sont implémentés dans la partie suivante. Toutefois, seuls les produits "prod_A" et "prod_B" sont retenus car les sources d'écarts sur les "prod_eur", "Prod_E1" et "prod_E2" mentionnées plus haut sont connues et acceptées.

2.4.2.3 Zoom sur les écarts d'expérience sur les produits "prod_A" et "prod_B"

Les produits "prod_A" et "prod_B" sont les produits d'épargne et de retraite individuelle phares d'Allianz. Sur ces produits, les taux de chocs appliqués permettent de réduire les écarts sur les rachats respectivement de l'ordre de 88% et 55% (Tableau 2.11 ci-dessus). Dans la suite de cette étude, des lois de rachats totaux sur ces produits sont calibrés à partir des algorithmes de Machine Learning. Les réductions d'écarts

d'expérience obtenues par ces lois seront comparées à celles obtenues par les chocs déterministes.

Tableau 2.11 – Décomposition de la provision de clôture "prod_A" et "prod_B" en millions d'euros

	"prod_A"			"prod_B"		
Postes	VIPR	R4 initial	R4 choc dét.	VIPR	R4 initial	R4 choc dét.
PM ouverture	80.39	80.39	80.39	45.05	45.05	45.05
premium	2.24	2.24	2.24	0.90	0.90	0.90
Arbitrages nets	0.09	0.00	0.00	-0.03	0.00	0.00
Rachats	-2.46	-3.63	-2.60	-1.17	-2.15	-1.61
autres sinistres	-5.26	-5.52	-6.00	-3.81	-3.34	-3.51
autres postes	3.68	3.23	3.62	2.47	2.52	2.34
PM cloture	78.68	77.37	78.31	43.41	43.35	43.53

Deuxième partie

Calibration des lois de rachats totaux via une approche machine learning

3

CONSTRUCTION DE LA BASE DE DONNÉES ET DESCRIPTION DU PORTEFEUILLE

3.1 Construction de la base de données

A Allianz, il n'existe pas de base complète directement exploitable pour la calibration des lois de rachats totaux sur les produits retenus. Un travail préliminaire consiste à trouver les sources de données, les extraire, et calculer les variables d'intérêt sur les produits étudiés. Des contrôles de cohérence sont par ailleurs effectués afin de s'assurer de la qualité des données.

3.1.1 Source de données

Les données de cette étude proviennent de l'infocentre GCP et du système Inventaire des produits d'épargne d'Allianz. Elles vont servir à récupérer les données à la maille contrat X mois relatives aux facteurs structurels influençant la décision pour l'assuré de racheter totalement son contrat ou non. Le logiciel SAS guide entreprise est utilisé pour se connecter aux bases de données. Essentiellement, quatre bases de données d'Allianz sont utilisées :

- **Les bases "MTT-MTO_SAP_CAP"** : Ces bases des mouvements enregistrent l'ensemble des mouvements mensuels effectués sur un support donné et à une date donnée, à la maille contrat. Ces mouvements correspondent aux arbitrages, aux rachats partiels et totaux, aux frais de gestion et charges sur prestations, participations aux bénéfices ... ;
- **Les bases "GARPRINC"** : elles contiennent les provisions de fin de mois par support (provisions de clôture) sur l'ensemble des contrats du périmètre épargne, ainsi que les caractéristiques des assurés et des contrats. Pour les supports euro, les provisions sont exprimées en euros ; tandis que sur les

supports en UC, les provisions sont exprimées en nombre de parts d'UC. Par ailleurs, pour un contrat et un support donné, la provision d'ouverture sur un mois considéré correspond à la provisions de clôture du mois précédent ;

- **La base "VL-Mensuelles"** : Dans cette base, sont stockées les valeurs liquidatives mensuelles de chaque support. La valeur liquidative d'un support correspond en effet à sa cotation quotidienne sur les marchés financiers. Cette base est utilisée pour convertir les provisions des supports UC des bases "GARPRINC" (exprimées en nombres de parts) en euros ;
- **Les bases "Info_Contrats"** : Ces bases permettent d'obtenir les informations micros des assurés :il s'agit notamment date d'effet du contrat associé à l'assuré, la périodicité de paiement des primes,le sexe, l'âge à la souscription et la classe d'âge à la souscription, la date de naissance, la catégorie socio-professionnelle ainsi que la date de naissance.

De nombreuses jointures sont effectuées entre ces bases à la maille contrat x produit x année d'inventaire. La base finale contient 3 185 202 lignes couvrant la période de 2015 à 2021 et 19 variables explicatives.

3.1.2 Traitement de la base de données

Les données extraites à la maille mensuelle sont agrégées à la maille mois afin d'estimer la probabilité pour un contrat d'être totalement racheté au cours de l'année. Précisons que seuls les rachats structurels sont modélisés. Cette base contient 4 997 471 lignes portant sur 530 082 contrats. Sur ces contrats, sont renseignées 10 variables quantitatives et 7 variables catégorielles. Il s'agit des caractéristiques qui pourraient avoir une influence sur la décision de racheter notamment le sexe de l'assuré, la catégorie socio-professionnelle (CSP), la périodicité de versement de la prime, l'option de gestion du contrat, le réseau de distribution ; il s'agit par ailleurs de l'encours du contrat, le type de contrat, la part d'UC, le nombre de support d'investissement, la fréquence et le montant de rachats partiels et d'arbitrages. Toutefois, il existe des valeurs manquantes sur certaines variables notamment la CSP et la périodicité qui présentent les taux les plus importants (figure 3.1 ci-dessous).

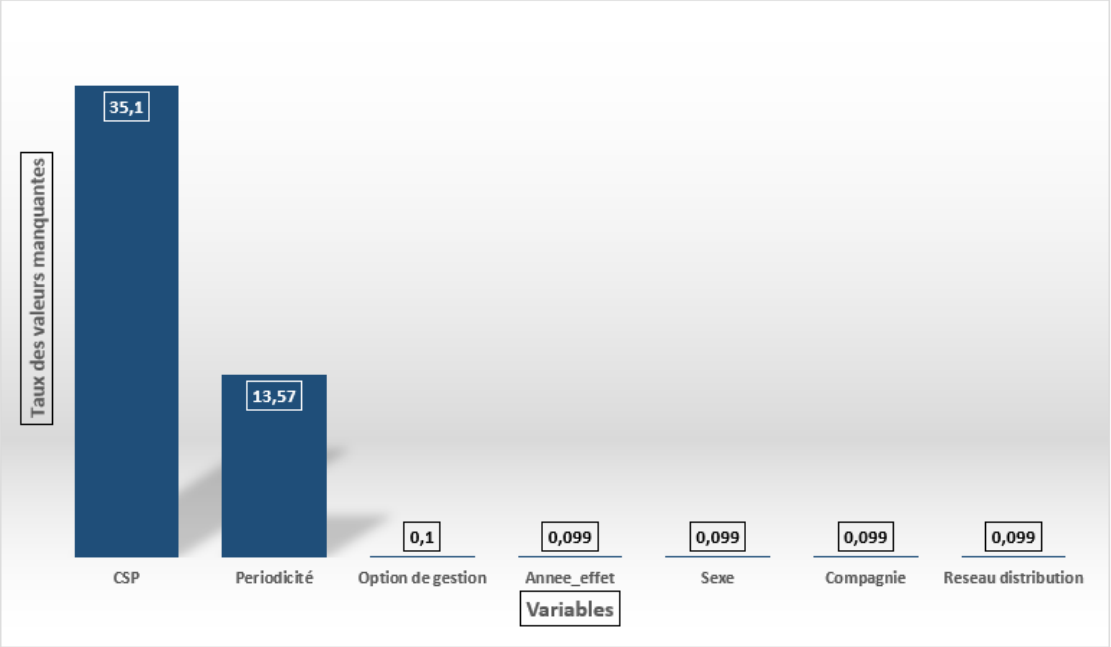


FIGURE 3.1 – Taux de valeurs manquantes par variable explicative

Ces valeurs manquantes sont imputées en utilisant les méthodes d'imputation avancées appliquées aux données mixtes suivantes : KNN, missforest et MICE.

3.1.2.1 Méthodes d'imputation des valeurs manquantes

- **Les forêts aléatoires (MissForest)**

Il s'agit d'une méthode de complétion basée sur les forêts aléatoires. Elle est capable de traiter des données de type mixte et en tant que méthode non paramétrique, elle permet de prendre en compte les interactions non linéaires entre les variables (Breiman 2019). Cette approche ne nécessite pas la présence d'une variable complète dans la base d'apprentissage. Les valeurs manquantes sont prédictes en utilisant des forêts aléatoires entraînées sur les parties observées des données.

Soit $X = (X_1, ..., X_p)$ une matrice de données à $n \times p$ dimensions. Pour une variable arbitraire X_s comprenant des valeurs manquantes, l'ensemble de données est séparée en quatre parties : les valeurs observées de la variable X_s (noté $Y_{obs}^{(s)}$), les valeurs manquantes $i_{mis}^{(s)}$ de la variable X_s (noté $Y_{mis}^{(s)}$), les valeurs correspondant aux observations $i_{obs}^{(s)} = \{1, .., n\} \setminus i_{mis}^{(s)}$ des autres variables (noté $X_{obs}^{(s)}$) et les valeurs correspondant aux observations $i_{mis}^{(s)}$ des autres variables (noté X_{mis}^{s}). L'algorithme Missforest se décrit comme suit.

Le critère d'arrêt α fonctionne en itérations, s'arrêtant lorsque la différence de performances entre l'itération i et $i+1$ des données imputées commence à augmenter pour les variables catégorielles et numériques,

Algorithme MissForest

1. Complétation naïve des valeurs manquantes

2. soit α un critère d'arrêt (à initialiser) et K vecteurs des indices de colonnes de X triées par quantité croissante d'observations manquantes.

3. Répéter :

A- X_{old}^{imp} : matrice précédemment imputée

B- Pour s dans K faire :

- Ajuster une forêt aléatoire : $Y_{obs}^{(s)} \sim X_{obs}^{(s)}$

- Prédire $Y_{mis}^{(s)}$ en utilisant les régresseurs $X_{mis}^{(s)}$;

-X_{new}^{imp} : matrice complétée à partir de $Y_{mis}^{(s)}$

C- Mettre à jour le critère α

4. Jusqu'à critère α atteint.

ou le nombre d'itérations défini par le paramètre "maxiter" est atteint.

- **La méthode des plus proches voisins (KNN)**

Cette méthode consiste, pour un individu donné, à sélectionner un échantillon d'individu qui lui sont les plus proches, au sens d'une certaine similarité (distance euclidienne pour les variables quantitatives ou la métrique de Gower pour les variables catégorielles par exemple). Ainsi, pour un individu ayant des valeurs manquantes, il devient possible de prédire la valeur de son observation à partir des observations de ses plus proches voisins. Le nombre de proches voisins considérés a donc une influence significative sur la performance de l'imputation. Il est donc fixé par validation croisée. Cet algorithme repose sur le même principe que celui des K-means. L'algorithme KNN par validation croisée se décrit comme suit :

Algorithme KNN

1. Choix du nombre d'ensemble de validation L et de l'ensemble K des proches voisins possibles

2.X^{CV}=imputation naïve des valeurs manquantes

3. Pour t dans L faire :

- $X_{mis,t}^{CV}$: introduction artificielle des valeurs manquantes dans X^{CV}

Pour k dans K faire :

- $X_{KNN,t}^{CV}$: Imputation de $X_{mis,t}^{CV}$ en utilisant les KNN ;

- $X_{k,t}$: erreur de l'imputation KNN pour k et t ;

Fin pour ;

Fin pour ;

4. k_{best} = Argmin $\frac{1}{l}\sum_{t=1}^{l} \varepsilon_{k,t}$;

5. X^{imp} : Imputation de X en utilisant les k_{best} plus proches voisins.

- **La méthode MICE**

En anglais "Multivariate Imputation by Chained Equations", la méthode MICE est basée sur la mise en place des équations chainées qui permet de spécifier le modèle d'imputation pour chaque variable incomplète fonction du type de variable (quantitative, binaire, catégorielle, ...). Soient des données incomplètes Y partiellement observées de distribution multivariée $P(Y|\Theta)$. Supposons que la distribution de Y est complètement spécifiée par Θ, un vecteur de paramètres inconnus. Il est question de trouver la distribution multivariée de Θ en manière explicite ou implicite. L'algorithme MICE obtient la distribution postérieure Θ en échantillonnant itérativement à partir des distributions conditionnelles de la forme : $P(Y_1|Y_{-1}, \Theta_1)...P(Y_p|Y_{-p}, \Theta_p)$

Les paramètres $\Theta_1, ..., \Theta_p$ sont spécifiques aux densités conditionnelles respectives. En partant d'un simple tirage des distributions marginales observées, la t^{ieme} itération des équations chaînées est la suivante :

$$\Theta_1^{*(t)} \sim P(\Theta_1|Y_1^{obs}, Y_2^{t-1}, ..., Y_p^{t-1})$$
$$Y_1^{*(t)} \sim P(Y_1|Y_1^{obs}, Y_2^{t-1}, ..., Y_p^{t-1}, \Theta_1^{*(t)})$$
$$\cdots$$
$$\Theta_p^{*(t)} \sim P(\Theta_p|Y_p^{obs}, Y_1^{t}, ..., Y_{p-1}^{t})$$
$$Y_p^{*(t)} \sim P(Y_p|Y_p^{obs}, Y_1^{t}, ..., Y_p^{t}, \Theta_p^{*(t)})$$

$Y_j^{(t)}$=$(Y_j^{obs}, Y_j^{*(t)})$ représente la $j^{ième}$ variable imputée à l'itération t. L' imputation $Y_j^{*(t-1)}$ n'est utilisée pour l'imputation $Y_j^{*(t)}$ qu'à travers sa relation avec les autres variables. En pratique, le nombre d'itérations optimal est assez faible variant de 10 et 20 itérations.

3.1.2.2 Imputation des valeurs manquantes

Afin d'évaluer la précision des différents algorithmes, deux critères sont utilisés en fonction de la nature de la variable concernée. Pour les variables quantitatives, le meilleur algorithme d'imputation est celui qui minimise la moyenne des RMSE (Root Mean Squared Error), définit comme suit :

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

Pour ce qui est des variables qualitatives, l'algorithme qui minimise la proportion de mauvais classement suivant est retenu : $PMC = \sum_{i=1}^{n} \frac{\#(y_i^{test} \neq \hat{y}_i)}{\#(y_i^{test})}$.

Les résultats des différents algorithmes d'imputation utilisés sont présentés dans le tableau ci-dessous :

TABLEAU 3.1 – Performance des différents modèles d'imputation

Indicateur : proportion de mauvais classement			
Variables qualitatives	KNN	MissForest	MICE
SEXE	0.46	0.21	0.50
CSP	0.91	0.86	0.92
PERIODICITE	0.34	0.24	0.42
MODE GESTION	0.29	0.20	0.24
RESEAU	0.38	0.32	0.47
Indicateur : RMSE			
Variables quantitatives	KNN	MissForest	MICE
Année_EFFET	18	20.08	19.75

L'algorithme MissForest est celui qui minimise l'erreur d'imputation aussi bien sur les variables quantitatives que qualitatives. Il est donc utilisé pour le traitement de données manquantes de la base de données de cette étude.

3.1.3 Calcul des taux de rachats totaux à Allianz

Dans cette section, la méthodologie détaillée de calibration des lois de rachats (structurels) utilisées dans le modèle interne est présentée ; Les lois associées aux produits "prod_A" et "prod_B" en font partie. Ainsi, la méthodologie suit les étapes décrites ci-dessous :

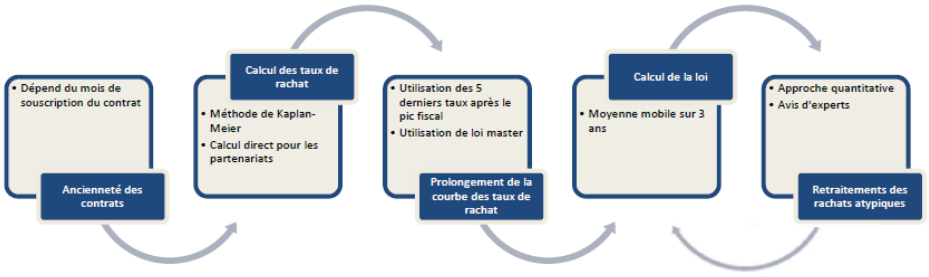


FIGURE 3.2 – Méthodologie de calcul des taux de rachats

La méthode de calcul varie en fonction de la taille des groupes de produits. Si ils sont assez grands, les taux de rachats sont calculés par année d'ancienneté suivant l'approche Kaplan-Meier. Pour les plus

« petits » groupes de produits, qui peuvent être déstabilisés par un flux, l'estimateur flat est utilisée ; il ne tient pas compte de l'ancienneté. Lorsque les résultats ne sont pas satisfaisants ; les taux sont calculés directement par ancienneté. Le tableau ci-après récapitule les formules de calcul suivant chaque approche :

TABLEAU 3.2 – Approche de calcul des taux

Approche	taille maille	formule
Kaplan-Meier	grand	$t_\kappa = 1 - \prod_{i=1}^{12}(1 - \frac{Rachats_{i,\kappa}}{PM_{i,\kappa}^{ouverture}})$
Estimateur flat	petit	$t = \frac{Rachats}{PM}$
Calcul direct	indifférent	$t_\kappa = \frac{Rachats_\kappa}{PM_\kappa^{ouverture}}$

La valeur d'ancienneté considérée correspond à la $i^{ème}$ année du contrat : l'année de souscription correspond à l'ancienneté 1. Les provisions sont ventilées proportionnellement sur les années d'ancienneté du contrat. Dans cette étude, la méthode de calcul direct est utilisée pour calculer les taux de rachats totaux par ancienneté dans la base de données.

Dans la suite de cette étude, il sera question de challenger les méthodes présentées ci-dessus en utilisant les algorithmes de Machine Learning pour calibrer les lois de rachats totaux sur les produits "prod_A" et "prod_B".

3.2 Description du portefeuille

3.2.1 Description des lois de rachats totaux sur les produits de l'étude

Les produits "prod_A" et "prod_B" objet de cette étude sont des produits d'épargne et de retraite individuelle. La construction de la maille de produit "prod_A" fait intervenir un produit d'épargne supplémentaire : il s'agit du produit "prod_C". Sur ces trois produits, les lois de rachats totaux réels issues de la base de données sont comparées à celles utilisées dans le modèle déterministe.

Les lois de rachats totaux sont calibrées par réseau de distribution des produits d'assurance. Il s'agit des personnes physique ou morale qui vendent des produits d'assurance aux particuliers. Ce sont donc les interlocuteurs privilégiés des assurés. Cette étude porte uniquement sur les contrats appartenant aux réseaux de distribution des salariés d'Allianz France "AF", des Agents Généraux "AG" et des courtiers "CT". Le réseau "partenariat" n'est donc pas pris en compte, mais représente peu de volume.

3.2.1.1 Lois de rachats totaux sur le produit "prod_A" et "prod_C"

Dans le modèle déterministe, les lois de rachats totaux sur les produits "prod_A" et "prod_C" sont identiques d'un réseau de distribution à l'autre. Sur le produit "prod_A", un peu plus de 95% des provisions proviennent du réseau des salariés "AF" ; Les taux de rachats réels augmentent avec l'ancienneté jusqu'à la 8^e année où on observe un pic de rachats totaux pour des raisons fiscales avantageuses. En effet, il existe un abattement fiscal en cas de rachats des contrats de plus de 8 ans de détention. Par ailleurs, les taux totaux pour des anciennetés supérieures à 29 ans sont identiques à cause du faible volume de provisions ; la moyenne des 5 dernières années est appliquée.

De manière générale, les taux de rachats totaux modélisés sur ce réseau sont deux fois plus importants que les rachats réels ; Le risque de rachats totaux est donc surestimé dans le modèle (Figure 3.3 ci-contre). Pour ce qui est des réseaux "AG" et "CT", ils possèdent très peu de volume, Les rachats réels sont assez volatiles ; on observe également une surestimation des taux de rachats dans le modèle.

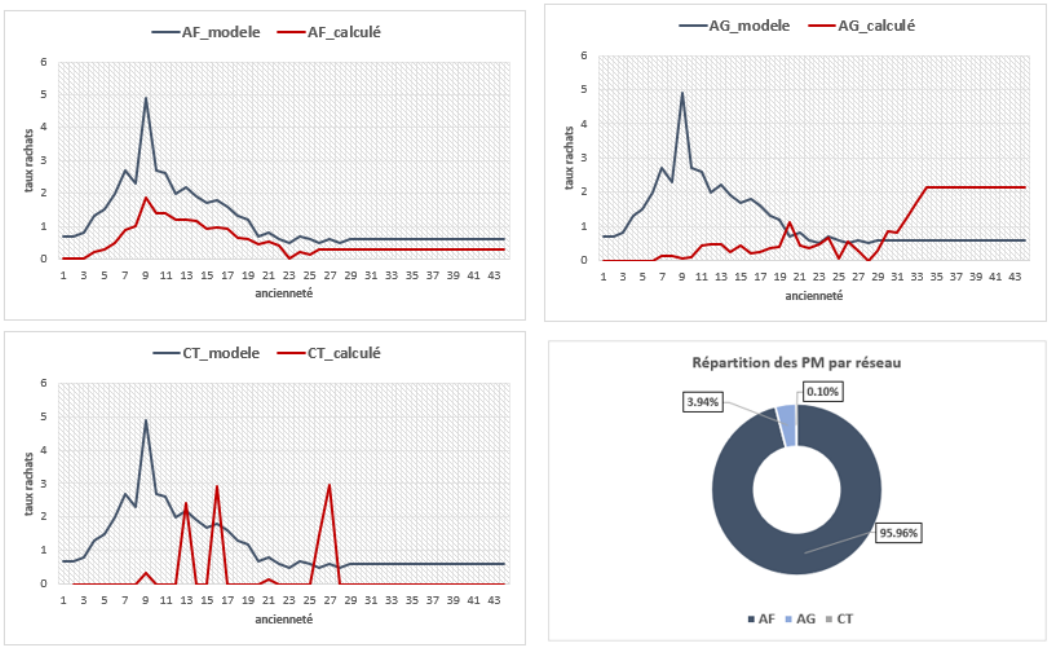


FIGURE 3.3 – Taux de rachats totaux réels et modélisés par ancienneté et par réseau sur le produit "prod_A"

Sur le produit "prod_C", la répartition des provisions par réseau est la même que sur le produit "prod_A". En revanche, les lois de rachats modélisées se rapprochent assez bien des taux réels sur le réseau des salariés "AF" (Figure 4.33 en annexe B). Les rachats totaux sur les contrats du réseau des agents généraux "AG" de moins de 15 ans d'ancienneté sont fortement surestimés dans le modèle. Les taux observés sur le réseau "CT" restent très volatiles à cause du très faible volume de provisions.

3.2.1.2 Lois de rachats totaux sur le produit "prod_B"

Le produit "prod_B" est un produit de retraite individuel qui permet au bénéficiaire de constituer un capital ou une rente dans le but d'améliorer la pension versée par les régimes de retraite obligatoires. La gestion pilotée est le mode de gestion par défaut sur ce produit. Elle vise à réduire progressivement le risque en réduisant l'épargne des bénéficiaires sur les supports UC (risqués) de la souscription jusqu'au terme du contrat. Chaque bénéficiaire fixe lui-même son horizon de placement et bénéficie ainsi d'un pilotage automatique et individualisé en fonction de son profil de risque.

Sur ce produit, la part la plus importante des provisions provient du réseau des agents généraux "AG" (82%) ; sur ce réseau, la loi de rachats du modèle déterministe se rapproche assez bien de la réalité. En revanche, les rachats totaux à la 8^e année d'ancienneté sont très surestimés dans le modèle interne. On observe en outre un pic de rachats précoces (sur les contrats jeunes de moins de 3 ans) causé par la défaillance de la politique marketing et de communication des agents généraux. Globalement, les rachats totaux diminuent avec l'ancienneté des contrats.

Sur le réseau des salariés "AF", les taux réels de rachats totaux sont deux fois moins élevés que ceux modélisés. Contrairement au pic de rachats habituel à la 8^e année d'ancienneté, ce pic est observé entre 17 et 18 ans d'ancienneté sur le réseau "AF". Ce qui se justifie par des rachats exceptionnels ainsi que des transferts liés au changement du type de support par les assurés.

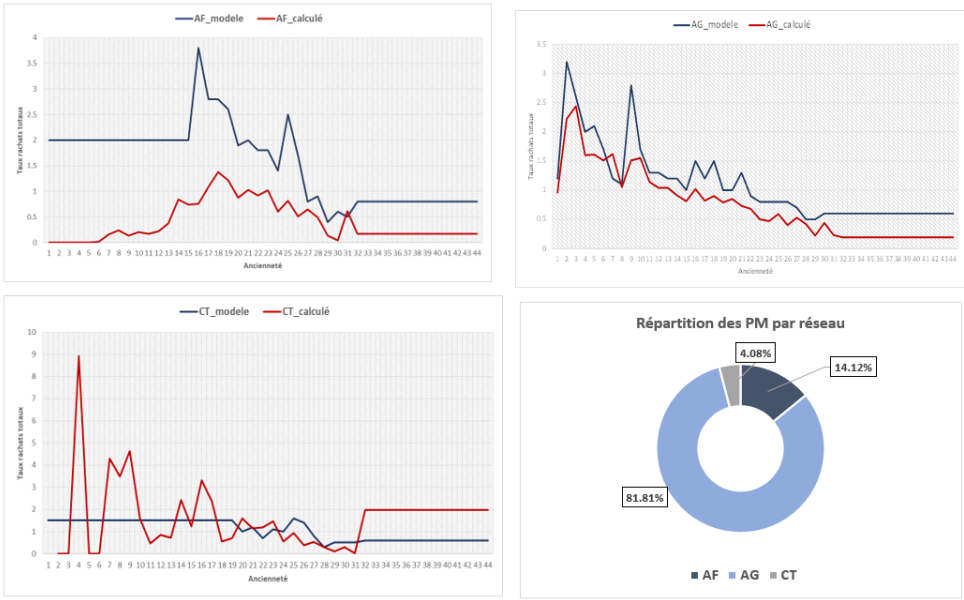


FIGURE 3.4 – Taux de rachats totaux réels et modélisés par ancienneté et par réseau sur le produit "prod_B"

Dans ce mémoire, des algorithmes de Machine Learning sont utilisés pour calibrer les lois de rachats totaux uniquement sur le réseau "AF" des produits "prod_A" et "prod_C" et les réseaux "AF" et "AG" du produit "prod_B" faute de volume de provisions sur les autres réseaux. Dans la suite, la base de données des rachats totaux enrichie des caractéristiques des assurés et des contrats, est étudiée en vue d'appréhender les caractéristiques du portefeuille dans un premier temps et analyser les associations entre ces caractéristiques et le comportement des assurés en matière de rachats totaux dans un second temps.

3.2.2 Caractéristiques générales des assurés et des contrats

- **Caractéristiques des assurés**

Les lois de rachats totaux utilisées dans le modèle déterministe et modélisées dans ce mémoire sont des lois structurels. Elles dépendent donc des caractéristiques des contrats et des assurés et non de la conjoncture économique. Le portefeuille de cette étude comporte 525 332 contrats dont 41% appartiennent au produit "prod_B". Les produits "prod_A" et "prod_C" représentent respectivement 43% et et 24% du portefeuille.

Les assurés sont en moyenne constitués de personnes âgées de 58 ans ; un assuré sur deux a plus de 60 ans. A la souscription, 50% des assurés sont âgés de plus de 50 ans ; en moyenne cet âge est de 47 ans, ce qui est normal puisqu'il s'agit des produits d'épargne et de retraite. Par ailleurs, sur l'ensemble du portefeuille étudié, les contrats sont assez vieux ; en effet, un contrat sur deux a plus de 12 ans d'ancienneté.

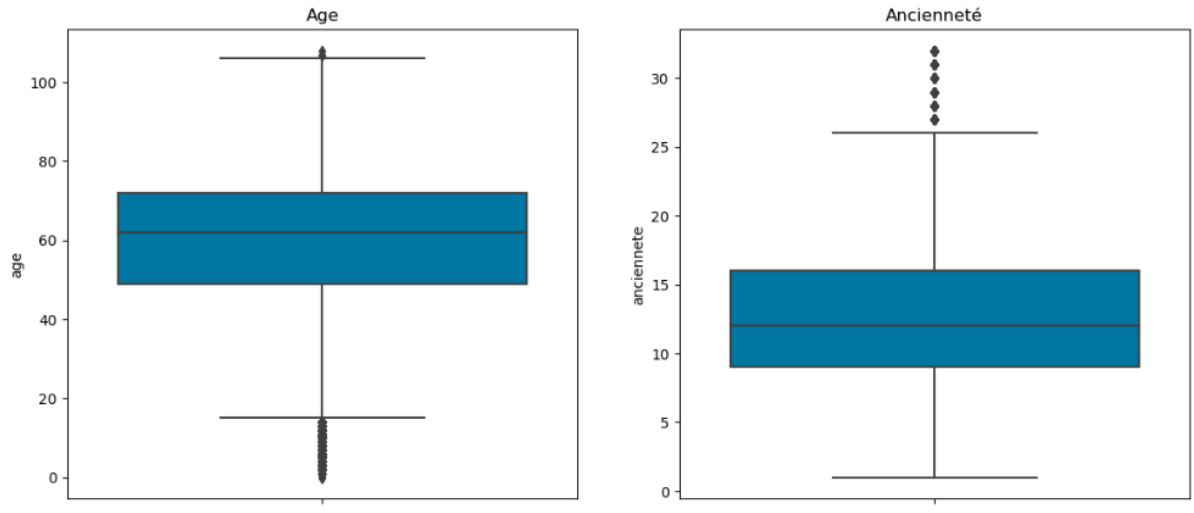


FIGURE 3.5 – Distribution de l'age et de l'ancienneté sur le portefeuille

La parité homme-femme est respectée dans le portefeuille car un souscripteur sur deux est une femme. Par ailleurs, 20,53% des souscripteurs sont des cadres et chefs d'entreprises, 30,4% sont des employés

ou exercent dans des professions intermédiaires, 24,37% sont des agriculteurs ou des ouvriers ; le taux d'inactivité dans ce portefeuille est assez élevé d'environ 14%.

- **Caractéristiques des contrats**

Pour ce qui est des contrats, l'encours moyen sur la période d'étude est de 38 543 € ; un assureur sur deux possède plus de 17 514 € d'épargne. En moyenne, les souscripteurs investissent 25,7% de leur encours sur les supports UC ; ils sont donc en majorité averse au risque. Par ailleurs, les contrats sont majoritairement de type multisupports Euro/UC (51,29%), seulement 8% des contrats sont des mono-supports UC. La moitié des contrats possèdent entre 2 et 5 supports d'investissement ;

Sur un peu plus de 8 contrats sur 10, la gestion est libre ; la majorité des souscripteurs gère donc leur épargne en toute autonomie. En outre, sur 71,62% des contrats, le versement des primes est unique (à la souscription). Sur les autres, le versement est mensuel, trimestriel, semestriel ou annuel. Le phénomène de rachats totaux est assez rare sur les produits de cette étude ; en effet, seulement 5,66% des contrats ont été totalement rachetés sur la période d'étude. Toutefois, en terme de volume, ces rachats totaux représentent près de 3% de l'encours sur l'ensemble du portefeuille.

