INSTITUT DES ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Alexandra GAUTRON

Titre du mémoire :

Refonte de l'indice client : modélisation du comportement des assurés et calcul de la valeur client épargne

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 — Signature

Entreprise :

Nom : AXA France

Signature :

Directeur de mémoire en entreprise :

Nom : Rafaël HISQUIN

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Membres présents du jury de la filière

Résumé

Afin de déterminer et de mieux piloter la rentabilité de leur portefeuille d'assurés, les directions métiers d'AXA France ont bâti un projet commun d'indice client. Un indice est affecté à chaque détenteur d'au moins un contrat d'assurance AXA France, représentant la valeur économique globale de l'assuré pour la compagnie. Cet indice est déterminé par l'agrégation des valeurs individuelles du client sur chacun de ses contrats. Il doit permettre d'orienter la stratégie de la compagnie afin de retenir les clients créateurs de richesse, en prévenant leur départ.

L'enjeu et l'apport de ce mémoire ont consisté à améliorer le modèle de la valeur client épargne, en changeant de méthodologie et en affinant le calcul de la valeur grâce à la segmentation du comportement client en assurance vie (pour les trois types de flux principaux : arbitrages, rachats et versements). Cette approche a mêlé des techniques statistiques poussées sur des grands volumes de données et a intégré des indicateurs actuariels de rentabilité à destination des réseaux de distribution et agents généraux particulièrement.

À l'issue des travaux, chaque client se voit affecter une valeur qui reflète mieux son comportement et ses caractéristiques, et permet d'améliorer l'orientation des stratégies commerciales. Au-delà d'un simple indicateur marketing stratégique permettant d'orienter les agents, la valeur client épargne pourra également servir à l'assureur pour contrôler sa rentabilité, mieux cibler les profils risqués dans son portefeuille, et compléter les indicateurs actuariels classiques de rentabilité.

Mots-clés :
Indice client, Épargne, Valeur, Comportement, Segmentation, *Machine Learning*

Abstract

In order to determine and better manage the profitability of their portfolio of policyholders, AXA France's business divisions have developed a joint customer index project. An index is assigned to each holder of at least one AXA France insurance policy, representing the policyholder's overall economic value to the company. This index is determined by aggregating the customer's individual values on each of its contracts. It aims to guide the company's strategy to retain wealth-creating clients by preventing their departure.

The challenge and contribution of this thesis was to improve the savings customer value model by changing the methodology and refining the calculation of value through segmentation of customer behaviour in life insurance (for the three main types of flows : fund switches, surrenders and premiums). This approach combined advanced statistical techniques on large volumes of data and integrated actuarial profitability indicators, particularly for distribution networks and tied agents.

At the end, each client is assigned a value that better reflects its behaviour and characteristics, and allows better orientation of sales strategies. Beyond a simple strategic marketing indicator to guide agents, the savings customer value can also be used by the insurer to control profitability, better target the risk profiles in its portfolio, and complement the traditional actuarial profitability indicators.

Key words :
Client Index, Savings, Value, Behavior, Clustering, Machine Learning

Remerciements

Je tiens tout d'abord à remercier Jean-Baptiste Peyre de m'avoir recrutée et permis de réaliser une alternance au sein de la direction technique Épargne d'AXA France.

Au sein de l'équipe, je souhaite particulièrement remercier Rafaël Hisquin, qui m'a encadrée et aidée tout au long de cette alternance. Il a su se rendre disponible et partager ses connaissances et expériences, ainsi que sa confiance et sa bonne humeur et je lui en suis très reconnaissante.

Je saisis également cette occasion pour remercier l'ensemble des autres membres de l'équipe, pour leur accueil chaleureux, leurs conseils éclairés et leur aide constante. Un grand merci également à Marine Terrasson et Ibtissam Benibrahim, notamment pour leur collaboration au projet commun d'indice client.

Par ailleurs, je voudrais remercier les professeurs et intervenants du Master 2 d'Actuariat de l'ISUP, ainsi que mon tuteur académique, Olivier Lopez, pour les connaissances qu'ils nous ont transmises, grâce auxquelles les travaux en entreprise, ainsi que ce mémoire, ont pu être sereinement abordés.

Enfin, je remercie toutes les personnes qui ont contribué de près ou de loin à ce mémoire, que ce soit par leurs conseils, leur soutien ou leurs encouragements.

Table des matières

Résumé **1**

Abstract **2**

Remerciements **3**

Introduction **6**

1 Contexte **9**
1.1 Contexte général et chiffres clés . . . 9
1.2 Spécificités d'un contrat d'assurance vie . . . 11
1.2.1 Éléments généraux . . . 11
1.2.2 Caractéristiques des contrats . . . 13
1.2.3 Opérations d'un client sur son contrat . . . 17
1.3 Quelques éléments de rentabilité et de valeur . . . 22
1.4 Assurance, actuariat et *Big Data* . . . 26

2 Objectifs de l'étude sur la valeur client épargne **28**
2.1 Indice client . . . 28
2.2 Modèle actuariel de la valeur client épargne . . . 30
2.3 Travaux et apport du mémoire . . . 32
2.4 Données et segmentation du portefeuille . . . 34
2.5 Déroulé du projet . . . 37

3 Présentation des données de l'étude **38**
3.1 Extraction . . . 38
3.2 Qualité des données . . . 41
3.3 Retraitements . . . 42
3.3.1 Données manquantes . . . 42
3.3.2 Valeurs aberrantes . . . 44
3.3.3 Enrichissement de la base de données . . . 46
3.4 Statistiques descriptives . . . 48
3.5 Contraintes réglementaires . . . 54

4 Segmentation et détermination des lois de comportement **55**
4.1 Présentation des modèles et des différentes méthodes de segmentation . . . 55
4.1.1 Notions utiles pour la suite . . . 55
4.1.2 Méthode 1 : segmentation directe de la base . . . 60
4.1.3 Méthode 2 : prédiction des lois puis segmentation selon les comportements . . . 71
4.2 Comparaison des résultats et choix de la méthode de segmentation . . . 76

5 Calcul de la valeur épargne, interprétation et déploiement **84**

5.1 Calcul de la valeur épargne 84
5.2 Analyse de la segmentation du portefeuille et de la valeur épargne 89
5.3 Limites des modèles et améliorations envisageables 98
5.4 Perspectives de déploiement de l'indice client 102

6 Valeur client épargne, rentabilité et risque **104**

Conclusion **106**

Bibliographie **109**

Annexe **112**

Introduction

Afin de déterminer et de mieux piloter la rentabilité de leur portefeuille d'assurés, les directions métiers d'AXA France ont bâti un projet commun d'**indice client**. Un indice est affecté à chaque détenteur d'au moins un contrat d'assurance AXA France, représentant la **valeur économique** globale de l'assuré pour la compagnie. Cet indice est déterminé par l'**agrégation** des valeurs individuelles du client sur chacun de ses contrats. Il doit permettre d'orienter la stratégie de la compagnie afin de retenir les clients créateurs de richesse, en prévenant leur départ. L'étude se concentre sur le périmètre de l'épargne/retraite individuelle.

Dans l'ancien modèle de valeur client épargne, une valeur fixe était attribuée à chaque contrat en fonction de son encours et de sa part UC. Jugée peu satisfaisante puisque ne permettant pas de différencier des assurés aux caractéristiques et comportements différents, elle a été remplacée par un modèle plus complexe.
Dans la littérature, la valeur client est généralement définie comme la **valeur actuelle** de tous les **revenus** futurs attendus d'un client sur sa durée totale de relation avec une entreprise, **diminués** des **coûts** associés.[9] Déterminée individuellement pour chaque client, elle permet de différencier ceux qui sont rentables des autres, plutôt que d'analyser une rentabilité moyenne globale. De plus, elle prend en compte la possibilité que le client interrompe la relation avec l'entreprise (rachat ou décès dans le cas de l'épargne/retraite).

La Direction Technique d'AXA France a alors retenu une nouvelle définition de la valeur client épargne :

$$\begin{matrix}\text{Valeur}\\ \text{épargne}\end{matrix} = (\text{PM } € \times \text{Coeff. } € + \text{PM UC} \times \text{Coeff. UC}) \times \begin{matrix}\text{Duration}\\ \text{contrat}\end{matrix} - \begin{matrix}\text{Frais}\\ \text{d'acquisition}\end{matrix}$$

avec :

$$\text{Coeff.} = \frac{\text{VIF}_{stoch} - \text{TVFOG} - \text{MVM} - \text{VAN FG}}{\text{PM} \times \text{Duration}}$$

La **PM €** représente l'encours détenu par l'assuré sur le fonds euro, et la **PM UC** celui sur des UC. La **TFVOG** (*Time Value of Financial Options and Guarantees*) mesure la valeur temps des options et garanties financières. La **MVM** (*Market Value Margin*) désigne la marge pour risque et représente le coût d'immobilisation des fonds propres réglementaires. L'estimation des **coefficients** fait également intervenir la valeur actuelle nette des frais de gestion (**VAN FG**). Les **frais d'acquisition** sont calculés pour chaque contrat via un forfait de frais d'acquisition. La **duration** des contrats peut être interprétée comme une **espérance de vie résiduelle de l'assuré en portefeuille**.

La **VIF** (*Value of In-Force*) représente la valeur actuelle des profits futurs issus du portefeuille acquis. Elle est surtout déterminée par le **comportement futur des assurés** et fait donc appel à des **lois de comportement** en épargne (arbitrages, rachats et versements). Déterminées à la maille produit dans le modèle interne, ces lois ne permettent pas de **différencier** les **assurés** d'un même produit selon leur **comportement**. En d'autres

termes, deux individus du même produit mais au comportement totalement différent (donc au potentiel de rentabilité également différent) pourraient se voir attribuer les mêmes lois, donc la même VIF. Puisque c'est la VIF qui permet principalement de déterminer les coefficients euro et UC, cet indicateur possède une grande importance.

Le but du mémoire est de **segmenter le portefeuille d'assurés** en classes de comportement, afin de leur affecter des lois différentes, ce qui impactera leur VIF respective et leurs coefficients euro et UC. À terme, cet affinement de la maille de calcul de la VIF et des coefficients doit permettre d'attribuer des **valeurs différentes** à des clients appartenant à des **classes de comportement différentes**, même s'ils se trouvent sur le même produit, à PM équivalente.

La première étape a consisté à prendre connaissance des travaux réalisés en interne par les équipes d'actuariat, puis à effectuer une recherche sur les études menées sur la valeur client, en général et d'un point de vue actuariel pour l'assurance ([1],[13]), ainsi que sur les techniques avancées de *machine learning* ([2], [10]).

Le mémoire a consisté à retenir les méthodes les plus avancées, compatibles avec les souhaits internes, et permettant de faire évoluer les outils en place.

Ainsi, pour réaliser la segmentation du portefeuille en différentes classes de comportement, deux méthodes ont été testées :

- **Méthode 1** : segmentation directe de la base de données, à partir des variables à disposition (telles que les caractéristiques du contrat, de l'assuré, de l'historique de ses flux par exemple). L'objectif est de rapprocher les clients qui se ressemblent en terme de comportement, selon ces variables ;

- **Méthode 2** : projection des montants de flux (arbitrages, rachats et versements) pour chaque contrat sur 30 ans pour construire des lois de comportement individuelles. Une classe de comportement ("*cluster*") est alors affectée à chaque client, qui permet de le rapprocher des autres assurés ayant une loi (donc un comportement prédit) proche.

Cette approche a mêlé des techniques **statistiques** poussées sur des **grands volumes de données** et a intégré des **indicateurs actuariels de rentabilité** à destination des réseaux de distribution et agents généraux particulièrement. Notamment, différents algorithmes de *machine learning* ont été testés pour des tâches de classification et de régression, et une phase de sélection de modèle a permis de déterminer le meilleur.
Les deux méthodes de segmentation ont ensuite été comparées, afin de retenir la plus performante. Pour ce faire, deux indicateurs reflétant une **vision métier** ont été créés : un indicateur d'écart de comportement entre les différents *clusters*, et un indicateur d'erreur de prédiction. Une fois la meilleure méthode de segmentation retenue, des **lois de comportements** ont été construites pour chaque type de flux (arbitrages, rachats et versements), et pour chaque classe de comportement. Elles ont alors permis d'affiner le calcul de la VIF.

À l'issue des travaux, chaque client se voit affecter une **valeur** qui **reflète mieux son comportement et ses caractéristiques**, et permet de mieux orienter les stratégies commerciales. Au-delà d'un simple indicateur marketing stratégique permettant d'orienter les agents, la valeur client épargne pourra également servir à l'assureur pour **contrôler sa rentabilité**, mieux **cibler les profils risqués** dans son portefeuille dont les motivations auront été mieux identifiées, et **compléter les indicateurs actuariels classiques** de rentabilité et de gestion des risques (dans le cadre de l'ORSA notamment).

Après une présentation générale du **contexte** et des enjeux actuels en assurance vie, ainsi qu'une description du **projet** et des aspects **techniques** sous-jacents, le mémoire s'attachera à illustrer l'**application** de ces **méthodes** pour améliorer le calcul de la **valeur client** en pratique. Enfin, après une **analyse** et une interprétation des nouvelles valeurs clients obtenues, une dernière partie expliquera comment ce nouvel indicateur pourra être utilisé par l'assureur pour **améliorer sa connaissance et sa gestion des risques**.

1 Contexte

1.1 Contexte général et chiffres clés

Dans cette partie, le cadre général de l'assurance vie en France est présenté. À travers une description de ses enjeux actuels, de son fonctionnement et de ses évolutions récentes, les fondamentaux sont exposés afin de comprendre les enjeux de l'étude sur l'indice client.

L'assurance vie occupe une place importante dans le système économique Français, et, malgré une image ternie par la baisse des taux, semble toujours susciter l'engouement des Français.
En effet, selon une revue de la Banque de France [31], elle représente environ 31% du total des placements financiers des ménages Français au dernier trimestre 2019, avec près de **1700 milliards d'euros d'encours** (premier poste de placements financiers).

Aussi, les activités Vie confirment leur place au cœur du système assurantiel Français. Une hausse de 4% du montant annuel de primes collectées par les sociétés d'assurance sur le marché de l'épargne Français a été observée fin 2019 par rapport à l'année précédente, avec un montant atteignant **145 milliards d'euros** [33].
Les versements ont été plus fréquemment effectués sur le fonds euro que sur des unités de compte ("UC"), qui comptabilisent 27% des versements, comme en 2018 [35]. Cependant, la part d'UC dans la collecte a connu une forte évolution tout au long de l'année 2019, passant d'environ 22% de la collecte en janvier 2019 à environ 41% en décembre 2019.
Les prestations versées sur la même période sont stables par rapport à l'année précédente, et la **collecte nette** s'élève à près de **26 milliards d'euros** sur l'ensemble de l'année, soit une hausse de 20%, meilleur résultat depuis 2010 (51 milliards d'euros). [33].

Cependant, une année 2019 exceptionnelle a laissé place à un début d'année **2020** très **perturbé** et incertain. L'environnement de **taux bas voire négatifs** semble perdurer, et le secteur a été frappé par la crise mondiale du **Coronavirus**. Le caractère atypique de ces premiers mois (confinement de mi-mars à mi-mai en France, puis déconfinement progressif et maintien de précautions sanitaires strictes) a restreint le nombre d'opérations et de placements en assurance vie et a engendré une **collecte mensuelle nette négative** au cours de plusieurs mois en 2020, alors que cette dernière ne l'avait pas été depuis décembre 2018. [34]

Enfin, la **situation démographique** suit la même évolution que celle des dernières années. La part des plus de 60 ans dans la population augmente toujours, et des gains d'espérance de vie ont été réalisés tout au long de ces dernières décennies. En parallèle, la majorité des contrats d'épargne-retraite est détenue par des clients de plus de 45 ans.
Ce vieillissement de la population joue un rôle prépondérant dans l'évolution des encours, des engagements des sociétés d'assurance vie, et des risques auxquels elles doivent faire face, avec un fort accroissement des prestations versées (coûts de plus en plus importants pour les restitutions en rente).

Dans le contexte économique actuel de taux bas voire négatifs, qui semble durable et per-

turbé par la crise du Coronavirus, associé au **contexte prudentiel et comptable** toujours plus prudent et exigeant avec l'apparition de nouvelles normes comptables (IFRS17), et à la situation démographique de moins en moins avantageuse, les sociétés d'assurance sont davantage contraintes et leurs fonds propres de plus en plus contrôlés.

Elles ont alors dû prendre certaines **mesures** afin de limiter les effets négatifs de cet environnement incertain, agité et défavorable, comme par exemple une diminution des rendements servis sur le fonds euro, ou une incitation forte des assurés à diversifier leurs investissements (dans le cas d'investissements sur des UC, c'est l'assuré qui supporte le risque et plus l'assureur).

Face à ces divers constats, les sociétés d'assurance sont plus que jamais appelées à une grande **prudence** dans la modélisation et la gestion de leurs fonds propres.
Dans le cadre des activités d'épargne-retraite, elles sont donc amenées à appréhender et modéliser au mieux leurs **engagements futurs**, afin de mieux piloter leur rentabilité et assurer leur solvabilité. Dans une telle logique, une meilleure approche des comportements des assurés semble nécessaire. Elle passe notamment par une compréhension approfondie des motivations et caractéristiques des épargnants, ainsi qu'une **modélisation plus fine** des flux associés aux contrats en portefeuille.

C'est dans ce contexte que les différentes directions métiers d'AXA France ont bâti un projet commun d'**indice client**. Construit, pour chaque client en portefeuille, à partir de l'agrégation des différentes valeurs client calculées sur chacun des différents contrats détenus par ce client, cet indice vise à en caractériser la **rentabilité** potentielle.
Ce mémoire traite de la **refonte du modèle de valeur client épargne**, dont la méthodologie sera approfondie dans la partie 2. La valeur épargne de chaque client est calculée en utilisant différents éléments de rentabilité, comme la VIF par exemple (valeur actuelle des profits futurs issus du portefeuille acquis). Pour répondre à un besoin d'affinement et d'**individualisation** de la valeur client épargne, le nouveau modèle proposé prend en compte les différences de **caractéristiques** et de **comportements** entre assurés, qui doivent se refléter dans la **VIF** notamment. Pour ce faire, une **segmentation du portefeuille** est réalisée selon ces caractéristiques, qui permet de calculer des lois de comportement propres à chaque segment. Ces lois interviennent directement dans le calcul de la VIF, dont la nouvelle maille de calcul sera donc plus précise. Ces étapes doivent permettre d'individualiser les valeurs épargnes finales, qui devraient être plus appropriées aux différences de caractéristiques et de comportement entre assurés.
Les sections suivantes présentent des généralités sur les contrats d'assurance vie, dont les différences de caractéristiques devront être prises en compte dans les modélisations. Aussi, quelques éléments de rentabilité et de valeur sont présentés, pour mieux comprendre par la suite le calcul de la valeur client épargne.

En résumé : Contexte général et chiffres clés

- Une année 2019 exceptionnelle ;
- Une année 2020 fortement perturbée ;
- Un contexte démographique peu favorable ;
- Un contexte réglementaire et comptable exigeant ;
- Un projet d'indice client visant à caractériser la rentabilité potentielle de chaque assuré en portefeuille.

1.2 Spécificités d'un contrat d'assurance vie

Les spécificités présentés ci-après illustrent des éléments qui seront pris en compte dans les travaux par la suite. Leur description permet également d'expliquer certains choix effectués dans la sélection des variables.

1.2.1 Éléments généraux

Un **contrat d'assurance** est une opération par laquelle, moyennant le paiement d'une prime, fixe ou variable, une partie, l'assureur, s'engage envers une autre partie, le preneur d'assurance, à fournir une prestation stipulée dans le contrat au cas où surviendrait un événement incertain que, selon le cas, l'assuré ou le bénéficiaire a intérêt à ne pas avoir réalisé.

L'assurance vie est une branche de l'assurance de personnes et couvre la prévoyance (risque décès), l'épargne et la retraite (ce mémoire se concentre sur le périmètre de l'épargne et de la retraite individuelle).

Les principaux objectifs de l'assurance vie consistent à :

- Constituer un capital pour préparer ou compléter sa retraite ;
- Profiter d'un placement fiscalement avantageux ;
- Transmettre un capital à ses proches en cas de décès ;
- Constituer un capital en vue d'un projet personnel (achat par exemple), ou professionnel ;
- Disposer d'une épargne de précaution à court terme ;
- Faire face aux imprévus et anticiper les risques liés à la vie (prévoyance, dépendance).

Plusieurs **parties** interviennent dans un contrat d'assurance vie.
L'**assureur** collecte les primes, réalise les placements financiers, paie les prestations et est également en général le gestionnaire du contrat. L'**assuré** est la personne sur laquelle repose le risque. L'**adhérent/souscripteur** réalise les demandes d'acte de gestion, effectue un (ou des) versement(s), rachat(s), souscrit un service, etc. Le **bénéficiaire** est désigné par l'adhérent/souscripteur et perçoit les prestations en cas de décès de l'assuré avant l'échéance du contrat. Enfin, le **distributeur** est un intermédiaire entre l'assureur et les clients, auxquels il propose des contrats (agents généraux, CGPI, banques ou assureur).

Dans la plupart des cas, le souscripteur se confond avec l'assuré. Cependant, il est possible que ces deux individus diffèrent : cette situation est toutefois marginale, et se produit pour seulement 1% des contrats de l'étude. Dans la suite du mémoire, le terme assuré sera donc parfois utilisé pour définir l'épargnant qui souscrit le contrat.
Dans le cadre du projet d'indice client, ce sont les actes d'arbitrages, de versements et de rachats qui sont étudiés. Ainsi, ce sont les caractéristiques de l'individu ayant un droit d'exercer ces options sur le contrat qui doivent être prises en compte. C'est pourquoi, lorsque l'assuré et le souscripteur sont deux personnes différentes, les données clients utilisées sont celles du souscripteur.

Il convient de distinguer deux types de contrats : les **contrats individuels** et les **contrats collectifs**.
Pour les premiers, l'épargnant (souscripteur) est directement lié à l'assureur par son contrat, alors que dans le deuxième cas, une association (comme ANPERE pour le produit Arpèges) ou un employeur fait le lien entre plusieurs épargnants et l'assureur, et représentera donc la personne morale qui souscrit le contrat auprès de l'assureur. C'est cet intermédiaire qui a en charge la gestion du contrat et qui peut exercer les différentes options du contrat. La personne physique, adhère au contrat collectif ("adhérent"). Cette deuxième forme de contrat présente un avantage dans la gestion des contrats pour l'assureur qui, pour une décision telle qu'une modification de conditions générales (CG) par exemple, n'a pas besoin de l'accord individuel de chacun des épargnants, mais uniquement de celui de l'association qui les représente. En cas de désaccord des épargnants, ceux-ci peuvent bien sûr exercer un droit de retrait (rachat du contrat ou action de groupe si un ensemble de clients jugent que la décision de l'association ne leur était pas favorable). L'assureur peut donc modifier les dispositions du contrat en traitant avec l'unique souscripteur, et des garanties collectives peuvent être proposées.

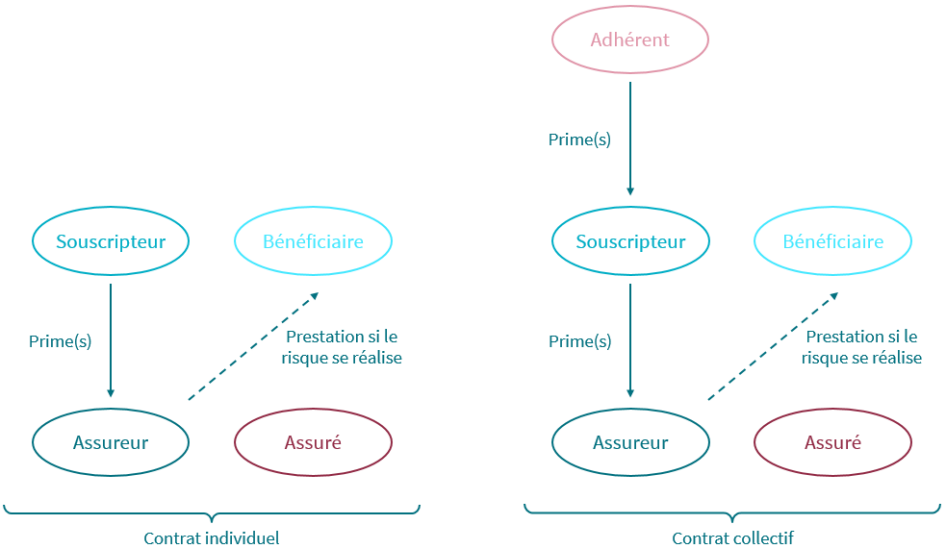

FIGURE 1 – Parties intervenant dans un contrat d'assurance vie

Deux risques complémentaires existent en assurance vie :

- le risque de **décès** (capital ou rente versé(e) en cas de décès de l'assuré) ;
- le risque de **survie** (capital ou rente versé(e) en cas de survie de l'assuré à l'échéance du contrat).

En épargne, l'aléa porte sur la durée de vie de l'assuré. La prime correspond à la cotisation payée par l'assuré. Elle peut être unique (lors de la création du contrat), ou périodique (programmée ou libre). La prestation, si elle existe, peut prendre la forme d'un capital ou d'une rente versé(e) au bénéficiaire.

Deux grandes catégories de produits épargne/retraite existent chez AXA France : les produits du ***Mass Market*** (produits grand public) et ceux du ***Wealth*** (produits sur-mesure, pour une clientèle possédant généralement un encours plus conséquent).

Ces produits sont distribués sur plusieurs **réseaux** selon leur catégorie :

- Produits du *Mass Market* :
  - Agents généraux ;
  - Agents prévoyance et patrimoine ;
  - Agents épargne et protection ;
  - Courtiers ;

- Produits du *Wealth* :
  - Partenaires bancaires ;
  - Conseillers en gestion de patrimoine ;
  - Gestion privée.

L'étude de la valeur client épargne se concentre sur les produits du ***Mass Market***, sans inclure les contrats distribués par les courtiers. En effet, l'indice client est un indicateur à destination des agents et salariés d'AXA France, distribuant des contrats sur ses réseaux propres.

1.2.2 Caractéristiques des contrats

L'épargne des contrats peut être investie sur deux types de supports : le **fonds euro** et les **unités de compté ("UC")**.

Le premier accorde à l'assuré une **garantie sur le capital investi**, historiquement nette de frais de gestion (donc l'assuré ne peut pas perdre d'argent), mais brute de frais de gestion dans la plupart des nouveaux contrats. Certains contrats disposent en plus d'un taux garanti :

- Défini contractuellement sur toute la durée de vie du contrat, pour certains anciens contrats : c'est le **Taux Minimum Garanti** ("**TMG**") ;

- Redéfini chaque année par l'assureur en fonction des performances financières des actifs associés au fonds euro, et actuellement nul pour les nouveaux contrats : c'est le **Taux Minimum Garanti Annuel** ("**TMGA**") .

Le capital (brut de frais de gestion) est garanti à tout moment, et les intérêts générés annuellement définitivement acquis : c'est l'effet cliquet. Le fonds euro est donc plutôt adapté aux profils averses au risque, recherchant un rendement sûr mais limité.
Pour respecter ces engagements forts de garantie en capital et un taux de valorisation minimum, l'assureur gère librement le fonds mais avec des contraintes réglementaires fortes et une répartition sur des actifs a priori non risqués (majoritairement des actifs obligataires, et une part plus faible d'immobilier et d'actions).

Contrairement au fonds euro, les **UC** n'offrent **pas de garantie sur le capital investi**. L'encours sur UC est alors exprimé en nombre de parts. Seul ce nombre est garanti, la valeur de chaque part évoluant en fonction des marchés financiers. Aucune garantie n'est proposée sur la revalorisation : le souscripteur supporte seul le risque.
Les UC peuvent regrouper des parts d'OPCVM (Organismes de Placement Collectif en Valeurs Mobilières) telles que des SICAV (Sociétés d'Investissement à Capital Variable) ou des FCP (Fonds Communs de Placement), des actions, des obligations, ou encore des parts de SCPI (Sociétés Civiles de Placement Immobilier) et autres placements immobiliers.
Enfin, des garanties peuvent être proposées sur certains contrats, comme la garantie plancher (garantie en capital au moment du décès) ou la garantie en cas de vie.

Du point de vue du souscripteur, la différence globale entre ces deux supports d'investissements peut être résumée par le couple rendement-risque. Le fonds euro délivre un rendement faible mais bénéficie d'un faible risque de perte du capital investi en raison des garanties offertes par l'assureur. Les UC, elles, offrent des perspectives de rendement beaucoup plus élevées, mais au prix d'un risque également beaucoup plus élevé pour le souscripteur.
Ces différences en termes de risques et de garanties (donc de rémunération) entre contrats sur fonds euro ou en UC peuvent expliquer des différences de comportements entre épargnants. Elles devront donc figurer dans les variables prises en compte pour l'étude effectuée dans ce mémoire.

Du point de vue de l'assureur, le fonds euro nécessite une immobilisation en capital plus forte d'après la réglementation Solvabilité II, puisqu'il représente un engagement plus fort envers les assurés (garantie sur le capital investi, alors que pour les UC, l'assureur s'engage uniquement sur un nombre de part, pas une valeur) et impacte sa marge de solvabilité (SCR).

Un nouveau type de fonds est apparu ces dernières années à mi-chemin entre le fonds euro et les UC : l'**Eurocroissance**, modernisé par la récente loi Pacte.
Un contrat dit Eurocroissance comporte un encours exprimé en euros et en parts de diversification. La partie exprimée en euros permet d'assurer la garantie du capital au terme (seulement à partir d'une échéance définie dans le contrat, au moins égale à huit ans à compter du premier versement). Les engagements exprimés en parts de diversification per-

mettent de "booster" les performances du support. Comme pour les UC, l'assureur ne garantit cette fois encore qu'un nombre de parts, la valeur de chacune étant évaluée en fonction des résultats techniques et financiers du fonds. Cependant, en raison de la garantie au terme, la part de diversification ne peut pas perdre trop de valeur puisqu'au terme du contrat, la somme de la partie en euros et de la partie de diversification doit au moins être égale à l'investissement initial.
Ce type de contrats possède a la fois des avantages pour l'épargnant (espérance de gain plus importante grâce à l'immobilisation du capital jusqu'au terme du contrat) et pour l'assureur (possibilité d'une stratégie de placement à long terme, et absence d'exposition à une exigence permanente de liquidité des fonds sur ce périmètre). La loi Pacte a simplifié le fonctionnement et propose une deuxième version dans laquelle tous les encours sont gérés sous forme de parts de provision de diversification.

Les contrats pour lesquels le capital est investi uniquement sur le fonds euro sont appelés **contrats monosupport**. Les **contrats multisupports**, qui représentent l'essentiel des contrats en portefeuille, répartissent l'épargne sur les différents types de supports (fonds euro, UC, Eurocroissance).

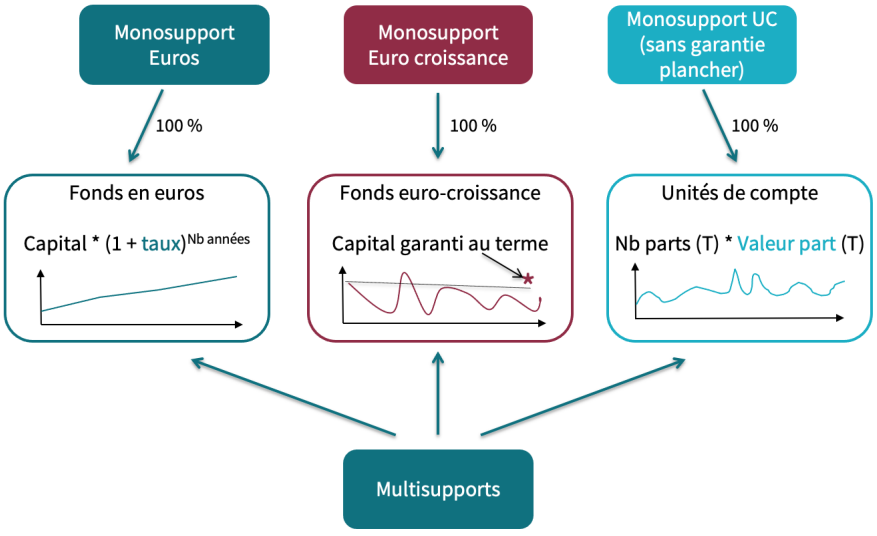

FIGURE 2 – Différents types de contrats, supports et garanties

La base de données utilisée dans l'étude sur la valeur client épargne contient à la fois des contrats mono- et multisupports, avec tous les types de supports d'investissement possibles (fonds euro, UC, Eurocroissance).

La réglementation impose aux assureurs de reverser aux assurés une partie des bénéfices réalisés : au moins **85%** des bénéfices **techniques** et **90%** des bénéfices **financiers**.

La **Participation aux bénéfices (PB)** se décompose alors en deux parties :

- la **PB contractuelle**, qui est définie au niveau d'un contrat ou d'un ensemble de contrats et représente la participation des assurés aux bénéfices des gestions technique et financière de l'assureur. L'ensemble des PB contractuelles versées par l'assureur ne peut pas être inférieur à la PB définie par la réglementation. Dans le cas contraire, l'assureur doit compléter les PB versées aux contrats pour atteindre a minima la PB réglementaire. ;

- la **PB discrétionnaire**, qui est donnée à la discrétion de l'assureur en plus de la PB contractuelle, et dont le but est de proposer une offre concurrentielle permettant à l'assureur de se replacer par rapport au marché (notamment pour éviter des vagues de rachats massifs l'année suivante, ou inciter les clients à orienter leur épargne vers les UC par exemple).

Enfin, la PB est affectée en fin d'année par un taux de PB définitif entraînant une augmentation de la valeur de rachat de l'assuré. L'indication de ce taux est essentielle pour comprendre le mécanisme de rémunération d'un contrat d'épargne retraite, il convient donc de la prendre en compte dans le cadre de la modélisation du comportement des épargnants.

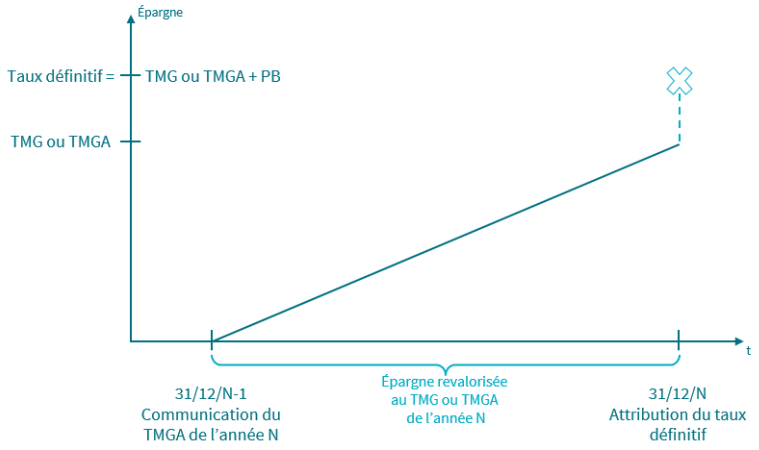

FIGURE 3 – Mécanisme de revalorisation de l'épargne sur une année

Le **Bonus "Euro +"** est un dispositif mis en place en 2010, qui récompense la diversification de l'épargne de certains contrats multisupport respectant des contraintes définies, au travers d'un taux bonifié de participation aux bénéfices (PB) sur le support en euros du contrat concerné.[22]
Selon la part de l'épargne investie sur des supports en UC, le montant total de l'encours et le type de gestion choisi pour le contrat, l'épargnant peut ainsi toucher jusqu'à 120% du taux de PB de son support en euros.

Ce bonus impacte donc fortement la rémunération des contrats qui y sont éligibles, c'est pourquoi il semble pertinent de le prendre en compte parmi les variables explicatives de l'étude.

L'**amendement Fourgous** de juillet 2005 permet aux épargnants de transférer la totalité de l'épargne de leur contrat monosupport vers des contrats multisupport tout en conservant l'antériorité fiscale du contrat monosupport originaire, auprès du même assureur et sous condition qu'au moins 20% de l'épargne soit placée sur des UC. Le transfert permet alors de profiter d'une fiscalité sur les produits plus intéressante au-delà de huit ans.
Il paraît alors raisonnable d'estimer qu'un épargnant ayant transféré son encours d'un contrat monosupport vers un contrat multisupport soit plus informé et ait une plus grande culture financière, mais surtout soit enclin à effectuer des arbitrages, qui lui étaient jusqu'alors impossibles. Une variable indiquant si le contrat a été "fourgoussé" semble donc pertinente car elle apporte une information sûrement significative sur le comportement de l'épargnant (au moins pour les arbitrages).

