

Mémoire présenté le : 2 mai 2023
pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires

Par : Lucie BULTEAU

Titre : Modélisation des comportements clients sur les versements périodiques en assurance vie individuelle et impact sur la rentabilité

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

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

*Membres présents du jury de l'Institut
des Actuaires :*

S. BORDELET

S. SIMRICK

Membres présents du jury de l'ISFA :

D. DOROBANTU

S. LOISEL

Entreprise :

Nom : AXA France

Signature :



Directeur de mémoire en entreprise :

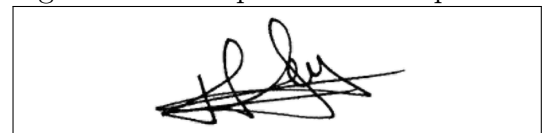
Nom : Feyza ERGUVEN

Signature :

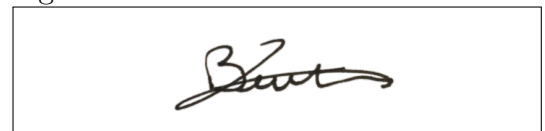


*Autorisation de publication et
de mise en ligne sur un site de
diffusion de documents actuariels
(après expiration de l'éventuel délai de
confidentialité)*

Signature du responsable entreprise



Signature du candidat



Résumé

L'assurance-vie est un produit d'épargne incontournable en France, souvent privilégiée pour sa souplesse, sa fiscalité avantageuse et pour la variété des investissements qu'elle propose. La rentabilité prospective de ce type de produit consiste à déterminer si les marges futures dégagées sont suffisantes pour couvrir les frais généraux. Le comportement des clients sur les mouvements effectués est un élément clé dans l'estimation des engagements des assureurs et des marges financières. Plusieurs études actuarielles ont déjà été menées sur la modélisation des versements, des arbitrages et des rachats, dans le cadre de problématiques réglementaires telles que Solvabilité 2 ou IFRS. Ce mémoire se concentre sur la modélisation des versements périodiques pour analyser l'incidence des volumes des primes sur la rentabilité attendue.

Les versements périodiques, en complément des versements initiaux et libres, représentent une partie non négligeable des montants collectés sur les contrats d'épargne. Les facteurs structurels influençant les comportements des assurés en matière de versements périodiques sont étudiés pour trois produits commercialisés par Axa France. La modélisation des versements périodiques proposée s'effectue à partir d'analyses statistiques et de méthodes de *Machine Learning*, afin de déterminer les montants de versements en fonction de différents facteurs. Ces estimations mettent en évidence l'importance de variables telles que le niveau de l'encours, le versement initial et l'âge. L'introduction de sous-groupes homogènes de clients permet de prédire la rentabilité attendue du produit et de souligner l'impact de certains profils d'assurés sur la rentabilité globale.

Mots clés : épargne, assurance-vie, versements périodiques, comportement client, Machine Learning, rentabilité

Abstract

Life insurance is an essential saving product in France, often favoured for its flexibility, its advantageous tax system, and the variety of investments it offers. The prospective profitability of the type of products allows to determine whether the future margins generated are sufficient to cover general expenses. As such, the behaviour of clients regarding the movements made on their contract is a key element in estimating insurers' commitments and the financial margins generated. Several actuarial studies have already been conducted on the modeling of premiums, arbitrations, and lapses in the context of regulatory issues such as Solvency 2 or IFRS. By contrast, this paper focuses on the modeling of periodic payments to analyze the impact of premium volumes on expected profitability.

In addition to the initial and free premiums, periodic premiums represent a part of the amounts collected on the savings contracts. The structural factors influencing policyholders' behaviour in terms of periodic premiums are studied for three products marketed by Axa France. The modeling of periodic premiums is carried out using statistical analysis and Machine Learning methods in order to identify the amounts of premiums as a function of different factors. These estimates highlight the importance of variables such as the level of outstanding amount, the initial premium, and age. The introduction of homogeneous subgroups of clients allows the prediction of the product's expected profitability. Additionally, it reveals the impact of certain policyholder profiles on the overall profitability.

Keywords : savings, life insurance, periodic payments, customer behavior, Machine Learning, profitability

Remerciements

Je tiens tout d'abord à remercier Feyza Erguven, ma tutrice d'entreprise, pour ses conseils et le temps qu'elle m'a consacré tout au long de mon alternance et de la rédaction de ce mémoire.

De plus, toute ma reconnaissance est portée envers Elodie Fourgous et Olivia Lafay qui m'ont accueillie dans l'équipe Rentabilité et Réassurance au sein d'Axa France. La confiance et l'autonomie qu'elles m'ont accordée ont contribué à la réalisation de mes missions dans un cadre bienveillant. Je remercie aussi mes collègues pour leur accueil et leur disponibilité qui m'ont permis de réaliser ce mémoire dans les meilleures conditions possibles.

Je tiens également à exprimer ma gratitude envers mon tuteur pédagogique, Pierrick Piette, pour son aide et le suivi du mémoire, et plus généralement l'ensemble de l'équipe pédagogique de l'ISFA pour l'enseignement dispensé.

Enfin, un grand merci à ma famille pour leurs encouragements tout au long de mes études, à Clément pour son soutien quotidien et ses précieuses relectures, et tous ceux qui ont contribué de près ou de loin à l'accomplissement de ce mémoire.

Table des matières

Remerciements	1
Introduction	5
1 Cadre de l'étude : contextualisation et enjeux	6
1.1 L'assurance-vie en France	6
1.1.1 Le patrimoine des Français	6
1.1.2 L'intérêt de l'assurance-vie	7
1.1.3 Le marché de l'assurance-vie en France	9
1.2 Principe du contrat d'assurance-vie	10
1.2.1 La gestion du contrat	10
1.2.2 Les types de support	12
1.3 Contexte	13
1.3.1 L'environnement économique	13
1.3.2 L'impact de la réglementation : la loi PACTE	14
1.3.3 Littérature actuarielle sur la modélisation des versements	15
2 Étude des données	17
2.1 Présentation des données à disposition	17
2.1.1 Périmètre et agrégation des données	17
2.1.2 Présentation des variables	18
2.2 Retraitement des variables	23
2.2.1 Données incomplètes	23

2.2.2	Regroupement de données	24
2.3	Analyse du portefeuille	29
2.3.1	Corrélations	29
2.3.2	Analyse bivariée	31
3	Modélisation des versements	39
3.1	Mise à jour des lois actuelles	39
3.1.1	Modélisation de la part de contrats générant des versements périodiques	40
3.1.2	Modélisation sur le montant des primes	42
3.1.3	Comparaison des deux modèles	44
3.2	Théorie des modèles de Machine Learning	47
3.2.1	Principe d'échantillonnage	48
3.2.2	Arbre de régression	49
3.2.3	Forêts aléatoires	51
3.2.4	L'Extreme Gradient Boosting	53
3.3	Application sur le portefeuille	55
3.3.1	Paramétrage des modèles	55
3.3.2	Prédiction et synthèse comparative	60
3.3.3	Validation du modèle	62
3.4	Regroupement des contrats	63
3.4.1	Méthode des k-means	63
3.4.2	Détermination des groupes de contrats	65
4	Rentabilité des produits	69
4.1	Modélisation de la rentabilité des produits	69
4.1.1	Présentation du modèle utilisé	70
4.1.2	Lois de comportement client	70
4.1.3	Déroulé et projection des réserves	71
4.1.4	Déroulé des résultats	72

4.2	Les indicateurs de rentabilité	73
4.2.1	Définition de la New Business Value	73
4.2.2	Le SCR	73
4.2.3	Le taux de rentabilité interne	75
4.2.4	Le ratio combiné	75
4.2.5	Un indicateur de solvabilité	75
4.3	Impact du modèle sélectionné sur la rentabilité	76
4.3.1	États des hypothèses techniques	76
4.3.2	<i>Model points</i> retenus	77
4.3.3	Résultats	80
	Conclusion	83
	Bibliographie	85
	Annexes	87
	A Méthodes de corrélation	87
	B Corrélation des variables	90
	C Les indicateurs de performance	92
	D Résultats de rentabilité sur un PER	94
	E Évolution des primes périodiques en fonction du temps	95

Introduction

La crise sanitaire a favorisé des taux d'épargne plus importants chez les ménages français. Parmi les différents supports d'épargne financière proposés sur le marché, les contrats d'assurance-vie ont connu une collecte plus importante depuis le début de l'année 2021. Cette popularité s'explique par la combinaison d'une garantie sur le capital investi, d'une rentabilité attractive et d'une grande flexibilité en termes de gestion du contrat.

Dans ce contexte, la projection des primes futures versées et des mouvements effectués est un enjeu majeur pour le calcul et le suivi de la rentabilité des produits d'assurance-vie. En particulier, la précision de ces projections est essentielle pour déterminer les principaux indicateurs de rentabilité, car le volume des primes perçues est la source des frais prélevés et des marges financières dégagées.

Les versements périodiques forment pour certains produits la part principale des flux de primes. Ils représentent en général un flux plus stable et donc plus facilement prévisible comparé aux versements libres. La détermination d'un montant de versement permanent par l'assuré peut être expliquée par des caractéristiques propres à l'assuré et au contrat (représentant l'ensemble des facteurs structurels) ou bien par un environnement économique plus ou moins favorable représenté par des variables conjoncturelles : indices du CAC40, taux d'inflation ou encore taux de chômage.

L'objectif de ce mémoire est de déterminer les facteurs structurels influençant les comportements des assurés sur les versements périodiques et d'estimer les montants de ces versements sur leur contrat d'épargne individuelle. L'influence des facteurs structurels sera déterminée par la modélisation des montants de versements sur les données de trois produits commercialisés au sein d'Axa France. Dans un premier temps, deux approches statistiques s'appuyant sur l'ancienneté des contrats seront proposées. Puis, des modèles prédictifs de *Machine Learning* tels que les arbres de régression, forêts aléatoires et *Extreme Gradient Boosting* seront construits sur les observations du portefeuille. Enfin, l'impact des volumes des primes sur la rentabilité des produits sera analysé à travers plusieurs profils d'assurés.

Chapitre 1

Cadre de l'étude : contextualisation et enjeux

Ce premier chapitre a vocation à présenter l'environnement d'étude de ce mémoire. Le marché de l'assurance-vie en France et les grands principes du contrat d'assurance-vie sont présentés avant de développer l'enjeu du mémoire dans le contexte économique et réglementaire actuel.

1.1 L'assurance-vie en France

1.1.1 Le patrimoine des Français

La crise sanitaire et économique de 2020 a permis aux ménages d'épargner davantage. En France, comme indiqué dans le dernier rapport [6] de la Banque de France sur l'épargne des ménages, les taux d'épargne ont atteint 21,7 % du revenu disponible brut ¹ au premier trimestre 2021. Cette masse importante d'épargne a été forcée par les mesures de confinement et par une incitation à la prudence face à l'instabilité de la situation.

Les flux d'épargne financière des ménages français se sont dirigés principalement vers des placements d'épargne liquide tels que les dépôts bancaires et les dépôts à vue. En 2021, les niveaux d'épargne diminuent mais restent supérieurs à ceux atteints avant la crise.

1. Le revenu disponible brut représente le revenu restant disponible pour les dépenses de consommation et l'épargne des ménages après déduction des charges fiscales et sociales.

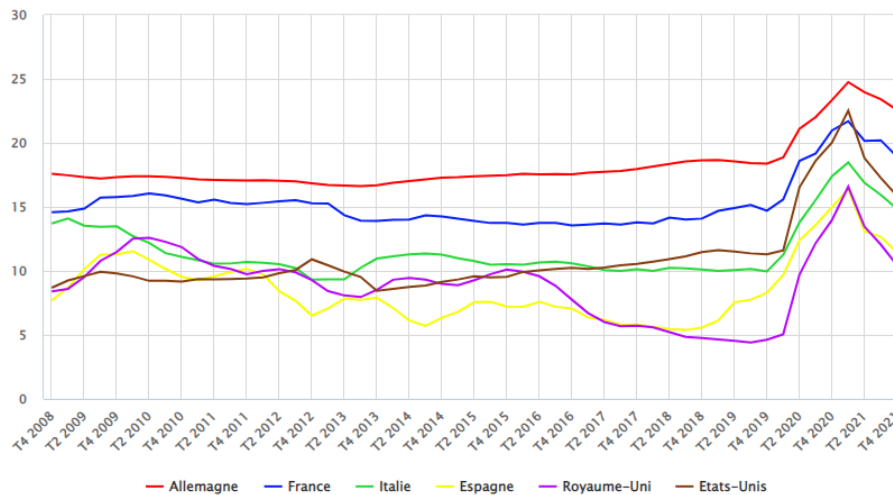


FIGURE 1.1 – Évolution des taux d'épargne depuis 2008 dans les principaux pays d'Europe et les États-Unis (Source : Banque de France)

L'instabilité de l'économie et de l'avenir du système de retraite continue de pousser les ménages français à privilégier des produits peu risqués plutôt que des produits à rendements plus élevés. Selon l'INSEE [16] [17], sur les 93 % des ménages détenant du patrimoine, la majorité de l'épargne est investie dans l'immobilier, puis 41 % représente des placements financiers. En particulier, les placements financiers se composent :

- d'assurance-vie (35 % des placements) ;
- de numéraire et de dépôts tels que le Livret A, le Livret Jeune et le Plan Épargne Logement (PEL) (29 % des investissements) ;
- d'actions et de parts de fonds d'investissement (27 % des placements financiers) ;
- dans une moindre mesure, le reste est constitué de titres de créances et crédits.

Ainsi, étant plus frileux sur les investissements en bourse, les ménages français privilégient les produits d'assurance-vie comme placement d'épargne financière.

1.1.2 L'intérêt de l'assurance-vie

Les contrats d'assurance-vie peuvent être assimilés à des contrats d'épargne bénéficiant d'une réglementation et d'une fiscalité plus avantageuses, rendant ces supports d'investissement plus intéressants qu'un contrat d'épargne classique.

En effet, l'assurance-vie est un placement présentant plusieurs intérêts concernant la transmission du patrimoine au décès. Il existe une clause bénéficiaire permettant de choisir librement les tiers qui percevront les capitaux en cas de décès. La transmission y est optimisée. La transmission des capitaux liés à un contrat d'assurance-vie présente peu de procédure administrative car elle ne nécessite pas d'acte notarié. D'après l'article L132-12 du Code des assurances, les assurances vie ne font pas partie de la succession du défunt. Il n'est donc pas obligatoire de déclarer les contrats d'assurance-vie au notaire.

Enfin en cas de décès, la fiscalité est plus favorable. Contrairement aux époux, partenaire de PACS et quelques cas particuliers qui bénéficient d'une exonération de droits de succession, l'assurance-vie permet aux autres bénéficiaires désignés d'appliquer les exonérations suivantes :

Versements avant 70 ans :	Versements à partir de 70 ans :
<ul style="list-style-type: none"> ■ Exonération jusqu'à 152 500€ par bénéficiaire ■ 20% jusqu'à 852 500€ ■ 31,25% au-delà. 	<ul style="list-style-type: none"> ■ Abattement de 30 500€ (à partager entre tous les bénéficiaires) puis application des droits de succession suivant le degré de parenté entre l'assuré et le bénéficiaire.

FIGURE 1.2 – Exonérations et abattements des droits de succession en assurance-vie

Aussi, l'assurance-vie est un moyen souple pour investir, puisqu'il dispose des caractéristiques suivantes :

- possibilité d'ouvrir plusieurs contrats ;
- disponibilité du capital avec la capacité d'effectuer un rachat à tout moment ;
- diversité des investissements en fonction des niveaux de risques auxquels l'épargnant souhaite s'exposer ;
- proposition d'options pour protéger les montants investis telles que la garantie décès plancher permettant en cas de moins-values d'assurer aux bénéficiaires au minimum les versements versés indépendamment des aléas financiers ;
- la fiscalité est également favorable en cas de rachat. Les revenus tirés d'un contrat d'assurance-vie sont soumis aux prélèvements sociaux (PS) à hauteur de 17,2 %. Les prélèvements sociaux sont composés de la CSG, la CRDS et du prélèvement de solidarité. Puis, lors d'un rachat, les gains sont imposés selon l'ancienneté du contrat par Prélèvement de l'impôt Forfaitaire Libérateur (PFL), ou bien intégrés à la déclaration de revenus. Les spécificités de la fiscalité applicable sur les plus-values acquises en fonction de l'ancienneté du contrat et de la date des versements des primes sont détaillées dans le tableau 1.3.

Âge du contrat	Primes versées avant le 27 septembre 2017 (et à partir du 1 janvier 1998)	Primes versées à partir du 27 septembre 2017
Avant 4 ans	52,2 % 35 % (PFL) + 17,2 % (PS) ou barème progressif	30 % PFU = 12,8% + 17,2 % (PS) ou barème progressif
Entre 4 et 8 ans	32,2 % 15 % (PFL) + 17,2 % (PS) ou barème progressif	
Après 8 ans	Abattement annuel de 4600 euros pour une personne célibataire ou 9200 euros pour un couple marié ou pacsé.	
	24,7 % 7,5 % (PFL) + 17,2 % (PS) ou barème progressif	24,7 % PFU = 7,5 % + 17,2 % (PS) pour les gains réalisés sur la part des primes inférieure à 150 000 euros (taux de 30 % au-delà). ou barème progressif

FIGURE 1.3 – Fiscalité des plus-values lors des rachats (Source : <https://avenuedesinvestisseurs.fr/fiscalite-assurance-vie-retrait-rachat/>)

Le taux d'imposition sur les plus-values est plus favorable pour les contrats de plus de 4 ans et est encore plus avantageux sur les contrats de plus de 8 ans avec une fiscalité plus faible et un abattement annuel. Ainsi, l'ancienneté du contrat influe sur les mouvements et en particulier sur les retraits des clients.

L'assurance-vie présente donc de nombreux avantages, la souplesse des contrats, les rémunérations proposées, la fiscalité allégée et la transmission simplifiée en font un placement apprécié des français.

1.1.3 Le marché de l'assurance-vie en France

Le marché de l'assurance-vie est concentré sur les bancassureurs, qui captent 82 % de l'activité. L'assurance-vie en ligne reste marginale. Bien qu'impactés par la crise sanitaire, les bancassureurs ont pu retrouver les niveaux de collecte d'avant crise sur l'assurance-vie. En 2021, la reprise de la collecte nette - prenant en compte les versements mais aussi les retraits - s'est établie à 18,3 milliards d'euros.

Néanmoins, la collecte sur les fonds en euros reste négative. Autrement dit, les fonds en euros subissent des retraits de capitaux. Ces retraits sont expliqués par des mouvements d'arbitrage vers les unités de compte. Selon l'ACPR, « En 2021, les contrats d'assurance-vie sur les fonds en euros ne représentent qu'un peu plus de la moitié des nouveaux versements sur les contrats d'assurance-vie (56 %), contre 85 % en 2011 ». Ainsi, la collecte positive est portée par les fonds en unités de compte en pleine croissance atteignant 30,6 milliards d'euros de collecte nette. Cet engouement pour les fonds en unités de compte peut s'expliquer d'une part par le contexte économique des taux bas. En effet, ce contexte favorise les ménages à s'orienter vers des produits présentant des rendements plus intéressants (ou du moins pousse les assureurs à proposer des stratégies d'investissement vers les fonds en unités de compte). Depuis 2019, les taux OAT sur dix ans fluctuent autour de zéro. Or les investissements sur les fonds en euros sont placés principalement sur des obligations d'entreprise ou d'État par les assureurs, le reste étant investi dans des actions et de l'immobilier. Ainsi, avec des obligations d'État sur 10 ans à taux négatifs et le maintien des taux de revalorisation minimums sur des contrats d'assurance-vie en euros, les fonds en euros sont devenus peu rentables.

D'autre part, l'intérêt des ménages pour les fonds en unités de compte est intrinsèquement lié aux performances des marchés actions. En 2021 le rebond des marchés financiers a pu être observé sur les contrats en unités de compte par une hausse de la collecte brute.

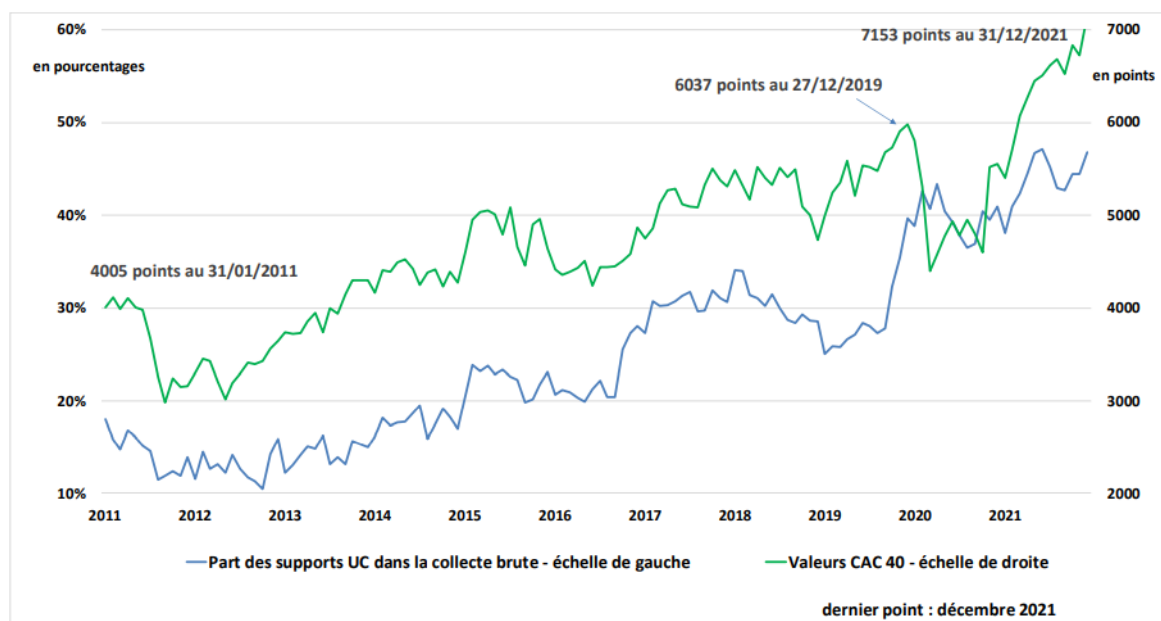


FIGURE 1.4 – Corrélation de l'investissement vers les unités de compte et les performances du CAC40 (Source : ACPR)

Fin 2021, avec un rebond de 30 % comparé au début d'année, le CAC40 atteint un niveau record depuis 1999 dépassant 7 000 points. Ainsi, la part des supports en unités de compte surperforme ses niveaux de collecte et atteint le taux de 47 %. A l'exception de 2020, l'intérêt pour les unités de compte semble corrélé aux performances des marchés financiers. En effet, investir dans des supports en unités de compte équivaut à investir indirectement sur des actions, obligations, fonds de placements ou encore

valeurs monétaires. Par conséquent, les performances de ces supports sont soumises aux fluctuations des marchés financiers.

Après avoir présenté les avantages de l'assurance-vie, les principes clés des contrats d'assurance-vie sont détaillés dans la section suivante.

1.2 Principe du contrat d'assurance-vie

L'assurance-vie est un support d'investissement qui permet de constituer une épargne qui sera versée en cas d'événements aléatoires dépendants de la durée de vie de l'assuré. Il existe des garanties en cas de vie, en cas de décès, ou mixte. L'épargne constituée est versée sous forme de capital ou d'une rente à l'assuré en cas de vie au terme du contrat, ou aux bénéficiaires en cas de décès avant le terme.

Aujourd'hui l'assurance-vie est constituée d'une majorité de contrats en cas de vie avec une garantie décès. On distingue en assurance-vie individuelle :

- les **contrats d'épargne** permettant de récupérer l'épargne investie revalorisée à l'échéance choisie. En cas de décès, le capital est versé aux bénéficiaires désignés ;
- les **contrats d'épargne retraite supplémentaire** ayant pour principal objectif d'apporter une compensation sous forme de capital ou de rente en vue de la retraite, en contrepartie de versement de prime en amont. En cas de décès, le capital est aussi versé aux bénéficiaires désignés.

Le montant versé au client dépend de plusieurs paramètres, principalement du montant des primes investies et du type de support utilisé, mais aussi du type de gestion choisi, des frais appliqués et des options souscrites. L'assurance-vie est donc très modulable en fonction du profil de risque de l'assuré.

1.2.1 La gestion du contrat

Les versements

Un contrat d'assurance peut être alimenté de trois manières :

- le versement initial marque l'ouverture du contrat. Il est unique et est défini au moment de la souscription ;
- les versements libres : au cours du contrat le souscripteur peut choisir d'approvisionner ses fonds à tout moment ;
- les versements périodiques : les montants et la périodicité (mensuel, trimestriel, biannuel ou annuel) sont fixés en amont. Certains contrats d'épargne ont contractuellement une obligation de versements périodiques avec un montant annuel investi minimal.

Le montant des versements dépend de multiples facteurs propres à chaque client tels que l'âge, l'ancienneté du contrat, les niveaux de revenu, l'environnement économique, etc. L'objectif de ce mémoire est donc de déterminer le comportement du client sur le versement régulier de prime et les variables qui influeraient sur le montant versé.

Le rachat

Le rachat est différent selon le type de contrat. Sur les contrats d'épargne, l'assuré a la possibilité, à tout moment, de récupérer une partie de l'épargne investie (acte qualifié de rachat partiel) ou bien sa totalité (rachat total). Ce dernier entraîne la fermeture du contrat et la perte de l'ancienneté associée au contrat. Tel qu'évoqué dans la section 1.1.2, tout rachat entraîne l'application de la fiscalité sur les

plus-values acquises. Alors que sur les contrats d'épargne retraite, l'épargne est en principe indisponible jusqu'au départ à la retraite. Il est possible de débloquer l'épargne de manière anticipée dans les cas suivants : le décès du conjoint ou du partenaire de PACS, l'invalidité (du titulaire, de son conjoint marié ou pacsé, de ses enfants), en situation de surendettement, à l'expiration des droits au chômage, cessation d'activité non salariée à la suite d'un jugement de liquidation judiciaire ou dans le cas de l'acquisition de la résidence principale (sauf pour les droits issus de versements obligatoires).

Les rachats impactent fortement la rentabilité d'un produit. Leur modélisation a fait le sujet de plusieurs mémoires. En effet, un rachat proche de la date de souscription du contrat implique des pertes pour l'assureur. Un nouveau contrat génère des coûts importants la première année nécessitant de garder le contrat le plus longtemps possible afin d'amortir les coûts sur plusieurs années (de rémunération des apporteurs de l'affaire, des commissions prélevées, du capital immobilisé).

Aussi, la fluctuation des rachats représente un risque de liquidité pour l'assureur. Celui-ci place une majorité des primes perçues sur des actifs afin de pouvoir servir des intérêts sur le capital déposé par l'épargnant. Or, une partie des rachats est liée à la dynamique des taux d'intérêts. Le mémoire de Nana Njoya (2016) [20] détaille l'impact de la demande de rachat en fonction de la fluctuation des taux d'intérêt :

- En cas de hausse des taux d'intérêt, la valeur des obligations baisse. Cette dégradation de valeur affecte le rendement des produits d'assurance-vie composés d'obligations d'États. Ainsi, face à de meilleures opportunités, les assurés sont incités à demander le rachat de leur contrat. Une demande massive de rachats dans ce contexte peut conduire l'assureur à revendre ses obligations en moins-values latentes si les provisions sont insuffisantes pour faire face à la demande de rachat.
- En cas de baisse des taux d'intérêt, l'assureur doit servir un rendement minimal avec un portefeuille d'actifs dégradés. Si les rachats ne sont pas suffisants, il est possible que les actifs ne puissent plus générer les rendements nécessaires pour couvrir les engagements de l'assureur, ce qui crée un risque de liquidité.

Les arbitrages

Un arbitrage consiste à réallouer de l'épargne investie sur un support vers un autre support financier. Cette réorientation résulte d'un changement stratégique ayant pour principal objectif de sécuriser et de dynamiser l'épargne face aux fluctuations des marchés, afin d'optimiser le rendement de l'assurance-vie en fonction du profil de risque de l'assuré et des opportunités des marchés. En général, des frais sont appliqués sur chaque transfert.

Les types de gestion

Il existe pour chaque produit, en fonction des objectifs choisis, différents types de gestion de portefeuille :

- Gestion libre : s'adresse aux assurés avec une certaine connaissance des marchés financiers, elle permet d'avoir la charge de leurs investissements et de choisir personnellement l'allocation des versements sur les différents supports.
- Gestion sous mandat également nommée gestion pilotée : l'assuré délègue la gestion de ses investissements à un gestionnaire qui effectue la sélection des supports et les réorientations d'épargne. Généralement l'assuré a le choix entre plusieurs profils de gestion définis à partir de catégories de supports et d'une exposition plus ou moins importante aux risques. Ce choix de gestion implique des frais supplémentaires prélevés mensuellement sur l'encours.
- Gestion par convention : permet une diversification de l'épargne sur plusieurs supports prédéfinis selon une clé de répartition. Chaque année civile, la totalité de l'épargne est réajustée selon la répartition de la convention choisie, de même lors de chaque nouveau versement net de frais.
- Gestion évolutive par horizon : l'assuré choisit un horizon d'investissement, autrement dit la clé de répartition est dépendante du temps (en fonction de l'âge de l'assuré ou de l'ancienneté du contrat). Cela permet de privilégier des investissements long terme orientés plutôt vers des actions pour les

profils d'assurés jeunes puis proposer une répartition plus prudente composée majoritairement de fonds en euros à partir d'un certain âge.

Les frais

Tout au long de la durée du contrat, les frais suivants sont prélevés sur les montants investis :

- les frais à l'acquisition : appliqués sur le versement initial et pouvant être différents des frais sur les versements complémentaires ;
- les frais sur les versements : prélevés sur chaque prime brute versée par l'assuré ;
- les frais sur encours : frais de gestion différents en fonction des supports qui viennent diminuer la rentabilité du fonds associé ;
- les frais d'arbitrage : applicables lors de réallocation des montants entre deux supports ;
- les frais d'arrérages : en cas de sortie en rente viagère, des frais sont prélevés sur chacun des arrérages.

Ces frais permettent de rémunérer l'apporteur de l'affaire, l'assureur mais aussi de pouvoir compenser les intermédiaires et les frais des services proposés. Ces frais peuvent jouer sur le comportement de versement du client si la rentabilité des investissements n'est pas suffisante face aux frais encourus.

1.2.2 Les types de support

Les placements sur des produits d'assurance-vie peuvent être investis sur différents types de supports. Il existe les fonds en euros, en unités de compte et les fonds eurocroissance. L'investissement sur ces supports ne présente pas les mêmes revalorisations.

Support en euros

Les montants investis sur les fonds en euros sont garantis à tout moment et les intérêts sont acquis une fois versés. Le taux de revalorisation des contrats est constitué du rendement garanti (au minimum 0 % sur ce support) et de la participation aux bénéfices au titre de l'exercice N.

Le taux technique assimilé au taux de rendement minimum sur lequel s'engage un assureur est limité par la réglementation (article A. 132-1 du Code des assurances [18]). En effet, il peut constituer une contrainte importante pour les assureurs, l'objectif de la réglementation étant d'encadrer l'utilisation des taux servis excessifs face à la situation des marchés financiers.

Il est possible de définir contractuellement un Taux Minimum Garanti (TMG) de rendement de l'épargne associé au fonds. Il représente le taux total de rémunération. Ce taux est brut de frais de gestion et de prélèvements sociaux et fiscaux. Le taux technique et éventuellement la part de participation aux bénéfices servis ne peuvent être inférieurs au TMG (Article A. 132-2).

Aujourd'hui les assureurs privilégient le Taux Minimum Garanti Annuel (TMGA) qui est plus flexible pour les assureurs ayant la possibilité de le réviser chaque année. De la même manière, cette revalorisation est encadrée par la loi.

Enfin, la participation aux bénéfices est déterminée en fonction des résultats techniques et financiers de l'exercice en respectant une distribution minimale à hauteur de 85 % des bénéfices financiers et 90 % des résultats techniques. La participation aux bénéfices peut ne pas être distribuée immédiatement. Le reliquat des bénéfices est affecté à la Provision pour Participation aux Excédents (PPE) permettant de constituer une réserve pour les résultats variables futurs. Cette réserve permet de lisser les rendements des contrats au cours du temps. Néanmoins, au bout de huit ans la PPE doit être intégralement redistribuée.

Support en unités de compte

Un support en unités de compte garantit seulement le nombre d'unités et non la valeur de l'unité. Chaque versement, net de frais, est converti en nombre d'unités de compte calculé en fonction de la valeur liquidative du fonds d'investissement au moment de l'opération. Les unités de compte sont, d'après l'article L.131-1 du Code des assurances, « constituées de valeurs mobilières ou d'actifs offrant une protection suffisante de l'épargne investie et figurant sur une liste dressée par décret en Conseil d'État ». L'article R. 131-1 fixe cette liste qui est composée d'un ensemble de produits financiers tels que :

- des actions ;
- des obligations d'entreprise ou d'État ;
- des valeurs mobilières (SICAV, FCP) ;
- des supports immobiliers (SCPI, SCI).

L'assureur n'a aucune obligation de revalorisation du portefeuille sur les unités de compte. Le risque est porté par l'assuré. L'épargne investie suit donc l'évolution de la valeur des actifs sous-jacents à la hausse comme à la baisse en fonction de la fluctuation des marchés financiers. Ainsi, la performance du fonds est potentiellement plus importante mais avec un risque de moins-values à l'échéance du contrat.

Support eurocroissance

Enfin le dernier support sur lequel il est possible d'investir est le fonds eurocroissance. Ce fonds a été mis en place à partir de 2014 assurant au terme fixé (durée de huit ans au minimum) une garantie totale ou partielle du capital net investi. Les investissements sont placés sur des actifs plus risqués que le fonds en euros. L'eurocroissance est donc un mixte entre un support euros et en unités de compte. Avant l'échéance de la garantie du fonds, les montants investis sont soumis aux fluctuations des marchés financiers et ne sont pas garantis. Il est toujours possible de racheter le capital placé sur le support eurocroissance avant le terme fixé, seulement le capital n'est pas garanti. Les produits d'assurance-vie AXA se sont transformés au fur et à mesure afin de pouvoir proposer un support eurocroissance. Ce support est plus flexible pour les assureurs : dans un premier temps une partie de l'épargne est placée sur des supports plus risqués et avec une espérance de rendements meilleure, puis à l'approche du terme les investissements sont sécurisés. Du point de vue client, ce support permet une perspective de rendement supérieur à celui du support euros par la diversification des investissements avec une garantie du capital net investi au terme fixé. Aujourd'hui malgré ces caractéristiques, le fonds eurocroissance reste un support peu utilisé par les épargnants.

1.3 Contexte

1.3.1 L'environnement économique

Le marché de l'assurance-vie évolue aujourd'hui dans un contexte de remontée des taux d'intérêts après quelques années marquées par des taux bas voire négatifs.

Pour la première fois en 2019 les taux d'emprunts d'État de la France (OAT) 10 ans sont devenus négatifs. Les obligations d'État français ne concernent pas directement les épargnants. En revanche, l'évolution des taux d'intérêts (qui a une incidence sur la rentabilité des contrats des assurés) repose en partie sur des obligations d'États dont la dette française. Les rendements des fonds en euros établis principalement sur des obligations souveraines ont été particulièrement atteints, entraînant plus de difficultés pour les assureurs d'honorer les taux minimums garantis. Cet environnement économique a poussé les assureurs et gestionnaires à revoir leur stratégie d'investissement. Une solution pour maintenir le taux minimum garanti est de distribuer une part plus importante de participation aux bénéficiaires.

Cela implique de diminuer les provisionnements sur la provision pour participation aux excédents pour les années futurs. Pour les nouveaux contrats et ceux dont le taux minimum garanti n'avait pas été fixé contractuellement, les assurances diminuent les taux techniques proposés. Aussi, les investissements peuvent se tourner vers des actifs plus risqués mais demandent, par la réglementation en vigueur sur la solvabilité, une mobilisation de capital plus importante. Enfin, la stratégie majeure mise en place est de réorienter les investissements vers les unités de compte.

Depuis fin mars 2021, l'OAT à 10 ans a franchi à nouveau à la hausse le seuil de 0 %. Après une période d'oscillation autour de 0 %, l'OAT 10 ans ne cesse d'augmenter. Cette tendance observée sur les taux d'intérêt a été accentuée en 2022, accompagnée par une hausse de l'inflation. Plusieurs facteurs sont responsables de l'augmentation des prix du marché. En premier lieu, cela peut être attribué aux répercussions de la pandémie qui ont entraîné des perturbations dans les chaînes d'approvisionnement à l'échelle mondiale en raison de la fermeture des frontières. De plus, l'inflation a été alimentée par la guerre en Ukraine. La crise entre ces deux pays aux ressources importantes en céréales ou gaz a créé une pression inflationniste et un stress sur les marchés financiers. Les banques centrales ont annoncé la remontée des taux pour contenir l'inflation.