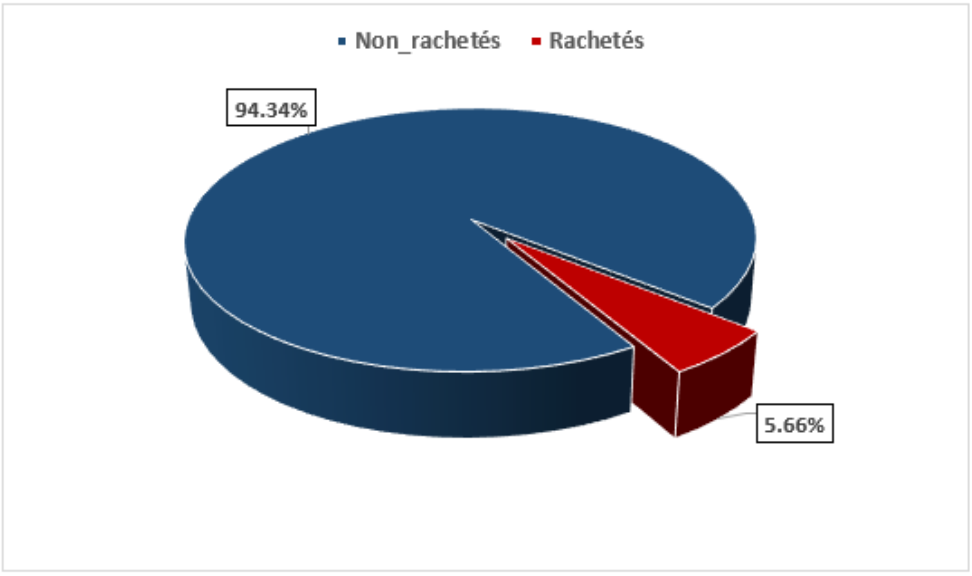


FIGURE 3.6 – Proportion de contrats totalement rachetés sur le portefeuille

La présentation des caractéristiques générales du portefeuille ainsi faite, il est important de mettre en évidence le lien entre ces caractéristiques et le comportement de rachats totaux des assurés.

3.2.3 Influence des caractéristiques des assurés et contrats sur les taux de rachats totaux par produit

3.2.4 Influence des caractéristiques des contrats

- **Influence de l'encours sur le contrat**

Sur le produit "prod_A", le niveau d'encours sur le contrat semble déterminer la décision et le taux de rachats totaux (la figure 3.7 ci-dessous). En effet, les contrats ayant un encours faible (moins de 10 000€) ont des taux de rachats totaux les plus importants atteignant le pic de 9% à la 8^e année d'ancienneté. Plus l'encours est élevé, moins les souscripteurs rachètent totalement leurs contrats ; ce qui s'explique par le fait que les assurés "riches" font moins souvent face à des besoins de liquidité. Toutefois, malgré le faible taux de rachats, le volume de provisions rachetées est deux fois plus élevé sur les gros contrats. Il s'agit donc des contrats les plus à risque.

Sur le produit "prod_B" et "prod_C", le constat similaire est fait. En revanche, les taux et montants rachetés sont plus faibles comme le montre la figure 4.34 en annexe B.

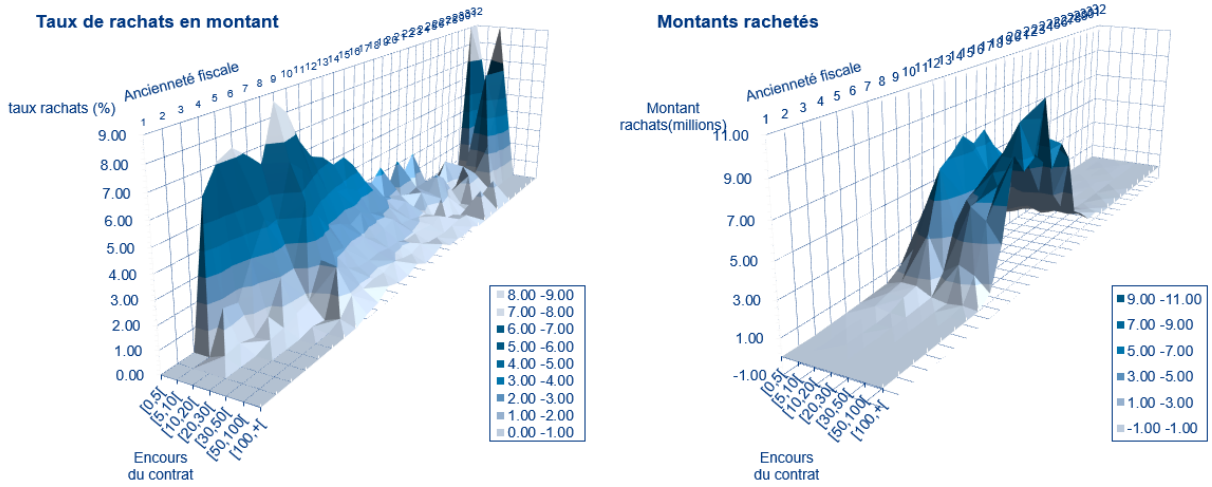


FIGURE 3.7 – Taux et montant de rachats totaux par ancienneté et par classe d'encours sur produit "prod_A"

- **Influence de la part d'UC**

Sur le produit "prod_A", plus la proportion de l'encours investie sur les supports en UC est importante, plus le taux de rachats totaux est élevé ; Sur les contrats dont cette proportion dépasse 60%, les taux de rachats atteignent leur pic (4,5%) à la 15^e année d'ancienneté. Le mauvais rendement des fonds UC sur ces dernières années explique ce constat. En effet, la valeur liquidative des support UC est fortement tributaire

des fluctuations du marché. Les assurés les plus prompts à prendre des risques vers les UC sont également les plus dynamiques dans leur gestion de portefeuille et leur propension à racheter.

Sur les contrats ayant moins de la moitié de leur encours sur les supports UC, les taux de rachats sont globalement assez bas, inférieurs à 2%. En revanche, les montants rachetés sont en moyenne 7 fois plus importants. Ce sont par conséquent les contrats les plus risqués en terme de rachats totaux pour l'assureur.

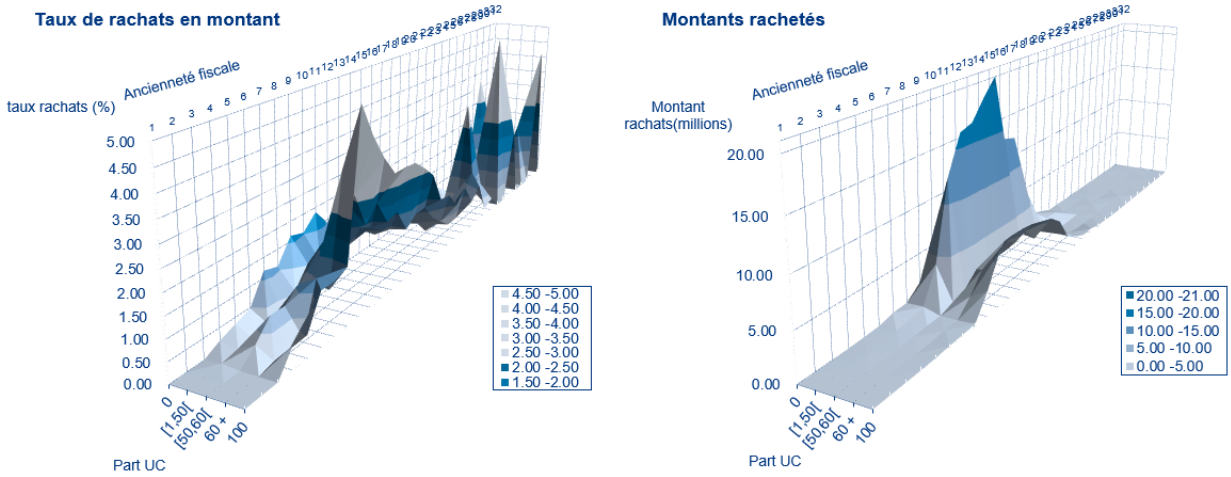


FIGURE 3.8 – Taux et montant de rachats totaux par ancienneté et part d'UC sur le produit "prod_A"

Sur le produit "prod_B", les constats similaires sont faits. En revanche, la relation entre part d'UC et taux de rachats semblent ambiguë sur le produit "prod_C" (figure 4.35 en annexe B).

- **Influence du mode de gestion des contrats**

Sur le produit "prod_A", les contrats en gestion libre (GL) représentent 80% du portefeuille ; ce sont les contrats les plus rachetés. Parmi ces derniers, ceux ayant entre 10 et 16 ans d'ancienneté sont les plus risqués car les montants rachetés excèdent 25 millions €. En revanche, les contrats en gestion sous mandat (GSM) présentent des taux et volume de rachats très bas (moins de 1% pour moins de 8 millions €) ; ceci s'explique par le faible volume de provisions associé à ce mode de gestion. Globalement sur les contrats de moins de 6 ans d'ancienneté, les rachats totaux sont très rares.

Sur les produits "prod_C" et "prod_B", les contrats en gestion libre (GL) sont également majoritaires et les plus rachetés (figure 4.36 en annexe B).

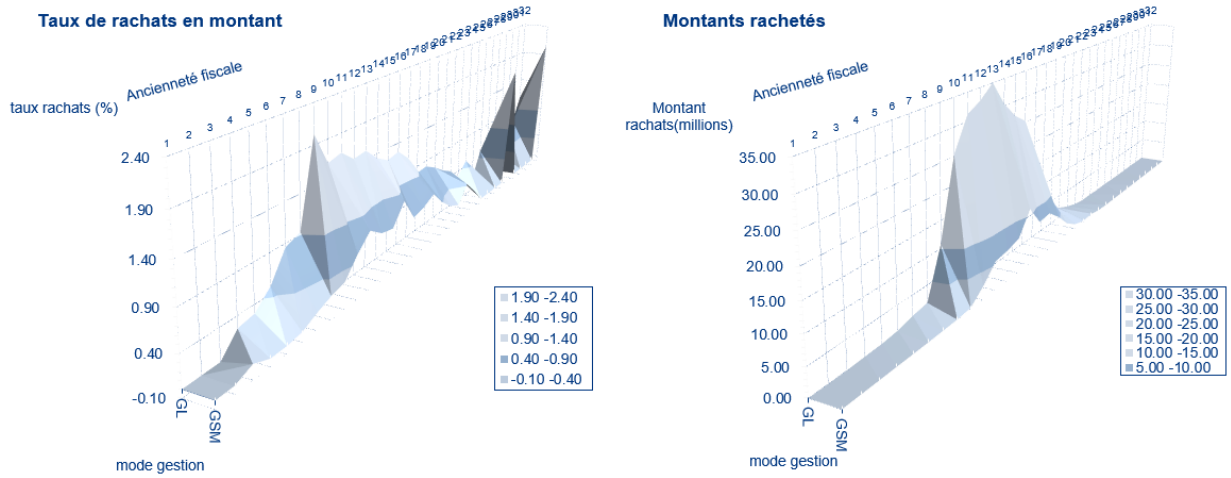


FIGURE 3.9 – Taux et montant de rachats totaux par ancienneté et par mode de gestion du contrat sur le produit "prod_A"

- **Influence de la périodicité de la prime**

La périodicité de versement de la prime semble avoir une influence sur la propension d'un assuré à racheter totalement son contrat. Lorsqu'il est périodique, le versement de la prime se fait de façon annuelle, semestrielle, trimestrielle ou mensuelle. En revanche, le versement unique se fait à la souscription du contrat. Sur le produit "prod_A", les contrats à versement périodique sont minoritaires (21,48%) ; il présente en revanche les taux de rachats totaux les plus élevés. Le taux de rachats totaux massifs est encore plus important à la 9^e année d'ancienneté.

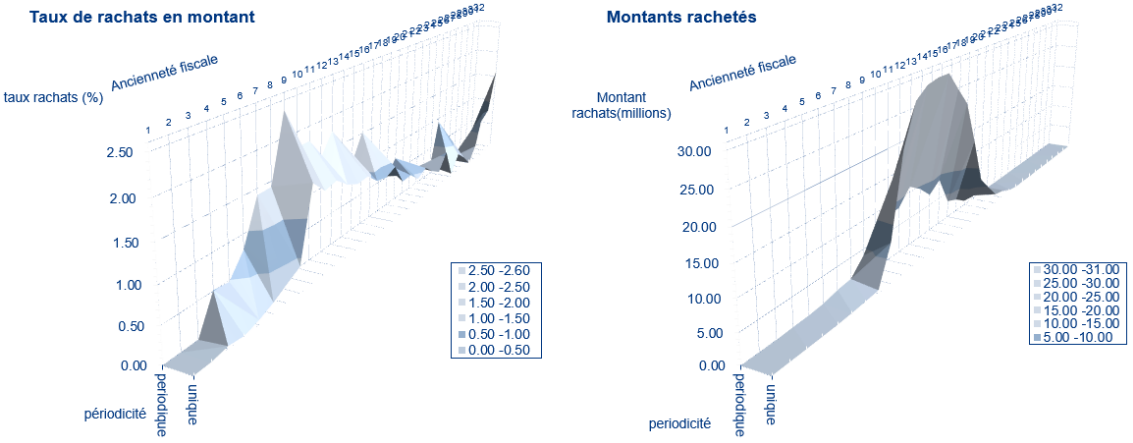


FIGURE 3.10 – Taux et montant de rachats totaux par ancienneté et par périodicité de la prime sur le produit "prod_A"

Malgré les taux de rachats globalement bas sur les contrats à versement unique, les volumes de provisions rachetées sont en moyenne 3 fois plus importants (atteignant 30 millions €) entre la 10^e et la 18^e année d'ancienneté ; ce qui s'explique par le gros volume de provisions (82% du portefeuille) associé à ces contrats. Il en ressort donc que les contrats à prime unique sont les plus risqués en terme de rachats totaux. Sur les produits "prod_C" et "prod_B", les constats similaires sont faits (figure 4.37 en annexe B).

- **Influence du type et du nombre de support d'investissement**

Les contrats d'assurance Vie peuvent être monosupport ou multisupports. Un contrat est dit monosupport lorsque l'épargne est investie uniquement sur les supports Euro ou UC ; lorsqu'elle est répartie simultanément sur les supports Euro et UC, le contrat est dit multisupport. Sur le produit "prod_A", les contrats multisupports Euro et UC représentent 76,73% du portefeuille. Ce sont les contrats les moins rachetés totalement ; les détenteurs de ce type de contrats sont plus enclins en effet à effectuer des arbitrages vers le support Euro en cas de mauvais rendements plutôt que de racheter. En revanche, les volumes de provisions rachetées sur ces contrats sont plus élevés. Sur le produit "prod_C", le même constat est fait. Sur le produit "prod_B", les montants de rachats totaux sont plus faibles sur les contrats multisupports ; ils sont en effet minoritaires sur ce portefeuille.

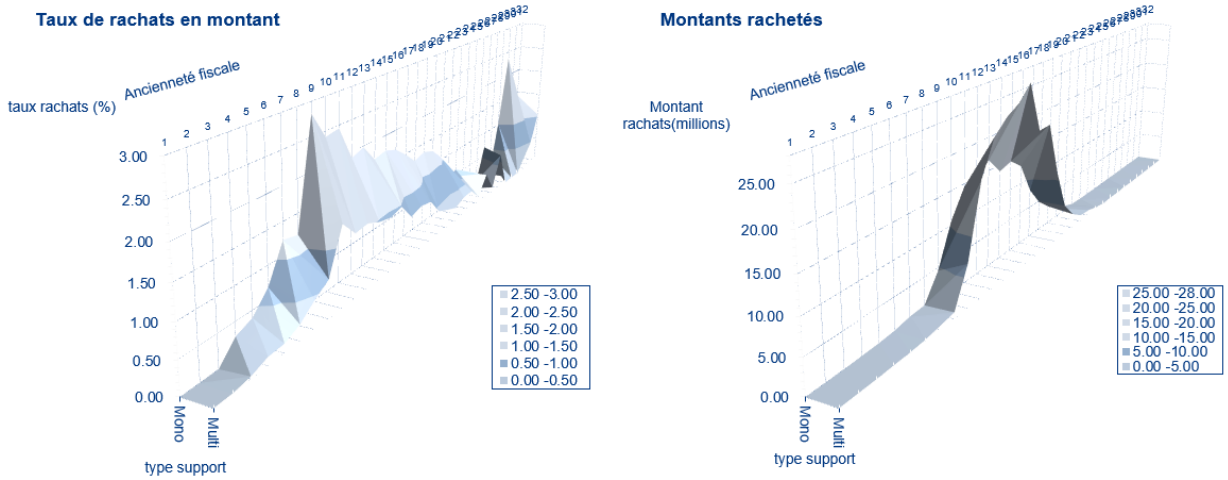


FIGURE 3.11 – Taux et montant de rachats totaux par ancienneté et par type de support sur le produit "prod_A"

L'épargne peut être investie sur un ou plusieurs supports d'investissement. On observe sur la figure 4.39 (en annexe B) que les taux et montants de rachats totaux diminuent avec le nombre de supports d'investissement. Les contrats qui possèdent un seul support d'investissement sont les plus risqués en terme de rachats totaux.

- **Influence des arbitrages et des rachats partiels**

L'objectif de ce mémoire est entre autres de quantifier l'influence des mouvements de rachats partiels et d'arbitrages sur le comportement de rachats totaux des assurés. La figure 3.12 ci-après montre des taux et montants de rachats totaux qui décroissent avec le nombre d'arbitrages antérieurs à l'acte de rachat total ; ainsi, plus les assurés arbitrent, moins ils sont enclins à racheter totalement leurs contrats. Le comportement passé du contrat en termes d'arbitrages semble donc affecter la décision future de rachat total.

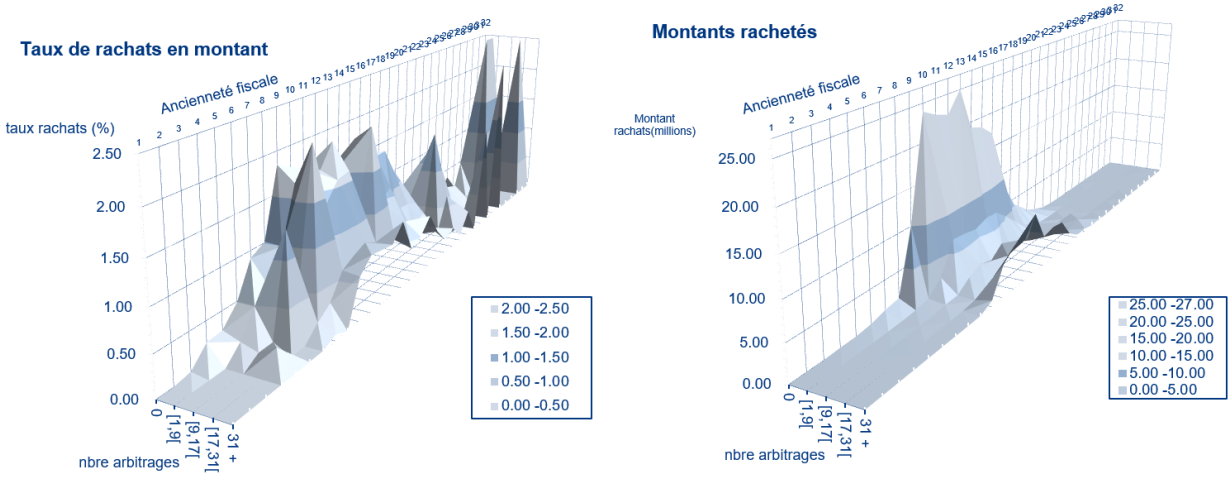


FIGURE 3.12 – Taux et montant de rachats totaux par ancienneté et nombre d'arbitrages sur le produit "prod_A"

Pour ce qui est des rachats partiels, aucun soupçon ne peut être émis à priori sur le sens et l'intensité de la relation avec les taux de rachats totaux. Toutefois, le volume de provisions rachetées sur les contrats sans aucun rachat partiel est très important à la 12^e année d'ancienneté ; ce sont donc les contrats les plus risqués.

3.2.5 Influence des caractéristiques des assurés

- **Influence de l'âge de l'assuré et de l'âge à la souscription**

L'âge de l'assuré au moment du rachat est un facteur de risque de rachats totaux. En effet, comme le montre la figure 3.13 ci-après sur le produit "prod_A", les taux et montants de rachats totaux décroissent à mesure que l'assuré vieillit. les jeunes seraient en effet plus confrontés à des besoins ponctuels de liquidité (achat d'une maison/voiture, mariage, études, voyage...) que les personnes plus âgées, qui quant à elles préparent leur succession/héritage en constituant une épargne ; le taux de rachats totaux assez faible observé chez les assurés de plus de 81 ans n'est donc pas surprenant. Les assurés jeunes sont donc les plus risqués

car plus enclins à effectuer des rachats totaux. Le constat similaire est fait sur les autres produits de l'étude (figure4.42 en annexe B).

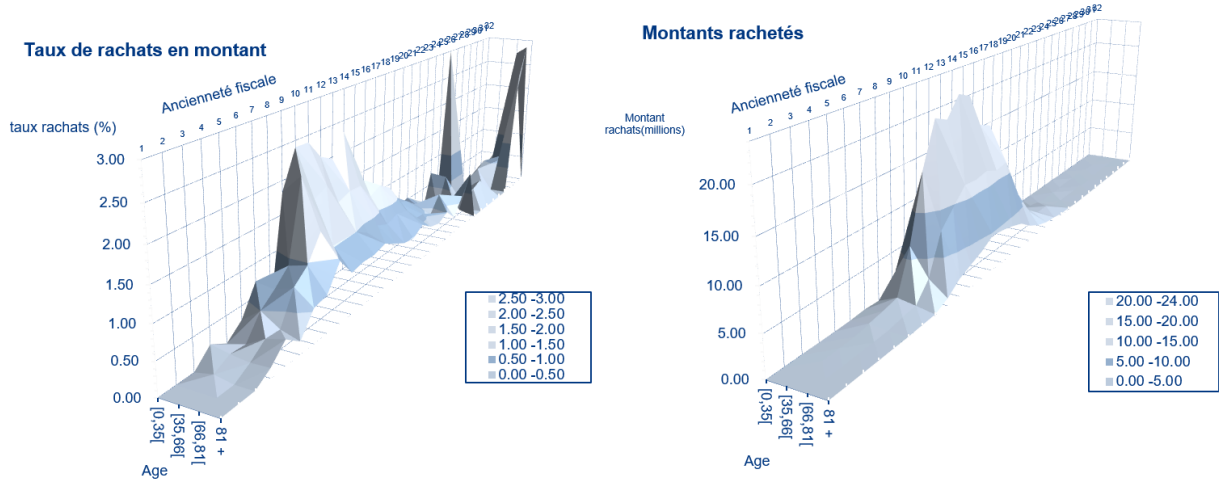


FIGURE 3.13 – Taux et montant de rachats totaux par ancienneté et par classe d'âge sur le produit "prod_A"

Pour ce qui est de l'âge à la souscription, On observe une décroissance moins importante des taux de rachats totaux avec cette variable. En revanche, plus l'assuré était jeune au moment de la souscription, plus le montant des rachats est faible ; le volume d'épargne est en effet moins important.

- **Influence du sexe de l'assuré**

Le sexe ne semble pas être un facteur discriminant des assurés en terme de comportement de rachats totaux. Comme le montre la figure 3.14, les taux et montants des rachats totaux sur le produit "prod_A" diffèrent très peu d'un assuré homme à une femme. Les hommes et les femmes sont en effet confrontés aux mêmes besoins de liquidité. La même conclusion est faite sur les produits "prod_C" et "prod_B" (figure 4.41).

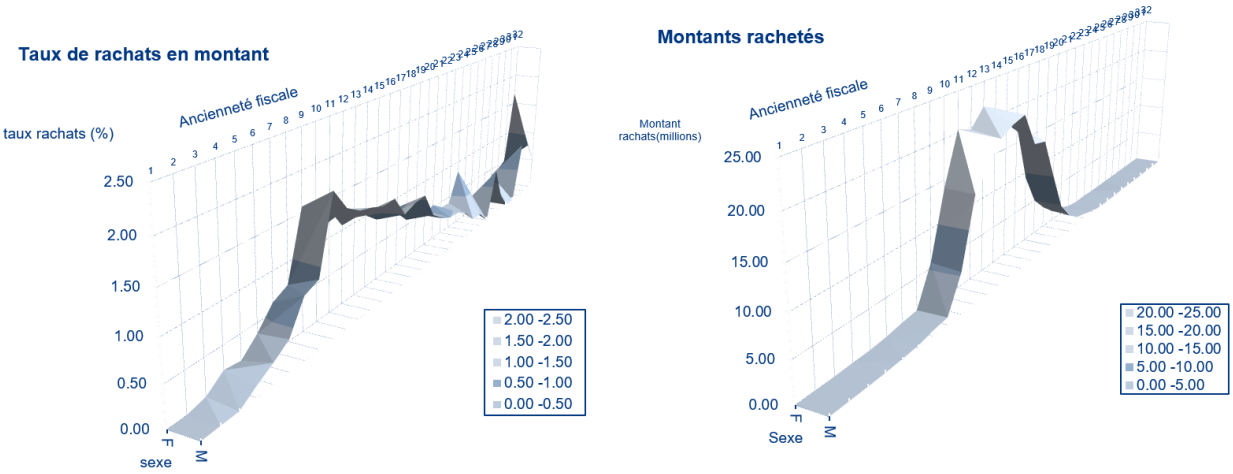


FIGURE 3.14 – Taux et montant de rachats totaux par ancienneté et par sexe sur le produit "prod_A"

- **Influence de la CSP de l'assuré**

D'une catégorie socio-professionnelle à l'autre, les taux de rachats totaux en montant varient assez faiblement sur l'ensemble des produits de l'étude (figure 3.15 ci-après et figure 4.43 en annexe B). En revanche, chez les inactifs, agriculteurs et ouvriers, artisans et commerçants, les montants rachetés sont plus faibles que ceux des cadres, professions libérales et intermédiaires, employés et fonctionnaires ; ce qui est normal car les revenus de ces derniers et par conséquent leurs épargnes sont plus élevés. Cette analyse présente toutefois des limites car les professions utilisées sont celles renseignées au moment de la souscription ; nous ne disposons pas des professions à jour des assurés.

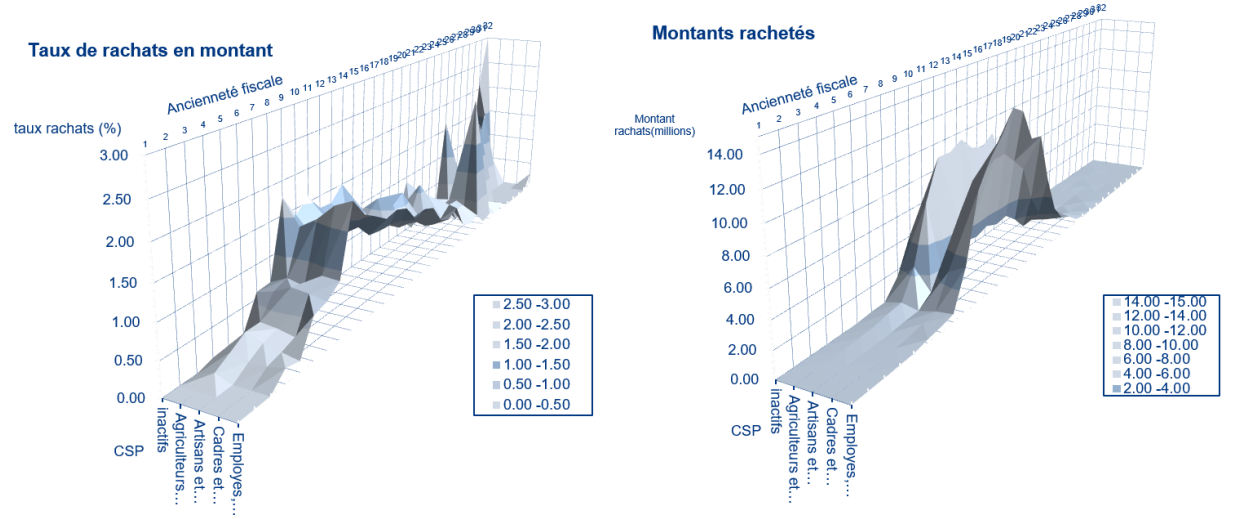


FIGURE 3.15 – Taux et montant de rachats totaux par ancienneté et par CSP sur le produit "prod_A"

Cette analyse descriptive a permis de mettre en évidence les caractéristiques socio-professionnelles des assurés, ainsi que les caractéristiques contractuelles et historiques des contrats qui pourraient expliquer les

comportements de rachats totaux sur le portefeuille de l'étude. Les caractéristiques les plus discriminantes seraient l'ancienneté, l'encours sur le contrat, la part de l'encours investie sur les supports en UC, le mode de gestion des contrats, le type de supports d'investissement et l'âge de l'assuré. Afin de quantifier l'influence de ces caractéristiques, des algorithmes de Machine Learning sont implémentés dans la partie suivante de ce mémoire. Cette partie s'attelle en effet à la calibration des lois de rachats totaux sur les trois produits de cette étude.

4 MODÈLES DE PRÉDICTION DES TAUX DE RACHATS TOTAUX

Dans ce chapitre, seront présentés d'une part les différents modèles mathématiques utilisés pour prédire les lois de rachats totaux par ancienneté d'une part et les différents résultats obtenus après modélisation d'autre part.

4.1 Présentation des outils de modélisation

Cette section s'attelle à présenter les différents modèles d'apprentissage statistique supervisé qui sont utilisés pour appréhender le comportement des assurés en matière de rachats totaux. Il s'agit de l'ensemble des méthodes pour lesquelles la variable cible est connue. Dans cette étude, Les méthodes classiques de régression et de classification ainsi que les algorithmes plus élaborés de Machine Learning qui sont utilisés. Soit X la matrice de variables explicatives et Y le vecteur de la variable à expliquer : $Y \in \mathbb{R}$ dans le cas d'une régression ou $Y \in \{1, 2, ..., K\}$ dans le cas d'une classification à K classes. Dans cette étude, la variable cible Y est quantitative : il s'agit des taux de rachats totaux. Il est donc question de rechercher une fonction f telle que : $Y = f(X) + \epsilon$ avec ϵ le bruit. La prédiction ou l'estimation faite par le modèle est : $\hat{Y} = f(X)$. Plusieurs métriques permettent d'apprécier les performances des modèles et de les comparer.

La description mathématique des algorithmes qui suivent est pour la plupart tirée de l'ouvrage "Introduction au Machine Learning" de Chloé-Agathe Azencott (Azencott 2018) et du cours d'apprentissage statistique de 3^e année de l'ENSAE de Arnak S. Dalalyan (Dalalyan 2022).