1.2.3 Opérations d'un client sur son contrat

D'après la réglementation en vigueur, le souscripteur détient une part de contrôle de son contrat d'assurance vie, via plusieurs opérations qui lui sont proposées contractuellement :

- **Arbitrages** : l'épargnant a la possibilité de réorienter son épargne d'un support vers un autre, pendant toute la vie de son contrat. En plus des arbitrages manuels, des arbitrages dits "automatiques" peuvent exister. Ils correspondent par exemple à la réorientation annuelle de montants sur différents supports afin de respecter l'allocation définie contractuellement, ou de dynamisation du profil de risque. Comme ils ne correspondent pas à un comportement propre au souscripteur, ces arbitrages automatiques ne seront pas étudiés dans le cadre de l'étude sur la valeur client.
  En cas d'arbitrages massifs de l'euro vers l'UC, une perte peut être réalisée par l'assureur, s'ils sont effectués dans une période où les obligations sont en moins-value, comme en période de hausse des taux obligataires par exemple. Cependant, les investissements sur UC demandent à l'assureur d'immobiliser moins de capital que ceux sur fonds euro et peuvent donc aussi lui être bénéfiques, notamment en période de baisse des taux de rendement du fonds euro, surtout si les chargements sur les UC et les fonds gérés sont plus importants que les chargements prélevés sur les fonds euros.
  En cas d'arbitrages massifs des UC vers le fonds euro, l'assureur se voit affaibli par la nécessité d'immobiliser un plus grand capital pour respecter les contraintes réglementaires. De plus, dans le contexte actuel de taux bas voire négatifs, les flux entrants sur le fonds euro se matérialisent par l'acquisition d'obligations à un taux de rendement qui est de plus en plus faible. Ceci a pour effet de "diluer" la rentabilité du portefeuille d'actifs, et donc finalement de diminuer le taux servi aux assurés. Ce risque peut entraîner l'insatisfaction des assurés qui rachèteraient alors leur(s) contrat(s) et pourrait mener l'assureur à faire face à un risque de rachat massif de

contrats. Dans ce contexte, il est important pour l'assureur de bien comprendre le comportement de ses assurés, afin de diminuer ces risques, par la prévention, les conseils ou l'accompagnement par exemple.
Par ailleurs, les arbitrages euro-euro et UC-UC sont minoritaires et ne représentent pas de risque majeur pour l'assureur. Ils ne seront donc pas étudiés dans le cadre de l'étude.

- **Versements** : à l'exception de certains contrats dits "à prime unique", cette option permet à l'épargnant d'augmenter son encours par l'injection de capitaux, de manière spontanée et ponctuelle (versements libres) ou programmée (primes périodiques).

- **Rachats** : afin de désinvestir tout ou partie de son épargne, l'épargnant a la possibilité de racheter partiellement ou totalement son contrat, à tout moment. Cette opération est cependant soumise à une fiscalité qui dépend surtout de l'ancienneté du contrat. Des rachats partiels programmés sont possibles, correspondant à des désinvestissements successifs réalisés suivant des paramètres choisis par le souscripteur (périodicité, montant, support(s) désinvesti(s)), ressemblant à une rente mais avec les avantages fiscaux offerts par l'assurance vie.
  En cas de rachats massifs de contrats, l'assureur est confronté au risque de rachat, qui se situe au coeur de la gestion de risques de la compagnie dans le cadre de Solvabilité II. Deux types de rachats sont alors à distinguer :

  - les **rachats structurels** sont modélisés par des lois de rachats d'expérience, correspondent à des taux de rachat en "rythme de croisière", que l'assureur peut observer dans un contexte économique "standard" ;

  - les **rachats conjoncturels** sont liés à un événement déformant le comportement dit "habituel" des assurés. Ils interviennent notamment dans un contexte concurrentiel lorsque l'assuré arbitre son contrat d'assurance au profit d'autres supports financiers. Ce type de rachat, par sa nature, rend impossible la construction d'une loi d'expérience. L'Autorité de Contrôle Prudentiel et de Résolution (ACPR) et les Orientations Nationales Complémentaires préconisent une fonction linéaire par morceaux qui dépend de l'écart entre un taux benchmark reflétant le marché et le taux servi par le contrat d'assurance vie.

Ces trois comportements doivent être correctement modélisés par chaque assureur, car ils influent sur son bilan et donc sa solvabilité. Ce sont ces trois comportements qui sont étudiés dans ce mémoire, afin de comprendre quelles caractéristiques clients et contrats peuvent les influencer, et donc d'aider l'assureur dans le pilotage d'actions stratégiques visant à protéger sa solvabilité. Dans cette optique, ce sont les flux structurels qui seront étudiés et pas conjoncturels.

Par ailleurs, différents modes de gestion sont proposés à l'assuré :

- la **gestion par convention** : permet à l'assuré de choisir une répartition de son épargne parmi les conventions proposées, en fonction de son profil d'investissement et de ses projets. L'épargne est réajustée chaque année grâce aux Réajustements Annuels Gratuits (RAG), pour respecter la répartition de la convention choisie.
  En cas de hausse des marchés financiers, ces réajustements permettent de sécuriser une partie des intérêts réalisés sur les supports en UC en les transférant vers le support en euros.
  En cas de baisse des marchés financiers, ils permettent de réinvestir une partie de l'épargne présente sur le support en euros vers les UC, afin de profiter d'un éventuel rendement plus élevé en contrepartie d'une diminution de l'épargne sécurisée.

  Voici des exemples de conventions pour le contrat Arpèges :

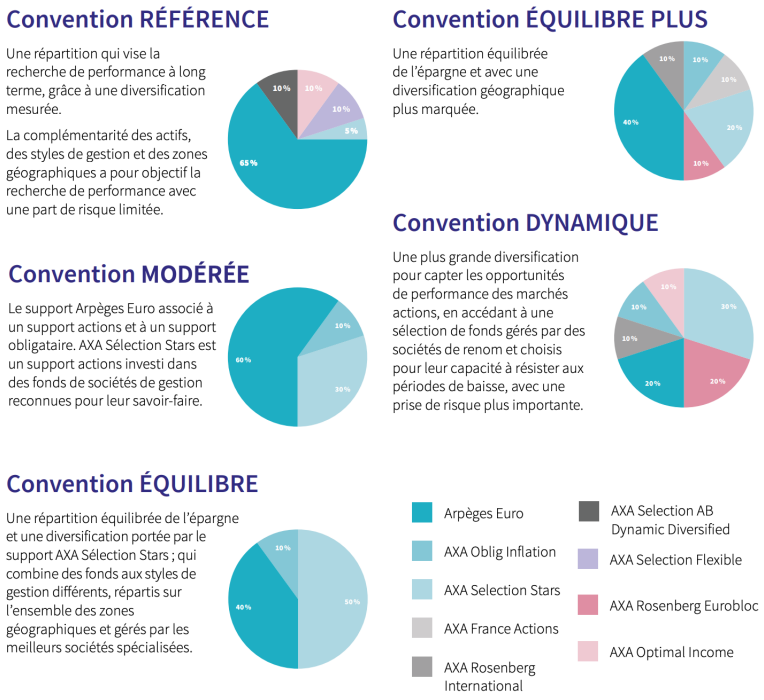

FIGURE 4 – Contrat Arpèges *(Source : Conditions Générales)*

- la **gestion sous mandat ou profilée** : permet à l'assuré de déléguer la gestion de son épargne par le biais d'un mandat de gestion donné à AXA France. Ses experts adaptent alors la répartition de l'épargne sur une large gamme de supports, en fonction des conseils et compte tenu des fluctuations et opportunités de marché. Quatre profils d'investissement sont possibles pour le contrat Arpèges par exemple :

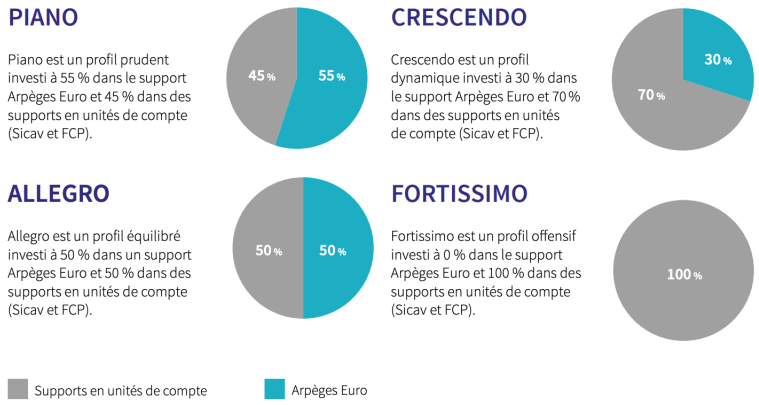

FIGURE 5 – Contrat Arpèges *(Source : Conditions Générales)*

- la **gestion évolutive** : permet à l'assuré de bénéficier des potentielles performances des marchés financiers tout en sécurisant progressivement son épargne. Celle-ci est répartie sur différents supports, en fonction de l'âge de l'assuré : de 30% sur le support en euros à 40 ans, la part euro de l'épargne augmente progressivement pour atteindre 85% à 70 ans. L'épargne est réajustée gratuitement et automatiquement chaque année, investie largement sur des supports de diversification en UC les premières années, puis progressivement orientée vers le support en euros.
  Voici un profil type d'évolution pour le contrat Arpèges :

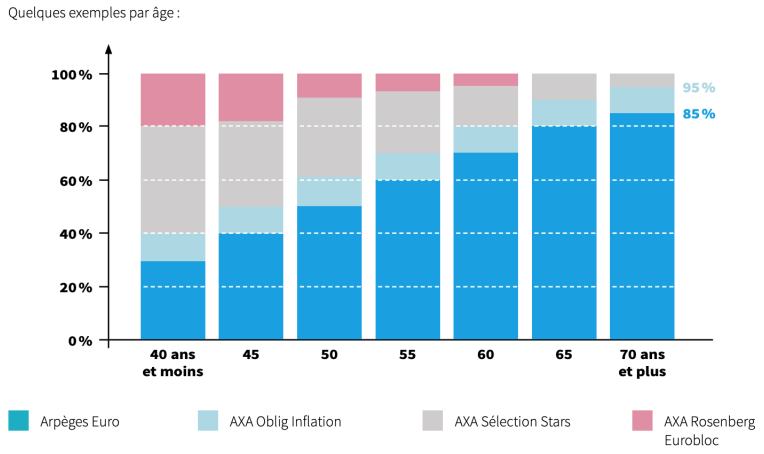

FIGURE 6 – Contrat Arpèges *(Source : Conditions Générales)*

- la **gestion personnelle** : permet à l'assuré de gérer en toute liberté la répartition de son épargne sur les différents supports d'investissement.

Le type de gestion informe donc sur la façon dont est géré un contrat par son souscripteur, sur l'évolution de la répartition de l'encours, et sur les types de flux permettant cette

répartition cible. Il apporte donc une information précieuse sur le souscripteur (potentiellement informé et actif dans le cas de la gestion personnelle, plutôt inactif dans le cas d'une gestion par convention ou sous mandat par exemple). Cette distinction entre types de gestion sera alors sûrement pertinente à retenir dans le cadre de la segmentation du portefeuille.

Tout au long de sa vie, un contrat d'assurance vie est soumis à certains **chargements**, retenus par la compagnie pour faire face à l'ensemble de ses frais. Ils se déclinent en plusieurs catégories :

- les **chargements de gestion** (aussi appelés frais sur encours) sont soustraits du montant de l'épargne constituée et financent le coût de la gestion et des placements ;
- les **chargements d'acquisition** sont prélevés lors des versements et réduisent le capital investi par l'assuré dans le but de couvrir le coût de commercialisation du contrat (y compris la rémunération des réseaux de distribution) ;
- les **chargements d'arbitrages** dans le cas des contrats multisupport sont prélevés sur l'épargne transférée ;
- les **autres frais** correspondent par exemple à des frais de gestion sous mandat, de garanties accessoires, de garantie plancher.

Le schéma suivant illustre ces différents types de frais appliqués à un contrat d'assurance vie type.

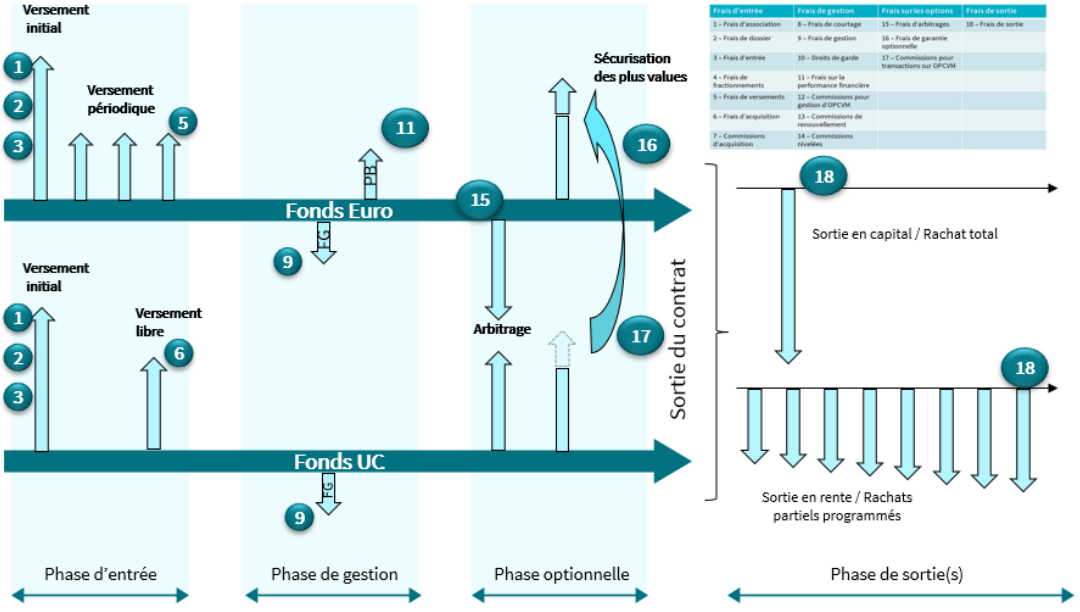

FIGURE 7 – Différents frais sur un contrat type d'assurance vie

Enfin, les produits d'assurance vie font l'objet d'impositions (prélèvements sociaux et fiscalité) selon des règles spécifiques, mais bénéficient surtout d'avantages fiscaux importants. La valeur de rachat (ou valeur acquise) d'un contrat d'assurance vie comprend les versements et les intérêts réalisés sur ces versements. Les prélèvements effectués sur cette valeur peuvent intervenir à différents moments de la vie du contrat : rachat(s) partiel(s) ou total(/totaux), échéance ou terme du contrat, ou encore décès par exemple.

Différentes lois régissent les principes de fiscalité s'appliquant aux produits d'assurance vie : loi de finances de 1998, de 2011, *flat tax* entrée en vigueur au 1er janvier 2018, par exemple.
Les règles de détermination des taux de prélèvement et d'imposition à appliquer aux contrats ne fait pas l'objet d'une modélisation particulière pour ce mémoire et ne seront donc pas détaillées ici. Il semble cependant important de noter que l'imposition dépend en grande partie de l'ancienneté du contrat
Ainsi, la fiscalité devient avantageuse pour l'épargnant à partir de la 4ème année de son contrat, et le devient encore plus après la 8ème année.
En conséquence, l'ancienneté du contrat permet souvent d'en expliquer la durée restante probable, et donc de prédire les comportements futurs du souscripteur. C'est pour ces raisons que la donnée de l'ancienneté devra être prise en compte dans les modélisations.

En résumé : Spécificités d'un contrat d'assurance vie

- Des options et garanties : garanties de taux (technique ou de revalorisation), planchers, options de PB ou de rachats entre autres ;
- Des particularités propres à chaque contrat, pouvant expliquer des différences de comportements : TMGA/TMGA, taux de PB, Bonus Euro+, mode de gestion par exemple ;
- Des frais à prendre en compte (de gestion ou d'acquisition notamment).

1.3 Quelques éléments de rentabilité et de valeur

Dans cette partie, des indicateurs de rentabilité ainsi que des éléments de valorisation sont décrits, qui sont utilisés dans le calcul de la valeur client.

L'*embedded value* est une **mesure de la valeur** de la compagnie, du point de vue de l'actionnaire. L'assurance vie connaît une problématique qui lui est propre : ses investissements se font sur le long terme, ce qui nécessite une modélisation prospective des contrats. Un paradoxe se dégage également : un assureur qui réalise beaucoup d'affaires nouvelles affiche souvent des résultats comptables inférieurs à ceux d'une compagnie en *run-off* (c'est-à-dire avec un arrêt de toute souscription d'affaire nouvelle, qui entraîne le déroulement, dans le temps, du stock des provisions techniques jusqu'à leur épuisement complet). En assurance, la valeur comptable n'est pas représentative de la valeur économique de la compagnie. L'*embedded value* est donc utilisée comme indicateur de performance par le marché, et pour le pilotage de la rentabilité du produit d'assurance au sein de l'entreprise.

Le *CFO forum* est un groupe de discussion créé en 2002 et composé des *Chief Financial Officiers* des principales compagnies d'assurance européennes. Son objectif est de progresser dans l'harmonisation des normes comptables et l'information financière, en accord avec l'IASB et les autorités de contrôle et de régulation des marchés. En mai 2004, il a publié les *EEV Principles* (*European Embedded Value*), qui ont fourni une base cohérente aux assureurs européens. Ils décrivent comment les entreprises doivent préparer leur rapport sur la valeur intrinsèque des performances de leurs opérations d'assurance vie. Ces principes ont été largement utilisés dans l'ensemble du secteur.

Depuis juin 2008, le *CFO Forum* fixe les règles de calcul de la MCEV au travers de 17 principes présentés dans les *MCEV principles & Guidance* [29] et les *MCEV basis for Conclusions* [30] (texte explicatif des principes).
Celle-ci répond à des principes de calculs bien spécifiques parmi lesquels :

- L'identification des affaires couvertes par le calcul ;
- Les hypothèses financières respectant l'approche *market consistent* ;
- Les modalités de prise en compte des affaires nouvelles ;
- Des paramètres de projections déterminés de façon *Best Estimate* ;
- Le format de publication.

La **MCEV** est la somme de deux éléments : la **NAV** (*Net Asset Value*) et la **VIF** (*Value of In-force*), comme l'illustre le schéma suivant.

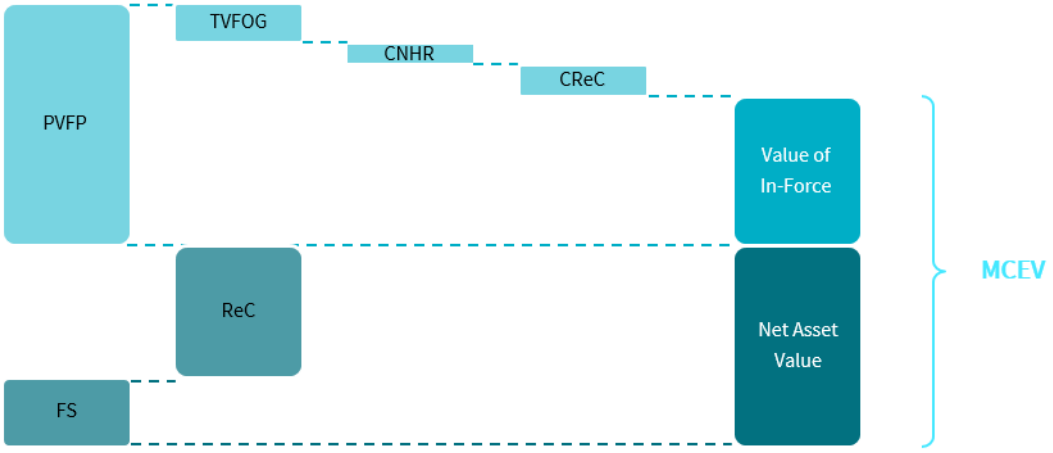

FIGURE 8 – Différents éléments de MCEV

La NAV et la VIF sont elles-mêmes constituées de plusieurs éléments : NAV = ReC + FS et VIF = PVFP − TVFOG − CNHR − CReC.

Le **ReC** (*Required Capital*) représente la valeur de marché des actifs détenus pour couvrir le portefeuille d'assurance selon les contraintes réglementaires.
Selon Solvabilité I, il s'agit en première approche de 4% des provisions mathématiques des

contrats libellés euros, 1% des provisions mathématiques des contrats libellés en UC et 0,3% des capitaux sous risques.
Sous Solvabilité II, la quantité de capital requise correspond à la perte de la valeur de la compagnie en cas de survenance du scénario de risque considéré avec une probabilité de $0,5\%$ à horizon 1 an. La formule standard Solvabilité II ou des modèles internes définissent les modalité de calcul du capital requis (SCR).

Le **FS** (*Free Surplus*) correspond au capital détenu au-delà du capital réglementaire. Il est égal à la valeur de marché des actifs alloués à l'activité couverte, au-delà du capital minimal et des réserves réglementaires liées au portefeuille constitué.

La **PVFP** (*Present Value of Future Profits*) est égale à la valeur actuelle des profits ou pertes futurs, nets d'impôts, générés par le portefeuille de contrats en cours. C'est la valeur actualisée des résultats futurs.

Considérons un contrat prenant fin à la date N. Soit R_k le résultat de l'année k sur le portefeuille observé et i le taux d'actualisation. La PVFP peut être calculée de la manière suivante :

$$\mathrm{PVFP} = \sum_{k=1}^{N} \frac{\mathrm{R}_k}{(1+i)^k}$$

Le taux d'actualisation représente le taux de retour attendu par l'investisseur. Deux approches sont alors possibles :

- L'approche ***Real World***, basée sur l'hypothèse que les actifs rapportent le taux sans risque et une prime de risque ;
- L'approche ***Market Consistent***, reposant sur le principe de valoriser les flux de l'activité d'assurance comme ils le seraient dans le cadre d'un instrument financier portant les mêmes risques et coté sur un marché financier.

Dans l'approche *Market Consistent* EV, le taux d'actualisation et le taux d'investissement sont supposés égaux à un même taux de référence.

La **TVFOG** (*Time Value of Financial Options and Guarantees*) est la valeur temps des options et garanties financières (arbitrages/rachats/versements, PB/Garantie plancher/TMG par exemple). Elle peut être décomposée en deux parties :

- La **valeur intrinsèque**, intégrée dans la modélisation déterministe ;
- La **valeur temps**, différence entre la valeur totale des options et garanties et la valeur intrinsèque. Elle est obtenue par différence de la PVFP déterministe intégrant la valeur intrinsèque des options et garanties, et la PVFP moyenne obtenue à partir des scénarios stochastiques.

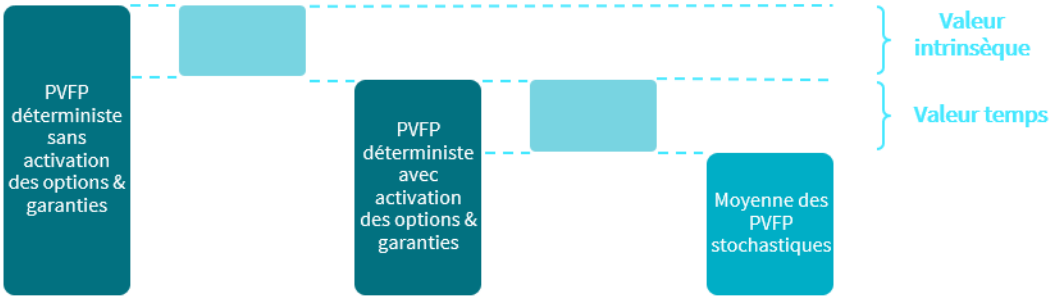

FIGURE 9 – Illustration de la PVFP

Une asymétrie des résultats existe selon les scénarios financiers considérés. En effet, lorsque le taux de rendement de l'actif est inférieur au taux de minimum garanti, l'assureur supporte entièrement les pertes et se doit de payer le différentiel de taux à l'assuré. Mais lorsque le taux de rendement de l'actif est supérieur au taux minimum garanti, l'assureur distribue au minimum le montant de participation aux bénéfices prévu dans chaque contrat. Il doit également s'assurer au niveau global de son actif général, qu'il redistribue au moins 85% de ses bénéfices financiers et 90% de ses bénéfices techniques.

En pratique, la valeur intrinsèque n'est pas calculée. Le coût des options et garanties est obtenu à partir d'une simulation de Monte-Carlo par différence entre la PVFP déterministe et la moyenne des PVFP stochastiques.
Cette approche permet de capter l'asymétrie des options et garanties.

Le **CNHR** (*Cost of Non-Hedgeable Risk*) modélise le coût des risques non couvrables, par exemple :

- Les risques non pris en compte dans la TVFOG, comme les risques opérationnels ;
- Le risque de modèle dans l'évaluation des scénarios *Best Estimate* ;
- Les risques de méthode d'extrapolation.

Il peut être tenu compte d'éventuels effets de diversification. Une méthodologie de calcul possible consiste à utiliser des sensibilités de choc aux risques étudiés (opérationnel, souscriptions, etc.).

Enfin, le **CReC** (*Cost of holding Required Capital*) correspond au coût d'immobilisation du capital requis, représentant le coût économique des investissements réalisés pour l'actionnaire.

Le coût d'immobilisation du capital est égal au capital initial investi diminué de la valeur actualisée des revenus de l'investissement du capital, et ajouté à celle de l'impôt sur les revenus de l'investissement du capital, des dotations/reprises du Capital, et des coûts de gestion de ce capital (coût de portage) :

$$\mathrm{CReC} = \mathrm{ReC}_0 + \sum_{i=0}^{N} \frac{-\mathrm{ReC}_i \times \mathrm{tx}_{inv} \times (1 - \mathrm{tx}_{tax}) + (\mathrm{ReC}_{i+1} - \mathrm{ReC}_i) + \mathrm{CI} \times (1 - \mathrm{tx}_{tax})}{(1 + \mathrm{tx}_{actu})^{(i+1)}}$$

avec CI le coût d'investissement, ReC_i le capital requis en i, tx_{inv} le taux d'investissement, tx_{tax} le taux d'imposition, et tx_{actu} le taux d'actualisation, défini comme le taux de rémunération attendu par l'actionnaire.

Dans l'approche *Market Consistent* EV, le taux d'actualisation et le taux d'investissement sont supposés égaux à un même taux de référence, et le CReC devient alors :

$$\text{CReC} = \sum_{i=0}^{N} \frac{\text{ReC}_i \times \text{tx}_{ref} \times \text{tx}_{tax} + \text{CI} \times (1 - \text{tx}_{tax})}{(1 + \text{tx}_{ref})^{(i+1)}}$$

Un autre élément de bilan important pour le calcul de la valeur épargne est la **MVM** (*Risk Margin*). Elle constitue, avec le *Best Estimate*, les provisions techniques inscrites au bilan en vision Solvabilité II. L'EIOPA [23] la définit comme le coût d'immobilisation d'un montant de fonds propres égal au SCR (*Solvency Capital Requirement*) nécessaire pour supporter les obligations d'assurance et de réassurance pendant toute leur durée. Le *Best Estimate*, lui, est défini comme la valeur actuelle probable des *cash flows* futurs, en tenant compte de la valeur temporelle de l'argent.

En résumé : Quelques éléments de rentabilité et de valeur

- L'*Embedded Value*, une mesure de la valeur de la compagnie ;
- Des indicateurs de rentabilité évalués sous certaines hypothèses, dont : VIF, MVM, TVFOG.

1.4 Assurance, actuariat et *Big Data*

À l'heure où les **données** sont qualifiées de nouvel "**or noir**", les acteurs du système assurantiel ont rapidement su se positionner et saisir cette révolution.
Habitués à utiliser des modèles de statistique classiques, ils ont commencé à en percevoir les limites. En effet, ces modèles font en général des hypothèses sur la distribution de probabilité et souffrent de faibles capacités prédictives. A contrario, le ***Machine Learning*** permet de ne faire aucune hypothèse sur la distribution, de modéliser des dépendances complexes, d'agréger plusieurs modèles plutôt que de n'en utiliser qu'un seul, et jouit de capacités prédictives particulièrement fortes.
Il convient cependant de noter que les grands principes sous-jacents ne sont en réalité pas vraiment nouveaux ; leurs bases avaient déjà été édifiées dans les fondamentaux de la statistique traditionnelle. La grande nouveauté aujourd'hui se retrouve en fait surtout dans les immenses quantités de données disponibles et exploitables, et dans les nouvelles puissances de calcul des machines. Il est en effet désormais possible à la fois de collecter des milliers voire des millions de données individuelles sur des milliers voire des millions de clients, et à la fois d'utiliser massivement des algorithmes d'apprentissage pour traiter ces informations.

Face à ces constats, et devant la **quantité** extraordinaire de **données** récoltées et disponibles dans le secteur assurantiel, la plupart des assureurs de la place ont pris le virage *Big Data* et initié des travaux dans ce domaine afin de mieux appréhender ses impacts métiers.
Les applications les plus fréquentes à ce jour concernent la **compréhension des comportements clients** (à des fins marketing et de gestion des risques), ou l'**adaptation des produits et des tarifs** (nouvelles techniques de tarification plus personnalisée, télématique par exemple). À ce titre, le *Big Data* semble introduire de profonds bouleversements à la fois dans l'assurance de personnes comme dans l'assurance dommages, et aussi bien pour les particuliers que pour les entreprises.

Bien que les applications du *Big Data* en assurance soient nombreuses et très prometteuses (révolution de la théorie du risque et de la science actuarielle, diminution de l'asymétrie de l'information, individualisation de la tarification, repoussement de la frontière de l'assurabilité), ces évolutions font face à de nombreux **défis** qui ne cessent de se renouveler.
Sur le plan **juridique**, elles devront suivre les évolutions réglementaires, avec notamment un grand sujet sur la protection des données (RGPD), mais également sur leur qualité avec l'application de Solvabilité II entre autres. Un équilibre pas toujours évident doit donc être trouvé entre pertinence, cohérence, exhaustivité, qualité et protection des données. Une opportunité semble certes être à saisir en matière de *Big Data*, mais prudence et sagesse doivent tout de même rester les mots d'ordre dans ce domaine.
Sur le plan des **produits**, des évolutions sont à envisager, en raison de lois nouvelles qui introduisent pour certaines de nouveaux produits (le PER avec la loi PACTE par exemple). Par ailleurs, en raison de leur complexité plus grande que celle des modèles de statistiques traditionnels, les modèles de *machine learning* sont souvent jugés trop compliqués et peu compréhensibles, et considérés comme des "boîtes noires". Un autre grand défi s'impose donc aux actuaires utilisant ce type de modèles : un devoir de **pédagogie**, de **communication** et de **clarté**, ainsi que de **compréhension** et d'**explicabilité** des modèles utilisés.

Enfin, il semble que ce nouveau champ d'analyse et ces nouveaux modèles ne doivent cependant **pas remplacer** les techniques actuarielles traditionnelles, mais plutôt être vues et utilisées comme des compléments permettant d'en améliorer sensiblement les performances. À cet effet il conviendrait peut-être d'appréhender le « *Big Data* » comme une évolution majeure de la science actuarielle, plutôt qu'une véritable révolution à part entière.

En résumé : Assurance, actuariat et *Big Data*

- Une évolution grâce à de nouvelles techniques toujours plus performantes ;
- Des outils traditionnels de l'actuaire pas abandonnés pour autant ;
- Une combinaison prometteuse de ces deux perspectives, associée à une grande quantité de données disponibles et exploitables ;
- Un secteur toutefois réglementé.

2 Objectifs de l'étude sur la valeur client épargne

Dans le contexte actuel de l'assurance vie, il devient nécessaire d'appréhender avec le plus de précision possible les **risques** auxquels la compagnie est soumise, afin de protéger sa rentabilité et de garantir sa solvabilité. Dans ce contexte, l'**indice client** semble être un baromètre pertinent, qui doit permettre d'orienter la **stratégie** de l'entreprise afin de retenir les clients créateurs de richesse, en prévenant le départ de ceux jugés comme profitables. Face à ce constat, il semble alors indispensable de **moderniser** et **personnaliser son calcul**, pour l'adapter à chaque profil, possédant ses propres caractéristiques.

L'**indice client**, décrit plus en détail dans les parties suivantes, est un **score** donné à chaque client d'AXA France, prenant en compte les **différentes valeurs** de ce client sur chacun de ses contrats souscrits dans la compagnie.
Dans cette partie du mémoire, le projet de refonte du modèle de valeur client épargne est exposé. Il complète les indicateurs actuariels classiques de rentabilité par des techniques de ***Machine Learning***, méthodes modernes et avancées de statistique sur de larges volumes de données afin de segmenter le portefeuille d'assurés et de déterminer au mieux la valeur associée à chaque client. Dans un second temps, des **méthodes d'analyse prédictive** sont employées, afin d'établir des lois de comportements pour chacune des opérations qu'un souscripteur peut effectuer sur son contrat, et qui impactent l'assureur (arbitrages / rachats / versements). Comme l'explique l'ouvrage de Patrick Thourot et Kossi Ametepe Folly [19], cet effort de modélisation prédictive remplace la logique d'extrapolation sur le futur des taux d'arbitrages/rachats/versements observés dans le passé.

Cette refonte permettra d'améliorer le modèle de valeur client épargne, ainsi que d'aboutir à un **ciblage** plus personnalisé des valeurs attribuées et à une **connaissance approfondie** des profils en portefeuille. À terme, une telle approche permettra sans doute d'atténuer le volume et la portée bilancielle des rachats notamment, grâce à une **prévention efficace** liée à la connaissance des motivations réelles des épargnants.[19]

2.1 Indice client

Afin de déterminer et de mieux **piloter la rentabilité** de leur portefeuille d'assurés, les directions métiers d'AXA France ont bâti ensemble un projet commun d'**indice client**. Un indice entre **1** et **3** est affecté à chaque détenteur d'au moins un contrat d'assurance AXA, représentant la "valeur" de l'assuré pour la compagnie, c'est-à-dire sa **rentabilité** (avec 3 pour un assuré très rentable, et 1 pour un assuré non rentable). L'indice client permet donc d'évaluer la **valeur économique globale** des clients détenteurs d'au moins un contrat.

Ce projet a pour objectif d'apporter aux **agents généraux** et aux **directions métiers** une vision plus précise de la rentabilité des assurés en portefeuille. Il doit permettre le déploiement ou l'adaptation de stratégies marketing dans le but d'**améliorer la rentabilité globale du portefeuille**.
A l'échelle de l'agent général, il doit servir à **optimiser les actions orientées clients** et

à mieux cibler les action commerciales telles que des avantages récompensant une bonne rentabilité. Il doit donc en quelque sorte permettre à l'agent de déterminer s'il est souhaitable de faire des efforts pour garder tel ou tel client en portefeuille ou non.
A l'échelle des directions métiers, cet indice doit permettre de **contrôler et surveiller la rentabilité** du portefeuille d'assurés, dans un contexte réglementaire strict et encadré.

Concrètement, l'indice global d'un client est déterminé par l'**agrégation** des valeurs individuelles de ce client sur chacun de ses contrats AXA France (auto, MRH, épargne, santé, prévoyance, etc.). La valeur d'un contrat est un montant en euros, représentant sa rentabilité prospective et pouvant être comparé à la valeur des autres contrats. Le périmètre de l'étude pour la direction technique épargne est donc la **détermination de la valeur client épargne**.

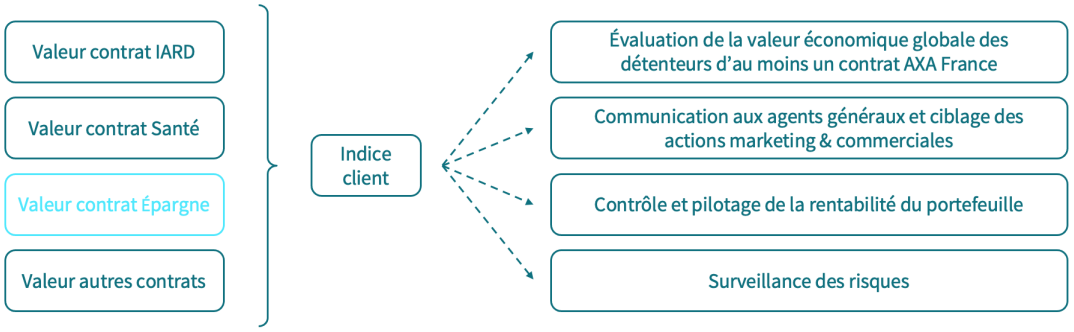

FIGURE 10 – Calcul et déploiement de l'indice client

La valeur client est généralement définie comme la **valeur actuelle** de tous les **revenus** futurs attendus d'un client sur sa durée totale de relation avec une entreprise, **diminués** des **coûts** associés.[9] Déterminée individuellement pour chaque client, elle permet de différencier ceux qui sont rentables des autres, plutôt que d'analyser une rentabilité moyenne globale. De plus, elle prend en compte la possibilité que le client interrompe la relation avec l'entreprise (rachat ou décès dans le cas de l'épargne/retraite).
Une première définition générale de la valeur client est donc [12],[3],[8] :

$$\mathrm{CLV} = \sum_{t=0}^{T} \frac{(\mathrm{r}_t - \mathrm{c}_t) \times \mathrm{p}_t}{(1+i)^t} - \mathrm{CA}$$

avec :

- CLV : valeur client (*Customer Lifetime Value*) ;
- r_t : revenus générés par un client en t ;
- c_t : coûts issus d'un client en t ;
- p_t : probabilité de présence du client en t ;
- i : taux d'actualisation ;
- CA : coûts d'acquisition.

Comme le rappellent Greg Firestone et Mohamad Hindawi dans un document de la *Casualty Actuarial Society* (CAS) [14], la valeur client est finalement plutôt un **cadre** qu'une mesure unique et de *nombreuses définitions* de valeur client existent, qui peuvent être utilisées suivant le domaine et les objectifs définis. Les auteurs notent également que le concept de valeur client diffère légèrement dans le domaine de l'assurance, qui possède certaines **particularités**. Parmi celles-ci, ils citent notamment l'allocation de capital requise, qui ajoute une certaine complexité dans les calculs. De plus, les revenus et coûts assurantiels varient par client et au fil du temps pour un même client : la valeur client est donc bien plus qu'une simple fonction de volume. Enfin, les auteurs notent que la structure de coûts très variable apporte une faible contribution aux marges, ce qui signifie que de petites variations de prix peuvent avoir un impact considérable sur la valeur client en assurance.

Dans ce cadre, la partie suivante explique la définition retenue par la Direction Technique d'AXA France, et la modélisation utilisée pour son estimation sur le périmètre de l'épargne/retraite individuelle.