Cette remontée des taux implique un risque de liquidité pour les organismes d'assurance en cas de rachats en masse. En effet, si le taux servi est inférieur aux taux proposés par la concurrence dans un contexte qui lui est plus favorable, l'épargnant cherchera à racheter son contrat. Bien que la vente des obligations avant leur remboursement pour pouvoir faire face aux engagements conduirait à réaliser des moins-values. Les rachats semblent être en partie limités par la fiscalité sur les contrats d'assurance-vie de moins de huit ans, les pénalités dissuadant les rachats des clients sur leur épargne.

De plus, l'inflation peut avoir des conséquences sur les marchés financiers. L'augmentation des taux d'intérêt par les banques centrales en réponse à la progression de l'inflation se répercute sur les entreprises à plusieurs niveaux. Tout d'abord, les entreprises subissent directement la hausse des prix, des salaires et des coûts d'emprunt, cela entraîne une baisse de leurs marges bénéficiaires et des dividendes versés aux investisseurs. En outre, l'augmentation des taux d'intérêt se traduit par une hausse des rendements obligataires, ce qui peut amener les investisseurs à privilégier les obligations au détriment des actions. Ainsi, en théorie, une menace d'inflation sur les marchés financiers peut entraîner une dépréciation de la valeur des actions ainsi que des fonds en unités de compte. Dans les faits, si l'inflation est maîtrisée, les entreprises maintiennent plus facilement leurs marges et les actifs progressent. Néanmoins, la collecte en assurance-vie risque d'être affaiblie par les changements de comportement des ménages français sur l'épargne en réponse à l'augmentation des prix étant donné que les salaires ne suivent pas nécessairement la même tendance.

1.3.2 L'impact de la réglementation : la loi PACTE

Entré en vigueur le 22 mai 2019, le Plan d'Action pour la Croissance et la Transformation de l'Entreprise (ou loi PACTE) a pour principal objectif d'aider à la transformation et à la croissance des entreprises en orientant les investissements vers leur développement. La loi PACTE entraîne des répercussions en assurance-vie sur plusieurs aspects. Le but de cette réforme est de simplifier l'offre d'épargne en assurance. Cette section détaillera de manière succincte et non exhaustive les changements apportés par cette loi. En effet, ce mémoire n'a pas pour objet d'étudier les conséquences de la loi PACTE en assurance-vie mais il est important de poser ce contexte qui a pu influencer les comportements clients sur la gestion de leurs contrats, ou bien favoriser le développement de certains produits au détriment d'autres.

Concernant les produits d'épargne retraite supplémentaire : l'épargne retraite a été simplifiée à travers la réforme de ses produits se résumant depuis par un unique produit d'épargne retraite : le PER (Plan d'Épargne Retraite) [19].

Les contrats PER sont disponibles depuis le 1er octobre 2019. L'épargne investie sur les anciens produits (PERP, Madelin, Article 83 ou PERCO) mais aussi depuis un contrat d'assurance-vie peut être

transférée vers un contrat PER. A partir du 1er octobre 2020, la commercialisation des anciens produits d'épargne retraite n'est plus autorisée. Seuls les versements sur des contrats déjà existants sont encore possibles, à l'exception des anciens produits d'épargne qui ont été mis à jour avant cette date afin d'être conformes aux nouvelles règles du PER.

Pour inciter les transferts, la loi prévoit jusqu'en 2023 des avantages fiscaux supplémentaires en cas de transfert d'épargne d'un produit d'assurance-vie vers un PER. Pour tout transfert d'un contrat d'assurance-vie de plus de huit ans, l'abattement fiscal est doublé (soit 9 200 € d'exonération sur les plus-values pour une personne seule et 18 400 € pour un couple) à condition que le rachat soit effectué au moins cinq ans avant le passage à la retraite. Les frais de transfert sont aussi limités par la loi. La portabilité des contrats dynamise l'épargne retraite et favorise la compétitivité entre assureurs.

Concernant les produits d'épargne en assurance-vie : en plus de la portabilité de l'épargne vers un PER détaillée précédemment, la loi PACTE cherche à moderniser le support eurocroissance. D'après la Fédération Française de l'Assurance (FFA) dans un bilan en 2017 seulement 0,13 % de l'encours total d'assurance-vie est orienté vers des fonds eurocroissance. L'objectif de la loi est de rendre ce support plus attractif afin d'orienter l'épargne accumulée sur le fonds en euros vers un support avec une part plus importante d'actifs risqués permettant de contribuer au financement des entreprises françaises.

Deux mesures sont proposées : un rendement unifié pour les épargnants à l'image du fonds en euros et une bonification des investissements les plus longs. Ainsi, des contrats avec la même maturité d'engagement auront les mêmes taux de rendement. En revanche, si le contrat bénéficie de la bonification, une différence de rendement est perçue pour des contrats avec des dates d'investissement différentes. Pour des maturités au dessus de 8 à 10 ans, un pourcentage plus important pourra être proposé pour la revalorisation du contrat.

En conclusion, sur le secteur de l'assurance, la loi PACTE a pour objectif d'orienter l'encours de l'épargne vers les produits de retraite supplémentaire et des supports plus diversifiés en assouplissant son cadre réglementaire.

1.3.3 Littérature actuarielle sur la modélisation des versements

Plusieurs études sur le comportement client face à la gestion de son épargne en assurance-vie ont été menées. Néanmoins, elles portent principalement sur les lois de rachats, les lois d'arbitrages et sur les versements libres. A ce jour, aucun sujet ne se focalise entièrement sur la modélisation des versements périodiques. La recherche porte principalement sur les versements libres car leur occurrence et leur montant versé sont plus volatiles. En effet, la modélisation des versements libres est plus difficile, car elle dépend du libre-arbitre de chaque assuré contrairement aux versements périodiques définis contractuellement sur une période donnée (l'assuré conserve la possibilité de modifier ou d'arrêter ses versements réguliers à tout moment). De manière générale, les primes périodiques sur un contrat d'épargne marquent la volonté d'investir sur le long terme de manière régulière, ce qui en fait une variable assez stable, donc plus facile à prédire. Certes moins volatiles, les primes périodiques en assurance-vie peuvent néanmoins fluctuer. Ce sont ces fluctuations que ce mémoire cherche à évaluer.

Plusieurs approches ont été utilisées pour estimer les versements libres. Les montants de versements libres ont été modélisés par Dieltiens (2021) [11] à travers une première approche utilisant les séries temporelles, avant de proposer une modélisation par *Machine Learning* qui introduit des variables explicatives et permet d'améliorer la prédiction des montants. D'autres auteurs se sont attachés à estimer les versements libres à partir de variables structurelles ou conjoncturelles en distinguant la fréquence du montant de versements. Les mémoires de Benabdelkrim (2017) [7] et Assaraf (2020) [5] s'attachent à modéliser par des méthodes de *Machine Learning* les versements libres en fonction de variables structurelles formées des spécificités du client ou encore du contrat. Le premier mémoire porte sur la probabilité d'effectuer un versement. Le deuxième mesure l'impact de l'intégration des versements libres sous IFRS 17 en modélisant la fréquence par *Machine Learning* et les montants moyens de versement

par *clustering*. Ces travaux ont permis de mettre en évidence l'importance de l'ancienneté, des niveaux de provision mathématique et des taux de frais d'acquisition et de gestion. De plus, leurs conclusions soulignent que les indicateurs structurels ne sont pas suffisants et insistent sur l'importance d'intégrer l'environnement économique dans lequel sont effectués les versements.

Le mémoire d'Andre (2019) [3], en plus d'étudier les variables structurelles pour prédire le taux et les montants de versements libres, met en avant l'importance des conjonctures économiques et propose une méthodologie de construction d'une loi conjoncturelle en complément du taux structurel modélisé. Aussi, le mémoire de Feniza (2019) [12] s'attache à la modélisation des arbitrages à partir de facteurs conjoncturels. Ainsi, à travers des séries temporelles, il introduit une dimension temporelle pour estimer les variations à court terme des montants arbitrés.

Enfin, la rentabilité des produits d'épargne a fait l'objet de nombreuses études, en particulier Chaudhry (2019) [10] s'intéresse à l'impact de la loi PACTE sur la rentabilité du PER individuel d'AXA France en effectuant des sensibilités sur les taux de sorties, de rachat, de primes et la clause de participation aux bénéfices. Dans la dernière partie de ce mémoire, seul l'impact de la fluctuation des primes sur les indicateurs de rentabilité est analysé.

Chapitre 2

Étude des données

L'objet de ce chapitre est de présenter les produits sur lesquels l'étude porte. La base de données utilisée sera analysée avant de pouvoir effectuer les retraitements, qui seront par la suite nécessaires à l'exploitation des variables dans la modélisation. Des analyses bivariées et de corrélations seront également étudiées afin de donner une première perspective des tendances entre les différentes variables qui pourraient expliquer les variations des flux de versements périodiques.

2.1 Présentation des données à disposition

2.1.1 Périmètre et agrégation des données

Le périmètre d'étude porte sur des produits d'épargne et de retraite individuelle commercialisés par Axa France et souscrits par des personnes physiques. La base utilisée est restreinte aux données de trois produits entre 2017 et 2021, soit sur une durée de cinq ans. Seuls les contrats pour lesquels les états sont en cours ou réduit, autrement dit qui sont susceptibles d'effectuer des versements, sont analysés.

Les trois produits étudiés ont les caractéristiques suivantes :

- Le **produit 1** est un contrat d'assurance-vie multisupports ouvert à la commercialisation depuis 2006. L'adhésion est sous condition d'un versement initial minimal et la gestion est unique pour toute l'épargne. Le souscripteur a le choix entre la gestion personnelle, la gestion par convention ou la gestion évolutive par âge. L'assuré a la liberté d'effectuer ou non des versements périodiques.
- Le **produit 2** est un contrat d'assurance-vie multisupports ouvert à la commercialisation depuis 2015. L'adhésion est sous condition d'un montant minimum annuel à verser et le choix de la fréquence des versements reste libre. Trois modes de gestion par convention sont proposés et deux gestions de type évolutive.
- Le **produit 3** est un contrat d'épargne retraite supplémentaire à adhésion facultative, fermé à la commercialisation depuis 2019. Seule la gestion évolutive est proposée sous l'option à horizon ou par âge.

Les données utilisées ont été récupérées à partir des informations propres au client, des flux du contrat et des caractéristiques du contrat. La base a été construite par agrégation des flux sur une année d'exercice par numéro de contrat. Chaque ligne représente donc les flux annuels sur un contrat avec les caractéristiques de l'individu et du contrat associé. Le choix d'une modélisation des montants de versements annuels se justifie par la méthode de projection des flux de la maquette utilisée en rentabilité. En effet, les flux sont projetés annuellement sur 60 ans, les lois de versement doivent être utilisables à un pas de temps annuel.

Dans la suite de l'étude, le numéro de contrat n'est pas conservé. Autrement dit, l'information passée des versements effectués n'est pas disponible. Néanmoins, la variable sur l'encours du contrat regroupe les montants totaux versés et non rachetés. Cela peut donner une indication sur le niveau d'activité de versement de l'assuré. L'intérêt est de se focaliser sur les caractéristiques de l'assuré et du contrat au moment du versement pour évaluer le volume des flux, sans prendre en compte le comportement passé de l'assuré.

La base de données est composée en grande partie de lignes concernant le produit 1 puis un tiers de données sur le produit 3, le reste étant constitué de contrats sur le produit 2.

Produit 1	Produit 2	Produit 3
48,3 %	19,4 %	32,2 %

TABLE 2.1 – Répartition des produits dans la base

Les contrats qui ne sont plus susceptibles d'effectuer des versements en 2021 ne sont pas pris en compte. Le nombre de contrats augmente donc d'une année sur l'autre en fonction des nouvelles souscriptions. L'augmentation est plus ou moins marquée en fonction du produit considéré.

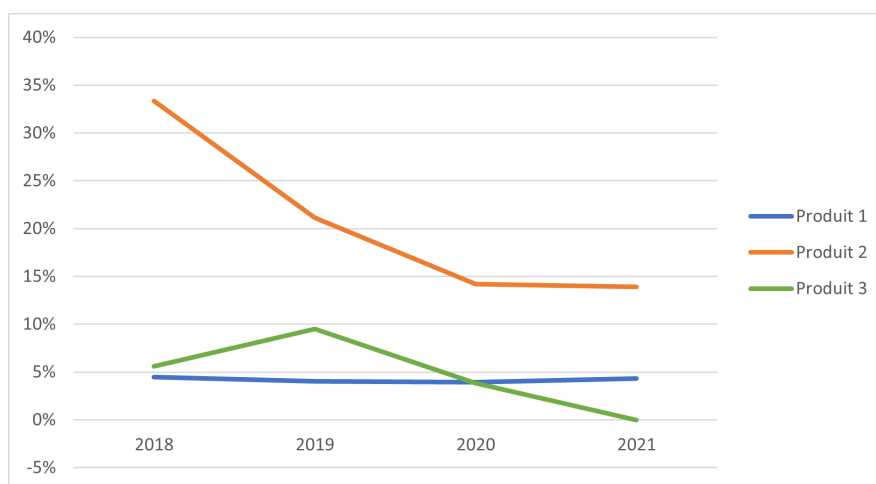


FIGURE 2.1 – Évolution année à année du nombre de contrats entre 2017 et 2021

Le produit 1 suit une augmentation linéaire de 5 % du volume de contrats chaque année. Après une forte commercialisation du produit 2, l'augmentation des nouvelles souscriptions ralentie et semble se stabiliser depuis 2020 à 15 %, ce qui reste un taux important. En effet, ouvert depuis six années, le produit 2 semble être en pleine phase de commercialisation. Enfin, concernant le produit 3, l'année de 2019 a été marquée par un dernier regain de nouvelles souscriptions avant la fermeture du produit à la commercialisation. En 2021 le nombre de contrats du produit 3 n'évolue pas par rapport à l'année 2020.

2.1.2 Présentation des variables

Cette partie présente en détail les données accessibles pour la suite de l'étude ainsi que les différentes variables qui seront potentiellement déterminantes dans l'estimation des versements périodiques.

Données du client

Les variables explicatives suivantes sont extraites de la base de données client. L'information la plus récente est gardée pour tout l'historique d'observation (hors l'âge qui est recalculé à chaque année d'exercice).

- Genre : indique si la personne physique est un homme ou une femme. La répartition est présentée ci-dessous :

	Produit 1	Produit 2	Produit 3
Femme	50 %	52 %	50 %
Homme	50 %	48 %	50 %

TABLE 2.2 – Répartition des individus selon leur genre

Dans l'ensemble, la répartition des hommes et des femmes est équivalente pour chacun des produits.

- Âge : calculé par différence de millésime entre la date d'anniversaire de l'assuré et l'année d'exercice N.

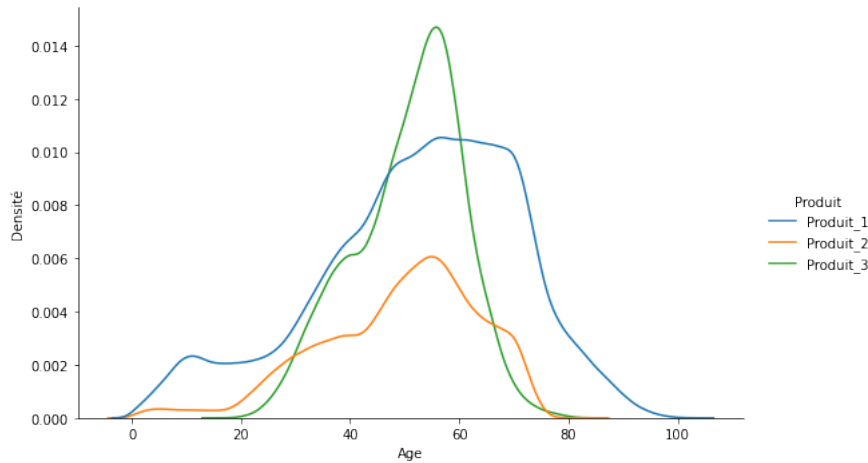


FIGURE 2.2 – Répartition des assurés par âge

Le portefeuille est constitué d'assurés entre 0 et 103 ans avec une moyenne d'âge à 53 ans. La majorité de l'effectif est incluse dans la classe d'âge 40-60 ans. Les produits 2 et 3 ont très peu d'individus de plus de 80 ans.

- Situation familiale : définie parmi les modalités suivantes *Célibataire*, *Concubinage*, *Divorcé(e)*, *Marié(e)*, *Pacsé(e)*, *Séparé(e)*, *Veuf(Ve)* et *Autre*.
- Code postal : le portefeuille présente 6 607 communes différentes.
- CSP : correspond à la Catégorie Socio-Professionnelle suivant la nomenclature définie par l'INSEE. Chaque profession est regroupée suivant huit catégories de métiers.

Code CSP	Catégorie
1	Agriculteurs exploitants
2	Artisans, commerçants, chefs d'entreprise
3	Cadres et professions intellectuelles supérieures
4	Professions intermédiaires
5	Employés
6	Ouvriers
7	Retraités
8	Inactifs
9	Non déterminés

TABLE 2.3 – Catégories des professions

- Profession : complète le code CSP avec le libellé complet de la profession.

Données du contrat

Cette section présente les variables propres au contrat.

- Réseau : correspond au type de réseau apporteur de l'affaire qui peut être réalisée par des agents généraux, à travers le réseau de courtage ou bien le réseau des Départements et Territoires d'Outre-Mer.
- Code gestion : fait référence au mode de gestion choisi pour le contrat. Les types de gestion ont été regroupés suivant les catégories de gestion : personnelle, sous convention, sous mandat, par horizon (répartition des versements selon un horizon d'investissement), par horizon par âge (évolution de la répartition des investissements en fonction de l'âge).
- Ancienneté : cette variable représente la différence de millésime entre la date de début d'effet du contrat et l'année d'exercice considéré. Ainsi, les contrats d'ancienneté nulle indique qu'ils ont été conclus en cours d'année. Le produit 3 étant le produit le plus longtemps commercialisé, il présente des années d'ancienneté jusqu'à 17 ans. Un creux au niveau des anciennetés 8 à 11 ans peut être observé sûrement dû à une diminution de commercialisation au profit du produit 1. Les contrats souscrits il y a dix ans se révèlent être l'ancienneté la plus représentée sur le produit 1. Cela est dû à une période forte de croissance sur les cinq premières années du produit. Enfin, le produit 2 a été commercialisé depuis moins longtemps et semble maintenir un rythme plus constant de nombre de contrats sur les trois dernières années.

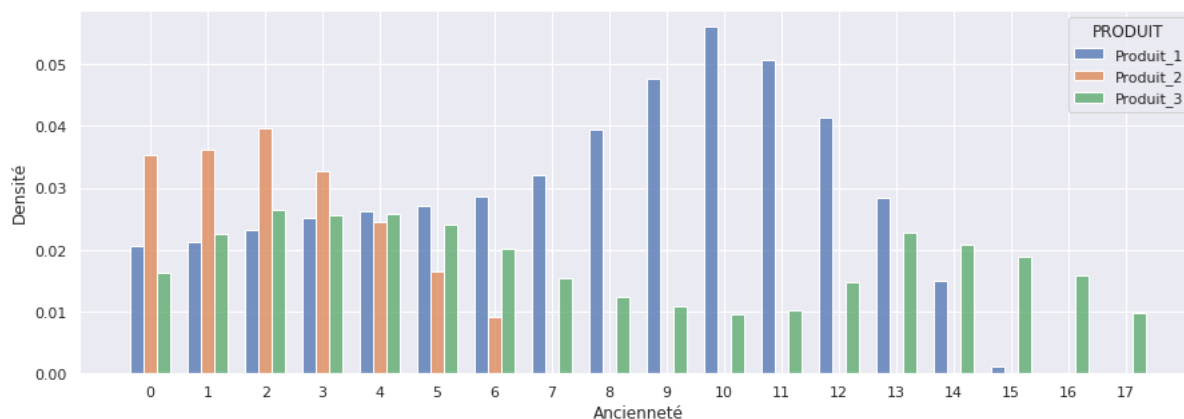


FIGURE 2.3 – Distribution de l'ancienneté par produit

- Provision mathématique (PM) : représente l'encours total du contrat au 1^{er} janvier de l'année d'exercice en fonction du support euros, UC, eurocroissance. L'encours est fixé au début d'année pour éviter d'inclure les versements effectués au cours de l'année. La répartition des PM par support est constante dans le temps quelque que soit le produit. Le fonds en euros est le support privilégié par rapport aux autres supports. Le fonds eurocroissance représente moins de 0,5 % de l'encours.

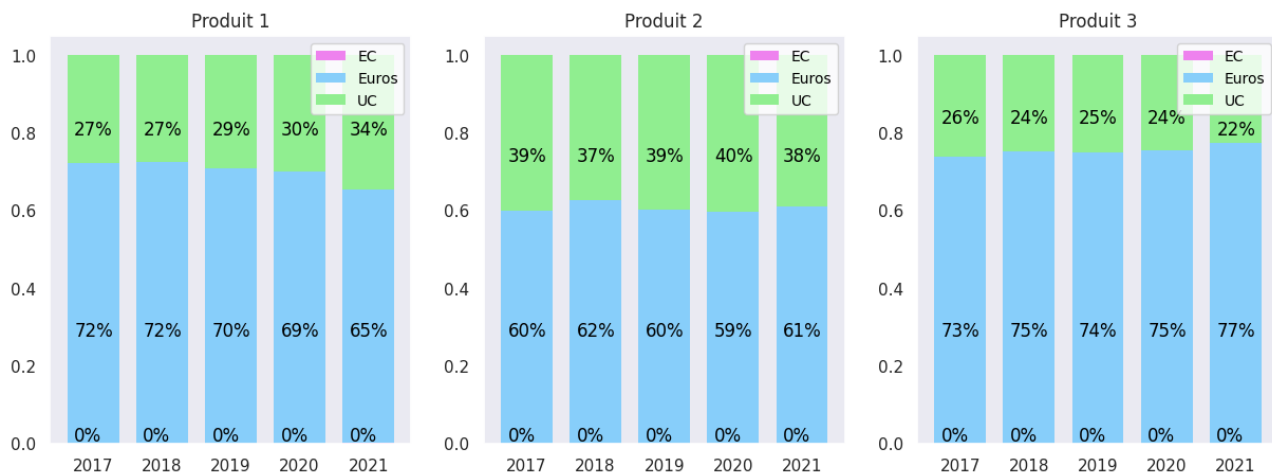


FIGURE 2.4 – Évolution de la part des PM par support

- Versements : trois variables indiquent le montant respectivement des versements initiaux (VI), libres (VL) ou périodiques (VP) annualisés.

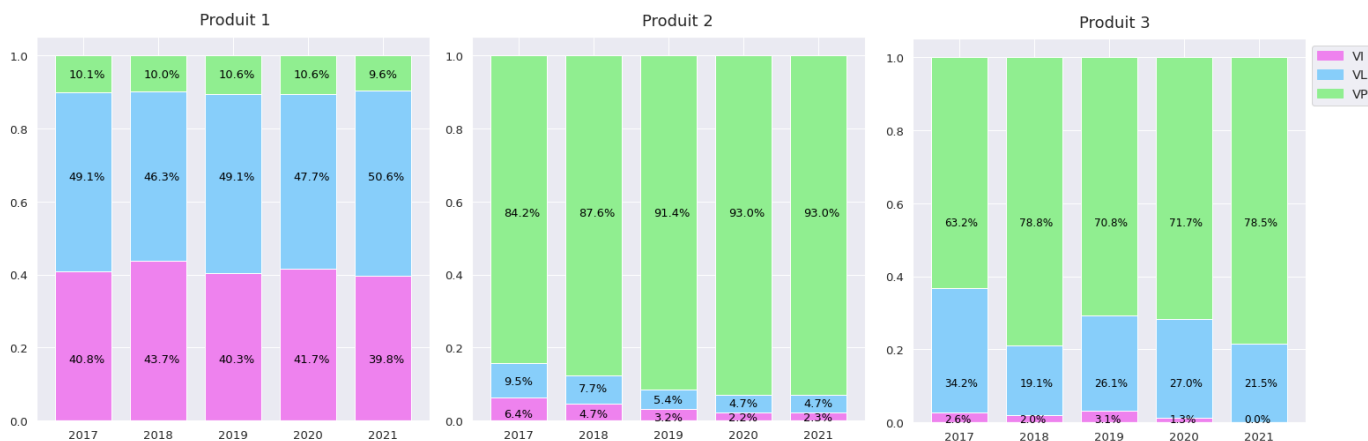


FIGURE 2.5 – Évolution de la proportion des montants de versements par type

Cette donnée est complétée par trois variables de comptage. La fréquence de versement par année par produit a été déterminée par des indicatrices représentant s'il y a eu au moins un versement au cours de l'année ou non.

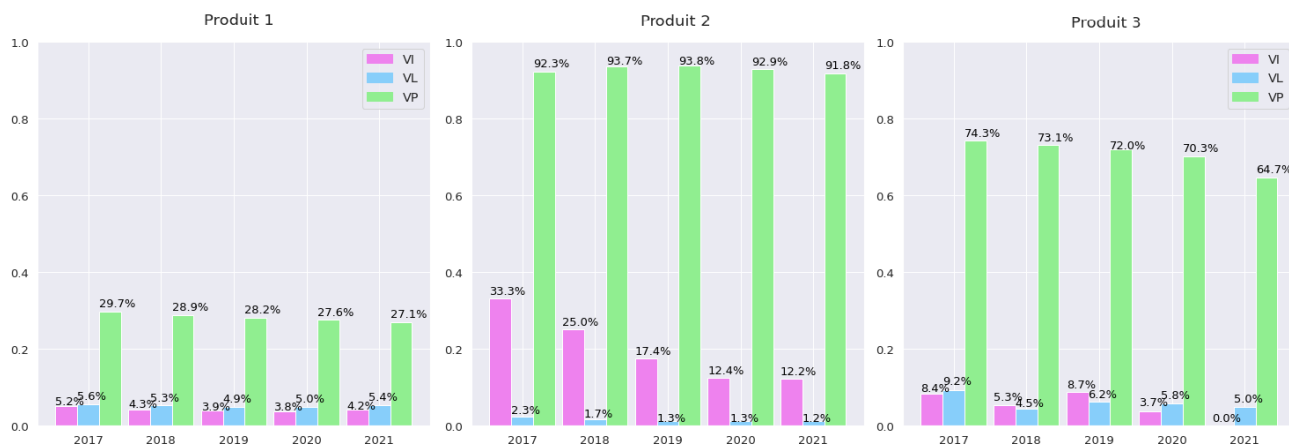


FIGURE 2.6 – Évolution de la part des contrats effectuant des versements par type

Les figures 2.5 et 2.6 mettent en évidence l'importance de modéliser les comportements client sur les versements en fonction des produits. En effet, les conditions générales du produit imposent ou non le versement d'un montant par an, ce qui influe sur la part des montants des versements périodiques sur le portefeuille. Ainsi, sur une année autour de 10 % des montants de versements sont périodiques concernant le produit 1 contre une part à plus de 70 % pour les produits 2 et 3.

En particulier, pour le produit 1 la répartition des montants est stable sur la période d'observation avec 10 % de versements programmés, 50 % de versements libres et 40 % de versements liés à l'ouverture de nouveaux contrats. Un peu moins de 30 % des assurés effectuent de façon périodique des versements (avec une légère diminution depuis cinq ans), 5 % de façon ponctuelle et autour de 4 % des contrats sont nouveaux. Bien que la proportion des contrats effectuant périodiquement des versements soit plus élevée que ceux de manière ponctuelle, les montants sont plus volumineux sur les versements libres.

Concernant le produit 2, les versements sont essentiellement constitués de versements périodiques : près de 93 % des contrats en effectuent. Cela représente entre 84 % à 93 % des montants des primes entre 2017 et 2021. Comme observé sur l'évolution du nombre de nouveaux contrats précédemment, la part des versements lors de la souscription diminue durant la période d'observation. Moins de 2 % des contrats effectuent des versements à titre ponctuel, les versements périodiques sont privilégiés.

Au sujet du produit 3, la fin de commercialisation de ce produit est surtout visible sur les versements initiaux qui deviennent nuls en 2021. Les montants de versements sont principalement tournés vers des versements périodiques à hauteur de 63 % jusqu'à 93 %. Les versements libres sont plus nombreux que sur le produit 2.

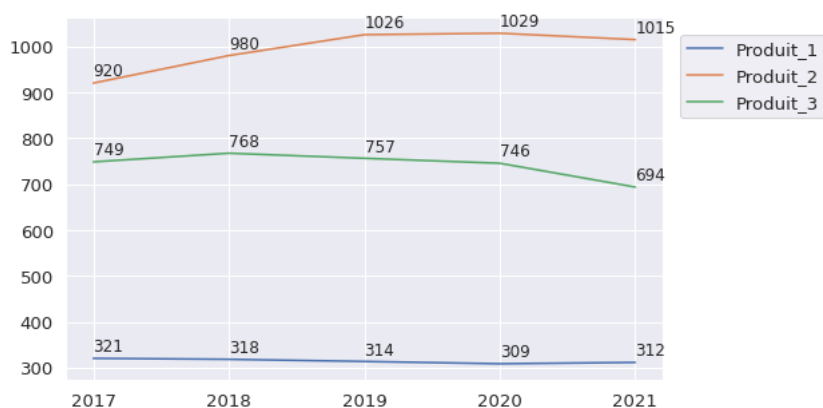


FIGURE 2.7 – Évolution du montant des versements périodiques par produit

Enfin, l'évolution du montant moyen par contrat de la variable cible est assez stable. Une grande disparité est visible entre les produits au niveau du montant moyen versé allant de 309 € pour le produit 1, en moyenne 743 € pour le produit 2 et 994 € pour le produit 3.

- Répartition des versements périodiques : la répartition en fonction du support euros, en unités de compte ou eurocroissance est détaillée pour chaque type de versement.

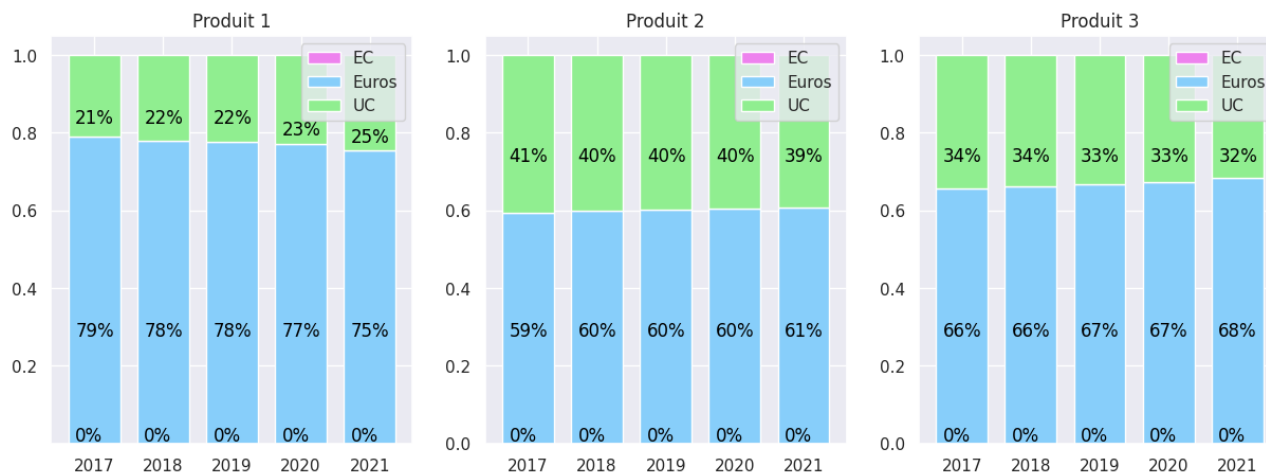


FIGURE 2.8 – Évolution de la part des versements périodiques par support

Moins de 1 % des versements périodiques sont orientés vers le support eurocroissance. Le support en euros représente deux tiers des versements permanents et les supports UC environ un tiers.

- Rachat partiel : indique le montant racheté partiellement (ce qui n’entraîne pas la fermeture du contrat). Une seconde variable compte le nombre de rachats partiels au cours de l’année.
- Arbitrage : cette variable est représentée à la fois par la variable de montant annuel d’arbitrages entrant sur un support et le nombre d’arbitrages effectués au cours de l’année.
- Taux acquisition : représente les frais appliqués sur le premier versement.
- Frais versement : sont calculés comme la moyenne des frais prélevés sur les versements au cours de l’année. A l’origine cette variable est nulle s’il n’y a eu aucun flux de versement. Elle a été retraitée afin de trouver la valeur des frais appliqués la plus proche. Si aucun frais n’est disponible sur la période d’observation, c’est le taux par défaut extrait des conditions générales du produit qui est affiché. En effet, la valeur nulle dans le cas où il n’y avait pas de flux au cours de l’année pouvait biaiser les estimations par la suite. Ce ne sont pas des frais de versements nuls qui entraînent l’absence de versement. Au contraire plus la part des frais est faible (souvent la conséquence d’un geste commercial face au montant investi) plus les montants de versements peuvent être importants.
- Exposition : variable entre 0 et 1 représentant le taux de présence sur l’année. Elle vaut 1 si l’ancienneté est supérieure à une année et est égale à des valeurs inférieures à 1 lorsque le contrat a été ouvert en cours d’année.

2.2 Retraitement des variables

2.2.1 Données incomplètes

Les données sur les versements sur le support eurocroissance représentaient une quantité d’information trop faible. Par conséquent, les variables concernant le support eurocroissance n’ont pas été gardées. Néanmoins, les montants propres au support eurocroissance restent inclus dans l’encours total, les versements de type initial, libre et périodique. Seule la distinction par support n’est pas complète.

Concernant les variables de versement pour les contrats réduits ou ne présentant pas de flux, la valeur vide a été remplacée par une valeur nulle pour indiquer qu’il n’y a pas eu de versement.

Chaque contrat dispose de l'information sur le versement initial payé à l'ouverture. Celle-ci a été dupliquée sur toutes les années pour un assuré, même si la souscription n'a pas été effectuée l'année d'exercice considérée.

2.2.2 Regroupement de données

Certaines variables présentent un nombre important de modalités. Pour éviter une surcharge de dimension à prendre en compte dans les modèles (certains modèles ne sont pas en mesure de traiter des variables catégorielles avec un grand nombre de classes) et surtout pour s'assurer de la représentativité de chaque modalité, un regroupement de ces dernières a été nécessaire. En effet, il est possible que les modalités d'une variable catégorielle présentent un nombre insuffisant d'individus ou qu'elles révèlent un déséquilibre, avec une classe beaucoup plus représentée que les autres. Effectuer un regroupement s'opère au détriment de l'exactitude de l'information. Il est donc crucial d'identifier les regroupements satisfaisants au regard de la variable cible. En particulier, les variables code postal, situation familiale et code de CSP ont été analysées.

Variable Région

Le choix a été porté sur un regroupement des codes postaux par région. Cela a pu être effectué à partir du code postal du client et d'une base de données détaillant pour chaque code postal de France : le nom de la commune, le département et la région associés. La jointure de ces deux bases a permis de récupérer le code de région. Les 6 607 modalités de code postal se réduisent à 15 modalités de région (France métropolitaine, Départements d'Outre-Mer, Collectivités d'Outre-Mer). En regroupant les Départements ou Régions d'Outre-Mer (DROM) et les Collectivités d'Outre-Mer (COM) pour former la modalité DROM-COM, nous obtenons la répartition suivante :

Code Région	Nom Région	Répartition
1	DROM-COM	2,09 %
11	Ile-de-France	9,86 %
24	Centre-Val de Loire	4,78 %
27	Bourgogne-France-Comté	5,18 %
28	Normandie	5,72 %
32	Hauts-de-France	6,87 %
44	Grand Est	9,19 %
52	Pays de la Loire	7,96 %
53	Bretagne	6,88 %
75	Nouvelle-Aquitaine	12,20 %
76	Occitanie	11,26 %
84	Auvergne-Rhône-Alpes	10,87 %
93	Provence-Alpes-Côte d'Azur	6,74 %
94	Corse	0,40 %

TABLE 2.4 – Codes des régions

Les codes de région ont été définis selon la nomenclature utilisée par l'INSEE entre 1 et 94. Les DROM-COM ont été rassemblés sous le code région 1.

En croisant les montants des versements programmés avec le code de région, les montants moyens par code de région mettent en évidence une disparité entre les versements effectués dans les territoires des DROM-COM (code 1) et les autres régions. Les montants moyens de ces dernières varient de manière plus progressive. La suite de l'étude déterminera si ces différences de montants sont suffisantes pour définir la variable région comme une variable discriminante.