4.1.1 Régression logistique et régression linéaire

- **Régression logistique**

La régression logistique est un cas particulier des modèles linéaires généralisés. La formulation mathé-

matique de ces modèles est de la forme : $E[Y/X] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ avec β_j un vecteur de p paramètres correspondant aux p variables explicatives. Lorsque Y est binaire ($Y \in \{0,1\}$: loi de Bernouilli), cette espérance conditionnelle est donnée par :$E[Y/X = x] = P(Y = 1/X = x) = q$.

La transformation Logit s'écrit comme suit : $log(\frac{q}{1-q}) = X'\beta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$. La probabilité q_i correspondant à l'évènement $y_i = 1$ s'obtient donc ainsi : $q_i = \Lambda(x_i\beta) = \frac{1}{1+e^{-x_i\beta}}$.

L'estimation des paramètres du modèle de régression logistique se fait par la méthode du maximum de vraisemblance. Sous l'hypothèse d'indépendance entre les observations, on a l'expression de la vraisemblance suivante : $L(y,\beta) = \prod_{i=1}^{n} q_i^{y_i}(1-q_i)^{y_i} = \prod_{i=1}^{n} \Lambda(x_i\beta)^{y_i}(1-\Lambda(x_i\beta))^{1-y_i}$. En pratique, la log-vraisemblance est utilisée afin de réduire la complexité du problème d'optimisation :

$$l(y,\beta) = \sum_{i=1}^{n} y_i log(\Lambda(x_i\beta))(1-y_i)log(1-\Lambda(x_i\beta))$$

Finalement, les paramètres β sont estimés en résolvant le problème d'optimisation suivant avec des méthodes numériques de descente de gradient, ou de type Newton-Raphson :

$$\hat{\beta} = Argmin \sum_{i=1}^{n} y_i x_i \beta - log(1 + exp(x_i\beta))$$

Seuls les signes des paramètres du modèle peuvent faire l'objet d'interprétation. Ce modèle est par ailleurs très sensible à la multi-colinéarité entre les variables explicatives.

- **Régression linéaire**

On appelle régression linéaire le modèle de la forme : $f : \vec{x} \mapsto \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ dont les coefficients sont obtenus par minimisation de la somme des moindres carrés, à savoir :

$$Argmin_{\vec{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (y^i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_j))^2$$

Sous forme matricielle, la somme des moindres carrés s'écrit : $S = (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$; si X est de rang colonne plein, alors la solution optimal au problème est unique : $\beta^* = (X^TX)^{-1}X^T\vec{y}$. La régression linéaire produit un modèle interprétable car les β_j permettent de comprendre l'importance relative des variables sur la prédiction. En effet, plus $|\beta_j|$ est grand, plus la $j^{ème}$ variable a un effet important sur la prédiction, et le signe de β_j nous indique la direction de cet effet. Cette interprétation n'est toutefois valide que si les variables ne sont pas corrélées, et que x_j peut être modifiée sans perturber les autres variables.

Lorsque les variables explicatives sont corrélées, ou trop nombreuses, la complexité d'un modèle de régression est bien souvent trop élevée, ce qui conduit à une situation de sur-apprentissage. Régulariser ces modèles permet donc de contrôler les coefficients de régression ; en effet, les poids affectés à chacune des variables dans leur combinaison linéaire sont ajustés.

- **Régularisation des modèles**

Les méthodes de régularisation présentées dans la suite s'appliquent aussi bien dans le cas de la régression linéaire que dans le cas de la régression logistique. On appelle régularisation le fait d'apprendre un modèle en minimisant la somme du risque empirique sur le jeu d'apprentissage et d'un terme de contrainte Ω sur les solutions possibles :

$$f = argmin_{h \in F} \frac{1}{n} \sum_{i=1}^{n} L(h(\vec{x}^i), y^i) + \lambda \Omega(h)$$

Le coefficient de régularisation $\lambda \in \mathbb{R}_+$ contrôle l'importance relative de chacun des termes. quand λ tend vers 0, le terme de régularisation devient négligeable devant le terme d'erreur, et $\vec{\beta}$ prendra comme valeur une solution de la régression non régularisée. Dans le cas d'un modèle de régression linéaire, la fonction de perte est la somme des moindres carrés est utilisée comme suit : $f = argmin_{\vec{\beta} \in \mathbb{R}^{p+1}} (\vec{y} - X\vec{\beta})^T (\vec{y} - X\vec{\beta}) + \lambda \Omega(h)$.

Les différentes techniques de régularisation sont les suivantes :

— **Régression RIDGE** : Cette méthode consiste à utiliser comme régulariseur la norme l_2 du vecteur $\vec{\beta}$ $\Omega_{ridge}(\vec{\beta}) = \left\|\vec{\beta}\right\|_2^2 = \sum_{j=1}^{p} \beta_j^2$. La solution du problème d'optimisation devient $\beta^* = (\lambda I_p + X^T X)^{-1} X^T \vec{y}$;

— **Régression LASSO** : Lasso (Least Absolute Shrinkage and Selection Operator) est une méthode de réduction de dimension qui utilise les valeurs absolues des coefficients (norme l_1) pour éliminer les variables ayant un coefficient nul. Le régulariseur est donc la norme l_1 du vecteur $\vec{\beta}$: $\Omega_{lasso}(\vec{\beta}) = \left\|\vec{\beta}\right\|_1 = \sum_{j=1}^{p} |\beta_j|$;

— **Régression ELASTIC NET** : La régularisation LASSO permet d'obtenir un modèle parcimonieux et donc plus facilement interprétable, tandis que la régularisation RIDGE permet elle d'éviter le sur-apprentissage ainsi que grouper les variables corrélées. R La régression Elastic Net combine ces deux approches grâce au régulateur suivant : $\Omega_{enet}(\vec{\beta}) = \left((1-\alpha)\left\|\vec{\beta}\right\|_1 + \alpha\left\|\vec{\beta}\right\|_2^2\right)$. La solution de l'elastic net est obtenue par un algorithme à directions de descente.

La solution de l'elastic net est parcimonieuse, mais moins que celle du lasso. Lorsque plusieurs variables fortement corrélées sont en effet pertinentes, le lasso sélectionne une seule d'entre elles, tandis que, les autres les sélectionnent toutes et leur affecte le même coefficient. Le chemin de régularisation permet de décrire l'évolution de la valeur du coefficient de régression d'une variable en fonction du coefficient de régularisation ; ce qui permet donc de comprendre l'effet de la régularisation sur les valeurs de β_j.

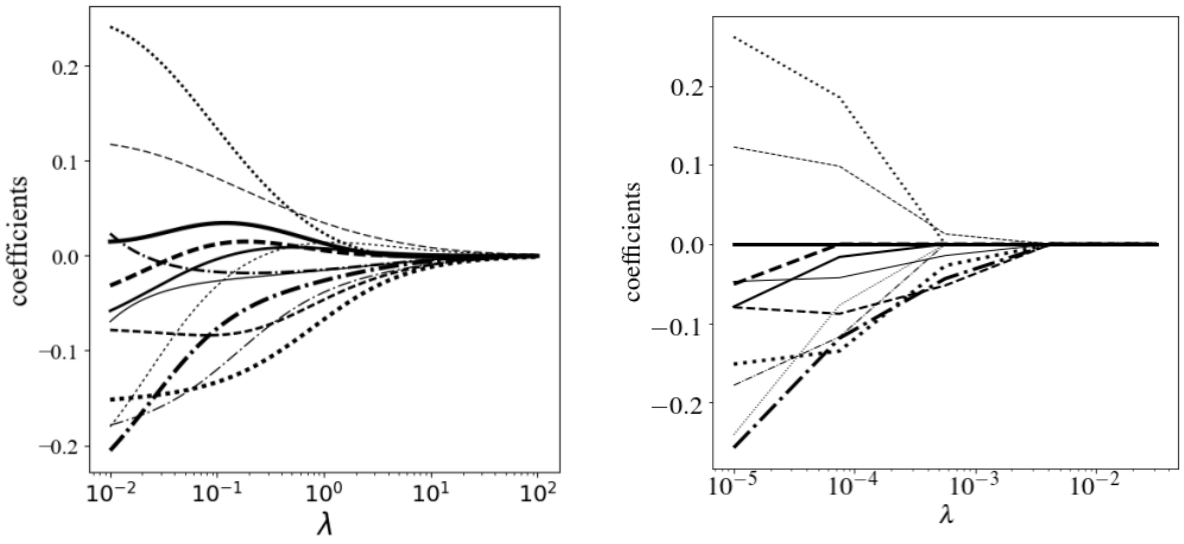


FIGURE 4.1 – Chemin de régularisation Ridge (à gauche) et Lasso (à droite)

4.1.2 Les supports vectors machines : SVM

Les machines à vecteurs de support ou SVM de l'anglais support vector machines sont de puissants algorithmes d'apprentissage automatique basé sur un algorithme linéaire proposé par Vladimir Vapnik et Aleksandre Lerner en 1963 (Vapnik V., 2014), mais permettent également d'apprendre des modèles non-linéaires grâce à l'astuce du noyau.

- **Le cas linéairement séparable**

En se plaçant dans le cas d'un problème de classification binaire, on suppose qu'il est possible de trouver un modèle linéaire qui ne fasse pas d'erreurs sur les données. Soit $D = \left\{ \left(\vec{x^i}, y^i \right)_{i=1,\dots,n} \right\}$ un jeu de données avec $\vec{x} \in \mathbb{R}^p$ et $y^i \in \{-1, 1\}$. On dit que D est linéairement séparable s'il existe au moins un hyperplan dans $\mathbb{R}^p$ tel que tous les points étiquetés +1 soient d'un côté de cet hyperplan et tous les points étiquetés -1 de l'autre. sur la figure 4.3 suivante, il existe une infinité d'hyperplans séparateurs qui ne font aucune erreur de classification et équivalents du point de vue de la minimisation du risque empirique.

On appelle vecteurs de support, les observations du jeu d'entraînement situés à une distance γ de l'hyperplan séparateur H ; elles soutiennent les hyperplans H+ et H- (ce sont les hyperplans parallèles à H et situés à une distance γ de part et d'autre de γ. La distance γ d'un hyperplan encore appelée *marge* est la distance de cet hyperplan à l'observation du jeu d'entraînement la plus proche. L'équation de l'hyperplan séparateur H que nous cherchons est de la forme $< \vec{w}, \vec{x} > +b = 0$ où $<,>$ représente le produit scalaire sur $\mathbb{R}^p$. Le problème revient à maximiser la distance γ sous n contraintes :

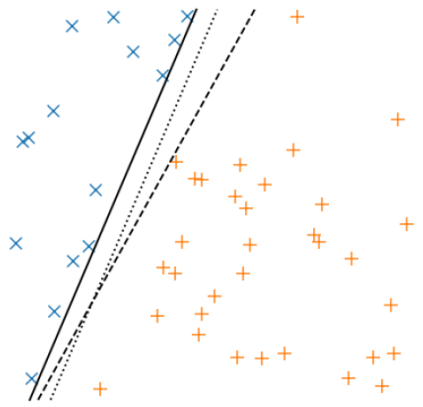


FIGURE 4.2 – Exemples d'hyperplans séparateurs linéaires

$$\begin{cases} argmin_{\vec{w}\in\mathbb{R}^P, b\in\mathbb{R}} & \frac{1}{2}\|\vec{w}\|_2^2 \\ s/c & y^i(<\vec{w},\vec{x}^i>+b)\geq 1 \quad i=1,...,n \end{cases}$$

Les contraintes permettent de s'assurer que chaque observation est soit positive soit négative. Supposons $\vec{w}^*$ et b^* les solutions du problème ci-dessus, la fonction de décision est donnée par : $f(\vec{x}) =< \vec{w}^*, \vec{x} > +b^*$

- **Le cas linéairement non séparable**

En pratique, les données ne sont généralement pas linéairement séparables. Dans ce cas, quel que soit l'hyperplan séparateur que l'on choisisse, certains des points seront mal classifiés ; le but est de trouver un compromis entre les erreurs de classification et la taille de la marge γ. Un hyperparamètre de coût C est introduit afin de contrôler l'importance relative de la marge et des erreurs du modèle sur le jeu d'entraînement. En posant $\xi_i = [1 - y^i f(\vec{x}^i)]_+$, le problème d'optimisation devient :

$$\begin{cases} argmin_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n} & \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\ s/c & y^i(<\vec{w},\vec{x}^i>+b)\geq 1 \quad i=1,...,n \\ & \xi_i \geq 0, \qquad i=1,...,n. \end{cases}$$

$\xi_i = [1 - y^i f(\vec{x}^i)]_+$ représente la fonction de coût qui permet s'assurer autant que possible que toute observation $\vec{x}$ d'étiquette y soit située à l'extérieur de la zone d'indécision.

- **Le cas non linéaire : SVM à noyau**

Il est fréquent qu'une fonction linéaire ne soit pas appropriée pour séparer nos données ; dans ce cas, on définit un espace H permettant au moyen d'une application ϕ, d'utiliser un algorithme SVM linéaire pour résoudre un problème non linéaire. Ainsi, on appelle SVM à noyau la solution du problème d'optimisation suivant :

$$\begin{cases} Max_{\vec{\alpha}\in\mathbb{R}} & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}\alpha_i\alpha_j y^i y^j k(\vec{x}^i,\vec{x}^l) \\ s/c & \sum_{i=1}^{n}\alpha_i y^i \\ & 0 \leq \alpha_i \leq C \qquad\qquad i = 1,...,n \end{cases}$$

$k(\vec{x},\vec{x}^{'}) =< \phi(\vec{x}),\phi(\vec{x}^{'}) >_H$ représente la fonction noyau et la fonction de décision est donnée par : $f(\vec{x}) = \sum_{i=1}^{n}\alpha_i^* y^i k(\vec{x}^i,\vec{x}) + b^*$. Le noyau polynomial est : $k(\vec{x},\vec{x}^{'}) = (< \vec{x},\vec{x}^{'} > +c)^d$, le noyau radial gaussien : $k(\vec{x},\vec{x}^{'}) = exp(-\frac{\left\|\vec{x}-\vec{x}^{'}\right\|^2}{2\sigma^2})$.

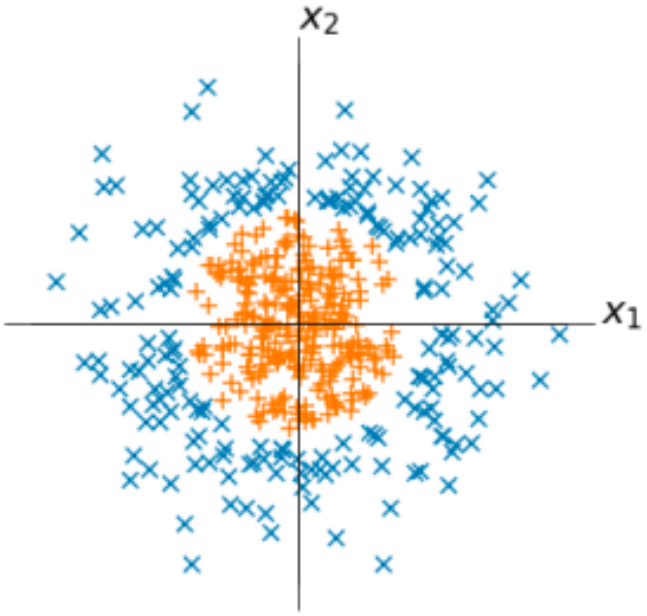


FIGURE 4.3 – Exemples séparation non linéaire

Lorsqu'il s'agit d'un problème de classification, on parle d'algorithme SVC (Support Vector Classifier) ; dans la cas d'un problème de régression, il s'agit de l'algorithme SVR (Support Vector Regression).

4.1.3 L'algorithme des plus proches voisins : KNN

Cet algorithme, dit des plus proches voisins, se base sur le principe de « qui se ressemble s'assemble» ; le principe consiste à utiliser les étiquettes des exemples les plus proches pour prendre une décision. Cet algorithme s'applique aussi bien à un problème de classification que de régression. Étant donné un jeu $D = \left\{\left(\vec{x^i}, y^i\right)_{i=1,...,n}\right\}$ de n observations étiquetées, une distance d sur χ et un hyperparamètre $k \in \mathbb{N}^*$; on appelle algorithme des k plus proches voisins, ou kNN pour k nearest neighbors en anglais, l'algorithme consistant à étiqueter une nouvelle observation x par l'étiquette des k points points du jeu d'entraînement dont elle est la plus proche suivant la distance d prédéfinie. En notant $N_k(\vec{x})$ l'ensemble des k plus proches voisins de $\vec{x}$ dans D :

- Pour un problème de classification, pour chaque classe c, on applique le vote de la majorité en d'autres termes $\vec{x}$ prend l'étiquette majoritaire parmi celles de ses k plus proches voisins : $f(\vec{x}) = argmax_c \sum_{i,\vec{x}^i \in N_k(\vec{x})} \delta(y^i, c)$;
- Pour un problème de régression, $\vec{x}$ prend comme étiquette la moyenne des étiquettes de ses k plus proches voisins : $f(\vec{x}) = argmax_c \sum_{i,\vec{x}^i \in N_k(\vec{x})} y^i$.

L'algorithme KNN est un exemple d'apprentissage non paramétrique car la fonction de décision s'exprime en fonction des données observées et non pas comme une formule analytique fonction des variables. Pour prendre en compte la notion selon laquelle les voisins véritablement proches sont plus fiables pour la prédiction que ceux plus éloignés, la contribution de chacun des voisins est pondérée en fonction de sa distance à l'observation à étiqueter. Les poids sont calculés comme suit :

$$w_i = \frac{1}{d(\vec{x},\vec{x}^i)} \text{ ou } w_i = e^{-(\frac{1}{2}d(\vec{x},\vec{x}^i))}$$

Dans le cas où $\chi = \mathbb{R}^{\mathbb{P}}$, on utilise le plus souvent la distance de Minkowski :

$$\|\vec{u} - \vec{v}\|_q = (\sum_{j=1}^{p} \vec{u} - \vec{v}|^q)^{\frac{1}{q}}$$

Lorsque $q = 2$, cette distance correspond à la La distance euclidienne : $d_2(\vec{u}, \vec{v}) = \sqrt{\sum_{j=1}^{p}(u_j - v_j)^2}$.

La frontière de décision de l'algorithme des k plus proches voisins est linéaire par morceaux et le choix de k se fera généralement en utilisant une validation croisée.

FIGURE 4.4 – Frontière de décision d'un algorithme des 5 plus proches voisins

4.1.4 L'arbre de décision CART et les forêts aléatoires

L'algorithme des plus proches voisins (KNN) permet de construire des modèles non paramétriques, c'est-à-dire qu'ils reposent sur la définition d'une distance ou similarité pertinente entre les observations. Les arbres de décision approchent le problème différemment. La classification ou la régression par arbre de

décision est une méthode d'apprentissage supervisée construite à partir d'une suite récursive de règles de division qui fournit une représentation graphique simple du prédicteur ; il permet de distinguer les variables qui contribuent le plus à la variation de la variable d'intérêt. Par conséquent, les résultats de ce modèle sont faciles à interpréter. L'algorithme CART popularisé par Breiman et al. (Breiman & Stone 1984) est le plus répandu, par conséquent utilisé dans le cadre de cette étude.

À chaque nœud d'un arbre de décision construit par CART correspond une variable séparatrice selon laquelle vont être partitionnées les données. Dans le cas où la variable de séparation est une variable discrète, sont définies les deux régions suivantes : $R_l(j,s) = \{\vec{x} x_j = s\}$ et $R_r(j,s) = \{\vec{x} \mid x_j \neq s\}$ | lorsque la variable de séparation est une variable réelle,on définit la valeur s de l'attribut par rapport à laquelle va se faire la décision comme suit : $R_l(j,s) = \{\vec{x} \mid x_j < s\}$ et $R_r(j,s) = \{\vec{x} \mid x_j \geq s\}$.

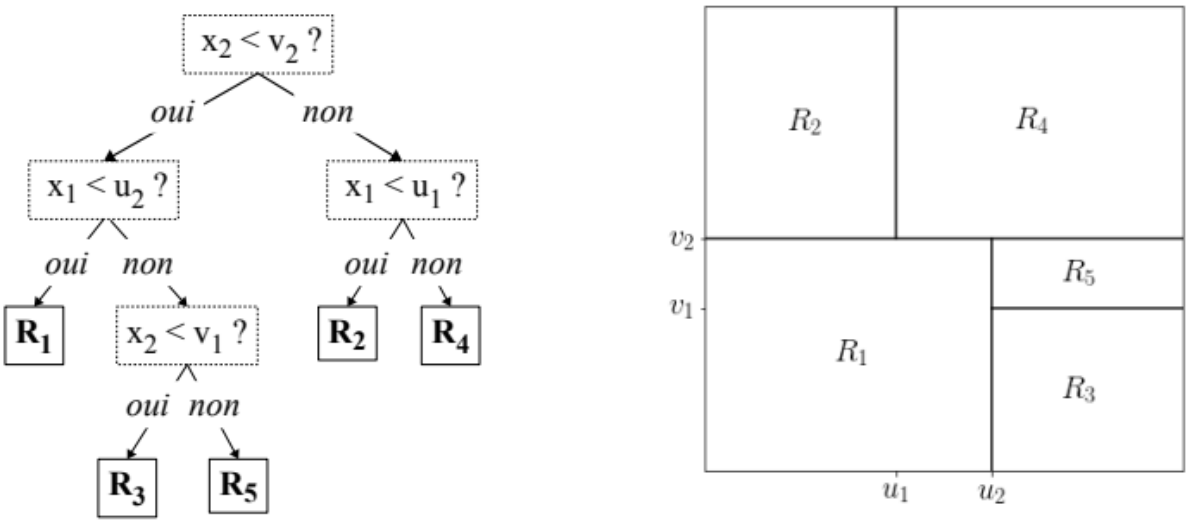


FIGURE 4.5 – L'arbre de décision (à gauche) partitionne $\mathbb{R}^2$ en 5 zones (à droite)

À chaque itération de cet algorithme, on itère sur toutes les valeurs possibles de s pour déterminer le couple (j,s) qui minimise un critère prédéfini. Finalement, le problème d'optimisation est le suivant :

$$\begin{cases} Argmin_{j,s} = \left(\sum_{i \mid \vec{x}^i \in R_l(j,s)} \left(y^i - y_l(j,s)\right)^2 + \sum_{i \mid \vec{x}^i \in R_r(j,s)} \left(y^i - y_r(j,s)\right)^2 \right) & regression \\ Argmin_{j,s} = \left(\frac{|R_l(j,s)|}{n} Imp(R_l(j,s)) + \frac{|R_r(j,s)|}{n} Imp(R_r(j,s)) \right) & Classification \end{cases}$$

$Imp(R)$ représente un critère d'impureté. L'impureté de Gini d'une région R est définie comme : $Imp(R) = \sum_{c=1}^{C} p_c(R)(1 - p_c(R))$ avec $p_c(R)$ qui indique la proportion d'exemples d'entraînement de la région R qui appartiennent à la classe c.

Bien que les arbres de décision présentent d'intéressantes propriétés, ils ont tendance à mal apprendre et à avoir de faibles propriétés de généralisation ; les méthodes ensemblistes comme les forêts aléatoires

permettent de résoudre ce problème.

- **Les forêts aléatoires**

L'idée des forêts aléatoires, proposée par Leo Breiman, est de construire des arbres individuels non seulement sur des échantillons différents, mais aussi en utilisant des variables différentes (Breiman, 2001). Plus précisément, à chaque nœud, on commence par sélectionner $q < p$ variables aléatoirement, avant de choisir la variable séparatrice parmi celles-ci. En classification, on utilise typiquement $q\sqrt{p}$, ce qui permet aussi de réduire considérablement les temps de calculs puisqu'on ne considère que peu de variables à chaque nœud ; Pour la régression, le choix par défaut est plutôt de $q\frac{p}{3}$. De cette manière, un nombre B d'arbres sont construits et les B prédictions sont ensuite combinées par vote de la majorité dans le cas d'un problème de classification ou en prenant la moyenne dans le cas d'un problème de régression.

Bien que les forêts aléatoires soient un des algorithmes les plus performants et les plus simples à mettre en place, cet algorithme à l'inconvénient d'être difficilement interprétable, il se comporte en effet comme une boite noire.

4.1.5 La validation croisée "Leave-one-out"

La séparation d'un jeu de données en un jeu d'apprentissage et un jeu de test est nécessairement arbitraire. Le risque est de créer aléatoirement des jeux de données qui ne sont pas représentatifs. Pour éviter ce problème, il est préférable de reproduire plusieurs fois la procédure, puis de retenir la moyenne des résultats et ainsi moyenner ces effets aléatoires. Pour ce faire, la validation croisée est la méthode la plus classique.

Étant donné un jeu D de n observations, et un nombre K, on appelle validation croisée la méthode décrite comme suit :

1. Partitionner D en K parties de tailles sensiblement similaires : $(D_1, D_2, ..., D_K)$;
2. Entraîner le modèle choisi pour chaque valeur de $k = 1, ..., K$, $\bigcup_{l \neq k} D_l$ et l'évaluer sur D_k.

Chaque observation étiquetée du jeu D appartient à un unique jeu de test, et à $(K1)$ jeux d'apprentissage. Cette procédure génère une prédiction par observation de D. Pour conclure sur la performance du modèle, on évalue la qualité de chacun des K prédicteurs sur le jeu de test D_k correspondant et on calcule une moyenne de leurs performances ; ce qui permet de se faire une meilleure idée de la variabilité de la qualité des prédictions au regard de l'écart-type de ces performances.

- **Le "Leave-one-out"**

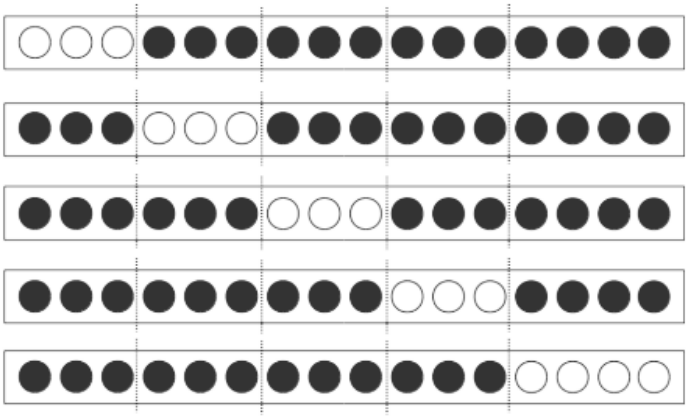


FIGURE 4.6 – Validation croisée avec 5 folds

Une validation croisée ""Leave-one-out" est celle dont le nombre de folds est égal au nombre d'observations dans le jeu d'apprentissage, et dont chaque fold est donc composé d'un jeu d'apprentissage de taille $n1$ et d'un jeu de test de taille 1 ; on met donc de côté, pour chaque fold, une unique observation. Cette approche est idéal lorsque le jeu de données est de petite taille.

Cette section nous a permis de passer en revue les différentes méthodes utilisées dans cette étude afin d'expliquer et de prédire au mieux les comportements de rachats totaux des assurés. La section suivante présente les résultats de ces modèles appliqués aux produits "prod_A", "prod_C" et "prod_B".

4.2 Résultats des modèles de prédiction et régression

Dans cette section, nous nous attellerons à présenter les résultats des modèles de calibration des lois de rachats totaux : il s'agit des modèles de régression des taux de rachats totaux par ancienneté sur la période de 2015 à 2021. L'étude se limite uniquement à la modélisation des rachats totaux structurels ; les effets structurels sont liés à la structure du portefeuille qui englobe les caractéristiques des contrats et des assurés, les comportements antérieurs en terme d'arbitrages et de rachats partiels, etc. Les facteurs de rachats totaux liés à la conjoncture en général notamment le contexte économique et financier (chômage, PIB, croissance, inflation, etc.), l'image de la compagnie, l'écart des taux avec la concurrence, l'évolution de l'offre, les stratégies de vente, ne sont pas pris en compte dans cette étude pour les raisons suivantes :

- Le caractère incertain du phénomène et l'absence d'historique sur les rachats conjoncturels. En effet, aucun rachat conjoncturel n'a été observé sur l'ensemble du portefeuille d'Allianz au cours des 25 dernières années ; ce qui s'explique par les taux durablement bas ces deux dernières décennies ;
- La non-prise en compte des rachats conjoncturels dans le modèle interne déterministe. Il s'agit en

effet du modèle qui est utilisé dans ce mémoire pour évaluer les différents écarts d'expérience.

4.2.1 Présentation de la base de données de modélisation

Afin de tenir compte de l'information à la maille contrat (caractéristiques des assurés et des contrats) utilisée dans les statistiques descriptives, la base de données utilisée pour la modélisation stocke pour chaque ancienneté et par année d'observation, la proportion de provisions associée aux catégories de variables explicatives. De cette manière, l'information micro est prise en compte. De façon pratique, la variable "sexe_F" par exemple indique la proportion de provisions ou d'encours correspondant aux assurés de sexe féminin pour un produit, une année d'observation et une ancienneté donnée.

Ainsi, la base de données pour la modélisation est agrégée à la maille ancienneté x année x produit, à laquelle s'ajoutent les variables explicatives suivantes :

- Les taux de rachats totaux, de rachats partiels et d'arbitrages nets par produit à la maille considérée ;
- La proportion de provisions associée à chaque classe de supports d'investissement en nombre, au type de supports, à la périodicité de versement de la prime et au mode de gestion des contrats ;
- La proportion de provisions associée aux classes d'encours sur contrats ;
- La proportion de provisions associée aux assurés femmes et à chaque catégorie de CSP (csp_bas, csp_mid, et csp_haut). Les CSP ayant des profils similaires en terme de comportements de rachats totaux sont regroupées. La csp_bas contient commerçants, ouvriers, agriculteurs et inactifs ; la csp_mid regroupe les employés, professions libérales et intermédiaires ; la csp_haut contient les cadres et chefs d'entreprises.
- La proportion de provisions associée à chaque classe d'âge actuariel, d'âge à la souscription ;
- La part d'UC, l'âge à la souscription et l'âge moyen à la maille de l'étude ;
- L'ancienneté des contrats qui oriente les comportements de rachats au travers des avantages fiscaux ;

De cette manière, trois bases de données sont construites respectivement pour les produits "prod_A", "prod_B" et "prod_C".

4.2.2 Calibration des lois de rachats totaux sur le produit "prod_A"

Dans le modèle de projection de cash-flows sur le produit "prod_A", la loi de rachats totaux associée au réseau des salariés Allianz France "AF" est également utilisée pour les réseaux de distribution des agents généraux "AG" et des courtiers "CT" ; le réseau "AF" est en effet majoritaire avec près de 96% de l'encours

total sur le portefeuille. Ainsi, dans cette étude, seule la loi de rachats associée au réseau "AF" est estimée. Dans la suite, l'abréviation "prod_A-AF" est utilisée pour designer le produit "prod_A" du réseau de distribution "AF".