En résumé : Indice client

- Un indice entre 1 et 3 reflétant la rentabilité prospective d'un assuré ;
- Une agrégation des valeurs individuelles sur les différentes branches ;
- Un objectif de pilotage des actions commerciales et stratégiques, et de contrôle de la rentabilité.

2.2 Modèle actuariel de la valeur client épargne

Dans l'ancien calcul de valeur client épargne, une valeur fixe en euros était affectée grâce à une grille de décision à chaque contrat, qui dépendait uniquement de l'encours du contrat et de sa part d'UC.
Cette vision ne reflétait pas correctement la réelle valeur économique des contrats, ni les différences propres à chacun, et a donc nécessité d'être remplacée par un modèle plus complexe.

En effet, la part des indicés 1 (clients non rentables) chez les détenteurs Épargne restait très élevée par rapport aux non détenteurs Épargne (55% VS 3%). Cela risquait de dissuader l'équipement en épargne des clients : la souscription d'un contrat épargne faisait baisser très fortement l'indice client global des individus en portefeuille, même ceux ayant un encours conséquent. De même, les clients effectuant des rachats (même faibles par rapport au volume très élevé d'encours restant sur le contrat) avaient toujours une valeur négative sur certains segments.
D'autres besoins ont émergé des discussions sur la refonte du modèle de valeur client épargne. Il a alors semblé nécessaire d'utiliser une méthodologie de segmentation du portefeuille, pouvant être expliquée aux agents et basée sur un grand nombre de variables influençant les comportements des épargnants.

En adaptant la définition générale, la formule suivante a donc été retenue pour estimer la valeur client sur le périmètre de l'épargne/retraite individuelle :

$$\begin{array}{c}\text{Valeur} \\ \text{épargne}\end{array} = (\text{PM } € \times \text{Coeff. } € + \text{PM UC} \times \text{Coeff. UC}) \times \begin{array}{c}\text{Duration} \\ \text{contrat}\end{array} - \begin{array}{c}\text{Frais} \\ \text{d'acquisition}\end{array}$$

avec :

$$\text{Coeff.} = \frac{\text{VIF}_{stoch} - \text{TVFOG} - \text{MVM} - \text{VAN FG}}{\text{PM} \times \text{Duration}}$$

La **PM €** représente l'encours détenu par l'assuré sur le fonds euro, et la **PM UC** celui sur des UC.

La **TFVOG** (*Time Value of Financial Options and Guarantees*) mesure la valeur temps des options et garanties financières.

La **MVM** (*Market Value Margin*) désigne la marge pour risque et représente le coût d'immobilisation des fonds propres réglementaires.

L'estimation des **coefficients** fait également intervenir la valeur actuelle nette des frais de gestion (**VAN FG**).

Les **frais d'acquisition** sont calculés pour chaque contrat via un forfait de frais d'acquisition. Comme la partie 5 l'explique plus précisément, les frais d'acquisition sur contrats épargne devront être **lissés** sur plusieurs années, en fonction du produit considéré, afin d'éviter que la valeur épargne ne soit très négative pour la première année de chaque contrat, et qu'elle ne dégrade donc trop l'indice client global.

La **duration** des contrats peut être interprétée comme une **espérance de vie résiduelle de l'assuré en portefeuille**. Elle représente la durée moyenne qu'il reste à chaque contrat avant sa sortie du portefeuille, qui peut avoir deux causes possibles : rachat total du contrat ou décès (l'hypothèse est faite à ce stade que le contrat n'a pas de date d'échéance prédéterminée). La méthodologie de calcul de la duration contrat sera approfondie en partie 5.

La **VIF** (*Value of In-Force*) représente la valeur actuelle des profits futurs issus du portefeuille acquis. Dans la formule du coefficient, la VIF stochastique est utilisée. Les principes techniques ont été exposés en première partie, et le calcul de la VIF se base sur des hypothèses utilisées par le modèle interne. Ainsi par exemple, la référence retenue pour le taux d'actualisation de la VIF est la courbe des taux risque neutre transmise par la direction des investissements. Le calcul de la VIF sera davantage approfondi dans la partie 5, qui détaille le calcul de la valeur client épargne.

La VIF est surtout déterminée par le **comportement futur des assurés** et fait donc appel à des **lois de comportement** en épargne (arbitrages, rachats et versements).
Ces lois correspondent à des suites de taux du flux concerné, exprimé en pourcentage

de la PM correspondante. Ainsi par exemple, une loi de comportement sur les rachats partiels donne, pour tout $t \in \{1, ..., 60\}$, le taux de rachats partiels de l'année t : $\frac{\text{RP}_t}{\text{PM}_{t-1}}$ (la méthodologie de construction de ces lois sera expliquée plus en détail dans la partie 5, dédiée au calcul de la valeur épargne).

Déterminées à la maille produit, ces lois ne permettent pas de **différencier** les **assurés** d'un même produit selon leur **comportement**. En d'autres termes, deux individus du même produit mais au comportement totalement différent (donc au potentiel de rentabilité également différent) pourraient se voir attribuer les mêmes lois, donc la même VIF.

Le but du mémoire est donc de **segmenter le portefeuille d'assurés** en classes de comportement, afin de leur affecter des lois différentes, ce qui impactera leur VIF respective, et donc leurs coefficients Euro et UC. À terme, cet affinement de la maille de calcul de la VIF et des coefficients doit permettre d'attribuer des valeurs différentes à des clients appartenant à des classes de comportement différentes, même s'ils se trouvent sur le même produit.

En résumé : Modèle actuariel de la valeur client épargne

- Une définition propre à la Direction Technique ;
- Des indicateurs actuariels de rentabilité pris en compte (VIF, TVFOG et MVM notamment) ;
- Une VIF influencée par les lois de comportements, qui doivent refléter les différences de caractéristiques entre assurés : une segmentation du portefeuille est donc nécessaire.

2.3 Travaux et apport du mémoire

Dans l'ancien modèle de la valeur client épargne, celle-ci dépendait uniquement de l'encours et de la part UC du contrat, et pouvait donc être identique pour deux clients aux caractéristiques et comportements très différents, et dont les perspectives de rentabilité pour AXA France étaient également très éloignées. Une première étape a donc consisté à utiliser la formule ci-dessus, qui devait permettre d'éviter ces écueils.
Dans une première tentative de modélisation par cette formule, les calculs de VIF et duration étaient effectués à une maille grossière (réseau de distribution / groupe de produits). En d'autres termes, deux épargnants ayant souscrit un contrat sur le même groupe de produits, via le même réseau de distribution, se voyaient attribuer la même valeur client. Ces éléments du calcul ne reflétaient donc toujours pas bien les **différences de comportements** entre assurés puisqu'ils ne prenaient pas en compte les caractéristiques clients et se basaient uniquement sur les comportements passés des assurés, observés à une maille assez large.
Une **nouvelle maille de segmentation** était donc essentielle afin de différencier les clients au sein d'un même réseau / groupe de produits. Il était également nécessaire de prendre en compte **tous les réseaux** de distribution, ainsi que de **nouvelles variables**

pour mieux expliquer les comportements des assurés.

L'**enjeu** et l'**apport** de ce mémoire sont donc d'**améliorer le modèle de la valeur client épargne**, en :

- Changeant de **méthodologie** (utilisation de la formule présentée précédemment à la place de la grille de valeurs en fonction de l'encours et de la part UC) ;

- Affinant la maille de détermination des différents éléments de la formule de calcul de la valeur épargne, tels que les coefficients, la VIF et la duration, grâce à la **segmentation du comportement client en assurance vie** (pour les trois types de flux principaux : arbitrages, rachats et versements).

Cette approche mêle des **techniques statistiques** poussées sur des **grands volumes de données** et intègre des **indicateurs actuariels** de rentabilité à destination des réseaux de distribution et agents généraux particulièrement.
Cet objectif d'amélioration a par ailleurs nécessité de commencer par retravailler en profondeur la base de données utilisée, notamment en l'alimentant de nouvelles **données internes** comme **externes**, ainsi qu'en veillant à la **qualité** des données utilisées.
Une première version d'extraction de données et de segmentation du portefeuille a été réalisée fin 2019 avec Rafaël Hisquin, mon tuteur en entreprise, dans le cadre de son mémoire pour la formation de *"Data Science pour l'actuariat"* dispensée par l'Institut des Actuaires. Le présent mémoire s'inscrit donc dans la continuité des travaux déjà initiés, en enrichissant considérablement la base de données utilisée et en la retraitant en profondeur, en lui appliquant la première segmentation déjà imaginée mais réadaptée, puis en imaginant une **autre méthode de segmentation** qui sera comparée à la première.

Comme expliqué plus tard dans ce mémoire, différentes perspectives de déploiement de l'indice client sont envisageables :

- sur le plan **marketing** : communication aux agents généraux et ciblage des actions commerciales tels que les "appels câlins" (à destination des clients rentables pour s'assurer de leur satisfaction) ou encore l'octroi d'avantages commerciaux par exemple ;

- sur le plan de la **gestion du risque** : prévention et gestion des risques de rachats ou d'arbitrages massifs de l'UC vers l'euro notamment, grâce à des facteurs explicatifs des comportements bien identifiés et à la modélisation prédictive.

En résumé, les travaux de ce mémoire doivent permettre d'**améliorer le calcul de la valeur client épargne**, en segmentant le portefeuille d'assurés selon leur caractéristiques comportementales afin de leur attribuer une valeur finale plus proche de leurs spécificités et plus personnalisée.

En résumé : Travaux et apports du mémoire

- Une ancienne grille d'affectation arbitraire et insatisfaisante, en fonction de l'encours et de la part UC ;
- Une première tentative de modélisation, selon une maille grossière, jugée insuffisante ;
- Une amélioration du modèle de valeur client épargne grâce à la segmentation du portefeuille selon le comportement des assurés.

2.4 Données et segmentation du portefeuille

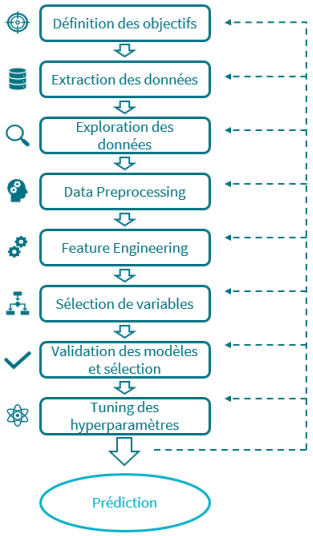

FIGURE 11 – Étapes d'un projet de *Machine Learning*

La première étape du projet a consisté à bien **identifier les besoins et objectifs**, afin de déterminer les travaux nécessaires ainsi que d'établir une feuille de route. Une fois cette étape réalisée, un important travail sur les données a alors pu commencer.

Ayant identifié la problématique, il a alors fallu **construire la base de données** qui devait ensuite servir à la segmentation du portefeuille.
Comme décrit dans la partie suivante, cette étape a notamment été l'objet de phases de compréhension des données à disposition, d'extractions de **données internes** de différents types ainsi que de **données externes** (de l'INSEE). Puis ces données ont été largement explorées, étudiées, retraitées quand nécessaire, puis préparées pour les travaux de segmentation.

Afin d'affiner la maille de calcul des coefficients déterminant la valeur client, une étape cruciale a été incorporée au modèle : la **segmentation du portefeuille d'assurés** en

fonction de leurs **caractéristiques comportementales**. L'objectif est d'avoir, au sein de chaque Réseau/groupe de produits, différentes classes dont le comportement (pour chacun des flux) est différent et pour lesquelles des lois de comportement seront calculées, qui leur seront propres.
Le projet consiste donc à rapprocher des assurés qui ont des lois de comportement similaires, mais la faible profondeur d'historique ne le permet pas immédiatement. Deux méthodes de segmentation ont alors été testées, qui seront expliquées en détail par la suite.

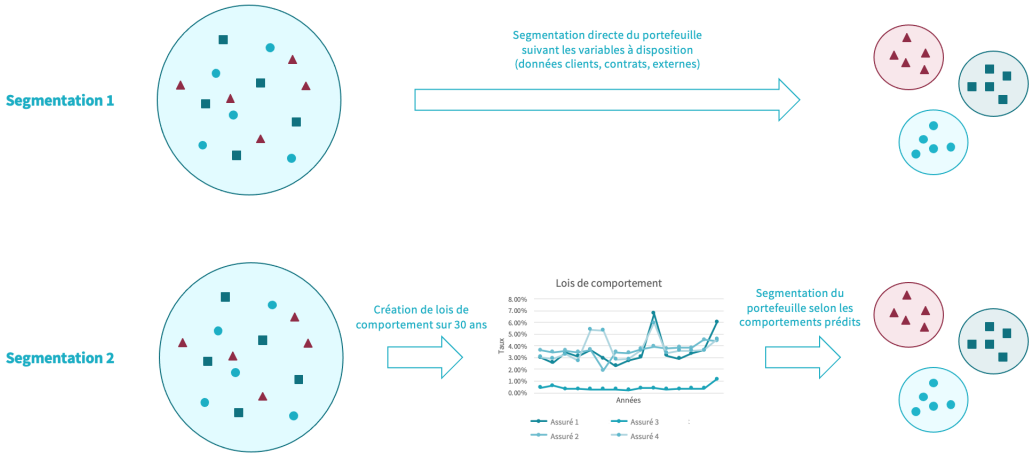

FIGURE 12 – Illustration des deux méthodes de segmentation pour un type de flux

Dans un premier temps, des assurés sont rapprochés s'ils se ressemblent dans leur comportement vis-à-vis du flux étudié.
Ainsi, la **première méthode** a consisté à affecter à chaque assuré de la base une **classe de comportement** pour chaque type de flux (après sélection de variables, entraînement et calibrage des modèles). Cette méthode de segmentation se base sur les **variables à disposition**, telles que les caractéristiques du contrat, de l'assuré, de l'historique de ses flux, comme détaillé par la suite. L'objectif de cette première méthode consiste à rapprocher des clients qui se ressemblent en terme de comportement, selon ces variables. À partir de cette base de clients classifiés, les lois de comportements sur 60 ans ont été établies, et les différents éléments de la valeur client ont ensuite pu être calculés.

La **deuxième** méthode de segmentation a utilisé une perspective quelque peu différente, se basant davantage sur des modèles prédictifs. Cette autre solution a consisté à reproduire les lois de comportement sur 30 ans (qui étaient indisponibles, faute de profondeur d'historique suffisante), pour rapprocher les assurés par rapport à leur loi.
À partir de la base client brute, contenant l'ensemble des variables explicatives décrites par la suite dans la partie dédiée, des modèles d'apprentissage ont été entraînés sur les montants (d'année $n+1$) de chaque flux. Ces modèles ont ensuite permis de prédire pas à pas, à horizon 1 an, les flux de chaque assuré pour chaque type de comportement, et

donc de **déterminer des lois de comportement** en itérant ce processus sur 30 ans. Ce n'est qu'une fois ces lois établies que la **segmentation du portefeuille** a été effectuée. Alors, pour chaque type de flux, une classe a été affectée à chaque assuré, permettant de le classer avec les autres assurés ayant une loi (donc un comportement prédit) proche pour le type de flux considéré. Une fois que ces lois et classes ont été déterminées, le calcul de la valeur client de chaque assuré a pu être effectué.

Enfin, il convient de noter qu'une étape a été nécessaire en amont de celle de segmentation. En effet, la base de données considérée était conséquente, tant en nombre d'observations qu'en nombre de variables disponibles. Toutes les observations devaient être conservées pour la segmentation puisque chaque client doit être associé à un *cluster* et finalement se voir attribuer une valeur, mais il était nécessaire de réduire le nombre de variables utilisées pour ne pas tomber dans ce que les spécialistes du domaine nomment le "fléau de la dimension". Pour ce faire, une étape de **sélection de variables** en amont de la segmentation s'est imposée. Comme les *clusters* finaux doivent séparer au mieux les clients selon leurs comportements, le choix a été fait de ne retenir que les variables les plus pertinentes de ce point de vue. Aussi, la sélection a été effectuée par une étape préliminaire consistant à prédire la survenance d'un flux (arbitrage/rachat/versement) sur l'année étudiée (apprentissage supervisé). Après entraînement et validation du modèle de prédiction, seules les variables qui avaient été les plus importantes dans la prédiction de la survenance du flux ont été retenues pour la segmentation. Leur importance a également été gardée en mémoire, afin de refléter l'apport relatif de chaque variable dans les modèles de segmentation.

En résumé : Données et segmentation du portefeuille

- Une identification des besoins et objectifs ;
- Un travail significatif réalisé sur les données (extraction, exploration, préparation) ;
- Une segmentation du portefeuille d'assurés réalisée en plusieurs étapes et selon deux méthodes différentes (sur la base brute d'une part, et sur les lois de comportements prédites, d'autre part), avant le calcul des valeurs finales.

2.5 Déroulé du projet

Différents acteurs interviennent dans le processus de détermination de la valeur client épargne :

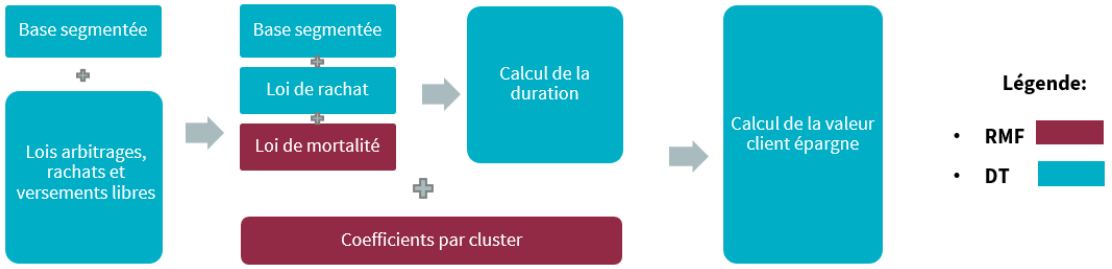

FIGURE 13 – Les différents acteurs du projet

Les travaux réalisés dans le cadre de ce mémoire consistent, au sein de la DT, à **segmenter la base client** en plusieurs classes de comportements ("*clusters*") et de **déterminer les lois** correspondantes. À partir de ces éléments et d'une loi de mortalité appliquée au portefeuille (loi d'expérience), il est ensuite possible de calculer la **duration** et les **coefficients** par *cluster*. Après intégration des frais d'acquisition, une valeur est alors calculée et affectée à chaque assuré.

Ce projet a suscité de nombreux échanges avec d'autres directions et services sur les travaux et contrôles effectués par chacun, avec une nécessité de les connaître et de les comprendre afin de ne pas laisser de zone d'ombre et de ne pas utiliser de "boîtes noires" provenant d'autres équipes lors des travaux. À cet effet, différents comités (de suivi, de pilotage, de gouvernance, etc.) ont régulièrement été planifiés afin de superviser les travaux et de se tenir à jour sur leur évolution.

En résumé : Déroulé du projet

- Un projet en plusieurs étapes, qui mêle des acteurs de différentes directions ;
- Des réunions et comités de pilotage pour répondre au mieux aux besoins et s'assurer de la mise à disposition des éléments.

3 Présentation des données de l'étude

La refonte du modèle de valeur client épargne a pour ambition de **moderniser** le calcul par contrat de la valeur économique des contrats épargne et retraite en intégrant à la fois des spécificités commerciales et des données clients révélatrices des potentiels d'arbitrages, de rachats et de versements.

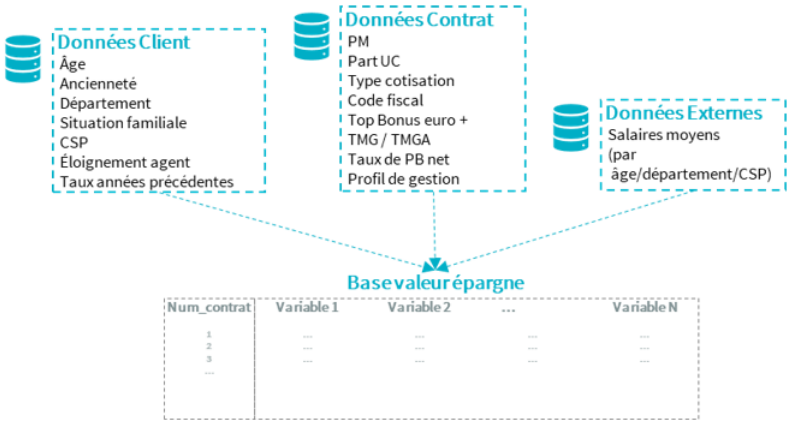

FIGURE 14 – Collecte des données

Le projet a donc nécessité d'**explorer** l'ensemble des données pouvant être utilisées à cette fin et de rechercher les informations pertinentes pour la modélisation, qui peuvent être regroupées en trois catégories :

- des caractéristiques sur les **clients** ;
- des informations sur les **contrats** ;
- des données **externes** (INSEE).

3.1 Extraction

Les données utilisées pour l'étude proviennent de différentes sources selon leur catégorie. La plus grosse source utilisée correspond aux **bases infocentre d'AXA France** (bases contrats et clients de la direction technique épargne). Provenant de l'ensemble des réseaux propriétaires de distribution des contrats d'épargne, elles contiennent toutes les informations de flux et de stocks sur les clients actuellement en portefeuille comme de ceux qui en sont sortis.

Les données contrat, dispersées dans une multitude de tables différentes (stock et flux, par année) et présentant des formats et caractéristiques propres ont dues être fortement retraitées et harmonisées en amont de l'étude et du travail de segmentation, afin d'obtenir une unique base contrat regroupant les informations de stock et de flux. Cette étape du projet a été une des plus chronophages.

Les données client proviennent des bases des équipes du **CRM** (*Customer Relationship Management*) d'AXA France, et ont été validées lors d'un comité afin de pouvoir être utilisées par la direction technique épargne.

De ces bases, plusieurs variables utilisées dans les modélisations ont donc été extraites :

- **Numctv** : numéro du contrat ;
- **Réseau** : réseau de distribution du contrat ;
- **Code produit** : code rattaché au contrat identifiant le produit d'épargne ou de retraite dont il est question ;
- **Code portefeuille** : code rattaché au contrat identifiant l'agence dans laquelle le contrat a été souscrit ;
- **Top SA** : indicatrice qui précise le statut juridique de l'agence commercialisant le contrat ;
- **Durée courue** : ancienneté du contrat ;
- **Âge** : âge de l'assuré au moment de l'extraction ;
- **CSP** : catégorie socio-professionnelle de l'assuré au moment de l'extraction ;
- **Département agent** : département de l'agent en charge du contrat ;
- **Département assuré** : département de résidence de l'assuré ;
- **Gestion** : type de gestion du contrat ;
- **Taux garanti moyen** : taux moyen garanti sur le contrat ;
- **Code fiscal** : code fiscal du contrat (par exemple, "capi"/"perp"/"madelin") ;
- **Situation familiale** : situation familiale de l'assuré ;
- **PM extract**, **PM euro extract** et **PM uc extract** : encours respectivement total, en euros, et en UC de l'année correspondant à l'extraction (pour la plupart des contrats, PM de décembre 2019) ;
- **PM histo 1an**, **PM euro histo 1an** et **PM uc histo 1an** : encours respectivement total, en euros et en UC, de l'année précédant les flux étudiés (pour la plupart des contrats, PM de décembre 2018) ;
- **Flux Euro** et **Flux UC** : montants de flux réalisés respectivement sur fonds Euro et sur UC, pour chaque type de flux (arbitrages, rachats et versements) et sur plusieurs années d'historique ;

Puisque l'étude se concentre sur la modélisation du comportement des assurés, il convient d'ajouter des variables qui peuvent fortement impacter et donc expliquer leurs décisions, afin de mieux les cerner et les prédire. C'est le cas notamment du **taux de TMG ou TMGA** appliqué au contrat pour la revalorisation de l'encours sur l'année concernée, ainsi que du **taux de PB** distribué en fin d'année précédant l'étude et reversé sur l'encours de chaque contrat. Ces deux variables donnent une indication forte sur la rémunération des contrats d'assurance vie et peuvent donc directement influencer les comportements des assurés. Des taux garantis ou de PB trop faibles réduisent la rémunération d'un contrat et donc la satisfaction de son détenteur, ce qui peut conduire à une baisse des versements, voire un rachat partiel ou total. Au contraire, des taux garantis ou de PB élevés augmentent l'encours et donc a priori la satisfaction de l'épargnant, qui sera plus enclin à effectuer les trois types de flux.

Ces deux informations ont été extraites des rapports de décisions de taux de TMGA et de PB des années précédentes, ainsi que des conditions générales des contrats pour la distinction entre TMG et TMGA. Disponibles à la maille produit, elles ont ensuite été agrégées à la base de données grâce au code produit de chaque contrat.
Pour la variable "Info TMG/TMGA", il a été décidé de transformer la donnée du taux et de la distinction entre TMG et TMGA en une unique variable exploitable par la suite, avec les modalités suivantes :

- **"TMG pos"** : si un TMG strictement supérieur à 0 est garanti contractuellement ;
- **"TMGA pos"** : si un TMGA strictement supérieur à 0 est garanti contractuellement pour l'année considérée ;
- **"TMG 0"** : si aucun TMG n'est garanti contractuellement ;
- **"TMGA 0"** : si le TMGA est à 0% pour l'année considérée ;
- **"TMG part"** : si différentes garanties sont proposées sur les multiples supports rattachés au contrat et qu'au moins 10% de l'encours présente une garantie de TMG.

En parallèle, des données externes pouvant aider à la modélisation ont été importées. En particulier, dans un objectif d'utiliser le plus de variables quantitatives possible dans les modèles de machine learning, la variable qualitative de la CSP a été remplacée par un **salaire moyen**, extrait des bases de données de l'INSEE.
Dans une étude de l'INSEE de 2016 (la plus récente à ce jour), des indicateurs de salaires nets horaires moyens sont disponibles en libre accès. Ventilés selon le sexe, l'âge et la catégorie socio-professionnelle (CSP), ces données peuvent être récoltées à différentes mailles. Dans le cadre de la construction de la base de données clients, les départements de l'assuré sont disponibles sous forme brute et l'information du salaire moyen a donc été extraite à la maille département. Ainsi, pour chaque assuré, un salaire moyen a pu être déterminé en fonction de sa tranche d'âge, sa catégorie socio-professionnelle, et son département de résidence. Pour tous les assurés dont une de ces variables viendrait à manquer, le salaire moyen selon les autres variables est affecté (par exemple, un assuré d'un certain âge et d'une certaine CSP mais dont le département est inconnu se verra attribuer le salaire moyen des travailleurs de même âge et même CSP, sur la France entière). Si l'information est indisponible, une estimation à partir de la PM du contrat a été effectuée. La construction d'un tel salaire moyen a finalement permis de croiser différentes variables en une seule (CSP/tranche d'âge/département/PM).

La base finale contient près de 2,2 millions de lignes (avec une ligne par contrat), ce qui représente un grand volume de données et répond bien aux critères du *Big Data*. Les techniques et algorithmes utilisés par la suite semblent donc appropriés, avec cependant un temps d'exécution qui sera sans doute très long, à cause de cette grande dimension.

En résumé : Extraction des données

- Des données contrats, des données clients, des donnés externes (INSEE) ;
- Des données soumises à vérification et validées ;
- Une grande dimension : de nombreuses observations et variables.

3.2 Qualité des données

Avant de commencer les travaux sur la base de données, il a été nécessaire de vérifier l'exactitude des donnés extraites. En effet, la directive de Solvabilité II, entrée en application en janvier 2016, a introduit des exigences fortes en ce qui concerne la qualité des données. Celles-ci sont résumées dans la figure suivante :

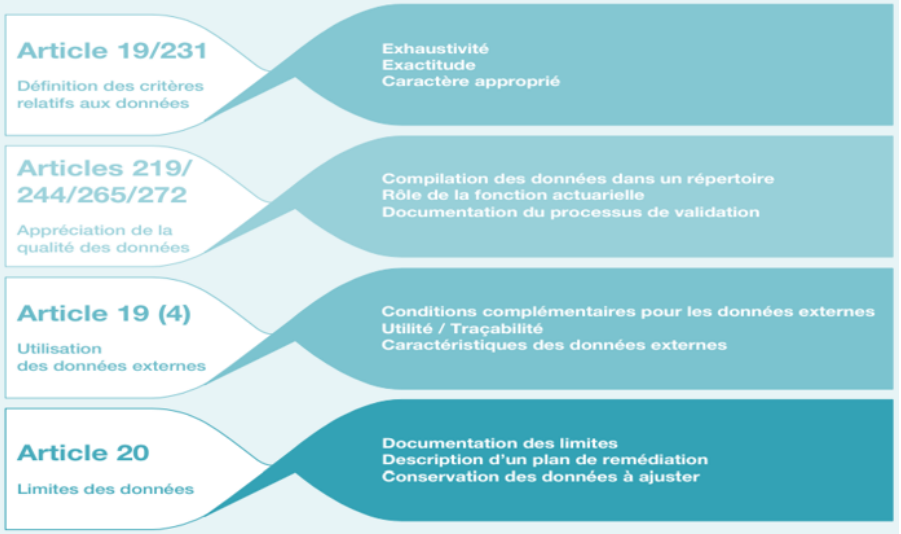

FIGURE 15 – Exigences réglementaires en matière de qualité des données dans le cadre de Solvabilité II (Source : étude PWC)

Afin de respecter ces critères et la législation en vigueur, un important travail de vérification des données, et notamment de leur exactitude, a donc été mené.

Pour ce faire, une comparaison des montants de PM et de flux extraits des bases infocentre (source brute) et retraités a été effectuée avec les données provenant de sources comptables. En ce qui concerne les montants de PM, des écarts très faibles avec la source comptable ont été notés, et considérés comme supportables et explicables (écarts en général de l'ordre de 0,3%).
De même, les montants de flux d'arbitrages, de rachats et de versements extraits ont été validés après comparaison avec la source comptable, et car leurs écarts restaient de nouveau contenus et rationalisables (de l'ordre de 1% à 2% en moyenne).

Par conséquent, les différents montants extraits pour l'étude sur la valeur épargne se sont révélés cohérents et ont donc été validés pour la suite de l'étude.

Par ailleurs, les autres informations extraites et provenant du CRM avaient été validées lors d'un précédent comité et ont donc été jugées comme valides et exploitables, après quelques vérifications.

Enfin, comme il sera expliqué dans la partie correspondante du mémoire, de nombreuses vérifications des lois de comportements ont été effectuées. Les écarts de méthodologie et de montants avec celles utilisées dans le Modèle Interne ont été analysés et rationalisés, et ont permis de conclure à l'exactitude des lois de comportements pour la suite de l'étude.

En résumé : Qualité des données

- De nombreuses vérifications de cohérence et d'exactitude effectuées ;
- Des écarts entre montants de PM et de flux extraits et source comptable, faibles et rationalisés ;
- Des comparaisons entre les lois de comportement calculées et celles utilisées dans le Modèle Interne.

3.3 Retraitements

Une autre étape importante du travail des donnés a été leur **retraitement**, qui constitue une phase de *Data preprocessing*. Lors de cette étape, les données brutes sont retravaillées afin d'être adaptées aux algorithmes qui seront utilisés par la suite. Aussi, de nombreuses variables ne présentaient pas les mêmes modalités selon les types de gestion et nécessitaient donc d'être retravaillées lors d'une phase d'harmonisation, afin d'obtenir des variables comparables et utilisables.
Enfin, une phase de nettoyage des données et de retraitement des valeurs aberrantes, particulièrement chronophage, a été réalisée afin de préparer les données à être exploitées.

3.3.1 Données manquantes

Le premier retraitement effectué concerne les **valeurs manquantes**. Parmi les variables extraites exposées précédemment, certaines présentaient un fort taux de valeurs manquantes, comme le montre le tableau ci-dessous.

Code Fiscal	Situation familiale	Taux garanti moyen	CSP	Département agent	TMG/TMGA	Taux PB	Région agent	Type de gestion
12%	22%	49%	21%	15%	40%	40%	15%	23%

FIGURE 16 – Parts de valeurs manquantes

Ces données peuvent être manquantes pour différentes raisons : parce qu'elles n'ont pas été renseignées par l'agent lors de la souscription du contrat (par oubli ou car les questionnaires n'existaient pas encore à l'époque), à cause d'une défaillance des systèmes d'information, parce qu'elles ont été perdues (mauvais encodage ou erreur de conversion des données par exemple), etc.
Bien souvent, ces informations manquantes sont importantes pour les modélisations et leur absence est donc problématique car elle peut biaiser les analyses.
Une première approche souvent utilisée lors de travaux sur des bases de données incomplètes consiste à ne conserver que les lignes de la base qui ne possèdent aucune valeur manquante. Cette méthode n'est cependant pas adaptée à l'étude sur la valeur client, dans laquelle chaque assuré doit se voir attribuer une valeur à la fin de l'étude et ne peut donc pas être supprimé de la base de données initiale. Une analyse plus approfondie des valeurs manquantes et de leurs retraitements est donc nécessaire.

Comme expliqué dans le manuel de Data Science d'Eric Biernat et Michel Lutz [2], s'appuyant sur la typologie proposée par Little et Rubin [16], il existe trois grandes catégories de données manquantes :

- les **données manquantes complètement aléatoires**, dites *MCAR* (*Missing Completely At Random*) : dans ce cas, l'absence de renseignement de la variable est considéré comme un pur hasard, qui n'a pas de réelle explication ni signification. La probabilité qu'une valeur d'une variable soit manquante ne dépend pas des valeurs prises par les autres variables, manquantes ou non. Il est alors impossible de tirer une conclusion sur les individus dont les données sont manquantes.

- les **données manquantes aléatoires**, dites *MAR* (*Missing At Random*) : dans ce cas, l'absence de donnée pour une variable ne dépend pas des valeurs prises par les autres variables manquantes, mais de leurs valeurs observées.

- les **données manquantes non aléatoires**, dites *MNAR* (*Missing Not At Random*) : dans ce dernier cas, l'absence de valeur pour une variable est considérée comme informative, la donnée est manquante pour une raison précise voulue. La probabilité qu'une variable soit manquante dépend alors des valeurs manquantes des autres variables observées.

Dans un premier temps, des **statistiques descriptives** sur les données manquantes ont été établies, afin de déterminer s'il existait des prédicteurs potentiels de données manquantes (par exemple si elles provenaient d'un mauvais renseignement de la part d'un réseau de distribution particulier, ou dans une région particulière).
Il en est ressorti que, pour toutes les variables présentant des valeurs manquantes, ces dernières se situent souvent sur les contrats dont l'agent et/ou le souscripteur viennent d'Île-de-France, avec un âge et une ancienneté souvent avancés. Cependant, ces conclusions sont sûrement biaisées par la sur-représentation des contrats d'Île-de-France, d'âge et d'ancienneté élevés parmi l'ensemble des contrats de la base, et aucun prédicteur de valeurs manquantes n'a pu être identifié.

La principale conclusion à ce stade sur les données manquantes est donc que l'**absence** de valeur pour certaines variables peut être **informative** pour certaines variables, et il a donc été jugé pertinent de conserver cette information. Pour ce faire, une classe dédiée **"inconnue"** a été attribuée aux valeurs manquantes de plusieurs variables : code fiscal, CSP, situation familiale, départements agent et souscripteur, type de gestion.

Pour d'autres variables cependant, l'absence de valeur a été jugée comme non informative et non impactante pour les modélisations, et a donc été remplacée par une valeur imputée selon une **règle métier**. Ainsi, dans le cas où l'information TMG/TMGA n'était pas renseignée, il a été établi qu'aucun taux n'était garanti sur ce type de contrat, et la valeur "TMGA_0" a donc été choisie. De même, pour les taux de PB non renseignés, la moyenne arithmétique du taux servi sur le Réseau-Produit a remplacé les données manquantes.

Enfin, la variable "Taux garanti moyen" présentant une trop grande part de valeurs manquantes, il a été décidé de ne pas la prendre en compte dans les modélisations, au profit de l'information TMG/TMGA et du taux de PB.

D'autres méthodes d'imputation de données existent, exposées dans le manuel de Data Science [2] (comme la régression ou les plus proches voisins), mais n'ont pas été utilisées dans cette partie.

En résumé : Données manquantes

- Des variables à fort taux de données manquantes identifiées (taux garanti ou de PB, CSP, par exemple) ;
- Une catégorie "inconnu(e)" créée pour le code fiscal, la CSP, la situation familiale, les départements agent et souscripteur, et le type de gestion ;
- Une imputation par règle métier pour l'information TMG/TMGA et le taux de PB.

3.3.2 Valeurs aberrantes

De nombreux contrôles de cohérence ont été réalisés sur les données afin de vérifier leur qualité. Il a ainsi par exemple été vérifié qu'aucun **flux** ou montant de **PM** n'était **négatif**, ou que les montants de PM et de flux extraits étaient cohérents avec les taux de flux observés et moyens sur l'historique disponible.

Les contrats pour lesquels l'**âge** et/ou l'**ancienneté** ont été jugés **aberrants** (ceux pour lesquels l'âge du souscripteur était négatif ou supérieur à 110 ans, l'ancienneté négative ou supérieure à 80 ans, et ceux pour lesquels l'âge extrait était inférieur à l'ancienneté du contrat) ont été **exclus** de la base de travail qui sera segmentée. Cependant, une valeur doit quand même leur être affectée en fin de processus. Elle le sera grâce à une méthode de type plus proches voisins une fois les valeurs épargne attribuées à tous les contrats de la base de travail.