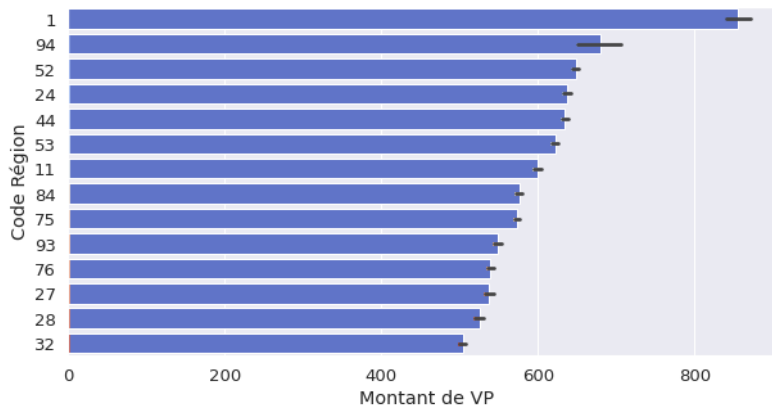


FIGURE 2.9 – Classement des codes de région par montant moyen de versements périodiques

Les montants moyens sont compris entre 500 € et 860 €. La région DOM-TOM représentée à 2 % effectue les versements les plus importants mais avec une grande volatilité dans les montants. La Nouvelle-Aquitaine, l’Occitanie et l’Auvergne-Rhône-Alpes constituent un tiers de la base de données. La Corse constitue une modalité peu représentée. Finalement il a été choisi de regrouper la modalité Corse (code 94) avec les DROM-COM (code 1). Ce regroupement diminue légèrement le montant moyen de la modalité. Néanmoins, cette région représente toujours la variable avec les montants de versements les plus importants quel que soit le produit considéré.

Variable catégorie socio-professionnelle

L’étude de la variable socio-professionnelle a consisté à analyser le libellé complet de la profession. La classe 9 représente les professions « Non déterminés ». Les libellés de cette classe comportent certains mots tels que : cadre, sans-profession, enfant, etc. Or ces professions devraient appartenir à une autre classe existante. La question est donc de savoir s’il est possible d’attribuer la bonne classe aux professions appartenant à la catégorie 9.

Pour chaque catégorie, une étude de la fréquence des mots a été faite afin d’en faire ressortir les mots les plus caractéristiques de la catégorie. Le libellé complet est formé d’un ensemble de mots. Cela peut poser des difficultés sur la mise en place d’une classification de texte puisque les mots « Agriculteur » et « Agriculteurs » seront considérés comme différents rendant la fréquence du terme sous-estimée. Un ensemble de techniques d’analyse de texte existe afin de réduire les formes d’un mot.

Ainsi, l’objectif est de transformer chaque libellé en liste de mots utilisables pour une classification. Une série de méthodes et de techniques est appliquée pour permettre une standardisation du texte. Ces méthodes sont empruntées au traitement automatique des langues (TAL), plus couramment appelé NLP en anglais pour *Natural Language Processing*. Le traitement de texte s’effectue en déroulant les étapes suivantes :

1. Le traitement débute par un nettoyage du texte en retirant la ponctuation, les accents, les lettres en majuscule, les espaces doublés et la tabulation inutile.
2. Puis, un processus de *tokenization* est appliqué permettant le découpage du texte en mot (appelé *token*). En général, concernant les langues latines, cela consiste à séparer le texte à chaque espace blanc.
3. Ensuite, les mots de liaison sont retirés même s’ils sont très peu présents dans notre cas. Les mots de liaison sont plus fréquents lors de l’analyse de texte avec des phrases plus complexes. Cette classe de mot n’apporte pas d’information sur la nature d’une catégorie. Cela concerne les mots tels que sur, dans, le, de, à, un, une, ...

Par exemple, après l'application des premières méthodes le libellé « Agriculteur sur grande exploitation » devient [« agriculteur », « grande », « exploitation »] et « Salarié cadre dans l'industrie » donne [« salarié, « cadre », « industrie »].

4. Enfin, la méthode dite de racinisation est appliquée. Ce processus recherche le radical le plus probable du mot. Cela permet d'accéder à la forme la plus basique du mot. En enlevant les suffixes et préfixes des mots, les accords et les formes conjuguées sont perdus évitant une sous-évaluation de l'apparition du mot dans une classe de CSP. Par exemple les mots « Agriculteur » et « Agriculteurs » sont réduits à leur racine « Agriculteur » et « exploitant » réduit à « exploite ». Cette étape peut transformer des mots en d'autres qui n'existent pas en français. Le package *Snowball Stemmer NLTK* a été utilisé pour cette étape se basant sur l'algorithme de Porter [21] développé en anglais au départ et proposant d'autres langues, et en particulier le français.

A la suite de ces quatre étapes, une liste de mots standardisés est obtenue pour chaque catégorie de CSP à partir du texte initial. Cette liste permet, pour chaque catégorie, de calculer l'occurrence des mots. Chaque catégorie est caractérisée par les dix premiers mots les plus fréquents.

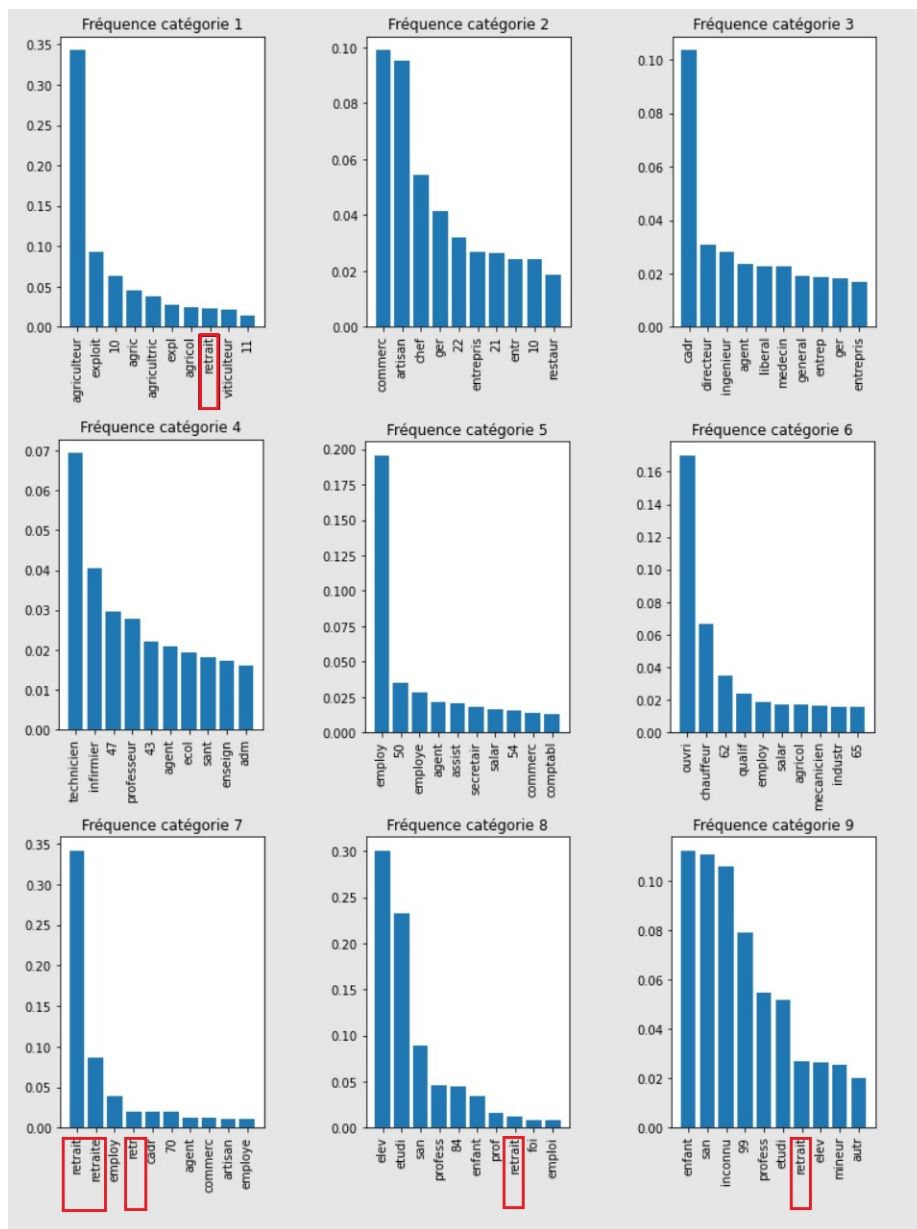


FIGURE 2.10 – Fréquence des mots dans chaque classe

Les observations des libellés complets des catégories socio-professionnelles ont mis en évidence des incohérences à l'égard de la catégorisation des professions. En particulier, les assurés avec la qualification de retraités se trouvent dans leurs anciennes catégories socio-professionnelles de vie active (encadré rouge dans la figure 2.10). Cela représente 0,58 % d'assurés mal classés.

L'objectif premier du retraitement des CSP était de reclasser entièrement la catégorie 9 avec le libellé. A la lecture du texte, des fautes d'orthographe et des tronquatures de la fin du libellé ont été relevées rendant difficile l'analyse de la fréquence d'apparition des mots. Les résultats ne sont pas concluants. Pour la catégorie 7 par exemple, le mot retraite semble se présenter sous trois formes. Cela limite les méthodes d'analyse de texte et pourrait conduire à une classification fautive. Ainsi, la base d'apprentissage n'est pas fiable pour reclasser les professions des individus de la catégorie 9. Finalement, le retraitement se limite à la reclassification des retraités dans la classe 7 par la recherche de la présence du mot « Retraite », « Retraités ».

CSP	Avant retraitement	Après retraitement
1	3,34 %	3,31 %
2	9,24 %	9,18 %
3	13,78 %	13,67 %
4	6,36 %	6,33 %
5	38,92 %	38,76 %
6	5,87 %	5,84 %
7	12,30 %	12,88 %
8	9,03 %	8,92 %
9	1,16 %	1,12 %

TABLE 2.5 – Fréquence des modalités de professions après retraitement

En conclusion, ce retraitement a eu peu d'impact sur la proportion finale de chaque modalité. Néanmoins, il a pour intérêt de présenter une meilleure qualité des données.

Variable situation familiale

Au vu des premières observations, les modalités *Pacsé(e)*, *Séparé(e)* et *Veuf(ve)* présentent trop peu d'information pour pouvoir être exploitées. En considérant les trois produits, 50 % des assurés du portefeuille étudié sont mariés.

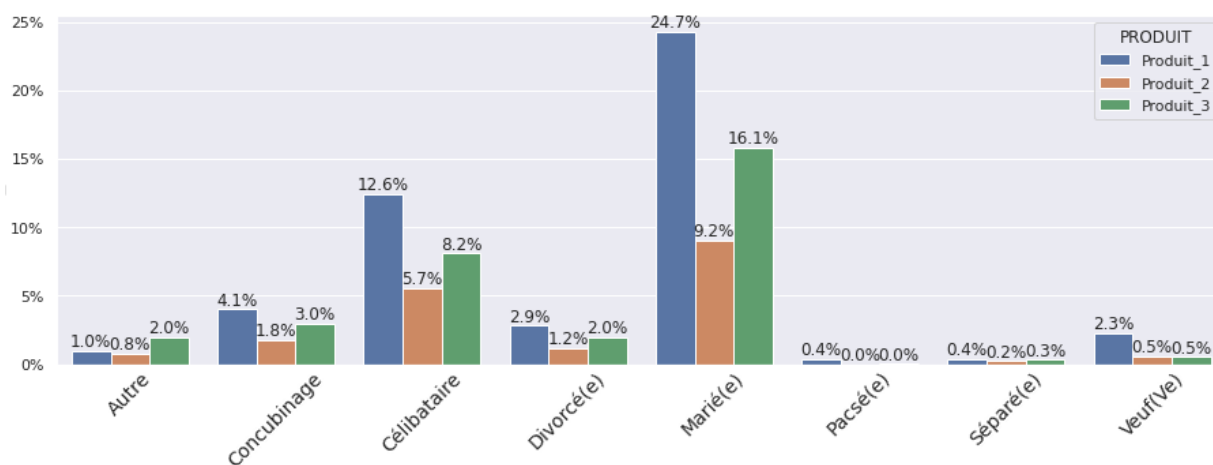


FIGURE 2.11 – Fréquence des modalités de situation familiale

Après croisement avec la variable à prédire, le regroupement de la modalité *Pacsé(e)* avec *Marié(e)* semblait judicieux. Les catégories *Concubinage* et *Célibataire* n'ont pas été modifiées. Enfin les situations *Divorcé(e)*, *Séparé(e)* et *Veuf(ve)* ont été regroupées avec la modalité *Autre*.

Les regroupements révèlent des montants moyens relativement proches en fonction du produit étudié. Néanmoins aucune catégorie ne se distingue par rapport à une autre : les écarts-types de chaque classe ne sont pas disjoints (cf. graphique B.1).

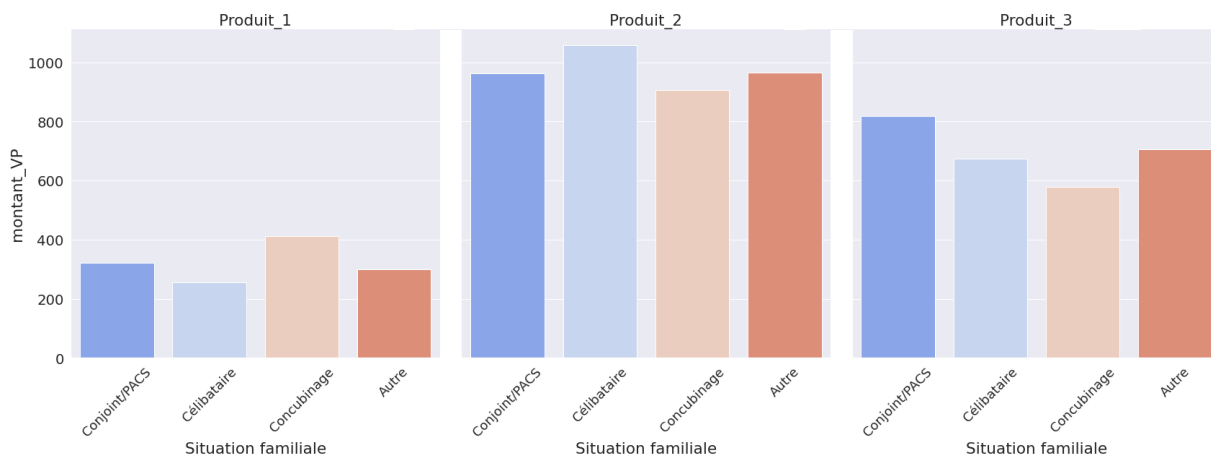


FIGURE 2.12 – Montant moyen des versements périodiques par modalités de situation familiale après retraitement

Pour le produit 1, les profils en concubinage versent un montant moyen de 400 €, ce qui est un peu plus important comparé aux montants moyens des autres catégories. Au contraire, pour le produit 2, la catégorie versant les primes périodiques les plus élevées sont les célibataires, avec un montant moyen au dessus de 1 000 €. Enfin, la différence de versements moyens en fonction de la situation familiale est plus marquée pour le produit 3 : les assurés pacsés ou mariés versent en moyenne des primes de 800 € contre des montants en dessous de 600 € pour les assurés en situation de concubinage.

Pour conclure, les regroupements effectués permettent de réduire le nombre de modalités à étudier, tout en essayant de garder une cohérence par rapport à la variable cible. Néanmoins, ces premières analyses croisées ne révèlent pas de modalité plus discriminante qu'une autre dans la prédiction du montant de versements. La section suivante permet d'identifier les variables qui ne contribuent pas significativement à la prédiction ou qui sont redondantes pour le modèle.

2.3 Analyse du portefeuille

2.3.1 Corrélations

L'étude des corrélations a pour objectif d'identifier les variables explicatives dépendantes entre elles, ce qui apporterait une redondance d'information et nuirait au modèle. Ainsi, supprimer de l'information permet aussi de réduire le nombre de variables à prendre en compte dans les modèles de *Machine Learning*.

Corrélation des variables qualitatives

La matrice de corrélation sur les données qualitatives donne une première évaluation de la dépendance entre les variables explicatives et la variable cible à modéliser. La matrice 2.13 a été réalisée en utilisant la méthode du τ de Kendall. Les coefficients mesurent la corrélation basée sur le rang entre deux variables. Dans la mesure où le caractère linéaire des corrélations n'est pas marqué, le τ de Kendall est une méthode plus robuste comparée à la méthode de Pearson utilisée par défaut. Les explications des méthodes de corrélation de Kendall et Pearson sont détaillées en annexes A.

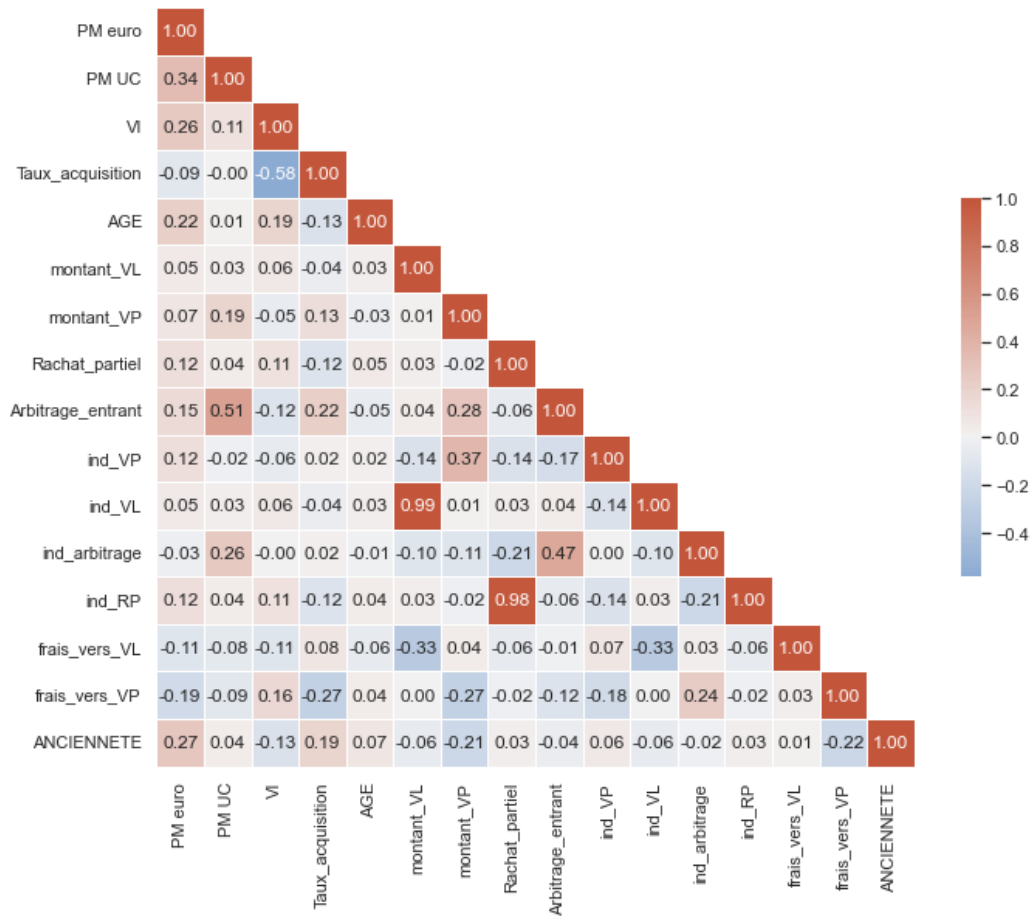


FIGURE 2.13 – Corrélation des variables quantitatives sur la base de données

La variable *montant_VP* est fortement corrélée aux nombres de versements par an, aux montants d'arbitrage et aux frais sur les versements périodiques. En effet, s'il y a eu un versement périodique, le montant de versements sera positif. Aussi, plus les montants sont élevés, plus il y a de possibilités d'effectuer un geste commercial pour réduire les frais sur les versements. Les variables *ind_VP* et *frais_vers_VP* sont forcément corrélées à la variable *montant_VP* et ne devront pas être considérées dans le modèle de prédiction.

Les variables de comptage et de montant, associées aux versements, aux arbitrages ou aux rachats sont aussi logiquement corrélées : si le montant est strictement positif alors il y a eu au moins un acte de versement (respectivement d'arbitrage ou de rachat) au cours de l'année. L'information sur le montant ou sur la fréquence devra être choisie dans les modèles pour éviter une redondance d'information et réduire la taille des données à analyser. A noter que si les variables des montants de versements libres et de rachats ont un coefficient de corrélation proche de 1 avec leur variable de comptage, la fréquence de primes périodiques est moins corrélée avec le montant annuel avec un τ de Kendall à 0,37. Autrement dit, effectuer des versements de manière plus régulière n'implique pas de verser des primes périodiques plus élevées.

D'autre part, les arbitrages semblent être corrélés à l'encours sur les supports en unités de compte. Le contexte économique des marchés financiers des dernières années (hors période de pandémie) a favorisé les arbitrages entrants vers les supports en unités de comptes, qui sont plus rentables que les supports en euros.

Corrélation des variables qualitatives

Le V de Cramer est une statistique utilisée pour tester l'indépendance entre deux variables qualitatives. La détermination du coefficient V de Cramer est détaillée dans l'annexe A. Plus le V de Cramer est proche de 1 plus les variables sont corrélées entre elles. A l'inverse, il n'y a pas de corrélation entre deux variables lorsque le coefficient est égal à 0.

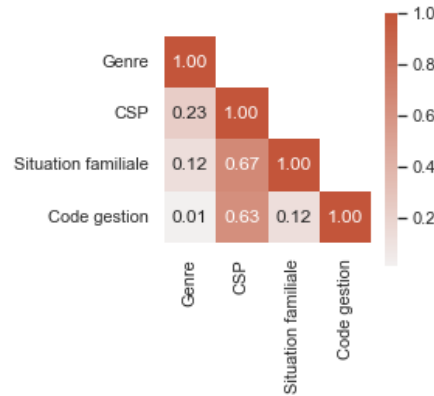


FIGURE 2.14 – Corrélation des variables qualitatives

Il ne semble pas exister de fortes corrélations entre les variables qualitatives. L'ensemble des coefficients calculés est en dessous de 0,7. Les variables CSP et la situation familiale de l'assuré apparaissent être plus corrélées comparées aux autres couples de variables observés avec le V de Cramer le plus élevé (0,67).

2.3.2 Analyse bivariée

Après l'analyse des corrélations, l'étude est portée sur une description statistique des montants de versements périodiques en fonction des différentes caractéristiques du contrat et de l'assuré au moment du versement. Cela permet de compléter les tendances mises en évidence par les résultats de corrélation. L'étude de corrélation entre les variables a pour intérêt de souligner l'information redondante qui alourdirait les modèles de prédiction, en plus de mesurer la relation entre deux variables. L'analyse bivariée, entre la variable cible et une variable explicative, aide à identifier la manière dont deux variables interagissent et les effets non-linéaires sous-jacents.

Montant de versements périodiques en fonction du genre de l'assuré

Les hommes et les femmes ont les mêmes comportements moyens sur le montant de versements. Cette variable ne semble pas être déterminante dans l'estimation des montants, et ce quel que soit le produit.

Montant de versements périodiques en fonction de l'âge et de l'ancienneté

De façon générale, les montants de versements décroissent avec l'ancienneté du contrat, de manière plus ou moins importante en fonction du produit. Par contre, les différents produits présentent une évolution des montants en fonction de l'âge très différente.

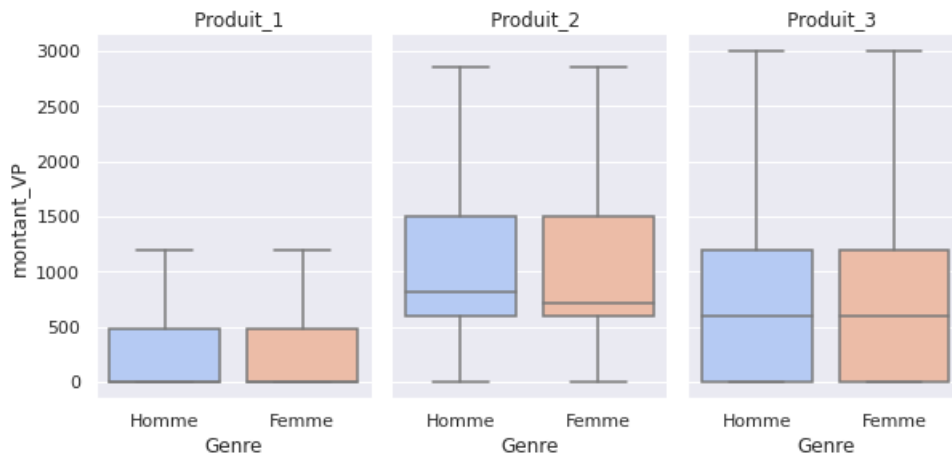


FIGURE 2.15 – Montant moyen des versements périodiques en fonction du genre de l'assuré

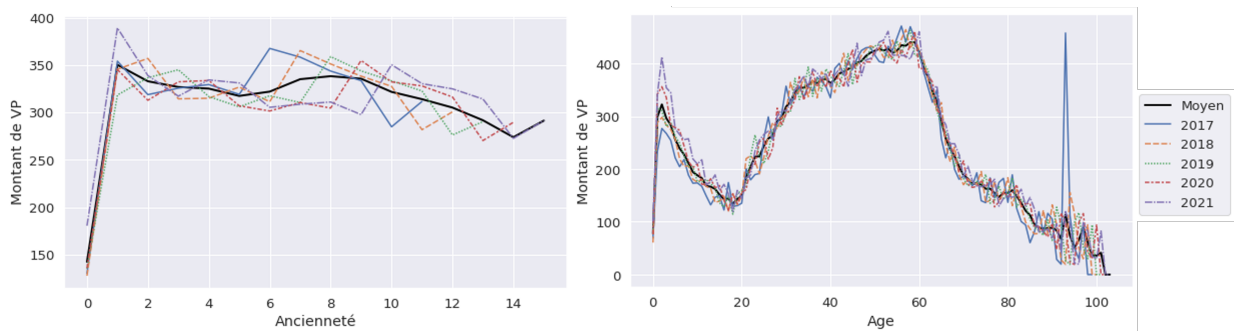


FIGURE 2.16 – Montant moyen des versements périodiques sur le produit 1 en fonction de l'ancienneté (graphique de gauche) et de l'âge (graphique de droite)

Plus particulièrement pour le produit 1, le montant moyen diminue légèrement en fonction de l'ancienneté. Un rebond peut être observé à partir de la sixième année qui semble être la conséquence d'un groupe de contrats souscrits en 2011 (les contrats ont six années d'ancienneté sur l'exercice 2017) qui présente des versements plus élevés que la moyenne. Ces pics de montants se retrouvent chaque année d'exercice à partir de 2017 et se décalent au vieillissement des contrats.

Concernant l'âge de l'assuré, les montants de versements augmentent à partir de 20 ans, ce qui correspond à une période de la vie où l'épargne augmente grâce à l'expérience du travail. Les contrats sont particulièrement actifs pour les assurés d'âge compris entre 20 et 60 ans avant une chute des versements à partir de 60 ans, et ce quelles que soient les années d'exercice. Cela correspond au début des départs à la retraite. Cette période peut nécessiter de nouveaux compléments de ressources financières qui poussent à réduire les versements sur l'assurance-vie.

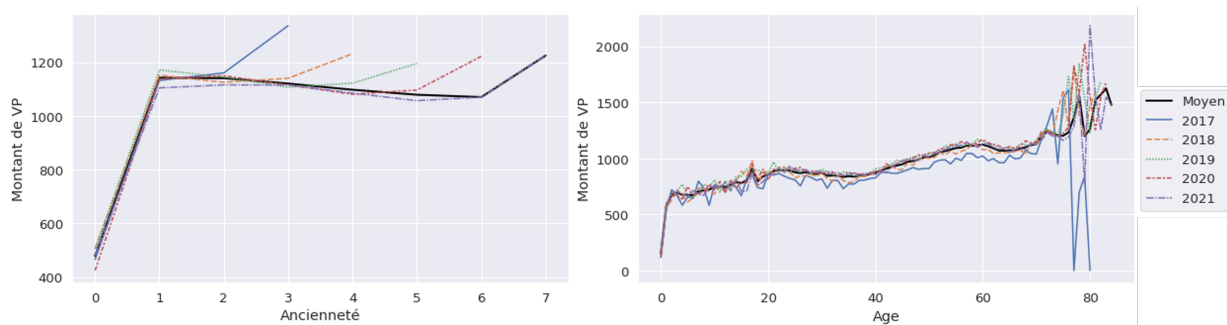


FIGURE 2.17 – Montant moyen des versements périodiques sur le produit 2 en fonction de l’ancienneté et de l’âge

La réduction des montants est plus lisse au cours de la vie des contrats du produit 2 passant de 1 200 € à 1 000 € de montant moyen. Les montants semblent augmenter linéairement avec l’âge de l’assuré. Le rebond observé sur la septième année d’ancienneté correspond à un groupe de contrats de faible effectif.

Aussi, ce produit demande un versement périodique minimal annuel. Conditionnés par ce plancher, les versements suivent une évolution légèrement à la hausse en fonction de l’âge.

Des pics de versements sur l’année d’exercice 2017 se détachent par rapport au montant moyen observé. Cet phénomène s’explique, d’une part, car le nombre de contrats avec des assurés ayant plus de 75 ans est faible. Ainsi, les montants moyens de versements sont plus volatiles. D’autre part, les assurés de 77 ans, et plus, sont principalement caractérisés par des contrats avec des anciennetés récentes (0 ou 1), ces contrats comptabilisent peu de versements périodiques voire aucun versement pour la première année de souscription. En 2018, des versements sont observés pour ce groupe de contrats, ce qui compense les nouvelles souscriptions.

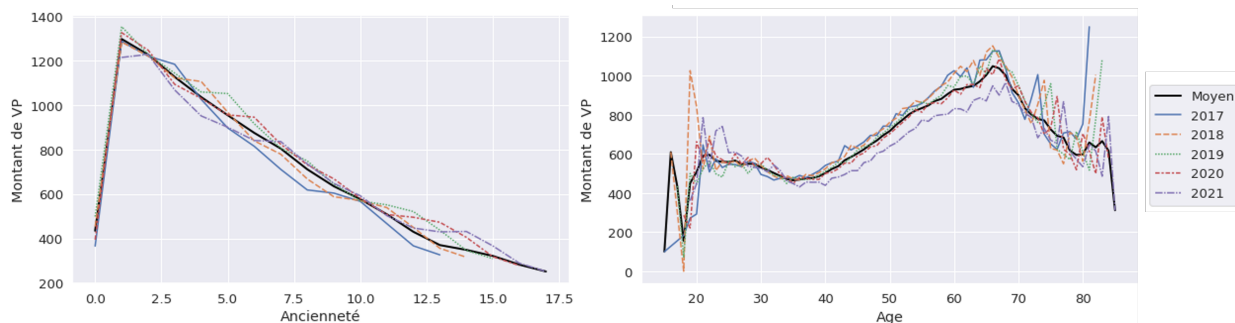


FIGURE 2.18 – Montant moyen des versements périodiques sur le produit 3 en fonction de l’ancienneté et de l’âge

Sur le produit 3, la diminution en fonction de l’ancienneté est beaucoup plus marquée. Concernant l’influence de l’âge de l’assuré, l’augmentation des montants est plus tardive et commence à partir de 40 ans avec une décroissance débutant autour de 65 ans. En effet, la nature du produit, qui est un contrat d’épargne retraite, permet de constituer un capital pendant la vie active puis de débloquent un complément de revenus sous forme de rente à partir de la retraite. Contrairement aux deux premiers produits qui proposent une transmission dans un cadre fiscal plus favorable du capital, le produit 3 est tourné vers un objectif de complément de ressources à partir de la retraite. Ainsi, ce produit tend à ne plus percevoir de versements périodiques à partir du moment où la phase de rente est déclenchée. Néanmoins, un épargnant retraité peut toujours verser sur ce produit pour profiter d’une défiscalisation de ses versements. Le montant moyen de versements chute à partir de 65-67 ans, tout en ayant une partie des assurés qui continuent leurs versements.

Des pics de versements concernant des assurés de plus de 80 ans sont aussi observables sur le produit 3. Contrairement au produit 2, aucun contrat n'a été récemment souscrit concernant ces assurés : l'ancienneté observée est supérieure à 4 ans. Ainsi, cela s'explique par un groupe de contrats qui en 2017 verse des montants importants de versements périodiques. Puis, ces versements diminuent petit à petit sur les années d'exercices suivantes. Cette génération est un cas particulier, les assurés de 80 ans sur les années d'exercices à partir de 2018 présentent des montants de versements plus faibles.

Montant de versements périodiques en fonction du type de gestion

Tous les modes de gestion ne sont pas proposés sur chaque produit. Différentes méthodes de gestion sont possibles, telles que la gestion personnelle, la gestion par horizon (basée sur le temps ou l'âge), la gestion par convention ou encore la gestion sous mandat. La fluctuation des montants en fonction du type de gestion diffère suivant les produits.

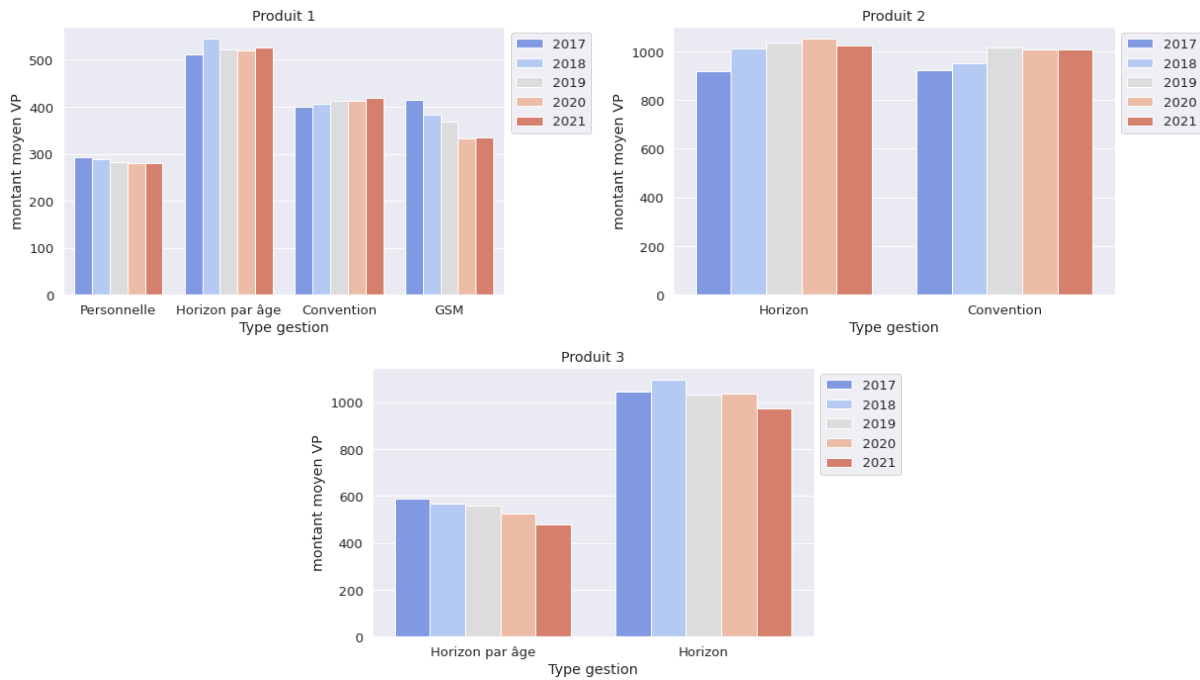


FIGURE 2.19 – Montant moyen des versements périodiques par type de gestion

Produit 1 : Les contrats en gestion par horizon d'âge semblent verser un montant plus important comparé aux autres modes de gestion.

Produit 2 : Le type de gestion a peu d'impact sur la prédiction des montants moyens du produit. Cette variable est peu discriminante sur la variable cible.

Produit 3 : Les montants versés sont deux fois plus importants en moyenne sur des contrats en gestion par horizon que des contrats en gestion par horizon d'âge.

En résumé, la variable type de gestion a plus d'impact sur les montants de versements des produits 1 et 3 que sur le produit 2.

Montant de versements périodiques en fonction des montants de rachats partiels

En considérant les rachats strictement positifs, plusieurs catégories de rachat partiel ont été formées puis croisées avec la variable cible.