4.2.2.1 Analyse des corrélations entre les variables : produit "prod_A-AF"

L'analyse des corrélations a pour but de réduire le nombre de variables explicatives ; ce qui permet de limiter les effets négatifs d'interactions de variables corrélées dans nos modèles d'une part éviter le sur-apprentissage d'autre part. Le coefficient corrélation de Spearman est calculé car il ne dépende pas de la distribution des données ; il est donc plus robuste. La matrice de corrélation entre les variables est la suivante :

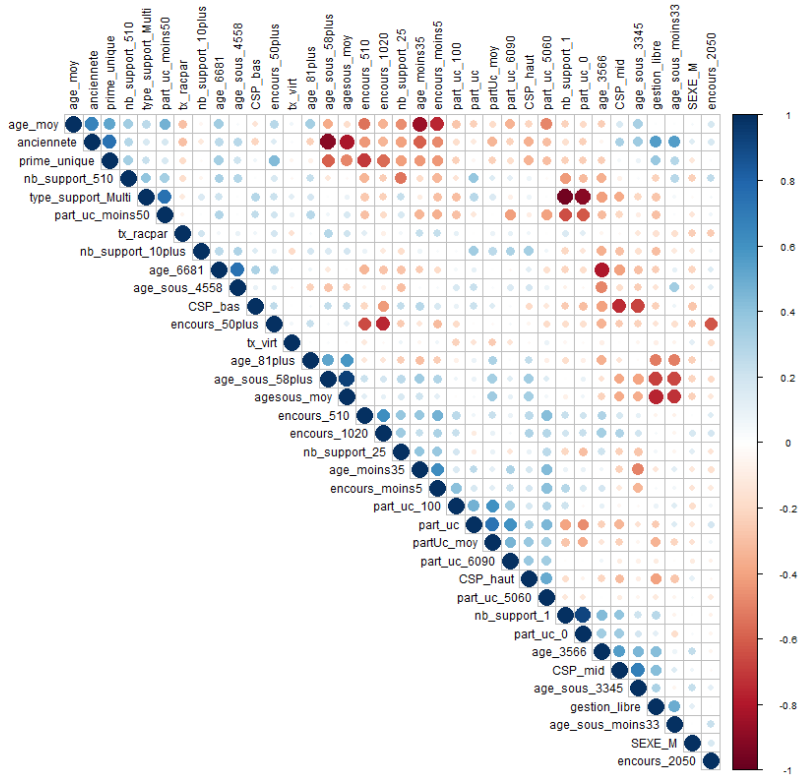


FIGURE 4.7 – Matrice de corrélation de Spearman entre des variables : produit "prod_A-AF"

De manière générale, les catégories et modalités associées à une même variable sont corrélées par construction. Par ailleurs, l'âge à la souscription s'obtient comme la différence entre l'âge et l'ancienneté ; ce qui entraîne une assez forte corrélation entre ces variables. On observe en outre une forte corrélation entre la part d'UC, l'âge à la souscription et l'âge moyen par ancienneté et respectivement les provisions associées aux classes de part d'UC, d'âge à la souscription et d'âge par ancienneté. Par conséquent, les variables classes d"âge à la souscription ainsi que la part d'UC, l'âge à la souscription et l'âge moyen sont

retirées de la base de données. En outre, la variable "type_support_multi" qui renseigne sur la proportion de provisions associée aux contrats multi-supports EUR/UC est fortement corrélée par construction aux variables "nbre_suport_1" et "part_uc_0" ; ces dernières sont donc retirées de la base.

- **Corrélation avec les taux de rachats**

Les corrélations entre les taux de rachats totaux et les variables explicatives confirment les résultats de l'analyse descriptive. L'encours faible sur les contrats, la CSP, l'âge et le taux de rachats partiels sont les variables plus corrélées positivement aux taux de rachats totaux sur le produit "prod_A-AF" ; ce qui signifie que plus les proportions de provisions associée aux "petits" contrats, aux assurés "cadres ou chefs d'entreprise" et aux assurés âgés de plus de 58 ans sont élevées, plus les taux de rachats totaux sont importants. Par ailleurs, il existe une assez forte corrélation négative avec l'ancienneté, ce qui est normal car les taux de rachats tendent à s'annuler au delà de 17 ans d'ancienneté.

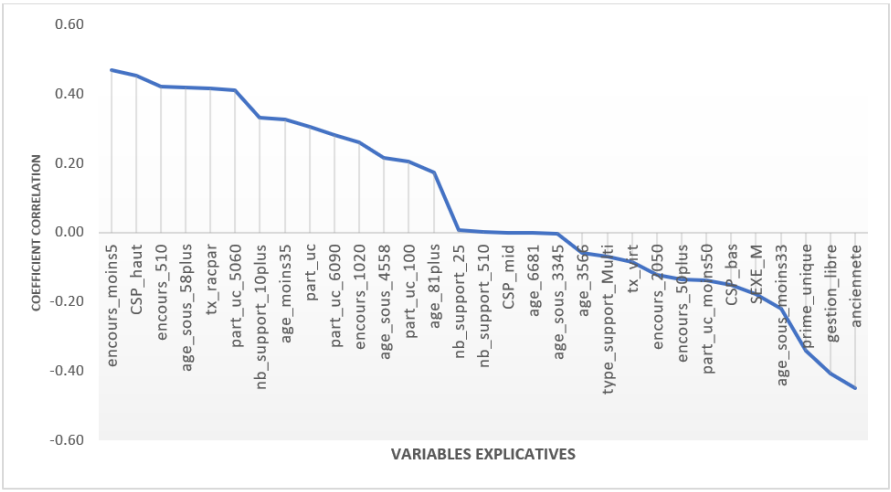


FIGURE 4.8 – Coefficient de corrélation entre le taux de rachats totaux et les variables explicatives

Les corrélations ainsi mises en évidence ne nous permettent pas de cerner les liaisons multiples entre les variables de l'étude. Il importe donc de recourir aux modèles de régression classiques et algorithmes de Machine Learning.

4.2.2.2 Résultats des modèles sur produit "prod_A-AF" : test sur 2020

Les modèles de régression pénalisés ainsi que les algorithmes d'arbre de décision simple, de forêts aléatoires, KNN (K-Nearest Neighbors) et SVM (Support-Vector Machines) sont utilisés pour calibrer les taux de rachats totaux. Ce sont en effet les algorithmes les mieux adaptés aux bases de petite taille. Les années 2015 à 2019 constituent la base d'apprentissage tandis que l'année 2020 sert de base validation.

Comme le montre le tableau suivant, près de la moitié des taux de rachats de la base d'apprentissage sont nuls. Par conséquent, des modèles de classification sont calibrés dans un premier temps afin de définir le statut de rachats pour une ancienneté donnée (statut_rachats=1 si le taux de rachats est non-nul et 0 sinon) ; en particulier, ils permettent de prédire la probabilité qu'il y ait au moins un rachat total à une ancienneté donnée. Par la suite, les modèles de régression sont calibrés sur les taux de rachats non-nuls. La prédiction finale est obtenue par agrégation des meilleurs modèles de classification et de régression.

TABLEAU 4.1 – Proportion de taux de rachats nuls sur le produit "prod_A-AF"

taux de rachats	base apprentissage	base test
0	46.79%	23.07%
>0	53.20%	76.92%

- **Performance des modèles de classification**

L'accuracy score est l'indicateur utilisé pour évaluer et comparer les différents modèles de classification. Il s'agit du taux de bon classement calculé comme suit :

$$score = \frac{Nombre\ d'observations\ bien\ classées}{taille\ de\ l'échantillon}$$

Lorsqu'une procédure d'apprentissage produit un modèle qui fait de bonnes prédictions sur les données d'apprentissage mais se généralise mal, on parle de sur-apprentissage ; la capacité de généralisation d'un modèle se traduit en effet par sa capacité à faire des prédictions correctes sur de nouvelles données qui n'ont pas été utilisées pour la construction dudit modèle. En pratique, le sur-apprentissage est observé lorsque les indicateurs de performance sont significativement différents sur les bases d'apprentissage et de test.

Afin d'éviter le sur-apprentissage, les hyper-paramètres de nos modèles sont optimisés par validation croisée ; en particulier, la validation "Leave-one-out" présentée plus haut est utilisée ; c'est la méthode la plus robuste lorsque la base de données est de petite taille.

Globalement, les modèles présentent d'assez bons scores de classification aussi bien sur la base d'apprentissage que de test. Toutefois, seul l'arbre de décision ne présente pas de sur-apprentissage. En effet, les autres algorithmes ont des scores test supérieurs de 15% à ceux de la base d'apprentissage. Le meilleur modèle de classification est donc l'arbre de décision. Cet algorithme permet de classer correctement 84% de nouvelles observations.

TABLEAU 4.2 – Performance des modèles de classification : produit "prod_A-AF"

Score	Lasso	Ridge	Elastic net	arbre de décision	RF	KNN	SVC
Train	96.79%	98.72%	96.79%	89.75%	100.00%	97.44%	99.36%
Test	80.77%	80.77%	80.77%	84.62%	76.93%	76.93%	76.93%

- **Performance des modèles de régression**

Une fois le meilleur modèle de classification retenu, il convient de calibrer des modèles de régression sur les observations ayant des taux strictement positifs. En moyenne, le taux de rachats totaux dans la base d'apprentissage et de test est respectivement de 1.06% et 0.80%. Pour évaluer nos modèles de régression, les mesures de performance utilisées sont la RMSE et la MAE :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\vec{x}^i) - y^i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f(\vec{x}^i) - y^i|$$

Le meilleur modèle est celui qui minimise ces métriques à la fois sur la base d'apprentissage et de test tout en évitant le sur-apprentissage ; le dernier critère est le plus important car il permet d'évaluer la capacité de généralisation du modèle considéré. Grâce à la fonction "GridSearchCV" de Python à laquelle est appliquée la validation croisée "leave-one-out", les paramètres optimaux de chaque algorithme sont calibrés. Ainsi, pour chaque algorithme, les performances du meilleur modèle sont stockées dans le tableau 4.3 ci-dessous. L'ensemble des modèles présentent des erreurs de prédiction de taux assez importants aussi bien sur la base d'apprentissage que sur la base de test. Par ailleurs, les écarts de performance observés sur les deux bases traduisent la mauvaise capacité de généralisation de nos modèles.

TABLEAU 4.3 – Performance des modèles de régression : produit "prod_A-AF"

MAE							
Base	Lasso	Ridge	Elastic net	arbre de décision	RF	KNN	SVR
Train	0.02	0.0007	0.02	0.13	0.07	0.20	0.09
Test	0.08	0.002	0.27	0.19	0.11	0.30	0.44
RMSE							
Base	Lasso	Ridge	Elastic net	Arbre de décision	RF	KNN	SVR
Train	0.12	0.02	0.15	0.13	0.13	0.45	0.12
Test	0.37	0.40	0.67	0.43	0.26	0.55	0.66

Bien qu'aucun algorithme ne soit très satisfaisant, nous retenons les forêts aléatoires (RF) qui reste le plus performant. Afin de prédire les lois de rachats sur l'année 2020, le meilleur modèle de classification par arbre de décision est utilisé dans un premier temps pour séparer les observations sans rachat total des autres. Ensuite, le meilleur modèle de régression par les forêts aléatoires est utilisé pour prédire les taux de rachats sur les observations retenues par le modèle de classification : il s'agit d'une agrégation de modèles.

Ainsi, la loi de rachats totaux prédite par nos modèles est comparée à la loi de rachats réelle observée sur l'année 2020 (figure 4.9 ci-dessous). Le modèle d'agrégation retenu a énormément du mal à prédire le pic observé à la 13^e année d'ancienneté ; par ailleurs pour des anciennetés au delà de 18 ans, les taux prédits sont nuls. Ces erreurs de prédiction sont normales car la base test présente des taux atypiques aux anciennetés susmentionnées ; en effet, l'exploration de la base de données met en évidence des rachats totaux exceptionnels effectués par 4 contrats à respectivement 22, 23, 25 et 26 ans d'ancienneté. Par ailleurs, deux gros contrats ont également été rachetés totalement à la 13^e année d'ancienneté sur l'année 2020, ce qui justifie le pic plus élevé que ceux des années 2015 à 2019.

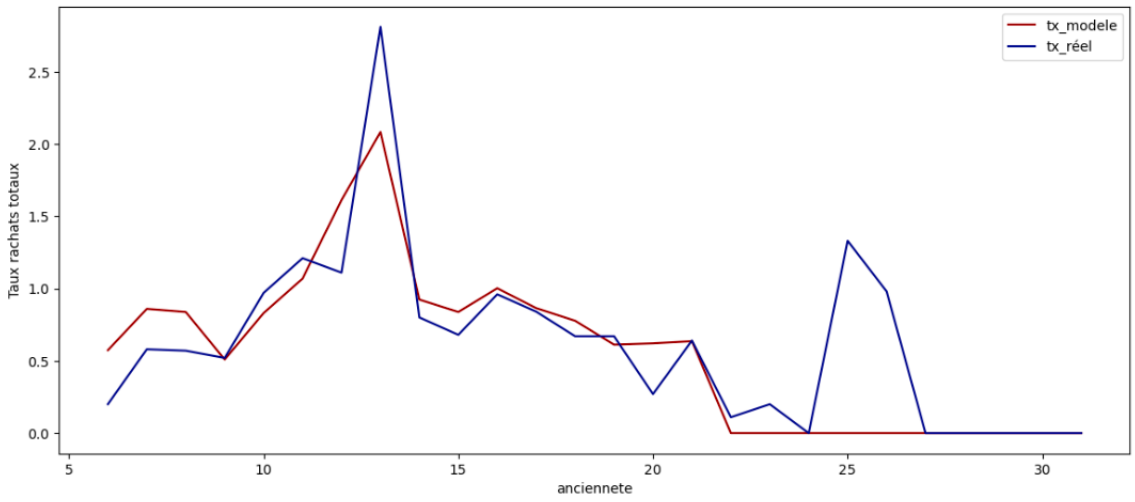


FIGURE 4.9 – Comparaison des lois estimée et réelle sur l'année 2020

Au regard de cette analyse, on conclut que l'année 2020 n'est pas appropriée pour tester nos algorithmes de Machine Learning. Ainsi dans la suite de cette étude, la base test est celle correspondante à l'année 2021.

4.2.2.3 Résultats des modèles sur produit "prod_A-AF" : test sur 2021

Sur l'année 2021, près de 40% des taux de rachats totaux des contrats du produit "prod_A-AF" sont nuls. Ainsi comme précédemment, une classification suivie d'une régression est effectuée. La base d'apprentissage couvre les années 2015 à 2019 tandis que la base de validation correspond à celle de l'année 2021. Pour rappel, l'année 2020 est écartée à cause du bruit qu'elle crée dans les modèles.

- **Performance des modèles de classification**

Aussi bien sur la base de test que d'apprentissage, le taux de bon classement des meilleurs modèles de chaque algorithme de classification excède 84% (Tableau 4.4 ci-dessous). Les modèles sont donc globalement satisfaisants. En particulier, les forêts aléatoires (RF) classent parfaitement les observations à rachats nuls et celles d'au moins un rachat total tout en évitant le sur-apprentissage. C'est donc le meilleur modèle de classification. Les hyper-paramètres optimaux qui ont permis d'avoir une telle performance sont les suivantes :

- **'criterion'= Gini** : cet indicateur permet d'évaluer la pureté des noeuds ; en effet, il mesure la fréquence à laquelle toute observation est mal étiquetée ;
- **'n_tree'= 27** : Il s'agit du nombre optimal d'arbres de la forêt aléatoire ; le modèle par les forêts aléatoires retenu comporte donc 27 arbres individuels ;
- **'max_depth'= 4** : Il s'agit de la profondeur optimal de chaque arbre ;
- **'max_features'= log2** : ce paramètre représente le nombre de variables à tester à chaque noeud des arbres individuels. Pour le modèle retenu, max_features optimal=log2(nbre_variables).

Tableau 4.4 – Performance des modèles de classification : produit "prod_A-AF"

Score	Lasso	Ridge	Elastic net	arbre de décision	RF	KNN	SVC
Train	99.24%	99.24%	99.24%	88.46%	100.00%	99.23%	99.23%
Test	84.62%	84.62%	84.62%	100.00%	99.90%	88.46%	88.46%

Le modèle retenu (RF) permet de quantifier l'importance des variables explicatives dans la classification. Comme le montre la figure 4.10 ci-après, l'ancienneté est la variable la plus importante ; ce qui est en accord avec la littérature, les analyses descriptives et l'analyse des corrélations. Par ailleurs, la proportion de provisions associée aux catégories socio-professionnelles (CSP) de l'assuré, à la périodicité de versement de la prime, aux classes d'âge et sexe de l'assuré, le comportement des assurés en matière de rachats partiels,

la part d'UC sur les contrats sont également les variables les plus importantes. En revanche, le nombre de supports et le type de supports d'investissement ont une contribution presque nulle (figure 4.10).

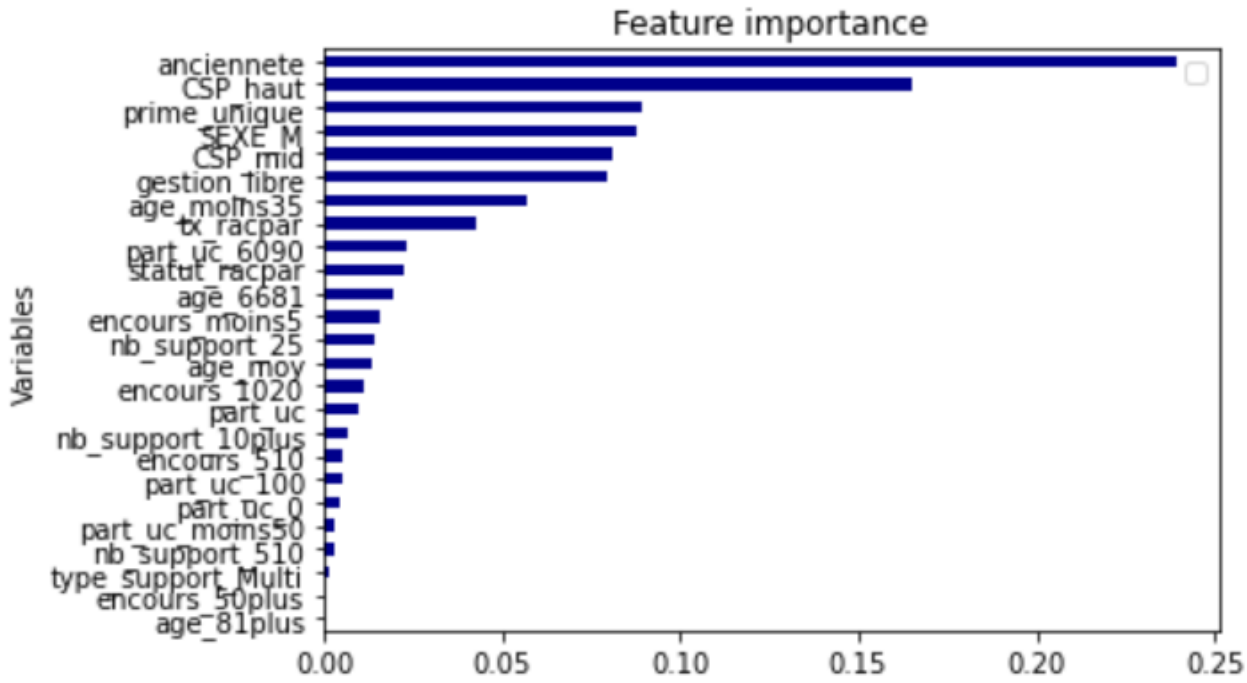


FIGURE 4.10 – Importance des variables dans la classification : produit "prod_A-AF"

Pour savoir comment le modèle optimal de forêts aléatoires utilise ces variables, la méthode de SHAP (SHapley Additive exPlanations) est appliquée. Cette méthode permet de donner le sens de la contribution de chaque variable sur les prédictions au global. Sur la figure 4.11 ci-après, la couleur rouge signifie que la caractéristique a une influence positive de la variable cible tandis que la couleur bleue met en évidence une influence négative. Les enseignements tirés de ce graphe sont les suivants :

- L'ancienneté augmente la probabilité d'avoir des taux de rachats totaux nuls ;
- Plus la proportion de provisions associée aux CSP "csp_haut" (cadres et chefs d'entreprises) et "csp_mid" (employés, professions libérales et intermédiaires) est élevée, plus la probabilité qu'il n'y ait pas de rachats totaux est importante ;
- La proportion de provisions associée aux contrats en gestion libre diminue la probabilité d'avoir au moins un rachat total ;
- Le comportement des assurés en terme de rachats partiels influence positivement le phénomène de rachats totaux ; ce qui est en accord avec nos analyses descriptives. plus les taux de rachats partiels sont élevés, plus la probabilité qu'il y ait au moins un rachat total est importante.

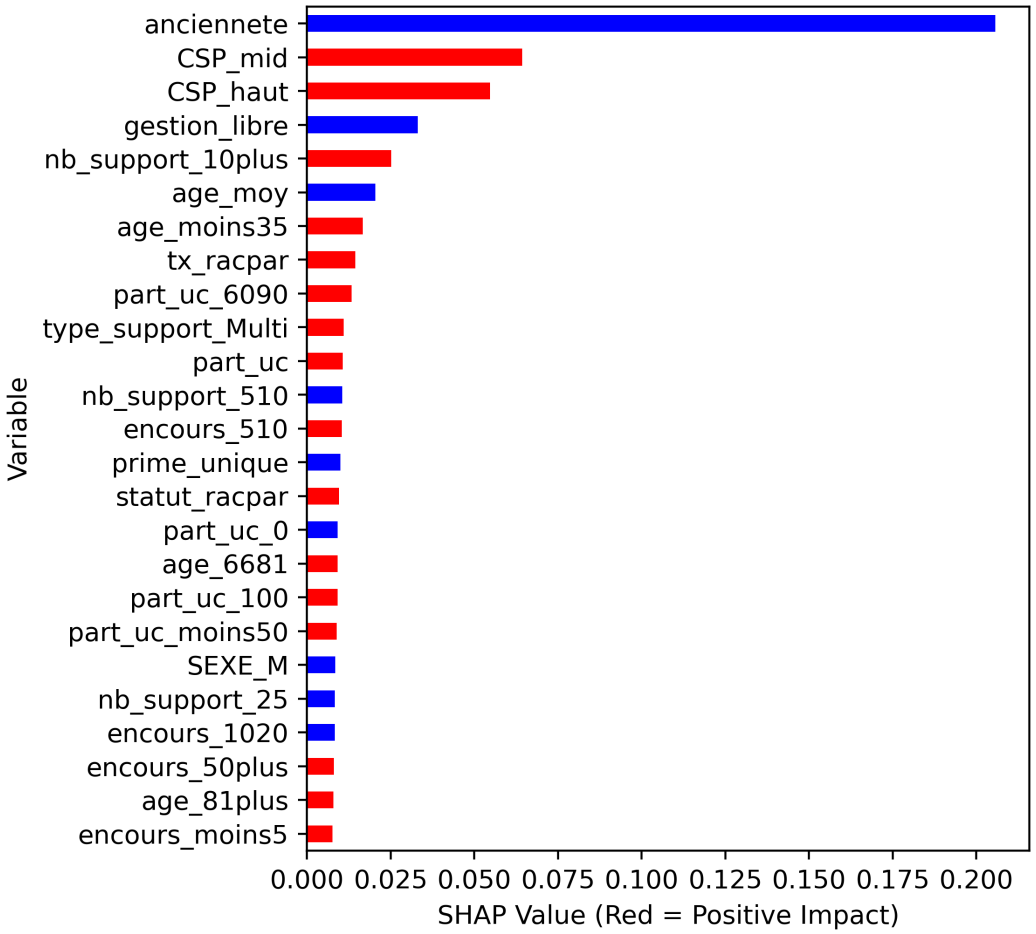


FIGURE 4.11 – Impact des variables dans la classification : produit "prod_A-AF"

- **Performance des modèles de régression : produit "prod_A-AF"**

Sur les observations ayant des taux de rachats totaux strictement positifs sur la période de 2015 à 2019, sont calibrées des modèles de régression ; comme précédemment, le test des modèles se fait sur l'année 2021. Les taux moyens de rachats totaux entre 2015 à 2019 et en 2021 sont respectivement de 0.83% est 1.12%. Le tableau 4.5 stocke les performances des différents modèles après calibration des hyper-paramètres optimaux. Les régressions pénalisées présentent énormément de sur-apprentissage ; les performances sur la base test sont en effet en moyenne 6 fois plus élevées que celles de la base d'entraînement. L'écart est encore plus important sur l'algorithme SVR. Ces modèles ont donc une mauvaise capacité de généralisation car s'ajustent mal à de nouvelles données, ce qui les rend inutilisables. En revanche, l'algorithme des plus proches voisins KNN performe le mieux et permet d'éviter le sur-apprentissage ; il est donc retenu pour effectuer les prédictions des taux de rachats totaux sur l'année 2021. Le nombre optimal de plus proches voisins pour ces prédictions est de 4.

TABLEAU 4.5 – Performance des modèles de régression : produit "prod_A-AF"

MAE						
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	0.14	0.14	0.14	0.11	0.19	0.14
Test	0.67	0.62	0.67	0.23	0.2	0.81
RMSE						
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	0.18	0.18	0.18	0.16	0.26	0.19
Test	0.75	0.69	0.74	0.29	0.24	0.95

Comme précédemment, la méthode de SHAP est utilisée pour mettre en évidence l'intensité et le sens de la contribution des variables à l'évolution des taux de rachats totaux au global. A partir de la figure 4.12 ci-après, les enseignements suivants sont retenus :

- Le mode de versement de la prime est la caractéristique la plus importante dans le modèle KNN ; plus la proportion de provisions associée aux contrats à versement unique est élevé, plus le taux de rachats totaux est important ;
- L'encours influence positivement ou négativement le comportement de rachats selon le volume ; en effet, plus la proportion de provisions associée aux contrats d'encours très élevé (supérieur à 50k) est importante, moins les taux de rachats totaux sont élevés ; l'influence est positive pour les encours faibles. Ces résultats sont cohérents avec nos analyses descriptives ;
- Le volume (en terme de provisions) de contrats en gestion libre influence positivement les taux de rachats de totaux ;
- Pour une ancienneté donnée, si le volume de contrats de plus de 5 supports d'investissement est important, les taux de rachats le sont moins ; en revanche, ces taux augmentent avec le volume de contrats multisupports de moins de 5 supports.
- Les rachats partiels et la part d'UC sont les caractéristiques les moins importantes dans la régression.

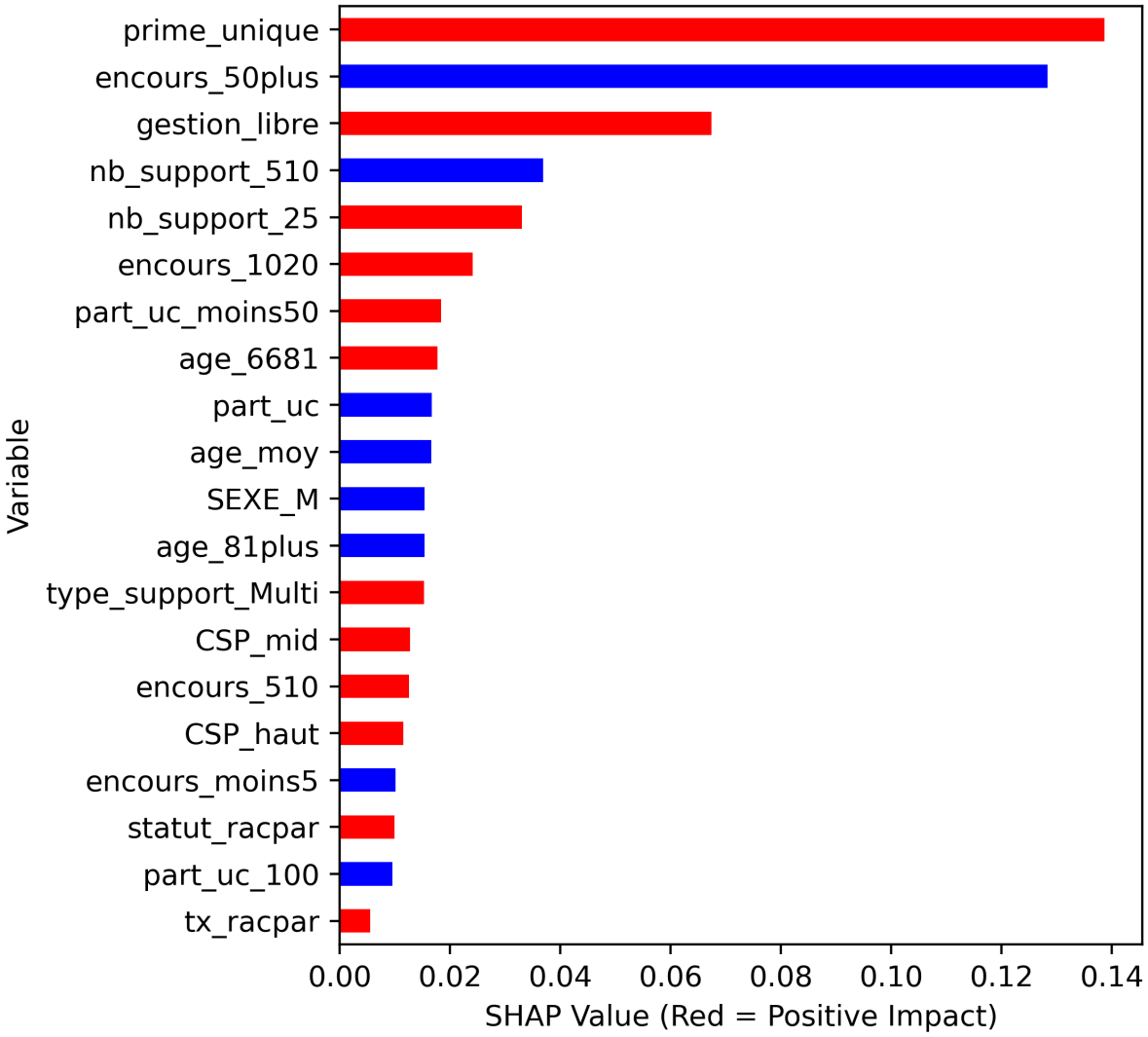


FIGURE 4.12 – Impact des variables dans la régression : produit "prod_A-AF"

- **Comparaison des lois modélisées, réels et de base : produit "prod_A-AF"**

Les meilleurs modèles de classification et de régression sont utilisés pour prédire les lois de rachats totaux sur l'année 2021. Afin d'évaluer nos prédictions, les taux modélisés (issus de nos modèles) sont comparés aux taux réels observés en 2021 (figure 4.13 ci-dessous) ; les deux lois étant relativement proches, nous pouvons conclure que nos algorithmes de Machine Learning modélisent assez bien les comportements de rachats totaux des assurés sur le produit "prod_A-AF". Par ailleurs, il en découle que la loi de base utilisée aujourd'hui dans le modèle interne ne reflète pas la réalité. Les taux de rachats sont en effet fortement surestimés sur les contrats de moins de 11 ans d'ancienneté, ce qui est à l'origine des écarts d"expérience.