Ensuite, des contrôles ont été effectués sur la cohérence des montants de flux avec les

montants de PM enregistrés, et des **bornes** supérieures ont été appliquées sur les taux observés et historiques.
Les différents taux de flux sont calculés de la manière suivante (en notant "RP" pour Rachat Partiel, "VL" pour Versement Libre, et "PM" pour Provision mathématique) :

$$\text{Taux Euro-UC}_i = \frac{\text{Montant sortant du fonds Euro}_i}{\text{PM Euro}_{i-1}}$$

$$\text{Taux UC-Euro}_i = \frac{\text{Montant entrant sur le fonds Euro}_i}{\text{PM UC}_{i-1}}$$

$$\text{Taux RP}_i = \frac{\text{Montant de rachat partiel}_i}{\text{PM}_{i-1}}$$

$$\text{Taux VL}_i = \frac{\text{Montant de versement libre}_i}{\text{PM}_{i-1}}$$

Il semble alors logique qu'un rachat partiel ne puisse pas être de montant supérieur à la PM avant rachat ou qu'un arbitrage du fonds euro vers des UC ne soit pas de montant supérieur à la PM Euro enregistrée à la fin de la période précédente. Cependant, les taux de flux étudiés sont calculés sur une période annuelle. Il se peut donc qu'un rachat ait été effectué à la suite de versements, arbitrages, ou revalorisation de l'épargne, et que son montant soit donc supérieur à la PM enregistrée, qui date de la fin de l'année précédente. Prenons par exemple un contrat de PM à 2000€ à fin décembre 2018, sur lequel un versement de 1500€ est effectué en février 2019, suivi d'un rachat partiel de 2500€ en novembre 2019. Le taux de rachat partiel alors observé vaut, d'après la méthodologie employée, $\frac{2500}{2000} = 125\%$.

Des taux de flux supérieurs à 100% ne semblent donc pas très logiques mais théoriquement possibles en considérant la méthodologie de calcul adoptée dans le cadre de l'étude.
Dans de tels cas, et afin d'éviter que ces taux ne cachent en fait des montants réellement aberrants et inexacts, des bornes ont été appliquées aux taux des différents flux, observés et moyens sur la période d'historique (voir partie suivante pour la description du calcul des taux moyens historiques).

Ainsi, les **taux de rachats partiels** ont été ramenés à 100% lorsqu'ils étaient supérieurs (cette situation est survenue pour seulement 1% des contrats de la base)
Une borne supérieure a également été appliquée aux taux d'arbitrages Euro-UC et UC-Euro, en considérant la réserve maximale d'euros (resp. UC) pouvant être arbitrée vers les UC (resp. le fonds euro). Lorsqu'ils étaient supérieurs à 100%, les **taux d'arbitrages** observés sur l'année étudiée ("i") ont donc été ramenés à :

$$\text{Taux Euro-UC}_i = \frac{\min(\text{Euro-UC}_i, \text{UC-Euro}_i + \text{PM Euro}_{i-1} + \text{VL Euro}_i)}{\text{PM Euro}_{i-1}}$$
$$\text{Taux UC-Euro}_i = \frac{\min(\text{UC-Euro}_i, \text{Euro-UC}_i + \text{PM UC}_{i-1} + \text{VL UC}_i)}{\text{PM UC}_{i-1}}$$

La même méthodologie de borne supérieure a été appliquée sur les taux moyens historiques, calculés sur 3 années et décrits ci-dessous.

En résumé : Valeurs aberrantes

- Des contrôles de cohérence effectués sur toute la base de données ;
- Un âge maximal fixé à 110 ans et une ancienneté maximale à 80 années ;
- Une borne supérieure appliquée aux taux de flux observés et moyens historiques.

3.3.3 Enrichissement de la base de données

À partir des donnés brutes extraites et exposées plus haut, il a ensuite été possible d'enrichir la base par des **variables supplémentaires** jugées **plus pertinentes** pour la modélisation du comportement des assurés en épargne :

- **Part UC histo 1 an** : part d'UC parmi l'encours total de l'année précédant celle des flux étudiés ;
- **Top éloignement agent** : indicatrice qui précise si l'assuré habite un département différent de celui de son agent ;
- **Top Bonus Euro+** : indicatrice précisant si le contrat est éligible au Bonus Euro+ ;
- **Top Fourgous** : indicatrice précisant si le contrat a été "fourgoussé" ;
- **Top Flux last** : indicatrice précisant, pour chaque type de flux, si au moins un mouvement a été effectué pendant l'année étudiée ;
- **NB Flux last** : nombre de mouvement effectués pendant l'année étudiée, pour chaque type de flux ;
- **Taux Flux last** : Taux de flux de l'année étudiée, pour chaque type de flux, calculée selon la définition exposée dans la partie précédente ;
- **Taux Flux moy old** : Moyenne sur 3 ans des taux de flux historiques, définie pour chaque type de flux par : $\text{Taux moy old}_{\text{Flux}}(N) = \frac{\text{Flux}_{N-1}+\text{Flux}_{N-2}+\text{Flux}_{N-3}}{\text{PM}_{N-2}+\text{PM}_{N-3}+\text{PM}_{N-4}}$;
- **Top NB Flux ext** : Indicatrice précisant, pour chaque type de flux, si le nombre de mouvements effectués lors d'au moins une des trois années d'historique est supérieur ou égal au nombre moyen de mouvements réalisés pour ce type de flux, sur une année, par les 10% des clients les plus actifs du portefeuille.

Le deuxième retraitement majeur effectué a consisté à convertir les variables quantitatives en variables qualitatives.

Ainsi, plusieurs variables ont été **discrétisées** par la création de tranches : **Tranche d'âge**, **Tranche d'ancienneté**, **Tranche de PM**, **Tranche de part UC**, par exemple.

Aussi, quand nécessaire, une **normalisation** et **mise à l'échelle** préliminaire à l'entraînement de modèles a été effectuée sur les données quantitatives. En effet, dans de nombreux algorithmes de *Machine Learning*, cette étape est essentielle pour ne pas avantager certaines variables au détriment d'autres et afin d'éviter que certaines n'aient un grand impact sur le modèle uniquement en raison de la forte magnitude dans ses valeurs. Par ailleurs, certains algorithmes convergent beaucoup plus rapidement lorsqu'ils traitent des données qui ont préalablement été mises à l'échelle.
Ainsi, les algorithmes reposant sur des calculs de distance (comme le *k-means* qui sera utilisé dans la suite de l'étude) nécessitent d'effectuer cette étape de normalisation, sans laquelle des variables contenant de grandes valeurs seraient jugées comme supérieures aux autres et domineraient donc lors des calculs de distance. Les autres algorithmes, basés sur des règles de décision (comme les arbres ou *Random Forests*) ne requièrent pas spécifiquement de passer par cette étape, mais ne seront pas pour autant biaisés par la normalisation.

Cette étape a alors été réalisée grâce à la fonction *RobustScaler*, une méthode de mise à l'échelle qui utilise des éléments statistiques robustes aux valeurs extrêmes (*outliers*, nombreuses sur certaines variables de la base, comme la PM par exemple). En effet, elle soustrait à chaque valeur de la variable la médiane, puis normalise les données en utilisant l'écart inter-quartile (différence entre le premier et le troisième quartile de la variable) :

$$\mathrm{X}_{i,scaled} = \frac{\mathrm{X}_i - \mathrm{med}(X)}{\mathrm{Q}_3(X) - \mathrm{Q}_1(X)}$$

La méthode de centrage-réduction la plus commune consiste à soustraire la moyenne et diviser par la variance, mais des *outliers* peuvent fortement influencer la moyenne et/ou la variance d'un jeu de données. C'est pourquoi la médiane et l'écart inter-quartile donnent souvent de meilleurs résultats plus robustes, et pourquoi cette méthode a donc été retenue pour l'étude.

Enfin, l'ajustement des modèles utilisés repose sur la construction d'une base ***train*** utilisée pour les entraîner, et d'une base ***test*** pour évaluer leur performance. Cette étape a été effectuée avec un tirage aléatoire de 80% de la base initiale pour la base *train* et les 20% restant pour la base *test*.

Notons que ces enrichissements ont été effectués afin d'obtenir une base complète contenant toutes les informations pertinentes pour les modélisations. En raison du très grand nombre de variables qu'elle contient, et de leurs dépendances/corrélation potentielles, toutes ne seront pas utilisées à chaque fois pour chaque modèle. Seule celles jugées plus judicieuses suite à une phase de **sélection de variables** seront conservées.

En résumé : Enrichissement de la base de données

- De nouvelles variables calculées et ajoutées car jugées plus pertinentes (part UC historique, nombres et taux de flux historiques, information sur l'éloignement avec l'agent, l'éligibilité au Bonus Euro+, l'existence d'un transfert Fourgous) ;
- Une étape de normalisation et mise à l'échelle nécessaire comme préliminaire à certains algorithmes de *Machine Learning* ;
- De nombreuses variables disponibles donc une nécessaire étape de sélection de variables.

3.4 Statistiques descriptives

Une fois les retraitements nécessaires effectués, la base de données a été explorée grâce à des statistiques descriptives.

L'étude s'intéresse à un portefeuille de clients d'AXA France, sur un certains nombre de réseaux, représentant une partie des engagements de la compagnie mais pas leur totalité. Pour des raisons de confidentialité, le nom des réseaux de distribution et des produits étudiés ont été volontairement modifiés. Ainsi, l'étude se concentre sur onze produits d'épargne/retraite individuelle, pouvant être distribués sur trois types de réseaux différents :

- **Réseau 1** : réseau spécialisé en offre d'épargne ;
- **Réseau 2** : réseau de démarchage ;
- **Réseau 3** : réseau multiple, non spécialisé en épargne et commercialisant également d'autres types de contrats (auto ou MRH par exemple).

Les graphiques ci-dessous illustrent la composition de la base de données selon plusieurs variables. Pour chacune d'entre elles, la part de contrats et de PM ainsi que les taux moyens des différents flux (versements, rachats partiels, arbitrages euro-UC et arbitrages UC-euro) sont représentés pour chaque modalité.

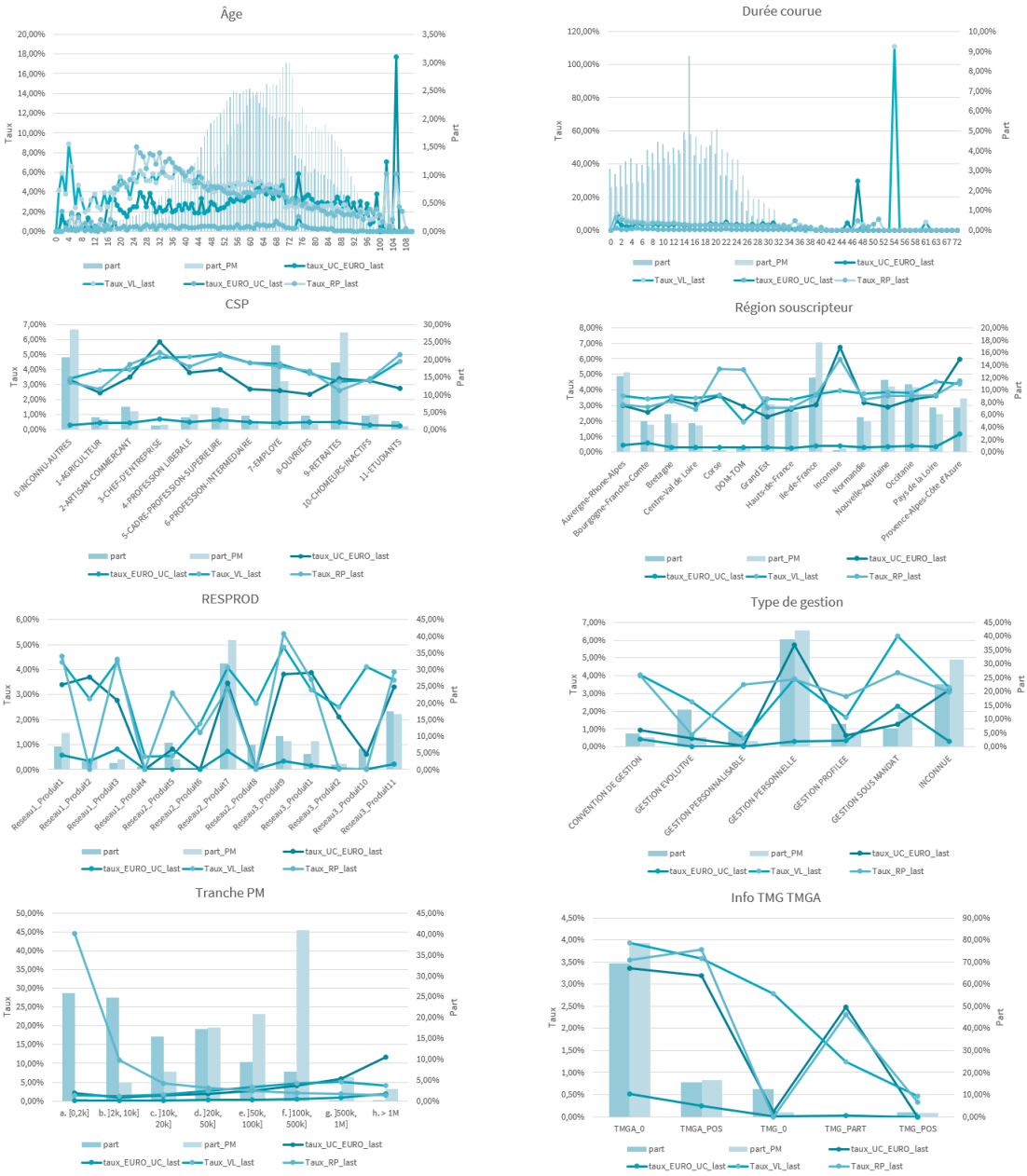

FIGURE 17 – Statistiques descriptives sur quelques variables de la base de données

Les contrats ne sont pas uniformément répartis (en nombre et en montants de PM) entre les différentes classes des variables, et certaines semblent même déséquilibrées. Cependant, certaines modalités présentent des taux de flux bien plus importants que les autres : il semble donc intéressant de prendre en compte ces variables dans les modélisations car leurs différentes modalités semblent refléter des écarts de comportement.

Ces différences de taux selon les catégories semblent correspondre à l'intuition qui pouvait exister lors du choix des variables de l'étude.
En effet, pour l'âge par exemple, des taux plus élevés sont observés entre 20 et 40 ans, années de début de la vie active, et sont beaucoup plus faibles pour les grands âges, à l'approche et lors de la retraite. De même, des taux plus élevés sont observés sur la CSP de chef d'entreprise (individus probablement mieux informés et plus actifs) que pour les retraités par exemple, qui ont moins intérêt d'effectuer des mouvements sur leur contrat. La région paraît également déterminante pour les taux de flux, puisque des écarts sont observés sur chaque type de flux entre les différentes régions. Le type de gestion semble, comme attendu, également expliquer une différence de comportement. En effet, les taux d'arbitrages (non automatiques) sont beaucoup plus élevés sur la gestion personnelle, ce qui peut s'expliquer par une information et une appétence aux concepts financiers sûrement plus fortes des individus possédant un contrat avec un tel type de gestion. Enfin, la distribution des taux de flux en fonction des différentes informations de TMG/TMGA semble également cohérente. En effet, pour les contrats sur lesquels un taux de TMG positif est garanti contractuellement, des taux de flux très faibles sont observés : sur de tels contrats, devenus rares aujourd'hui et garantissant une rémunération plus élevée que la plupart des autres contrats commercialisés actuellement, l'épargnant a plutôt intérêt de sécuriser son épargne et de ne pas effectuer de rachat(s) partiel(s) par exemple.

Un pic de contrats de durée courue égale à 15 ans est observé sur le deuxième graphique. Celui-ci est expliqué par le lancement d'un nouveau produit en 2004, qui a induit une hausse des affaires nouvelles cette année-là.
D'autres pics dans les taux de flux sont observés pour les grands âges et/ou grandes anciennetés. Ils s'expliquent par le faible effectif de contrats présentant cette caractéristique dans la base de données, donc la faible PM cumulée. Ainsi, un flux réalisé par un seul de ces contrats implique un taux correspondant très fort sur l'ensemble de la tranche considérée, d'où les pics observés.

Les graphiques ci-dessous représentent les mêmes parts et taux en fonction des tranches taux de flux moyens historiques.

FIGURE 18 – Répartition des contrats et des taux de flux en fonction des tranches de flux moyens historiques

Puisque les flux observés sont assez rares, la plus grande partie des contrats de la base se situent dans la tranche de flux moyens historiques à 0%. En effet, sur l'ensemble de la base de donnée, seulement $2,41\%$ des clients ont effectué au moins un arbitrage lors de l'année étudiée, $6,21\%$ des clients au moins un rachat partiel et $4,77\%$ des clients au moins un versement.
Par ailleurs, il semble que plus le taux moyen historique est élevé, plus le taux observé lors de l'année d'étude est grand. Une relation existe donc sûrement entre les taux moyens historiques et les taux actuels observés, et ces variables paraissent donc pertinentes dans le cadre de la modélisation du comportement des assurés.

Par ailleurs, l'analyse du rapport entre localisation géographique et taux de flux a été approfondie par la construction de cartes, fournissant un résumé graphique des taux moyens observés dans chaque département :

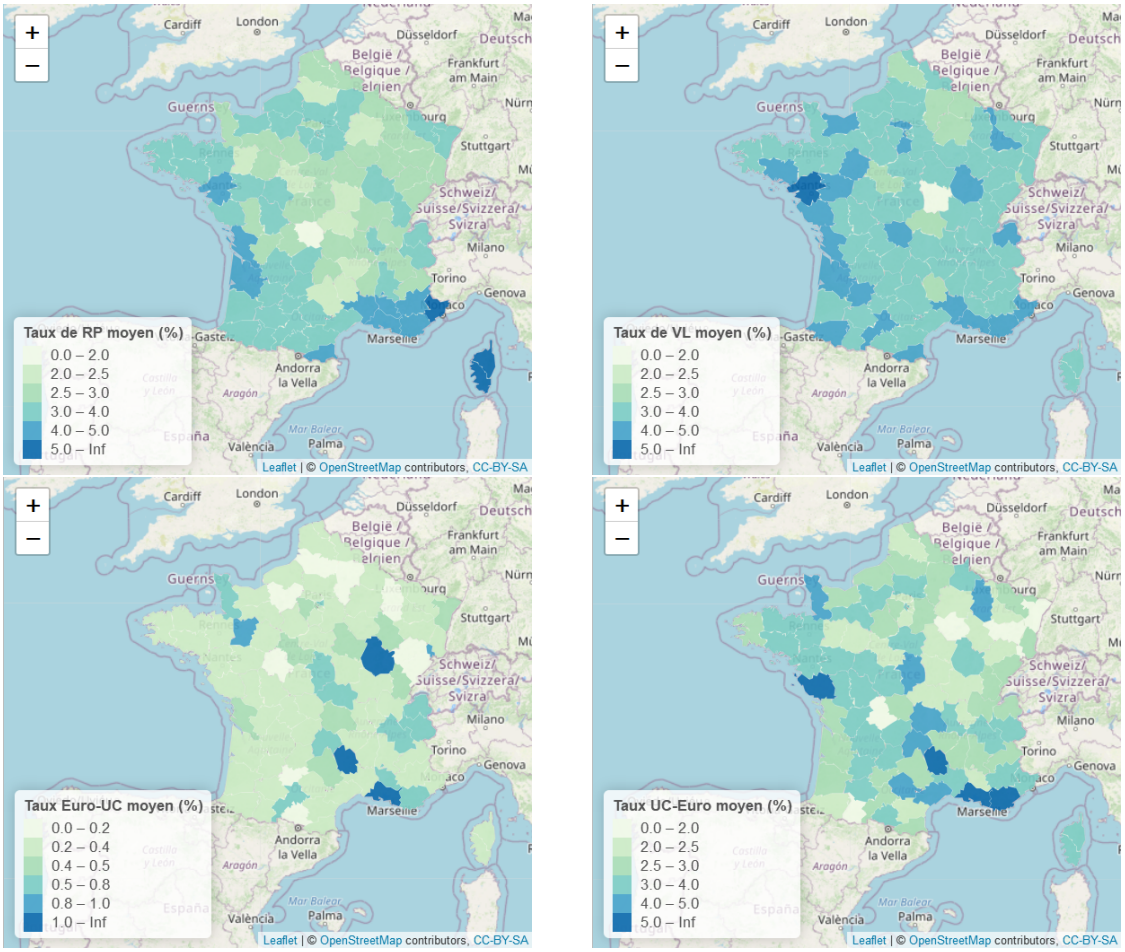

FIGURE 19 – Taux moyens observés de rachats partiels (RP), versements libres (VL), arbitrages Euro-UC, et arbitrages UC-Euro par département de l'assuré

Les taux des différents flux semblent distribués de manière hétérogène selon les départements Français. En effet, ils paraissent bien plus élevés dans le Sud-Est pour les rachats partiels, et dans l'Ouest pour les versements par exemple. Les taux d'arbitrages euro-UC et UC-euro semblent également fortement dépendre du département de l'épargnant. Ces observations permettent donc de valider la prise en compte du critère de localisation géographiques dans les variables permettant d'expliquer les comportements des épargnants.

Les corrélations entre variables explicatives d'une part, et entre variables explicatives et variable cible d'autre part ont également été étudiées, afin de déterminer si certaines variables sont redondantes et doivent donc être écartées des modélisations, et si certaines semblent plus appropriées pour étudier le comportement des assurés. La matrice ci-dessous illustre la corrélation entre chaque variable explicative (estimée grâce au coefficient de Pearson)

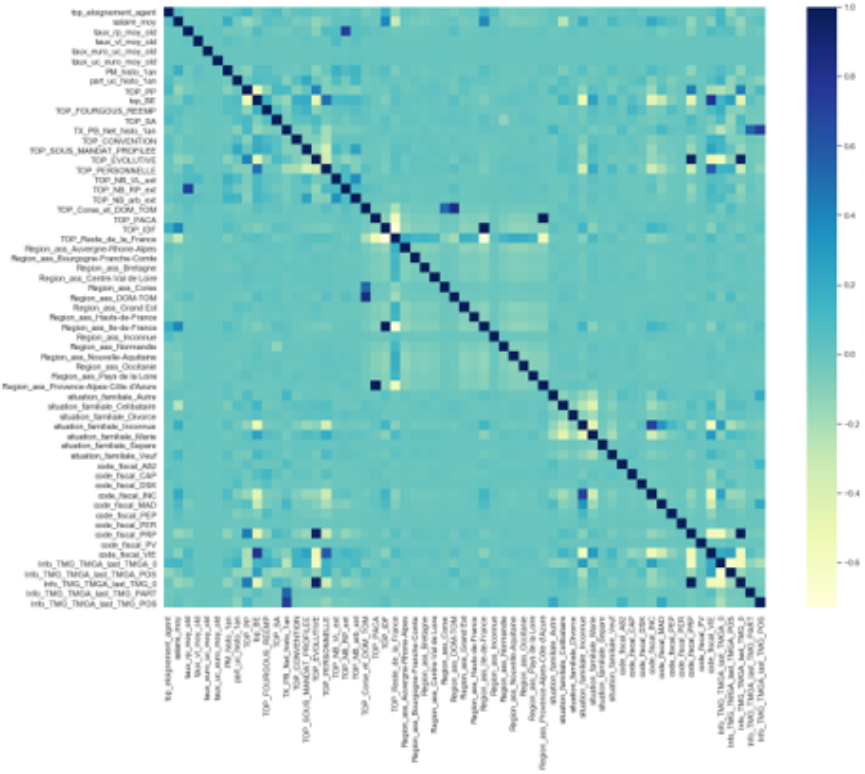

FIGURE 20 – Matrice de corrélation entre variables explicatives

Il convient de noter que la plupart des variables ont une corrélation linéaire quasi nulle avec les autres variables de la base de données. Certaines doivent être étudiées car présentent une corrélation beaucoup plus élevée et il faut donc vérifier qu'il ne s'agit pas de redondance inutile d'information. Enfin, une autre analyse a permis de déterminer qu'il n'existe pas de corrélation à 100% entre une de ces variables et la variable indicatrice précisant la survenance d'un flux lors de l'année étudiée (qui sera utilisée comme variable cible lors de l'étape de sélection de variables).

En résumé : Statistiques descriptives

- Des taux de flux moyens qui varient fortement selon les modalités des variables considérées (informations sur le client ou son contrat) ;
- Une présence de variables corrélées entre elles ;
- Des variables plus ou moins corrélées à la variable cible, mais aucune à 100%.

3.5 Contraintes réglementaires

Tout d'abord, le Règlement Général de Protection des Données (**RGPD**), adopté en 2016, impose des exigences de conformité fortes sur les données utilisées par les assureurs. En effet, elles doivent par exemple être protégées par des mesures telles que la **pseudonymisation**, qui rend impossible l'attribution des données à une personne physique sans avoir recours à des informations supplémentaires. Aussi, seules les données à caractère personnel nécessaires sont traitées et elles ne sont pas mises à disposition d'un nombre indéterminé de personnes sans que l'individu concerné n'en soit conscient. À cet effet, la base de données créée dans le cadre de l'étude sur la valeur client ne contient **aucune donnée sensible**, de santé ou médicale telles que décrites dans le RGPD.

L'un des objectifs principaux de l'indice client consiste à orienter les stratégies et actions envers certains clients du portefeuille, et à accorder des avantages aux "meilleurs" souscripteurs. À la manière des techniques de *pricing* qui n'utilisent pas de donnée sexuée pour éviter des **discriminations de genre**, et afin d'éviter tout biais dans les algorithmes (qui doivent rester les plus neutres possibles), l'étude sur la valeur client n'utilise à aucun moment la variable "Sexe" dans les méthodes employées, que ce soit au niveau de la segmentation comme de la création de lois de comportement.

Enfin, l'article 82 de la Directive 2009/138/CE (Solvabilité II) [24] introduit des exigences en matière de qualité des données utilisées pour le calcul des provisions techniques. Trois critères sont précisés dans les articles 19 à 21 du Chapitre III section 2 du Règlement délégué [25] afin d'en apprécier la qualité. Ainsi, les données utilisées doivent, selon ces textes, être **exhaustives**, **exactes**, et de **caractère approprié**. C'est notamment dans l'objectif de remplir ces exigences qu'un maximum de données et variables ont été prises en compte pour l'étude, et qu'une longue étape de vérifications de cohérence et de correction des valeurs manquantes et/ou aberrantes a été réalisée.

Ces nombreux impératifs peuvent apparaître comme une limitation au *Big Data*, ce qui pourrait expliquer l'utilisation pour l'instant encore marginale des techniques de *machine learning* pour compléter les méthodes actuarielles. Cependant, ces exigences permettent de protéger les données personnelles des clients, dont AXA a choisi de faire sa priorité ces dernières années. Par ailleurs, les analyses comportementales permises par la *Data Science* et dont un exemple est fourni dans ce mémoire semblent prometteuses car elles permettent la constitution de nouveaux groupes de risques homogènes via une approche *data driven*. Finalement, l'étude a été menée de manière sereine et complète, aucune contrainte réglementaire n'ayant empêché les prédictions et analyses.

En résumé : Contraintes réglementaires

- Un contexte réglementaire strict sur l'utilisation des données ;
- Des textes aux exigences précises (RGPD, Solvabiilité II) ;
- Des contraintes respectées dans le cadre de l'étude.

4 Segmentation et détermination des lois de comportement

4.1 Présentation des modèles et des différentes méthodes de segmentation

Cette partie approfondit l'étape de la segmentation, notamment en expliquant la démarche générale et les différents modèles utilisés.
Dans un premier temps, une méthode de **segmentation directe** du portefeuille après sélection des variables les plus pertinentes a été effectuée.
Dans un second temps, ce modèle a été challengé par une autre méthode consistant à commencer par prédire des **lois de comportements**, à partir desquelles la segmentation a ensuite pu être réalisée.

4.1.1 Notions utiles pour la suite

Après avoir préparé la base de données comme expliqué dans la partie précédente, il convient de sélectionner les variables les plus appropriées pour la segmentation. Cette étape préliminaire à la segmentation est nécessaire afin que les modèles de *clustering* soient plus performants et est réalisée en apprentissage supervisé.

Cette **sélection de variables** a été effectuée grâce à un modèle d'apprentissage supervisé, sur l'indicatrice qui précise si au moins un flux (rachat partiel par exemple) a été effectué pendant l'année d'étude (et qui vaut donc 1 si au moins un rachat partiel a été effectué sur le contrat considéré pendant l'année d'étude, et 0 sinon). Concrètement, cela revient à entraîner, pour chaque type de flux, un **modèle de classification binaire** sur la base des indicatrices de flux, et à analyser quelles ont été les **variables les plus importantes** lors de ce travail de prédiction. Ces variables sont alors sélectionnées pour constituer les caractéristiques selon lesquelles la base client sera segmentée.

Avant de détailler les modèles utilisés lors de l'étude, quelques notions doivent être expliquées.

Validation croisée

Un classifieur (ou régresseur) entraîné sur une base de données doit être généralisable : il doit être capable de segmenter (ou prédire) correctement à partir de données différentes de celles de la base d'entraînement.
Apprendre les paramètres d'une fonction de prédiction et la tester sur les mêmes données constitue donc une erreur méthodologique : un modèle qui se contenterait de répéter les étiquettes des échantillons qu'il vient de voir aurait un score parfait mais ne pourrait rien prédire d'utile sur des données nouvelles. Cette situation s'appelle le **surapprentissage** (*overfitting*). Pour l'éviter, il est courant, lors d'une expérience d'apprentissage supervisé, de considérer une partie des données disponibles comme une base de test, afin d'entraîner et de valider le modèle sur deux bases de données différentes.

D'une part, la base de données est séparée en base d'entraînement et base de test. La première servira à entraîner le modèle retenu, et la seconde à évaluer sa performance prédictive. D'autre part, il convient de sélectionner le modèle à entraîner, et d'en déterminer les meilleurs hyper-paramètres. Pour ce faire, des techniques de recherche par grille peuvent être employées.

Lors de l'évaluation des différents hyper-paramètres des estimateurs, il existe toujours un risque de surapprentissage de la base de test, car les paramètres peuvent être ajustés jusqu'à ce que l'estimateur fonctionne de manière optimale. Pour résoudre ce problème, une autre partie de la base de données peut être considérée comme une base de validation : l'apprentissage s'effectue sur la base d'entraînement, puis l'évaluation est faite sur l'ensemble de validation. Lorsque l'expérience semble être réussie et que les hyper-paramètres optimaux sont déterminés, l'évaluation finale peut être faite sur la base de test.

Toutefois, en divisant les données disponibles en trois ensembles, le nombre d'échantillons pouvant être utilisés pour l'apprentissage du modèle se voit considérablement réduit, et les résultats peuvent dépendre d'un choix aléatoire particulier de bases d'entraînement et de validation.

Une solution à ce problème est une procédure appelée **validation croisée**, qui s'insère dans la modélisation comme l'illustre le schéma ci dessous.

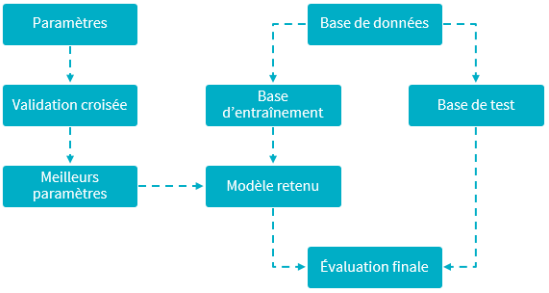

FIGURE 21 – Principes de la validation croisée

Une base de test est toujours mise à part pour l'évaluation finale, mais la base de validation n'est plus nécessaire pour la validation croisée. Dans l'approche la plus commune, appelée validation croisée *k-folds*, la base d'entraînement est divisée en k bases plus petites (d'autres approches [26] peuvent aussi être utilisées, mais suivent généralement les mêmes principes). La procédure suivante, illustrée par le schéma ci-dessous, est alors suivie pour chacune des k sous-bases :

- un modèle est entraîné en utilisant $k - 1$ sous-bases en tant que données d'entraînement ;
- le modèle résultant est validé sur la partie restante des données, utilisée comme base de test pour calculer une mesure de performance telle que la précision.

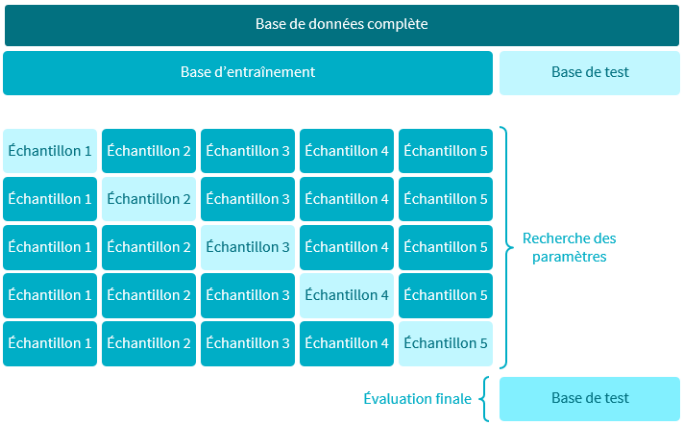

FIGURE 22 – Validation croisée *k-folds* (avec $k = 5$)

La mesure de performance affichée par la validation croisée *k-folds* est alors la moyenne des valeurs calculées dans la boucle. Cette approche peut être coûteuse en termes de calcul et donc en temps d'exécution, mais ne gaspille pas trop de données (comme c'est le cas lors de la fixation d'un ensemble de validation arbitraire). Il n'existe pas de règle fixe pour choisir la valeur de k mais les validations croisées *5-folds* ou *10-folds* sont les plus courantes. Au vu de la grande taille de la base utilisée dans le cadre de l'étude sur la valeur client (plus de $2,2$ millions de contrats), le choix a été fait ici de fixer $\boldsymbol{k = 5}$ pour les travaux.
Enfin, l'avantage de cette méthode plutôt que celle du *bootstrap* (sous-échantillonage aléatoire) est que toutes les observations sont utilisées à la fois pour l'entraînement et la validation, et chaque observation est utilisée exactement une fois pour la validation.

Évaluation de la performance d'un modèle et fonction de *scoring*

Le critère d'évaluation est un facteur clé à la fois dans l'évaluation de la performance du modèle, dans la recherche de ses hyper-paramètres, et dans la comparaison de différents modèles.
Dans un problème de classification binaire, la **matrice de confusion** représente les résultats des individus correctement et incorrectement reconnus de chaque classe :

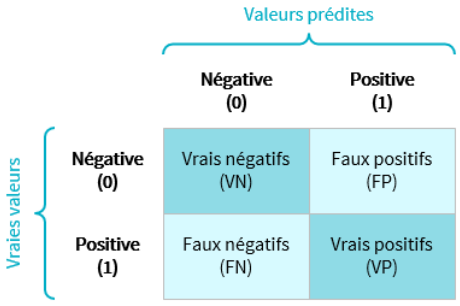

FIGURE 23 – Matrice de confusion

Traditionnellement [11], l'**exactitude** (*accuracy*) a été la mesure empirique la plus utilisée. En reprenant les éléments de la matrice de confusion, elle est définie ainsi :

$$\text{Exactitude} = \frac{\text{VN} + \text{VP}}{\text{VN} + \text{FP} + \text{FN} + \text{VP}}$$

Autrement dit, l'exactitude mesure la part d'individus correctement classés, parmi tous les individus.
Cependant, dans le cadre de l'étude de bases de données déséquilibrées (comme c'est le cas ici), l'exactitude n'est plus une mesure appropriée car elle ne fait pas la distinction entre le nombre d'individus correctement classés de différentes classes. Elle peut donc conduire à des conclusions erronées.
Ainsi par exemple, si une base de données est déséquilibrée de telle sorte que l'étiquette cible comporte 98% de "0" et 2% de "1", alors un classifieur qui prédirait un "0" pour tous les individus serait jugé très performant selon cette métrique puisque son score serait de 98%. Mais alors 2% de la base comporte des individus étiquetés "1" qui seraient prédits dans la classe "0", et selon le but de la classification, le coût de cette erreur de prédiction peut être élevé (prédiction de maladies graves, de rachats massifs, par exemple).

C'est pour cette raison que, lors de l'étude d'une base de données déséquilibrée, il faut prendre en compte une métrique plus appropriée que l'exactitude.
En particulier, une façon de combiner les différents éléments de la matrice de confusion est d'étudier la courbe **ROC** (*Receiver Operating Characteristic*). Elle permet de visualiser le compromis entre les bénéfices (taux de vrais positifs) et les coûts (taux de faux positifs), et démontre ainsi qu'aucun classifieur ne peut augmenter le nombre de vrais positifs sans augmenter le nombre de faux positifs. L'aire sous la courbe ROC (**AUC**, *Area Under the ROC Curve*) fournit une mesure unique de la performance d'un classifieur pour déterminer quel modèle est meilleur en moyenne. La figure ci-dessous illustre la forme générale d'une courbe ROC.

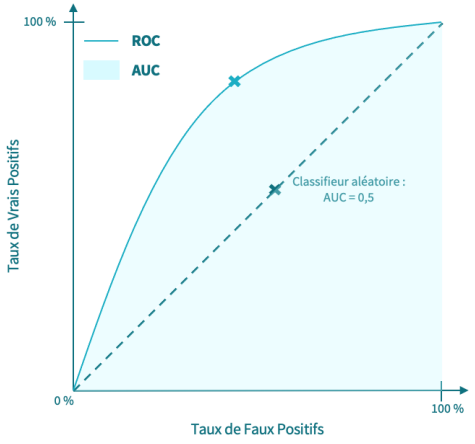

FIGURE 24 – Courbe ROC et AUC

Les points (0 ;0) et (1 ;1) sont des classifieurs triviaux pour lesquels la classe prédite est toujours respectivement négative et positive. Au contraire, le point (0 ;1) représente la classification parfaite. L'AUC est calculée en obtenant simplement la surface sous la courbe ROC.