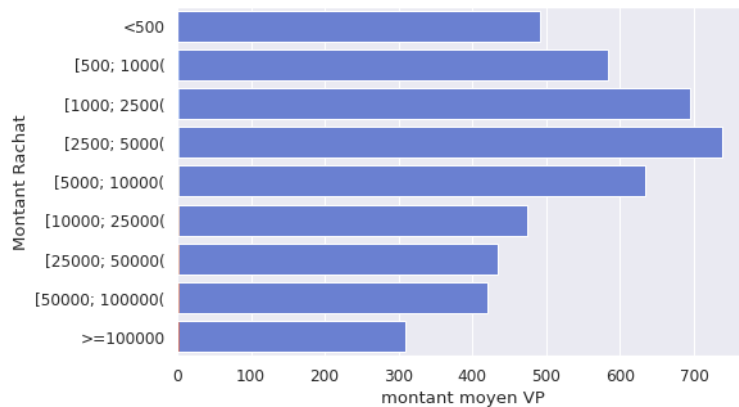


FIGURE 2.20 – Montant moyen des versements périodiques par catégorie de montant de rachats

Pour des rachats partiels entre 1 000 € et 10 000 €, le montant de versements moyen dépasse le seuil de 600 €. Plus les rachats sont importants, plus les montants moyens de versements augmentent jusqu'au seuil de 5 000 €. Puis à partir de ce seuil, les montants sont corrélés négativement. Cette tendance s'accroît pour les rachats de plus de 10 000 €. En effet, un rachat peut être effectué par un assuré en cas de besoin de capitaux, et par conséquent, peut impliquer des montants de versements plus faibles ou nuls en fonction du produit.

Montant de versements périodiques en fonction des montants d'arbitrages

De la même façon, les flux d'arbitrage strictement positifs ont été répartis par intervalle avant de déterminer le montant moyen de versements.

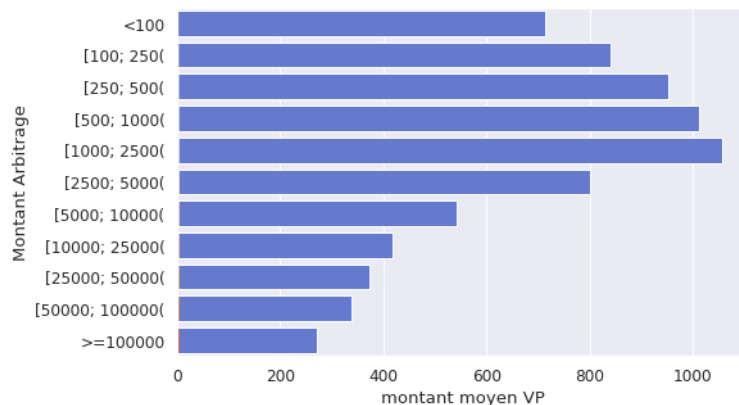


FIGURE 2.21 – Montant moyen des versements périodiques par catégorie de montant d'arbitrages

Les montants de versements sont croissants avec les montants d'arbitrages jusqu'à des flux de 2 500 €, puis sont décroissants au-delà. Les arbitrages supérieurs à 5 000 € sont souvent orientés vers un support en euros de façon à sécuriser les montants investis. La finalité est par la suite d'effectuer un rachat. Ce contexte de récupération de l'épargne accumulée est donc moins propice à la continuité de versements périodiques. Ce phénomène est d'autant plus accentué au cours du vieillissement des assurés qui de manière automatique (avec la gestion par horizon par exemple) ou volontaire (besoin de compléments financiers) cherchent à limiter les risques encourus et à stabiliser leur épargne.

Montant de versements périodiques en fonction des frais

Les différentes catégories formées par les frais de versements d'acquisition, libres ou périodiques sont présentées dans ce paragraphe.

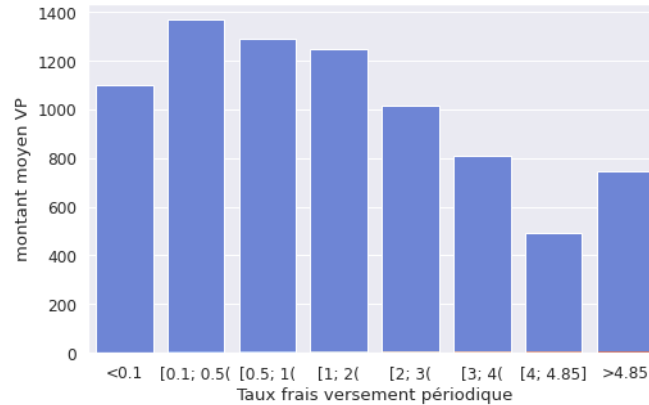


FIGURE 2.22 – Taux de frais sur les versements périodiques

Les frais sur les versements périodiques sont inversement corrélés aux montants. Plus les montants de versements sont élevés, plus la part de frais prélevés diminue. Cette variable ne sera pas intégrée dans les modèles de prédiction des versements étant trop dépendante des montants passés. Il est intéressant de noter que ce facteur peut être un levier pour inciter les assurés à augmenter leurs versements si les frais sont faibles (dans la limite de leur épargne disponible).

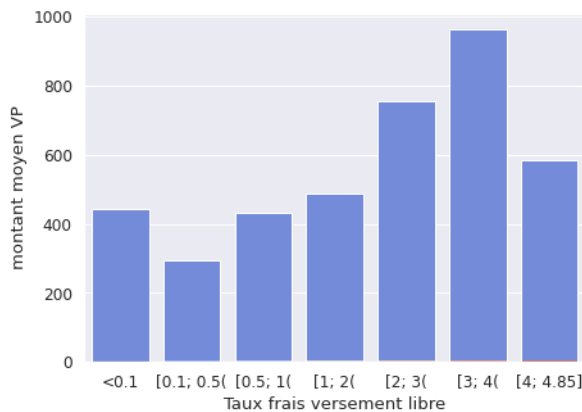


FIGURE 2.23 – Taux de frais sur les versements libres

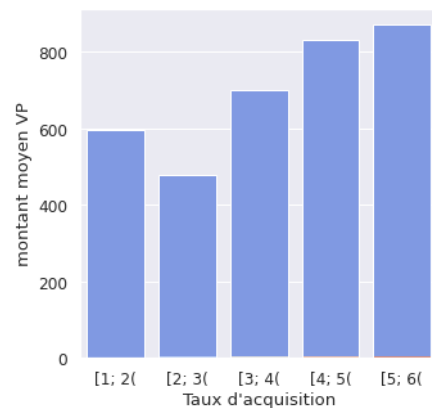


FIGURE 2.24 – Taux de frais d'acquisition

Les taux de frais d'acquisition semblent être corrélés au montant de versements périodiques. Le même phénomène est observé avec les taux sur les versements libres. En ce qui concerne ces derniers, des taux de versements libres élevés peuvent dissuader les assurés d'effectuer des versements plus importants, ce qui peut expliquer pourquoi l'épargne est plus susceptible d'être constituée à partir de versements périodiques.

Montant de versements périodiques en fonction de l'encours

Plusieurs catégories de provisions de l'encours du contrat et de la proportion de provisions constituées sur le support en euro ont été formées pour observer les relations possibles avec le montant des primes versées.

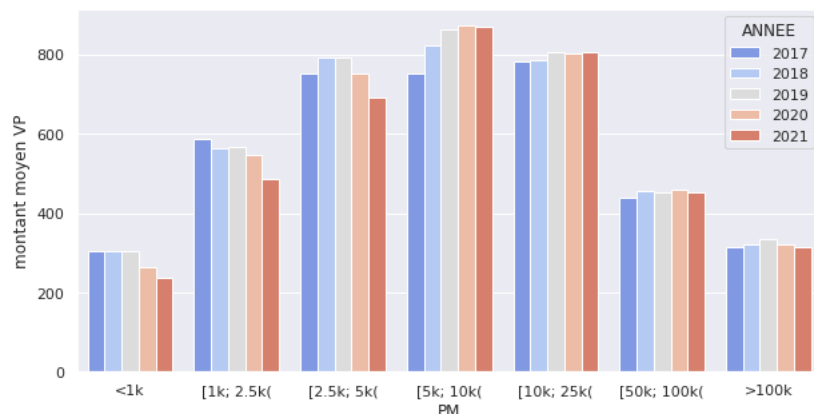


FIGURE 2.25 – Montant moyen des versements périodiques en fonction du niveau de provisions mathématiques

Les montants de versements sont répartis de façon centrée autour des encours de 10 000 €. L'encours du contrat est fortement corrélé aux montants de versements. En effet, à chaque versement l'encours est augmenté du montant de la prime versée.

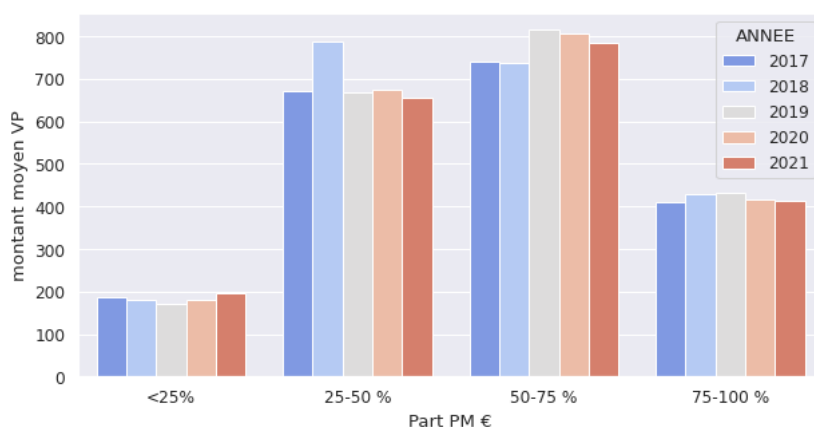


FIGURE 2.26 – Montant moyen des versements périodiques en fonction de la proportion de l'encours orientée vers un fonds en euro

La majorité des contrats est investie sur le fonds en euros comme mis en évidence par les graphiques 2.4 et 2.8. Les contrats les plus diversifiés, sous-entendu tournés vers des supports en unités de compte, présentent les montants de versements périodiques les plus faibles. Finalement, les contrats avec une part de l'encours orientée vers de l'euro à hauteur de 50 % versent en moyenne les montants les plus importants. Ces contrats présentent de la diversification tout en assurant une partie de l'épargne sur le fonds en euros.

Conclusion

Cette première analyse permet d'identifier les principales tendances qui déterminent le comportement des clients en matière de versements périodiques. Les résultats indiquent que le type de produit est un facteur déterminant dans les montants moyens versés. Le contrat d'épargne classique représenté par le

produit 1 est moins propice à des versements périodiques et s'alimente principalement à travers des versements libres. En revanche, les contrats d'épargne supplémentaires pour la retraite, représentés par le produit 3, sont caractérisés par des versements périodiques plus importants. Quant au produit 2 d'assurance-vie, les conditions incitent les assurés à effectuer un montant minimal de versements, ce qui favorise les versements périodiques.

Par ailleurs, des relations entre certaines variables du portefeuille sont mises en évidence, notamment l'âge, l'ancienneté et les arbitrages qui jouent un rôle significatif dans l'estimation des versements périodiques.

Ces résultats sont pris en compte lors de l'élaboration de modèles permettant d'estimer les versements périodiques, en particulier sur les modélisations qui sont réalisées de manière distincte pour chaque produit étudié.

Chapitre 3

Modélisation des versements

L'objet de ce chapitre est de confronter plusieurs méthodes de prédiction des primes périodiques en fonction des variables explicatives décrites précédemment. Le but est de définir le modèle le plus performant en fonction de différents critères sur la qualité de prédiction.

La première partie se concentre sur la détermination d'un montant moyen de versements en fonction de l'ancienneté des contrats. Deux méthodes sont étudiées. La loi de versements périodiques utilisée dans le modèle actuel de rentabilité est formée par une approche basée sur la variation de la part des contrats effectuant des versements. Cette méthode a été appliquée sur les dernières données disponibles. Puis, une seconde méthode statistique modélisant le montant moyen versé par produit et sa fluctuation en fonction de l'ancienneté est confrontée à cette dernière.

Par la suite, trois méthodes de *Machine Learning* sont paramétrées, afin de tester l'importance de l'ancienneté dans la fiabilité de prédiction et cela en comparant l'influence des autres variables dans la prédiction. L'objectif est de déterminer si les variables de prédiction sont autant décisives sur chacun des produits de nature différente. Ces modèles permettent, dans un second temps, de comparer l'efficacité de prédiction. Enfin, le modèle le plus performant sera choisi pour estimer la rentabilité des produits.

3.1 Mise à jour des lois actuelles

Tout d'abord, dans la suite du mémoire, le terme loi d'évolution est utilisé pour définir les variations à la hausse ou à la baisse des montants de versements d'une année sur l'autre.

Les lois d'évolution actuelles sont construites pour chaque produit en calculant, selon l'ancienneté des contrats, un montant moyen de versements et la proportion de contrats effectuant des versements périodiques. Ces lois d'évolution ont été modélisées à partir d'une base historique d'observations sur la période 2012-2016. A l'origine, les versements étaient modélisés par une loi de réduction et une loi d'augmentation des primes. Les études statistiques ont montré que les montants de primes ont tendance à diminuer en fonction de l'ancienneté du contrat. Ainsi, dans le modèle de rentabilité, les lois d'évolution utilisées n'étaient que des lois de réduction de primes.

Actuellement, seuls les produits¹ 2 et 3 présentent des lois de réduction de versements périodiques. Le produit 2 a la particularité de demander un montant minimal à verser annuellement contrairement aux autres produits. Quant au produit 3, près de 70 % des contrats ont mis en place des versements périodiques. Ainsi, l'épargne des produits 2 et 3 est principalement alimentée par des versements périodiques. De

1. Les produits 1 et 2 sont des produits d'assurance-vie, tandis que le produit 3 est un produit d'épargne retraite.

par l'importance des versements périodiques, ces produits ont été modélisés avec une loi d'évolution sur ces versements. Concernant le produit 1, les versements périodiques représentent des montants moins importants. Une moyenne du montant est considérée comme suffisante pour représenter l'ensemble des montants dans la modélisation de la rentabilité du produit. La pertinence de ce choix de modélisation, tout comme le choix de l'ancienneté comme seule variable explicative sont éprouvés dans ce mémoire.

Dans cette première partie, les versements périodiques ont été modélisés pour chaque produit en fonction de l'ancienneté des contrats de deux façons :

- à partir de la variation de la part de contrats versant des primes périodiques, ce qui correspond à une application de la modélisation actuelle sur une base de données plus récente ;
- à partir de la fluctuation des montants de versements périodiques.

Dans un premier temps, les fréquences et le montant moyen de versements ont été calculés à partir des données historiques entre 2017 et 2020. Ces données constituent la base d'apprentissage. Puis, les contrats associés à l'année 2021 serviront à mesurer les écarts de prédiction avec les montants observés, formant la base de test.

3.1.1 Modélisation de la part de contrats générant des versements périodiques

L'application de la méthode utilisée dans le modèle de rentabilité sur la base de données de 2017 à 2021 est présentée dans cette section. Pour rappel les lois utilisées actuellement sont des lois de réduction basées sur la variation du nombre de contrats générant des versements périodiques venant diminuer la probabilité annuelle de verser. La probabilité de verser est calculée selon l'ancienneté par différence sur deux années consécutives de la proportion de contrats générant des primes périodiques.

Pour la modélisation des lois d'évolution décrite par la suite, la méthode consiste à regrouper en un unique vecteur les fluctuations à la hausse comme à la baisse de la probabilité de verser afin d'être au plus proche des mouvements observés. Les variables discriminantes retenues sont le produit pour le montant et le produit et l'ancienneté pour la fréquence.

Pour chaque ensemble de contrats d'ancienneté n , le taux de versements périodiques (VP) est mesuré par la formule :

$$\text{Taux de } VP_n = \frac{\text{Nombre de contrats effectuant des } VP_n}{\text{Nombre de contrats } n} \quad (3.1)$$

Ainsi, pour chaque produit, la loi de versement est calculée comme le rapport entre deux années consécutives d'ancienneté des taux de versement.

Le montant de versements périodiques moyen du produit est calculé uniquement sur la première année comme suit :

$$VP_{moyen} = \frac{\text{Montant de } VP_1}{\text{Nombre de contrats } 1} \quad (3.2)$$

Le montant global est évalué pour chaque produit i par la formule suivante à partir d'une année complète d'ancienneté, sur les N années d'ancienneté :

$$\text{Montant } VP \text{ Total}_i = \sum_{n=1}^N \text{Nombre de contrats}_n^i * VP_{moyen}^i * \prod_{\tau=1}^n (1 + l_\tau) \quad (3.3)$$

avec $l_\tau = \frac{\text{Taux de } VP_\tau}{\text{Taux de } VP_{\tau-1}} - 1$ et $l_0 = 1$.

Ces formules supposent que les fluctuations du montant total versé par produit sont dues aux variations de la part de contrats générant des versements périodiques.

Ces formules ont été ensuite appliquées sur les données de 2017 à 2020 afin de pouvoir valider et comparer les capacités de prédiction sur les données de 2021.

Les résultats des taux de versement par ancienneté, estimés pour chaque produit, sont présentés ci-dessous :



FIGURE 3.1 – Modélisation de la probabilité de verser en fonction du produit et de l'ancienneté

L'intervalle de confiance à 95 % autour de la moyenne observée de la probabilité de verser est construite selon la formule issue de la loi des grands nombres :

Où :

$$IC_n = \left[\frac{f_n + \frac{u^2}{2k_n} - \frac{u}{\sqrt{k_n}} \sqrt{\frac{u^2}{4k_n} + f_n(1-f_n)}}{1 + \frac{u^2}{k_n}}, \frac{f_n + \frac{u^2}{2k_n} + \frac{u}{\sqrt{k_n}} \sqrt{\frac{u^2}{4k_n} + f_n(1-f_n)}}{1 + \frac{u^2}{k_n}} \right] \quad (3.4)$$

- u est le quantile de la loi normale centrée réduite d'ordre $1 - \alpha/2$ (avec $\alpha = 0,05$) ;
- k_n le nombre de contrats total d'ancienneté n ;
- f_n est le *taux de VP_n*.

L'année de souscription est marquée par une plus faible part de contrats générant des primes périodiques car en moyenne les affaires nouvelles arrivent en milieu d'année.

En particulier, pour les contrats avec une forte part de primes périodiques (produits 2 et 3), la part de contrats générant des primes diminue fortement au cours de l'ancienneté. Cela est dû à des contrats qui ne sont plus actifs, qualifiés de réduits. Concernant le premier produit, la part de contrats émettant des versements périodiques évolue à la hausse jusqu'à la neuvième année en moyenne pour ensuite diminuer. Concernant le produit 2, le nombre d'observations de la sixième année d'ancienneté est trop faible pour estimer un taux de versement satisfaisant. Une diminution constante de la probabilité de verser sera supposée à partir de la cinquième année d'ancienneté.

L'ancienneté 0 ne constituant pas une année entière de versement, les montants de versements et le taux de versement sur cette ancienneté ne sont pas représentatifs des volumes futurs. Cela est provient de la façon de définir l'ancienneté comme la différence entre l'année de souscription et l'année d'exercice observé. Le taux de versement est donc pris en compte qu'à partir de l'ancienneté 1. Ainsi, les montants moyens versés sont estimés sur les contrats d'ancienneté 1 ce qui donne :

	Produit 1	Produit 2	Produit 3
Montant moyen VP_1	342	1149	1309

TABLE 3.1 – Montant moyen de versements après une année d'ancienneté

Les tables de loi finales sont formées à partir des fluctuations entre deux taux consécutifs de versements. Lors de la projection des montants totaux estimés, les années de projection sont complètes. Le montant moyen estimé versé par l'assuré lors de la première année est le montant moyen défini par la formule 3.2. Puis, les années suivantes, il évolue en fonction des fluctuations des taux de versement. Aussi, l'hypothèse suivante est supposée : à partir d'une certaine ancienneté, la diminution des taux de versement est constante. Ainsi, le dernier taux calculé est gardé pour les prochaines années d'ancienneté qui n'ont pas encore été observées.

Les résultats obtenus avec la première méthode sont combinés dans la section 3.1.3, avec ceux produits par la méthode suivante.

3.1.2 Modélisation sur le montant des primes

Une deuxième méthode pour la modélisation des versements périodiques est confrontée. Ainsi, il ne s'agit plus de s'intéresser à la part de contrats effectuant des versements mais aux évolutions des montants seulement. Cette méthode est de type coût fréquence - coût moyen. Puisque seule la variable ancienneté joue dans la détermination des montants de versements, la fréquence de versements sur une année d'ancienneté multipliée par le montant moyen des contrats ayant versé revient à calculer le montant moyen par année d'ancienneté.

Une particularité dans la détermination des montants moyens est introduite. La comparaison des versements effectués entre deux années consécutives est regardée sur des contrats à iso-caractéristiques.

Cela permet de distinguer des générations de contrats ayant une tendance à verser davantage par rapport à d'autres. L'évolution des montants est focalisée sur l'évolution du temps à travers l'ancienneté et l'âge des assurés et ne prend pas en compte les autres caractéristiques des individus.

Pour ce faire, les montants de primes ont été récupérés pour chaque produit selon des triangles par génération et par ancienneté. Le triangle suivant est donné à titre d'exemple (pour des raisons de confidentialité les données ont été modifiées). Le triangle est composé en colonne des années d'ancienneté et en ligne des années de souscription. Par construction, le nombre de contrats par année de génération reste constant tout au long de l'observation.

Montant VP par année de police

Génération	Nb contrat	0	1	2	3	4	5	6	7	8	9	10
2006	1 175											
2007	15 017											4 280 181
2008	14 403									4 808 793	4 721 757	
2009	14 083								4 844 994	4 767 111	4 697 257	
2010	10 086							3 618 481	3 545 924	3 469 403	3 352 433	
2011	6 962						2 562 027	2 546 085	2 501 412	2 471 838	-	
2012	5 916					1 889 668	1 843 183	1 838 115	1 803 608	-		
2013	5 528					1 822 702	1 808 225	1 757 332	1 717 710	-		
2014	6 173				2 012 117	1 946 045	1 889 797	1 862 820	-			
2015	6 250			1 994 243	1 966 513	1 981 459	1 920 397	-				
2016	5 460		1 935 579	1 951 486	1 886 063	1 823 883	-					
2017	4 963	645 213	1 717 310	1 672 263	1 650 867	-						
2018	4 310	549 240	1 373 666	1 349 808	-							
2019	4 067	548 904	1 405 439	-								
2020	4 097	554 859	-									
2021	1 175	-										
Données inférieures		1 653 004	4 496 415	4 973 558	5 503 444	5 751 387	5 618 419	5 463 335	6 101 911	7 850 944	10 708 352	12 771 447
Données supérieures		1 743 358	5 026 555	5 617 993	5 864 693	5 750 207	5 587 690	6 162 542	8 002 682	10 892 330	13 045 308	13 699 195
Taux d'évolution			157,9%	-1,1%	-2,0%	-1,9%	-2,3%	-2,2%	-1,0%	-1,9%	-1,7%	-2,1%

FIGURE 3.2 – Triangle des versements périodiques par produit

Le taux de versement est obtenu en calculant le rapport des montants de versements périodiques pour des générations où l'information existe sur deux anciennetés consécutives (cellules encadrées dans la figure 3.2). Cela permet de comparer l'évolution des montants de versement par ancienneté sur les mêmes contrats : les caractéristiques générales du contrat et de l'individu restent les mêmes. Cette méthode contribue à mesurer de manière plus fine l'effet de l'ancienneté sur le comportement de versement des assurés.

Enfin, les fluctuations des montants de versements sont modélisées à travers un unique taux afin de lisser les variations obtenues. Le taux annuel global est obtenu par moyenne des taux de versement calculés. Le taux global est établi à partir de la première année d'ancienneté jusqu'à l'ancienneté maximale du produit. L'ancienneté 0 qui représente l'année de souscription n'est pas prise en compte afin d'avoir un taux plus cohérent. Le taux initial observé est celui de la première ancienneté. L'évolution des taux d'une année d'ancienneté sur l'autre est présentée dans le tableau 3.2.

Ancienneté	Taux d'évolution des VP		
	Produit 1	Produit 2	Produit 3
1	-1,1 %	-1,0 %	-5,5 %
2	-2,0 %	-2,0 %	-8,2 %
3	-1,9 %	-1,9 %	-7,4 %
4	-2,3 %	-2,3 %	-7,2 %
5	-2,2 %	2,2 %	-7,8 %
6	-1,0 %		-6,5 %
7	-1,9 %		-6,8 %
8	-1,7 %		-7,1 %
9	-2,1 %		-6,3 %
10	-1,4 %		-6,2 %
11	-2,0 %		-5,4 %
12	-2,2 %		-4,4 %
13	-0,2 %		-3,9 %
14			-4,5 %
15			-10,2 %
Taux Global	- 1,70 %	- 1,00 %	- 6,50 %

TABLE 3.2 – Évolution des montants de versement d'une année sur l'autre

Les taux d'évolution des montants sont négatifs, autrement dit les montants diminuent au cours de la vie du contrat pour chaque produit.

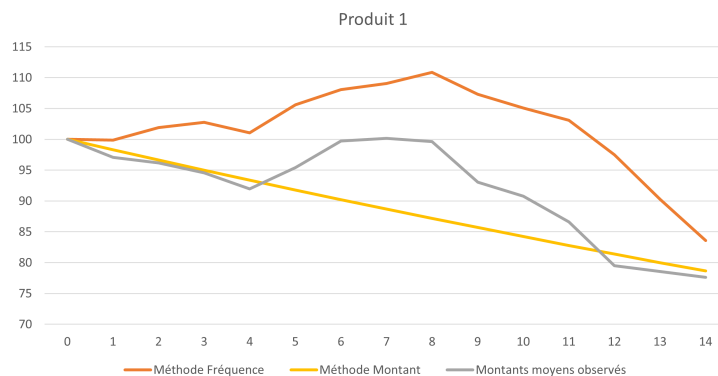
L'année de souscription n'étant pas prise en compte, cette méthode demande de calculer le montant moyen par produit sur l'ensemble des contrats avec exactement un an d'ancienneté. Le montant moyen de versements périodiques est exactement celui calculé dans la première méthode (tableau 3.1).

$$Montant\ VP\ Total_i = \sum_{n=1}^N \text{Nombre de contrats}_n^i * VP_{moyen}^i * \prod_{\tau=1}^n (1 + k^i) \quad (3.5)$$

avec k^i le taux d'évolution global retenu pour le produit i .

3.1.3 Comparaison des deux modèles

Pour comparer les deux modèles présentés, les lois d'évolution ont été tracées en partant d'un montant moyen de versements égal à 100. La courbe de l'évolution des montants moyens par ancienneté, sur tout l'historique confondu, sans effectuer de retraitement est aussi ajoutée (modélisée en gris sur les graphiques).



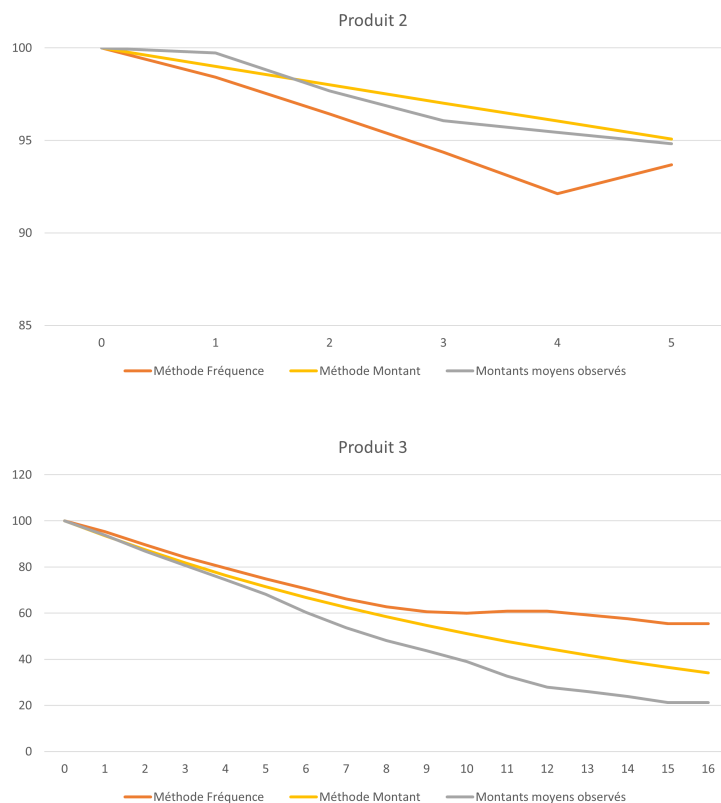


FIGURE 3.3 – Modélisation de l'évolution d'un montant moyen égal à 100 en fonction du produit et de l'ancienneté

La courbe déterminée à partir des évolutions via la méthode 2 sur les montants ne ressemble pas à une courbe de tendance sur les montants moyens observés (courbe en grise). En effet, la méthode 2 se concentre sur l'évolution des montants périodiques pour des contrats avec les mêmes caractéristiques. Bien que la méthode se base exclusivement sur les montants de versements, elle ne prend pas en compte les mêmes montants globaux pour chaque année d'ancienneté que ceux qui déterminent l'évolution observée en fonction de l'ancienneté.

Pour le produit 1, une augmentation de la proportion des contrats est observée à partir de la cinquième année d'ancienneté. Ainsi, les montants moyens observés augmentent parce qu'il y a plus de contrats actifs en ce qui concerne les versements périodiques. Au contraire, en regardant les taux d'évolution des montants périodiques des mêmes générations de contrats entre deux années d'ancienneté, une diminution des montants est constatée.

Sur le produit 2, la différence d'évolution est peu marquée entre les méthodes. La méthode basée sur la fréquence de versement tend à déterminer des montants moyens périodiques relativement plus faibles.

Sur le produit 3, une tendance à la baisse est estimée pour les deux méthodes. La méthode de fréquence détermine des montants moyens plus constants à partir de la onzième année.

Validation des premiers modèles :

Les modèles déterminés précédemment ont été appliqués pour estimer les montants moyens obtenus en 2021 en fonction de l'ancienneté. Pour chaque année d'ancienneté, un montant moyen de versements est calculé auquel sont appliquées les lois d'évolution des montants. La comparaison des estimations des deux méthodes est faite avec les montants observés de 2021 pour chaque produit.

La méthodologie de validation des modèles est basée sur les montants connus des années passées pour prédire les années suivantes à partir des taux d'évolution estimés. Ainsi, avec une profondeur de cinq ans sur nos données d'observations, il est possible d'appliquer au maximum quatre taux consécutifs pour reconstituer les montants observés en 2021. Tout d'abord, les montants moyens par ancienneté sur l'exercice 2020 sont calculés pour estimer les montants enregistrés en 2021 et ce jusqu'aux montants moyens de 2017 pour estimer les montants de 2021.

Soit N l'année maximale d'ancienneté observée, sur l'année d'exercice 2021, en considérant k années de profondeur pour calculer les montants périodiques versés de 2021, le total s'obtient par la formule :

$$Montant\ VP\ Total^i = \sum_{n=1}^N Nombre\ de\ contrats_n^i * VP_{moyen\ (n-k)}^i * \prod_{\tau=1, n-\tau > 1}^k (1 + l_{n-\tau}^i)$$

avec $l_{n-\tau}$ le taux d'évolution entre $n - \tau - 1$ et $n - \tau$

Les montants moyens de 2020 et 2017 ont été utilisés pour estimer les montants de 2021.



FIGURE 3.4 – Modélisation des montants moyens sur les données de 2021 en fonction de l'ancienneté et des montants moyens de 2020 (gauche) et 2017 (droite)

Deux métriques sont utilisées pour évaluer les performances du modèle sur la base qui a servi à calculer les lois d'évolution. Le RMSE mesure l'écart quadratique moyen entre la valeur prédite et la valeur réellement observée. Tandis que l'estimation du montant total indique si au global le total des versements est surestimé (valeur positive) ou bien sous-estimé (valeur négative). Les mêmes métriques ont été calculées sur les données en excluant les écarts observés avec l'ancienneté 0. L'ensemble des indicateurs de performance et de mesures d'erreurs qui seront utilisés tout au long de l'étude sont détaillés en annexe C.

Les métriques sont calculées en sommant par ancienneté le produit du montant moyen obtenu par le nombre de contrats :

	Indicateurs	Produit 1	Produit 2	Produit 3
Modèle 1 (fréquence)	RMSE	56 919	89 586	110 442
Modèle 1 (fréquence)	Estimation du montant total	-2,08 %	0,83 %	7,14 %
Modèle 2 (montant)	RMSE	30 798	129 585	93 897
Modèle 2 (montant)	Estimation du montant total	-1,89 %	1,83 %	5,47 %

TABLE 3.3 – Projection des modèles à partir des montants moyens de 2020 sur les données de 2021

Le montant total de versements périodiques du produit 1 est sous-estimé autour de 2 % pour les deux modèles. Tandis que le montant total du produit 3 est sur-estimé à 7,14 % pour le modèle 1 et 5,47 % pour le modèle 2. Enfin, le produit 2 affiche un montant total estimé assez proche de l'observé en 2021. Au global, le montant total de versements périodiques est surestimé pour le produit 2 et 3 quel que soit le modèle et sous-estimé pour le produit 1. Le modèle 2 est plus performant au niveau de l'estimation des montants totaux des produits 1 et 3.

Sur une profondeur d'une année, les montants moyens estimés sont assez proches des montants moyens observés. Mais les écarts sur les montants moyens ramenés au nombre de contrats total produisent des erreurs d'estimation pouvant être importantes. Notamment, 7,14 % du montant pour le produit 3 est surestimé. Sur une période d'estimation plus longue, les écarts d'estimation s'accumulent. Dans certains cas, cela peut permettre de compenser les fluctuations. La projection des montants moyens de 2017 sur quatre ans met en avant des erreurs de plus en plus significatives qui n'étaient pas visibles sur un an de projection.

L'approche globale, considérant un montant moyen par ancienneté, reste une modélisation simple. Surtout concernant le modèle 1 sur l'évolution de la part de contrats effectuant des versements, il ne permet pas de capter des phénomènes de fluctuation des montants permanents autres que la décision d'effectuer un versement.

Dans le but de réduire les erreurs de manière plus fine, une approche ligne-à-ligne est considérée à travers les algorithmes de *Machine Learning*. Ces méthodes alternatives permettent d'intégrer d'autres facteurs pouvant influencer sur les versements périodiques afin d'aboutir à un modèle plus complexe.

3.2 Théorie des modèles de Machine Learning

Le *Machine Learning* regroupe un ensemble d'algorithmes permettant à partir d'une base de données d'apprentissage de prédire des résultats en fonction de paramètres d'entrée. Le processus d'apprentissage consiste à déterminer la fonction qui génère une sortie à partir de données en entrée (souvent sous la forme d'un vecteur de variables explicatives). Il existe deux principaux types de méthodes : supervisé ou non-supervisé. Un algorithme d'apprentissage dit supervisé développe ses capacités de prédiction à partir de données associées à une solution connue. L'apprentissage non-supervisé apprend, contrairement au supervisé, sur des données sans solution connue. Les algorithmes non-supervisés se focalisent sur les relations entre les variables pour permettre de partitionner les données en des groupes de mêmes caractéristiques.

Dans le cadre de ce mémoire, trois algorithmes supervisés sont testés pour déterminer les montants de versements périodiques : les arbres de décision CART, les forêts aléatoires et l'*Extreme Gradient Boosting*. Avant de présenter les différentes méthodes d'apprentissage, les principes généraux de *Machine Learning* sont introduits.

3.2.1 Principe d'échantillonnage

Notion de sur/sous-apprentissage

La phase d'apprentissage d'un modèle de *Machine Learning* est le processus permettant à partir de données d'entraînement d'aboutir à un modèle de prédiction le plus précis possible sur des données nouvelles. Néanmoins, plus le modèle optimisé est complexe, plus le risque de sur-apprentissage est élevé. Un modèle trop proche des données d'entraînement aboutit à une solution trop particulière, ce qui conduit à prédire de mauvais résultats sur de nouvelles entrées. Au contraire, construire un modèle trop général qui n'introduit pas assez de complexité conduit à une situation de sous-apprentissage. Si le modèle paramétré ne s'adapte pas assez aux données d'apprentissage, les erreurs de prédiction seront importantes. Le modèle obtenu ne permettra pas d'être performant sur de nouvelles données.

L'erreur de prédiction en x d'un modèle \hat{f} se mesure à travers la formule :

$$Err(x) = Biais^2 + Variance + Erreur irréductible$$

avec $Biais(\hat{f}(x)) = E(\hat{f}(x) - f(x))$ et $Variance((\hat{f}(x)) = E(\hat{f}(x)^2) - E(\hat{f}(x))^2$.

Le biais est l'écart entre la variable réellement observée et la valeur moyenne de la prédiction. Réduire le biais conduit à générer des modèles très complexes qui sont par la suite capables d'expliquer la majorité des variations de la variable cible par rapport aux données d'apprentissage. Par opposition, la variance représente la capacité de généralisation du modèle. Réduire la variance conduit à des modèles plus indépendants de la base d'apprentissage, cela permet de généraliser les estimations. En rendant le modèle moins complexe, le biais augmente. Inversement, réduire le biais augmente la variance. Estimer le modèle optimal consiste à minimiser l'erreur en déterminant le meilleur compromis biais / variance du modèle.