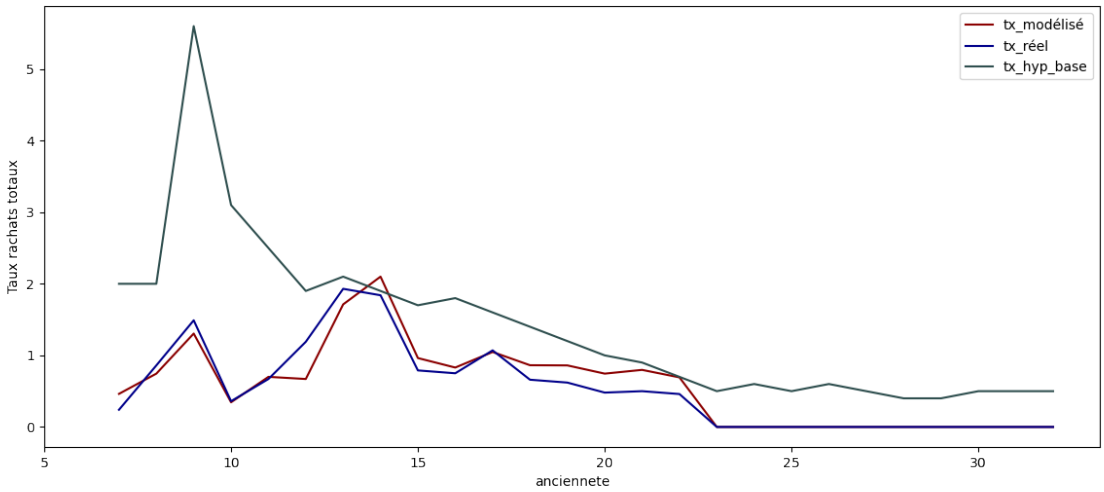


FIGURE 4.13 – Comparaison des lois modélisée, réelle et de base sur l'année 2021 : produit "prod_A-AF"

Les pics de rachats totaux observés à la 8^e et la 14^e année d'ancienneté peuvent être expliqués par nos modèles (figure 4.14 ci-après). Il en ressort que le taux en volume de contrats à versement unique et de contrats d'encours importants a un impact particulièrement important sur les pics de rachats totaux modélisés ; Le waterfall ci-après montre que ces variables permettent d'augmenter les prédictions respectivement de 0.07% et 0.55% par rapport à la moyenne respectivement à la 8^e et 14^e année d'ancienneté ; le nombre de supports d'investissement sur les contrats a également une assez forte contribution sur les prédictions élevées.

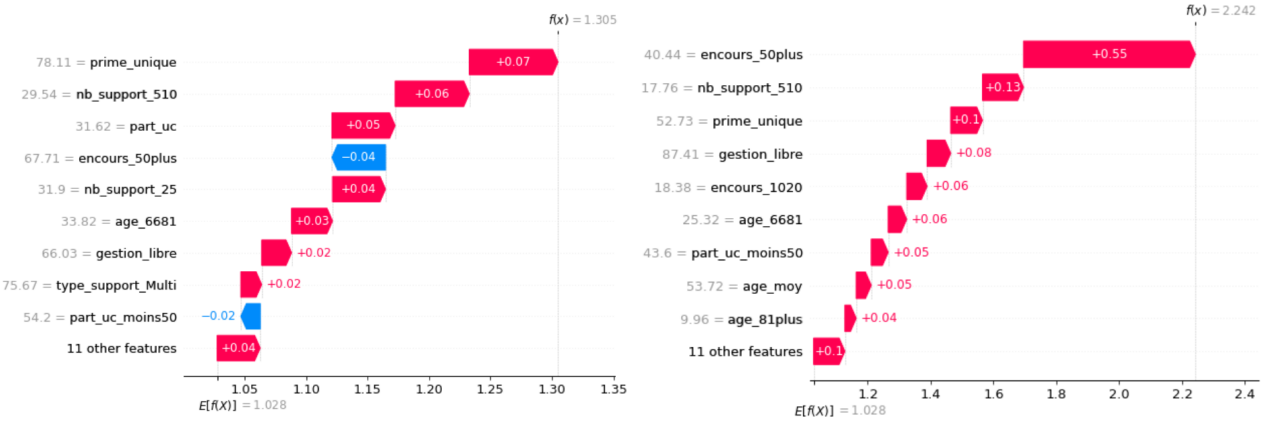


FIGURE 4.14 – Shape waterfall associé à la 8^e (à gauche) et la 14^e année d'ancienneté (à droite)

Étant donné le faible volume de provisions (moins de 5%) associé aux réseaux "CT" et "AG" sur le produit "prod_A", des lois de rachats totaux ne sont pas modélisées sur ces réseaux. La loi prédite sur le réseau "AF" est celle utilisée sur les autres réseaux.

4.2.3 Calibration des lois de rachats totaux sur le produit "prod_B"

Sur le produit "prod_B", le réseau des agents généraux "AG" est largement majoritaire avec 82% du volume de provisions du portefeuille. C'est par conséquent le seul réseau de distribution sur lequel une loi de rachats totaux est modélisée par l'approche Machine Learning. Comme pour le produit "prod_A", l'abréviation "prod_B-AG" est utilisée pour désigner les contrats sur le produit "prod_B" appartenant au réseau de distribution "AG". la base d'entraînement des modèles couvre la période 2015-2019 et la base de test est celle de 2021. Le tableau 4.6 présente la répartition des taux de rachats totaux sur le produit "prod_B-AG" ; globalement, les taux sont plus élevés sur le base "train" avec une moyenne de 1.04% contre 0.71% ; ce qui nous donne de bonnes raisons de penser que les modèles de ML auront tendance à prédire des taux plus élevés que ceux réellement observés en 2021.

TABLEAU 4.6 – Répartition des taux de rachats totaux : produit "prod_B-AG"

tx(%)	min	25%	median	mean	75%	max
Train	0.09	0.63	0.93	1.04	1.33	4.25
Test	0.13	0.47	0.67	0.71	0.91	1.73

- **Performance des modèles de régression : produit "prod_B-AG"**

Sur ce produit, tous les taux de rachats sont strictement positifs ; seuls les modèles de régression sont implémentés dans ce cas. Ils sont testés sur la base de 2021 et les indicateurs de performance des modèles optimaux sont stockés dans le tableau 4.7 ci-après ; il en ressort que les régressions pénalisées Lasso, Ridge et Elastic net présentent une mauvaise capacité de généralisation car les erreurs de prédiction MAE et RMSE sur la base test sont 4 fois plus importantes que celles de la base d'entraînement des modèles ; ce qui les rend inexploitable. Pour ce qui est des modèles KNN et SVR, ils sont très instables car ils performent mieux sur le nouveau jeu de données test que sur les données utilisées pour les construire. Finalement, le modèle par les forêts aléatoires est le modèle de régression retenu pour la prédiction des lois de rachats totaux sur le produit "prod_B-AG". Les hyper-paramètres optimaux retenus pour ce modèle sont les suivants : "max_depth"=4, "max_features"='sqrt', "n_tree"=12.

TABLEAU 4.7 – Performance des modèles de régression : produit "prod_B-AG"

MAE						
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	0.14	0.18	0.15	0.18	0.77	0.59
Test	0.41	0.52	0.4	0.21	0.2	0.28
RMSE						
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	0.3	0.25	0.22	0.27	0.88	0.9
Test	0.64	0.7	0.63	0.31	0.45	0.53

A partir de ce modèle, plusieurs enseignements sont mis en évidence au regard du sens et l'importance des contributions des variables à la régression (figure 4.15 ci-dessous et figure 4.44 en annexe B) :

- l'ancienneté reste la variable la plus importante pour modéliser les comportements des assurés en matière de rachats totaux sur le produit "prod_B-AG". L'influence de cette variable est négative ; les taux de rachats diminuent donc avec l'ancienneté. Cette conclusion est cohérente avec l'analyse descriptive ;
- La proportion de provisions associée aux contrats dont les assurés sont des employés, professions libérales et intermédiaires ("csp_mid") impacte négativement les taux de rachats ; plus c'est élevé et moins il y a de rachats totaux ;
- L'âge moyen des assurés est également une caractéristique importante. plus les assurés sont vieux, plus les taux de rachats sont faibles ; ils sont en effet moins confrontés aux besoins de liquidité que les jeunes ;
- le volume de "petits" contrats (d'encours faible) a un impact positif sur les taux de rachats ; Plus les "petits" contrats sont représentés (en volume), plus les rachats sont importants. Il est donc important de prendre en compte la taille des contrats dans la calibration des lois de rachats.
- La part d'UC a une influence moindre dans le modèle de régression.

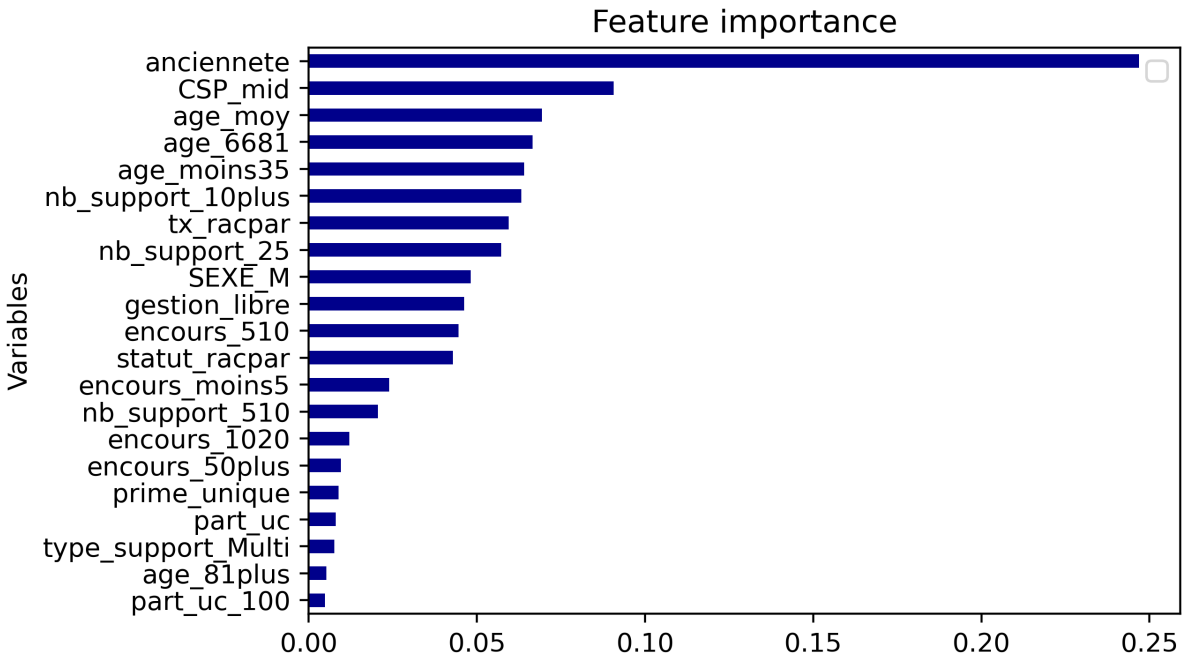


FIGURE 4.15 – Impact des variables dans la régression : produit "prod_B-AG"

- **Comparaison des lois modélisées, réels et de base : produit "prod_B-AG"**

Le meilleur modèle de régression par les forêts aléatoires retenu est donc utilisé pour prédire les taux de rachats sur le produit "prod_B-AG". La figure 4.28 permet de comparer la loi de rachats modélisée sur 2021 aux lois réels et de base ; la loi de base représente en effet la loi de rachats utilisée en "inputs" dans le modèle interne.

Les craintes émis (en début de section) sur la capacité de prédiction de nos modèles sont confirmées. En effet, on peut remarquer une légère surestimation des taux réels par le meilleur modèle. La loi prédite reflète toutefois mieux la réalité que la loi de base ; ainsi, cette loi devrait par conséquent nous permettre de réduire les écarts d'expérience sur le produit "prod_B-AG". Remarquons que le pic de rachats totaux est tardif sur ce produit ; il est observé à la 11^e année d'ancienneté plutôt qu'à la 8^e année. La méthode SHAP permet de mettre en évidence la contribution de chaque caractéristique à ce niveau élevé de rachats totaux.

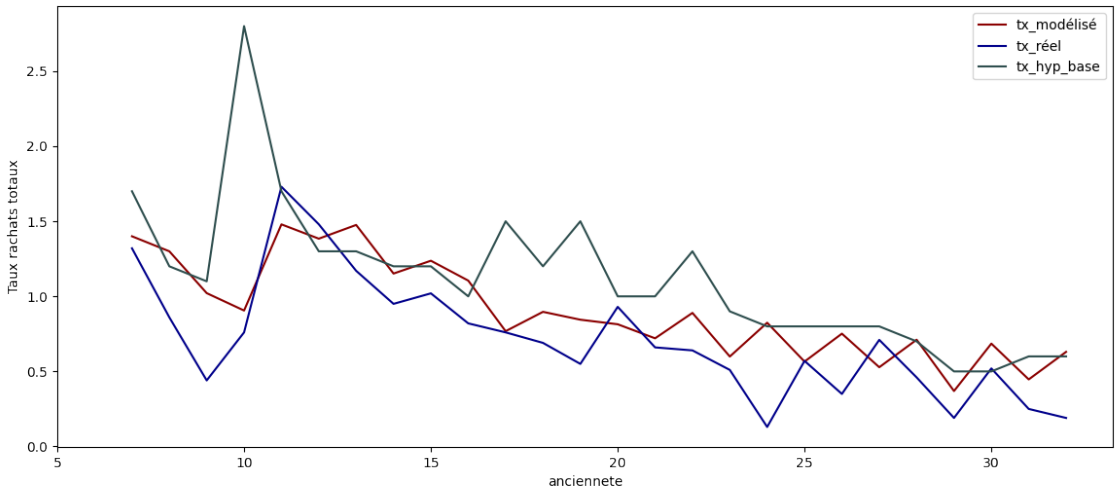


FIGURE 4.16 – Comparaison des lois modélisée, réelle et de base sur l'année 2021 : produit "prod_B-AG"

Le waterfall ci-après montre que l'ancienneté permet d'augmenter le taux de rachats totaux modélisé de 0.26% par rapport à la moyenne ; la contribution de l'âge moyen est de 0.14%. Le taux en volume de contrats pour lesquels les assurés sont jeunes de moins de 35 ans contribue très faiblement au pic de rachats observé. Il en ait de même pour les contrats de faible encours. En revanche, le nombre de supports sur les contrats et le sexe des assurés viennent atténuer légèrement ces effets en réduisant le taux modélisé. Toutes ces contributions ne sont pas captées par l'approche actuelle de calibration des lois de rachats.

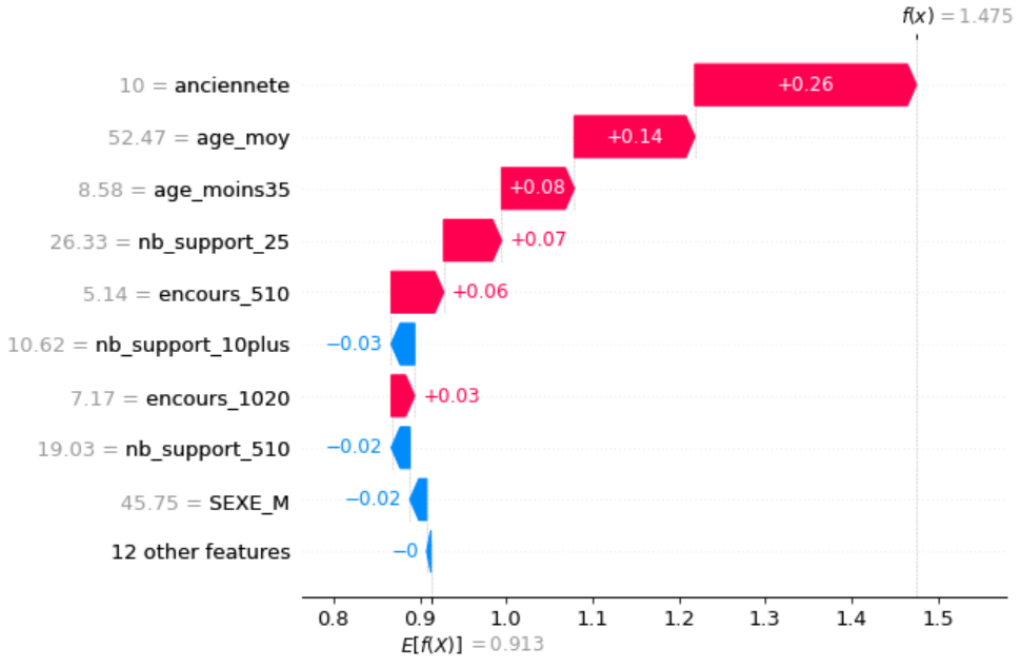


FIGURE 4.17 – Shape waterfall associé à la 6^{e} année d'ancienneté

4.2.4 Calibration des lois de rachats totaux sur le produit "prod_C"

Comme sur le produit "prod_A", le réseau "AF" est largement majoritaire sur le produit "prod_C" avec près de 96% du volume de provisions du portefeuille. Comme le montre le tableau 4.8, un peu plus de 60% des taux de rachats sur la période de 2015 à 2019 (base d'apprentissage) sont nuls contrairement aux taux de 2021 qui sont en grande majorité strictement positifs (près de 89%). Cet important déséquilibre est à l'origine des performances très médiocres de nos algorithmes. Avec des scores inférieurs à 49%, nos modèles sont pires qu'une affectation aléatoire. L'approche de calibration des lois de rachats totaux par les algorithmes de Machine Learning n'est donc pas appropriée dans le cas du produit "prod_C". Les lois de base sont donc maintenues sur ce produit.

TABLEAU 4.8 – Proportion de taux de rachats nuls sur le produit "prod_C"

taux de rachats	base apprentissage	base test
0	63.84%	11.53%
>0	36.16%	88.46%

4.2.5 Évaluation des écarts d'expérience avec les taux modélisés ; 2021

Les nouvelles lois de rachats totaux calibrées par les méthodes de Machine Learning sur les produits "prod_A" et "prod_B" sont utilisées dans le modèle interne afin d'évaluer leur impact sur les écarts d'expérience. Sur le produit "prod_A" et "prod_B", les nouvelles lois permettent de réduire les écarts d'expérience sur les rachats respectivement de 75% et 45%. En revanche, sur les provisions de clôture, les écarts ne sont réduits que respectivement de 10% et 33%, ce qui est du au fait que sur 2021 ce sont les arbitrages qui contribuent majoritairement aux écarts observés sur les provisions. En effet, l'étude réalisée sur les arbitrages (mémoire (Miralles 2021)) n'a pas été prise en compte lors de la calibration des lois d'arbitrages. Précisons qu'en 2021, les écarts d'expérience sur les rachats sont moins importants que ceux observés en 2020 grâce à la mise à jour des hypothèses de base.

TABLEAU 4.9 – Évaluation des écarts d'expérience avec les nouvelles lois (en millions €)

	prod_A			prod_B		
CF	VIPR	R4 old	R4 new	VIPR	R4 old	R4 new
Arbitrages nets	-7.81	0.001	0.001	-1.73	0.0003	0.0003
Rachats	-2.07	-3.14	-2.33	-1.09	-1.83	-1.50
PM cloture	66.12	72.66	72.00	39.75	40.87	40.49
	Écarts d'expérience par poste					
	Écarts old	Écart new	Écarts réduits	Écarts old	Écart new	Écarts réduits
Arbitrages nets	-7.81	-7.82	0.003	-1.73	-1.73	-0.00003
Rachats	1.08	0.26	0.81	0.74	0.41	0.33
PM cloture	-6.53	-5.87	-0.66	-1.11	-0.74	-0.37

4.3 Nouvelle loi de rachats totaux sur le produit "prod_A-AF"

Les écarts d'expérience observés sur les rachats remettent en question les taux de rachats en montant utilisés dans le modèle interne. Par ailleurs, les résultats des analyses descriptives et des modèles calibrés montrent que les taux de rachats totaux en montant des "petits" contrats sont nettement supérieurs à ceux des "gros" contrats. Il en est de même pour les taux de rachats totaux en nombre (figure 4.18 ci dessous). Par conséquent, ne pas tenir compte de la taille des contrats pour calibrer les lois de rachats minimise les sorties des "petits" contrats ; ce qui entraîne la surestimation des coûts unitaires sur ces contrats et donc du résultat technique de l'assureur. D'où la nécessité de tester une approche alternative en considérant les lois de rachats en nombre calibrés par segment de contrats. Dans le modèle interne, ces nouvelles lois devraient être appliquées au nombre total de contrats et non plus à l'encours sur les contrats.

De façon pratique, la nouvelle approche consiste à estimer au moyen des algorithmes de Machine Learning, la proportion de contrats qui pourraient être totalement rachetés séparément sur les "petits", "moyens" et "gros" contrats. La segmentation des contrats se fait par avis d'experts de la manière suivante : les "gros" contrats sont ceux dont l'encours est supérieur au $9^{ème}$ décile de la distribution des encours sur le portefeuille, les "petits" contrats sont ceux dont l'encours est inférieur au $2^{ème}$ décile.

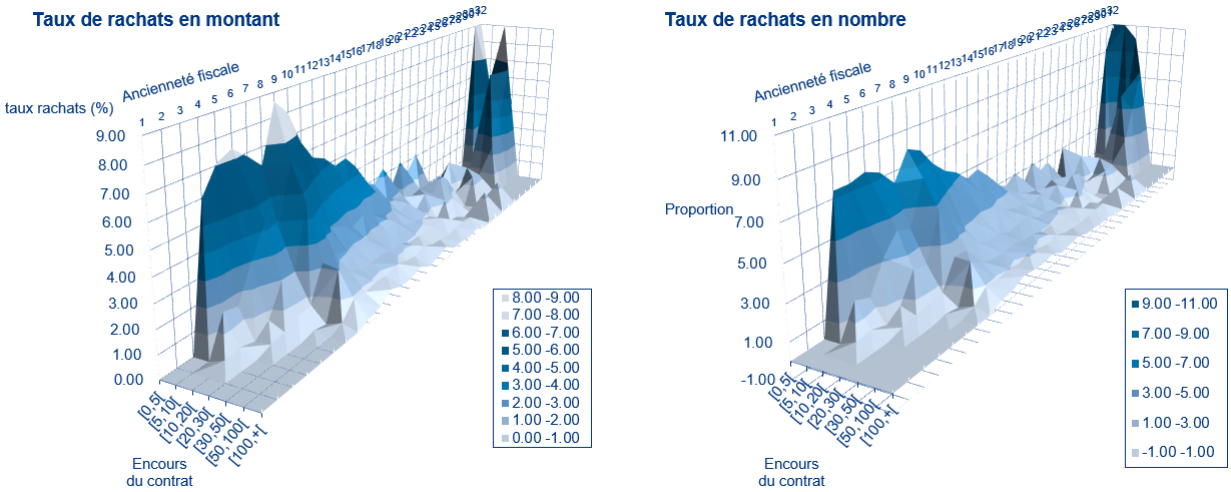


FIGURE 4.18 – Taux et montant de rachats totaux par ancienneté et par classe d'encours sur produit "prod_A-AF"

Cette approche est testée uniquement sur le produit "prod_A" et le réseau de distribution des salariés Allianz France "AF". Sur ce portefeuille, les "petits" contrats ont un encours inférieur à 7300 € et représentent 40% des rachats totaux en nombre, tandis que les "gros" contrats regroupent les contrats ayant un encours supérieur à 54 000 € et ne représentent que 6% des rachats en nombre.

4.3.1 Modélisation de loi de rachats en nombre : cas des "gros" contrats

Les bases de données utilisées dans cette étude sont construites de manière identique à celles de la modélisation précédente. La seule différence réside dans le calcul des variables explicatives ; dans ce cas, elles correspondent à la proportion de contrats en nombre par modalité de variables. Les bases d'entraînement et de test des modèles couvrent également les périodes 2015-2019 et 2021 respectivement. La répartition des taux de rachats en nombre est similaire à celle des taux en volume. En moyenne, 0.79% de contrats effectuent un rachat total ; un peu plus de la moitié des taux sont en effet nuls. Par conséquent, des modèles de classification et de régression sont calibrés.

Les modèles de classification implémentés sont les régressions logistiques pénalisées, les algorithmes KNN, SVC et les forêts aléatoires. Une fois optimisée, ces modèles performent très bien sur la base des "gros" contrats ; ils permettent en effet de séparer parfaitement les observations sans rachat total des autres. En revanche, les modèles de régression présentent de moins bonnes performances comme le montre le tableau 4.10. Toutefois, le modèle de régression par les forêts aléatoires est celui qui minimise les erreurs de prédiction MAE et RMSE aussi bien sur la base d'entraînement que de test.

TABLEAU 4.10 – Performance des modèles de régression sur les "gros" contrats

	RMSE					
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	0.3	0.29	0.29	0.24	0.18	0.24
Test	0.34	0.32	0.34	0.28	0.41	0.388

Finalement, les meilleurs modèles de classification et de régression permettent de prédire la loi de rachats totaux en nombre des "gros" contrats sur l'année 2021. Comme le montre la figure 4.19, la loi modélisée ne reflète pas la réalité pour les contrats de moins de 10 ans d'ancienneté ; en effet, le modèle présente des difficultés à prédire les pics de taux de rachats. Sur les autres anciennetés, les taux prédits sont surestimés mais gardent toutefois la même dynamique que les taux réels.

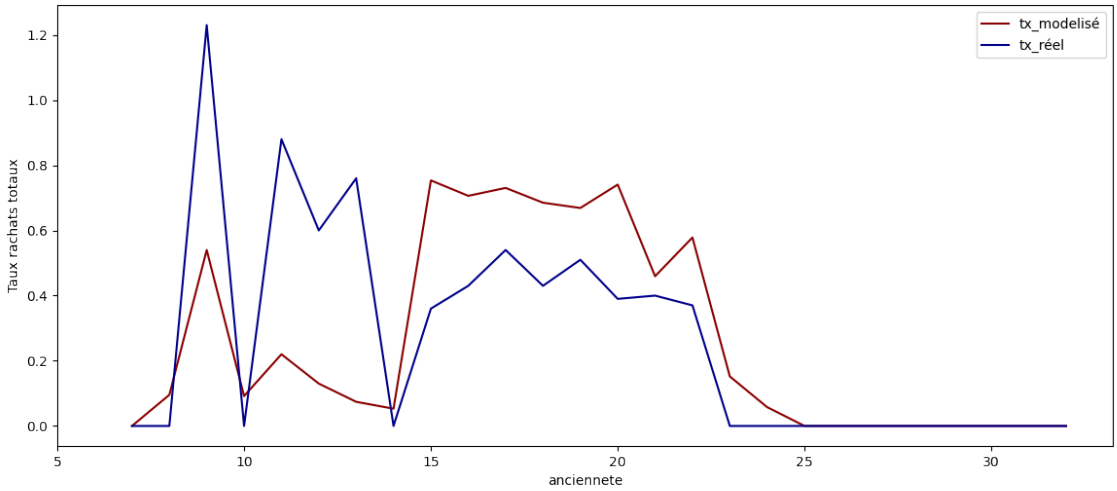


FIGURE 4.19 – Comparaison des lois modélisée et réelle sur l'année 2021 : cas des "gros" contrats

4.3.2 Modélisation de loi de rachats en nombre : cas des "petits" contrats

Sur les "petits" contrats, le taux moyen de rachats totaux en nombre est de 3.5% soit 5 fois plus élevé que sur les "gros" contrats. Globalement, les performances de nos modèles de régression sur l'année 2021 sont peu satisfaisantes (voir tableau 4.11 ci-après). Ces contre-performances peuvent être attribuées à la taille de la base d'entraînement des modèles ; les "petits" contrats ne représentent en effet que 10% du portefeuille. Le modèle de régression linéaire pénalisée RIDGE est retenu pour les prédictions.

TABLEAU 4.11 – Performance des modèles de régression sur les "petits" contrats

	RMSE					
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	1.55	1.34	1.63	1.4	1.62	1.47
Test	3.09	2.29	2.55	2.51	2.5	2.48

La loi prédite par ce modèle est globalement surestimée ; les taux de rachats en nombre prédits sur les "petits" contrats sont en effet supérieurs aux taux réels de 2021. Une fois testée dans le modèle interne, cette loi pourrait toutefois permettre de réduire les écarts d'expérience sur les provisions.

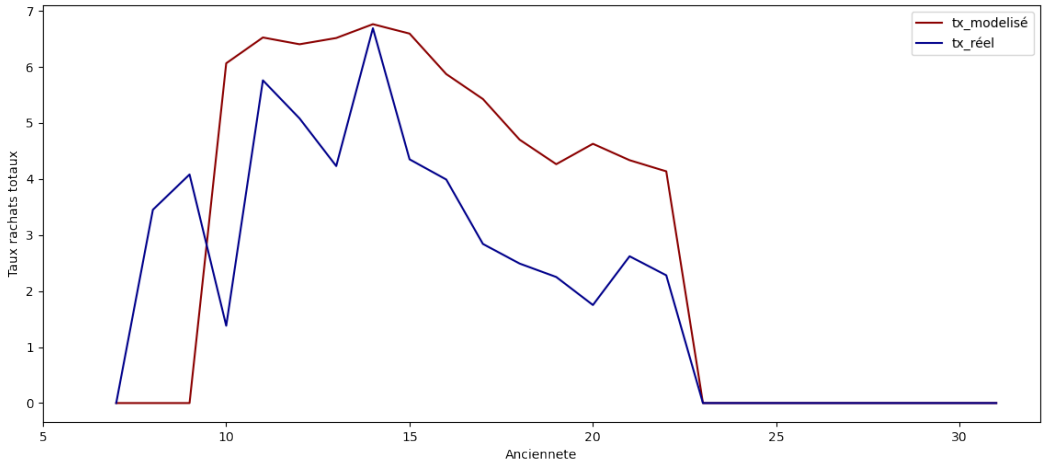


FIGURE 4.20 – Comparaison des lois modélisée et réelle sur l'année 2021 : cas des "petits" contrats

4.3.3 Modélisation de loi de rachats en nombre : cas des contrats "moyens"

Sur les contrats "moyens" (les autres contrats), le taux moyen de rachats totaux en nombre est de 1.26%. Une fois optimisés, les modèles de régression pénalisés sont peu satisfaisants ; en effet sur la base test, les erreurs de prédiction sont au moins deux fois plus élevées que sur la base d'apprentissage. Un tel sur-apprentissage rend ces modèles inexploitables. Seul le modèle SVR minimise l'erreur de prédiction aussi bien sur la base d'entraînement que de test ; c'est donc celui retenu pour la prédiction de la loi de rachats en nombre sur les contrats "moyens".

TABLEAU 4.12 – Performance des modèles de régression sur les autres contrats

	RMSE					
Base	Lasso	Ridge	Elastic net	RF	KNN	SVR
Train	1.15	1.22	0.57	1.41	0.7	0.66
Test	4.22	2.01	3.79	1.06	2.27	1.05

Comme le montre la figure 4.29, La loi prédite reflète mieux la réalité comparativement aux lois modélisées sur les "petits" et les "gros" contrats ; ce qui se justifie par la taille des données d'entraînement des modèles qui représentent 80% du portefeuille considéré. Toutefois, le modèle SVR retenu présente également des difficultés à modéliser les pics de rachats totaux (Figure 4.29 ci-après).