D'autres indicateurs peuvent être obtenus à partir de la matrice de confusion :

- **la précision** mesure la proportion de vrais positifs parmi tous les positifs :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

  Intuitivement, c'est la capacité du classifieur à ne pas étiqueter comme positif un individu qui est négatif ;

- **le rappel** mesure la proportion d'individus correctement prédits positifs parmi tous ceux qui sont vraiment positifs :

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

  Intuitivement, c'est la capacité du classifieur à trouver tous les échantillons positifs.

Ces deux indicateurs sont alors résumés en une métrique, le **F1 Score**. Il s'agit de la moyenne harmonique de la précision et du rappel, qui donne une meilleure mesure des individus incorrectement classifiés que l'exactitude :

$$\text{F1 Score} = \left(\frac{\text{Précision}^{-1} + \text{Rappel}^{-1}}{2} \right)^{-1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Ainsi, dans le cas où les mauvaises prédictions sont cruciales (comme dans l'étude sur la valeur client, dans laquelle il convient de n'oublier aucun individu dans la classe "1" lors de la prédiction), le F1 Score semble plus pertinent que l'exactitude. [27] Ce dernier est surtout une meilleure métrique que l'exactitude pour travailler avec des bases de données déséquilibrées.

Deux métriques semblent donc pouvoir être utilisées à la place de l'exactitude dans notre cas de classification binaire sur une base de données très déséquilibrée. Après avoir comparé les résultats en utilisant successivement l'AUC et le F1 Score comme fonction de score dans le *tuning* des hyper-paramètres par validation croisée, il semble que les résultats diffèrent peu et que l'optimisation effectuée en cherchant les paramètres donnant le meilleur F1 Score permette d'obtenir *in fine* des performances de prédiction plus élevées. C'est pour ces raisons que le **F1 Score** est retenu pour toute la suite de l'étude comme **fonction de score** dans la validation croisée, et comme **métrique privilégiée** pour la comparaison de modèles (les autres indicateurs seront également considérés, dans un deuxième temps, pour confirmer les conclusions).

En résumé : Notions utiles

- Une étape de validation croisée $5-$*folds* pour déterminer les meilleurs hyper-paramètres de chaque modèle et éviter le surapprentissage ;
- Différentes métriques possibles pour évaluer la performance d'un modèle et servir de fonction de *scoring*, dont une privilégiée : le F1 Score.

4.1.2 Méthode 1 : segmentation directe de la base

Modèles utilisés pour la sélection de variable

Plusieurs types de modèles, décrits ci-dessous, ont été testés afin de prédire la survenance d'un flux (arbitrage, rachat, versement) et sélectionner les variables : une régression logistique et des modèles de *machine learning* (Random Forest, Extremely Randomized Trees, XG Boost). Celui jugé le plus performant lors de l'étape de sélection de modèle a ensuite été retenu ainsi que ses variables les plus importantes.

Régression logistique

Le premier modèle testé pour modéliser et prédire la survenance de flux a été la **régression logistique**. Celui-ci offre un bon compromis entre **performance** du modèle et **pouvoir explicatif**, et est souvent utilisé comme premier essai "simple" de modélisation, avant de tester des modèles plus complexes et afin de constituer un benchmark de performance.
Il s'agit d'un cas particulier de Modèle Linéaire Généralisé (GLM), dans le cas où la variable cible à prédire est **binaire** : ici une indicatrice qui vaut 1 si le flux considéré a été effectué pendant l'année observée, et 0 sinon.

En notant $Y \in \mathbb{R}^m$ la variable cible et $X \in \mathbb{R}^{n \times m}$ l'ensemble des n variables explicatives, la régression logistique repose sur les hypothèses suivantes :

- Y est binaire ;
- $Y \mid X, \Theta$ est caractérisée par la fonction sigmoïde g : $\mathbb{P}(Y = 1 \mid x_i, \Theta) = g(x_i\Theta) = g(\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_n x_{in})$ avec $g(t) = \frac{1}{1+e^{-t}}$, pour tout $t \in \mathbb{R}$.

Et la classification s'effectue alors selon le critère de décision suivant :

$$\begin{cases} 1 \text{ si } g(x_i\Theta) \geq 0,5 \\ 0 \text{ sinon} \end{cases}$$

Le meilleur vecteur de coefficients Θ est déterminé, comme pour la régression linéaire, par minimisation d'une certaine fonction de coût. Étant donné le problème de classification étudié et le caractère binaire de la variable cible, un choix légitime est alors de minimiser la fonction de perte logarithmique (*"log loss"*). Ainsi, la régression logistique considérée revient à résoudre le problème d'optimisation suivant :

$$\min_{\Theta} \frac{1}{m} \sum_{i=1}^{m} -y_i \log\left(\frac{1}{1+e^{-x_i\Theta}}\right) - (1-y_i)\log\left(1 - \frac{1}{1+e^{-x_i\Theta}}\right)$$

Il n'existe pas de solution analytique pour trouver le vecteur Θ optimal, mais cette fonction de coût est convexe, donc il est possible d'en trouver le minimum global grâce à un algorithme d'optimisation.
Dans le cadre de la présente étude, c'est l'algorithme *"saga"* qui a été utilisé. Il s'agit d'une variante de ***Stochastic Average Gradient descent***, qui converge donc plus rapidement qu'un algorithme classique sur de grandes bases de données (avec à la fois un grand nombre d'observations et de variables), et qui supporte la régularisation L_1. Cet algorithme est souvent considéré comme le meilleur solveur dans le cas de régression logistique éparse sur une grande base de données. C'est pourquoi il a été retenu dans le cadre de l'étude.

Enfin, seulement 6,21% des contrats de la base de travail ont subi un rachat partiel lors de l'année étudiée, donc la variable cible de la classification binaire contient seulement 6,21% de 1 et est remplie de 0 sinon. Le problème de classification binaire est donc réalisé sur une base de données **déséquilibrée**. Dans ce cadre, un autre argument de la fonction *LogisticRegression* a mérité une attention particulière : l'argument ***class weight***. Dans le cas d'une classification sur base (très) déséquilibrée, cet argument est fixé à la valeur "*balanced*", qui permet d'utiliser les valeurs de Y pour ajuster automatiquement les poids de manière inversement proportionnelle aux fréquences des classes. [5]
Le poids alors attribué à une prédiction sur la classe minoritaire est beaucoup plus important que celui affecté à une prédiction sur classe majoritaire, ce qui permet de pénaliser plus lourdement les erreurs effectuées sur la classe minoritaire (ici, les clients ayant effectué un rachat partiel).

Cette méthode a pour effet de **pénaliser** davantage les **erreurs de prédiction** effectuées sur la **classe minoritaire**, et donc d'adapter les coefficients de la régression en conséquence. Une comparaison de régression logistique avec et sans cet argument a d'ailleurs montré qu'il permettait d'augmenter sensiblement le nombre de prédictions de la classe minoritaire.

Random Forest
Face à des premiers résultats peu concluants, il a ensuite été décidé d'utiliser des modèles plus poussés de *Machine Learning*, à commencer par un algorithme de ***Random Forest***, appartenant à la famille des méthodes dites ensemblistes.

Le principe fondamental derrière ce type d'algorithme est de remplacer un unique estimateur complexe par plusieurs de moindre qualité individuelle, disposant chacun d'une vision restreinte du problème d'apprentissage à résoudre, puis réunis pour former une vision globale. C'est ce dernier rassemblement qui les rend très performants.

Les *Random Forests* ont été introduits par Leo Breiman [4], qui avait précédemment élaboré la méthode CART (*Classification And Regression Tree*) : des arbres de décision permettant d'expliquer une variable cible à partir d'une série de variables discrètes ou continues.
Un arbre de décision **partitionne** les individus en groupes les plus **homogènes** possible du point de vue de la variable à prédire. Plusieurs itérations sont nécessaires à la construction d'un arbre entier, chacune consistant à diviser les individus en k classes (en

général, $k = 2$) pour expliquer la variable cible. La première séparation est effectuée en sélectionnant la variable qui fournit la meilleure partition des observations, et définit des sous-groupes représentés pas les "noeuds" de l'arbre. L'opération de séparation est répétée sur chaque sous-groupe d'individus, jusqu'à ce qu'aucune division ne soit possible. Cela aboutit aux noeuds terminaux de l'arbre, appelés "feuilles", chacune caractérisée par un chemin spécifique dans l'arbre.

Utilisé seul, un arbre de décision peut souffrir d'une **performance** trop **dépendante** de **l'échantillon choisi**, c'est pourquoi les *random forests* ont été introduits. En utilisant une série d'arbres de décision, disposant chacun d'une vision restreinte du problème (double échantillonnage, tiré aléatoirement, des observations et des variables utilisées), la performance de l'algorithme dépend moins de l'échantillon utilisé. Dans le cadre d'un problème de classification, le résultat final est obtenu en faisant classer chaque arbre et en retenant la catégorie majoritaire (pour un problème de régression, la moyenne des prédictions serait retenue).

Les *random forests* reposent donc sur deux principes :

- le *tree bagging* : tirage aléatoire des observations ;
- le *feature sampling* : tirage aléatoire des variables explicatives.

Considérons une matrice X de m individus, décrits par n variables explicatives, et un vecteur cible $Y \in \mathbb{R}^m$.
La construction des arbres de décision par *tree bagging* consiste à :

- tirer aléatoirement et avec remplacement des échantillons de (X, Y) ;
- entraîner un arbre de décision sur chaque couple de (X, Y) tiré aléatoirement.

Chacun des arbres construits est ensuite appliqué à de nouvelles données, et la catégorie ayant comptabilisé la majorité des prédictions est retenue comme prédiction finale de l'assemblage des arbres.

En supplément de ce tirage aléatoire par *tree bagging* effectué sur les observations, un tirage aléatoire sur les variables à utiliser dans chaque arbre de décision est réalisé par *feature sampling*. Ce deuxième échantillonnage permet de réduire la variance de l'ensemble créé. En effet, la moyenne de B variables indépendantes et identiquement distribuées, chacune de variance σ^2, a une variance de $\frac{\sigma^2}{B}$. En excluant l'hypothèse d'indépendance des variables et en notant ρ le coefficient de corrélation des paires de variables, la variance de l'ensemble des arbres de décision est alors :

$$V_{\text{forest}} = \underbrace{\frac{1-\rho}{B}\sigma^2}_{\text{Bagging}} + \underbrace{\rho\sigma^2}_{\text{Feature sampling}}$$

Ainsi, en augmentant le nombre B d'arbres de décision (en prenant garde au surapprentissage) utilisé dans l'étape de *bagging*, et en diminuant la corrélation entre les arbres grâce au *feature sampling*, la **variance** du problème d'apprentissage peut être nettement **réduite**.

De nombreux critères de *split* existent lors de la construction d'un arbre, pour déterminer la façon d'effectuer la séparation des individus considérés en deux sous-populations les plus homogènes possibles. La plus courante est le **critère de Gini**, qui se concentre sur la classe la plus représentée dans le jeu de données, en cherchant à la séparer le plus rapidement possible. Critère le plus utilisé car permettant généralement de construire des arbres de grande qualité, c'est celui qui a été retenu dans le cadre de l'étude. De plus, il présente l'avantage (par rapport au critère de gain d'information basé sur l'entropie par exemple) de ne pas calculer des fonctions logarithmiques, qui nécessitent beaucoup de calculs.

Moins facilement explicable qu'un modèle linéaire pour lequel l'algorithme prend une décision particulière car le poids affecté à une variable a une valeur précise, l'algorithme de *random forest* peut quand même être expliqué grâce aux **importances des variables**. Plus une variable est utilisée par un grand nombre d'arbres de la forêt, et plus elle est utilisée "haut" dans ces arbres, plus son importance sera grande.

L'importance d'une variable correspond à la réduction moyenne d'impureté sur l'ensemble des nœuds où cette variable intervient, au travers de l'ensemble des arbres de la forêt aléatoire. Cette réduction moyenne est pondérée par l'importance de chaque nœud, puis normalisée pour que la somme des importances soit égale à 1.

Pour chaque arbre, l'importance d'un nœud est calculée en utilisant la méthode d'impureté de Gini, en supposant qu'il n'y a que deux nœuds enfants (arbre binaire) :

$$Imp_j = w_j C_j - w_{gauche,j} C_{gauche,j} - w_{droite,j} C_{droite,j}$$

avec :

- Imp_j l'importance du nœud j ;
- w_j le nombre pondéré d'échantillons atteignant le nœud j ;
- C_j la valeur d'impureté du nœud j.

L'impureté de Gini est la probabilité de classer incorrectement un élément choisi au hasard dans la base de données s'il était étiqueté au hasard selon la distribution des classes dans la base :

$$C_j = \sum_{l=1}^{\text{nombre de classes}} p(l) \times (1 - p(l))$$

avec $p(l)$ la probabilité de tirer aléatoirement un élément de la classe l.

L'importance de chaque variable sur un arbre est alors calculée ainsi :

$$Imp_{var_i} = \frac{\sum_{j \text{ t.q. le nœud } j \text{ split sur la variable } i} Imp_j}{\sum_{k \in \text{tous les nœuds}} Imp_k}$$

Cette importance est ensuite normalisée pour aboutir à un chiffre entre 0 et 1 en divisant la quantité ci-dessus par la somme de toutes les valeurs d'importances de variables.

L'importance finale d'une variable, au niveau du *random forest* total, est ensuite la moyenne de toutes les importances sur les différents arbres de la forêt.

Enfin, de la même manière qu'une méthode de **pondération** a permis la pénalisation des erreurs effectuées sur la classe minoritaire dans le cadre de la régression logistique, la même technique a été appliquée dans l'algorithme de *Random Forest*, en utilisant le même paramètre de *class weight* disponible sur *Scikit Learn*.
Les résultats se sont alors révélés largement meilleurs qu'avec un algorithme de ***Random Forest*** classique, comme le suggéraient Chao Chen, Leo Breiman et Andy Liaw dans un papier de 2004. [5]

Extremely Randomized Trees (Extra Trees)
L'algorithme d'*Extra Trees* constitue une **variante** de *random forest*, en rajoutant une **composante aléatoire supplémentaire**, et en comporte donc à trois niveaux : lors de l'étape de *bagging*, lors de celle de *feature sampling*, et lors d'une étape de ***random split***. Pour rappel, le *random forest* sélectionne la meilleure variable parmi les $\sqrt{n}$ en fonction de celle qui réalise le meilleur *split* au sens du critère de Gini. L'algorithme d'*extra trees* réalise des *splits* de manière aléatoire et sélectionne ensuite la meilleure variable sur ce résultat uniquement. Plutôt que de rechercher les meilleurs seuils possibles, l'algorithme d'*extra trees* rend donc les arbres encore plus aléatoires en utilisant un seuil aléatoire pour chaque variable. Cette méthode permet généralement d'obtenir de meilleurs résultats et de réduire un peu plus la variance du modèle, au prix d'une augmentation légèrement plus importante du biais. Elle permet surtout de gagner en temps d'exéctution (par rapport à l'entraînement d'une forêt aléatoire "simple"), car la recherche du meilleur seuil possible pour chaque variable, à chaque nœud, est une des étapes les plus chronophages lors de la construction d'un arbre.

Comme pour les modèles précédents, une **version pondérée** de l'*extra trees* peut être implémentée, qui permet de prendre en compte le déséquilibre des classes dans la variable à prédire. Encore une fois, cette version pondérée s'est révélée **plus performante** que la version classique d'*extra trees*.

XG Boost
Le *XG Boost* est une version extrême du ***Gradient Boosting***, détaillé ci-dessous.
Le *gradient boosting* est une méthode ensembliste non linéaire très performante. En reprenant le fonctionnement général du *boosting* (qui construit itérativement un algorithme corrigeant à chaque étape les erreurs effectuées précédemment), il le généralise à l'utilisation de plusieurs fonctions de coût. À chaque étape, l'erreur sera corrigée cette fois en calculant le gradient de la fonction de coût choisie en amont. Concrètement, le *gradient boosting* fonctionne par ajouts successifs de prédicteurs à un ensemble, chacun d'eux corrigeant son prédécesseur, en ajustant chaque nouveau prédicteur aux erreurs résiduelles du précédent.

Soient :
- X les variables explicatives et m le nombre d'observations : $X = \{x_1, x_2, ..., x_m\}$;
- Y la variable cible : $Y = \{y_1, y_2, ..., y_m\}$;

- h la fonction qui cherche à approximer Y ;
- J la fonction de coût : $J(h) = \sum_{i=1}^{m} j(y_i, h(x_i))$.

L'algorithme cherche à approximer le vecteur cible par une méthode de *boosting* : il va construire itérativement B fonctions hypothèses h, qui seront finalement assemblées en une fonction H pour approximer le vecteur cible.

Tout d'abord, h_1 réalise un premier apprentissage des données et sera la seule fonction à constituer H à la fin de la première étape. Lors de la deuxième étape, h_2 essaie de corriger l'erreur effectuée par h_1. On cherche donc h_2 telle que :

$$\forall i \in \{1, ..., m\}, h_2(x_i) \approx y_i - H(x_i)$$

avec $y_i - H(x_i)$ qui représentent les résidus de l'étape précédente.
Le principe du *gradient boosting* consiste à exprimer ces résidus comme un gradient négatif d'une certaine fonction de coût. Ainsi, en considérant par exemple les moindres carrés pour un problème de régression, on cherche à minimiser :

$$J = \sum_{i=1}^{m} j(y_i, h(x_i)) = \sum_{i=1}^{m} \frac{(y_i - H(x_i))^2}{2}$$

Le problème que doit résoudre la nouvelle fonction h_2 est donc de minimiser la fonction de coût J, dont le gradient par rapport à $H(x_i)$ s'écrit :

$$\frac{\partial J}{\partial H(x_i)} = \frac{\partial \sum_k j(y_k, H(x_k))}{\partial H(x_i)} = H(x_i) - y_i$$

Ainsi, à chaque itération, la fonction finale H qui approxime la variable cible est modifiée, en lui ajoutant un élément corrigeant les erreurs effectuées à l'étape précédente :

$$\forall i \in \{1, ..., m\}, H(x_i) := H(x_i) - \frac{\partial J}{\partial H(x_i)}$$

C'est cette fonction H qui sera finalement utilisée, à la fin de toutes les itérations, pour approximer Y. Elle est donc une combinaison linéaire de chacune des fonctions intermédiaires h_i, dans laquelle les fonctions réalisant la plus faible erreur sont sur-pondérées.
Puisque les résidus de chaque étape sont exprimés comme le gradient d'une fonction de coût J, le grand avantage du *gradient boosting* est qu'il peut être utilisé avec n'importe quelle fonction de coût définie au préalable. Comme expliqué précédemment, son choix est primordial et dépend en partie du problème d'apprentissage considéré.

Dans le cas d'un problème de classification binaire comme celui qui se pose ici, celle qui a été choisie est la fonction de perte logarithmique. Elle quantifie le prix payé pour l'inexactitude des prédictions. Plus précisément, elle pénalise les fausses classifications en prenant en compte les probabilités de chacune, exacte ou inexacte. En notant Y la variable

cible comportant n observations (donc pour tout i, y_i vaut 0 ou 1), et p_i la probabilité de prédiction de la $i^{\text{ème}}$ observation sachant y_i, alors :

$$\text{Log Loss}(y_i, p_i) = -\frac{1}{n}\sum_{i=0}^{n}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

L'algorithme utilisé dans le cadre de l'étude est le ***XG Boost***, une version "extrême" du *gradient boosting*. Cette variante apporte deux améliorations par rapport au *gradient boosting* classique :

- elle est implémentée en parallèle et permet donc d'être exécutée plus rapidement ;
- elle permet d'implémenter différents algorithmes sous-jacents plutôt que des arbres de régression comme le *gradient boosting* classique.

Enfin, comme pour les algorithmes précédents, une version pondérée a été implémentée afin de sur-pénaliser les erreurs effectuées sur la classe minoritaire. L'idée est ici de pondérer plus fortement le gradient d'erreur sur les éléments de la classe minoritaire, puisqu'un gradient de valeur plus grande indique une erreur plus coûteuse, incitant donc le modèle à la sur-corriger.

Le lecteur désireux d'avoir plus de détails sur le *Gradient Boosting* (ou sur tout autre modèle d'apprentissage statistique) peut se référer au chapitre 10 (resp. au chapitre approprié) du manuel *The Elements of Statistical Learning* [10].

Par ailleurs, les différents modèles qui viennent d'être présentés peuvent également être adapté pour réaliser des tâches de régression. Ainsi par exemple, les méthodes ensemblistes telles que les forêts aléatoires réalisent une moyenne de toutes les prédictions qui détermine la décision finale, plutôt qu'un vote à la majorité comme c'est le cas pour une tâche de classification. Ces modèles seront donc utilisés dans la méthode 2 de segmentation, sans être détaillés de nouveau.

En résumé : Modèles utilisés pour la sélection de variables

- Régression logistique ;
  - un algorithme de classification binaire très répandu du fait de son fort pouvoir explicatif ;
  - une fonction de coût appropriée : la *log loss*.
- Random Forest ;
  - un assemblage d'arbres de décision indépendants ;
  - une étape de *tree bagging* : tirage aléatoire des observations ;
  - une étape de *feature sampling* : tirage aléatoire des variables explicatives ;
  - des *splits* réalisés selon le critère de Gini.
- Extra Trees ;
  - une variante de *random forest*, rajoutant une composante aléatoire, permettant généralement d'obtenir plus rapidement de meilleurs résultats et d'augmenter la variance du modèle.
- XG Boost ;
  - une version extrême du *gradient boosting*, qui en reprend les principes ;
  - une méthode ensembliste non linéaire très performante ;
  - une construction par itération, chacune visant à corriger les erreurs effectuées à la précédente ;
  - une multitude de fonctions de coût utilisables, dont une particulièrement adaptée au problème de classification binaire : la *log loss*.
- Pour chacun de ces modèles, une version pondérée implémentée pour pallier le déséquilibre des classes dans la variable cible.

Modèles utilisés pour le *clustering*

Une fois les variables sélectionnées, il faut choisir le modèle de classification non supervisé à appliquer à la base de données pour le travail de segmentation. Plusieurs modèles ont cette fois encore été testés (et expliqués ci-dessous), afin de déterminer le plus performant au regard de différents scores.

k-means

L'algorithme des k-moyennes ou *k-means* est un algorithme de **classification**, qui cherche à séparer les individus en k groupes appelés ***clusters*** (k étant précisé en avance) de même variance, en minimisant un critère d'inertie intra-classe.

Concrètement, l'algorithme se déroule en trois étapes :

- Tout d'abord, des centroïdes initiaux sont choisis en tirant aléatoirement k observations de la base de données. Après cette initialisation, une boucle est effectuée

sur les deux autres étapes ;

- La première associe chaque observation à son centroïde le plus proche pour former des *clusters* ;

- La seconde crée de nouveaux centroïdes qui sont les moyennes de toutes les observations associées aux *clusters* précédents. L'écart entre les anciens et les nouveaux centroïdes est calculé, et les deux dernières étapes sont répétées tant que cet écart ne descend pas sous un seuil précédemment défini, ou que le nombre maximum d'itérations défini n'a pas été atteint.

La distance entre chaque observation et un centroïde est calculée en utilisant la distance euclidienne :

$$\text{dist}(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

En théorie, le nombre k de *clusters* déterminés est un hyper-paramètre du modèle, qui peut être sélectionné de manière optimale en utilisant par exemple la règle du coude. Cette technique consiste à représenter sur un graphique la somme des erreurs au carré (où l'erreur pour un point représente l'écart entre ce point et le centroïde associé) pour différentes valeurs de k. Le "meilleur" k sera alors celui pour lequel la diminution de l'erreur n'est pas significative pour des valeurs plus grandes de k (cf. illustration sur la figure ci-dessous). Ce k optimal donne au graphique une apparence de "coude", d'où le nom de la méthode. En pratique, et pour des raisons opérationnelles, le nombre de *clusters* dans le cadre de l'étude sur la valeur client a été fixé au préalable à $k = 3$ (un cluster par type de flux doit être attribué à chaque assuré, ce qui peut conduire à un très grand nombre de combinaisons possibles). Toutefois, en analysant le graphique de la somme des erreurs au carré en fonction de différentes valeurs de k ci-dessous, cette valeur de 3 semble cohérente et pertinente.

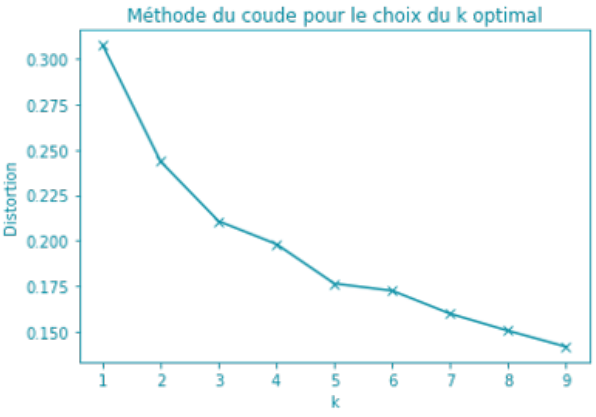

FIGURE 25 – Méthode du coude pour déterminer le k optimal

Avec assez de temps, l'algorithme converge toujours, en un nombre fini d'itérations (pour des raisons mathématiques non approfondies ici), bien que cela puisse être vers un minimum local. Cette convergence dépend fortement de l'initialisation des centroïdes. Par conséquent, l'algorithme est exécuté plusieurs fois, avec différents centroïdes initiaux.

***k-means* pondéré**

Un algorithme de *k-means* classique nécessite une étape de normalisation préalable comme expliqué précédemment dans la partie de *data preprocessing*, qui revient à traiter toutes les variables environ de la même manière. Cependant, la sélection de variables effectuée grâce à l'étape d'apprentissage supervisé a montré que certaines variables étaient plus importantes que d'autres dans la prédiction d'un comportement de rachat partiel (/d'arbitrage /de versement). L'idée est donc d'effectuer une mise à l'échelle des données pour refléter ces différences d'importance de variables (par une méthode de pondération des distances, plusieurs fois citée dans la littérature sur le sujet [6] [21] [17]). Comme expliqué précédemment, il a été possible d'affecter à chaque variable une "importance" chiffrée en fonction de son impact sur les prédictions de comportements. En multipliant les valeurs prises par chaque variable par l'importance correspondante, il est alors possible d'éloigner ou de rapprocher les observations et donc de rendre une variable avec une importance élevée plus discriminante pour la segmentation. En effet, imaginons par exemple que nous disposions de deux variables et que chaque observation soit donc représentée par deux coordonnées. Prenons trois observations $a = (0; 0)$, $b = (0; 1)$ et $c = (1; 0)$. Dans cette configuration, a est équidistant de b et de c. Supposons maintenant que la deuxième variable soit d'importance $0,5$. Alors en multipliant la deuxième coordonnée de chaque observation par $0,5$, on obtient : $a' = (0; 0)$, $b' = (0; 0,5)$ et $c' = (1; 0)$. Dans cette nouvelle configuration, la deuxième variable permet effectivement de rapprocher le point b' de a', qui n'est dans ce cas plus équidistant des deux autres points.

Les deux versions du *k-means* (classique et pondérée) seront donc testées afin d'évaluer l'impact de la sélection de variables et de la pondération.

k*-prototypes : variante du *k-means

L'algorithme de k-prototypes est une variante du *k-means*, pour des données mixtes (à la fois continues et qualitatives). L'étape de dummisation de toutes les données qualitatives n'est donc plus nécessaire, seules les données continues sont normalisées et les variables qualitatives sont utilisées telles quelles.

L'algorithme de k-prototypes est le même que celui du *k-means* classique, seules la définition de la distance entre deux individus et la méthodologie de calcul des nouveaux centroïdes diffèrent.
En effet, le nouveau centroïde correspond au couple formé par la moyenne des observations du *cluster* associé pour les variables quantitatives (comme pour le *k-means* classique) et la modalité la plus représentée dans le *cluster* pour les variables qualitatives. Ce nouveau centroïde "double" est nommé "prototype".

La mesure de distance entre un point x et un prototype p est alors définie comme :

$$\text{dist}(x, p) = E(x, p) + \lambda C(x, p)$$

où :

- $E(x, p)$ représente la distance euclidienne entre les variables continues de x et celles de p ;
- $C(x, p)$ est le nombre de variables qualitatives pour lesquelles la modalité diffère entre le point x et le prototype p ;
- λ est la pondération des variables qualitatives.

Cet algorithme a été utilisé pour la segmentation et comparé aux deux essais de *k-means* classique et pondéré.

k-prototypes pondéré

Tout comme le *k-means* a été décliné en une version pondérée (par les *feature importances*) pour refléter les importances relatives des différentes variables, le k-prototypes a également été testé en version pondérée. Celle-ci consiste à pondérer les variables quantitatives normalisées par leur *feature importance* respective, afin de donner plus de poids à celles jugées comme plus impactantes sur la segmentation des comportements.

En résumé : Modèles utilisés pour le *clustering*

- *k-means* ;
  - un algorithme de *clustering* qui sépare les individus en k groupes, construits par étapes successives ;
  - un choix de $k = 3$ pour des raisons opérationnelles, qui semble juste et cohérent.
- *k-means* pondéré ;
  - une variante du *k-means* classique qui fonctionne de la même manière, avec une étape préliminaire de pondération des variables par leur importance (déterminée lors de l'étape de sélection de variables).
- k-prototypes ;
  - une variante du *k-means*, utilisant une nouvelle mesure de distance, permettant de traiter les variables qualitatives sans les retraiter au préalable.
- k-prototypes pondéré ;
  - une variante du k-prototypes, qui pondère les variables quantitatives selon leur importance relative.

4.1.3 Méthode 2 : prédiction des lois puis segmentation selon les comportements

Cette deuxième méthode de segmentation diffère de la première dans la stratégie adoptée pour construire les *clusters*. Contrairement à la première méthode qui consistait à regrouper les individus de caractéristiques proches parmi les variables déterminées comme les plus impactantes sur le comportement, la présente méthode consiste à entraîner un modèle de prédiction sur les montants de flux observés, afin de prédire ces montants chaque année pour les 30 prochaines années en faisant vieillir le portefeuille pas à pas. Ce modèle de projection permet de déterminer des lois de comportement sur 30 ans, sur la base desquelles la segmentation du portefeuille est ensuite effectuée.
Comme la sélection de variables de la première méthode de segmentation a permis de déterminer que les modèles de *random forest*, *extra trees* et *XG Boost* semblent les plus performants pour étudier le comportement des assurés du portefeuille, ce sont ces modèles qui sont testés en version régression pour prédire les montants de flux.

Sélection de modèle

Afin de déterminer le modèle le plus approprié pour prédire les montants de flux chaque année, deux régresseurs sont testés et comparés. Comme pour la sélection de modèle effectuée précédemment lors de l'étape de classification binaire, il convient d'abord de déterminer les meilleurs hyper-paramètres (par validation croisée *5-folds*), puis de comparer les deux modèles.

L'indicateur classique pour comparer différents modèles de régression est l'erreur quadratique moyenne (***RMSE***, *Root Mean Square Error*), qui représente l'écart-type des résidus (erreurs de prédiction) et mesure donc leur dispersion :

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

avec y_i les valeurs observées de la variable cible et $\hat{y_i}$ celles prédites.

Cependant, cet indicateur peut être très instable et ne doit donc pas être l'unique critère de décision du meilleur modèle. Un deuxième indicateur est donc plutôt considéré, l'erreur absolue moyenne (***MAE***, *Mean Absolute Error*), définie comme :

$$\mathrm{MAE}(y_i, \hat{y_i}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

où y_i et $\hat{y_i}$ sont toujours les valeurs respectivement observées et prédites de la variable cible.

Elle mesure l'ampleur moyenne des erreurs sur l'ensemble des prévisions. Comme l'erreur quadratique moyenne, elle exprime l'erreur moyenne de prédiction dans la même unité que

la variable cible, est positive, et meilleure quand elle est plus faible. Cependant, puisque les erreurs sont mises au carré avant d'appliquer la racine dans l'indicateur de RMSE, ce dernier accorde un poids plus grand aux erreurs de plus grande amplitude.
Concrètement, si deux modèles sont comparés, dont le premier présente beaucoup de faibles erreurs de prédictions, et le deuxième modèle une seule erreur de prédiction mais de très forte amplitude, alors le deuxième aura la même MAE que le premier, mais une RMSE beaucoup plus importante.

En effet la MAE est plutôt stable alors que la RMSE augmente à mesure que la variance associée à la distribution de fréquence des amplitudes d'erreur augmente, et est donc très sensible aux *outliers*. C'est pourquoi l'indicateur utilisé pour la sélection de modèle, sera la **MAE**.

Enfin, un critère d'**explicabilité** est pris en compte dans le choix du modèle. En effet, la valeur client est un indicateur à destination des agents généraux, qui doivent en comprendre le processus de calcul, et à qui les modèles utilisés doivent être clairement expliqués. À ce titre, un modèle de *Random Forest* pourrait être privilégié face à un *XG Boost* aux erreurs similaires par exemple.

Projections des montants de flux pour construire les lois de comportement

Une fois le meilleur modèle déterminé, il est entraîné sur les informations et montants de flux de l'année étudiée. Ensuite, une boucle est itérée sur 30 ans pour construire les lois de comportements prédites :

- le modèle préalablement entraîné est appliqué aux données de l'année N afin de déterminer les flux de $N+1$;
- le portefeuille est vieilli d'une année pour atteindre l'état $N+1$ (chaque variable explicative est vieillie selon un modèle de projection préalablement défini et spécifique à chaque variable, explicité ci-dessous) ;
- puis la boucle est de nouveau parcourue : le modèle de régression est appliqué aux variables explicatives de l'année $N+1$ pour prédire les flux $N+2$;
- de la même manière, le portefeuille est vieilli d'une année en incorporant les flux déterminés pour $N+1$ afin d'obtenir les données de l'année $N+2$;
- et ainsi de suite.

Cette boucle est effectuée de manière à prédire les taux de chaque type de flux sur 30 ans.

Hypothèses de vieillissement du portefeuille

Lors de chaque itération de la boucle, les variables explicatives doivent être vieillie d'une année, afin de pouvoir prédire les montants de flux associés. Différents modèles de projection sur 30 ans de ces variables explicatives ont été mis en place.

Chaque itération de la boucle représente une année supplémentaire. L'**âge** et l'**ancienneté** sont donc mis à jour en augmentant de 1. Une borne est fixée afin de permettre les sorties de portefeuille, comme pour les retraitements effectués initialement sur la base de données. Ainsi, pour les âges et les anciennetés respectivement strictement supérieurs à 110 et 80 ans, les montants de flux ont été considérés comme nuls afin de refléter la sortie du portefeuille, et de même des taux de flux associés.

Compte tenu du contexte actuel, une projection simple et réaliste des **TMG/TMGA** a pu être mise en place. Il a été considéré que les TMGA positifs sur la première année passent tous à 0 les années suivantes, et que les TMG, définis contractuellement (et difficilement modifiables par l'assureur sans intervention de l'organisme de contrôle prudentiel ou du gouvernement) restent inchangés.

Les **taux de PB nets** ont nécessité un modèle de projection légèrement plus complexe. Déterminés à la maille produit, ils sont calculés en fonction des résultats techniques et financiers réalisés chaque année, ainsi que des frais de gestion annuels du support sur lequel les fonds sont investis. Une étape de lecture des conditions générales de chaque produit permet donc d'identifier, pour chacun, le taux de performances financières et de frais de gestion à retenir pour le calcul des taux de PB nets, considérés comme fixes au cours du temps car définis contractuellement. Pour chaque produit, il convient ensuite de retenir le support d'investissement associé, afin de sélectionner la projection des performances financières correspondante. Celle-ci a été fournie pour les dix prochaines années par la direction des investissements (DI), en charge de son calcul, résultant de la calibration de différents scénarios. Pour des raisons opérationnelles, et puisque ce mémoire ne se concentre pas sur la méthodologie de calcul des taux de PB, une prolongation de ces trajectoires sur les années suivantes de projection a été retenue en suivant la même évolution que celle observée sur les trajectoires fournies par la DI, et en bornant les taux à 0%.
Une modélisation plus poussée prenant en compte différents scénarios économiques aurait sans doute été plus précise pour déterminer les performances financières futures des supports considérés, mais l'approche retenue était suffisante et déjà satisfaisante puisque le but final de l'étude est de séparer au mieux les individus. Ainsi, ce sont plus les écarts entre les différents taux de PB que les taux en eux-mêmes qui importent pour l'étude, et la présente méthodologie de projection s'est révélée satisfaisante.

Chaque année, la **PM** est recalculée selon un modèle de projection prenant en compte les flux entrants et sortants. Puisque les montants de flux sont prédits chaque année par contrat (arbitrages, versements, rachats), la PM peut être reconstituée à partir de ces montants et de la PM de l'année précédente. Ainsi, la PM de fin d'année N s'écrit :

$$\begin{aligned} \mathrm{PM}(N) &= \mathrm{PM}(N-1) + \text{Flux entrants}_N - \text{Flux sortants}_N \\ &= \mathrm{PM}(N-1) + \mathrm{VL}(N) - \mathrm{RP}(N) \end{aligned}$$

La revalorisation de l'épargne due au TMG/TMGA et au taux de PB n'a pas été prise en compte car les taux minimum garantis n'ont pas été projetés. Comme pour les performances financières, une projection plus fine des PM prenant en compte le mécanisme de

revalorisation aurait peut être permis d'obtenir des montants de PM plus précis mais la modélisation choisie a déjà permis d'obtenir des montants satisfaisants. Cette approche se justifie d'ailleurs bien dans une période de taux bas voire négatifs donc de faibles taux de revalorisation, dont l'impact sur l'évolution des encours est par conséquent assez faible.
Cette projection des PM se décline pour la partie en euros et la partie en UC de l'encours, et les $\text{PM}_{€}$ et PM_{UC} ont été projetées de la même manière. Ainsi, la **part uc** peut également être recalculée année après année pour chaque contrat, en divisant l'encours projeté en UC par le total de l'encours projeté du contrat.