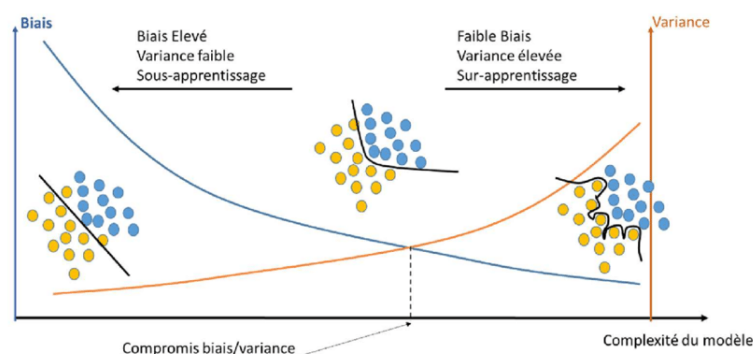


FIGURE 3.5 – Compromis biais variance (Source : Introduction au Data Mining - Pascal Scalart)

Echantillonnage

Sur des bases de données suffisamment importantes, les données sont réparties en trois groupes :

- les données qui constituent la base d'apprentissage pour ajuster les paramètres du modèle ;
- une base de validation qui est utilisée pour comparer les modèles ;
- un échantillon test qui permet d'évaluer la performance et la robustesse du modèle sur de nouvelles observations.

Dans certains cas, la taille de l'échantillon n'est pas suffisante pour constituer une base d'apprentissage. La technique de validation croisée divise de façon aléatoire les données d'entraînement en sous-échantillons. Pour k sous-échantillons, k modèles sont générés au cours de la validation croisée. Le processus effectue l'apprentissage du modèle sur les $k - 1$ segments de la base, puis le k -ème segment est utilisé comme base de validation pour mesurer différents indicateurs de performance. Le modèle de validation croisée retourne l'erreur moyenne de chacun des k modèles.

Optimisation du modèle

Afin d'obtenir de meilleures prédictions, les paramètres des modèles ont besoin d'être ajustés. Une méthode appelée *Grid Search* est utilisée pour trouver les paramètres optimaux. Parmi une série prédéfinie de valeurs possibles pour chaque paramètre, l'algorithme détermine la combinaison optimale qui minimise l'erreur de prédiction. La performance du modèle est évaluée pour chaque combinaison de valeurs possibles. Les paramètres optimaux sont obtenus à partir de la combinaison qui aboutit aux meilleurs résultats de performance. L'algorithme *Grid Search* est un moyen simple pour trouver les paramètres optimaux d'un modèle, mais il peut être très coûteux en temps de calcul pour des modèles complexes avec de nombreux paramètres.

3.2.2 Arbre de régression

Les arbres de régression de type CART (*Classification And Regression Trees*) ont été introduits par Breiman et al. (1984) [9]. Le principe des arbres CART est de prédire une réponse Y (ou variable cible) à partir d'un ensemble de variables explicatives X_1, X_2, \dots, X_p . Les arbres de décision sont qualifiés d'arbres de régression lorsque la variable d'intérêt est quantitative et d'arbres de classification lorsque la variable à prédire est qualitative. Nous nous intéressons à la construction des arbres de régression.

L'intérêt des arbres de décision est de pallier les limites des prédicteurs linéaires utilisant une unique formule pour décrire la relation entre la variable d'intérêt et l'ensemble des données. Les modèles linéaires sont moins efficaces face à des données présentant beaucoup de modalités interagissant entre elles et avec des interactions non linéaires.

Le but des arbres CART est de partitionner les données afin de former des groupes d'individus ou de contrats avec les mêmes caractéristiques. Puis, pour chaque groupe déterminé, le modèle prédit une réponse concernant la valeur de la variable d'intérêt. Le partitionnement s'effectue de manière binaire sur les variables explicatives afin de former un ensemble de sous-partitions qui aboutira à une prédiction optimale. L'implémentation du modèle débute par la construction de l'arbre maximal, qui est ensuite élagué afin de former le sous-arbre optimal. Cette construction permet d'éviter le surapprentissage du modèle et de former une solution générale et robuste sur d'autres bases de données.

Construction de l'arbre de régression

Un arbre de décision est formé d'une racine, de nœuds et de feuilles. Ces éléments sont reliés par des branches. Le nœud initial (racine) représente l'ensemble des individus. Chaque nœud correspond à une condition à vérifier permettant de segmenter la population. Les branches relient la condition à une sous-arborescence à droite (fils de droite) et une autre à gauche (fils de gauche). Une feuille représente alors un nœud terminal qui n'a pas été segmenté. Les sous-partitions optimales trouvées par le modèle sont représentées par les feuilles de l'arbre qui déterminent les classes de sous-populations et fournissent une estimation de la variable cible. Enfin, la profondeur de l'arbre est évaluée par la distance maximale entre la racine et une feuille.

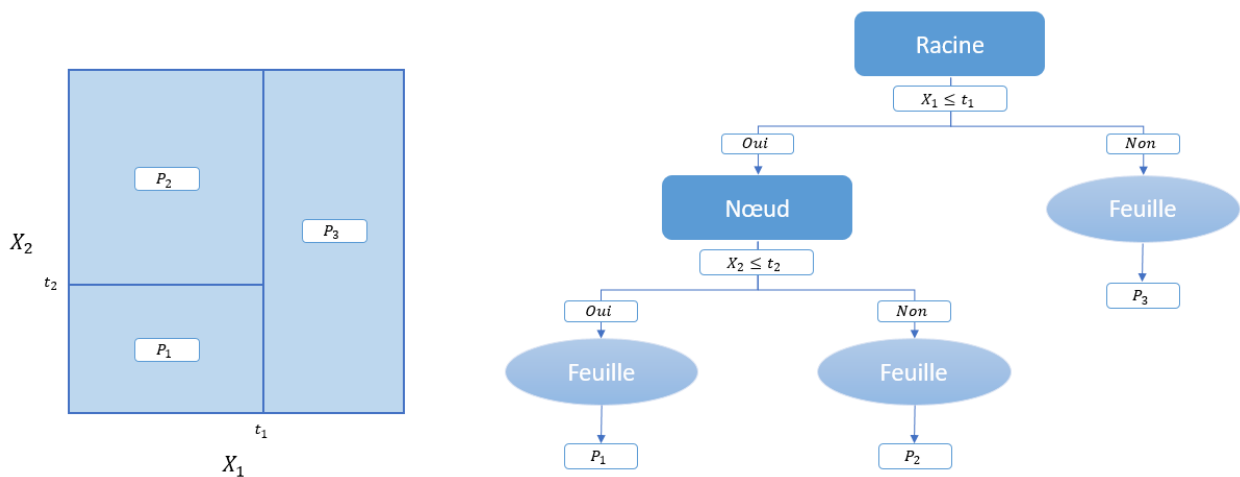


FIGURE 3.6 – Exemple d'un arbre CART sur un partitionnement en dimension 2

L'approche par CART repose sur le partitionnement binaire des individus de façon récursive. Sans critère d'arrêt, l'arbre maximal est construit jusqu'à ce qu'il ne reste plus qu'un individu par nœud. Il est possible de mettre en place une règle d'arrêt spécifique :

- sur la profondeur maximale ;
- sur un poids de feuilles minimal qui impose une proportion minimale d'individus par feuille ;
- en pénalisant la progression de l'arbre par un paramètre de complexité, la construction de l'arbre s'arrête si le nombre de nœuds formés est supérieur au critère de complexité.

L'algorithme cherche à chaque itération la meilleure segmentation caractérisée par le couple (variable, seuil) qui minimise une certaine fonction de perte. En régression, la fonction de perte utilisée est l'erreur quadratique moyenne, ce qui revient à minimiser la variance intra-groupes des deux nœuds fils N_L et N_R issus du nœud parent.

En considérant un jeu de données de n observations (ou individus), chaque observation i dans $1, \dots, n$ est caractérisée par le couple (X_i, Y_i) avec Y_i la réponse observée et X_i le vecteur des variables explicatives de l'individu i . Ce vecteur est composé de p variables, et est noté $X_i = (X_i^1, \dots, X_i^p)$. La variance d'un nœud N_m est définie par :

$$V(N_m) = \frac{1}{n_m} \sum_{i \in N_m} (Y_i - \bar{Y}_m)^2$$

$$\text{avec } \bar{Y}_m = \frac{1}{n_m} \sum_{i \in N_m} Y_i.$$

\bar{Y}_m représente la moyenne des observations présentes dans le nœud N_m et n_m le nombre d'observations à l'intérieur de ce même nœud N_m .

La division d'un nœud est représentée par un couple (j, d) formant l'ensemble $\{X_j \leq d\} \cup \{X_j > d\}$ avec X_j la variable de segmentation $j \in 1, \dots, p$ et d le seuil de coupure de la variable. Toutes les observations qui ont une valeur de X_j plus petite que d sont attribuées au nœud fils de gauche $N_L = \{X_j \leq d\}$ et celles dont la valeur est plus grande que d dans le nœud fils de droite $N_R = \{X_j > d\}$.

La fonction à minimiser pour déterminer le meilleur couple de variables explicatives X_j et de seuil d associé est la variance intra-nœuds :

$$V_{intra}(N_L, N_R) = \frac{1}{n} \sum_{i \in N_L} (Y_i - \bar{Y}_L)^2 + \frac{1}{n} \sum_{i \in N_R} (Y_i - \bar{Y}_R)^2 = \frac{n_L}{n} V(N_L) + \frac{n_R}{n} V(N_R)$$

La recherche de la meilleure solution par minimisation de variance intra-nœuds est répétée sur chaque nœud fils à partir de la racine jusqu'à atteindre un critère d'arrêt défini.

Élagage de l'arbre

La procédure d'élagage débute après avoir préalablement construit l'arbre maximal. Elle permet de trouver le meilleur sous-arbre élagué au sens de l'erreur de généralisation. Cette phase consiste à sélectionner parmi l'ensemble des sous-arbres élagués de l'arbre maximal celui qui présentera le meilleur compromis entre biais et variance. Un arbre réduit à sa racine (qui estime une valeur constante pour l'ensemble des individus) présente un fort biais et une faible variance. Tandis que l'arbre maximal possède une grande variance et un faible biais. Le principe est donc de supprimer des nœuds pour éviter un surapprentissage tout en gardant une certaine complexité pour garder une capacité de prédiction suffisante. La complexité du modèle est représentée par le nombre de feuilles de l'arbre.

L'idée est de réduire l'arbre en introduisant un paramètre de coût de complexité. L'élagage pénalise l'erreur d'ajustement d'un sous-arbre T par une fonction proportionnelle au nombre de feuilles $|T|$. En effet, sans paramètre sur la taille de l'arbre, la meilleure solution retournée serait l'arbre maximal avec l'erreur d'ajustement minimal. Le critère d'erreur s'écrit :

$$C_\alpha(T) = E(T) + \alpha|T|$$

avec $E(T)$ défini comme l'erreur de mauvais classement du nœud terminal et α le paramètre de complexité.

L'objectif de l'élagage est de trouver le sous-arbre qui minimise C_α . Plus α est grand, plus le nombre de feuilles diminue. L'arbre maximal est retrouvé par la valeur nulle du paramètre α .

Avantages et limites

D'un point de vue opérationnel, l'arbre CART reste un modèle facile à implémenter et à interpréter par la présentation claire de sa segmentation. Ce modèle d'arbre est adapté aux larges jeux de données composés d'un grand nombre de variables.

Cependant, cet algorithme ne présente pas de solution stable. En effet, une faible variation de l'échantillon d'apprentissage peut faire modifier la forme de l'arbre optimal. De plus, l'arbre optimal est construit par une succession de minimisations d'erreur locale en choisissant la meilleure variable explicative à chaque nœud. Ainsi, la solution finale ne représente pas forcément l'optimum global.

L'agrégation de plusieurs modèles est une solution pour pallier l'instabilité d'un unique modèle. Des méthodes d'agrégation telles que le *bagging* ou le *boosting* permettent de réduire la variance du modèle de prédiction et améliorer la robustesse des arbres. Ces algorithmes sont détaillés dans la suite du mémoire à travers l'implémentation des forêts aléatoires et du *gradient boosting*.

3.2.3 Forêts aléatoires

Pour comprendre le fonctionnement des forêts aléatoires, il est nécessaire de présenter le *bootstrap aggregating*, appelé *bagging*.

Bagging

Le *bagging* est une méthode d'agrégation développée par Breiman en 1994 [8] et reposant sur le principe de bootstrap. C'est une technique de rééchantillonnage permettant de former plusieurs échantillons de manière aléatoire à partir de la base d'apprentissage. Un algorithme de classification est appliqué sur chaque échantillon obtenu. Le modèle final est la forme combinée des estimateurs déterminés : la solution retournée est la moyenne des prédictions individuelles. Le principe est de diminuer la variance du modèle en agrégeant plusieurs estimateurs indépendants par l'utilisation d'échantillons sur la base d'apprentissage.

Une forêt aléatoire est un cas particulier de *bagging* appliqué aux arbres CART. Une composante aléatoire est ajoutée à chaque construction d'un arbre : seule une partie des variables explicatives est tirée aléatoirement dans la formation d'un nœud.

Construction des forêts aléatoires

La précision d'une forêt aléatoire augmente avec le nombre d'arbres estimés et la complexité du modèle.

Pour construire une forêt aléatoire, l'algorithme se déroule à partir des B arbres à estimer et du nombre k de variables sélectionnées à chaque nœud. Les étapes sont les suivantes :

1. Construire B échantillons à partir de la base d'apprentissage de taille n .
2. Pour chaque échantillon b de $1, \dots, B$, construire un arbre CART $T_{b,k}$ tel que :
 - (a) pour chaque nœud de l'arbre, k variables sont tirées aléatoirement parmi toutes les variables explicatives ;
 - (b) la segmentation optimale du nœud est choisie ;
 - (c) la construction de l'arbre continue jusqu'au critère d'arrêt.

L'estimation finale pour chaque individu X est la moyenne des décisions des arbres obtenus :

$$\hat{f}_{RF}(X, B, k) = \frac{1}{B} \sum_{b=1}^B T_{b,k}(X)$$

Importance des variables

La forêt aléatoire est une méthode qui ne permet pas d'interprétation visuelle des segmentations. Néanmoins, il est possible de mesurer l'importance de l'influence des variables explicatives dans la prédiction de la variable cible.

L'importance de chaque variable utilisée pour l'apprentissage est déterminée à l'aide de l'erreur *Out Of Bag* (OOB). L'algorithme de forêts aléatoires construit des arbres de prédiction à partir d'échantillons de données. Ainsi, il existe un ensemble d'échantillons d'apprentissage composés de données excluant l'observation i . L'erreur *Out Of Bag* est l'erreur moyenne de prédiction de l'observation i par l'ensemble des arbres formés à partir des échantillons ne contenant pas l'observation i .

L'erreur de généralisation de l'algorithme est mesurée par l'erreur *Out Of Bag* sur l'ensemble des observations :

$$Erreur_{OOB} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - Y_i)^2$$

avec \mathcal{A}_B l'ensemble des échantillons ne contenant pas l'observation i et

$$\hat{y}_i = \frac{1}{\text{Card}(\mathcal{A}_B)} \sum_{b \in \mathcal{A}_B} \hat{f}_{RF}(X_i, b, k)$$

L'importance de chaque variable est aussi basée sur l'erreur OOB. Soit OOB_b l'échantillon *Out of Bag* associé à l'arbre b , l'erreur de prédiction associée s'écrit :

$$\text{Erreur}_{OOB_b} = \frac{1}{\text{Card}(OOB_b)} \sum_{i \in OOB_b} (T_{b,k}(X_i) - Y_i)^2$$

Pour mesurer l'importance de la variable j , une perturbation est générée dans l'échantillon : les valeurs de la $j^{\text{ème}}$ variable sont permutées de manière aléatoire dans l'échantillon OOB_b . L'erreur *Out of Bag* est calculée sur les observations perturbées X_i^j :

$$\text{Erreur}_{OOB_b^j} = \frac{1}{\text{Card}(OOB_b^j)} \sum_{i \in OOB_b^j} (T_{b,k}(X_i^j) - Y_i)^2$$

Enfin, si la différence entre $\text{Erreur}_{OOB_b^j}$ et Erreur_{OOB_b} est importante alors cela signifie que l'influence de la variable j est déterminante dans la prédiction du modèle.

$$\text{Importance}_j = \frac{1}{B} \sum_{b=1}^B (\text{Erreur}_{OOB_b^j} - \text{Erreur}_{OOB_b})$$

Avantages et limites

En utilisant plusieurs arbres de prédiction, la méthode de forêts aléatoires présente généralement de meilleures performances en terme de généralisation de modèle. Les forêts aléatoires permettent de former un estimateur performant et robuste.

Malgré une meilleure capacité de prédiction, les forêts aléatoires fournissent peu d'outils d'interprétation hormis une hiérarchisation de l'importance des variables. Les temps de calculs sont aussi plus onéreux comparés à ceux d'un arbre un arbre CART, d'autant plus si la base de données est large ou que les variables explicatives sont nombreuses.

3.2.4 L'Extreme Gradient Boosting

Boosting

Le principe du *boosting* consiste, tout comme le *bagging*, à améliorer la prédiction en agrégeant plusieurs modèles de prédictions. La différence réside dans la manière de former les modèles : il n'y a plus d'échantillons bootstrap, la base d'apprentissage est utilisée dans sa globalité à chaque itération. L'idée est qu'à chaque nouvelle itération, le modèle s'améliore en mettant plus de poids sur les observations avec les plus mauvaises prédictions. Le modèle cherche à corriger les erreurs des modèles précédents en se concentrant sur les observations les plus difficiles à prédire. Ce qui permet au modèle de réduire les erreurs de prédictions au fur et à mesure. Le modèle final est l'agrégation selon une pondération basée sur la qualité de prédiction de chacun des modèles. Ainsi, à la différence du *bagging*, les modèles sont construits en série et non de manière parallèle, le but étant de présenter des solutions plus performantes de manière séquentielle.

En considérant un premier estimateur optimisé T_1 , l'étape suivante du *boosting* consiste à former le prochain estimateur plus performant T_2 tel que la variable à expliquer s'écrit :

$$Y = T_1(X) + \epsilon_1(X) = T_1(X) + T_2(X) + \epsilon_2(X)$$

où ϵ_1 représente le résidu de la première estimation, soit ici l'erreur de prédiction. Le second modèle est déterminé à partir du résidu $Y - T_1(X)$.

En considérant la succession de B modèles à estimer et T_k un arbre de décision construit à l'étape k , le modèle final est une combinaison linéaire des T_k pondérée par un poids α_k :

$$\hat{Y} = f_b(X) = \sum_{k=1}^b \alpha_k T_k(X)$$

Un algorithme de *boosting* très utilisé pour sa vitesse d'exécution est l'*Extreme Gradient Boosting* (XGBoost) qui se base sur le modèle du *Gradient Boosting*.

Gradient Boosting

Le *Gradient Boosting* est un cas particulier de *boosting*, qui ajuste les poids des modèles par un algorithme de descente de gradient. Introduit par Friedman (2001) [13], le processus de descente de gradient cherche à minimiser une fonction de perte choisie, telle que l'erreur quadratique pour le cas de la régression.

L'objectif est de minimiser une fonction de perte globale L calculée de manière additive sur l'ensemble des observations.

$$L(Y, f) = \sum_{i=1}^n l(Y_i, f(X_i))$$

où l est la fonction d'erreur entre la valeur observée et la prédiction du modèle.

L'algorithme de descente de gradient repose sur le principe itératif où pour tout $k \leq B$:

$$f_k(X) = f_{k-1}(X) - \eta * \nabla l(Y, f_k(X)) \text{ avec } \nabla l(Y, f_k(X)) = \frac{\partial l(Y, f_k(X))}{\partial f_k(X)}$$

∇ est le gradient de la fonction d'erreur du modèle, autrement dit sa dérivée partielle première. Le paramètre η est le coefficient d'apprentissage, en d'autres termes le paramètre de pénalisation évitant le sur-apprentissage.

Enfin, le procédé de l'algorithme est le suivant :

1. Un premier arbre de décision est construit. L'opposé du résidu est calculé pour l'ensemble des observations $i \in \llbracket 1, \dots, n \rrbracket$: $-\nabla l(Y_i, f) = Y_i - f(X_i)$.
2. Un nouvel arbre de décision T_k est modélisé sur les résidus $-\nabla l(Y, f)$. Le gradient devient la nouvelle variable à prédire.
3. Le poids associé au nouvel estimateur est calculé par $\gamma_k = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(Y_i, f_{k-1}(X_i) + \gamma T_k(X_i))$.
4. Enfin, le prédicteur est déduit suivant la formule : $f_k(X_i) = f_{k-1} + \eta \gamma_k * T_k$.

Le *Gradient Boosting* permet de réduire le biais d'un modèle par la technique de *boosting* et s'adapte facilement à différentes fonctions de perte. Néanmoins, cet algorithme présente certaines faiblesses. En particulier, le temps d'apprentissage est très lent par rapport aux méthodes par arbres CART ou forêts aléatoires.

L'algorithme d'Extreme Gradient Boosting

L'*Extreme Gradient Boosting* (XGBoost), développé par Chen (2016) [15], est une variante du modèle précédent plus performante en temps de calcul grâce à la parallélisation des calculs des estimations de manière séquentielle. Cet algorithme présente de nombreux paramètres permettant d'être plus flexible sur les objectifs d'optimisation. L'introduction de régularisations à travers le paramétrage de pénalisations sur la complexité du modèle formé permet d'éviter le surapprentissage. Il est possible de régler les paramètres d'ajustement alpha (régularisation en norme L1), lambda (régularisation en norme L2) ou gamma pour limiter la profondeur des arbres si celle-ci n'apporte pas de meilleures performances. Cependant, l'ajustement des paramètres peut s'avérer assez long à optimiser.

Les différents algorithmes étant définis, l'utilisation de ces méthodes dans la prédiction des montants de versements périodiques est explorée par la suite.

3.3 Application sur le portefeuille

Tout d'abord, la base de données recueillies est divisée en trois échantillons :

- le premier est destiné à constituer la base d'apprentissage pour entraîner les modèles et ajuster les paramètres ;
- un échantillon de validation est utilisé pour comparer les erreurs de prédictions ;
- un échantillon de test sera utilisé pour vérifier les capacités du modèle à généraliser les résultats.

Les versements entre 2017 et 2020 serviront à 80 % pour la base d'apprentissage et à 20 % pour la base de validation. Le découpage est réalisé par un échantillonnage aléatoire simple sans remise. La robustesse du modèle sera évaluée par des prédictions du modèle final sur les versements de 2021.

3.3.1 Paramétrage des modèles

Afin de prendre en compte le montant de versements périodiques de manière homogène, l'exposition a été considérée afin d'ajuster les montants d'ancienneté nulle. Cela permet d'estimer le montant versé sur une année complète. Dans un premier temps, le montant de versements périodiques a été ajusté en le rapportant au taux de présence sur l'année d'exercice considérée. Une deuxième approche a été envisagée en incluant l'exposition en tant que variable explicative. Après avoir comparé les résultats obtenus avec les deux approches, il a été constaté que les résultats étaient meilleurs avec l'ajout de la variable exposition. En conséquence, la deuxième méthode a été adoptée.

Ainsi, le montant de versements périodiques annuel est estimé à travers les variables explicatives suivantes :

- de flux : le montant de versement initial, le montant de l'encours, la part d'euro sur l'encours total, le montant de versements libres annuel, les montants d'arbitrages, les rachats partiels ;
- de contrat : l'ancienneté, le taux d'acquisition, le réseau, le type de gestion ;
- des caractéristique de l'assuré : l'âge, le sexe, la situation familiale, la catégorie socioprofessionnelle, le code région.

L'ajustement des paramètres de chaque modèle est détaillé pour les données du produit 1. Puis, une synthèse des résultats des modèles de prédiction sur chaque produit est présentée.

Arbres de régression

Le premier arbre construit par la fonction *DecisionTreeRegressor* de la librairie *sklearn.tree* de python est l'arbre maximal sans optimisation des paramètres. Cet arbre présente un grand nombre de nœuds et de feuilles mais s'avère peu robuste dans la prédiction de montants de versements sur d'autres bases de données.

Ainsi, la première étape consiste à optimiser le paramètre *ccp_alpha* représentant la complexité utilisée pour l'élagage de l'arbre. Le sous-arbre avec la plus grande complexité de coût qui est inférieure à *ccp_alpha* est choisi. Le paramètre *alpha* optimal pour la détermination de l'arbre CART sur les données du produit 1 est fixé par la valeur 293.

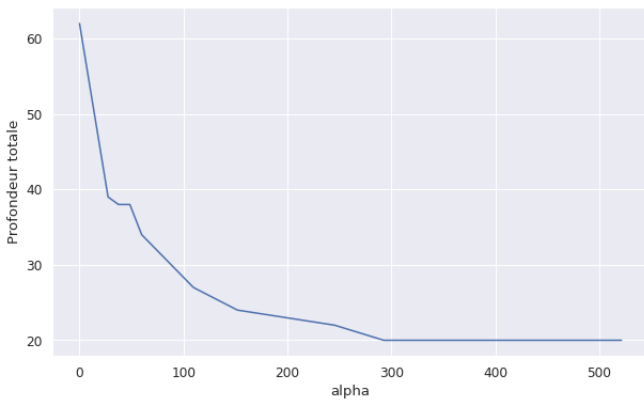


FIGURE 3.7 – Profondeur de l'arbre en fonction du paramètre de complexité

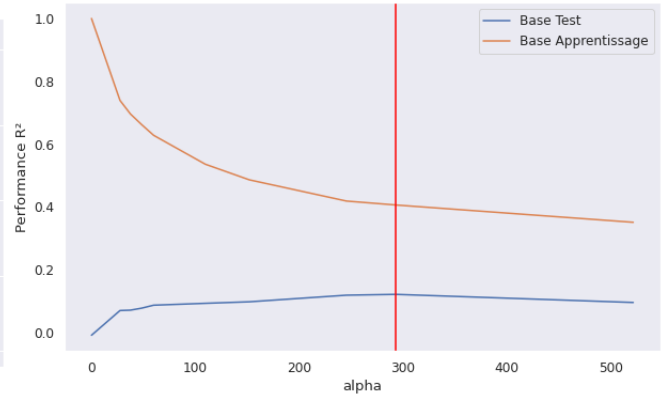


FIGURE 3.8 – Score de prédiction en fonction du paramètre de complexité sur la base d'apprentissage et de test

Plus le paramètre de complexité est grand, plus la profondeur de l'arbre est réduite (figure 3.7). La réduction de la taille de l'arbre à travers le paramètre *alpha* fait augmenter l'erreur de prédiction sur la base d'apprentissage : le coefficient de détermination R^2 diminue. Cela permet de mieux s'adapter à de nouvelles données (graphique 3.8), comme l'indique l'augmentation du score de performance sur la base de validation. L'amélioration de la prédiction est valable jusqu'à un certain seuil, au-delà duquel une nouvelle augmentation du paramètre de complexité ne permet plus de réduire les erreurs de prédiction et constitue un modèle trop général.

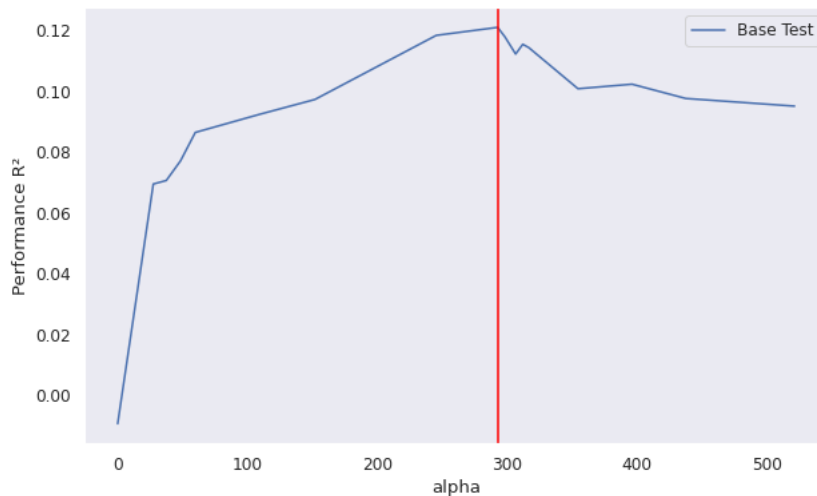


FIGURE 3.9 – Score de prédiction du test en fonction du paramètre de complexité

L'arbre CART optimisé pour le produit 1 contient 60 feuilles pour une profondeur de 17 nœuds, lorsque la complexité α est définie à 293. L'arbre présenté ci-dessous a été réduit à l'affichage de 16 feuilles en ajoutant la condition d'un poids minimum pour chaque feuille d'au moins 0,5 %. Grâce à cette condition, chaque sous-groupe de l'arbre présente un effectif supérieur à 0,5 % de la base d'apprentissage.

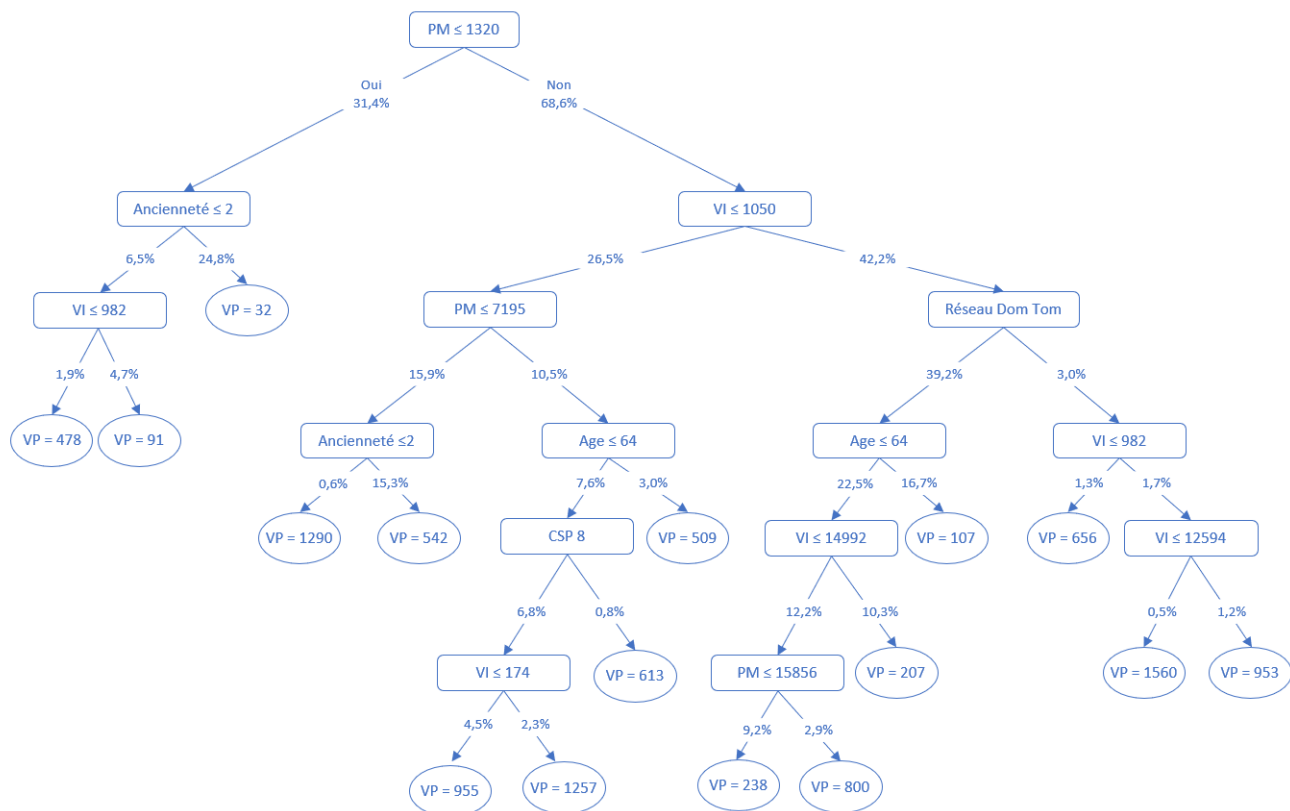


FIGURE 3.10 – Arbre optimal

Chaque nœud est représenté par un encadré avec le montant probable de versement, la part de population représentée et la condition de séparation des sous-arbres. À gauche se trouve le sous-arbre si la condition est vérifiée et à droite si elle ne l'est pas. Ainsi, les contrats avec un encours de moins de 1 352 € (condition de la première feuille) versent en moyenne moins de primes périodiques.

À partir du calcul de l'importance de chaque variable dans l'explication de la variable cible, l'encours du contrat est identifié comme étant la variable la plus importante dans la détermination du montant de versements. Le montant de versement initial y contribue aussi fortement. Les autres variables sont moins déterminantes pour le modèle.

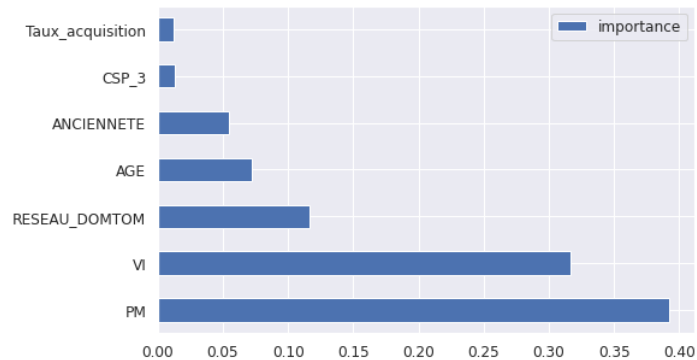


FIGURE 3.11 – Importance des variables selon le modèle CART optimisé sur le produit 1

Performances du modèle : Différentes métriques mesurant les capacités de prédiction de l'arbre optimal sont calculées sur la base de validation. Ces mêmes indicateurs sont calculés pour chaque modèle afin de pouvoir les comparer.

	RMSE	MAE	R ²
Validation	778,67	315,58	15,84 %

TABLE 3.4 – Performance de la prédiction du montant de versements par le modèle CART

En résumé, les capacités de prédiction du modèle CART sont assez faibles. Seulement 16 % des variations du montant sont expliquées par le modèle. Les forêts aléatoires sont utilisées pour tenter d'obtenir de meilleurs résultats en terme d'estimation.

Forêts aléatoires

Le modèle de forêts aléatoires combine plusieurs arbres de décisions construits sur un sous-ensemble de variables explicatives tirées aléatoirement. Chaque arbre est entraîné avec un sous-ensemble aléatoire de *features* (variables choisies aléatoirement) pour former la meilleure répartition des variables à chaque nœud.

Le modèle de forêts aléatoires a été construit en optimisant les hyperparamètres suivants à l'aide de la méthode de validation croisée :

- $\text{max_depth} \in \{10, 15, 20, 50, 75, 100\}$: ce paramètre détermine la profondeur maximale des arbres ;
- $\text{n_estimators} = 100$: pour calculer l'erreur moyenne, le modèle utilise 100 arbres ;
- $\text{max_features} \in \{2, 3, 4, 5\}$: à chaque division d'un nœud ce paramètre définit le nombre maximal de variables testées ;
- $\text{nfolds} = 5$: pour la validation croisée, ce paramètre fixe le nombre de divisions de la base de données en sous-échantillons.

Le modèle optimal construit présente une profondeur maximale de 50 et teste à chaque nœud la meilleure division parmi 5 variables.

Le choix de 100 arbres pour calculer l'erreur moyenne est justifié par la stabilisation de l'erreur *Out Of Bag* au delà de 75 arbres. En d'autres termes, ajouter plus d'arbres n'aurait pas un impact significatif sur l'estimation de l'erreur de généralisation du modèle.

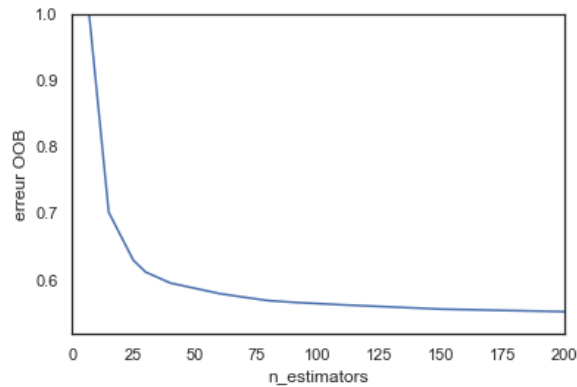


FIGURE 3.12 – Valeur de l’erreur OOB en fonction du nombre d’arbres

L’algorithme des forêts aléatoires présente un effet "boîte noire" puisque les relations entre les variables ne sont pas mesurables. Néanmoins, à partir de l’erreur *Out of bag*, il est possible de représenter l’importance des variables dans la détermination des montants de versements.