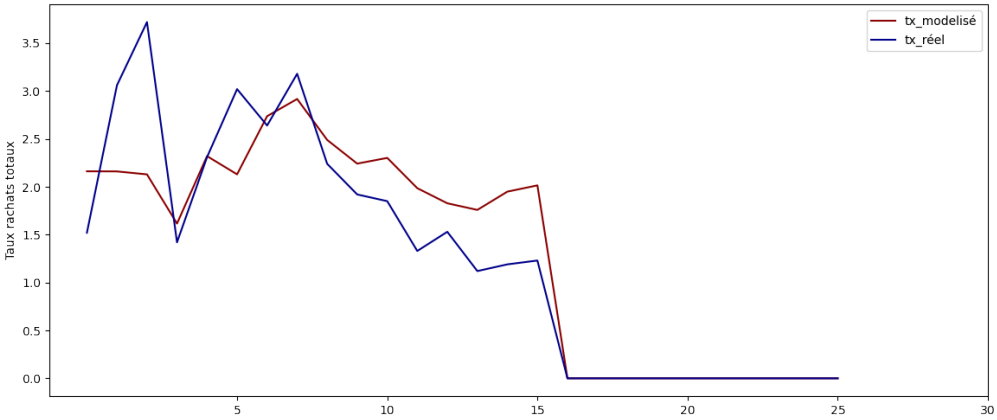


FIGURE 4.21 – Comparaison des lois modélisée et réelle sur l'année 2021 : cas des contrats "moyens"

Finalement, les algorithmes de Machine Learning seraient moins adaptés à la modélisation des lois de rachats en nombre sur les "petits" et les "gros" contrats. En revanche, ces algorithmes performent mieux lorsque le portefeuille est constitué de contrats "moyens". Ces écarts de performances se justifient par :

- La faible variabilité des caractéristiques des contrats et des assurés dans les bases d'apprentissage des modèles sur les "petits" et "gros" contrats ;
- La taille des différents segments de contrats ; en effet, les "petits" et "gros" contrats ne représentent que 10% du portefeuille (taille calculée en nombre de contrats).

CONCLUSION GÉNÉRALE

La contrainte principale des compagnies d'assurance est de respecter leurs engagements vis-à-vis des assurés. Les nouveaux référentiels comptables (IFRS 17) et prudentiels (Solvabilité II) obligent les acteurs du secteur d'assurance à développer modèles mathématiques dits modèles actuariels pour valoriser les risques futurs inhérents à leur activité. Toutefois, il découle de ces modèles, des écarts d'expérience entre les comportements modélisés et réels des assurés qui proviennent de sources diverses notamment des erreurs de calibration des hypothèses de projection de cash-flows. Ce qui pourrait mettre en péril l'équilibre financier et donc conduire à la faillite de l'assureur si les écarts sont importants. Dans cette étude, il était question de réduire les écarts d'expérience sur les provisions (calculées sous solvabilité 2) en utilisant une approche Machine Learning.

A cet effet, une maquette Excel est développée afin d'automatiser le calcul des écarts d'expérience ainsi que le choix des postes à forte contribution. Cette étude porte sur les écarts observés entre 2017 et 2020 sur 288 produits ; à cause de la très forte contribution (55%) des rachats partiels et totaux aux écarts observés sur les provisions, l'étude s'est limitée à réduire les écarts d'expérience uniquement sur les actes de rachats. Pour ce faire, des groupes homogènes de produits sont construits grâce aux méthodes de clustering (K-médoïdes et DBSCAN) afin de calibrer des taux de chocs moyens à appliquer aux hypothèses de projection des rachats totaux, de manière à augmenter les rachats si le modèle les sous-estime et vice-versa. Cette approche a permis de réduire les écarts d'expérience sur les rachats et les provisions respectivement de l'ordre de 85% et 50%. Ces résultats ont permis de mettre en évidence les limites des mailles de construction des lois de rachats.

Afin de proposer une approche Machine Learning pour la calibration de nouvelles lois de rachats totaux, deux produits sur lesquels les écarts d'expérience sont les plus importants sont sélectionnés d'un point de vue statistique et non métier. Sur ces produits, les caractéristiques des assurés et des contrats sont analysées

sur la période de 2015 à 2021 afin de prédire leurs comportements en termes de rachats totaux. Les taux de rachats totaux en montant et en nombre sont calibrés par ancienneté et par réseau de distribution. L'ancienneté, l'encours sur le contrat, la CSP et le mode de gestion des contrats sont les caractéristiques les plus importantes dans nos modèles de régression des taux. Contrairement à l'année 2020 durant laquelle des comportements atypiques des assurés sont observés, nos modèles de Machine Learning approchent mieux la réalité observée en 2021. Une fois testée dans le modèle interne, les nouvelles lois de rachats calibrées permettent de réduire de manière significative les écarts d'expérience sur les rachats et les provisions.

Afin de limiter la surestimation des coûts unitaires liée aux rachats importants en nombre sur les contrats d'encours faible, des lois de rachats en nombre sont modélisées par segment de portefeuille (fonction du volume d'encours). Les algorithmes d'apprentissage supervisé marchent moins bien sur les "petits" et les "gros" contrats à cause de la faible variabilité des caractéristiques des contrats et des assurés dans les bases d'apprentissage des modèles.

Ce mémoire présente un double intérêt. Il permet de mettre en évidence l'apport du Machine Learning pour la réduction des écarts d'expérience sur les provisions d'une part et de quantifier la contribution des caractéristiques des contrats et des assurés au niveau de taux de rachats par ancienneté d'autre part. Par ailleurs, il propose une nouvelle approche de calibration des lois de rachats afin de pallier aux limites de la méthodologie actuelle.

Malgré les résultats assez satisfaisants découlant de cette étude, elle présente toutefois des limites. La première limite est liée au fait que l'étude ne prend pas en compte l'aspect dynamique du phénomène étudié. En effet, tout au long des analyses, nous faisons l'hypothèse selon laquelle les comportements des assurés en matière de rachats totaux ne change pas au cours du temps ; toutefois, cette hypothèse a découlé des analyses descriptives. De plus, nos algorithmes éprouvent des difficultés à modéliser correctement des pics de rachats ; des modèles plus adaptés à cette structure de taux pourraient être utilisés pour améliorer les prédictions. En outre, d'autres méthodes supplémentaires d'interprétabilité de "boites noires" comme les méthodes "LIME" ou "ALE" pourraient être testées en vue de mieux comprendre nos modèles et d'en cerner les limites. Enfin, une perspective de ce travail consisterait à analyser la contribution de la politique de participation aux bénéfices aux écarts observés sur les Provisions.

BIBLIOGRAPHIE

1. *1 Articles L. 331-3, A. 331-3, A. 331-4 et R. 351-2* (n.d.), Code des Assurances.

2. Azencott, C.-A. (2018), *Introduction au Machine Learning*.

3. Breiman, L. (2019), *Random forests. Mach. Learn.*

4. Breiman, L., F. J. H. O. R. A. & Stone, C. J. (1984), 'Classification and regression trees', *Wadsworth International Group* .

5. Chamroukhi, F. (2016), *Algorithme des centres moblies (K-means)*, University of Caen-Normandy, Statistics & Data Science.

6. Courjault-rade, V. (2018), Ballstering : un algorithme de clustering dédié à de grands échantillons, PhD thesis, université de Toulouse.

7. Dalalyan, A. S. (2022), *Cours* 3^e *année ENSAE, Apprentissage statistique*.

8. Ducos, F. (2020), *Droits financier et des assurances*, EURIA.

9. Ester, M. (1996), 'A density-based algorithm for discovering clusters in large spatial databases with noise.', *Kdd.* **34**, 226–231.

10. Gower (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* **4**(27), 859.

11. Han, Kamber M., P. J. (2012), 'Data mining - concepts and techniques', *Amsterdam : Morgan Kaufmann Publishers* **3**.

12. Hastie T., Tibshirani R., F. J. (2009), 'The elements of statistical learning : Data mining, inference, and prediction', *Springer-Verlag* **2**.

13. Hastie Trevor, Tibshirani Robert, J. F. (n.d.), *The Elements of Statistical Learning Data Mining, Inference, and Prediction*.

14. Little, R. & Schluchter, M. (1985), 'Maximum likelihood estimation for mixed continuous and categorical data with missing values', *Biometrika* (72), 497–512.

15. Miralles, R. (2021), *Analyse des grandeurs explicatives des arbitrages des contrats d'assurance vie en mode de gestion libre par méthodes d'apprentissage statistique*, EURIA.

16. Raghunathan, T. e. a. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Surv. Methodol.* (27), 85–96.

17. *Règlement Délégué Article 35* (n.d.), Groupes de risques homogènes d'engagements d'assurance vie, Solvabilité2.

18. TUFFERY (2012), *data Mining et statistique décisionnelle : L'intelligence des données*, Technip.

19. Vapnik V., L. A. (2014), 'Pattern recognition using generalized portrait method. automation and remote control', *Telecommunication Systems and Networks* **3**(24), 774–780.

20. Winter, A. (2008), *Solvabilité 2 : Présentation générale*, ISFA.

NOTE DE SYNTHÈSE

Cette note présente brièvement les travaux réalisés dans le cadre de mon mémoire d'actuariat à l'ENSAE. Il est principalement question d'utiliser des algorithmes de Machine Learning pour tenter de réduire les écarts d'expérience sur les provisions d'un portefeuille fictif d'épargne et retraite individuelle et collective d'Allianz France.

Contexte général

Un contrat d'épargne en assurance vie est un contrat par lequel l'assureur garanti à l'assuré ou au bénéficiaire désigné par l'assuré, le versement d'une prestation, d'un capital ou d'une rente si l'événement garantit par le contrat survient. La contrainte principale des compagnies d'assurance est donc de respecter leurs engagements vis-à-vis des assurés. A cet effet, les nouveaux référentiels comptables prudentiels obligent les acteurs du secteur d'assurance à développer des modèles mathématiques dits modèles actuariels pour valoriser les risques futurs inhérents à leur activité. Toutefois, il découle de ces modèles un risque de sur-provisionnement ou de sous-provisionnement lorsqu'il existe des écarts d'expérience entre les comportements modélisés et réels des assurés ; ce qui pourrait mettre en péril l'équilibre financier de l'assureur.

Le provisionnement en assurance vie se base sur des modèles internes dont le but est d'actualiser le bilan et donc les engagements de l'assureur jusqu'à maturité des contrats du portefeuille. Pour ce faire, des hypothèses de projection servent à faire évoluer l'ensemble des flux de trésorerie dans le temps : il s'agit des hypothèses sur l'évolution du marché financier et la conjoncture économique, les rachat et arbitrages, les coûts et commissions, etc. Finalement, on parle d'écarts d'expérience lorsque la valeur réelle d'un flux ou d'un indicateur (observée l'année N) diffère significativement de la valeur attendue par le modèle de projection (calculée l'année N-1). Ces écarts découlent entre autres des erreurs de calibration des hypothèses définies sur le passif et sur l'actif.

Le but de ce mémoire est de proposer une approche Machine Learning pour la réduction des écarts

d'expérience observés sur les provisions.

Approche Clustering pour la réduction des écarts d'expérience sur PM

Le portefeuille utilisé dans cette approche est constitué de 288 produits souscrits en 2017, 2019 et 2020 sur lesquels les écarts d'expérience sur les provisions sont analysés. L'année 2018 est écartée à cause de la qualité des données peu fiable. Une analyse de mouvements est faite afin de mettre en évidence les postes qui ont une contribution importante aux écarts observés sur les PM. Sur les rachats et arbitrages nets, les contributions sont respectivement de 55% et 3.39% ; à cause de la très forte contribution des rachats partiels et totaux aux écarts observés sur les provisions, l'étude se limite à réduire les écarts d'expérience uniquement sur ce poste. L'approche clustering utilisée consiste à construire des groupes homogènes de produits en fonction de l'intensité des écarts relatifs et de leurs caractéristiques métier ; ceci dans le but de calibrer des taux de chocs moyens à appliquer aux hypothèses de projection des rachats totaux de la manière suivante :

$$\begin{cases} coef\_cor\_ractot\_final = coef\_cor\_ractot\_initial * (1 + taux\_choc\_moyen) \\ taux\_rachat\_total\_final = taux\_rachat\_total\_initial * coef\_cor\_ractot\_final \end{cases}$$

A partir de 20 clusters construits grâce à l'algorithme K-médoïdes, les hypothèses de rachats associées aux produits sont mises à jour dans le modèle interne, de manière à augmenter les rachats si le modèle les sous-estime et vice-versa ; cette approche permet de réduire de manière significative les écarts d'expérience sur les rachats et les provisions respectivement de l'ordre de 85% et 50%.

Tableau 4.13 – Comparaison des flux modélisés avant et après les chocs

Ecarts d'expérience au global (en millions d'€)			
flux	Ecart initial	Ecart après choc	Montants réduits
PM cloture	20.65	11.14	9.51
Rachats	11.42	1.69	9.73
Arbitrages nets	-0.73	-0.97	0.24

Cette approche permet par ailleurs de mettre en évidence les limites de la méthodologie de calibration des lois de rachats en interne. En effet, ces lois sont calibrées à une maille agrégée de produits nommée "groupe rachats" car elle représente des groupes de produits auxquels sont attribués une unique et même loi. Afin de challenger cette méthodologie et ainsi tenter de réduire les écarts d'expérience, ce mémoire

propose une approche machine learning pour la calibration de nouvelles lois de rachats à la maille la plus fine (maille produit).

Approche Machine Learning pour la réduction des écarts d'expérience sur PM

Les produits retenus dans cette approche sont ceux sur lesquels les écarts d'expérience sont les plus importants sur les rachats. Le seuil de choix correspondant au 98^e centile de la loi de Gilbrat (loi log-normale) calibrée sur la distribution des écarts en valeur absolue, permet de retenir le produit d'épargne "prod_A" et de retraite collective "prod_B".

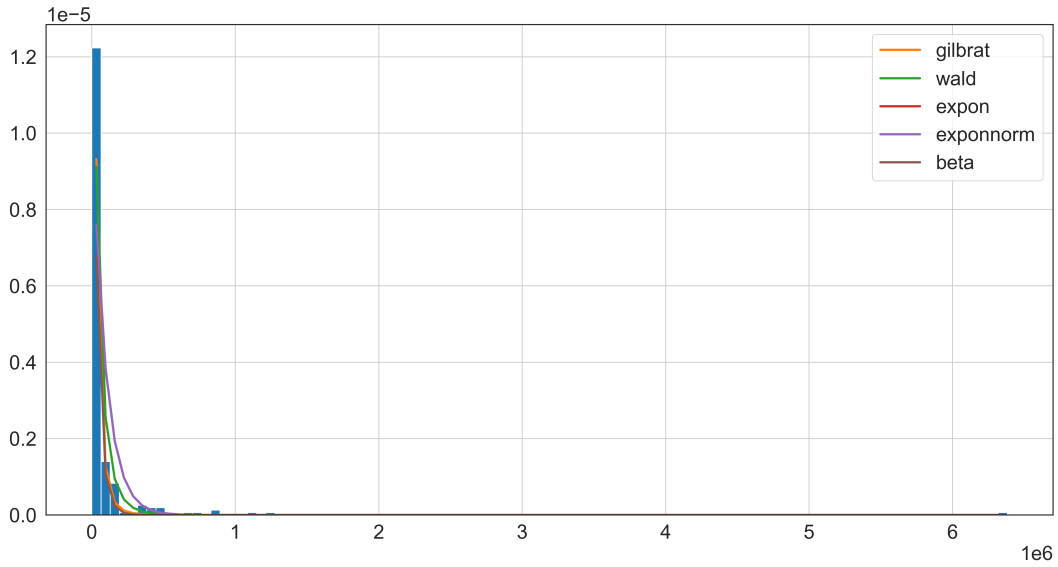


FIGURE 4.22 – Densité des lois qui s'ajustent le mieux aux écarts

Sur ces produits, les taux de rachats totaux en montant sont calibrés grâce aux modèles de régression pénalisées Ridge, Lasso, Elastic Net et aux algorithmes d'apprentissage statistique supervisé KNN, SVM et les forêts aléatoires (RF). Après traitement, la base de données utilisée porte sur près de 450 000 contrats couvrant la période de 2015 à 2021 (soit 3 185 202 lignes) sur lesquels sont renseignées les caractéristiques des assurés (sexe, âge et CSP) et des contrats (ancienneté, périodicité de la prime, mode de gestion, type et nombre de supports, part d'UC), le comportement antérieur des contrats en matière de rachats partiels et d'arbitrages.

- **Nouvelles lois de rachats en montant sur le produit "prod_A"**

Sur le produit "prod_A" et le réseau de distribution de salariés d'Allianz France "AF", les lois de rachats totaux de base du modèle interne sont en moyenne deux fois plus importantes que les taux réels. Par

ailleurs, près de la moitié des taux de rachats sont nuls sur ce produit, ce qui nous contraint à calibrer dans un premier temps des modèles de classification puis des modèles de régression (sur les taux strictement positifs) sur la base couvrant 2015 à 2019. Ces modèles sont testés sur la base de 2021 ; l'année 2020 est en effet écartée à cause des comportements atypiques des assurés qui ont eu un impact significatif sur les performances des modèles de prédiction.

Grâce à l'agrégation des meilleurs modèles de classification (Forêts aléatoires) et de régression (KNN), la loi prédite (ou modélisée) reflète nettement mieux la réalité que la loi de base du modèle interne. La méthode "SHAP" nous révèle que l'ancienneté, l'encours, le mode de gestion des contrats et la périodicité de la prime sont les variables les plus importantes dans la prédiction. En particulier, le taux en volume de contrats à versement unique et de "gros" contrats (contrats d'encours importants) a une contribution importante aux pics de rachats observés sur ce produit, respectivement de 0.07% et 0.55% par rapport à la moyenne respectivement à la 8^e et 14^e année d'ancienneté ; ce qui n'est pas capté par l'approche actuelle de calibration des lois de rachats.

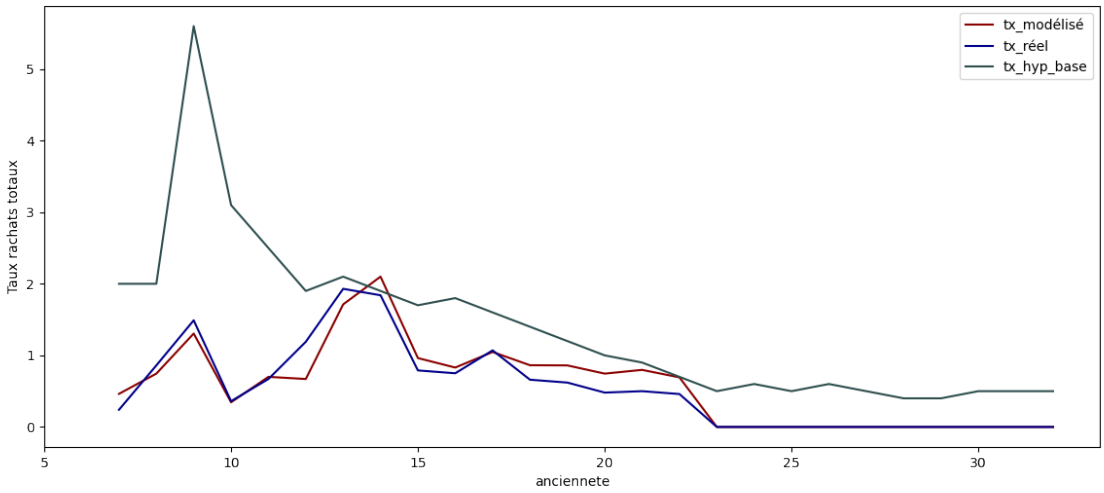


FIGURE 4.23 – Comparaison des lois modélisée, réelle et de base sur l'année 2021 : produit "prod_A-AF"

- **Nouvelles lois de rachats en montant sur le produit "prod_B"**

Sur le produit "prod_B" et le réseau de distribution des agents généraux "AG", le meilleur modèle de régression par les forêts aléatoires révèle que l'ancienneté, la CSP et l'âge moyen des assurés, la part de la prime investie sur les supports en UC sont les caractéristiques structurelles les plus importantes dans la prédiction. Les taux de rachats prédits par ce modèle est sur-estimés mais se rapproche plus des taux réels comparativement à la loi de base. Par ailleurs, l'ancienneté des contrats et l'âge moyen des assurés

contribuent fortement au pic de rachats observé à la 10^e année d'ancienneté ; ils permettent d'augmenter les prédictions respectivement de 0.26% et 0.14% par rapport à la moyenne.

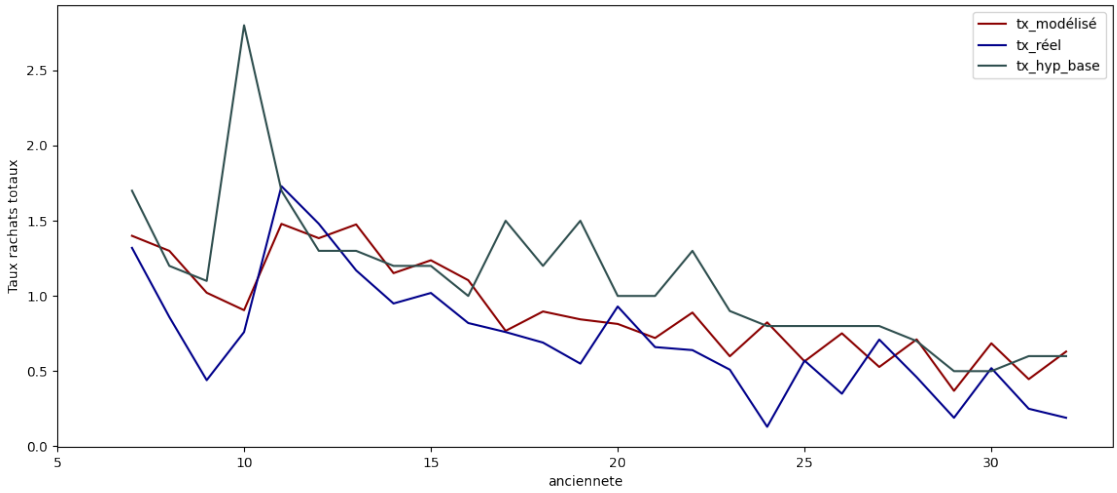


FIGURE 4.24 – Comparaison des lois modélisée, réelle et de base sur l'année 2021 : Réseau "AG"

Une fois mises à jour dans le modèle interne, les nouvelles lois de rachats totaux calibrées par nos algorithmes de Machine Learning permettent de réduire les écarts d'expérience sur les rachats respectivement de 75% et 45% respectivement sur les produits "prod_A" et "prod_B". En revanche, la réduction des écarts sur les provisions est relativement faible à cause de la forte contribution des arbitrages.

- **Nouvelles lois de rachats en nombre**

Suite aux écarts importants constatés sur certains produits et aux conclusions de nos modèles de prédiction, ce mémoire se propose de calibrer de nouvelles lois de rachats par segment de portefeuille ; cette approche a pour but de pallier aux limites de la loi actuelle ; elle minimise en effet les sorties des "petits" contrats ; ce qui entraîne la surestimation des coûts unitaires sur ces contrats et donc du résultat technique de l'assureur. De façon pratique, la nouvelle approche consiste à estimer au moyen des algorithmes de Machine Learning, la proportion de contrats qui pourraient être totalement rachetés séparément sur les "petits", "moyens" et "gros" contrats. Les seuils de segmentation sont définis par avis d'experts. Il découle de nos modèles de prédiction que les algorithmes de Machine Learning seraient moins adaptés à la modélisation des lois de rachats en nombre sur les "petits" et les "gros" contrats. En revanche, ces algorithmes performent mieux lorsque le portefeuille est constitué de contrats "moyens". La taille de ces différents segments de portefeuille ainsi que la faible variabilité des caractéristiques des contrats justifient sans doute ces écarts de performances observés.

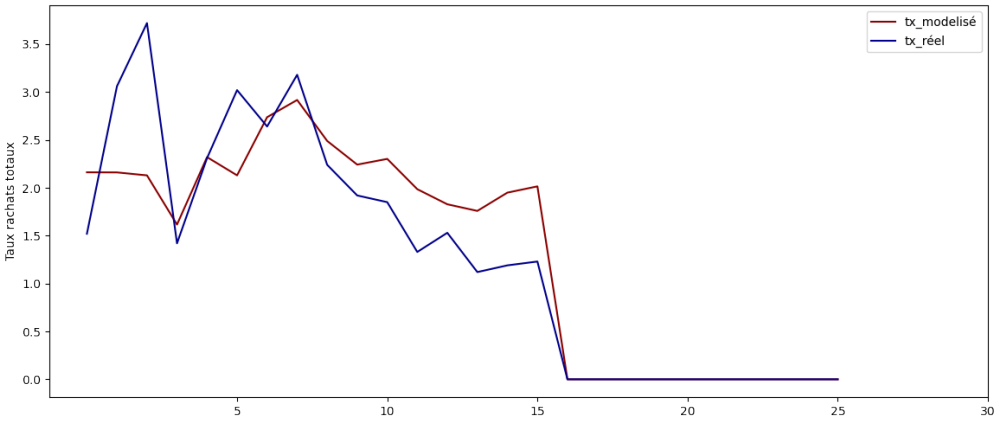


FIGURE 4.25 – Comparaison des lois modélisée et réelle sur l'année 2021 : cas des contrats "moyens"

Limites et perspectives de l'étude

Malgré les résultats assez satisfaisants découlant de cette étude, elle présente toutefois des limites liées à la non-prise en compte de l'aspect dynamique du phénomène étudié. Par ailleurs, nos algorithmes éprouvent des difficultés à modéliser correctement des pics de rachats ; des modèles plus adaptés à cette structure de taux pourraient être utilisés pour améliorer les prédictions. Une autre perspective de ce travail consisterait à analyser la contribution de la politique de participation aux bénéfices aux écarts observés sur les provisions.

EXECUTIVE SUMMARY

This Executive summary briefly presents the work done for my actuarial thesis at ENSAE. It mainly deals with the use of Machine Learning algorithms to try to reduce the experience gaps on the reserves of a fictitious portfolio of individual and group savings and pensions of Allianz France

General context

A life insurance savings contract is a contract by which the insurer guarantees to the insured or to the beneficiary designated by the insured, the payment of a benefit, a capital or an annuity if the event guaranteed by the contract occurs. The main constraint for insurance companies is therefore to respect their commitments to the insured. To this end, the new prudential accounting standards obliged the actors in the insurance sector to develop mathematical models, known as actuarial models, to value the future risks inherent in their activity. However, these models entail a risk of over-provisioning or under-provisioning when there are gaps in experience between the modeled and actual behavior of policyholders, which could jeopardize the insurer's financial equilibrium.

Life insurance provisioning is based on internal models whose purpose is to update the balance sheet and therefore the insurer's commitments until the portfolio contracts reach maturity. For this purpose, projection assumptions are used to develop all cash flows over time : these include assumptions on the development of the financial market and the economic situation, lapses and arbitrages, costs and commissions, etc. Finally, experience gaps occur when the actual value of a cash-flow or indicator (observed in year N) differs significantly from the value expected by the projection model (calculated in year N-1). These deviations result, amongst others, from calibration errors in the assumptions defined for liabilities and assets.

The aim of this thesis is to propose a Machine Learning approach for the reduction of experience gaps observed on life insurance reserves.

Clustering approach for reducing experience gaps on reserves

The portfolio used in this approach consists of 288 products underwritten in 2017, 2019, and 2020 on which experience variances on reserves are analyzed. An analysis of movements is performed in order to highlight the items that have a significant contribution to the experience gaps observed on the reserves. On lapses and net switches, the contributions are respectively 55% and 3.39%. due to the very high contribution of partial and total lapses to the experience gaps observed on reserves, the study was limited to reducing the gaps only on lapses. The clustering approach used is to construct homogeneous groups of products according to the intensity of the relative gaps and their business characteristics ; this is done with the aim of calibrating the average shock rates to be applied to the projection hypotheses for total redemptions in the following way :

$$\begin{cases} coef\_cor\_ractot\_final = coef\_cor\_ractot\_initial * (1 + taux\_choc\_moyen) \\ taux\_rachat\_total\_final = taux\_rachat\_total\_initial * coef\_cor\_ractot\_final \end{cases}$$

From 20 clusters built with the K-medoids algorithm, the lapses assumptions associated to products are updated in the internal model ; this approach reduce significantly the experience gaps on lapses and reserves respectively by about 85% and 50%.

Tableau 4.14 – Comparison of modelled cash-flows before and after shocks

Overall experience gaps (in € million)			
cash-flow	Initial gap	Gap after shock	Reduced amounts
Closing_reserves	20.65	11.14	9.51
Lapses	11.42	1.69	9.73
switches nets	-0.73	-0.97	0.24

This approach also highlights the deficiencies of the methodology used to calibrate the lapses laws internally. Indeed, these laws are calibrated at an aggregated mesh of products called "lapses group" because it includes groups of products to which a unique and same law is attributed. In order to challenge this methodology and thus try to reduce the experience gaps, this thesis proposes a machine learning approach for the calibration of new lapses laws at the finest mesh (product mesh).

Machine Learning approach to reduce experience gaps on reserves

The products used in this approach are those on which the experience gaps are the largest on lapses. The selection threshold corresponding to the 98^e percentile of the Gilbrat distribution (lognormal distribution) calibrated on the distribution of the experience gaps in absolute value on lapses. The products finally selected are the savings product "prod_A" and the collective retirement product "prod_B".

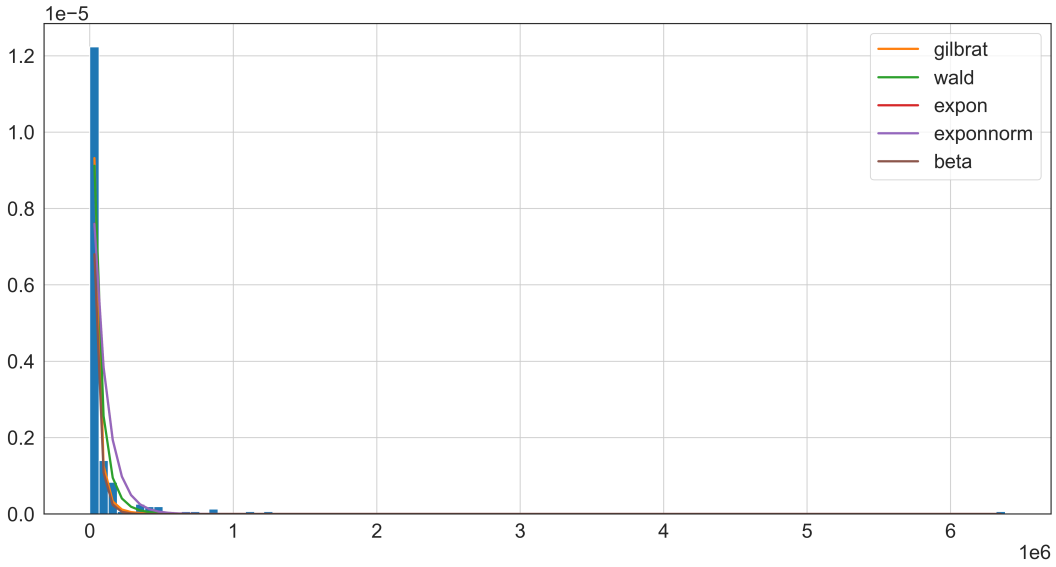


FIGURE 4.26 – Density of laws that best fit the gaps

On these products, the total lapses rates are calibrated using Ridge, Lasso and Elastic Net penalized regression models and KNN, SVM and random forest (RF) supervised statistical learning algorithms. After processing, the database used covers nearly 450,000 contracts covering the period from 2015 to 2021 (i.e. 3,185,202 lines) on which are provided the characteristics of the policyholders (gender, age and social class) and contracts (age, premium frequency, management method of contracts, type and number of investment funds, share of UC), and the past behavior of the contracts in terms of partial lapses and switches.