Les **taux de flux moyens historiques** sont progressivement recalculés, en prenant en compte les montants de flux prédits sur les trois années précédentes. En reprenant la formule explicitée en partie 3, le taux de flux moyen historique sera égal, pour une nouvelle année N à :

$$\text{Taux moy old}_{\text{Flux}}(N) = \frac{\text{Flux}_{N-1} + \text{Flux}_{N-2} + \text{Flux}_{N-3}}{\text{PM}_{N-2} + \text{PM}_{N-3} + \text{PM}_{N-4}}$$

Comme c'était le cas lors de l'étape de retraitements de la base de données, des bornes sont appliquées à ces taux.

Par ailleurs, le **salaire moyen** déterminé initialement grâce aux données externes de l'INSEE évolue selon la trajectoire moyenne observée sur le portefeuille. Sur la première année, un salaire moyen est calculé à la maille CSP/région/tranche d'âge/tranche d'ancienneté/tranche de PM, puis, chaque année suivante, le salaire moyen de chaque individu est déterminé par celui calculé sur la première année, pour la maille CSP/région/tranche d'âge/tranche d'ancienneté/tranche de PM correspondante.

Afin de déterminer la trajectoire moyenne de la **situation familiale** sur la base de données, des statistiques ont été réalisées par région et âge. Il en est ressorti que la situation la plus fréquente pour chaque tranche d'âge est la même d'une région à une autre. Une recherche plus fine a donc été réalisée, en isolant la situation la plus fréquente à chaque âge, indépendamment de la région de résidence.
Une règle d'évolution a alors pu être dégagée selon l'âge de l'épargnant :

- de 0 à 42 ans : "célibataire" (sauf si déjà marié, dans ce cas : reste "marié") ;
- de 43 à 94 ans : "marié" (sauf si déjà "divorcé" ou "séparé" ou "veuf", dans ce cas : reste dans la situation précédente)
- de 95 ans à 110 ans : "veuf".

Enfin, certaines variables ont été considérées comme fixes sur les 30 prochaines années : le département (et donc la **région**) de résidence de l'épargnant, ainsi que de son agence, le **type de cotisation**, le **code fiscal** (inchangé au cours de la vie d'un contrat), le **Top Bonus Euro+**, le **Top Fourgous** et le **type de gestion**. Il en a résulté que le **Top éloignement agent** est aussi resté inchangé (puisque les départements du souscripteur et de l'agent le sont restés).
Ces hypothèses de stabilité des variables explicatives semblent faibles et donc raisonnables.

Segmentation selon les comportements et *clustering* de séries temporelles

Une fois les lois de comportements ainsi déterminées sur 30 ans pour chaque contrat de la base de données et pour chaque type de flux, un algorithme de segmentation est utilisé afin de regrouper les épargnants aux comportements les plus proches.

Cette approche consiste à considérer les lois définies par le modèle précédent comme des séries temporelles. Ainsi, à chaque client et pour chaque type de flux, une série temporelle est affectée, qui représente son comportement prédit vis-à-vis de ce type de flux pour les années à venir. L'objectif dans cette partie est de rapprocher les clients dont le comportement prédit est proche, donc d'effectuer une classification sur les séries temporelles.

Pour regrouper les séries temporelles proches, les techniques classiques de classification déjà présentées précédemment pourraient être utilisées mais ne tiendraient pas compte des spécificités liées aux séries temporelles (différences d'échelle, décalage temporels, etc.). L'idée est donc d'adapter ces algorithmes usuels pour qu'ils soient capables de traiter correctement des séries temporelles.

Comme le précisent différents auteurs ([18], [20]), plusieurs adaptations pertinentes existent pour traiter des séries temporelles :

- le ***clustering* à partir des séries brutes** ("*temporal-proximity based clustering*") consiste à appliquer les méthodes usuelles directement aux séries temporelles, mais en adaptant la mesure de distance utilisée. Par exemple, la distance DTW (*Dynamic Time Warping*) distort l'axe du temps en considérant un temps non-linéaire afin de permettre le rapprochement de deux séries décalés dans le temps mais pourtant similaires ;

- le ***clustering* à partir de la modélisation des séries** ("*model-based clustering*") repose sur l'application d'un modèle aux séries brutes (de type ARMA par exemple) et le rapprochement ensuite de séries de modèle similaire ;

- le ***clustering* à partir de métriques d'évaluation des séries** ("*characteristic-based clustering*") consiste à extraire des indicateurs à partir des séries temporelles brutes qui permettent de les qualifier, puis à appliquer un algorithme classique de *clustering* pour regrouper les séries d'indicateurs similaires.

L'approche par "*temporal-proximity based clustering*" utilisant la distance DTW a été considérée dans les premiers moments de construction de la méthode de segmentation, mais n'a pas été retenue. Comme l'indiquent Wang, Smith-Miles et Hyndman [20], cette méthode offre une pauvre performance pour traiter des séries assez longues puisque la notion même de la similarité est douteuse dans un espace de grande dimension, d'autant plus que les projections sur 30 ans selon la méthode adoptée peuvent accumuler des erreurs dans les montants prédits au fur et à mesure du temps.

Dans ce contexte, c'est la troisième approche qui a été privilégiée. Ainsi, à partir des lois construites grâce à la modélisation précédente, des **indicateurs** statistiques sont extraits

et associés aux numéros de contrats correspondants :

- taux moyen de flux sur les 30 années de projection ;
- taux moyen de flux sur les 10 premières années de projection ;
- taux moyen de flux sur les 10 dernières années de projection ;
- variance sur les 30 années de projection ;
- pourcentage de valeurs de la série qui sont supérieures à la moyenne de la série (donc au taux moyen sur les 30 années de projection) ;
- moyenne des différences absolues entre les valeurs des séries ultérieures, c'est à dire : $\frac{1}{n-1}\sum_{i=1}^{n}|x_{i+1}-x_i|$

Le premier indicateur a été choisi afin de refléter la tendance générale de comportement de l'épargnant sur les 30 années de projection. Les deux suivants servent à refléter la tendance qui se dégage respectivement en début et en fin de projection. La variance sur les 30 années permet de quantifier la stabilité dans le comportement de l'épargnant, ou au contraire son absence de stabilité. Le quatrième indicateur a été choisi afin de repérer les clients qui auraient un comportement très agressif sur une petite période (et qui auraient donc une très faible part de leur loi qui comporterait des taux supérieurs, et de beaucoup, à leur moyenne sur 30 ans). Enfin, le dernier indicateur permet de mesurer les grands écarts de comportement qui pourraient exister d'une année sur l'autre dans la loi prédite.

Ensuite, un algorithme classique de type *k-means* est appliqué à la nouvelle base ainsi crée, afin de segmenter le portefeuille.

En résumé : Méthode 2 de segmentation

- Une sélection de modèle pour la prédiction des montants de flux, parmi le *random forest*, l'*extra trees* et le *XG Boost*, selon le critère de la MAE ;
- Un vieillissement du portefeuille sur 30 ans en prenant des hypothèses de projection des variables explicatives ;
- Une segmentation du portefeuille par un *clustering* de séries temporelles basé sur des indicateurs extraits des lois de comportement prédites.

4.2 Comparaison des résultats et choix de la méthode de segmentation

Dans le but de sélectionner la méthode la plus performante, les modèles présentés dans la partie précédente ont été testés et implémentés au travers d'algorithmes développés en langage Python, et leurs résultats comparés.

Résultats et sélection de modèle pour la méthode 1 de segmentation

Afin de déterminer le meilleur modèle de classification binaire permettant de prédire la survenance de flux et d'en déduire les variables les plus impactantes, différents modèles (présentés dans la partie précédente) ont été testés sur l'ensemble de la base de données. Les résultats présentés ci-dessous et dans les parties suivantes correspondent à la prédiction

du rachat partiel, les mêmes modèles ayant été appliqués aux deux autres types de flux. Après une phase de détermination des hyper-paramètres par validation croisée pour chacun, les scores de performance ont pu être obtenus et résumés dans le tableau ci-dessous :

	AUC	Classement AUC	Accuracy	Precision	Recall	F1	Classement F1	Nombre de Faux Négatifs	Classement nombre de faux négatifs	Explicabilité	Classement meilleur modèle
Régression Logistique	82,58%	5	93,74%	45,80%	2,52%	4,78%	8	25926	8	✓	6
Régression Logistique pondérée	81,51%	6	86,04%	24,78%	60,88%	35,22%	4	10405	2	✓	4
Random Forest	83,58%	4	93,71%	48,24%	11,25%	18,25%	7	23604	7	✓	5
Random Forest pondérée	85,52%	1	90,79%	34,70%	53,99%	42,24%	1	12237	4	✓	1
Extra Trees	80,54%	7	93,14%	36,88%	13,99%	20,29%	6	22875	6	✓	8
Extra Trees pondéré	83,77%	3	86,68%	25,95%	61,27%	36,45%	2	10301	1	✓	2
XG Boost	80,03%	8	92,79%	35,49%	19,19%	24,91%	5	21492	5		6 bis
XG Boost pondéré	83,82%	2	86,53%	25,24%	59,13%	35,38%	3	10870	3		2 bis

FIGURE 26 – Sélection de modèles et de variables pour la Segmentation 1

Pour chaque score, un classement est donné, puis un classement final permettant de déterminer le meilleur modèle suivant les critères défini a été réalisé.

Comme expliqué dans la présentation de la méthode 1 de segmentation, les scores statistiques retenus pour comparer les modèles sont l'**AUC** et le **F1 Score**. Les scores d'AUC sont globalement très bon pour l'ensemble des modèles testés. Le F1 Score, qui donne une meilleure mesure des individus incorrectement classifiés, est plus volatile selon les modèles. Selon les deux critères, le modèle de *random forest* pondéré semble le meilleur.
Un troisième indicateur a été observé pour mieux comprendre la volatilité dans le F1 Score. Il s'agit du **nombre de faux négatifs** prédits par chaque modèle, obtenu grâce à la matrice de confusion. En prenant l'exemple du rachat partiel, il s'agit du nombre d'individus pour lesquels le modèle prédit qu'ils n'effectueront pas de rachat partiel sur l'année étudiée, alors qu'ils en ont en réalité effectué au moins un. Ce type d'erreur de classification est le plus grave que les modèles peuvent réaliser (puisqu'ils conduisent à sous-estimer le risque encouru par la compagnie), et c'est pourquoi le nombre de faux négatifs a donc également été étudié.
Selon ce critère, c'est l'*extra trees* pondéré qui semble se tromper le moins. De plus, cet indicateur permet de visualiser l'intérêt d'utiliser des modèles pondérés, qui visent à pénaliser les erreurs effectuées sur la classe minoritaire. En effet, parmi les huit modèles testés, ce sont ceux qui ont recours à cette pondération qui permettent d'obtenir les plus faibles nombres de faux négatifs (environ 11 000 en moyenne, contre environ 23 000 en moyenne pour les modèles en version classique). Cette méthode de pondération semble donc être une alternative pertinente à celles de ré-échantillonnage afin d'entraîner des modèles et d'effectuer des prédictions sur des bases de données déséquilibrées.

Enfin, il convient de se rappeler que la segmentation doit permettre *in fine* d'attribuer à chaque client un indice qui représente sa rentabilité prospective pour la compagnie. Cet indice sera donc utilisé par les agents généraux, qui doivent comprendre précisément la construction de l'indice, et pouvoir expliquer comment le portefeuille a été segmenté et

pourquoi un client en particulier a reçu un indice précis.
Dans cette optique, les modèles utilisés doivent être facilement explicables et compréhensibles, même par des acteurs qui n'ont que très peu voire aucune connaissance en *machine learning*. Ainsi, un critère d'explicabilité a été pris en compte dans le choix du modèle, qui avantage ceux facilement explicables.

Finalement, en prenant en compte tous ces critères, c'est le modèle de ***random forest* pondéré** qui a été sélectionné comme meilleur modèle de prédiction de survenance de flux.

L'étape suivante a consisté à s'intéresser aux variables jugées comme les plus impactantes lors du travail de prédiction réalisé par ce modèle. Pour ce faire, les *feature importances* associées à ce modèles ont été étudiées :

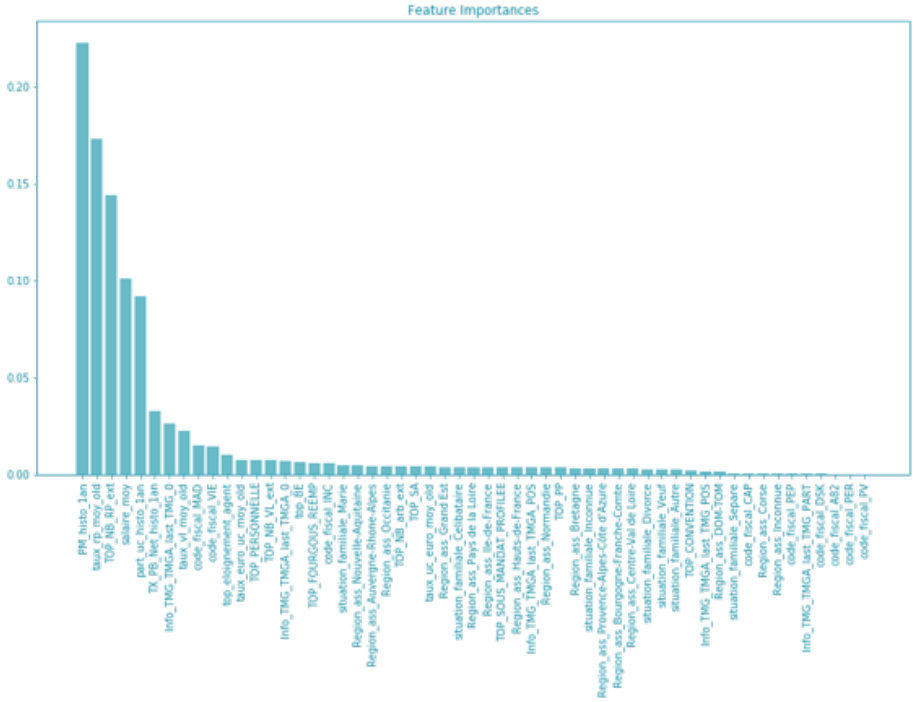

FIGURE 27 – *Feature importances* du modèle de *random forest* pondéré pour la survenance

Ainsi, les dix variables les plus impactantes pour prédire la survenance d'un rachat partiel sont, dans l'ordre :

- la PM de fin d'année précédente ;
- le taux moyen historique de rachat partiel ;
- l'indicatrice qui précise si le client a effectué un nombre de rachat supérieur à celui de 90% du portefeuille ;
- le salaire moyen ;
- la part UC de fin d'année précédente ;
- le taux de PB Net ;

- l'information du TMG/TMGA ;
- le taux moyen historique de versement libre ;
- le code fiscal.

Les variables qui ressortent comme les plus impactantes sont donc celles qui semblent les plus cohérentes vis-à-vis du but de la prédiction. En effet, il semble intuitif que la PM et le comportement passé des épargnants reflètent bien leur comportement futur par exemple. Aussi, il convient de remarquer la présence de la PM et de la part UC dans cette liste, qui étaient les deux critères d'attribution de la valeur épargne selon l'ancienne grille de décision, ce qui confirme que l'ancienne grille n'était pas totalement aberrante bien que non satisfaisante.

Résultats et sélection de modèle pour la méthode 2 de segmentation

Puisque la méthode 2 de segmentation fait intervenir une phase de prédiction pas à pas des montants de flux sur 30 ans avant la segmentation, une étape de sélection du modèle de régression a été effectuée.

Afin de déterminer le meilleur modèle de régression permettant de prédire les montants de flux, plusieurs ont été entraînés et comparés sur la base entière, avec les montants de rachats partiels. Les résultats suivants ont alors été obtenus :

	MAE	Classement MAE	Explicabilité	Classement meilleur modèle
Random Forest	1629,24	3	✓	3
Extra Trees	1597,74	2	✓	2
XG Boost	1384,35	1		1

FIGURE 28 – Sélection de modèles et de variables pour la Segmentation 2

C'est finalement le ***XG Boost*** qui a été sélectionné (malgré un travail plus complexe d'explicabilité sûrement nécessaire a posteriori), car il permet d'effectuer l'erreur moyenne absolue la plus faible sur les montants prédits.

Comme pour les modèles de classification, il est possible d'étudier les *feature importances* :

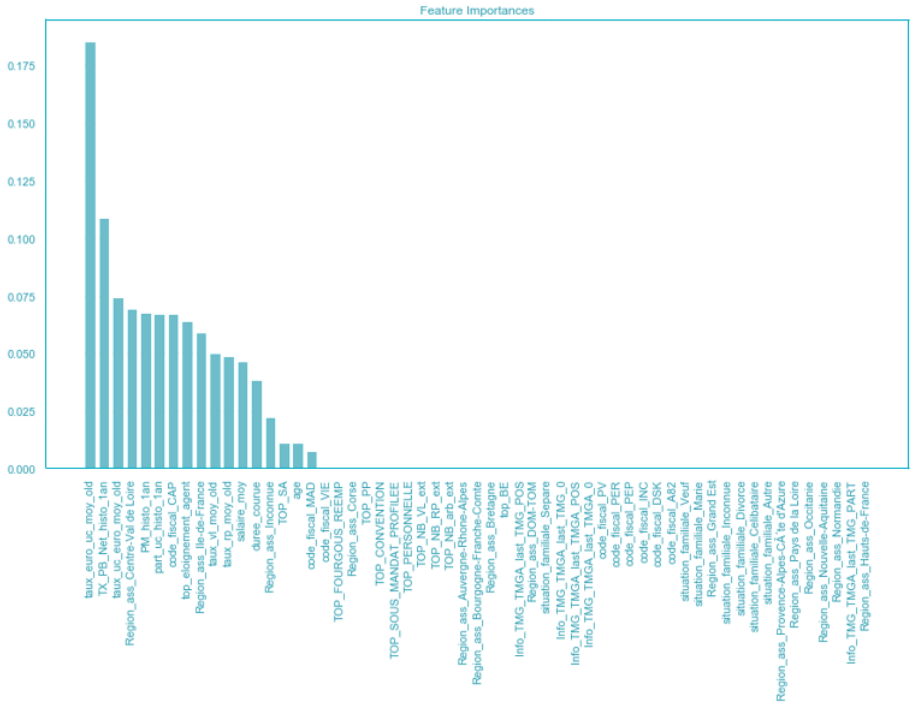

FIGURE 29 – *Feature importances* du modèle de *XG Boost* pour le montant

Les variables les plus importantes pour prédire les montants de rachats partiels sont alors :

- le taux d'arbitrage euro-UC moyen historique ;
- le taux de PB net ;
- le taux d'arbitrage UC-euro moyen historique ;
- la région de l'épargnant ;
- la PM de fin d'année précédente ;
- la part UC de fin d'année précédente ;
- le code fiscal ;
- l'indicatrice qui précise si le souscripteur réside dans un département différent de celui de son agence ;
- le taux moyen historique de versement libre ;
- le taux moyen historique de rachat partiel.

Grâce à ce modèle de régression, des lois de comportement sont calculées pas à pas, puis une un algorithme de *clustering* est appliqué sur les comportements prédits (indicateurs extraits des lois), conformément à la méthodologie de la segmentation 2.

Une fois les meilleurs modèles déterminés pour chaque méthode de segmentation, ils peuvent être mis en place afin de créer les différents *clusters* qui partitionnent le portefeuille, et les deux méthodes peuvent être comparée pour que la meilleure soit retenue.

Comparaison des deux méthodes de segmentation et sélection de la meilleure

Après avoir segmenté le portefeuille de clients selon les deux méthodes expliquées précédemment, et avec les modèles jugés les meilleurs lors de la phase de sélection, la difficulté réside dans la comparaison de ces deux méthodes pour déterminer la meilleure.

En effet, la seule mesure classique de distance intra- et inter-classes ne suffit pas car elle ne permet pas totalement de juger de la qualité de notre segmentation, dont l'objectif principal est de séparer au mieux les individus aux comportements différents, et de rapprocher ceux aux comportements similaires.
Un autre objectif à garder en mémoire est l'absence de grand déséquilibre de poids entre les différents *clusters* : une méthode de segmentation qui aboutirait à la construction de trois *clusters* dont un comporte seulement 2% des individus par exemple, et dont les deux autres partageraient équitablement les clients restants, ne pourrait pas être jugée comme satisfaisante.
Enfin, il convient de vérifier grâce à un *backtesting* que les classements effectués ne sont pas aberrants au regard des montants de flux réellement observés.

Pour remplir ces objectifs, plusieurs indicateurs ont été créés, qui s'appuient sur les lois de comportement calculées pour chaque Réseau-groupe de produits / *Cluster*. (Dans toute la suite, on notera "RESPROD" la maille de calcul "Réseau-groupe de produits". Elle rassemble tous les contrats distribués sur le même réseau, et correspondant à un même groupe de produits).
Une fois les *clusters* affectés à chaque client, des lois de comportement sont donc construites comme défini précédemment (et comme expliqué en détail dans la partie 5) sur la base des flux passés avec un historique de 3 ans, pour pouvoir calculer les indicateurs appropriés.

Tout d'abord, une **règle métier** qualitative permet une première sélection : si une méthode de segmentation aboutit à la construction de lois aberrantes pour au moins deux *clusters* (vis-à-vis de repères tels que les lois utilisées par le modèle interne par exemple), alors la méthode est écartée d'emblée.

Ensuite, l'**écart de comportement entre *clusters*** ("EC") permet de quantifier la séparation des comportements effectuée entre les différents *clusters*. Un premier écart de comportement entre les différents *clusters* d'un même RESPROD est d'abord calculé :

$$\mathrm{EC}_{\mathrm{RESPROD}} = \frac{1}{60}\sum_{n=1}^{60} \frac{|\tau_{n,0} - \tau_{n,1}| + |\tau_{n,1} - \tau_{n,2}| + |\tau_{n,0} - \tau_{n,2}|}{3}$$

où $\tau_{n,k}$ représente le taux de flux prédit pour l'année n et le *cluster* k.

Puis, l'écart moyen total de comportement est calculé comme la moyenne des écarts de comportement sur tous les RESPROD :

$$\mathrm{EC} = \frac{1}{\text{nombre de RESPROD}} \sum_{i=1}^{\text{nombre de RESPROD}} \mathrm{EC}_{\mathrm{RESPROD}}$$

Ainsi, un indicateur EC trop faible signifierait que les *clusters* construits ne séparent pas assez les comportements, puisque les taux de flux seraient trop proches d'un *cluster* à l'autre. Cependant, un EC beaucoup trop élevé pourrait aussi signifier qu'un des *clusters* présente des taux prédits qui "explosent", ce qui peut survenir pour des cas non validés par la règle métier par exemple.

Enfin, l'**écart de prédiction** ("EP") doit permettre de vérifier que le *cluster* associé à un client, et donc les taux qui y sont liés, ne sont pas aberrants avec les taux réels observés. Pour ce faire, une étape de *backtesting* a été réalisée : le taux de flux prédit pour 2019, associé à un *cluster* a été comparé au taux réel de flux observé en 2019 sur ce *cluster*. Ainsi, on définit l'écart de prédiction pour un RESPROD comme :

$$\text{EP}_{\text{RESPROD i}} = \frac{\sum_{j=0}^{2} |mr_{i,j} - mp_{i,j}|}{\sum_{j=0}^{2} \text{PM 1812}_{i,j}}$$

avec :

- $mr_{i,j}$ le montant réel de flux observé sur le *cluster* j du RESPROD i en 2019 ;
- $mp_{i,j}$ le montant prédit de flux sur le *cluster* j du RESPROD i pour 2019 ;
- $\text{PM 1812}_{i,j}$ la PM de fin d'année 2018 du *cluster* j du RESPROD i.

Comme pour l'EC, c'est l'erreur moyenne sur l'ensemble des RESPROD qui est finalement observée :

$$\text{EP} = \frac{1}{\text{nombre de RESPROD}} \sum_{i=1}^{\text{nombre de RESPROD}} \text{EP}_{\text{RESPROD}}$$

Puisque cet indicateur mesure une erreur sur le taux prédit vis-à-vis du taux réellement observé, alors plus il est faible, meilleure sera la méthode de segmentation.

Ces indicateurs peuvent être calculés pour l'ensemble des différents flux. Pour illustrer le choix de la meilleure méthode, ils ont ici été calculés sur les lois de rachats partiels :

	Règle métier	EC	Classement EC	EP	Classement EP	Classement meilleure méthode
Segmentation 1 - k-means classique	✗	8,86 %	2	1,51 %	4	✗
Segmentation 1 - k-means pondéré	✓	1,11 %	4	0,79 %	1	1
Segmentation 1 - k-prototypes	✗	9,04 %	1	1,49 %	3	✗
Segmentation 1 - k-prototypes pondéré	✓	0,96 %	5	0,81 %	2	2
Segmentation 2	✗	7,54 %	3	1,85 %	5	✗

FIGURE 30 – Comparaison des différentes méthodes de segmentation

Trois méthodes ont pu être éliminées d'office car les lois qui en découlaient ne respectaient pas les critères métiers. En effet, elles présentaient des lois de rachats partiels qui ne cor-

respondaient pas aux lois habituellement utilisées (par exemple, un taux de 95% pour le premier *cluster* ne comportant que deux contrats, 21% pour le deuxième qui n'en comporte qu'une cinquantaine, et 6% pour le troisième *cluster*, rassemblant tout le reste du portefeuille). Comme on pouvait s'y attendre, ce sont également ces méthodes qui présentaient les plus forts taux d'écart de comportement (EC) et d'écarts de prédiction (EP).
Finalement, la meilleure méthode qui a été sélectionnée a donc été la **segmentation 1** avec le modèle sous-jacent de ***kmeans* pondéré**. C'est celle, parmi les deux méthodes qui restaient à comparer, qui sépare au mieux les comportements entre les différents *clusters* et qui commet la plus faible erreur vis-à-vis des taux de flux affectés.

En résumé : Comparaison des résultats et choix de la méthode de segmentation

- Un modèle retenu pour la sélection de variables de la méthode 1 : le *random forest* pondéré ;
- Un autre modèle retenu pour la projection des montants de flux de la méthode 2 : le *XG Boost* ;
- Une comparaison des méthodes de segmentation qui a permis de sélectionner la meilleure : méthode 1 avec *k-means* pondéré.

5 Calcul de la valeur épargne, interprétation et déploiement

5.1 Calcul de la valeur épargne

Dans cette partie, le calcul de la valeur épargne est détaillé, élément par élément. Pour rappel, la formule retenue est la suivante :

$$\begin{matrix}\text{Valeur}\\ \text{épargne}\end{matrix} = (\text{PM } € \times \text{Coeff. } € + \text{PM UC} \times \text{Coeff. UC}) \times \begin{matrix}\text{Duration}\\ \text{contrat}\end{matrix} - \begin{matrix}\text{Frais}\\ \text{d'acquisition}\end{matrix}$$

avec :

$$\text{Coeff.} = \frac{\text{VIF}_{stoch} - \text{TVFOG} - \text{MVM} - \text{VAN FG}}{\text{PM} \times \text{Duration}}$$

Afin de déterminer les coefficients Euro et UC, les lois de comportements pour chaque type de flux ont été calculées.

Construction des lois de comportement

Les lois de comportement sont utilisées dans les modules de calculs de coefficients (€ et UC), qui permettent ensuite de calculer la valeur de chaque contrat. Ainsi, pour chaque contrat, des lois de comportement sur 60 ans pour les trois types de flux (arbitrages, rachats et versements) doivent être déterminées.
Elles correspondent à des suites de taux du flux concerné, exprimé en pourcentage de la PM correspondante, et sont donc calculées, pour chaque année t, en divisant le montant de flux estimé pour l'année t par la PM de fin d'année $t-1$. Ainsi par exemple, une loi de comportement sur les rachats partiels donne, pour tout $t \in \{1, ..., 30\}$, le taux de rachats partiels de l'année t : $\frac{\text{RP}_t}{\text{PM}_{t-1}}$.

Dans les deux méthodes de segmentation, un ***cluster*** pour chaque type de flux est affecté à tous les assurés, puis les lois de comportement sont ensuite déterminées sur cette base segmentée.
Ces lois se basent sur une **hypothèse forte** : le présent est représentatif du futur. La méthodologie employée suppose donc qu'un individu d'âge x et d'ancienneté a aujourd'hui se comportera dans n années comme un individu qui a aujourd'hui $x+n$ ans et $a+n$ années d'ancienneté. Alors, grâce à l'observation des taux (d'arbitrages / de rachats / de versements) de l'année d'extraction des données, les lois de comportements de chaque client peuvent être reconstruites sur 60 ans, dans la limite d'un âge de 100 ans et d'une ancienneté de 80 années.
Les lois de comportement de chaque assuré labellisé par un *cluster*, sur les 60 prochaines années, peuvent alors être déterminées.

Notons que le calcul de duration nécessite de connaître les **lois de rachats totaux**. Ces lois ne sont pas calculables directement selon le même procédé à partir de la base segmentée, puisque les clients ayant effectué un rachat total en 2019 ne se trouvent plus dans les bases de données lors de l'extraction de fin 2019.
Ces lois de rachats totaux sont donc calculées séparément et de manière à correspondre à la

méthodologie employée par les modélisations du Modèle Interne. Le principe se rapproche de la celui suivi pour calculer les lois de comportement des autres flux, mais sans prendre en compte une différence de *cluster*. Ainsi, pour chaque *RESPROD*, un taux est déterminé en fonction des montants de rachats totaux et de PM observés, pour chaque tranche d'âge et tranche d'ancienneté. En reprenant l'hypothèse utilisée pour calculer les autres lois de flux (invariance dans le temps du comportement pour un âge et une ancienneté donnés), il est alors possible de construire les lois de rachats totaux selon la même méthodologie, mais sans la disctinction du *cluster*.

Enfin, il convient de noter que la méthodologie générale de construction des lois de comportement expliquée ici est également suivie pour la méthode de segmentation 2. En effet, les lois brutes déterminées par projection des montants de flux sont nécessaires pour effectuer la segmentation du portefeuille, mais pas utilisées telles quelles pour la détermination finale des lois de comportement (incertitude sur les montants prédits et obligation de cohérence avec la méthodologie employée par le Modèle Interne).

Calcul de la VIF

Les principes techniques ont été présentés en première partie. Calculée en environnement risque neutre, la VIF est la somme des flux futurs de résultats statutaires actualisés. C'est la valeur du portefeuille une fois le produit lancé (à partir de l'année 1) sur une période de 60 ans.

Elle est calculée de deux manières :

- VIF déterministe : obtenue en utilisant le scénario central CE (Certainty Equivalent) ;
- VIF stochastique : déterminée comme la moyenne des VIFs calculées sur 4000 scénarios stochastiques risque neutre.

On appelle alors *Time Value of Options and Guarantees* (TVOG), l'écart entre les deux.

Pour le calcul de la valeur épargne, une estimation de la VIF stochastique à la maille RESPROD / *Clusters* est nécessaire afin de déterminer les coefficients. La référence retenue pour le choix du taux d'actualisation est la courbe des taux risque neutre transmise par la direction des investissements.

Plusieurs éléments sont intégrés dans le calcul de la VIF :

- la marge financière ;
- la rétrocession AXA Investment Managers ;
- la marge d'acquisition.

Ces éléments sont basés sur les données de la segmentation (à la maille Réseau/Groupe de produits/Classe de comportement) pour les hypothèses de rachats et versements ainsi que pour les niveaux d'encours, et sur des hypothèses EEV (*European Embedded Value*), utilisée par le Modèle Interne. Ils nécessitent donc d'avoir au préalable établi les lois de

comportements pour chacun des types de flux (arbitrages, rachats et versements), pour chaque réseau/groupe de produits/classe de comportement.

Calcul de la duration contrat

La duration des contrats peut être interprétée comme une **espérance de vie résiduelle en portefeuille**.
Elle représente la durée moyenne qu'il reste à chaque contrat avant sa sortie du portefeuille, qui peut avoir deux causes possibles : rachat total du contrat ou décès. L'hypothèse est faite à ce stade que le contrat n'a pas de date d'échéance prédéterminée.
Le calcul de la duration contrat s'appuie donc sur deux données principales : les **lois de rachats totaux** par *RESPROD* et la **table de mortalité** adaptée. Ici, le choix a été fait d'utiliser une table de mortalité d'expérience représentant le portefeuille d'AXA et certifiée par Mr. Planchet.

Dans un premier temps, les **probabilités de décès** sont calculées à partir de la table de mortalité choisie, qui contient pour chaque âge x le nombre de survivants à cet âge, noté l_x (pour x allant de 0 à 119 ici). Ainsi, pour une personne d'âge x, la probabilité de décès dans l'année, q_x, s'écrit comme le rapport entre le nombre de décès annuel et le nombre de survivants à cet âge :

$$\mathbb{P}_x(\text{"Décès dans l'année"}) = q_x = \frac{l_x - l_{x+1}}{l_x}$$

Ensuite, il faut utiliser les **lois de rachat** par Réseau/Produit ("*RESPROD*"), en les aplatissant sur les dernières années : les lois sont initialement calculées sur un horizon de projection de 60 ans, il faut ici considérer que de 60 à 120 ans, le taux de rachat est constant, égal à celui de la $60^{\text{ème}}$ année de projection.

Pour chaque *RESPROD*, une matrice est alors calculée, avec en lignes le nombre d'années d'ancienneté et en colonnes l'âge. Le terme (i, j) de cette matrice de duration donne alors pour le *RESPROD* considéré, la duration restante pour le contrat d'un assuré d'ancienneté i et d'âge j :

$$\text{Duration(i,j)} = \sum_{k=1}^{+\infty} k \times \overbrace{\mathbb{P}(\text{"l'individu n'est pas décédé et n'a pas racheté jusqu'en } k\text{"})}^{\mathbb{P}_1} \times \underbrace{\mathbb{P}(\text{"l'individu quitte le portefeuille en } k\text{"})}_{\mathbb{P}_2}$$

avec :

$$\begin{aligned}\mathbb{P}_1 &= \prod_{l=0}^{k-1} (1 - \mathbb{P}_j(\text{"Décès à } j+l \text{ ans"})) \times (1 - \mathbb{P}_i(\text{"Rachat à } i+l \text{ années d'ancienneté"})) \\ &= \prod_{l=0}^{k-1} (1 - q_{j+l}) \times (1 - \text{"taux de rachat en } i+l\text{"})\end{aligned}$$

et :

$$\begin{aligned}\mathbb{P}_2 &= 1 - (1 - \mathbb{P}_j(\text{"Décès à } j+k \text{ ans"})) \times (1 - \mathbb{P}_i(\text{"Rachat à } i+k \text{ années d'ancienneté"})) \\ &= 1 - (1 - q_{j+k}) \times (1 - \text{"taux de rachat en } i+k\text{"})\end{aligned}$$

Des hypothèses sur l'âge et l'ancienneté qu'il est possible d'atteindre au maximum sont prises. Ainsi : $q_x = 1$ pour tout $x \geq 110$, et le taux de rachat est égal à 1 pour toute ancienneté $a \geq 80$.

Grâce à ces matrices, une **duration restante** peut donc être affectée à chaque individu de la base client segmentée, grâce à la donnée de son *RESPROD*, de son âge et de son ancienneté.

Modélisation des frais d'acquisition

Le calcul de rentabilité des produits prend en compte deux éléments principaux : la **marge** réalisée par l'assureur (avant prise en compte des frais), et les **frais** associés.
Les **frais de gestion** sont compris dans le calcul de la **VIF** comme expliqué précédemment.
Les **frais d'acquisition** sont calculés pour chaque contrat via un **forfait** de frais d'acquisition. La masse des frais d'acquisition représente la masse de frais de l'année constatés par les services de comptabilité. Cette masse (calculée à la maille produit) est ensuite ramenée au nombre de contrats des affaires nouvelles et des contrats ayant effectué des versements libres, pour obtenir un forfait d'acquisition qui sera appliqué à l'intégralité du portefeuille, par produit.

La particularité des produits d'assurance vie se traduit par un coût de démarrage très élevé. Au fil de la vie d'un tel contrat, il existe toujours un certain coût de maintenance, aussi appelé coût de "gestion", mais qui est beaucoup moins élevée que le coût dit "d'acquisition" au départ. Ainsi, les produits d'assurance vie sont **rentables** pour l'assureur sur le **long terme** grâce aux frais prélevés qui absorbent progressivement la dette constituée initialement. Plus l'assuré reste longtemps en portefeuille, plus ce coût initial est absorbé, et plus l'assureur gagne en marge.

Ce phénomène est délicat à manipuler et à modéliser dans le cadre d'une étude comme l'indice client, qui agrège des valeurs de contrats épargne avec des valeurs de contrats **IARD**, qui ne fonctionnent pas forcément de la même manière.
Aussi, les frais d'acquisition sont comptabilisés dans la valeur épargne après calcul des coefficients et de la duration grâce à la base segmentée. Il convient donc de porter une attention particulière à la modélisation des frais d'acquisition et à leur incorporation dans le calcul de la valeur finale, afin de ne pas détériorer le travail d'amélioration de la valeur effectué grâce à la nouvelle segmentation du portefeuille.