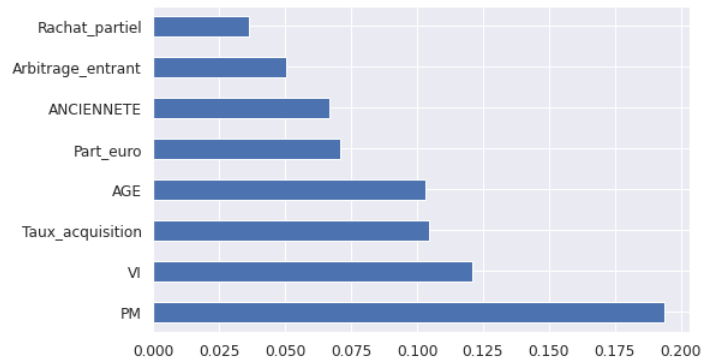


FIGURE 3.13 – Importance des variables selon le modèle de forêts aléatoires optimisé sur le produit 1

Le modèle optimal de forêts aléatoires sur le produit 1 permet de conclure que les deux variables les plus contributives sont l’encours et le versement initial. Nous retrouvons le même ordre de répartition de l’importance des deux premières variables que sur le modèle CART. Puis, le taux d’acquisition et l’âge de l’assuré sont deux autres variables discriminantes pour le modèle.

Performances du modèle :

	RMSE	MAE	R ²
Validation	579,92	212,52	53,32 %

TABLE 3.5 – Performance de la prédiction du montant de versements par le modèle de forêts aléatoires

La méthode de forêts aléatoires est nettement plus performante en terme de prédiction sur la base de validation avec un coefficient de détermination de 53 % contre 16 % pour le modèle CART.

L’Extreme Gradient Boosting (XGBoost)

L’*Extreme Gradient Boosting* est une technique d’agrégation des modèles par *boosting*. Le principe étant de pénaliser les prédictions mal classées tandis que les prédictions bien classées ont l’attribution d’un poids plus léger.

Pour limiter l'apprentissage et obtenir le modèle optimal, une étape d'ajustement des paramètres est essentielle. Les différents paramètres sont :

- `learning_rate` $\in \{0,01; 0,1\}$: représente le taux d'apprentissage compris entre $[0,1]$, il est utilisé pour contrôler la vitesse de convergence lors de la descente de gradient ;
- `max_depth` $\in \{15; 20; 30; 50\}$: pour fixer la profondeur maximale de chaque arbre ;
- `subsample` $\in \{0,5; 0,6; 0,7; 0,8\}$: indique la fraction des observations sélectionnées pour former un arbre ;
- `colsample_bytree` $\in \{0,5; 0,6; 0,7; 0,8\}$: fixe le pourcentage d'échantillonnage aléatoire sur les variables utilisées à chaque construction d'un arbre ;
- `gamma` $\in \{0; 5; 10; 100; 1000\}$: représente le paramètre de pénalisation pour régulariser les profondeurs des arbres ce qui empêche de construire des arbres trop profonds sans apport de performance suffisante.

Le modèle XGBoost optimal a pour paramètre : `learning_rate = 0,1` ; `max_depth = 20` ; `subsample=0,7` ; `colsample_bytree= 0,8` et `gamma = 0`.

Performances du modèle :

	RMSE	MAE	R ²
Validation	552,37	201,24	57,65 %

TABLE 3.6 – Performance de la prédiction du montant de versements par le modèle XGBoost

L'algorithme XGBoost est le modèle le plus performant sur les données du produit 1 : 57,65 % des variations de la variable cible sont expliquées. Les écarts lignes-à-lignes sont les plus faibles avec un RMSE à 552.

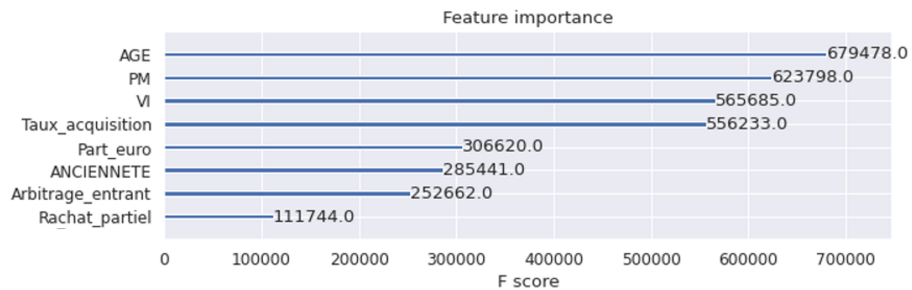


FIGURE 3.14 – Importance des variables selon le modèle XGBoost optimisé sur le produit 1

Les variables qui étaient les plus discriminantes pour les deux premiers modèles sont à nouveau présentes dans les variables les plus importantes pour le modèle XGBoost. La variable âge apparaît pour la première fois comme étant la variable qui contribue le plus fortement à la détermination du montant prédit. Puis, la PM et le versement initial sont à nouveau présents parmi les variables importantes.

3.3.2 Prédiction et synthèse comparative

Cette section présente une synthèse des capacités de prédiction de tous les modèles construits sur chaque produit afin de sélectionner le modèle le plus performant. Aussi, les variables les plus discriminantes du modèle considéré comme le plus performant sont présentées. La détermination de l'ajustement des paramètres des modèles estimés sur les observations du produit 1 a déjà permis de considérer des premiers résultats.

Produit 1

	RMSE	MAE	R ²
CART	778,67	315,58	15,84 %
Forêt aléatoire	579,92	212,52	53,32 %
XGBoost	552,37	201,24	57,65 %

TABLE 3.7 – Synthèse des indicateurs de performance des modèles sur les données du produit 1

Le coefficient de détermination de prédiction est plus satisfaisant pour les deux derniers modèles que celui obtenu avec le modèle CART. Le score passe en effet de 16 % à 50 % pour les modèles d'agrégation. La meilleure qualité de prédiction obtenue est de 50 %, autrement dit 50 % des variations des montants observés sont expliquées par le modèle. De plus, les écarts de prédiction sont plus faibles comparés aux premières modélisations statistiques.

En résumé, le modèle XGBoost est le modèle le plus performant selon tous les critères d'erreurs. Pour rappel, les quatre principales variables déterminantes pour prédire le montant de versements grâce à ce modèle sont l'âge, l'encours, le versement initial et le taux d'acquisition. L'encours est apparu fréquemment parmi les variables les plus contributives. En effet, c'est une donnée assez corrélée au montant de versements car il constitue en quelque sorte l'historique des versements passés. Le versement initial donne aussi une première idée des volumes futurs des primes périodiques.

Produit 2

	RMSE	MAE	R ²
CART	268,9	107,46	85,2 %
Forêts aléatoires	247,29	107,38	87,47 %
XGBoost	239,31	95,29	88,27 %

TABLE 3.8 – Synthèse des indicateurs de performance des modèles sur les données du produit 2

Les prédictions se révèlent bien meilleures concernant un produit avec des versements périodiques obligatoires tel que le produit 2. Le fait d'avoir une part plus importante d'assurés avec des montants positifs de primes périodiques améliore la représentativité des individus dans la base de données et permet aux modèles d'obtenir de meilleures prédictions. Le modèle XGBoost est encore celui qui présente les meilleures performances au niveau des métriques considérées.

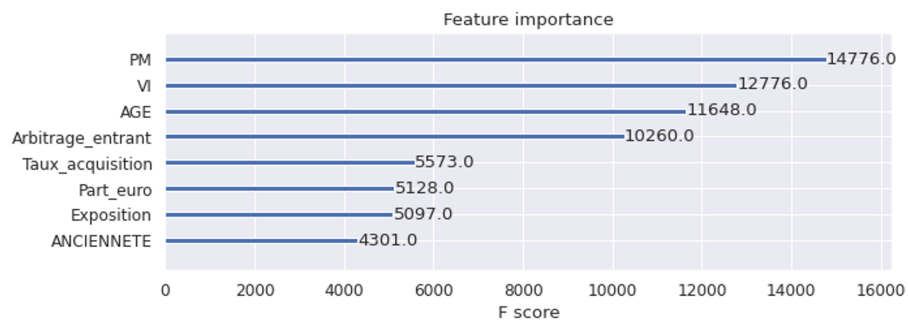


FIGURE 3.15 – Importance des variables selon le modèle XGBoost optimisé sur le produit 2

L'encours, le versement initial, l'âge et les arbitrages sont de loin les variables les plus contributives pour la prédiction des montants de versements du produit 2. Globalement, ce sont les mêmes variables explicatives que celles déterminées sur le produit 1, bien que les primes soient plus fréquentes concernant ce produit.

Produit 3

	RMSE	MAE	R ²
CART	570,78	284,79	61,97 %
Forêts aléatoires	488,07	250,21	72,19 %
XGBoost	457,72	223,42	75,54 %

TABLE 3.9 – Synthèse des modèles sur les données du produit 3

Enfin, sur ce dernier produit, le modèle XGBoost est une nouvelle fois meilleur en terme de prédiction sur la base de validation avec un score de 75 %. Les performances des modèles sur ce produit sont plus faibles que sur le produit 2 bien que la proportion de contrats versant des primes périodiques est proche. Le produit 3 est commercialisé depuis plus longtemps et présente plus de contrats qui ont diminué leurs primes contrairement au produit 2 qui est en pleine phase de commercialisation et de collecte. Les premières analyses montraient aussi que les variations des montants périodiques sur le produit 2 sont plus stables en fonction de l'ancienneté. Cette différence pourrait expliquer l'écart de performance sur la prédiction des montants. Il est possible qu'une autre explication de la différence de performance entre les produits réside dans la nature du contrat elle-même. Le produit 2 est un contrat d'épargne pure avec des montants minimums à verser, tandis que le produit 3 est un contrat d'épargne conçu spécifiquement pour une retraite supplémentaire avec des versements qui diminuent à partir de 65 ans.

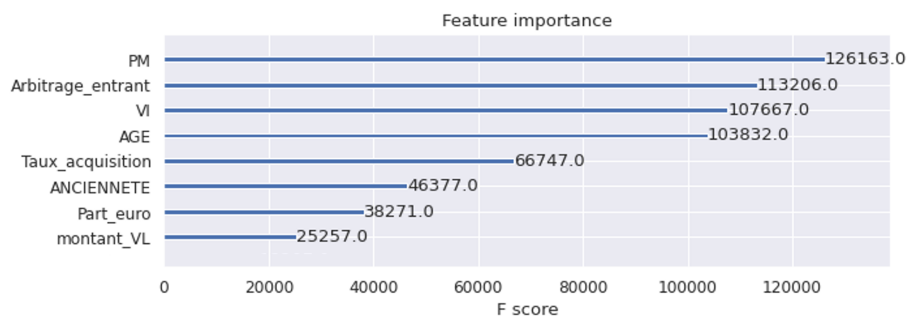


FIGURE 3.16 – Importance des variables selon le modèle XGBoost optimisé sur le produit 3

L'encours est à nouveau la variable la plus importante dans la prédiction des montants de primes périodiques. Pour le produit 3, l'arbitrage apparaît fortement dans les variables les plus contributives.

3.3.3 Validation du modèle

La validation du modèle consiste à vérifier que le modèle de prédiction obtenu est stable dans le temps. Après avoir entraîné et testé les différents modèles sur les données de 2017 à 2020, les données de 2021 servent de base de validation pour analyser la précision et la pérennité du modèle.

Pour chaque contrat, un montant de versements est calculé puis le montant global prédit est comparé au montant total observé en 2021.

	RMSE	MAE	R ²	Estimation du montant total
Produit 1 (XGBoost)	253,61	64,87	91,80 %	-0,02 %
Produit 2 (XGBoost)	187,57	77,31	92,81 %	0,03 %
Produit 3 (XGBoost)	257,21	122,62	92,34 %	-0,04 %

TABLE 3.10 – Indicateurs de performance sur les projections des montants sur les données de 2021

Les montants prédits totaux sont très bien estimés par la méthode XGBoost. Les modèles de prédiction ont permis de réduire les erreurs de prédiction comparés aux premiers modèles en fonction de la variable

ancienneté. En conclusion, les variables représentant l'encours, l'âge, le versement initial et l'arbitrage sont plus déterminantes que l'ancienneté. Finalement, l'ordre d'importance des variables varie peu entre les produits.

La prédiction ligne-à-ligne des montants de versements périodiques, par des algorithmes de *Machine Learning*, se révèle être performante. En pratique, l'utilisation d'un modèle XGBoost est plus difficile. D'une part, le caractère "boîte noire" de ces outils est un frein à leur application. Le manque de transparence limite la compréhension des résultats obtenus et peut poser problème face aux principes d'audit et de contrôle interne. D'autre part, au sein de l'équipe Axa France, cela demande d'adapter l'outil de calcul de rentabilité. Néanmoins, les résultats obtenus sont utilisables pour déterminer des groupes auxquels est affecté un montant probable de versements périodiques. Cela permet de distinguer les contrats qui vont verser des montants plus importants de primes périodiques des contrats qui vont en verser moins.

3.4 Regroupement des contrats

La modélisation des montants de versements est réalisée afin de pouvoir projeter les flux futurs des contrats pour calculer la rentabilité d'un produit. Pour chaque produit, le modèle de prédiction retenu met en avant les variables les plus significatives dans la détermination du montant de versements. La provision mathématique est une variable déterminante dans la prédiction du montant de versement. Cette variable est un indicateur sur les versements passés effectués, relativement aux plus ou moins-values constatées. Sur les affaires nouvelles, cette variable est égale au versement initial fixé à la souscription du contrat. Étant donné que le calcul de la rentabilité sera effectué sur les affaires nouvelles, la variable PM n'est pas gardée dans le processus de regroupement des contrats, au profit de la variable versement initial. Ainsi, la formation des groupes de contrats est effectuée en fonction des versements initiaux et de l'âge. La méthode *k-means* est utilisée pour créer des groupes de contrats avec des similitudes au regard des deux variables choisies.

Dans un premier temps, la méthode de construction des classes est présentée. Les regroupements des contrats obtenus sont utilisés pour déterminer les lois de versements périodiques propres à chaque groupe. Enfin, l'impact des lois sur la rentabilité des produits est étudié.

3.4.1 Méthode des k-means

Il existe plusieurs méthodes de regroupement proposées par la littérature, appelées aussi méthodes de *clustering*. L'algorithme *k-means* est une méthode de classification assez simple à implémenter qui permet de partitionner des observations en K classes. Cet algorithme est particulièrement apprécié pour le partitionnement de données volumineuses, mais requiert de trouver le nombre optimal de classes, ce qui peut s'avérer difficile.

Théorie de l'algorithme

Le principe de la méthode *k-means* est de diviser les observations notées X en K groupes, dans le but de rassembler les données similaires et de façon à ce que les groupes soient les plus distincts possible. Chaque partie est caractérisée par son centre de gravité, que l'on appelle centroïde. L'objectif de la méthode est de minimiser la distance intra-groupe, ce qui consiste à trouver la répartition des points dans chaque classe telle que la somme des distances par rapport à leur centre est minimale.

La distance intra-groupe est représentée ici par la distance euclidienne, qui mesure le degré de similarité entre deux données. Deux individus sont d'autant plus semblables que la distance euclidienne est petite.

Chaque individu i est représenté dans l'espace par un point X_i , avec (x_i^1, \dots, x_i^p) l'ensemble des p informations quantitatives propres à l'individu. La distance euclidienne entre deux points X_i, X_j est définie par la mesure :

$$d_2(X_i, X_j) = \|X_i - X_j\| = \sqrt{\sum_{k=1}^p (x_i^k - x_j^k)^2}$$

La somme des distances intra-groupe est une mesure de la qualité du partitionnement obtenue par la méthode *k-means*. Également appelée inertie, elle reflète la somme des distances des points à leur centre de groupe respectif. Une valeur d'inertie plus faible indique que les groupes formés sont homogènes. Plus précisément, pour chaque groupe K , l'inertie est définie comme la somme des distances euclidiennes entre chaque point par rapport à leur centroïde. L'inertie pour K groupes est définie par la formule suivante :

$$inertie = \sum_{k=1}^K \left(\sum_{X \in C_k} \|X - \mu_k\|^2 \right)$$

où C_k représente la k -ème classe avec X une observation et μ_k le centroïde du groupe C_k .

Description de l'algorithme

1. L'initialisation consiste à sélectionner K points aléatoirement parmi les données pour former K centres initiaux.
2. Puis, les observations sont associées une à une à un groupe dont le centroïde est le plus proche. Pour toute nouvelle attribution d'un point à un groupe, le centroïde du groupe est recalculé. A noter que le centre du groupe n'est pas forcément représenté par une observation de la population étudiée.
3. Chaque observation est à nouveau parcourue afin de recalculer ses distances avec les nouveaux centres de groupe. Dans le cas où le centre de groupe avec lequel l'observation est la plus proche est différent de celui qui était attribué, alors l'observation est associée à ce groupe et les centroïdes des deux groupes sont mis à jour.
4. L'étape 3 est répétée jusqu'à la stabilisation des observations dans chaque groupe. Lorsque les centres des groupes ne changent plus, la convergence de l'algorithme est atteinte. Un nombre maximum d'itération peut aussi être fixé à l'avance. Dans ce cas, la convergence n'est pas forcément atteinte.

La méthode de *k-means* est rapide et simple d'application. Une des limites de cet algorithme itératif est qu'il converge vers un optimum local, qui n'est pas nécessairement l'optimum global. De plus, l'algorithme est sensible à l'initialisation. En effet, le positionnement initial des centres de classes conditionne la formation des classes finales. Pour réduire l'aléa du positionnement dans le choix du nombre optimal de classes, l'algorithme est lancé plusieurs fois avec des initialisations différentes. Différentes répartitions des centres de classes sont testées pour un même nombre de classes déterminé. Ainsi, l'inertie finale est la moyenne de l'ensemble des inerties calculées pour chaque répartition initiale des centres de classe.

Une amélioration de cet algorithme a été proposée par David Arthur et Sergei Vassilvitskii (2007) [4]. La méthode, nommée *k-means++*, propose une initialisation des centroïdes plus efficace. En effet, comme la première étape est déterminante dans le résultat obtenu, l'algorithme permet d'initialiser les centres des classes de manière à être les plus distants des uns des autres.

1. Le premier centre est choisi aléatoirement parmi les observations.
2. Pour chaque observation non assignée comme centroïde, la distance minimale au centroïde le plus proche est calculée. Le prochain centre d'un groupe est choisi à partir du nuage d'observations avec une probabilité proportionnelle à la distance au carré calculée pour chaque observation. Ainsi,

l'observation la plus éloignée des centroïdes déjà affectés aura une probabilité plus importante d'être le nouveau centroïde.

3. La dernière étape est répétée jusqu'à atteindre le nombre de centroïdes fixés.

La suite de l'algorithme reprend les mêmes étapes que la méthode *k-means* en utilisant ces centres initiaux. Comparée à une initialisation aléatoire, l'article [4] montre que l'initialisation par la méthode *k-means++* aboutit à de meilleurs résultats.

Détermination du nombre de groupes optimal

Le critère de décision du nombre optimal de classes à choisir s'est porté sur la méthode du coude. A partir d'un graphique déterminant l'inertie moyenne en fonction du nombre de classes, le nombre optimal est le point représentant le coude de la fonction tracée. Ce point représente le nombre de groupes à partir duquel la variance ne se réduit plus significativement.

Cette méthode graphique est simple d'utilisation mais ne permet pas d'aboutir sur une solution unique et claire. Néanmoins, en fonction du contexte, elle permet de fournir de précieuses indications.

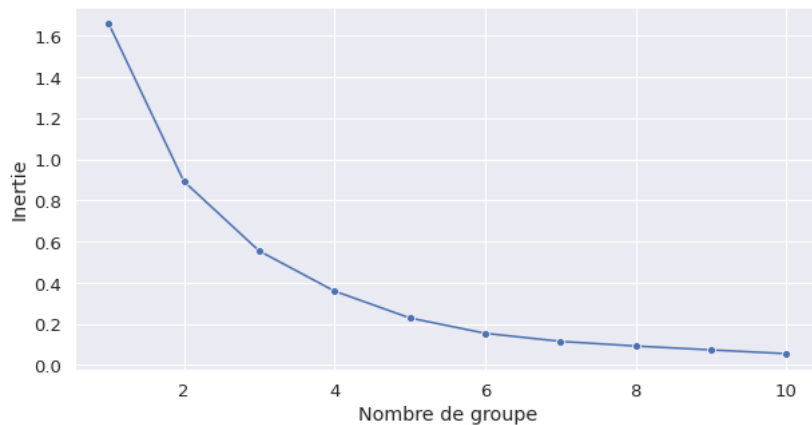


FIGURE 3.17 – Méthode du coude sur les observations du produit 1

En général, lorsque le nombre de groupes augmente, l'inertie diminue. A partir d'un certain point, l'inertie globale ne diminue plus significativement. En représentant l'inertie en fonction du nombre de classes, le point où l'inflexion est la plus forte est appelé le coude. Le nombre de groupes correspondant au coude est considéré comme le nombre optimal de classes.

Ainsi, pour le premier produit étudié, la méthode *k-means* a été appliquée aux individus en prenant en compte les variables sur le montant de versement initial et l'âge de l'assuré, qui ont été au préalable standardisées. Le nombre optimal de groupes est choisi en le représentant en fonction de l'inertie. Le graphique permet de mettre en évidence que deux ou trois groupes suffisent à être pertinents pour distinguer les individus. En effet, la perte d'inertie entre trois et quatre groupes est moins marquée qu'entre deux et trois. Trois est le nombre de groupes retenu pour le produit 1 avec cette méthode.

3.4.2 Détermination des groupes de contrats

Le nombre de groupes a été déterminé pour chaque produit par la méthode de *clustering*, ce qui permet de former des répartitions homogènes. Puis, le montant moyen de versements périodiques est calculé sur chaque classe formée. Ainsi, pour déterminer le montant de versements périodiques d'un nouvel individu, celui-ci sera tout d'abord affecté à un groupe en fonction de ses caractéristiques. Ensuite, le montant moyen de versements périodiques de ce groupe lui sera attribué.

Produit 1

Pour le premier produit, la modélisation des versements périodiques par trois profils différents a été choisie à partir du graphique 3.17. Le tracé des montants moyens versés par sous-groupes en fonction de l'ancienneté révèle que cette dernière influe sur la valeur des versements périodiques.

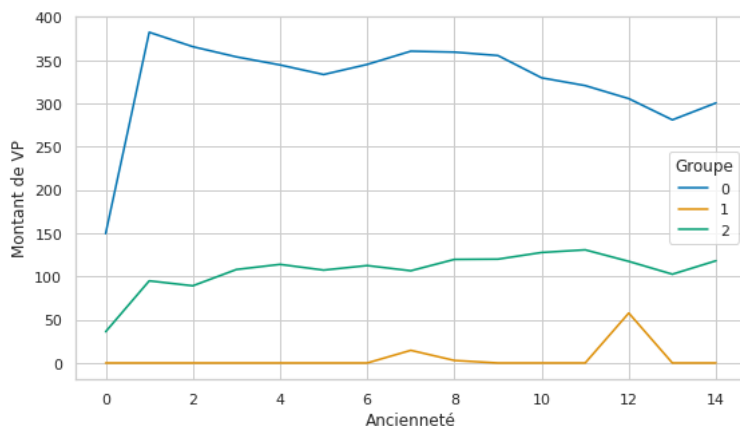


FIGURE 3.18 – Versement périodique moyen par groupe en fonction de l'ancienneté pour le produit 1

Pour caractériser les groupes, quelques statistiques sur les variables sont mises en avant.

Groupe	Répartition	Montant moyen VP	VI moyen	VL moyen	Âge moyen
0	92,29 %	333	7 413	1 316	52
1	7,24 %	103	104 790	3 538	64
2	0,47 %	6	445 681	7 324	66

TABLE 3.11 – Résultats des groupes formés sur le produit 1

La majorité des individus (92 %) est classée dans le groupe 0, qui est caractérisé par des versements périodiques plus élevés. En moyenne, ce groupe verse 333 €, avec un versement à l'ouverture du contrat de 7 413 € et des versements libres égaux à 1 316 €. Ce groupe a une moyenne d'âge de 52 ans.

Le groupe 1 est caractérisé par des versements périodiques de 100 €. Ce montant est trois fois plus faibles par rapport au groupe 0, mais accompagné par un montant à l'ouverture beaucoup plus important (104 790 €) avec des versements libres en moyenne trois fois plus élevés (3 538 €). L'âge moyen du groupe est de 64 ans.

Enfin, le dernier groupe peut sembler peu significatif avec moins de 1 % des individus du portefeuille. Ce groupe est formé d'assurés en moyenne plus âgés et avec des versements périodiques faibles. Ce groupe opte pour des versements libres plutôt que des versements périodiques : les versements libres ont pour montant moyen 7 324 €.

Ainsi, les groupes 1 et 2 sont caractérisés par un âge au dessus de 60 ans, soit plus de dix ans d'écart avec le dernier groupe. Trois niveaux de versements périodiques se distinguent : les montants élevés (groupe 0), les montants modérés (groupe 1) et les montants faibles (groupe 2).

Produit 2

A partir des critères sur le versement initial et l'âge, les individus ayant souscrit au produit 2 ont été regroupés en trois classes significatives.

Groupe	Répartition	Montant moyen VP	VI moyen	VL moyen	Âge moyen
0	10,41 %	2 101	468	86	52
1	51,31 %	589	122	60	47
2	38,28 %	1 259	256	68	51

TABLE 3.12 – Résultats des groupes formés sur le produit 2

Tout d’abord, les individus sont plus équitablement répartis dans les différents groupes que pour le produit 1 :

- Le groupe 0, regroupant 10 % des individus, est caractérisé par des montants plus élevés (montant de versements périodiques de 2 101 € et un versement initial de 468 €) pour un âge moyen de 52 ans.
- Le groupe 1 est formé par 51 % des individus ayant souscrit à ce produit. Ce groupe est caractérisé par des versements périodiques et initiaux faibles, pour une catégorie de personne ayant une moyenne d’âge de 47 ans.
- Le groupe 2 est un groupe intermédiaire avec un montant moyen de versements périodiques de 1 259 € pour un versement initial de 256 €. L’âge moyen du groupe est de 51 ans.

Contrairement au produit 1 pour lequel les montants se compensaient (si les versements périodiques sont plus élevés cela se fait au détriment du versement initial, et inversement) les groupes du produit 2 représentent plutôt des niveaux de richesses et de capacité à épargner. Les âges moyens des groupes sont proches. Trois profils se dégagent avec une capacité d’épargner qualifiée de faible, moyenne ou élevée.

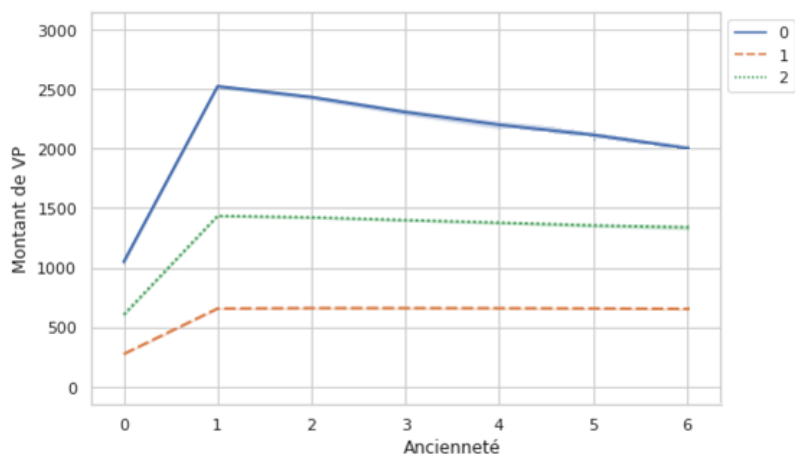


FIGURE 3.19 – Versement périodique moyen par groupe en fonction de l’ancienneté pour le produit 2

En matière d’évolution des montants en fonction de l’ancienneté, les trois groupes présentent des variations similaires. La diminution des versements périodiques au cours du temps est plus marquée pour le groupe 0.

Produit 3

Les assurés versant leur épargne sur le produit 3 ont été classés en deux groupes distincts.

Groupe	Répartition	Montant moyen VP	VI moyen	VL moyen	Âge moyen
0	97,67 %	725	196	260	51
1	2,33 %	1 434	3 356	498	58

TABLE 3.13 – Résultats des groupes formés sur le produit 3

Le premier groupe, rassemblant la majorité des individus, est caractérisé par un âge moyen de 51 ans, avec un versement initial de 196 € et un montant périodique de 725 €. Les individus du second groupe sont plus âgés (en moyenne 58 ans) et effectuent des versements plus élevés avec un versement initial de 3 356 € et un montant périodique deux fois plus important (1 434 €) que le premier groupe.

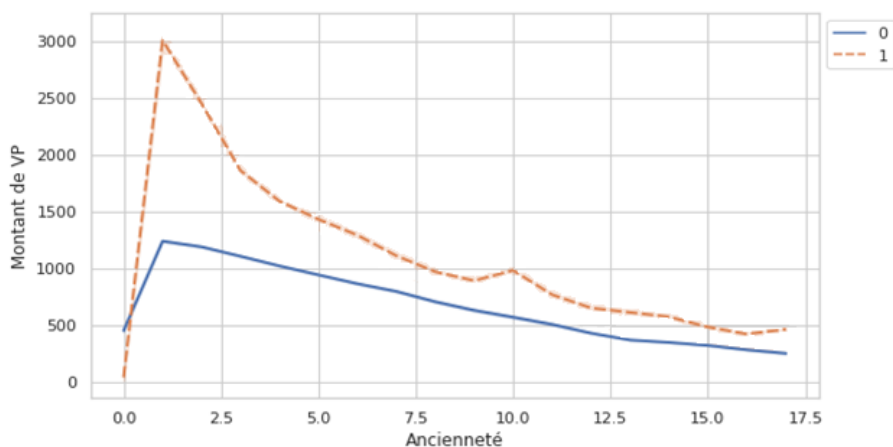


FIGURE 3.20 – Versement périodique moyen par groupe en fonction de l’ancienneté pour le produit 3

Les individus du groupe 1 versent en moyenne des montants supérieurs à ceux du groupe 0, quelle que soit l’ancienneté considérée. Ce groupe représente 2,3 % de la population globale observée.

Pour conclure, les modèles de *Machine Learning* n’étaient pas compatibles avec le modèle de projection de rentabilité. En effet, à chaque pas de projection, l’algorithme de prédiction détermine pour chaque assuré un montant estimé de versements. De plus, plusieurs scénarios financiers sont testés lors de la projection des flux, ce qui au global crée une complexité de calcul plus importante. Ainsi, la modélisation des flux de versements périodiques à travers le regroupement des individus a été privilégiée. Cette méthode permet d’ajouter de la précision sur le comportement des primes périodiques à partir des variables sélectionnées par le modèle de *Machine Learning*. Pour évaluer la rentabilité, les lois d’évolution des flux de versements périodiques qui sont retenues se composent d’un montant moyen calculé sur la première année d’ancienneté pour chaque groupe d’un produit, ainsi qu’un vecteur qui décrit les évolutions de ce montant en fonction de l’ancienneté. Les vecteurs d’évolutions sont construits à partir des variations observées sur les différents graphiques (3.18, 3.19, 3.20).

Chapitre 4

Rentabilité des produits

L'objet de ce mémoire est dans un premier temps de modéliser le comportement client sur les versements complémentaires périodiques, et puis dans un second temps d'étudier l'impact de la mise à jour des lois d'évolution des versements périodiques sur la rentabilité estimée des produits. Ainsi, cette partie présente les lois utilisées pour la projection des primes et des engagements de l'assureur afin de déterminer les résultats attendus et les indicateurs de rentabilité associés.

Le suivi de la rentabilité des produits existants permet de s'assurer que les profits générés sont suffisants face aux coûts portés par l'assureur et que le provisionnement couvre les engagements futurs envers les assurés. Les études de rentabilité interviennent lors :

- du lancement de nouveaux produits ;
- des modifications ou ajouts d'options sur des produits existants ;
- du suivi de rentabilité sur les produits en fonction des hypothèses prises en compte sur des scénarios financiers, des niveaux de frais appliqués, de clause de participation aux bénéfices, etc. En effet, les hypothèses du modèle de rentabilité peuvent changer au cours du temps à l'issue d'un choix du groupe ou d'une évolution de la réglementation.

Ces études de rentabilité permettent de mesurer l'impact de nouvelles stratégies sur la profitabilité estimée des contrats. Différents indicateurs permettent de s'assurer de la production d'une certaine richesse. Rapportés au chiffre d'affaires, ils peuvent par exemple permettre de comparer les produits entre eux. L'objectif principal pour la compagnie est de s'assurer de générer une marge suffisante pour compenser les frais généraux de l'entreprise.

Ainsi, il est nécessaire d'utiliser des lois robustes, en particulier sur le développement des versements futurs afin de mettre en évidence les potentielles sources de marges durant la durée de projection (60 ans).

4.1 Modélisation de la rentabilité des produits

Les produits d'assurance-vie et de retraite individuelle commercialisés par AXA France sont regroupés dans un même modèle de rentabilité. Pour diminuer le temps de calcul du modèle de rentabilité sur le portefeuille, les contrats sont rassemblés pour former des model points représentant différents profils d'assurés. Un model point est construit de façon à regrouper des contrats avec des probabilités de sortie (décès ou résiliation), des caractéristiques contractuelles et des probabilités de versements qui sont similaires. Ainsi, le modèle final permettra de distinguer des lois de versements en fonction des groupes d'individus qui auront des comportements semblables.

4.1.1 Présentation du modèle utilisé

Chaque *model point* est caractérisé par des paramètres qui lui sont propres (âge moyen, montant moyen de versements initiaux, périodiques et libres moyens, répartition UC/Euro de l'investissement sur les contrats). Des lois statistiques calibrées à partir de l'expérience du portefeuille sont ensuite appliquées à chaque pas de projection en fonction de l'ancienneté du contrat ou de l'âge moyen.

Les hypothèses suivantes sont utilisées dans le modèle de rentabilité à un pas de temps annuel :

- Hypothèses techniques :
 - Tables de mortalité
 - Lois de rachats partiels et totaux
 - Lois de versements libres
 - Lois des versements périodiques construites et présentées dans le chapitre 3
 - Lois d'arbitrages euros vers UC et UC vers euros
 - Coûts unitaires moyens d'acquisition et de gestion
 - Frais de gestion
 - Garanties associées au produit
- Hypothèses financières :
 - Tables de gestion des encours
 - Clause de participation aux bénéfices
 - Chroniques de taux de produits financiers des segments du fonds euros
 - Performances des indices de marchés (Eurostoxx 50, Footsie, Nikkei, Inflation, S&P500)
 - Prix des zéro-coupons à différentes maturités et pour différentes notations.

Les trois dernières hypothèses financières sont utilisées pour déterminer les performances attendues des supports d'investissement.

La projection des performances des fonds euros est issue de chroniques de taux de produits financiers fournis par les équipes du modèle interne.

Concernant les performances des supports UC, elles sont construites à travers une combinaison d'OPCVM actions avec un indice immobilier et d'OPCVM de taux. Les OPCVM actions sont modélisées par des générateurs de scénarios économiques sur plusieurs indices (Eurostoxx 50, Footsie, Nikkei, S&P500 et Property). Tandis que les OPVCM de taux sont construits à partir de chroniques de prix des zéro-coupons.

L'évolution des flux des *model points* est modélisée à travers 4 000 scénarios financiers. La projection des provisions mathématiques permet d'estimer les chargements et marges financières. Chaque scénario engendre des résultats annuels futurs qui sont actualisés pour former la *Value of Inforce* (VIF). La VIF globale est la moyenne des VIF obtenues sur chaque scénario.

4.1.2 Lois de comportement client

Au cours de la vie d'un contrat, les montants épargnés sont susceptibles d'évoluer en fonction des choix de l'assuré. Des lois sur le comportement de l'assuré en fonction de variables (ancienneté du contrat, âge de l'assuré, etc.) évaluent la probabilité de survenance d'un événement susceptible d'affecter un contrat, tel qu'un rachat ou un versement de prime par exemple. Ces lois permettent d'estimer l'engagement de l'assureur tout au long de la projection.

Taux de rachats

Les lois de rachats partiels et rachats totaux viennent diminuer la provision mathématique à chaque pas de projection. Les rachats totaux ont également pour effet de réduire la présence probable des individus dans le portefeuille, ce qui réduit à son tour la probabilité d'effectuer des versements futurs. En revanche, les rachats partiels ne réduisent que la partie de la provision correspondant à la fraction rachetée.

Arbitrages

Les mouvements d'arbitrages entre supports sont modélisés par des lois d'arbitrages du support euros vers les supports en UC et des supports UC vers le support euros.

Mortalité

Chaque année projetée, le nombre de contrats est réduit du taux de mortalité correspondant à une personne dont l'âge est égal à l'âge moyen du *model point*.