- **New lapses laws in amount on the product "prod_A"**

On the "prod_A" product and the distribution network of Allianz France employees "AF", the basic total lapses laws of the internal model are on average twice as large as the actual rates. Moreover, almost half of the lapses rates are zero on this product, which forces us to calibrate first classification models and then regression models (on strictly positive rates) on the database covering 2015 to 2019. These models are tested on the basis of 2021 ; the year 2020 is indeed discarded because of the atypical behaviour of policyholders which had a significant impact on the performance of the prediction models.

Through the aggregation of the best classification (Random Forests) and regression (KNN) models, the predicted (or modeled) distribution reflects reality much better than the basic distribution of the internal

model. The "SHAP" method reveals that the seniority, the outstanding amount, the contract management mode and the premium periodicity are the most important variables in the prediction. In particular, the volume rate of single-payment contracts and "large" contracts (contracts with large amounts outstanding) has a significant contribution to the observed lapses peaks on this product, respectively by 0.07% and 0.55% compared to the average at the 8^{th} and 14^{th} year of seniority ; this is not highlighted in the current approach to calibrating the redemption laws.

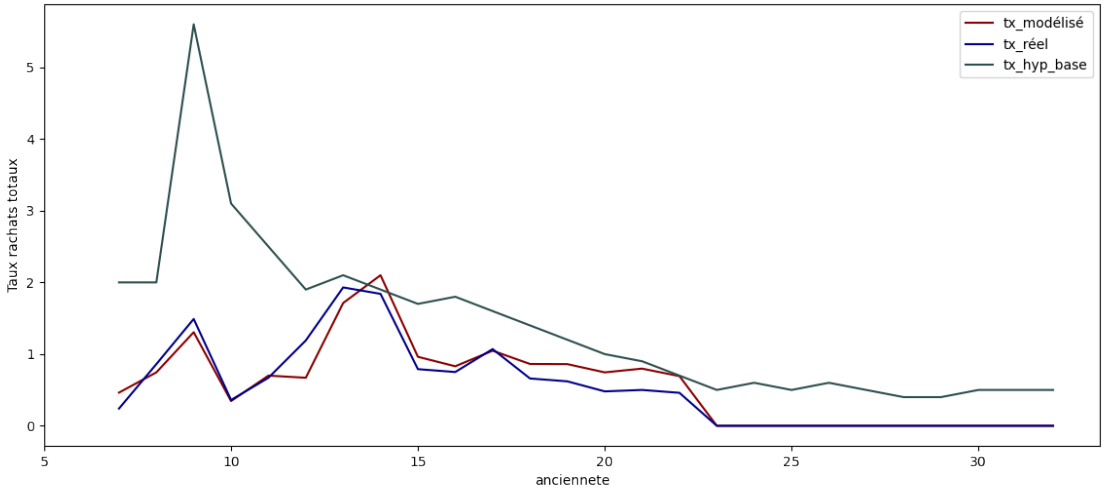


FIGURE 4.27 – Comparison of modeled, actual and base laws for the year 2021 : product "prod_A-AF

- **New laws of lapses in amount on the product "prod_B**

For the product "prod_B" and the distribution network of general agents "AG", the best random forest regression model reveals that the seniority, the CSP and the average age of the policyholders, the share of the premium invested in UC are the most important structural characteristics in the prediction. The law predicted by this model is overestimated but is closer to the actual rates than the base law. Moreover, the length of the contracts and the average age of the insureds contribute strongly to the peak in lapses observed in the 10^{th} year of seniority ; they increase the predictions by 0.26% and 0.14% respectively compared to the average.

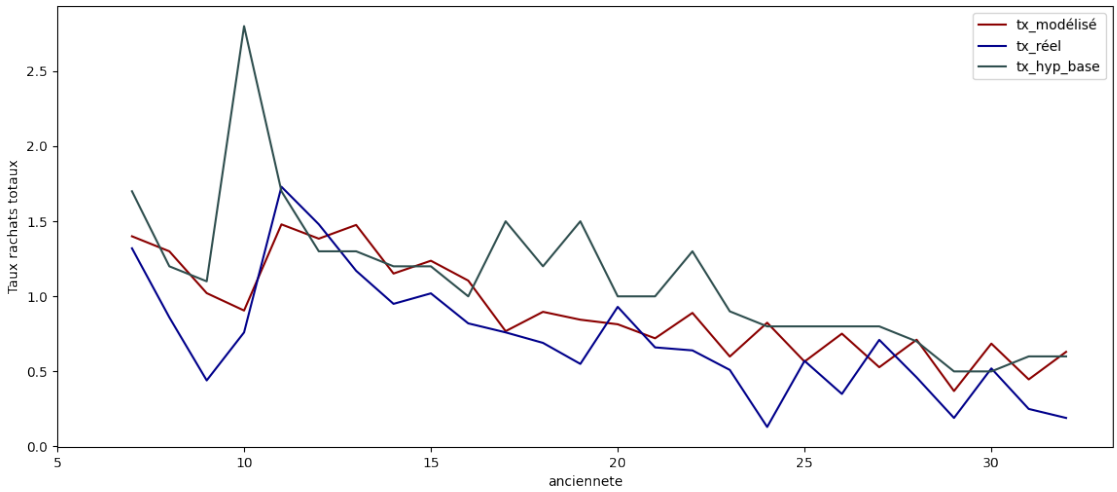


FIGURE 4.28 – Comparison of the modeled, actual and base laws for the year 2021 : "AG" network

Once updated in the internal model, the new total lapses laws calibrated by our Machine Learning algorithms reduce the experience gaps on lapses by 75% and 45% respectively on the "prod_A" and "prod_B" products. On the other hand, the reduction of the gaps on reserves is relatively weak because of the strong contribution of switches.

- **New lapses laws in number**

Following the important experience gaps observed on some products and the conclusions of our prediction models, this study proposes to calibrate new lapses laws by portfolio segment ; this approach aims to overcome the limitations of the current law ; it minimizes lapses in number of the "small" contracts ; which leads to the overestimation of the unit costs on these contracts and thus the insurer's technical result. In practice, the new approach is to estimate by Machine Learning algorithms, the proportion of contracts that could be fully surrendered separately on "small", "medium" and "large" contracts. The segmentation thresholds are defined by expert opinion. It follows from our prediction models that Machine Learning algorithms would be less suitable for modeling the laws of lapses in numbers on "small" and "large" contracts. On the other hand, these algorithms perform better when the portfolio includes "medium" contracts. The size and the variability in characteristics of these different portfolio segments probably justifies the differences observed in model performance.

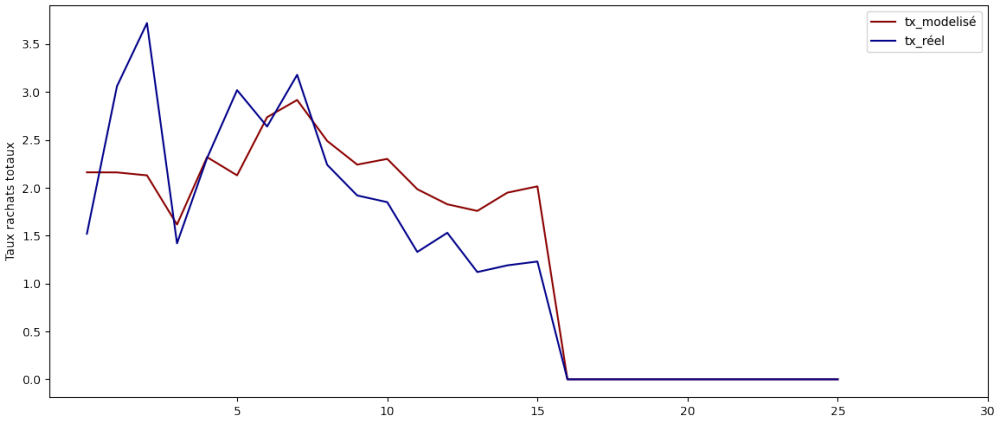


FIGURE 4.29 – Comparison of modeled and actual laws for the year 2021 : case of "average" contracts

limitations and perspectives of the study

Although the results of this study are quite satisfactory, it has some limitations due to the fact that it does not take into account the dynamic aspect of the phenomenon studied. Furthermore, our algorithms have difficulties in correctly modeling peaks in lapses ; models better adapted to this rate structure could be used to improve predictions. Another perspective of this work would be to analyze the contribution of the profit sharing policy to the experience gaps observed on life insurance reserves.

ANNEXE A

FIGURE 4.30 – Les 3 piliers de Solvabilité II

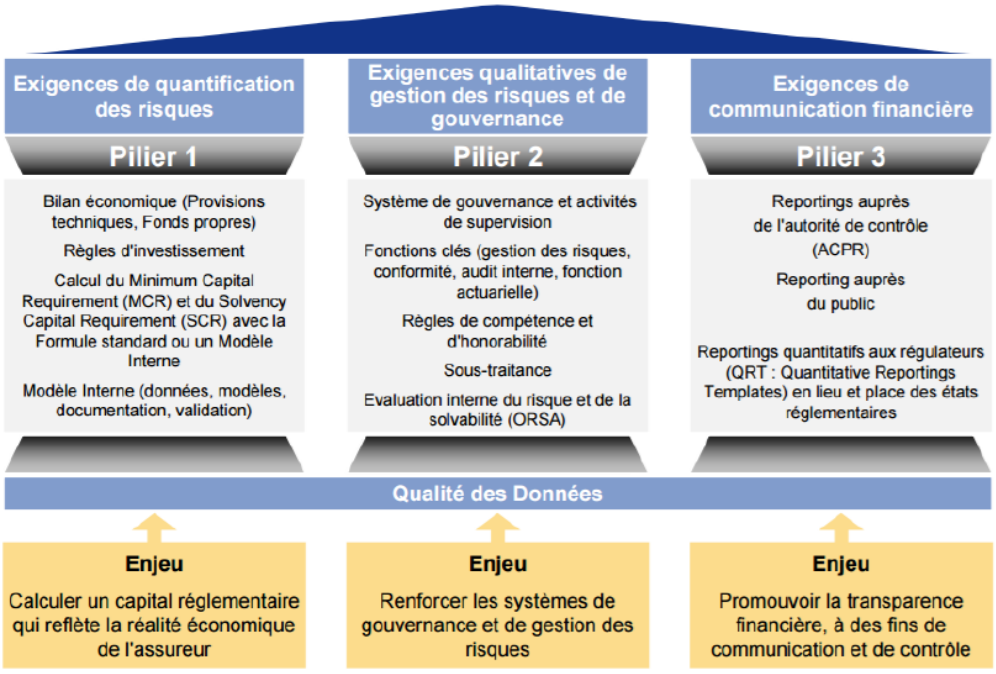


Tableau 4.15 – Exemple Corrections et correspondances R4 et VIPR

Maille ALIM	Maille traitement
A83_TELLUS_0_EUR	A83_TELLUS_xxx_EUR
A83_TELLUS_35_EUR	A83_TELLUS_xxx_EUR
A83_TELLUS_45_EUR	A83_TELLUS_xxx_EUR
A83_TELLUS_0_UC	A83_TELLUS_0_UC
A83_TELLUS_35_UC	A83_TELLUS_0_UC
A83_TELLUS_45_UC	A83_TELLUS_0_UC
ASAC_INF	ASAC_xxx
ASAC_SUP	ASAC_xxx
ASAC_SUP_ER1	ASAC_xxx_ER1
CREA_025	CREA_xxx
CREA_TIT_MOY	CREA_xxx

Figure 4.31 – Distribution des écarts sur les PM de cloture

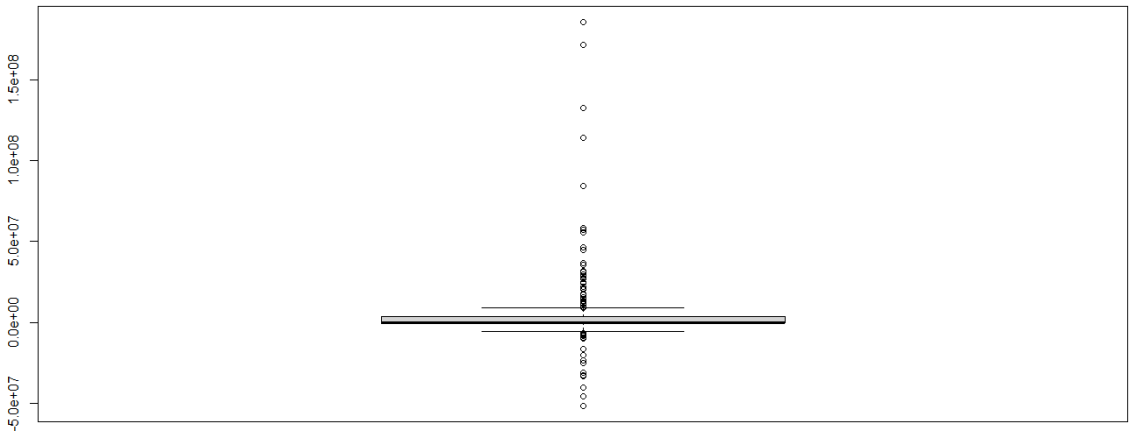


Tableau 4.16 – Description des taux de chocs par année

	Min	1st Q.	Median	Mean	3rd Q.	Max
2020	-100	-63.33	-16.9	-11.95	15.73	331.08
2019	-100	-57.48	0	-1.91	35.490	210.06
2017	-100	-47.83	0	0.27	36.26	175.73

FIGURE 4.32 – Choix du nombre de clusters : K-médoïdes

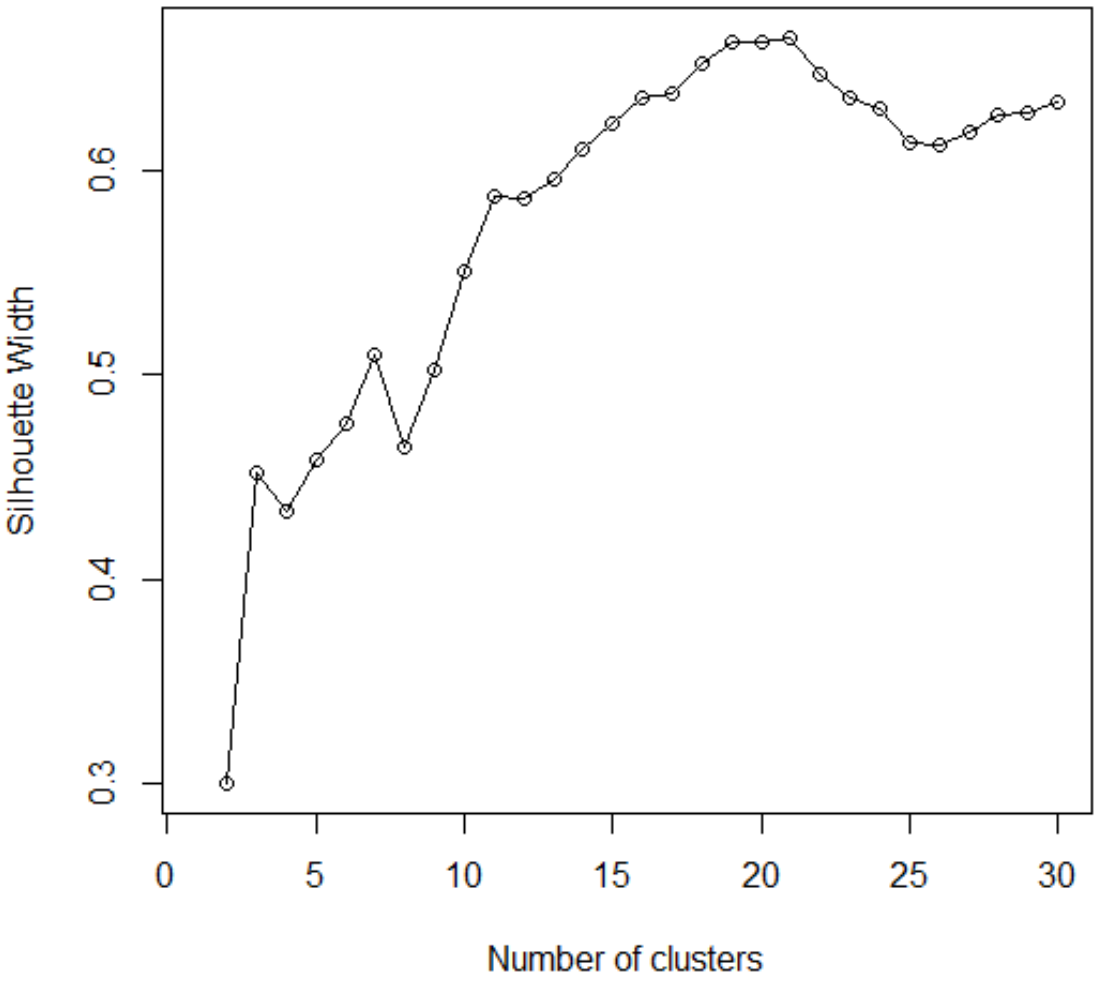
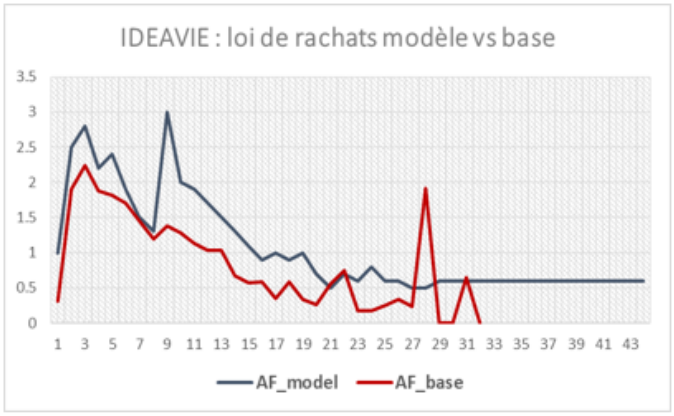
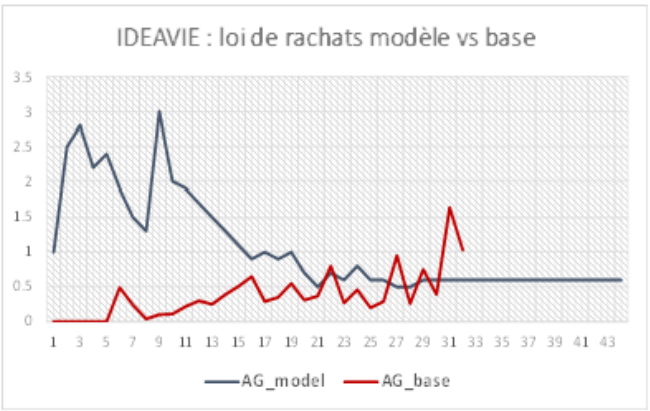


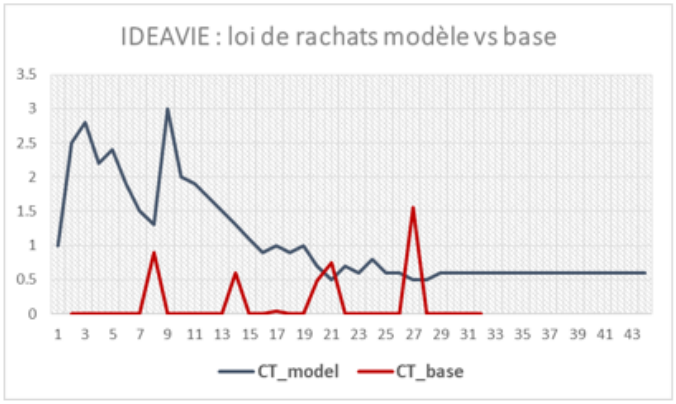
TABLEAU 4.17 – Nombre de groupe rachats par clusters

N° clusters	nb produits	nb groupe rachats	N° clusters	nombre produits	nb groupe rachats
1	4	2	11	14	8
2	3	2	12	9	4
3	3	1	13	6	2
4	9	5	14	21	11
5	7	3	15	15	7
6	5	2	16	18	6
7	24	12	17	4	2
8	10	5	18	11	5
9	7	4	19	21	15
10	29	18	20	2	2

ANNEXE B







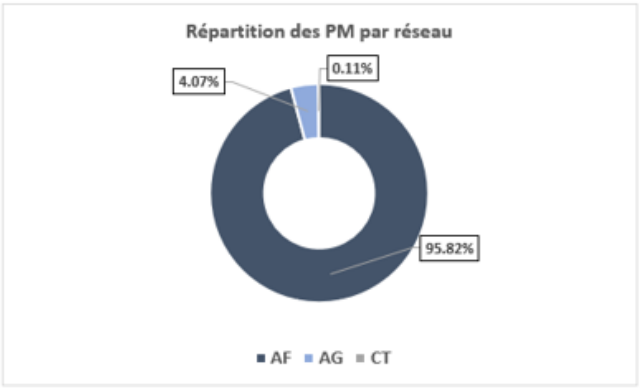


FIGURE 4.33 – Taux de rachats totaux réels et modélisés par ancienneté et par réseau de distribution sur le produit "prod_C"

FIGURE 4.34 – Taux et montant de rachats totaux par ancienneté et par classe d'encours sur les produits "prod_C" et "prod_B"

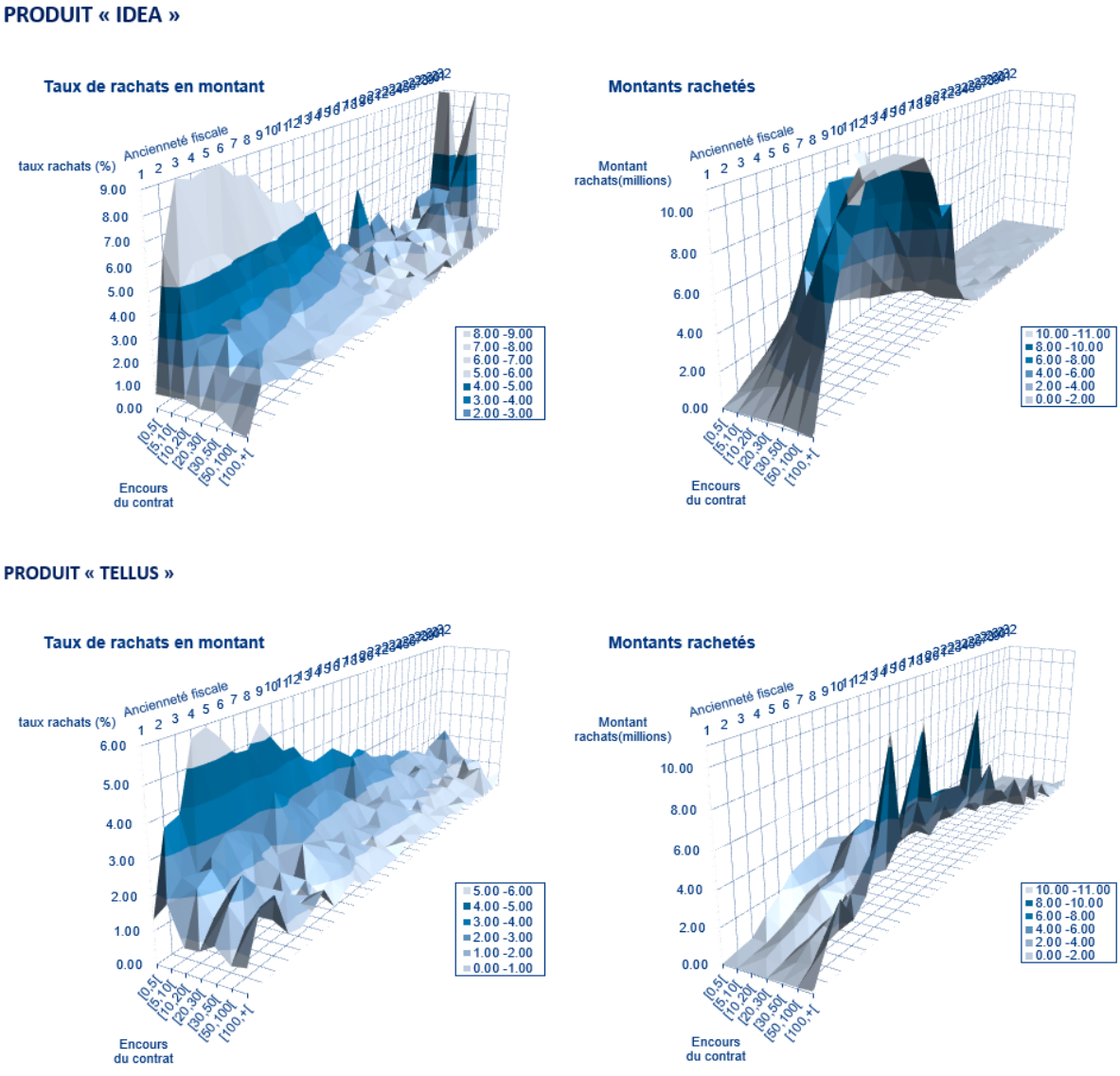


FIGURE 4.35 – Taux et montant de rachats totaux par ancienneté et part d'UC sur les produits "prod_C" et "prod_B"

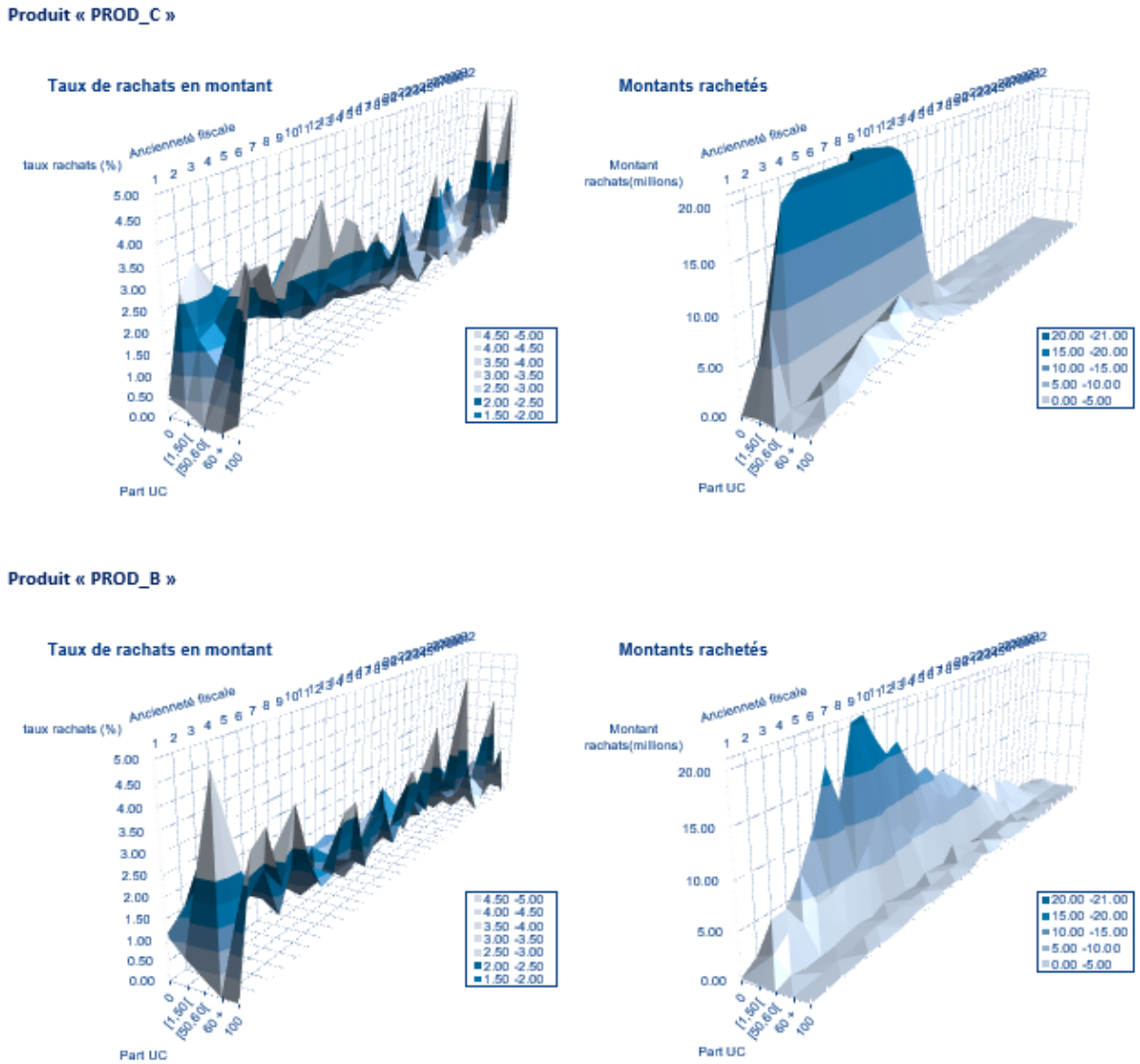


FIGURE 4.36 – Taux et montant de rachats totaux par ancienneté et par mode de gestion du contrat sur les produits "prod_C" et "prod_B"