Une première solution consiste à **lisser** les frais d'acquisition sur la duration moyenne du produit pour les produits en cours, et sur 15 ans pour les produits en run-off, et à plafonner ce lissage à la durée évoquée.
Ainsi, la duration de chaque contrat est calculée et si elle est inférieure au plafond correspondant, une portion des frais d'acquisition est retranchée à la valeur épargne, qui

correspond à l'**amortissement** de ces frais sur la durée correspondante. Si la duration est supérieure au plafond considéré, alors aucune portion de ces frais n'est retranchée. Pour ce faire, les frais d'acquisition finalement retranchés à un contrat correspondent au forfait de frais d'acquisition (déterminé en amont par type de produit), multiplié par le ratio de la duration moyenne (ou 15 ans pour les produits en *run-off*) moins l'ancienneté, sur la duration moyenne (ou 15 ans le cas échéant) :

$$\text{Frais d'acquisition} = \begin{cases} \max\left(\text{forfait frais d'acquisition} \times \frac{\text{duration moyenne} - \text{ancienneté}}{\text{duration moyenne}}, 0\right) & \text{Produits en cours} \\ \max\left(\text{forfait frais d'acquisition} \times \frac{15 - \text{ancienneté}}{15}, 0\right) & \text{Produits en run-off} \end{cases}$$

De cette manière, le forfait de frais d'acquisition n'est en fait pas appliqué à l'intégralité du portefeuille mais uniquement aux contrats ayant une duration inférieure à la duration moyenne du produit considéré (ou 15 ans pour les produits en *run-off*). L'idée de ce plafonnement du lissage des frais d'acquisition est de ne pas désavantager les clients de grande duration (donc les clients fidèles), dont la valeur épargne ne se voit plus diminuée de frais d'acquisition.

Cette approche de lissage des frais permet d'éviter une trop grande différence de valeur d'un contrat entre sa première année en portefeuille et les suivantes, qui serait constatée si les frais d'acquisition n'étaient comptabilisés que sur la première année. Une telle différence pourrait faire passer la valeur d'un contrat de négative la première année à positive les années suivantes (ce qui aurait pour conséquence, par exemple, d'inciter les agents à ne réaliser aucune affaire nouvelle, au risque de détériorer la valeur de leur portefeuille). Ainsi, en amortissant les frais d'acquisition sur la duration moyenne (ou 15 ans pour les produits en run-off) plutôt que de ne les appliquer que sur la première année, cet écueil est évité.

Un problème apparaît alors : si les frais d'acquisition sont amortis sur la duration moyenne (ou 15 ans pour les produits en *run-off*), plus de frais sont appliqués sur le portefeuille que ce qui est réellement observé sur une année. Afin de limiter cet effet et de garder une certaine cohérence avec les masses de frais déterminées par les services de rentabilité et du Modèle Interne, un **ratio** est appliqué a posteriori de telle sorte que les frais d'acquisition appliqués sur la base segmentée et conduisant à la détermination de la valeur client correspondent aux frais d'acquisition *New Business* du *Risk Management*. Cette méthode est cohérente avec celles utilisées dans les autres branches, permet d'éviter l'impact très négatif des frais d'acquisition sur la première année, et permet de conserver l'incitation à garder les assurés en portefeuille.

Enfin, il convient de noter que la modélisation des frais d'acquisition a été un sujet d'attention particulier tout au long de l'étude sur la valeur client épargne, qui a fait l'objet de nombreux tests de sensibilité, d'un suivi particulier et d'une validation lors de différents comités.

Une fois tous ces éléments déterminés, la valeur épargne peut être calculée pour chaque contrat. Afin d'éviter des valeurs extrêmes aberrantes, et pour s'aligner à la méthodologie suivie en IARD, les valeurs épargne finales sont bornées aux premier et neuvième déciles.

En résumé : Calcul de la valeur épargne

- Construction des lois de comportement ;
  - Des lois calculées sur un horizon de projection de 60 ans pour chaque type de flux (arbitrages, rachats et versements) ;
  - Une hypothèse forte : le présent est représentatif du futur ;
  - Des bornes sur l'âge et l'ancienneté pour refléter les sorties de portefeuille.
- Calcul de la VIF et de la duration contrat ;
  - Une VIF qui dépend des nouvelles lois de la base segmentée et d'hypothèses *EEV* issues du Modèle Interne ;
  - Une duration contrat calculée en fonction du RESPROD, de l'âge et de l'ancienneté de l'assuré (espérance de vie résiduelle en portefeuille de l'assuré d'âge et d'ancienneté donnés, au sein de son RESPROD).
- Modélisation des frais d'acquisition ;
  - Des frais d'acquisition qui doivent être modélisés, de façon comparable aux méthodes utilisées en IARD, sans pour autant pénaliser l'épargne ;
  - Un amortissement sur la duration moyenne du produit (ou 15 ans pour les produits en *run-off*).

5.2 Analyse de la segmentation du portefeuille et de la valeur épargne

Analyse des différences entre les *clusters*

Une fois la segmentation du portefeuille réalisée et la valeur épargne de chaque client calculée, une analyse plus approfondie des *clusters* permet de comprendre ce qui les différencie, et comment la segmentation a été réalisée. Les analyses se sont concentrées sur les contrats du *Réseau3-Produit9*, qui rassemble environ 10% du portefeuille total et est représentatif de l'ensemble.

Des statistiques descriptives permettent d'étudier la répartition des variables de *feature importances* les plus élevées, dans les différents *clusters* associés aux flux de rachat partiel. Ainsi, pour chaque *cluster* du *Réseau3-Produit9*, la distribution des variables parmi les *clusters* a été étudiée. Afin de distinguer les spécificités de chaque *cluster*, la distribution globale de ces variables au sein du RESPROD global a également été observée pour comparaison.

Les graphes résultants se trouvent en Annexe. Pour rappel, le nombre de trois classes de comportement a été retenu, puisqu'une classe par type de flux doit être attribuée à chaque assuré, ce qui peut conduire à un très grand nombre de combinaisons possibles. Ce nombre de *cluster* semblait d'ailleurs cohérent avec les résultats de la partie 4.1. Voici un résumé des différentes observations pour les dix variables principales, et pour les trois classes de comportement (pour le rachat partiel) :

	Cluster 0	Cluster 1	Cluster 2
Tranche de PM de fin d'année précédente	Sur-représentation des PM entre 50 000 € et 500 000 €	Sur-représentation des PM entre 0 € et 50 000 €	Sur-représentation des PM supérieures à 100 000 €
Tranche de taux de RP moyen historique	Sous-représentation des taux à 0% et sur-représentation des taux entre 0% et 30%	Suit la distribution du RESPROD	Sous-représentation des taux à 0% et sur-représentation forte des taux entre 2% et 6% puis entre 10% et 20%
Top nombre de RP extrême	Sous-représentation des 0 et sur-représentation des 1	Suit la distribution du RESPROD	Sous-représentation forte des 0 et sur-représentation forte des 1
Tranche de salaire moyen	Sous-représentation des faibles salaires (jusqu'à 15€/h) et sur-représentation des salaires élevés	Suit la distribution du RESPROD	Sous-représentation forte des faibles salaires (jusqu'à 15€/h) et sur-représentation forte des salaires élevés
Taux de PB net de l'année précédente	Sous-représentation légère des grands taux de PB et sur-représentation légère des faibles taux de PB	Suit la distribution du RESPROD	Sous-représentation forte des grands taux de PB et sur-représentation forte des faibles taux de PB
Tranche de part UC de fin d'année précédente	Sur-représentation des parts entre 0% et 20% ainsi qu'entre 40% et 70%	Suit la distribution du RESPROD	Sur-représentation des parts entre 0% et 70%
Info TMG / TMGA	Sur-représentation des TMGA strictement positifs	Suit la distribution du RESPROD	Sur-représentation forte des TMGA strictement positifs
Tranche de taux de VL moyen historique	Sous-représentation des taux à 0% et sur-représentation de toutes les autres tranches, surtout 10%-20% et 30%-50%	Suit la distribution du RESPROD	Sous-représentation forte des taux à 0% et sur-représentation de toutes les autres tranches, surtout 10%-20% et 30%-50%
Âge	Sur-représentation des grands âges	Suit la distribution du RESPROD	Sur-représentation des grands et très grands âges
Ancienneté	Sur-représentation des faibles et des grandes anciennetés	Suit la distribution du RESPROD	Sur-représentation des moyennes et des grandes anciennetés

FIGURE 31 – Distribution des dix variables principales selon les *clusters* de rachat

Les statistiques réalisées permettent de mettre en avant les sur-représentations et sous-représentations de certaines variables au sein de chaque *cluster*. Ainsi, le ***cluster* 1** semble être un groupe **moyen**, qui suit la distribution globale du RESPROD. Les individus affectés au ***cluster* 0** sont en moyenne ceux qui ont une PM entre 50 000 € et 500 000 €, un taux moyen historique de RP entre 0% et 30%,un nombre élevés de rachats partiels dans le passé, un salaire souvent élevé, un faible taux de PB net, une part d'UC entre 0% et 20% ou entre 40% et 70%, un TMGA positif, un taux moyen historique de VL entre 10% et 20% ou entre 30% et 50%, un âge élevé et une ancienneté faible ou très élevée. Le ***cluster* 2** présente les mêmes sur- et sous-représentations que sur le *cluster* 0, mais sont encore plus marquées.

De plus, des lois de comportements ont été calculées pour chaque *cluster* de ce RESPROD. Une illustration des lois de rachats partiels se trouve ci-dessous, en fonction des *clusters* de rachat :

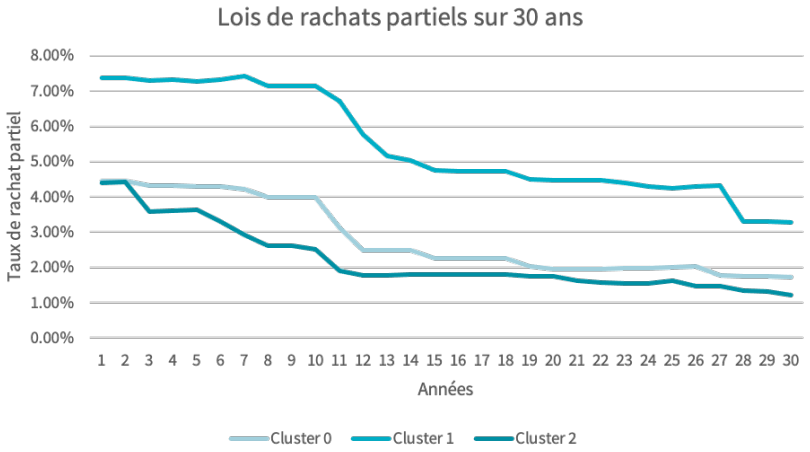

FIGURE 32 – Lois de rachats partiels pour les différents *clusters* de rachat du Réseau3-Produit9

Ces lois donnent, pour chaque *cluster*, le taux de rachat partiel calculé pour chaque année calendaire à venir.
Deux observations méritent d'être soulevées :

- Les lois sont **décroissantes** au cours du temps : ceci s'explique par le vieillissement du portefeuille étudié. En effet, aucune nouvelle entrée en portefeuille n'est modélisée par la méthodologie de construction des lois de comportement. Ainsi, plus les années calendaires augmentent, plus les contrats correspondent à des âges et anciennetés élevés, sur lesquels de faibles taux de flux sont observés ;

- La séparation du RESPROD en **trois *clusters*** n'était peut être pas nécessaire, **deux** auraient sans doute pu suffire puisque les lois des *clusters* 0 et 2 semblent assez proches, donc les comportements des clients classés dans ces deux groupes doivent l'être également.

Analyse de la nouvelle valeur épargne et comparaison avec l'ancienne grille

Après avoir calculé les nouvelles valeurs obtenues suite à la segmentation du portefeuille, la valeur que chaque contrat aurait obtenu en utilisant l'ancienne grille de décision (en fonction de l'encours et de la part UC) a été calculée. Dans cette sous-partie, les différences entre ces deux valeurs sont exposées. Les chiffres présentés ont été légèrement modifiés pour des raisons de confidentialité et ne sont donc pas les vraies valeurs, mais toutes les observations et analyses restent valables. En effet, les valeurs relatives sont les mêmes, la transformation a été uniforme entre l'ancienne et la nouvelle méthode : seuls les montants ne sont pas représentatifs.
Une première étape a consisté à comparer les distributions des valeurs sur deux produits différents, distribués sur deux réseaux différents :

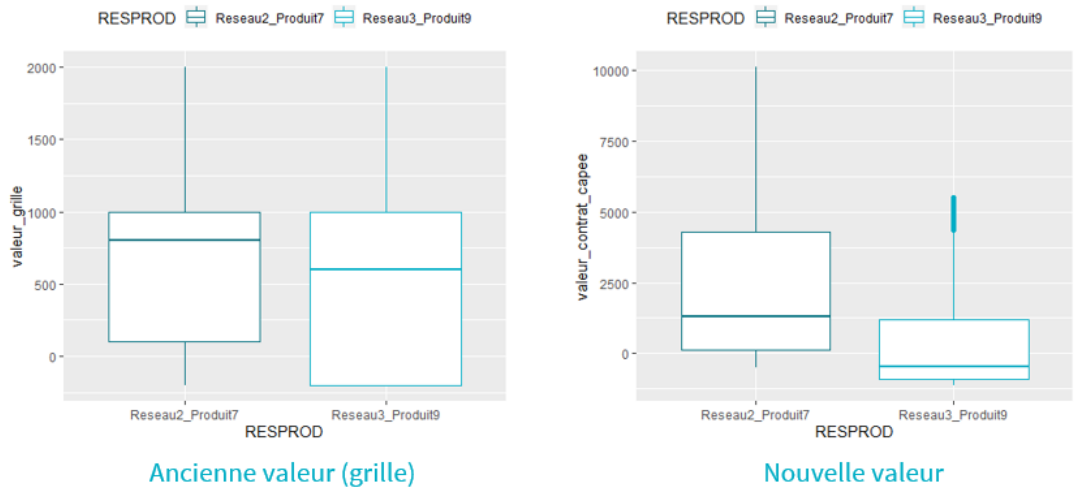

FIGURE 33 – Distribution des valeurs issues de la grille et des nouvelles valeurs épargne sur deux RESPROD différents

Plusieurs constatations ressortent de ces distributions :

- La **médiane** des valeurs du Reseau2-Produit7 est **supérieure** à la médiane des valeurs du Reseau3-Produit9, que ce soit avec l'ancienne valeur ou avec la nouvelle. Ceci est rassurant car signifie que l'ordre général a été conservé ;

- La **dispersion des nouvelles valeurs** issues de la segmentation est **plus grande** que celle des anciennes valeurs issues de la grille. Ceci semble indiquer que la nouvelle méthodologie incluant l'étape de segmentation permet de mieux différencier les caractéristiques des assurés et donc de leur attribuer une plus large gamme de valeurs différentes.

Afin d'analyser plus en détail les différences de dispersion entre ancienne et nouvelle valeur, un *zoom* a été effectué sur le Réseau3-Produit9 :

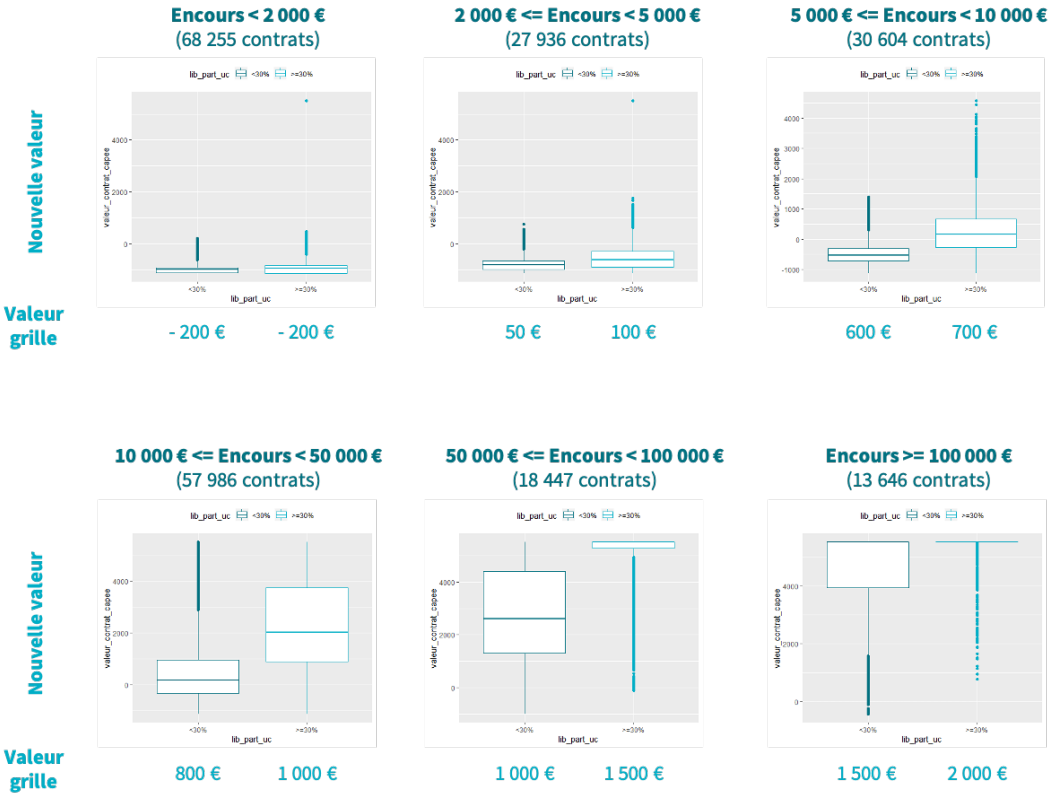

FIGURE 34 – Distribution des nouvelles valeurs épargne sur le Réseau3-Produit9 en fonction des tranches d'encours et de part UC de l'ancienne grille

Deux boîtes à moustache apparaissent sur chaque graphique correspondant à une tranche d'encours particulière : celle de gauche (bleu-vert foncé) correspond à une part UC strictement inférieure à 30%, et celle de droite (bleu turquoise clair) à une part UC supérieure à 30%. Ce croisement entre la tranche d'encours et la part UC fixait le critère d'affectation de la valeur épargne selon l'ancienne méthode (grille).

Il convient alors de remarquer que les nouvelles **valeurs épargne** sont de façon générale **croissantes** avec l'**encours** et la **part UC**. Ceci semble bien cohérent avec l'ancienne grille de valeurs et avec l'intuition.
À **encours égal**, une **plus grande valeur** est attribuée lorsque la **part UC est plus grande**. Encore une fois, ceci est bien cohérent.
Par ailleurs, une plus grande dispersion est constatée sur les valeurs attribuées aux contrats d'encours entre 10 000 € et 100 000 €. Ceci reflète les différences de caractéristiques entre individus qui impliquent une valeur finale différente après segmentation, ce qui n'était pas le cas avec la valeur de la grille (valeur unique pour une tranche d'encours et une tranche de part UC fixées). De plus, des valeurs négatives sont observables, qui sont dues à une VIF négative (à cause de l'environnement de taux bas) et aux frais d'acquisition.

Enfin, une valeur extrême est constatée sur le premier graphe tout à gauche, pour un contrat d'encours inférieur à 2 000 € et de part UC supérieure à 30%. Un zoom sur ce contrat s'impose pour comprendre pourquoi une valeur si élevée et atypique lui a été attribuée :

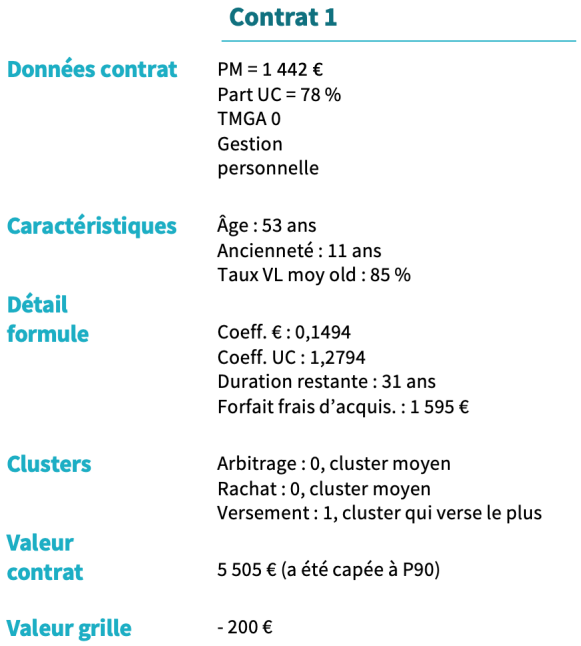

FIGURE 35 – Zoom sur un contrat particulier du Réseau3-Produit9

Cet assuré n'est **pas très âgé** et possède déjà une certaine **ancienneté**. L'**encours** est **faible** mais la **part UC élevée**, et l'assureur n'a **aucun engagement de taux minimum de revalorisation** sur ce contrat. De plus, ce client a été classé dans les *clusters* moyens d'arbitrages et rachats, mais dans celui qui **verse le plus**. Il a d'ailleurs effectué des versements libres lors des trois dernières années, à hauteur (en moyenne) de **85%**. Ces deux derniers éléments expliquent pourquoi les coefficients Euro et UC utilisés dans la formule de calcul de la valeur sont élevés.
Son faible encours aurait conduit ce contrat à recevoir une valeur négative selon l'ancienne grille de décision. Cependant, la nouvelle méthode de calcul de la valeur après segmentation du portefeuille permet de mettre en avant les caractéristiques positives de ce contrat et de refléter sa rentabilité future qui semble élevée.

Un autre point d'attention a été approfondi : comme illustré en figure 34, il existe des contrats d'**encours supérieur à 100 000 €** qui ont **pourtant une valeur négative**. Ces contrats ont été étudiés et présentent des caractéristiques similaires :

- Une part UC nulle ;
- Un âge compris entre 89 et 99 ans ;

- Une ancienneté comprise entre 0 et 8 années ;
- Un *cluster* moyen de rachats, moyen d'arbitrages et moyen de versements.

Malgré leur encours important, ces contrats n'ont pas encore amorti leurs frais d'acquisition à cause de leur faible ancienneté. De plus, ils sont âgés et font preuve d'inactivité, ou alors d'une activité non particulièrement bénéfique pour la compagnie (pas de versements importants par exemple). Enfin, ils sont sur un support 100% euro, ce qui est risqué pour l'assureur en ce contexte de taux bas voire négatifs et donc peu apprécié. Deux de ces contrats sont illustrés en détail dans la figure suivante :

	Contrat 2	Contrat 3
Données contrat	PM = 114 265 € Part UC = 0 % TMGA 0 Gestion personnelle	PM = 101 463 € Part UC = 0 % TMGA 0 Gestion personnelle
Caractéristiques	Âge : 95 ans Ancienneté : 3 ans Taux VL moy old : 0 %	Âge : 99 ans Ancienneté : 8 ans Taux VL moy old : 0 %
Détail formule	Coeff. € : 0,0021 Coeff. UC : 0,0087 Duration restante : 5 ans Forfait frais d'acquis. : 1 595 €	Coeff. € : 0,0021 Coeff. UC : 0,0087 Duration restante : 4 ans Forfait frais d'acquis. : 1 595 €
Clusters	Arbitrage : 0, cluster moyen Rachat : 0, cluster moyen Versement : 0, cluster moyen	Arbitrage : 0, cluster moyen Rachat : 0, cluster moyen Versement : 0, cluster moyen
Valeur contrat	- 257 €	- 247 €
Valeur grille	1 500 €	1 500 €

FIGURE 36 – Zoom sur deux contrats particuliers du Réseau3-Produit9

Après ces analyses à la maille RESPROD, une analyse à la maille RESPROD/*clusters* a été réalisée. Plus précisément, les valeurs de contrats du segment qui **"rachète le plus et verse le moins"** ont été comparées à celles du segment qui **"rachète le moins et verse le plus"**. Ces combinaisons n'étant en fait pas représentées dans le Réseau3-Produit9, deux contrats du Réseau2-Produit7 ont été observés en détail :

	Contrat 4	Contrat 5
Données contrat	PM = 266 855 € Part UC = 0 % TMGA 0 Gestion personnelle	PM = 2 225 € Part UC = 59 % TMGA 0 Gestion sous mandat
Caractéristiques	Âge : 84 ans Ancienneté : 19 ans Taux VL moy old : 24 % ; Taux RP moy old : 0 %	Âge : 81 ans Ancienneté : 1 an Taux VL moy old : 0 % ; Taux RP moy old : 0 %
Détail formule	Coeff. € : 0,0117 Coeff. UC : 0,0308 Duration restante : 9 ans Forfait frais d'acquis. : 1 231 €	Coeff. € : 0,0079 Coeff. UC : 0,0219 Duration restante : 10 ans Forfait frais d'acquis. : 1 231 €
Clusters	Arbitrage : 2, cluster qui arbitre le plus Rachat : 1, cluster qui rachète le moins Versement : 2, cluster qui verse le plus	Arbitrage : 0, cluster moyen Rachat : 0, cluster qui rachète le plus Versement : 1, cluster qui verse le moins
Valeur contrat	10 106 € (a été capée à P90)	- 520 €
Valeur grille	1 500 €	100 €

FIGURE 37 – Zoom sur deux contrats particuliers du Réseau2-Produit7

Ces contrats permettent de mettre en valeur l'intérêt de la segmentation du portefeuille selon les comportements dans le cadre de la détermination de la valeur client épargne : les **coefficients Euro et UC** ainsi que les **valeurs épargne** finales sont **plus élevés** pour la combinaison de *clusters* qui **"rachète le moins et verse le plus"** relativement à la combinaison qui "verse le moins et rachète le plus". C'est cohérent, et la partie suivante montre de manière plus poussée la plus-value des travaux de segmentation réalisés.

Plus-value des travaux réalisés et avantages du nouveau modèle de valeur épargne

Afin d'illustrer l'intérêt des travaux de segmentation du portefeuille, ainsi que leur apport au projet de valeur client épargne, trois contrats sont examinés :

	Contrat 6	Contrat 7	Contrat 8
Données contrat	PM = 43 302 € Part UC = 60 % TMGA 0 Gestion sous mandat	PM = 11 215 € Part UC = 42 % TMGA 0 Gestion profilée	PM = 46 487 € Part UC = 87 % TMGA 0 Convention de gestion
Caractéristiques	Âge : 63 ans Ancienneté : 2 ans Taux VL moy old : 0 % ; Taux RP moy old : 0 %	Âge : 77 ans Ancienneté : 12 ans Taux VL moy old : 0 % ; Taux RP moy old : 73 %	Âge : 55 ans Ancienneté : 4 ans Taux VL moy old : 97 % ; Taux RP moy old : 0 %
Détail formule	Coeff. € : 0,0022 Coeff. UC : 0,0091 Duration restante : 23 ans Forfait frais d'acquis. : 1 595 €	Coeff. € : 0,0021 Coeff. UC : 0,0087 Duration restante : 13 ans Forfait frais d'acquis. : 1 595 €	Coeff. € : 0,0430 Coeff. UC : 0,0286 Duration restante : 29 ans Forfait frais d'acquis. : 1 595 €
Clusters	Arbitrage : 2, cluster qui arbitre le plus Rachat : 2, cluster qui rachète le moins Versement : 2, cluster qui verse le moins	Arbitrage : 1, cluster qui arbitre le moins Rachat : 1, cluster qui rachète le plus Versement : 1, cluster qui verse le plus	Arbitrage : 0, cluster moyen Rachat : 1, cluster qui rachète le plus Versement : 1, cluster qui verse le plus
Valeur contrat	4 832 €	- 131 €	5 506 €
Valeur grille	1 000 €	1 000 €	1 000 €

FIGURE 38 – Zoom sur trois contrats particuliers du Réseau3-Produit9

Selon l'ancienne grille d'attribution de la valeur, ces trois contrats auraient reçu la même valeur épargne, puisqu'ils se trouvent dans la même tranche d'encours et de part UC. Cependant, leurs caractéristiques diffèrent, et leur rentabilité pour la compagnie également. Ainsi, le **contrat 6** est un contrat **"moyen"**. Le **contrat 7** a un **historique de forts rachats partiels**. Il dispose d'un **faible encours**, malgré une ancienneté qui commence à être élevée. Ce client a beaucoup racheté par le passé et **pas versé**, donc ne semble pas rentable. Ses caractéristiques le poussent à recevoir une **valeur négative**. Enfin, le **contrat 8** est à l'opposé : il présente un historique de **fort taux de versements libres** et d'**aucun rachat partiel**. De plus, il comporte un encours assez important, avec une grande part UC, et pas de taux de revalorisation minimum garanti. En plus de ces caractéristiques, puisqu'il a beaucoup versé par le passé et pas racheté, il est jugé comme rentable et reçoit donc une valeur épargne positive.

L'analyse de deux autres contrats permet de démontrer une autre face de la plus-value apportée par le nouveau modèle de valeur épargne :

	Contrat 9	Contrat 10
Données contrat	PM = 9 675 € Part UC = 100 % TMGA 0 Gestion personnelle	PM = 1 906 € Part UC = 42 % TMGA 0 Gestion profilée
Caractéristiques	Âge : 31 ans Ancienneté : 3 ans Salaire moyen : 32 € net / heure Taux VL moy old : 0 % ; Taux RP moy old : 0 %	Âge : 57 ans Ancienneté : 35 ans Taux VL moy old : 0 % ; Taux RP moy old : 0 % Taux RP last : 94 %
Détail formule	Coeff. € : 0,0021 Coeff. UC : 0,0087 Duration restante : 47 ans Forfait frais d'acquis. : 1 595 €	Coeff. € : 0,0022 Coeff. UC : 0,0097 Duration restante : 23 ans Forfait frais d'acquis. : 1 595 €
Clusters	Arbitrage : 1, cluster qui arbitre le moins Rachat : 1, cluster qui rachète le plus Versement : 1, cluster qui verse le plus	Arbitrage : 1, cluster qui arbitre le moins Rachat : 1, cluster qui rachète le plus Versement : 1, cluster qui verse le plus
Valeur contrat	2 537 €	277 €
Valeur grille	700 €	- 200 €

FIGURE 39 – Zoom sur trois contrats particuliers du Réseau3-Produit9

Le contrat 9 est certes récent et avec une faible PM, mais le client qui y est rattaché est jeune et a un salaire moyen particulièrement élevé par rapport aux autres clients du portefeuille. Son encours est réparti entièrement sur des UC et il ne présente pas de taux minimum garanti. Toutes ces caractéristiques lui permettent d'obtenir une valeur épargne (très) positive. Celle de la grille était déjà positive, mais ce contrat était moins bien valorisé selon cette ancienne méthodologie (classé avec les inactifs de faible encours et donc rentabilité faible).

Le contrat 10 a effectué des rachats récents et exceptionnels sur son contrat, et dispose donc d'un très faible encours. Cependant, le modèle arrive à ne pas le catégoriser comme fort racheteur, grâce à ses caractéristiques (grande fidélité récompensée, pas de 100 % euro, pas de taux minimum garanti).

Le **nouveau modèle** développé **arrive** donc bien à **capter les différences de comportements et de caractéristiques** entre assurés : il attribue une meilleure valeur aux clients présentant des caractéristiques avantageuses, versant beaucoup et rachetant peu (voire pas), et une plus faible (carrément négative) à des clients présentant des caractéristiques inverses.

La partie 6 approfondit la plus-value du modèle mis en place, en démontrant les gains réalisés par l'assureur lorsqu'il cible de manière appropriée les clients les plus rentables et ce qui le sont moins, et arrive à diriger les actions stratégiques appropriées envers ces deux types d'assurés.

En résumé : Analyse de la segmentation du portefeuille et de la valeur épargne

- Analyse des différences entre les *clusters* ;
  - Des différences de caractéristiques sous- et sur-représentées entre les *clusters*.
- Analyse de la nouvelle valeur épargne et comparaison avec l'ancienne grille ;
  - Une nouvelle valeur plus dispersée que l'ancienne mais toujours croissante avec l'encours et la part UC ;
  - Des contrats particuliers analysés pour mieux comprendre leurs valeurs, liées à leurs caractéristiques : meilleures en général pour les contrats d'ancienneté plus élevée, comptabilisant plus de versements et peu de rachats, ou d'assuré jeune par exemple.
- Plus-value des travaux réalisés et avantages du nouveau modèle de valeur épargne ;
  - Un nouveau modèle qui arrive à capter les différences de caractéristiques et de comportement entre épargnants, et à leur affecter une valeur représentative de leur rentabilité prospective ;
  - Des valeurs plus faibles pour les forts racheteurs aux caractéristiques peu avantageuses ;
  - Des valeurs plus élevées pour ceux qui versent beaucoup et présentent des caractéristiques avantageuses (jeunes, prometteurs car au salaire élevé, sur du 100% UC, par exemple).

5.3 Limites des modèles et améliorations envisageables

Bien que les modèles utilisés dans le cadre de l'étude sur la valeur épargne aient démontré leur performance, leur utilisation a aussi mis en lumière certaines limites.

La première a concerné les données utilisées. Outre les problématiques de **qualité des données**, essentielles mais aussi parfois difficiles à gérer, la **base** utilisée par les modèles de *machine learning* était très **déséquilibrée**. En effet, la majorité des clients en por-

tefeuille n'effectuent pas de mouvement sur leur contrat : en 2019, seulement 2,41% ont effectué au moins un arbitrage (manuel), 6,21% au moins un rachat partiel, et 4,77% au moins un versement. Ce grand déséquilibre a alors dans un premier temps entaché la performance des algorithmes de classification supervisée, incapables de reproduire un tel déséquilibre et effectuant de grosses erreurs de prédiction.
Une solution partielle a consisté à **modifier la fonction de coût** de ces modèles en surpondérant les erreurs effectuées sur la classe minoritaire, afin d'en diminuer l'ampleur. Ce type d'erreur serait grave pour l'assureur, puisque cela reviendrait à prédire une absence de flux pour des contrats sur lesquels il y aurait en réalité bien des flux, ce qui ne serait pas prudent.
Cependant, cette alternative présente elle aussi une limite : sur-pénaliser les erreurs effectuées sur la classe minoritaire de la sorte revient presque à doubler les observations de cette classe, et la méthode n'est donc pas très stable. Une autre alternative envisagée et qui serait plus stable, consisterait à utiliser la théorie du *bootstrap* en entraînant les modèles sur de petits échantillons aléatoires de la base, avec une contrainte sur la proportion d'observations de la classe minoritaire (au moins 20% ou 30% de "1" pour la classification binaire par exemple). Cette méthode permettrait d'introduire de la variance, et d'obtenir des résultats de classification encore meilleurs.

Ce **déséquilibre** se retrouve **dans la répartition des montants de flux** également, puisque la majorité des contrats présente des montants de flux nuls. Les algorithmes de régression qui tentent de prédire ces montants sont donc eux aussi appauvris par ce déséquilibre et leur performance s'en trouve diminuée.
En effet, l'erreur moyenne absolue (MAE) sur les montants de rachats partiels du modèle retenu (*XG Boost*) était d'environ 1 000 €. Cette erreur semble faible en comparaison avec les montants d'encours parfois très important, mais beaucoup plus élevée quand les montants moyens de flux sont analysés. En effet, sur l'ensemble de la base, le montant moyen de rachat partiel est d'environ 1 020 €. Parmi la faible part de contrats sur lesquels des rachats partiels sont effectués, ils le sont en moyenne de 16 000 €. Le modèle de projection des flux (utilisé dans la méthode 2 de segmentation) a donc commis des erreurs dans les montants prédits, qui ont conduit à l'obtention de lois de comportement aberrantes, et dont les sources ont été identifiées et détaillées ci-dessous.
Les taux obtenus sur l'année 0 correspondaient aux montants de flux et d'encours observés sur la base à la date d'extraction.
Ceux de l'année 1 commençaient à être légèrement plus élevés que ceux normalement observés sur d'autres lois de comportement. Ils correspondent à la première itération de la boucle, donc à la première prédiction et ne sont pas encore aberrants. Mais la MAE était quand même d'environ 1 000 €, le **modèle** fait donc des **erreurs**, et **surtout en fréquence** : il prédit un montant nul pour seulement 271 contrats, soit $0,01\%$ de la base (alors que les montants de flux nuls représentaient 94% de la base en année 0). Au lieu de prédire des 0, le modèle a prédit beaucoup de faibles montants. De plus, il a légèrement sous-estimé les montants élevés, puisque la somme globale de tous les flux du portefeuille pour l'année 1 est assez proche du montant total observé en année 0.
Cependant, dès l'année 2, les erreurs semblent encore plus grandes, en fréquence et en mon-

tant (seulement 4557 montants nuls prédits, le montant total de flux sur toute la base est largement supérieur au total des années 0 ou 1, et les taux calculés supérieurs aux taux observés normalement par ailleurs). Ceci peut s'expliquer par la très **faible part de contrat pour lesquels un montant nul a été prédit en année 1**. En effet, lorsque les taux moyens historiques (sur trois ans) utilisés comme variables explicatives sont recalculés, ils intègrent le montant de l'année 1, et deviennent strictement positifs pour quasiment tous les contrats de la base. Or, la base initiale sur laquelle le modèle a été entraîné (en année 0) comporte une très faible part de contrats pour lesquels ces taux de flux moyens historiques sont différents de 0, puisque la majorité des contrats sont "inactifs". Ainsi, pour quasiment tous les contrats, le modèle reprend comme exemple les contrats de la base initiale d'entraînement qui avaient des taux moyens historiques strictement positifs, et qui correspondaient à des contrats avec de gros flux (donc taux élevés, puisque la quasi totalité de la base est inactive, donc cela signifie que les contrats qui rachètent ont eux des taux de rachats très élevés), et prédit donc des montants de flux élevés.