Projection des primes

Concernant la projection des primes, elles sont supposées être versées en début de période. Ce choix est plus prudent au regard du provisionnement (le montant provisionné sera plus grand en considérant les primes en début de période). Le versement futur des primes complémentaires qu'elles soient libres ou périodiques dépendra de la probabilité de présence définie par les lois de mortalité et des rachats totaux. Dans le modèle utilisé, la projection des primes ne dépend pas de scénarios financiers.

- Le versement initial est modélisé par moyenne et est versé de façon sûre la première année.
- Les versements périodiques sont calculés jusqu'à la fin de la projection en considérant le taux de personnes présentes dans le portefeuille en fin de période précédente et un montant moyen fixe de versements. Le montant total fluctue en fonction de lois complémentaires basées sur le taux des contrats générant des versements périodiques en observant l'historique des versements par produit. Ces lois complémentaires dépendent de l'ancienneté des contrats. Plusieurs sensibilités sont effectuées avec nos modèles pour modifier les lois de comportement sur ces primes.
- Les versements libres sont modélisés par un montant moyen de versement en fonction du produit et un taux moyen de versement libre déterminé par ancienneté et par produit.

Pour comparer les taux de rentabilité de différents produits d'assurance comportant des versements de primes différents, deux indicateurs de niveaux de primes sont utilisés. Il s'agit de :

- L'**APE** (Annualized Premium Equivalent), qui équivaut à une année de primes. Celui-ci est calculée en prenant en compte la moyenne des primes versées par l'assuré sur les dix premières années.
- La **PVEP** (Present Value of Expected Premiums), qui représente la somme actualisée de l'ensemble des primes versées par l'assuré.

4.1.3 Déroulé et projection des réserves

Les provisions mathématiques sont établies en début et en fin de période. Le déroulé des provisions mathématiques commence par la dernière provision connue :

$$\text{PM au début : } PM_{début} = \begin{cases} 0 & t = 0 \\ PM_{fin}(t-1) & t \geq 1 \end{cases}$$

Pour chaque support i , une réserve est associée en fonction de la répartition (notée R_i) de l'épargne sur le support.

$$PM_{début,i}(t) = PM_{fin}(t-1) \times R_i(t)$$

La PM en fin de période est calculée selon les étapes suivantes.

PM après versement des primes : avec $Primes(t)$ l'ensemble des versements initiaux, périodiques et libres à chaque pas de temps t , la $PM_{début}$ est augmentée de la part de prime en fonction de la répartition sur chaque support et est diminuée des chargements prélevés lors du versement.

$$PM_{Primes,i}(t) = PM_{début,i}(t) + R_i(t) \times Primes(t) \times (1 - Chargements_i)$$

PM après arbitrages : les arbitrages sont modélisés entre le support en euros et les fonds en unités de compte. Les arbitrages entrants sur le support sont nets de frais, et les flux d'arbitrages sortants sont bruts de frais. La provision est calculé sur chaque support UC modélisé k .

$$PM_{Arbitrages}(t) = PM_{Arbitrages,Euro}(t) + \sum_k PM_{Arbitrages,UC_k}(t)$$

avec $PM_{Arbitrages,Euro}(t) = PM_{Primes,Euro}(t) + PM_{Arbitrages\ Nets,UC \rightarrow Euro}(t) - PM_{Arbitrages\ Bruts,Euro \rightarrow UC}(t)$

$$PM_{Arbitrages,UC_k}(t) = PM_{Primes,UC_k}(t) + PM_{Arbitrages\ Nets,Euro \rightarrow UC_k}(t) - PM_{Arbitrages\ Bruts,UC_k \rightarrow Euro}(t)$$

PM après prestations : les prestations dues aux assurés, autrement dit les rachats partiels, les rachats totaux et les décès sont pris en compte et viennent diminuer les réserves totales.

$$PM_{Prestations,i}(t) = PM_{Arbitrages,i}(t) - Prestations_i(t) - RachatsTot_i(t) - RachatsPart_i(t)$$

PM après revalorisation : la provision finale est revalorisée en fonction des scénarios financiers.

$$PM_{Fin,i}(t) = PM_{Prestations,i}(t) + Revalo_i(t)$$

4.1.4 Déroulé des résultats

Les résultats statutaires ou résultats nets sont l'ensemble des marges et des coûts liés au produit. Autrement dit, ils comprennent l'ensemble des chargements prélevés, les marges sur la performance financière, les commissions d'acquisition, les coûts de gestion, les impôts et une participation aux frais généraux de l'entreprise. Les résultats statutaires générés par le produit déterminent en grande partie les indicateurs de rentabilité.

Les marges et les coûts à calculer s'expriment en fonction des primes et des réserves, il suffit donc de dérouler les primes et les PM du produit pour calculer les résultats statutaires de chaque année.

Pour ce qui est des résultats, un contrat est caractérisé par :

- une première année déficitaire due aux coûts d'acquisition importants du nouveau contrat ;
- des résultats généralement positifs les années suivantes permettant de dégager des profits.

Le fonctionnement de la projection des réserves dans le modèle ayant été détaillé, la prochaine section permet de définir les indicateurs de rentabilité qui seront utilisés.

4.2 Les indicateurs de rentabilité

Pour comparer différents produits entre eux, une entité peut s'appuyer sur plusieurs indicateurs de rentabilité, qui se présentent comme des taux pouvant être rapportés aux primes, au chiffre d'affaires, etc.

4.2.1 Définition de la New Business Value

La *New Business Value* (NBV) représente les gains attendus par l'actionnaire. Elle est calculée comme la somme actualisée des résultats statutaires à partir du lancement du produit. Elle est mesurée dans un environnement dit "risque neutre" où tous les actifs rapportent le même rendement moyen, à savoir celui du taux d'emprunt d'État, considéré comme le taux sans risque.

Ainsi, cet indicateur est composé :

- du *Strain*, soit les résultats de l'année 0, généralement négatifs dus aux frais prélevés la première année qui sont insuffisants face aux coûts engendrés et aux commissions redistribuées à l'apporteur de l'affaire ;
- de la VIF (*Value of Inforce*) stochastique qui est calculée à partir de la somme des flux futurs actualisés. La VIF représente la valeur du portefeuille une fois le produit lancé.

La VIF peut-être calculée de deux façons. La VIF stochastique est obtenue par la moyenne des VIF sur 4 000 scénarios dans un environnement risque neutre. Quant à la VIF déterministe, cette valeur est déterminée sur un unique scénario central équivalent certain.

L'écart entre la VIF déterministe et la VIF stochastique est mesuré par la valeur temps des options et garanties intégrées dans un contrat telles que le taux minimum garanti, la participation aux bénéfices ou le rachat. Cette valeur est appelée TVOG (*Time Value of Options and Guarantees*). La TVOG est calculée par la différence entre le prix de l'option et sa valeur intrinsèque.

Ainsi, la NBV s'écrit :

$$NBV = Strain + VIF_{stochastique} = Strain + VIF_{déterministe} + TVOG \quad (4.1)$$

Enfin, la *New Business Value margin* (NBVm) est la NBV rapportée à une année de primes, soit :

$$NBV_m = \frac{NBV}{APE} \quad (4.2)$$

Cet indicateur est une mesure importante de la rentabilité d'un produit. Étant rapporté à la première année de primes, il permet de comparer différents produits entre eux.

Les autres indicateurs nécessitent de calculer le capital immobilisé par la réglementation, et plus généralement les résultats distribuables.

Les résultats dits distribuables intègrent en plus des résultats nets le coût du capital immobilisé. Le capital immobilisé représente le capital minimum requis détenu par une entreprise d'assurance selon la directive Solvabilité 2. Le coût du capital immobilisé est calculé comme la marge minimal qui aurait pu être dégagée en investissant ce capital.

4.2.2 Le SCR

Le premier pilier de la directive Solvabilité 2 impose des exigences quantitatives dont le calcul des provisions techniques, d'un capital minimal requis (MCR) au niveau des fonds propres et du SCR (*Solvency*

Capital Requirement). Le SCR est le montant de capital à immobiliser pour faire face aux risques auxquels l'assureur est exposé. Le capital de solvabilité requis correspond au montant minimal de fonds propres permettant de maîtriser la probabilité de ruine à un an à 0,5 % ou moins.

Le SCR en formule standard est calculé à partir d'une grille de risques auxquels des chocs bicentennaires sont appliqués. L'agrégation des capitaux requis pour limiter les pertes liées à chaque facteur de risque est effectuée via une matrice de corrélation. Cette matrice prend en compte les interactions entre les différents facteurs de risque. Le BSCR (*Basic SCR*) est composé des risques de marché, de souscription santé, de contrepartie, de souscription vie, de souscription non-vie et d'actifs intangibles.

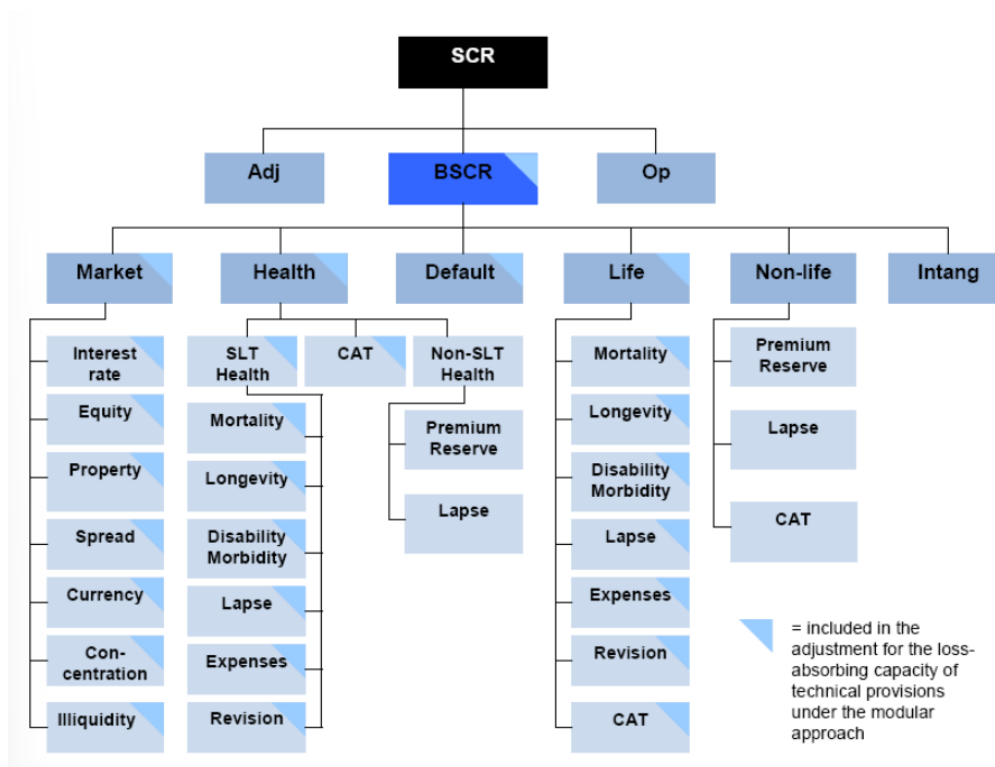


FIGURE 4.1 – Module des risques du SCR en formule standard (source : ACPR)

Le SCR d'un sous-module i est calculé par la variation de la marge entre la valeur à l'actif et au passif (appelée NAV pour *Net Asset Value*) suite à l'application du choc correspondant au risque sur la méthode classique.

$$SCR_i = \max(0, NAV_i^{Central} - NAV_i^{Choc})$$

Le SCR global est obtenu à partir de la somme via une matrice de corrélation du BSCR, d'un ajustement et de l'ajout du SCR risque opérationnel. L'ajustement permet de prendre en compte la capacité d'absorption des pertes par les provisions techniques et les impôts différés.

Enfin, le coût d'immobilisation du SCR doit être pris en compte à travers la provision de la *Risk Margin* (RM) ou *Market Value Margin* (MVM). Ce coût représente le montant supplémentaire de capital qu'un tiers demanderait pour reprendre le stock de provisions. Le coût d'opportunité de l'immobilisation du capital est estimé à 6 %. La MVM est calculée à partir du coût des capitaux à immobiliser actualisés en considérant un engagement long terme n , en notant r_k le taux sans risque de maturité k , la formule s'écrit :

$$MVM_t = 6 \% \times \sum_{k=t}^n \frac{SCR_k}{(1 + r_k)^k}$$

Le Groupe AXA dispose d'un modèle interne pour calculer le SCR. La mise en place d'un modèle interne après approbation de l'ACPR permet d'adapter la formule standard afin de mieux refléter la structure de l'entreprise. La matrice de corrélation et les chocs utilisés pour le calcul des SCR par module sont adaptés aux risques et à la diversification du portefeuille d'AXA France.

4.2.3 Le taux de rentabilité interne

Le taux de rentabilité interne (TRI) est le taux d'actualisation des résultats distribuables qui permet de compenser l'investissement initial. Le TRI représente le taux de rendement global du produit.

En notant CF_t les flux en année t , l'indicateur se définit par la formule :

$$\sum_{t=0}^T \frac{CF_t}{(1 + TRI)^t} = 0$$

Pour rappel, le premier flux est généralement fortement négatif, puis les flux suivants viennent compenser positivement les coûts d'acquisition et de gestion.

Le TRI présente l'inconvénient de ne pas prendre en compte le coût des options et garanties mais il considère le coût du capital immobilisé (qui est ignoré par la NBV).

Enfin, un indicateur lié au TRI est la période de récupération ou *payback period*. Cet indicateur représente le temps nécessaire pour que les flux entrants équilibrent le coût total de l'investissement.

4.2.4 Le ratio combiné

Le principe général du ratio combiné est de la forme : $1 - (NBV/PVEP)$ à laquelle est ajouté un terme représentant le coût du retour attendu par l'actionnaire sur le capital immobilisé. La totalité des profits futurs (NBV) est rapportée à la totalité des primes futures (PVEP).

Le ratio combiné mesure la performance économique du produit en calculant le rapport entre les gains de l'assureur et les coûts engendrés. Si le ratio est inférieur à 100 %, cela signifie que le produit est suffisamment rentable pour absorber les dépenses de l'assureur.

4.2.5 Un indicateur de solvabilité

Une variante du ratio de solvabilité, appelée K-light, est utilisée pour déterminer si le produit est consommateur en capital ou autosuffisant. Pour des raisons de confidentialité, la formule exacte de cet indicateur n'est pas détaillée. Le ratio dépend positivement de la NBV et négativement du besoin en capital.

Le produit est dit de :

- « Capital léger » si le ratio est au-dessus de 130 % ;
- « Capital modéré » lorsque le ratio est entre 100 % et 130 % ;
- « Capital intensif » pour les ratio en dessous de 100 %.

L'objectif est de détenir des produits avec un ratio au-dessus de 130 %.

4.3 Impact du modèle sélectionné sur la rentabilité

Après avoir mis en avant les variables déterminantes dans la prédiction du montant moyen de versements périodiques, l'objet de cette dernière partie est d'appliquer les résultats dans la détermination de la rentabilité des produits. Pour cela, des *model points* sont créés afin de regrouper les individus avec des caractéristiques similaires. En effet, la modélisation retenue dans le chapitre précédent calcul les montants moyens versés tête par tête. Or, la projection globale de tous les assurés serait trop longue à dérouler. Pour diminuer le temps de calcul la projection des flux pour étudier la rentabilité des produits s'effectue à travers plusieurs *model points*. Dans le modèle actuel, il en existe pour chaque produit en fonction du type de gestion choisi et du mode de sortie (en capital ou en rente). La rentabilité d'un produit est calculée à partir de la moyenne de la rentabilité des *model points*, pondérée par le nombre de contrats le constituant.

La création de plusieurs groupes pour modéliser les versements périodiques implique de dupliquer les distinctions déjà présentes. Par exemple, pour un produit modélisé avec deux *model points* : le premier représentant l'ensemble des assurés ayant choisi une gestion sous mandat ; le second pour ceux en gestion personnelle ; et en considérant deux profils (intense / faible) de versements, cela conduit finalement à projeter les flux de quatre *model points*. Or, multiplier des *model points* augmente le temps de projection du modèle. Cela peut donc complexifier le modèle général.

De plus, créer de nouveaux *model points* conduit à mettre à jour les paramètres des groupes en matière de montant de versement initial, de montants moyens de versements libres et d'âge moyen du groupe. Dans la suite, les paramètres de chaque sous-groupe déterminés à partir de la méthode *k-means* sont présentés. Puis, l'estimation des montants moyens de versements périodiques est effectuée. Enfin, les nouveaux souscripteurs de l'année 2021 sont regroupés selon leurs caractéristiques déterminées par le *clustering*, afin d'effectuer une analyse de sensibilité sur la rentabilité des affaires nouvelles.

4.3.1 États des hypothèses techniques

L'étude de rentabilité est établie sur les affaires nouvelles de 2021 pour pouvoir projeter les flux futurs perçus par AXA France et déterminer les niveaux de marges des produits à partir de 2021 sur les nouveaux souscripteurs. L'impact du regroupement des individus sur le produit 3 de retraite n'est pas présenté. Ce produit n'est plus ouvert à la commercialisation, l'approche considérant les nouvelles affaires n'est donc pas cohérente. De plus, le produit présentait deux sous-groupes ayant des effectifs assez déséquilibrés.

Avant de présenter la rentabilité des produits, les différentes hypothèses techniques utilisées sur chacun des produits sont passées en revue. Cela permet de comprendre les écarts de rentabilité entre les deux produits. Au sein d'un même produit, les hypothèses techniques retenues sur les *model points* sont identiques à celles présentées ci-dessous. Les montants moyens sur les versements, les lois d'évolution des versements et la répartition entre le support en euros et ceux en UC sont différents en fonction du type de gestion.

- Table de mortalité : une table certifiée est utilisée pour calculer la probabilité de décès des assurés.
- Lois de rachats : les lois sont dépendantes du produit.
- Lois d'arbitrages : contrairement au produit 1, le produit 2 ne présente pas d'hypothèses sur les arbitrages effectués au cours du temps. En effet, le produit 2 présentent deux types de gestion qui n'impliquent pas d'arbitrage, les provisions mathématiques suivent une loi prédéfinie selon le support.
- Chargements : des taux sont appliqués sur les primes versées par l'assuré et sur l'encours, ces taux composent une partie des marges.

- Coûts : les coûts d'acquisition et de gestion se décomposent en une partie fixe et une autre variable. La part d'acquisition fixe est plus importante sur le produit 1 que sur le produit 2 mais la part de variable est plus élevée sur le produit 2 que sur le produit 1.
- Garantie : les deux produits sont modélisés avec une garantie plancher jusqu'aux 80 ans de l'assuré. En cas de moins-value, la garantie permet aux bénéficiaires de récupérer un montant égal au cumul des capitaux versés net de frais et des éventuels rachats.

Enfin, la loi d'évolution des montants de versements libres varie en fonction du produit. En particulier, le produit 1 présente deux lois d'évolution de versements qui dépendent du type de gestion. Il y a une loi pour les contrats sous mandat et une deuxième pour les autres types de gestion. Pour le produit 1, la loi d'évolution des versements libres est plus conservatrice lorsque le contrat est en gestion sous mandat. Autrement dit, les versements libres sont en moyenne plus élevés au cours du temps comparés à ceux effectués sur les autres supports. Les lois d'évolution des versements libres sur les autres types de gestion ne sont pas assez distinctes. Une unique loi sur les versements libres suffit pour représenter les évolutions sur les contrats en gestion autre que la gestion sous mandat.

4.3.2 *Model points* retenus

Les groupes déterminés sur l'ensemble du portefeuille, toutes anciennetés confondues, permettent de retenir un montant moyen de versements libres et la fluctuation des montants périodiques en fonction de l'ancienneté.

Pour chaque sous-groupe de nouvelles affaires souscrites, les variables âge moyen et versement initial sont calculées afin de représenter au mieux les caractéristiques des individus. Les montants moyens de versements périodiques et des versements libres ont été calculés à partir de montants historiques. Aussi, pour pouvoir comparer l'intérêt de séparer les individus en fonction de certaines caractéristiques, des *model points* sans *clustering* sont aussi paramétrés. Les lois d'évolution des montants périodiques appliquées sont celles construites à partir du taux global de réduction des versements (développé dans la section 3.1.2). De la même façon, l'âge moyen et le versement initial moyen sont caractéristiques des *model points* alors que les autres montants proviennent de l'observation des montants versés sur la base historique.

Produit 1

Pour rappel, le produit 1 a permis de distinguer trois sous-groupes différents. Puis, pour chaque type de gestion, l'âge moyen et le versement initial sont déterminés. Aucun individu du sous-groupe 2 n'a souscrit au produit 1 en gestion par horizon par âge. Les descriptions par groupe de la section 3.4.2 se retrouvent comme données d'entrée des *model points*. Le sous-groupe 0 est caractérisé par des individus en moyenne plus jeunes et des montants à la souscription plus faibles (en moyenne entre 3 500 € et 22 600 €).

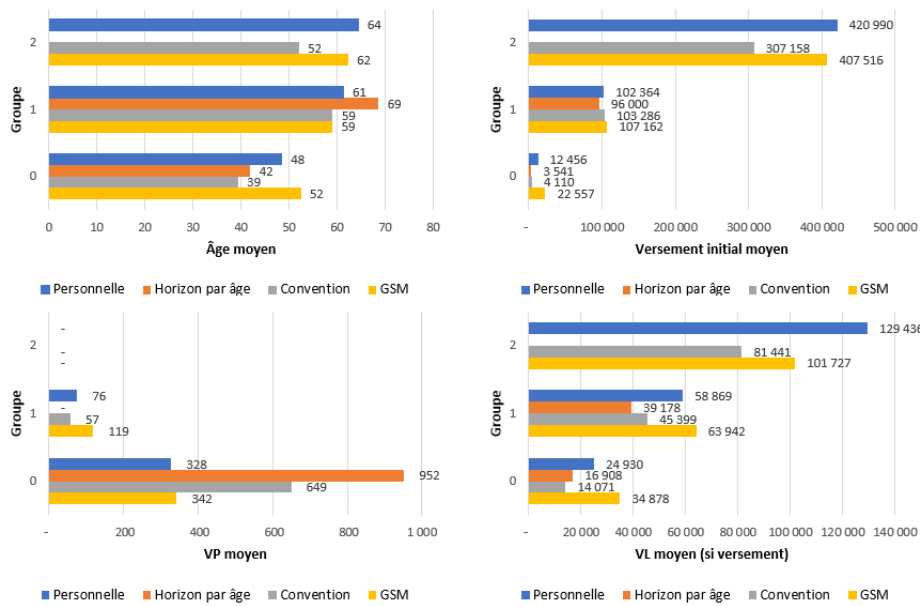


FIGURE 4.2 – Paramètres retenus pour les sous-groupes du produit 1

Au global, les paramètres par type de gestion sont les suivants :

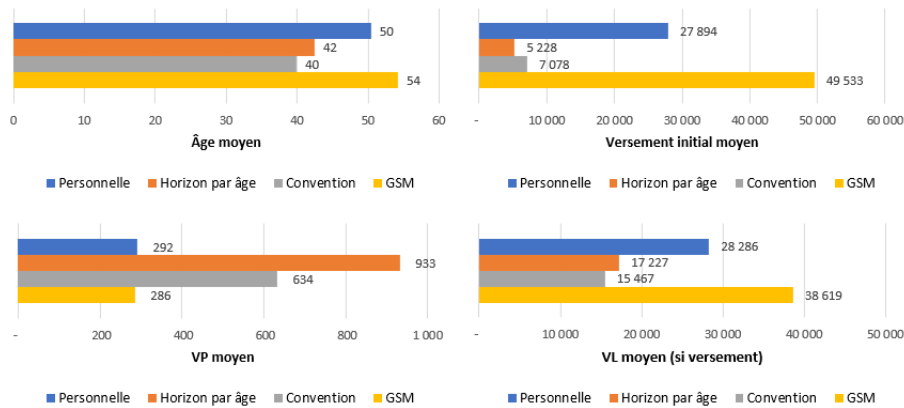


FIGURE 4.3 – Paramètres retenus dans l'approche globale du produit 1 (sans *clustering*)

Les paramètres utilisés dans la modélisation sans *clustering* se rapprochent de ceux retenus pour modéliser le sous-groupe 0 de l'autre méthode. Les versements initiaux sont inférieurs à 50 000 € et les montants périodiques moyens sont situés entre 286 € et 933 € en fonction du type de gestion. En effet, la répartition par sous-groupe est inégale : le sous-groupe 0 étant majoritairement représenté pour chaque type de gestion.

Produit 2

Les sous-groupes sont appliqués au produit 2 avec deux types de gestion :

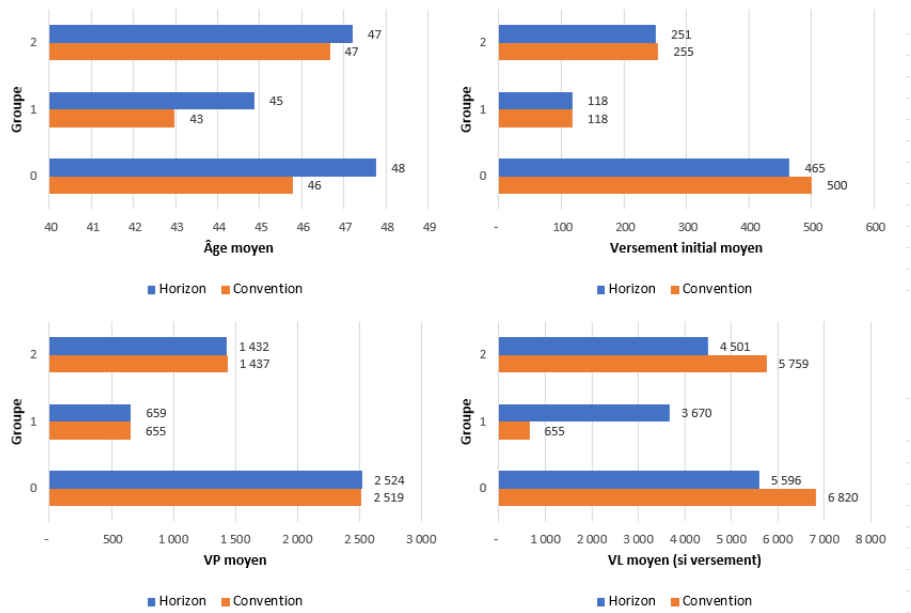


FIGURE 4.4 – Paramètres retenus pour les sous-groupes du produit 2

Quel que soit le type de gestion, les montants de versements initiaux et périodiques sont très proches entre les sous-groupes. Ceux-ci restent similaires au regard des caractéristiques, indépendamment du type de gestion. Le sous-groupe 1 a un comportement différent sur les versements libres : les individus en gestion par horizon versent en moyenne 3 670 € lors d'un versement libre contre 655 € en gestion par convention. En matière de rentabilité, il sera intéressant de regarder si le montant moyen plus élevé de versements libres en gestion par horizon est plus rentable qu'un versement plus faible en convention au regard des autres caractéristiques.

Au global, la similitude des versements se retrouve bien entre les types de gestion. La différence de montant de versements libres est moins marquée. En effet, l'écart observé sur le sous-groupe 1 est compensé par les deux autres groupes où le phénomène inverse est visible.

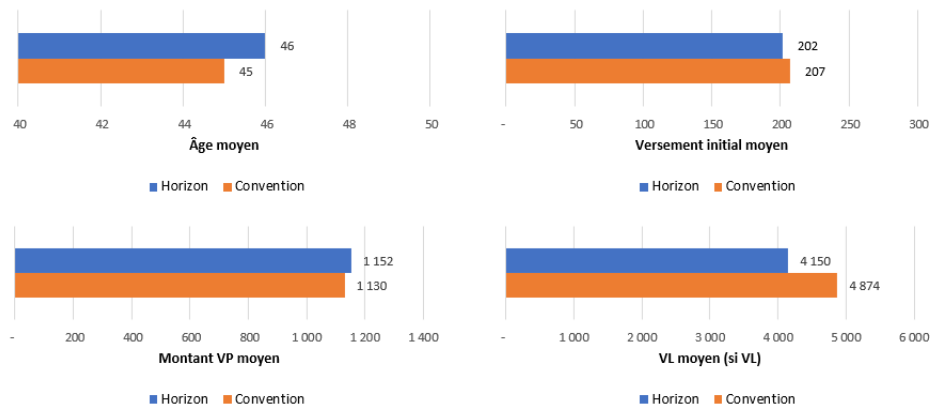


FIGURE 4.5 – Paramètres retenus dans l'approche globale du produit 2

4.3.3 Résultats

Pour comparer l'impact des dernières analyses, le portefeuille est projeté à partir des nouvelles affaires de l'année 2021 en fonction des groupes et des différentes lois modélisées. Les marges dégagées à partir des assurés avec plusieurs années d'ancienneté ne sont pas prises en compte. Seuls les nouveaux assurés qui ont souscrit en 2021, donc sans ancienneté, sont projetés.

Dans un premier temps, la rentabilité des produits 1 et 2 est présentée en prenant en compte la mise à jour des lois d'évolution de montant de versements périodiques à partir de la réduction des montants de versements en fonction de l'ancienneté. Puis, ces rentabilités sont comparées à la rentabilité estimée à partir des sous-groupes constitués pour la projection des versements périodiques, pour lesquels les lois d'évolution des primes ont aussi été adaptées dans la section 3.4.2.

Produit	NBVm	Ratio Combiné	K-light
Produit 1	54,8 %	97,3 %	145,2 %
Produit 2	22,3 %	98,8 %	82,2 %

TABLE 4.1 – Rentabilité des produits en prenant en compte la mise à jour des lois de versements périodiques

Le produit 1 présente la meilleure *NBVmargin*, avec le ratio combiné le plus bas à 97,3 %. La *NBVmargin* dépend en partie négativement du ratio combiné. Enfin, le produit 1 est moins consommateur en capital avec un K-light à 145,2 %.

Le produit 2 présente des indicateurs de rentabilité assez faibles par rapport aux seuils minimums ciblés par AXA, avec une *NBVmargin* inférieure à 25 % et un K-light inférieur à 100 %.

Le produit 1 est majoritairement constitué de capital provenant du versement initial puis des versements libres, ce qui explique que la *NBV margin* est meilleure même si les versements périodiques sont plus faibles. L'APE, qui représente la prime moyenne versée par l'assuré sur les dix premières années, est de 9 082 € sur le produit 1 contre 906 € sur le produit 2. Cet écart sur les montants de primes explique la différence de rentabilité. Les montants de versements viennent compenser plus rapidement les coûts élevés de l'acquisition du contrat pour le produit 1. Il faut en moyenne 7 ans pour le produit 1 et 23 ans pour le produit 2 pour compenser les coûts d'investissement.

Rentabilité du produit 1

Produit 1	Répartition	NBVm	Ratio Combiné	K-light
Personnelle	40,46 %	17,1 %	99,0 %	62,5 %
Horizon par âge	0,75 %	36,8 %	99,1 %	82,7 %
Convention	19,84 %	15,1 %	101,4 %	31,4 %
GSM	38,95 %	69,4 %	96,8 %	158,0 %
Global	100 %	54,8 %	97,3 %	145,2 %

TABLE 4.2 – Rentabilité de produit 1 par type de gestion

Le détail par type de gestion révèle que la rentabilité du produit est portée par les contrats en GSM. Ce type de gestion reçoit les versements les plus importants à la souscription et lors des versements libres. Les individus ayant souscrit au produit 1 se tournent principalement vers de la gestion personnelle et la gestion sous mandat. La gestion par horizon présente une rentabilité satisfaisante : tous les indicateurs sont au dessus des seuils minimums cibles. En matière de paramètres, la gestion par horizon est adossée à des montants périodiques élevés. Ces montants sont de 933 € en moyenne. Néanmoins, avec une faible part sur la répartition des nouvelles affaires de 2021, ce type de gestion n'impacte par la rentabilité globale du produit.

	Répartition	NBVm	Ratio Combiné	K-light
Personnelle	40,46 %	13,7 %	99,2 %	53,3 %
<i>Personnelle G0</i>	<i>35,39 %</i>	<i>5,1 %</i>	<i>102,0 %</i>	<i>20,3 %</i>
<i>Personnelle G1</i>	<i>4,90 %</i>	<i>25,8 %</i>	<i>97,9 %</i>	<i>137,8 %</i>
<i>Personnelle G2</i>	<i>0,17 %</i>	<i>28,1 %</i>	<i>97,5 %</i>	<i>172,3 %</i>
Horizon par âge	0,75 %	25,9 %	99,3 %	77,4 %
<i>Horizon par âge G0</i>	<i>0,72 %</i>	<i>27,1 %</i>	<i>100,5 %</i>	<i>63,8 %</i>
<i>Horizon par âge G1</i>	<i>0,03 %</i>	<i>21,1 %</i>	<i>98,1 %</i>	<i>161,3 %</i>
Convention	19,84 %	5,1 %	102,1 %	26,8 %
<i>Convention G0</i>	<i>19,23 %</i>	<i>-0,9 %</i>	<i>106,5 %</i>	<i>-115,2 %</i>
<i>Convention G1</i>	<i>0,58 %</i>	<i>29,8 %</i>	<i>97,5 %</i>	<i>170,0 %</i>
<i>Convention G2</i>	<i>0,02 %</i>	<i>42,7 %</i>	<i>96,3 %</i>	<i>278,6 %</i>
GSM	38,95 %	69,5 %	96,8 %	161,7 %
<i>GSM G0</i>	<i>28,79 %</i>	<i>75,4 %</i>	<i>96,9 %</i>	<i>153,3 %</i>
<i>GSM G1</i>	<i>9,55 %</i>	<i>64,0 %</i>	<i>96,7 %</i>	<i>170,4 %</i>
<i>GSM G2</i>	<i>0,61 %</i>	<i>54,6 %</i>	<i>96,5 %</i>	<i>186,0 %</i>
Produit 1 - Groupes	100 %	54,8 %	97,3 %	149,9 %

TABLE 4.3 – Rentabilité du produit 1 par type de gestion et par sous-groupes

Les résultats de rentabilité totaux sont similaires à ceux agrégés des sous-groupes. La rentabilité estimée pour la gestion personnelle, par horizon et convention est plus faible avec l'approche par regroupement, alors que la rentabilité projetée de la gestion sous mandat est meilleure. Au global, la rentabilité est similaire car 74,4 % de l'APE des sous-groupes proviennent de l'APE de gestion sous mandat contre 72,3 % sans effectuer de regroupement.

L'indicateur de K-light est plus faible avec l'approche au global. En effet, le capital à immobiliser est légèrement plus important ce qui vient diminuer le K-light. Cette différence n'est pas visible dans le ratio combiné. L'indicateur de consommation de capital est plus élevé pour le total selon l'approche de *clustering* comparé au total global car le taux est différent pour les contrats en GSM : 161,7 % avec l'utilisation de sous-groupes et 158,0 % pour l'autre approche. Les trois sous-groupes de contrats en GSM présentent en moyenne sur la projection une part d'UC de 26 %, 28 % et 30 %. Une part plus importante d'UC implique une part de montant garanti sur le support en euros moindre. Ainsi, les contrats avec une part d'UC plus élevée nécessitent un capital à immobiliser moins important, ce qui fait augmenter le K-light.

En conclusion, l'ajout de complexité dans la modélisation des versements périodiques a très peu d'impact sur la rentabilité calculée du produit. Tout d'abord, cela est expliqué par le fait que le produit est majoritairement constitué de versements à la souscription et de versements libres. Les versements périodiques ont peu d'impact sur la rentabilité du fait de leur montant et du nombre réduit d'assurés effectuant ce type de versement. De plus, les sous-groupes constitués sont peu significatifs dans la répartition par type de gestion, le sous-groupe 0 est trop représentatif. Enfin, les sous-groupes viennent augmenter les versements sur les contrats en gestion sous mandat. Cet effet est globalement absorbé par des résultats plus faibles sur les autres supports.

Rentabilité du produit 2

Produit 2	Répartition	NBVm	Ratio Combiné	K-light
Horizon par âge	48,2 %	4,3 %	99,9 %	12,6 %
Convention	51,8 %	39,3 %	97,9 %	165,8 %
Global	100 %	22,3 %	98,8 %	82,2 %

TABLE 4.4 – Rentabilité du produit 2 par type de gestion

La répartition des types de gestion dans le portefeuille du produit 2 est équilibrée entre la gestion par horizon et par convention. Une nette différence de rentabilité est visible entre les deux types de gestion : la gestion par horizon, avec une NBVm de 4,3 %, vient dégrader la rentabilité générale du produit. La différence des primes moyenne sur 10 ans est faible, l'APE sur un contrat en gestion par convention est de 943 €, et par horizon de 951 €. Ainsi, l'écart sur la NBVm provient de la VIF déterministe. Or, la TVOG et le *Strain* sont similaires pour un contrat sur les deux types de gestion. La différence de VIF déterministe est due à la différence de répartition de l'encours entre support en euros et en UC et l'écart d'une année sur l'âge moyen. Un contrat en convention de gestion est paramétré avec un âge moyen de 45 ans et une part d'encours UC autour de 40 %. Tandis qu'un contrat avec une gestion par horizon a 46 ans d'âge moyen et une proportion de seulement 21 % sur les supports UC. Les profits futurs actualisés sur ces contrats compensent moins bien la première année déficitaire et la valeur temps des options et garanties.