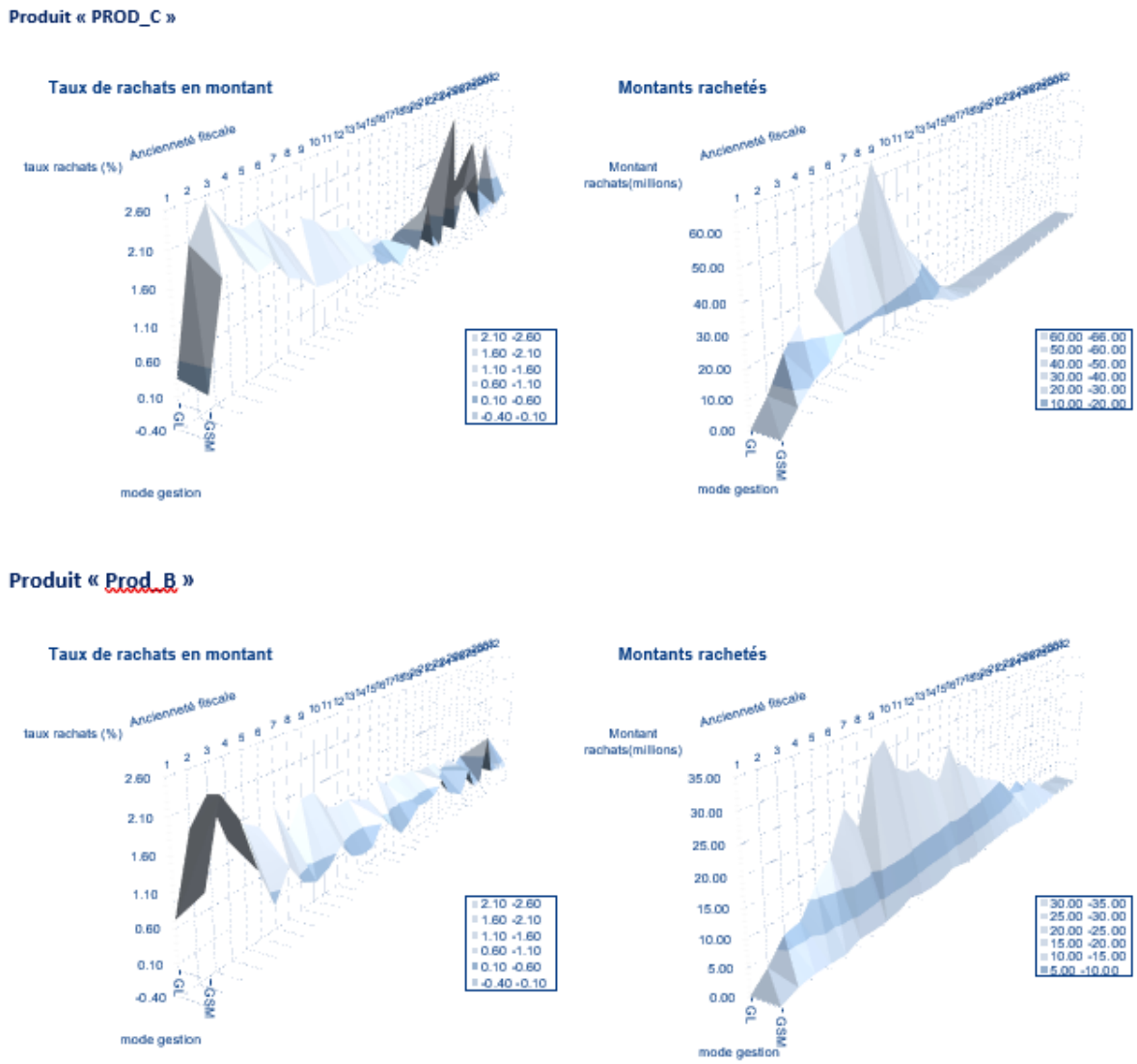


FIGURE 4.37 – Taux et montant de rachats totaux par ancienneté et par périodicité de la prime sur les produits "prod_C" et "prod_B"

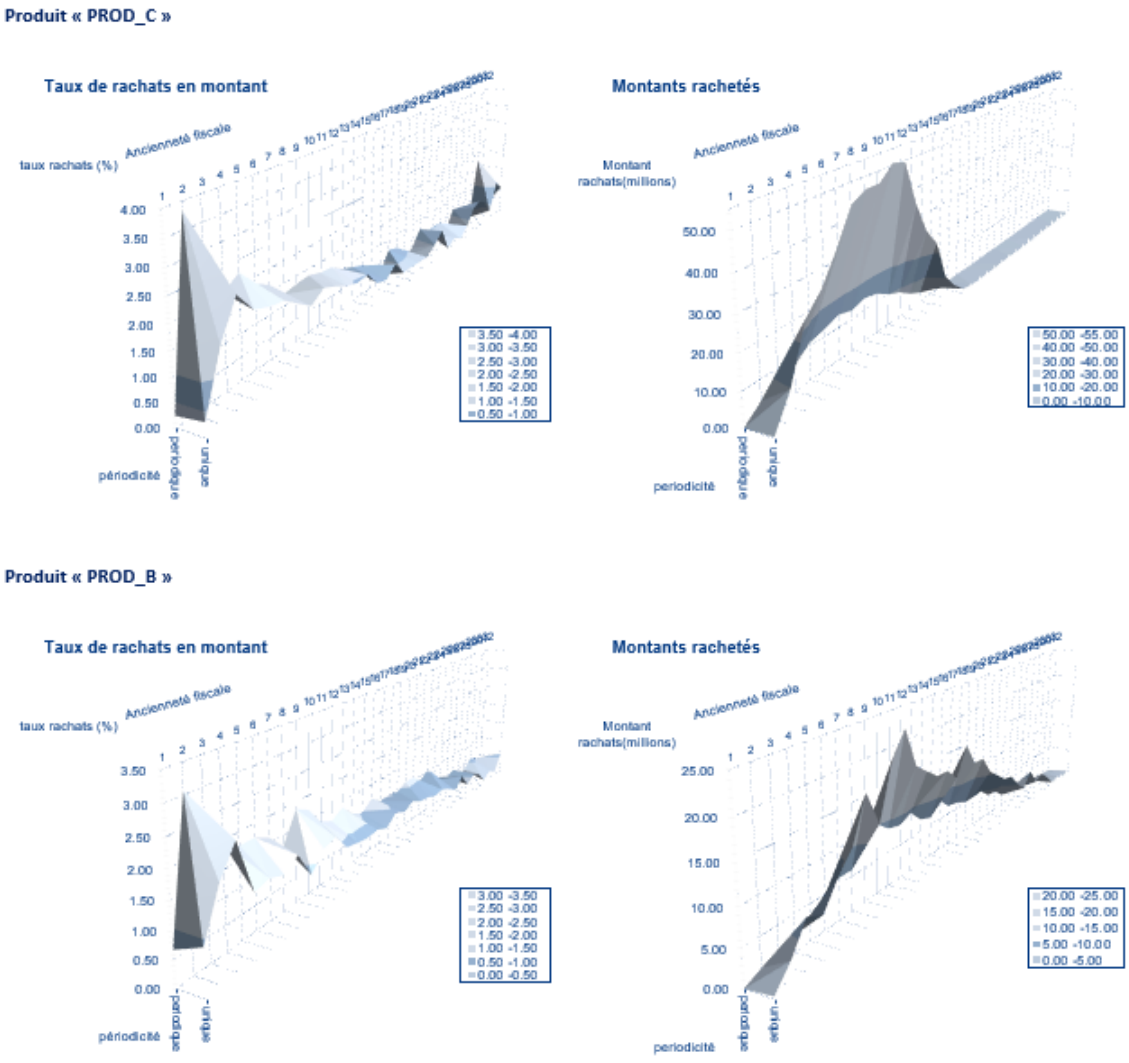


FIGURE 4.38 – Taux et montant de rachats totaux par ancienneté et par type de support sur les produits "prod_C" et "prod_B"

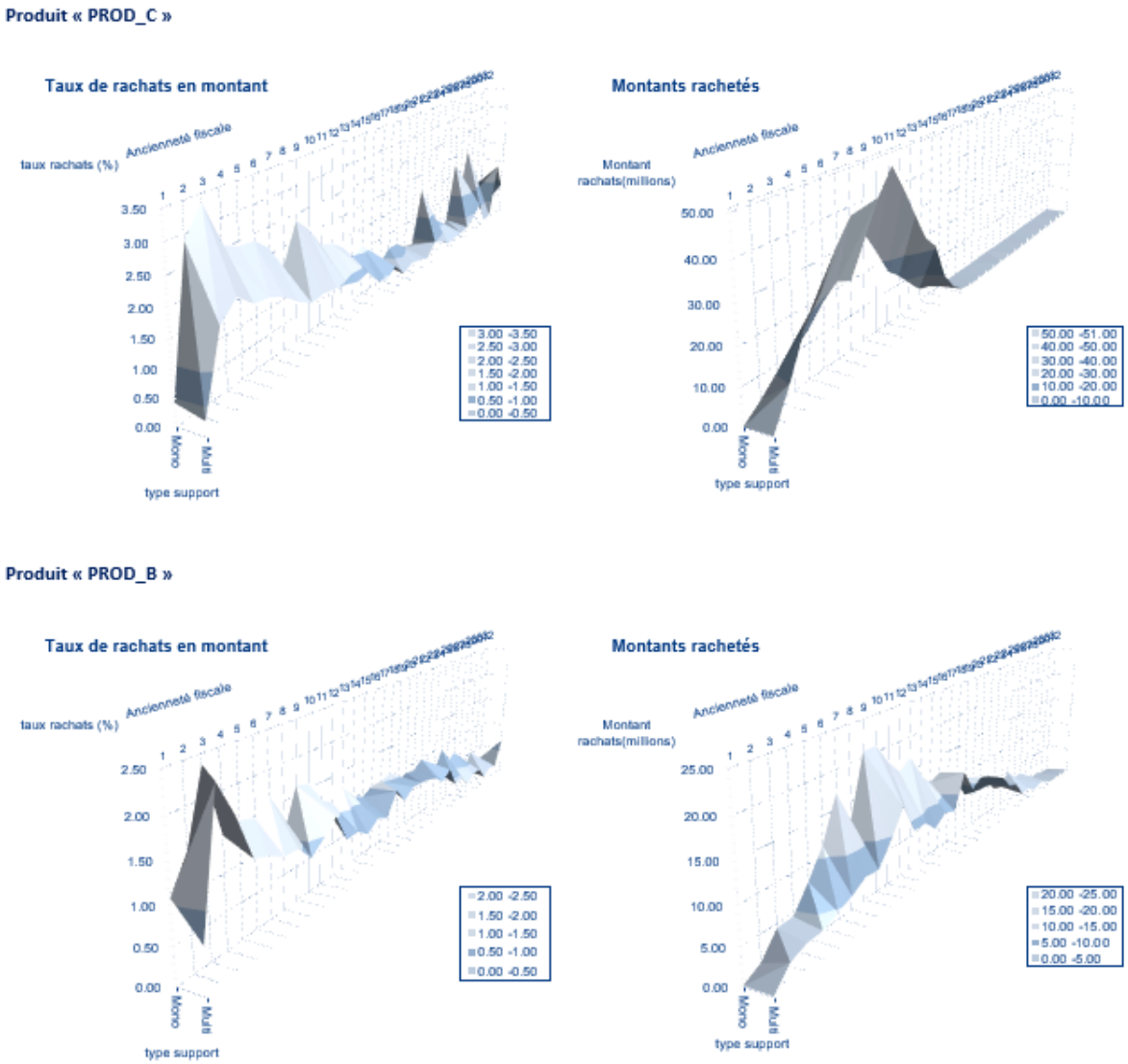


FIGURE 4.39 – Taux et montant de rachats totaux par ancienneté et par nombre de support sur le produit "prod_A"

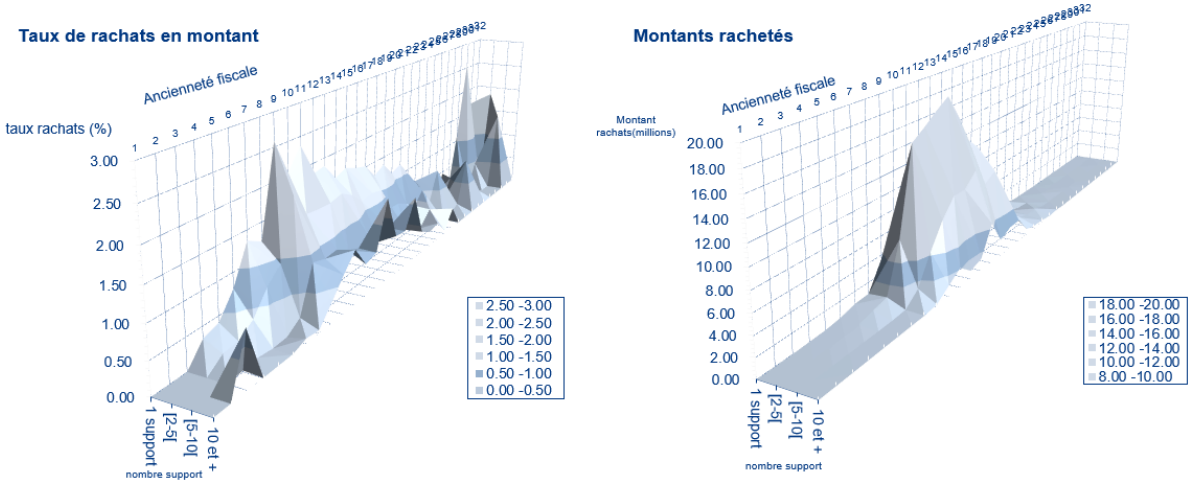


FIGURE 4.40 – Taux et montant de rachats totaux par ancienneté et par nombre de support sur le produit "prod_C" et "prod_B"

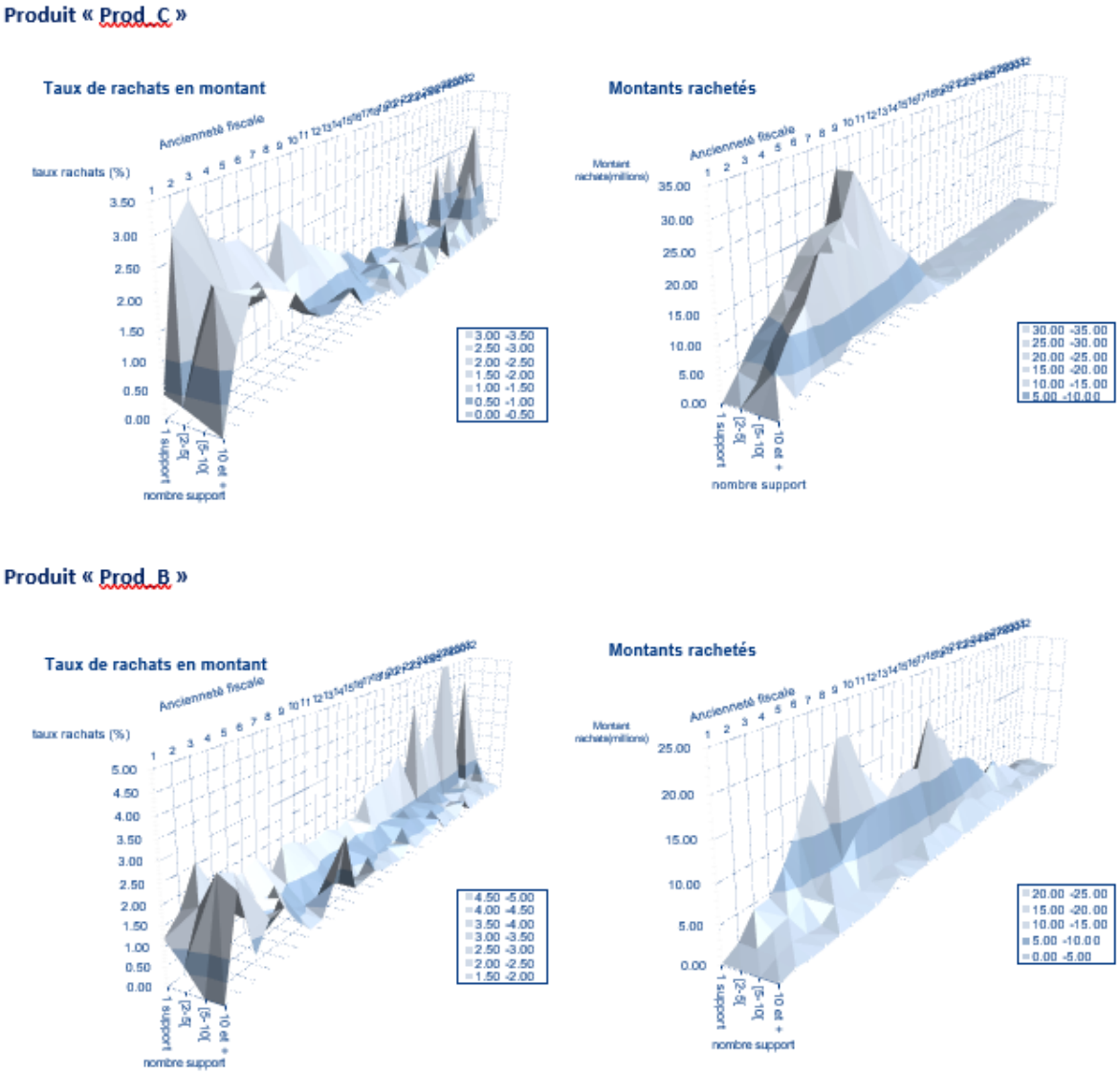


FIGURE 4.41 – Taux et montant de rachats totaux par ancienneté et par sexe sur les produits "prod_C" et "prod_B"

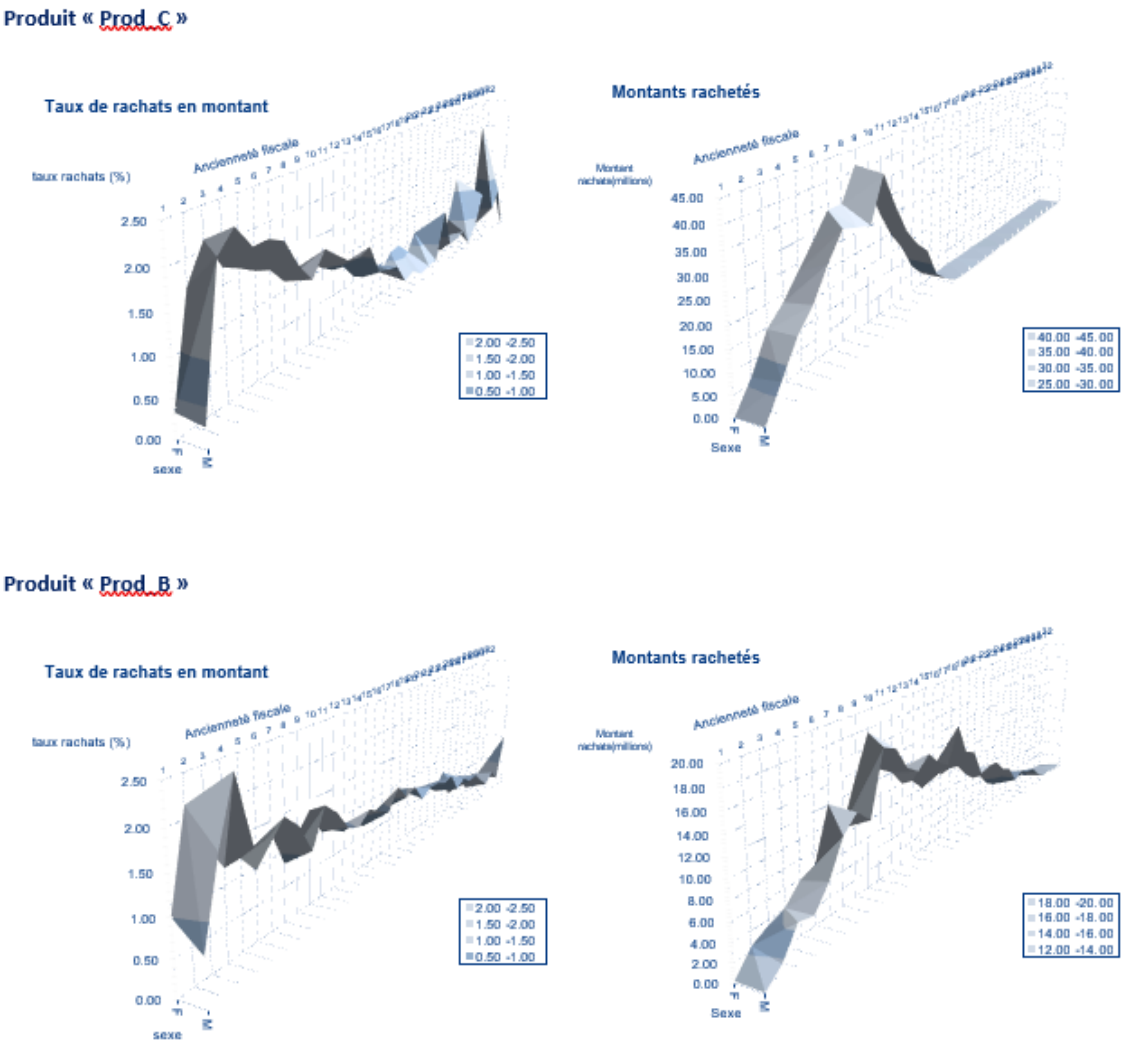


FIGURE 4.42 – Taux et montant de rachats totaux par ancienneté et par classe d'âge sur les produits "prod_C" et "prod_B"

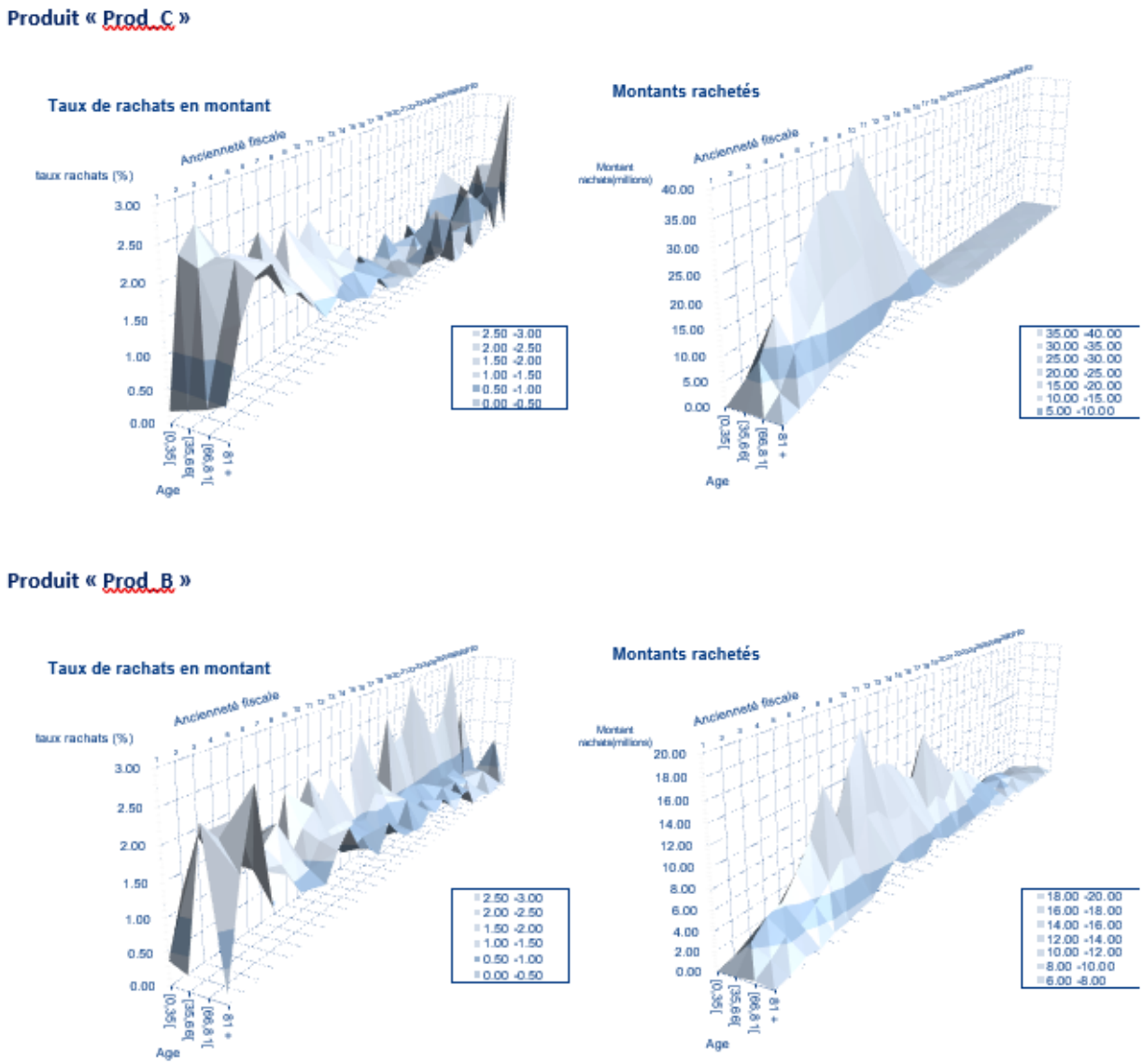
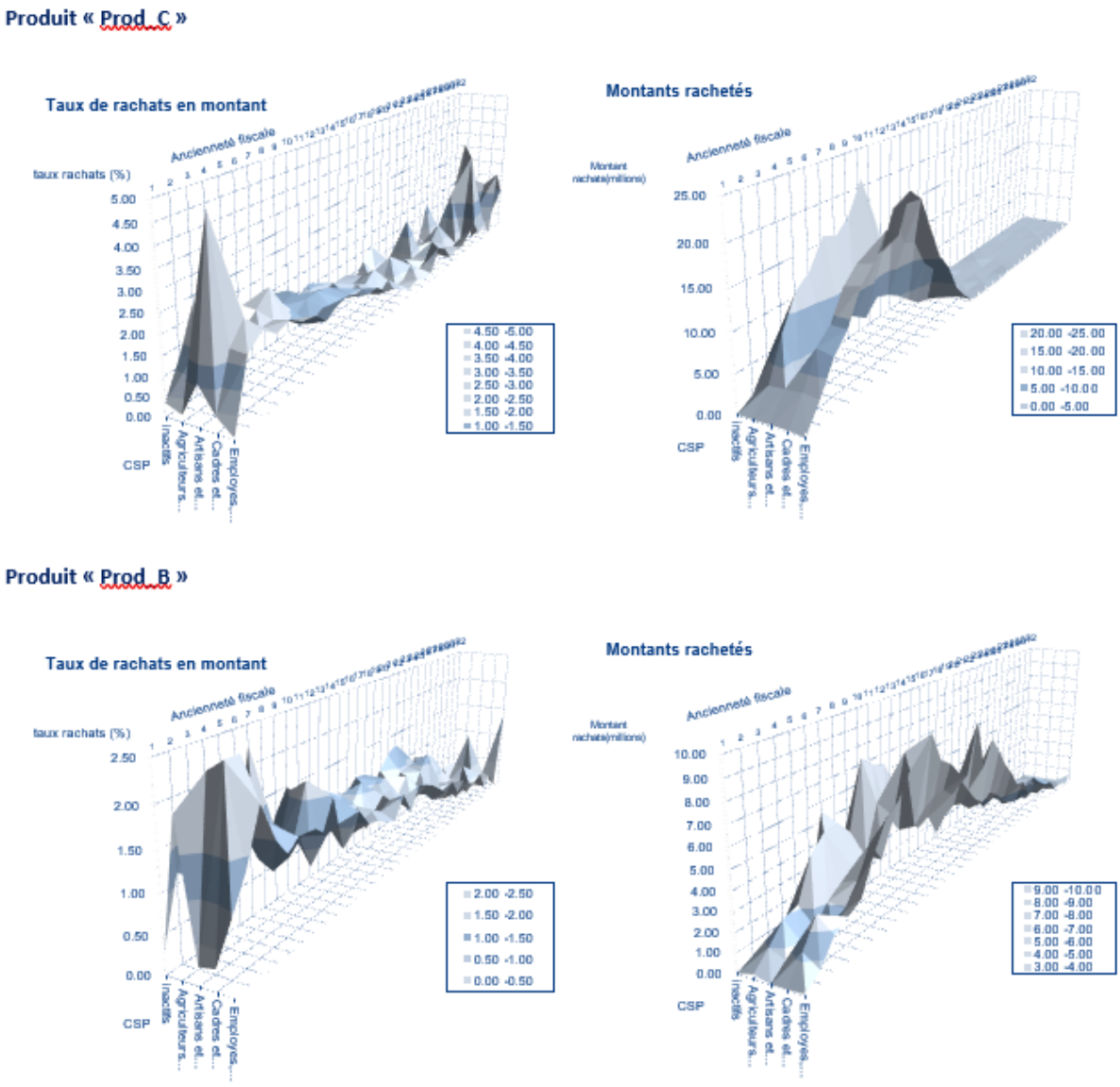


FIGURE 4.43 – Taux et montant de rachats totaux par ancienneté et par CSP sur les produits "prod_C" et "prod_B"



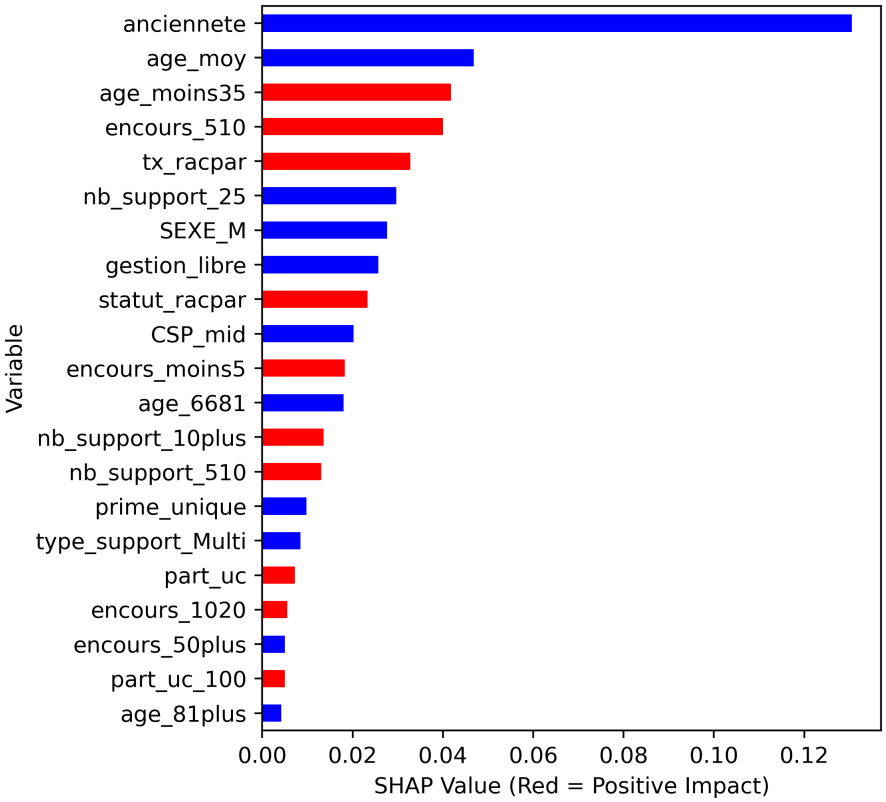


FIGURE 4.44 – Impact des variables dans la régression : produit "prod_B"

SIGLES ET ABRÉVIATIONS

CFO : Chief Financial Officer

CSP : catégorie socio-professionnelle

MAE : Mean Absolute Error

MCEV : Market Consistency Embedded Value

ML : Machine Learning

OAT : Obligation Assimilable au Trésor

OPCVM : Organisme de Placement Collectif en Valeurs Mobilières

PB : Participation aux Bénéfices

PM : Provision Mathématique (sous solvabilité 2)

PME : Petites et moyennes entreprises

PVFP : present value of future profits

PVSF : Prévoyance et santé

RF : Random Forest

RM : Risk Margin

RMSE : Root Mean Squared Error

S2 : Solvabilité 2

SCI : Société Civile Immobilière

SCPI : Société Civile de Placement Immobilier

SICAV : Société d'Investissement à Capital Variable

TMG : Taux Minimum Garanti

UC : Unité de Compte

VIEP : Epargne Individuelle

LISTE DES TABLEAUX

LISTE DES FIGURES