	Répartition	NBVm	Ratio Combiné	K-light
Horizon par âge	48,2 %	-4,0 %	100,4 %	-17,4 %
<i>Horizon par âge G0</i>	5,0 %	32,5 %	97,8 %	163,9 %
<i>Horizon par âge G1</i>	26,0 %	-57,9 %	108,0 %	-182,5 %
<i>Horizon par âge G2</i>	17,2 %	19,5 %	98,9 %	68,6 %
Convention	51,8 %	29,6 %	98,4 %	131,0 %
<i>Convention G0</i>	5,3 %	61,0 %	95,9 %	374,6 %
<i>Convention G1</i>	27,5 %	-25,2 %	101,6 %	-114,4 %
<i>Convention G2</i>	19,0 %	53,1 %	97,1 %	240,9 %
Produit 2 - Groupes	100 %	13,3 %	99,3 %	50,3 %

TABLE 4.5 – Rentabilité du produit 2 par type de gestion et par sous-groupes

La projection des flux en fonction des sous-groupes déterminés sur le produit 2 vient diminuer la rentabilité estimée du produit : la NBVm globale entre les deux approches diminue de 9 points. Les deux types de gestion sont impactés par cette diminution.

Les sous-groupes 0 et 2 présentent des rentabilités satisfaisantes. Les assurés du groupe 1 sont caractérisés par des versements initiaux et libres beaucoup plus faibles que les deux autres groupes. Le groupe 1 est aussi plus jeune comparé aux autres groupes. Ainsi, la réduction de l'âge, qui conduit la projection des flux à partir d'une probabilité plus faible de sortie des assurés liée au décès, n'est pas suffisante pour combler les différences de marges prélevées sur des montants plus importants.

Au global, l'APE et la NBV sont plus faibles avec la modélisation par *clustering* comparée à celle sans sous-groupes. La variation à la baisse de la NBV est plus importante que celle de l'APE avec des coûts d'acquisition similaires pour les deux méthodes. Ces variations créent une NBVm plus faible lorsque les sous-groupes sont modélisés. Ainsi, les montants moyens retenus avec l'ajout des sous-groupes d'assurés ne permettent pas d'acquérir autant de primes comparés aux montants paramétrés dans l'approche sans groupe.

En conclusion, l'ajout de complexité dans les *model points* influe différemment sur la rentabilité estimée des deux produits étudiés. Le produit 1, avec une part des primes périodiques moins volumineuse sur les primes totales versées, présente peu de variation dans la rentabilité estimée. Au contraire, la multiplication des *model points* sur le produit 2 révèle des conséquences plus importantes sur la rentabilité calculée. La rentabilité d'une partie des assurés du produit 2 est fortement compensée par la part complémentaire du portefeuille. Ce phénomène est moins visible sans les groupes. Il peut être intéressant de cibler davantage l'acquisition de nouveaux profils similaires aux assurés du sous-groupe 2 plutôt que du sous-groupe 1 pour permettre d'avoir au global une rentabilité plus satisfaisante.

Conclusion

En assurance-vie, il est essentiel d'appréhender de manière la plus fiable possible le comportement client qui déterminera les niveaux des engagements de l'assureur. La conjoncture économique passée, qui a été marquée par des taux d'intérêt historiquement bas, a rendu les rendements des contrats d'épargne sur les fonds en euros de moins en moins attractifs et a poussé les assureurs à utiliser leurs réserves pour permettre de servir les taux garantis passés. Les versements périodiques représentent une part importante des montants collectés sur les contrats d'épargne. Il est donc nécessaire d'estimer au plus juste les volumes des primes utiles à l'exercice de projection des flux, pour assurer le suivi de rentabilité des produits commercialisés. Il existe déjà un certain nombre de mémoires qui présentent les facteurs influençant les comportements des assurés sur les versements libres. Ce mémoire a permis de compléter les précédentes études par l'analyse des versements périodiques. Il a été remarqué qu'une proportion plus importante de contrats effectuait des versements de manière récurrente allant de 10 % jusqu'à 90 % sur certains produits. La fréquence des flux n'a donc pas été un enjeu de l'étude, comme elle peut l'être pour estimer les versements libres.

Plusieurs modélisations ont été proposées sur le portefeuille mis à disposition. Trois produits ont pu être étudiés sur deux périmètres différents : l'assurance-vie et l'épargne retraite. Bien que les versements périodiques représentent des montants annuels moyens assez constants dans le temps pour chaque produit, il s'est avéré plus compliqué de suivre correctement les évolutions des décisions des assurés sur le montant versé au cours de la vie du contrat. Les premiers modèles statistiques ont permis d'obtenir des résultats selon une approche simple et rapide à mettre en place. Ces résultats ont montré une diminution des versements en fonction de l'ancienneté du contrat. La diminution des montants de versements périodiques est plus marquée sur les contrats d'épargne retraite, ce qui est intrinsèquement lié à la nature du produit.

Sur le long terme, les résultats des premiers modèles ne semblaient pas assez satisfaisants. Par la suite, des modélisations plus complexes avec l'introduction d'algorithmes de *Machine Learning* ont permis d'obtenir de meilleures estimations des montants totaux estimés. De plus, les modèles ont mis en avant d'autres variables déterminantes dans la prédiction des comportements des assurés. Ils ont notamment souligné l'importance du niveau de l'encours, du versement initial, et de l'âge comme indicateurs sur les versements périodiques futurs probables, et ce quel que soit le type de contrat d'épargne. Enfin, la complexité des modèles ligne-à-ligne de *Machine Learning* n'a pas permis l'utilisation de leurs prédictions dans l'étude de rentabilité des produits. Le coût de calcul pour prédire l'ensemble du portefeuille multiplié par le nombre de scénarios financiers à projeter et la durée de projection rendent l'emploi des algorithmes trop coûteux en temps de calcul. Aussi, l'interprétation des résultats sur des modèles de forêts aléatoires ou XGBoost s'avère être un frein à leur utilisation dans la pratique. Néanmoins, l'introduction de nouveaux facteurs à travers l'âge de l'assuré et le versement initial permet d'ajouter de la précision dans le comportement des assurés sur les versements périodiques. À partir d'une méthode de *clustering*, deux à trois sous-groupes homogènes avec des versements distincts de primes périodiques ont été déterminés. Puis, ces groupes d'individus ont été utilisés pour prédire sur les affaires nouvelles la rentabilité attendue du produit. L'impact de l'introduction des sous-groupes sur le produit 1 est très faible sur la rentabilité globale calculée. Finalement, le produit 2 qui dépend majoritairement des versements périodiques s'est révélé être plus sensible à l'introduction de sous-groupes dans la modélisation. La rentabilité estimée sur chaque sous-groupe a aussi montré que certains profils d'assuré impactaient négativement la rentabilité

globale du produit. Cette étude souligne l'importance de la prédiction des comportements des assurés sur les versements périodiques, et en particulier l'utilité de suivre certains profils.

Le modèle final proposé, bien qu'il introduise de la précision sur la projection des flux des affaires nouvelles, admet tout de même quelques faiblesses. Tout d'abord, l'utilisation de l'âge dans la détermination des profils d'assurés peut être contraignante sur la projection des flux des contrats. Il n'est pas impossible qu'un assuré change de groupe au cours de la projection. La force de rattachement au sous-groupe en fonction du niveau de versement initial (plutôt que de l'âge) n'a pas été étudiée.

Pour aller plus loin, il s'avère intéressant d'analyser les comportements des assurés sur les versements libres des différents groupes. Aussi, les variables structurelles ne sont pas les seuls facteurs influençant le comportement des assurés. Il serait nécessaire d'étudier dans quelle mesure les facteurs conjoncturels influent sur le comportement des assurés. L'étude du contexte économique actuel marqué par une baisse des taux garantis sur les fonds en euros et par une importante inflation observée depuis 2022 permettrait de mesurer l'impact de variables économiques (inflation, taux mensuel des emprunts d'État, etc.) sur l'évolution future des montants de versements.

Bibliographie

- [1] ACPR. Le marché de l'assurance vie pendant la crise sanitaire. *Analyses et synthèses*, 121, 2021.
- [2] ACPR. Le marché de l'assurance-vie en 2021. *Analyses et synthèses*, 133, 2022.
- [3] S. Andre. *Comportement de versement libre en épargne individuelle : approche conceptuelle et modélisation*. Mémoire d'actuariat, ISFA, Lyon, 2019.
- [4] D. Arthur and S. Vassilvitskii. k-means++ : The Advantages of Careful Seeding . <https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf>, 2007.
- [5] K. Assaraf. *Modélisation de versements libres sous IFRS 17 par des méthodes de machine learning*. Mémoire d'actuariat, Paris Dauphine, 2020.
- [6] Banque de France. Placements et patrimoine des ménages au 4e trimestre 2021. <https://www.banque-france.fr/statistiques/epargne-des-menages-2021t4>, 2022.
- [7] F. Z. Benabdelkrim. *Modélisation des versements libres en assurance non vie : utilisation de méthodes de scoring*. Mémoire d'actuariat, ISUP, Paris, 2017.
- [8] L. Breiman. Bagging predictors. *Machine Learning*, 24 :123–140, 1996.
- [9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.
- [10] A. Chaudhry. *Étude de rentabilité du PER individuel d'AXA France*. Mémoire d'actuariat, ENSAE, Paris, 2019.
- [11] B. Dieltiens. *Contributions à la gestion des risques en assurance vie*. Gestion et management, Université de Lyon, 2021.
- [12] S. Feniza. *Modélisation du comportement d'arbitrage en assurance vie*. Mémoire d'actuariat, ENSAE, Paris, 2019.
- [13] J. H. Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, 29 :1189–1232, 2001.
- [14] R. Genuer and J.-M. Poggi. Arbres CART et Forêts aléatoires, Importance et sélection de variables. *HAL*, 2017.
- [15] T. C. . C. Guestrin. *Xgboost : A scalable tree boosting system*. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. Association for Computing Machinery, 2016, August.
- [16] INSEE. 10% des ménages détiennent près de la moitié du patrimoine total. <https://www.insee.fr/fr/statistiques/4265758>, 2019.
- [17] INSEE. Le patrimoine économique national en 2020. <https://www.insee.fr/fr/statistiques/5430978>, 2021.

- [18] LegiFrance. Article A132 - Code des assurances. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000035514601/, 2017.
- [19] Ministère de l'Economie et des Finances. Le nouveau plan epargne retraite (PER). <https://www.economie.gouv.fr/PER-epargne-retraite>, 2019.
- [20] E. S. Nana Njoya. *Prédiction des comportements de rachat en épargne individuelle : une approche de machine learning*. Mémoire d'actuariat, ENSAE, 2016.
- [21] M. Porter. *An algorithm for suffix stripping*, volume 14. Program : electronic library and information systems, 1980.
- [22] Vijaya Rani. Nlp tutorial for text classification in python. <https://github.com/vijayaiitk/NLP-text-classification-model>, 2021.

Annexe A

Méthodes de corrélation

Cette annexe présente les méthodes qui ont été utilisées pour mesurer le lien entre deux variables.

Corrélation de Pearson

Le coefficient de corrélation de Pearson mesure la relation entre deux variables quantitatives. Le coefficient est défini par la relation :

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

où $Cov(X, Y)$ représente la covariance entre X et Y , et $Var(\cdot)$ désigne la variance de la variable.

Il mesure la corrélation linéaire entre deux variables X, Y . Néanmoins, cette formule ne rend pas bien compte de la relation entre deux variables dans le cas où la corrélation n'est pas linéaire ou encore sur des variables très hétérogènes. Par exemple dans le premier cas, si X suit une loi uniforme sur $[-1, 1]$, $\rho_{X, X^2} = 0$. Autrement dit la corrélation entre X et X^2 est nulle, pourtant X et X^2 sont bien dépendantes.

Corrélation de Kendall

Le τ de Kendall permet de pallier les défauts de la première mesure. Il mesure la corrélation de rang entre deux variables. Le coefficient de corrélation de Kendall varie entre -1 et 1. Plus le τ est proche de 1 en valeur absolue, plus les variables sont liées. A l'inverse, plus la corrélation tend vers 0, plus les variables sont indépendantes.

Le τ de Kendall pour deux variables aléatoires X, Y est défini par la mesure :

$$\tau(X, Y) = \mathbb{P}((X - \tilde{X})(Y - \tilde{Y}) > 0) - \mathbb{P}((X - \tilde{X})(Y - \tilde{Y}) < 0)$$

où (\tilde{X}, \tilde{Y}) un couple de variables aléatoires indépendant de (X, Y) et identiquement distribué.

$\mathbb{P}((X - \tilde{X})(Y - \tilde{Y}) > 0)$ représente la probabilité de concordance, et $\mathbb{P}((X - \tilde{X})(Y - \tilde{Y}) < 0)$ la probabilité de discordance.

Soient (x_1, y_1) et (x_2, y_2) deux couples d'observations de (X, Y) , il y a :

- concordance lorsque $(x_1 < x_2 \text{ et } y_1 < y_2)$ ou $(x_1 > x_2 \text{ et } y_1 > y_2)$;
- discordance lorsque $(x_1 < x_2 \text{ et } y_1 > y_2)$ ou $(x_1 > x_2 \text{ et } y_1 < y_2)$.

Soient n observations du vecteur (X, Y) , le principe est de comparer pour toutes combinaisons de i et j possibles, les paires (x_i, y_i) et (x_j, y_j) afin de déterminer le nombre de paires concordantes et discordantes.

Le τ de Kendall empirique est donné par :

$$\tau_n(X, Y) = \frac{(\text{Nombre de paires concordantes}) - (\text{Nombre de paires discordantes})}{\frac{1}{2}n(n-1)}$$

Les formules présentées précédemment ne sont utilisables que sur des variables quantitatives. Les coefficients suivants permettent de compléter les outils de corrélation en mesurant les relations entre deux variables qualitatives.

Statistique du Khi-deux

Le test d'indépendance du Khi-deux détermine si deux variables qualitatives sont indépendantes. Sur un échantillon de n individus, soient X et Y deux variables qualitatives, avec respectivement k_1 et k_2 le nombre de modalités de chacune des variables, le test mesure les écarts entre les effectifs théoriques dans le cas où les variables seraient indépendantes et les effectifs empiriques.

Pour une modalité X_i de X et une modalité Y_j de Y , l'effectif conjoint se note n_{ij} avec $i \in \{1, \dots, k_1\}$ et $j \in \{1, \dots, k_2\}$.

Pour tout i dans $\{1, \dots, k_1\}$, le nombre d'individus pour lesquels $(X = X_i)$ se calcule par :

$$n_{i+} = \sum_{j=1}^{k_2} n_{ij}$$

De même, pour chaque modalité Y_j de Y avec j dans $\{1, \dots, k_2\}$, l'effectif s'écrit :

$$n_{+j} = \sum_{i=1}^{k_1} n_{ij}$$

L'échantillon n peut aussi s'écrire comme suit :

$$n = \sum_{i=1}^{k_1} n_{i+} = \sum_{j=1}^{k_2} n_{+j}$$

L'hypothèse du test est la suivante : les deux variables X et Y sont indépendantes (H0). La statistique du khi-deux est définie par :

$$D = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}}$$

La statistique du test de khi-deux suit, lorsque (H0) est vérifiée, une loi de probabilité χ^2 à $(k_1 - 1)(k_2 - 1)$ degrés de liberté.

La valeur de la statistique de test est comparée à la valeur critique dans la loi du χ^2 de degrés de liberté $(k_1 - 1, k_2 - 1)$, pour un seuil fixé α (usuellement à 5 %). α représente la probabilité de rejeter l'hypothèse (H0) quand celle-ci est vraie.

La p-value associée au test est $P(\chi_{1-\alpha}^2((k_1 - 1)(k_2 - 1)) > D)$. Par conséquent, si $D > \chi_{1-\alpha}^2((k_1 - 1)(k_2 - 1))$ avec $\chi_{1-\alpha}^2$ le quantile d'ordre $(1 - \alpha)$, l'hypothèse (H0) est rejetée et les variables X et Y sont liées.

La statistique de khi-deux est dépendante de la taille de l'échantillon n et du nombre de modalités k_1 et k_2 .

V de Cramer

Le V de Cramer est un indicateur permettant de mesurer la corrélation entre deux variables qualitatives X et Y . Cet indicateur est à privilégier sur des grands échantillons, là où le test de khi-deux est influencé par la taille de l'échantillon de données. Le coefficient de Cramer est mesuré pour un échantillon de taille n par la formule :

$$V = \sqrt{\frac{\chi^2}{\chi_{max}^2}} = \sqrt{\frac{\chi^2}{n \times \min(k_1 - 1, k_2 - 1)}} \quad (\text{A.1})$$

Avec k_1 et k_2 le nombre de modalités respectivement de X et Y .

Le terme $\chi_{max}^2 = n \times \min(k_1 - 1, k_2 - 1)$ représente la valeur maximale que peut prendre la statistique de test du χ^2 . Le V de Cramer est un indicateur compris entre 0 et 1. Plus il est proche de 0, plus les variables sont indépendantes. À l'inverse, plus la valeur est proche de 1, plus les variables sont corrélées.

Annexe B

Corrélation des variables

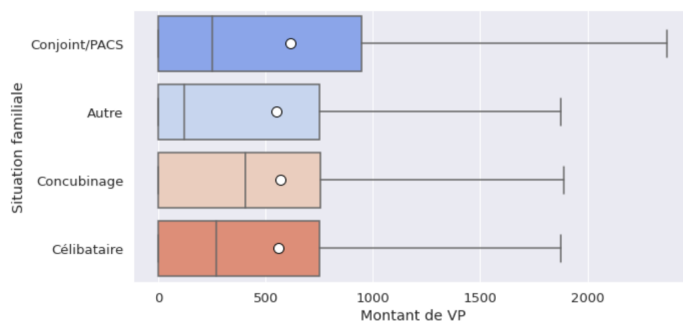


FIGURE B.1 – Montant des versements périodiques moyen par modalités de situation familiale après retraitement sur l'ensemble du portefeuille

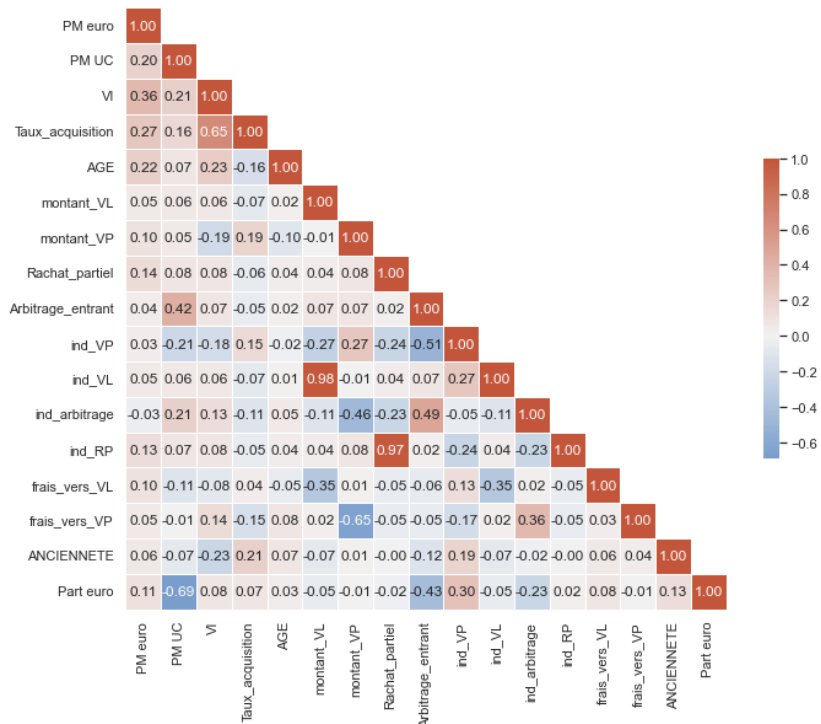


FIGURE B.2 – Corrélation des variables quantitatives sur le produit 1

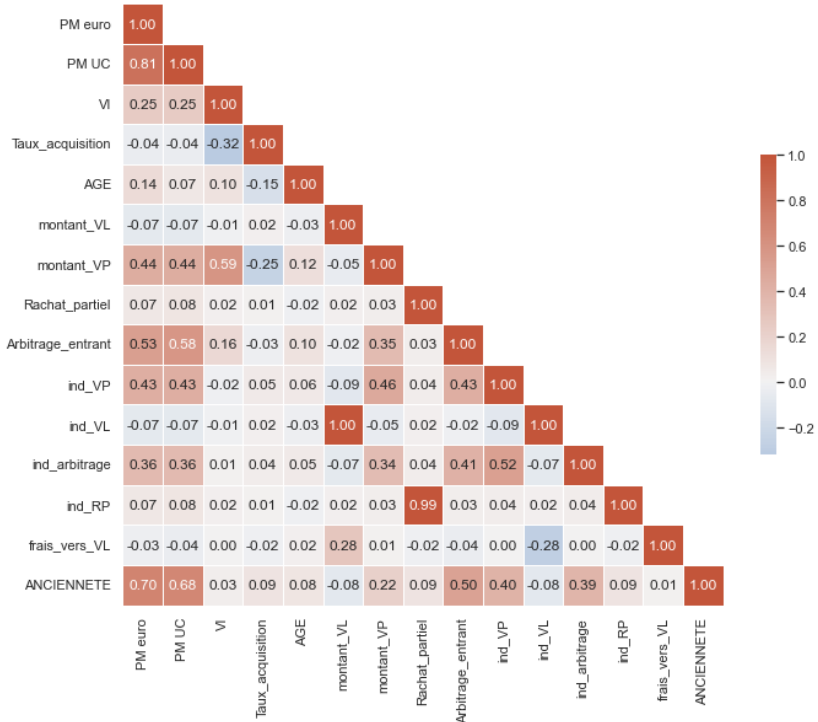


FIGURE B.3 – Corrélation des variables quantitatives sur le produit 2

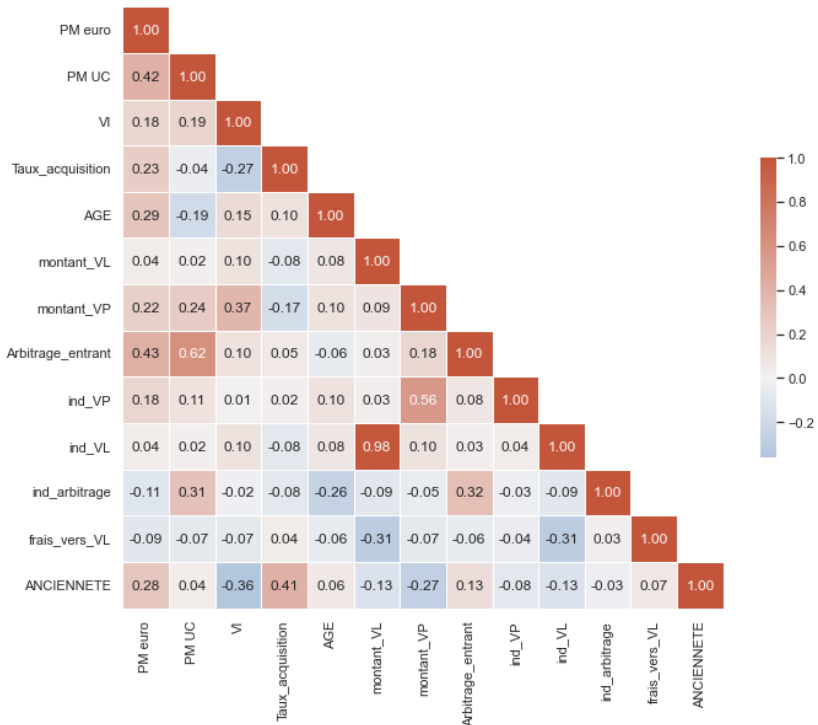


FIGURE B.4 – Corrélation des variables quantitatives sur le produit 3

Annexe C

Les indicateurs de performance

Plusieurs critères permettent d'évaluer la pertinence d'un modèle de prédiction. Les formules des indicateurs utilisés dans ce mémoire sont explicitées avec les notations suivantes, notons :

- y_i la valeur observée à prédire ;
- \hat{y}_i la valeur estimée par le modèle ;
- n le nombre d'observations de l'échantillon.

MSE

L'erreur quadratique moyenne ou MSE (*Mean Square Error*) mesure la moyenne des écarts quadratiques entre les prévisions du modèle et les observations réelles.

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

RMSE

Le RMSE (*Root Mean Square Error*) est la racine de l'erreur quadratique moyenne. Ce critère a la même unité que la variable à prédire et est donc plus facilement interprétable.

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

Des valeurs élevées de RMSE et MSE indiquent que le modèle est peu performant en terme de prédiction. Par la présence du carré, ces deux métriques pénalisent plus fortement les grandes erreurs. Elles sont donc très sensibles aux valeurs aberrantes.

MAE

La MAE (*Mean Absolute Error*) est la moyenne des valeurs absolues des erreurs. Elle est définie par la formule :

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

R²

Le score R^2 appelé aussi coefficient de détermination est le rapport entre la somme des carrés des résidus et la somme des distances entre les valeurs à prédire et leur moyenne \bar{y} .

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Le R^2 peut être vu comme l'erreur du modèle normalisé par une erreur raisonnable qui serait acceptée, représentée par la somme des distances entre chacune des valeurs à prédire et leur moyenne. Un score R^2 proche de 100 % signifie qu'il y a très peu d'erreurs de prédiction et que le modèle est performant. Un modèle présentant un score R^2 à 0 % signifie qu'il n'apporte pas plus de complexité que de prédire la valeur moyenne à chaque estimation. Il est possible d'obtenir un R^2 négatif ce qui signifie que les prédictions sont moins bonnes qu'un modèle retournant la valeur moyenne. Cette métrique permet de comparer facilement la performance entre deux modèles.

Annexe D

Résultats de rentabilité sur un PER

Étant donné que le produit 3 est un produit en *run-off*, les projections de flux futurs n'ont plus de pertinence car ce produit n'est plus en production. Cependant, les analyses effectuées sur ce produit permettent d'obtenir des informations sur les comportements de versements qui sont applicables au nouveau produit PER.

Ce nouveau produit PER individuel qui vient remplacer le produit 3 propose deux types de gestion : par horizon et par gestion personnelle. Pour rappel, le produit 3 proposait seulement une gestion évolutive par âge ou par horizon. Il existe d'autres différences entre les deux produits de retraite, telles que les frais ou les supports d'investissement, ce qui vient modifier les profils d'investissement et donc créer des différences sur les montants moyens de versements initiaux et complémentaires. Par conséquent, l'utilisation de la méthode de *clustering* n'est pas appropriée car les montants des versements sont différents. Néanmoins, le taux global de réduction peut être appliqué au montant moyen de versements périodiques observés lors de la première année de vie des contrats. La rentabilité estimée sous ces différentes hypothèses devra être ajustée en fonction des comportements des assurés sur le nouveau produit de retraite. Une fois que l'historique du portefeuille sera plus développé, il sera possible d'ajuster les lois de comportements au nouveau produit.

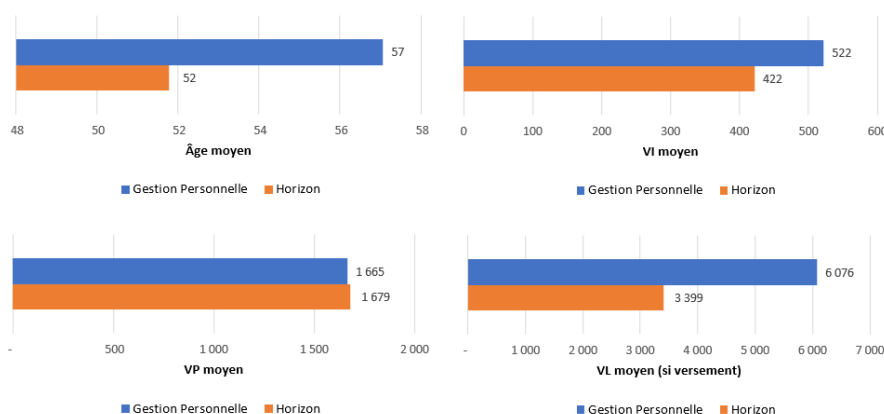


FIGURE D.1 – Paramètres retenus pour la rentabilité du nouveau produit PER

Produit 3	Répartition	NBVm	Ratio Combiné	K-light
Personnelle	3,3 %	12,2 %	98,9 %	44,5 %
Horizon par âge	96,7 %	45,5 %	96,1 %	138,2 %
Global	100 %	44,4 %	96,2 %	135,6 %

TABLE D.1 – Rentabilité de produit 3 par type de gestion (sans *clustering*)

Annexe E

Évolution des primes périodiques en fonction du temps

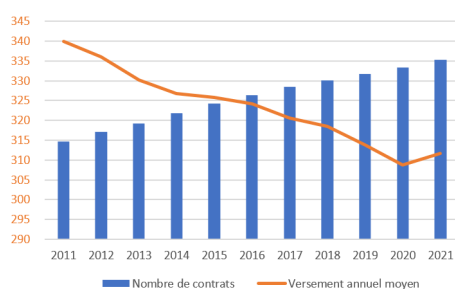


FIGURE E.1 – Évolution du montant moyen des versements périodiques sur le produit 1 depuis 2011

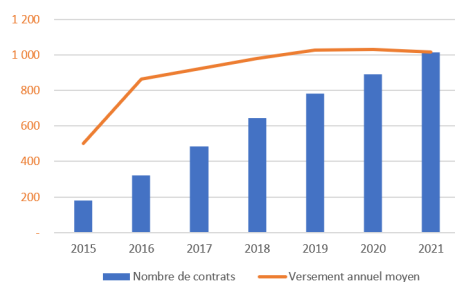


FIGURE E.2 – Évolution du montant moyen des versements périodiques sur le produit 2 depuis 2015

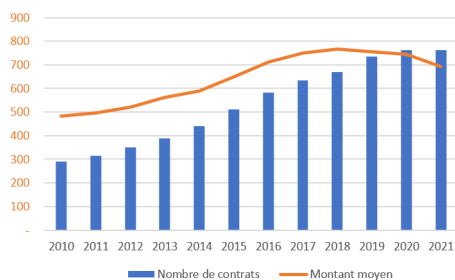


FIGURE E.3 – Évolution du montant moyen des versements périodiques sur le produit 3 depuis 2010

Table des figures

1.1	Évolution des taux d'épargne depuis 2008 dans les principaux pays d'Europe et les Etats-Unis	7
1.2	Exonérations et abattements des droits de succession en assurance-vie	8
1.3	Fiscalité des plus-values lors des rachats	8
1.4	Corrélation de l'investissement vers les unités de compte et les performances du CAC40	9
2.1	Évolution année à année du nombre de contrats entre 2017 et 2021	18
2.2	Répartition des assurés par âge	19
2.3	Distribution de l'ancienneté par produit	20
2.4	Évolution de la part des PM par support	21
2.5	Évolution de la proportion des montants de versements par type	21
2.6	Évolution de la part des contrats effectuant des versements par type	21
2.7	Évolution du montant des versements périodiques par produit	22
2.8	Évolution de la part des versements périodiques par support	23
2.9	Classement des codes de région par montant moyen de versements périodiques	25
2.10	Fréquence des mots dans chaque classe	27
2.11	Fréquence des modalités de situation familiale	28
2.12	Montant moyen des versements périodiques par modalités de situation familiale après retraitement	29
2.13	Corrélation des variables quantitatives sur la base de données	30
2.14	Corrélation des variables qualitatives	31
2.15	Montant moyen des versements périodiques en fonction du genre de l'assuré	32
2.16	Montant moyen des versements périodiques sur le produit 1 en fonction de l'ancienneté (graphique de gauche) et de l'âge (graphique de droite)	32

2.17	Montant moyen des versements périodiques sur le produit 2 en fonction de l'ancienneté et de l'âge	33
2.18	Montant moyen des versements périodiques sur le produit 3 en fonction de l'ancienneté et de l'âge	33
2.19	Montant moyen des versements périodiques par type de gestion	34
2.20	Montant moyen des versements périodiques par catégorie de montant de rachats	35
2.21	Montant moyen des versements périodiques par catégorie de montant d'arbitrages	35
2.22	Taux de frais sur les versements périodiques	36
2.23	Taux de frais sur les versements libres	36
2.24	Taux de frais d'acquisition	36
2.25	Montant moyen des versements périodiques en fonction du niveau de provisions mathématiques	37
2.26	Montant moyen des versements périodiques en fonction de la proportion de l'encours orientée vers un fonds en euro	37
3.1	Modélisation de la probabilité de verser en fonction du produit et de l'ancienneté	41
3.2	Triangle des versements périodiques par produit	43
3.3	Modélisation de l'évolution des montants moyens en fonction du produit et de l'ancienneté	45
3.4	Modélisation des montants moyens sur les données de 2021 en fonction de l'ancienneté et des montants moyens de 2020 (gauche) et 2017 (droite)	46
3.5	Compromis biais variance	48
3.6	Exemple d'un arbre CART sur un partitionnement en dimension 2	50
3.7	Profondeur de l'arbre en fonction du paramètre de complexité	56
3.8	Score de prédiction en fonction du paramètre de complexité sur la base d'apprentissage et de test	56
3.9	Score de prédiction du test en fonction du paramètre de complexité	56
3.10	Arbre optimal	57
3.11	Importance des variables selon le modèle CART optimisé sur le produit 1	58
3.12	Valeur de l'erreur OOB en fonction du nombre d'arbres	59
3.13	Importance des variables selon le modèle de forêts aléatoires optimisé sur le produit 1	59
3.14	Importance des variables selon le modèle XGBoost optimisé sur le produit 1	60
3.15	Importance des variables selon le modèle XGBoost optimisé sur le produit 2	61

3.16	Importance des variables selon le modèle XGBoost optimisé sur le produit 3	62
3.17	Méthode du coude sur les observations du produit 1	65
3.18	Versement périodique moyen par groupe en fonction de l'ancienneté pour le produit 1 . .	66
3.19	Versement périodique moyen par groupe en fonction de l'ancienneté pour le produit 2 . .	67
3.20	Versement périodique moyen par groupe en fonction de l'ancienneté pour le produit 3 . .	68
4.1	Module des risques du SCR en formule standard	74
4.2	Paramètres retenus pour les sous-groupes du produit 1	78
4.3	Paramètres retenus dans l'approche globale du produit 1	78
4.4	Paramètres retenus pour les sous-groupes du produit 2	79
4.5	Paramètres retenus dans l'approche globale du produit 2	79
B.1	Montant des versements périodiques moyen par modalités de situation familiale après retraitement sur l'ensemble du portefeuille	90
B.2	Corrélation des variables quantitatives sur le produit 1	90
B.3	Corrélation des variables quantitatives sur le produit 2	91
B.4	Corrélation des variables quantitatives sur le produit 3	91
D.1	Paramètres retenus pour la rentabilité du nouveau produit PER	94
E.1	Évolution du montant moyen des versements périodiques sur le produit 1 depuis 2011 . .	95
E.2	Évolution du montant moyen des versements périodiques sur le produit 2 depuis 2015 . .	95
E.3	Évolution du montant moyen des versements périodiques sur le produit 3 depuis 2010 . .	95

Liste des tableaux

2.1	Répartition des produits dans la base	18
2.2	Répartition des individus selon leur genre	19
2.3	Catégories des professions	19
2.4	Codes des régions	24
2.5	Fréquence des modalités de professions après retraitement	28
3.1	Montant moyen de versements après une année d’ancienneté	42
3.2	Évolution des montants de versement d’une année sur l’autre	44
3.3	Projection des modèles à partir des montants moyens de 2020 sur les données de 2021	47
3.4	Performance de la prédiction du montant de versements par le modèle CART	58
3.5	Performance de la prédiction du montant de versements par le modèle de forêts aléatoires	59
3.6	Performance de la prédiction du montant de versements par le modèle XGBoost	60
3.7	Synthèse des indicateurs de performance des modèles sur les données du produit 1	61
3.8	Synthèse des indicateurs de performance des modèles sur les données du produit 2	61
3.9	Synthèse des modèles sur les données du produit 3	62
3.10	Indicateurs de performance sur les projections des montants sur les données de 2021	62
3.11	Résultats des groupes formés sur le produit 1	66
3.12	Résultats des groupes formés sur le produit 2	67
3.13	Résultats des groupes formés sur le produit 3	67
4.1	Rentabilité des produits en prenant en compte la mise à jour des lois de versements périodiques	80
4.2	Rentabilité de produit 1 par type de gestion	80
4.3	Rentabilité du produit 1 par type de gestion et par sous-groupes	81

4.4	Rentabilité du produit 2 par type de gestion	81
4.5	Rentabilité du produit 2 par type de gestion et par sous-groupes	82
D.1	Rentabilité de produit 3 par type de gestion	94