Ces erreurs n'ont pas été jugées comme bloquantes puisque le but de l'étude, à cette étape, est de séparer au mieux les comportements, pas de prédire avec la plus grande précision possible les flux futurs. Mais dans une optique de développement de tels modèles de projection, et pour essayer de résoudre les problèmes soulevés, deux solutions sont envisageables :

- Après les prédictions faites sur année 1, re-entraîner le *XG Boost* sur la base de l'année 1. On s'attend alors à ce que les montants prédits soient nombreux en fréquence, comme observé sur l'année 1 déjà, mais moins grands en montants, surtout pour les faibles flux ;

- Insérer une étape de classification de survenance de flux, puis n'entraîner et n'effectuer la prédiction de montant de flux par régression que sur les individus pour lesquels on a prédit la survenance d'un flux.

Ces deux alternatives devraient permettre d'obtenir des prédictions de montants de flux plus précises, et donc des lois de comportement moins aberrantes. Ceci pourrait aussi jouer en faveur de la méthode 2 de segmentation, qui pourrait donc être améliorée et surpasser la méthode 1 en terme de performance. Surtout, cela permettrait de développer une méthode plus fiable de projection des montants de flux par contrat.

Par ailleurs, il existe d'autres limites propres à la méthode 2, qui pourraient être dépassées. En effet, les **hypothèses de vieillissement** du portefeuille et d'**évolution des variables explicatives** semblent améliorables.
Par exemple, la prise en compte de différents scénarios économiques pour projeter les taux de performances financières utilisées dans le calcul des **taux de PB** permettrait notamment d'obtenir une modélisation plus fine et plus précise.
De même, le **modèle d'évolution de la PM** pourrait être amélioré en construisant un modèle de projection plus complexe, prenant en compte la revalorisation et des hypothèses de taux de rachats totaux notamment.

Enfin, une dernière alternative qui a été imaginée (et introduite dans la partie sur le *clustering* de séries temporelles) mais n'a pas pu être développée, serait de réaliser un ***k-means* directement sur les lois de comportement construites** par le modèle de projection itératif, en utilisant une distance appropriée (**DTW** par exemple).

Une autre limite concerne les caractéristiques des contrats. Il existe des **variables** très intéressantes car **pertinentes** pour modéliser les comportements et qui restent pour l'instant encore **inaccessibles** (par exemple pour modéliser les rachats, les variables "cause du rachat", "le client est propriétaire de son habitation", ou encore "nombre de contrats total chez AXA" devraient être très discriminantes). En effet, il semble que de telles variables refléteraient bien mieux les motivations des épargnants à effectuer des flux et donc permettraient de bien mieux prédire les comportements. Avec l'ambition actuelle d'AXA de construire un très large entrepôt de données bien organisé et fiable, ces informations devraient bientôt être facilement accessibles et pourraient être utilisées à des fins d'amélioration des modèles de prédiction.

Par ailleurs, la **construction des lois de comportement** par RESPROD/*cluster* après le travail de segmentation semble également comporter quelques limites.
Seulement trois années d'historique sont disponibles pour le calcul des taux, qui reproduisent donc un peu trop ce qui a été observé sur les années disponibles. Pour améliorer ce point, il faudrait déjà disposer d'un historique de données plus profond et construire les taux prédits en utilisant plusieurs années. Ceci est aujourd'hui encore impossible, mais encore une fois, la construction de l'entrepôt de données pourrait y remédier.
Aussi, si un historique plus profond de taux de flux pour chaque contrat sur plusieurs années venait à être exploitable, alors une segmentation du portefeuille par rapprochement direct des épargnants en fonction de leurs lois de comportement passées pourrait aussi être imaginable.

Enfin, au-delà de tout le travail de segmentation effectué et de ses limites, il convient de garder en mémoire que les **frais d'acquisition** jouent un **rôle non négligeable** dans le calcul de la valeur épargne finale. Ainsi, deux individus aux caractéristiques et comportement proches, mais appartenant à deux RESPROD différents se voient retrancher des frais d'acquisition très différents, puisque ces derniers sont calculés à la maille RESPROD (et donc obtiennent deux valeurs épargne éloignées).
Il est donc finalement très important de bien modéliser les frais d'acquisition, ce qui justifie notamment le grand nombre de tests de sensibilité et le suivi particulier effectués tout au long de l'étude, sur cette partie du modèle.

En résumé : Limites des modèles et améliorations envisageables

- Une base de données déséquilibrée qui complique le travail d'apprentissage ;
- Des erreurs de prédiction de montants de flux qui pourraient être diminuées en introduisant une étape de classification sur la survenance de flux, ou de ré-entraînement du modèle de régression après la première itération ;
- Des hypothèses d'évolution des variables explicatives qui pourraient être améliorées (projection des taux de PB, des PM, par exemple) ;
- Un modèle de *clustering* sur les lois brutes à tester ;
- Des variables explicatives encore plus pertinentes à prendre en compte (nombre total de contrats chez AXA, cause du rachat, propriétaire ou locataire, notamment) ;
- Une modélisation des frais d'acquisition à effectuer avec précaution.

5.4 Perspectives de déploiement de l'indice client

Dans l'environnement actuel mouvementé, instable et fortement concurrentiel, l'enjeu est de mieux **cibler les actions** et **orienter les stratégies** en combinant une approche commerciale centrée sur le client, et une **maîtrise de la rentabilité**, afin de pérenniser la solvabilité de la compagnie.
À cet effet, l'indice client doit permettre de caractériser l'ensemble des clients possédant au moins un contrat chez AXA France en reflétant leur rentabilité prospective estimée.

Il permet d'évaluer la **valeur économique globale** des clients détenteurs d'au moins un contrat en prenant en compte les différents qu'il possède. Construit comme la somme des résultats attendus dans le futur pour chacun des contrats valorisés actuellement détenus par chaque client, il permet donc d'avoir une vision globale du client correspondant.
Les clients indicés **1** sont les clients non rentables et donc **à redresser**. Ceux d'indice **2** sont les clients à rentabilité moyenne et donc à **développer ou multi-équiper**. Enfin, les indicés **3** sont les clients les plus rentables et donc **à protéger**.

Langage commun entre tous les acteurs, l'indice client permet une reconnaissance simple et partagée du statut du client pour les agents (entre les différents réseaux de distribution), les engagements IARD, le service client, le marketing, et tous les autres interacteurs potentiels.
Enfin, il doit permettre l'**optimisation des actions orientées Client**, parmi lesquelles quelques exemples :

- Ciblage des actions marketing et commerciales ;
- "Appels câlins" pour les clients bien indicés ;
- Octroi de crédit en banque ;
- Versement pro-actif de la prime de naissance aux indicés 3.

Le but de l'indice est ainsi toujours d'avantager les bons clients, jamais de pénaliser les mauvais.

Enfin, au-delà d'un simple indicateur marketing permettant d'orienter les agents, la valeur client épargne pourrait également servir à l'assureur pour **contrôler sa rentabilité** et **compléter les indicateurs actuariels classiques** de rentabilité tels que la VIF : c'est ce qui est expliqué dans la partie suivante.

En résumé : Perspectives de déploiement de l'indice client

- Un indice proposant une vision globale de chaque client et reflétant sa rentabilité prospective estimée ;
- Un meilleur ciblage des actions orientées Clients et des stratégies à développer ;
- Une maîtrise de la rentabilité ;
- Un complément aux indicateurs actuariels de rentabilité traditionnels.

6 Valeur client épargne, rentabilité et risque

Comme l'expliquait la partie précédente, le nouveau modèle de valeur client épargne permet à présent de prendre en compte les différences de caractéristiques et de comportements entre assurés pour leur affecter une valeur représentant leur rentabilité future potentielle. Ceci permet à l'assureur de fournir aux agents des indicateurs techniques, basés sur une approche actuarielle et statistique, leur permettant de **mieux orienter et cibler leurs actions**. Mais, au-delà d'un simple indicateur marketing, cette nouvelle valeur client épargne pourra permettre à l'assureur de **contrôler sa rentabilité** et **pérenniser sa solvabilité**, en **optimisant l'orientation de ses stratégies en épargne**. À ce titre, l'assureur pourra ajouter cet indicateur dans son rapport ORSA et suivre son évolution au cours du temps, globalement et par secteur d'activité. Ainsi, cette valeur pourrait s'ajouter aux indicateurs traditionnels de rentabilité et de valorisation, pour les compléter, sans bien sûr les remplacer pour autant. C'est ce qu'illustre cette partie du mémoire.

Puisqu'une valeur en euros a été attribuée à chaque client en portefeuille, l'assureur est capable de distinguer les segments rentables de ceux qui le sont moins. Grâce à l'étape intermédiaire de segmentation qui a affecté un *cluster* (pour chaque type de flux) à tous les épargnants, il est même en mesure d'**identifier les profils à risque** (et de l'indiquer, comme le demande la réglementation Solvabilité II, dans ses rapports narratifs à destination du Public (SFCR) et/ou de l'ACPR (RSR)), de mieux cibler les clients, et pourra mieux **anticiper l'impact sur la rentabilité** des actions stratégiques et commerciales personnalisées à mener. En effet, la nouvelle valeur client doit pouvoir l'aider dans son approche du **risque** qu'il encourt (notamment sur les rachats massifs, les arbitrages massifs vers l'euro, la diminution de la collecte nette par un anéantissement des versements libres).

Imaginons par exemple que cette valeur client soit utilisée comme levier de la rentabilité du portefeuille, de telle sorte que la compagnie décide de tout faire pour avantager les 10% de clients de plus grande valeur (ceux jugés comme très rentables, versant souvent et des gros montants en général) : actions commerciales favorables, protections et offres diverses. Et au contraire, qu'elle choisisse de ne rien offrir de plus que les engagements contractuels aux 10% des clients ayant reçu les valeurs les plus basses (jugés comme risqués car rachètent souvent, entre autres). Trois scénarios sont alors envisagés, pour tester la sensibilité de la nouvelle valeur du portefeuille :

- **Scénario 1** : les détenteurs des moins bons contrats ne sont pas satisfaits et rachètent donc à hauteur de 10% l'année suivante, tandis que les clients de valeurs plus élevées sont satisfaits et effectuent des versements libres sur leurs contrats, à hauteur de 10% également ;
- **Scénario 2** : les détenteurs des moins bons contrats ne sont pas satisfaits et rachètent donc à hauteur de 25% l'année suivante, tandis que les clients de valeurs plus élevées sont satisfaits et effectuent des versements libres sur leurs contrats, à hauteur de 25% également ;
- **Scénario 3** : les détenteurs des moins bons contrats ne sont pas satisfaits et

rachètent donc à hauteur de 40% l'année suivante, tandis que les clients de valeurs plus élevées sont satisfaits et effectuent des versements libres sur leurs contrats, à hauteur de 40% également.

Tous les autres flux sont négligés. Les PM sont projetées à horizon 1 an en se basant sur ces hypothèses après avoir simulé les flux par contrat et les avoir agrégés au niveau du portefeuille. Des travaux de recalcul des VIFS correspondantes ont été effectués.
Le total de l'encours et la VIF du portefeuille associé sont alors comparés, entre la date de calcul des valeurs client épargne ("scénario 0"), et un an après la mise en place de ces stratégies, en ayant fait évoluer le portefeuille en conséquence :

Scenario	Taux de rachat des moins bons contrats	Taux de versement des meilleurs contrats	PM du portefeuille (en millions d'euros)	VIF du portefeuille (en millions d'euros)	Évolution de la VIF	VIF / PM
0	-	-	31 493	3 797	-	12,06 %
1	10 %	10 %	32 826	3 972	+ 4,61 %	12,10 %
2	25 %	25 %	34 826	4 233	+ 6,57 %	12,15 %
3	40 %	40 %	36 827	4 495	+ 6,19 %	12,21 %

FIGURE 40 – Zoom sur trois contrats particuliers du Réseau3-Produit9

Tout comme les valeurs client présentées dans la partie précédente, les valeurs présentées dans le tableau de la figure 39 ont aussi été modifiées. De nouveau, les valeurs relatives sont toujours les mêmes et les conclusions restent identiques.

Quels que soient la stratégie et le scénario envisagés, trois conclusions peuvent alors être dressées, en comparaison avec le scénario 0 représentant l'année passée :

- L'**encours** du portefeuille **grossit** ;
- La **VIF** (qui mesure la valeur globale du portefeuille) **s'améliore** significativement ;
- Le ratio VIF / PM supérieur à celui de l'année de mise en place de la stratégie, donc **la VIF augmente plus vite que l'encours**.

Ainsi, la valeur client épargne a permis à l'assureur d'identifier les clients au potentiel de rentabilité future le plus prometteur, de les distinguer des clients plus risqués et moins rentables, et de mettre en place une stratégie visant à améliorer la rentabilité et la valeur de son portefeuille.
En ceci, la valeur client épargne semble donc être un élément pertinent à retenir en complément des indicateurs actuariels traditionnels.

En résumé : Valeur client épargne, rentabilité et risque

- Un indicateur permettant d'identifier les clients à risque et d'orienter les stratégies d'amélioration de la rentabilité ;
- Un baromètre prometteur pour pérenniser le portefeuille et en augmenter la valeur.

Conclusion

Dans le cadre du projet d'indice client, une valeur client doit être déterminée pour chaque assuré, sur chacun des contrats d'assurance qu'il détient chez AXA France. Le rôle de la Direction Technique Épargne dans ce projet est de déterminer la valeur client épargne : une **mesure de la rentabilité future** estimée de chaque contrat en portefeuille. Plus précisément, il a été question dans ce mémoire de l'**amélioration du modèle de valeur client épargne**, qui se basait auparavant sur les seules données de l'encours et de la part UC de chaque contrat, et pouvait donc attribuer la même valeur à deux contrats de tranches d'encours et part UC similaires, bien qu'ils puissent présenter des potentiels de rachats, versements et arbitrages très différents (donc de rentabilité potentielle différente).

Le modèle de la valeur épargne a donc été modifié, en **affinant la maille de détermination des coefficients** intervenant dans la formule de calcul, de la **VIF** et de la **duration** contrat grâce à la **segmentation du comportement client en assurance vie** (pour les trois types de flux principaux : arbitrages, rachats et versements). Cette approche a mêlé des **techniques statistiques poussées** sur des **grands volumes de données** et a intégré des **indicateurs actuariels** de rentabilité à destination des réseaux de distribution et agents généraux particulièrement.

Elle a été l'occasion de développer et tester plusieurs **modèles de *machine learning*** à des fins de prédiction et classification, sur une base de données longuement et méticuleusement construite en amont. **Déséquilibrée** en raison du caractère exceptionnel des flux (en 2019, seulement 2,41% des contrats ont effectué au moins un arbitrage manuel, 6,21% au moins un rachat partiel, et 4,77% au moins un versement), elle ne se prêtait pas directement aux algorithmes et fonctions classiques. Des **scores de performance adaptés** ont donc été utilisés pour la sélection de modèles (F1 Score, AUC), et une **méthode de pénalisation des erreurs** commises sur la classe minoritaire, par pondération dans la fonction de coût, a été testée.
Par ailleurs, la définition d'**indicateurs** propres à la vision métier a été nécessaire pour déterminer la meilleure méthode de segmentation du portefeuille. En effet, cette dernière avait pour objectif de séparer au mieux les assurés selon leur comportement et donc leur rentabilité future, pas uniquement de maximiser un critère statistique de variance inter- ou intra-classes par exemple. Un premier indicateur reflétant l'**écart de comportement** entre les différents *clusters* a donc été construit, puis une phase de *backtesting* a permis de calculer un **écart** entre les **prédictions** et les observations.

Cette phase de sélection a alors permis de déterminer les meilleurs modèles parmi ceux testés :

- ***Random Forest* pondéré** pour la classification sur la **survenance** de flux (utilisé pour la sélection et la pondération de variables) ;
- ***XG Boost*** pour la prédiction des **montants** de flux ;

- ***k-means* pondéré** pour la méthode finale de **segmentation** du portefeuille en classes d'assurés de caractéristiques et comportements proches.

Grâce à ces algorithmes, le portefeuille a été segmenté en **trois groupes d'assurés pour chaque type de flux**, qui ont été analysés. Ainsi, des profils types ont été dressés pour chaque groupe. Par exemple, le cluster 2 de rachat partiel rassemble en moyenne plutôt des clients âgés, de grande ancienneté, possédant un encours supérieur à 100 000 €, et ayant plutôt tendance à effectuer des mouvements sur leur contrat.

Des **lois de comportement** ont été construites sur 60 ans pour chaque type de flux, à la maille RESPROD/*Clusters* (où "RESPROD" représente la maille réseau de distribution/groupe de produits). À partir de ces lois, des premiers éléments du calcul de la valeur client épargne ont alors pu être déterminés : les **coefficients Euro et UC**, qui englobent notamment la VIF. Par ailleurs, la **duration contrat** a été calculée en fonction du RESPROD, de l'âge de l'assuré et de l'ancienneté de son contrat. Un modèle de **frais d'acquisition** a également été mis en place, à partir d'un forfait de frais à la maille RESPROD, et d'un lissage sur la duration moyenne.

Ainsi, des **nouvelles valeurs épargne** ont été attribuées à chaque contrat du portefeuille, qui ont été analysées et comparées à celles que ces mêmes contrats auraient obtenu en suivant l'ancienne grille d'attribution.
Elles se sont révélées globalement cohérentes avec les anciennes, et avec l'intuition (croissantes avec l'encours et la part UC, meilleures pour des individus versant plus et rachetant moins, moins bonnes pour des assurés aux caractéristiques et comportements peu avantageux).
Ces analyses ont permis de vérifier que les **objectifs avaient été atteints** : le nouveau modèle de valeur client épargne, incluant une étape de segmentation du portefeuille selon les comportements, réussit à présent à capter les **différences de caractéristiques** entre assurés et à attribuer des meilleures ou moins bonnes valeurs en conséquence.

De plus, ce nouveau modèle de valeur client permet à présent à l'assureur d'**identifier** les clients au **potentiel** de **rentabilité** future le plus prometteur, de les **distinguer** des **clients plus risqués** et moins rentables, et de mettre en place une **stratégie** visant à améliorer la rentabilité et la valeur de son portefeuille, comme l'a montré la dernière partie du mémoire.
En ceci, la valeur client épargne semble donc être un élément pertinent à retenir en **complément des indicateurs actuariels traditionnels**.

Cependant, il convient de garder en tête que ces méthodes, comme tout modèle, ne sont pas parfaites, et présentent des **limites** :

- Sur les **données** utilisées : la qualité des prédictions est étroitement liée à la qualité des données, pas toujours facile à assurer. De plus, la base d'apprentissage, construite à partir d'un historique peu profond, était fortement déséquilibrée avec beaucoup de contrats "dormants". Enfin, certaines informations pertinentes sur les contrats n'étaient pour l'instant pas disponibles (multi-équipement, cause du ra-

chat, par exemple).

- Sur les **modèles** de *machine learning* employés : ils n'étaient pas forcément tous adaptés au déséquilibre dans les données ;

- Sur les **hypothèses** admises : les évolutions de PM et taux de PB admises dans la méthode 2 étaient acceptables mais améliorables.

Dans une optique de **perfectionnement** du modèle de valeur client épargne, plusieurs pistes seraient envisageables :

- Enrichir la base de données d'**années d'observations supplémentaires** et de **variables pertinentes** pour la modélisation de survenance ou de montant de flux : motivation du flux, muti-équipement chez AXA France, Top contrat Eurocroissance, satisfaction du client par exemple.

- Améliorer les modèles de prédiction sur la base **déséquilibrée** : les entraîner sur des échantillons *bootstrap* avec une contrainte sur la part d'observations de la classe minoritaire pour la classification binaire, et insérer une étape intermédiaire de classification avant la prédiction de montants de flux, qui serait alors uniquement réalisée sur les individus identifiés comme effectuant des flux ;

- Complexifier les **modèles de projection des variables explicatives** pour les rendre plus précis et performants : prendre en compte la revalorisation et les rachats totaux pour la PM, et plusieurs scénarios de performances financières pour les taux de PB.

En conclusion, le modèle de valeur client épargne est encore perfectible sur quelques points, mais globalement jugé satisfaisant puisqu'il **répond aux objectifs initiaux**. Il est d'ailleurs déjà en phase d'implémentation et d'opérationnalisation pour une utilisation dès 2021, afin de construire les nouvelles valeurs client épargne qui seront utilisées pour affecter un indice client global à tous les assurés en portefeuille.

Une **veille** devra cependant être effectuée, afin de vérifier que les modèles utilisés et présentés dans ce mémoire ne deviennent pas obsolètes et restent bien adaptés aux objectifs fixés ainsi qu'aux nouvelles données à disposition.

Enfin, un **travail d'explication et de communication** reste encore à mener, afin de permettre aux agents généraux de comprendre au mieux les valeurs client épargne et leur impact dans le projet plus global d'indice client.

Bibliographie

Références

[1] BICHOT A., (2016) *Assurance vie, approche économique de la Valeur Client*. Mémoire d'actuariat.

[2] BIERNAT E., LUTZ M. (2016) *Data science : fondamentaux et études de cas*. Nanterre : Eyrolles.

[3] BOHARI A.M., RUSLAN R. et ak. (2011) *Customer Lifetime Value Model in Perspective of Firm and Customer : Practical Issues and Limitation on Prospecting Profitable Customers of Hypermarket Business*. International Journal of Business and Management 6.8, 161-168.

[4] BREIMAN L., FRIEDMAN J., OLSHEN R., C. STONE C. (1984) *Classification and regression Trees*. Pacific Grove.

[5] CHEN C., LIAW A., BREIMAN L. (2004) *Using Random Forest to Learn Imbalanced Data*

[6] CORDEIRO DE AMORIM, R. (2011) *Learning feature weights for K-Means clustering using the Minkowski metric*, Thèse de PhD au département de l'informatique et des systèmes d'information de l'Université de Londres.

[7] CORLOSQUET-HABART M., JANSSEN J. (2017) "Le big data pour les compagnies d'assurance", ISTE Editions

[8] DONKERS B., VERHOEF P.C., DE JONG M.G. (2007) *Modeling CLV : A test of competing models in the insurance industry*. Quantitative Marketing and Economics 5, 163–190.

[9] DWYER F.R. (1997) *Customer lifetime valuation to support marketing decision making*. Journal of Direct Marketing 11.4, 6-13.

[10] FRIEDMAN J., HASTIE T., TIBSHIRANI R. (2001) *The Elements of Statistical Learning*. Springer.

[11] GALAR M., FERNANDEZ A. et al. (2012) *A Review on Ensembles for the Class Imbalance Problem : Bagging-, Boosting-, and Hybrid-Based Approaches*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463-484

[12] GUPTA S., HANSSENS D. et al. (2006) *Modeling Customer Lifetime Value*. Journal of Service Research 9.2, 139-155.

[13] HENNOM D., (2016) *Création d'un indicateur de valeur client en assurance non vie*. Mémoire d'actuariat.

[14] HINDAWI M., FIRESTONE G. (2012) *Customer Lifetime Value, Opportunities and challenges*

[15] HISQUIN R., (2020) *Segmentation du comportement client en assurance vie*. Mémoire pour la formation d'expert en Data Science pour l'actuariat.

[16] LITTLE RJA., RUBIN DB. (2002) *Statistical analysis with missing data*. 2ème édition, Wiley-Blackwell

[17] MODHA D.S., SPANGLER W.S. (2003) *Feature Weighting in k-Means Clustering*. Machine Learning, 52, 217–237.

[18] RANI S., SIKKA G. (2012) *Recent techniques of clustering of time series data : a survey*. International Journal of Computer Applications, 52 :15, p.1-9.

[19] THOUROT P., FOLLY K.A., (2016) *Big Data : Opportunité ou menace pour l'assurance ?*. Paris : Eyrolles.

[20] WANG X., SMITH K., HYNDMAN R. (2006) *Characteristic-based clustering for time series data*. Data Mining and Knowledge Discovery, 13 :3, p.335-364.

[21] XU Y., FU X. et al. (2018) *A K-means Algorithm Based On Feature Weighting*. MATEC Web of Conferences, vol. 232

[22] `https://www.axa.fr/epargne-retraite/assurance-vie/bonus-euro.html`
Description du Bonus Euro+ d'AXA France, site consulté le 12 mai 2020

[23] `https://register.eiopa.europa.eu/Publications/Reports/QIS5_Report_Final.pdf`
Rapport QIS5 de l'EIOPA, site consulté le 13 mai 2020

[24] `https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:335:0001:0155:fr:PDF`
Directive 2009/138/CE, site consulté le 14 mai 2020

[25] `https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R0035&from=EN`
Règlement délégué, site consulté le 14 mai 2020

[26] `https://scikit-learn.org/stable/modules/cross_validation.html`
Documentation sur la validation croisée, site consulté le 5 juin 2020

[27] `https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2#:~:text=Accuracy%20is%20used%20when%20the,as%20in%20the%20above%20case.`
Article démontrant que le F1 Score est plus pertinent que la mesure d'*accuracy*, site consulté le 9 juin 2020

[28] https://www-axa-com.cdn.axa-contento-118412.eu/www-axa-com%2F46b0ccd2-7730-490a-9986-cfe9b9adca07_axa_eev_2019.pdf
Rapport annuel EEV d'AXA en 2019, site consulté le 15 juin 2020

[29] http://www.cfoforum.eu/downloads/CFO-Forum_MCEV_Principles_and_Guidance_April_2016.pdf
Principes de MCEV dictés par le CFO Forum, site consulté le 15 juin 2020

[30] http://www.cfoforum.eu/downloads/CFO-Forum_MCEV_Basis_for_Conclusions_April_2016.pdf
Bases de la MCEV dictées par le CFO Forum, site consulté le 15 juin 2020

[31] banque-france.fr/statistiques/epargne-et-comptes-nationaux-financiers/epargne-des-menages
Bilan statistique des placements des ménages au 1er trimestre 2020, site consulté le 1er août 2020.

[32] http://alexminnaar.com/2014/04/16/Time-Series-Classification-and-Clustering-with-Python.html
Classification et Clustering de séries temporelles, site consulté le 07 juillet 2020.

[33] https://www.argusdelassurance.com/epargne/assurance-vie/assurance-vie.159109
Assurance-vie : 2019, une année exceptionnelle, site consulté le 1er août 2020.

[34] https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurance-vie-collecte-nette-negative-en-juin-2020
Assurance vie : collecte nette négative en juin 2020, site consulté le 1er août 2020.

[35] https://www.axa.com/fr/presse/publications/rapport-annuel-2019
AXA : Rapport annuel 2019, document consulté le 1er août 2020.

Annexe

Distribution des variables principales selon les différents *clusters* de rachat du Réseau3-Produit9

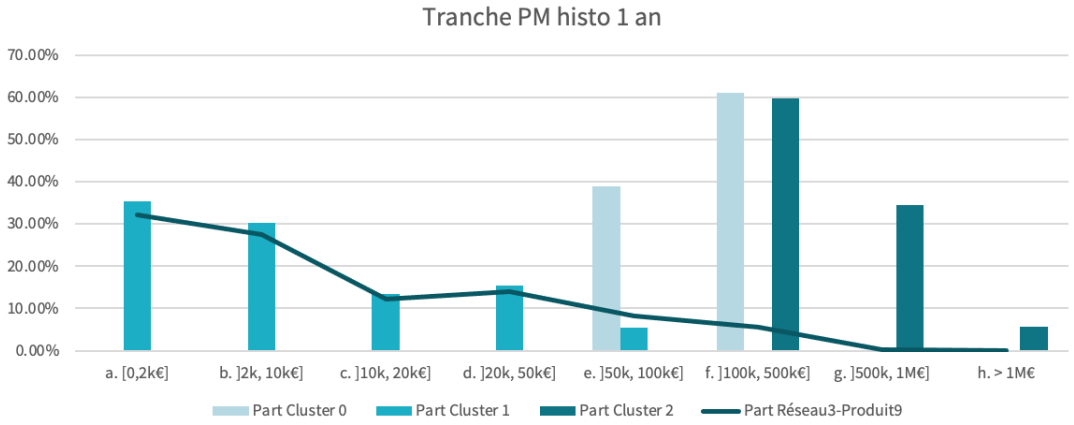

FIGURE 41 – Distribution des PM selon les *clusters* du Réseau3-Produit9

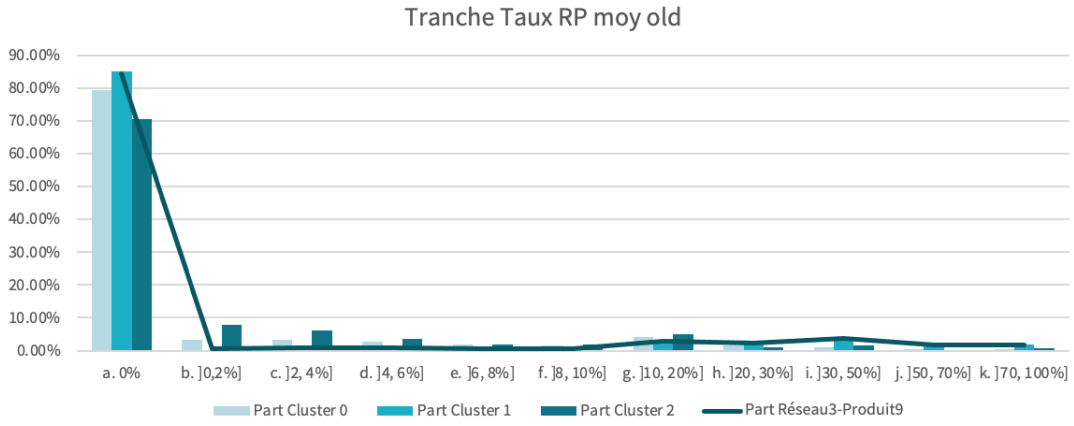

FIGURE 42 – Distribution des taux moyens historiques de RP selon les *clusters* du Réseau3-Produit9

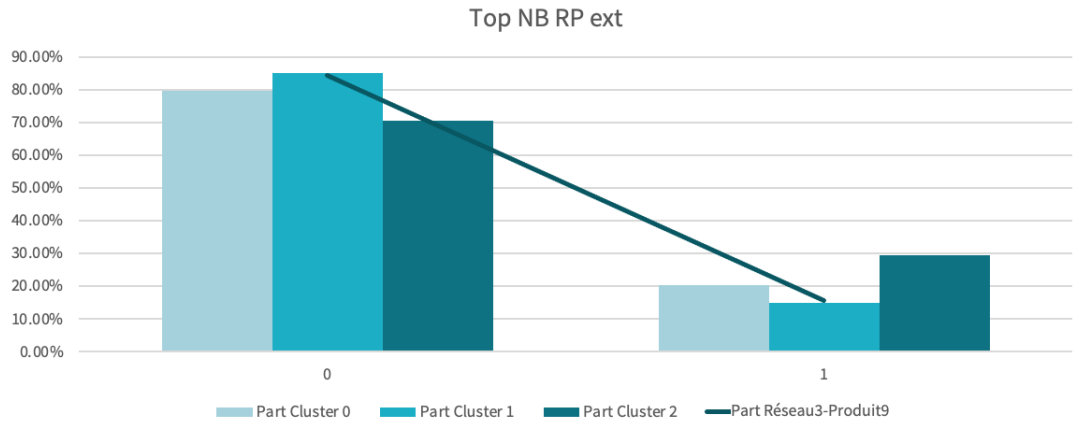

FIGURE 43 – Distribution du Top nombre de RP extrême selon les *clusters* du Réseau3-Produit9

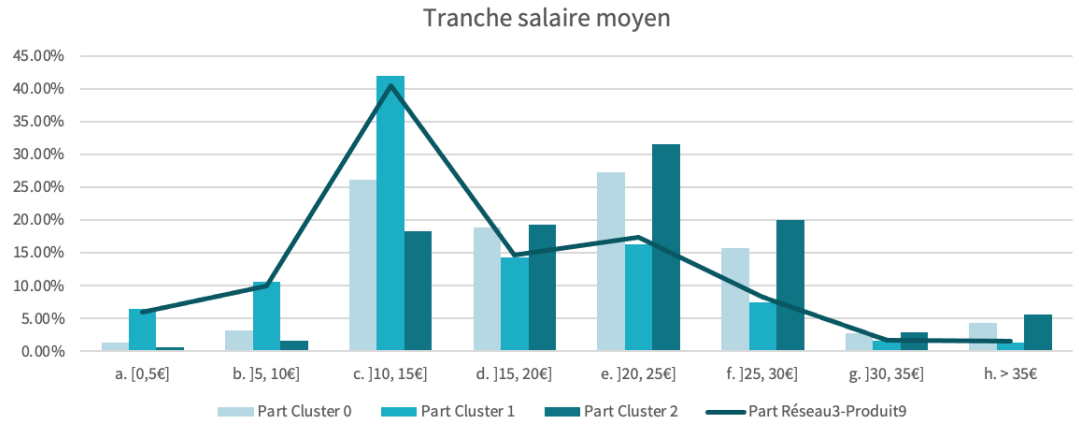

FIGURE 44 – Distribution des salaires moyens selon les *clusters* du Réseau3-Produit9

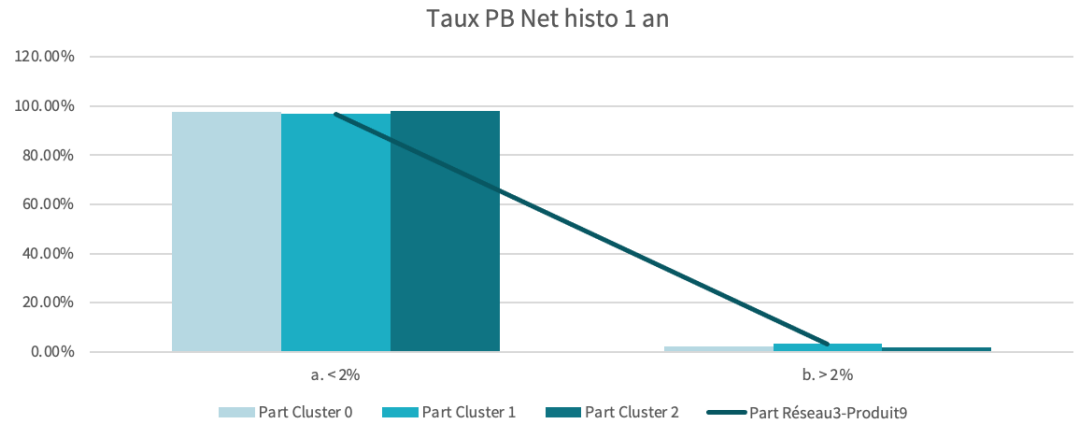

FIGURE 45 – Distribution des taux de PB net selon les *clusters* du Réseau3-Produit9

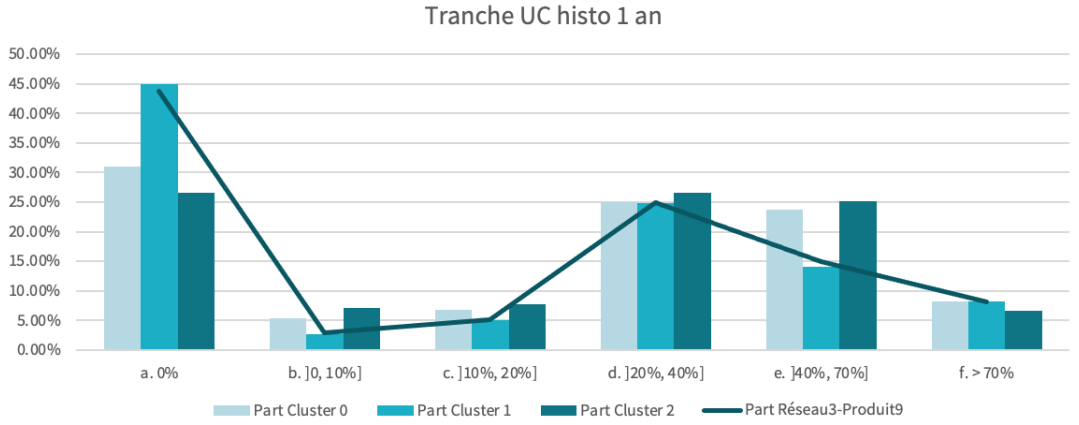

FIGURE 46 – Distribution des parts UC selon les *clusters* du Réseau3-Produit9

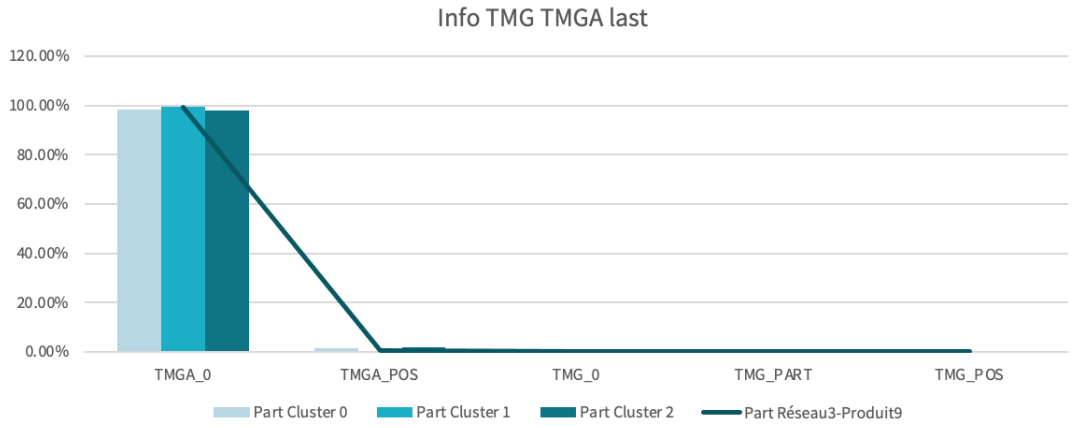

FIGURE 47 – Distribution des TMG/TMGA selon les *clusters* du Réseau3-Produit9

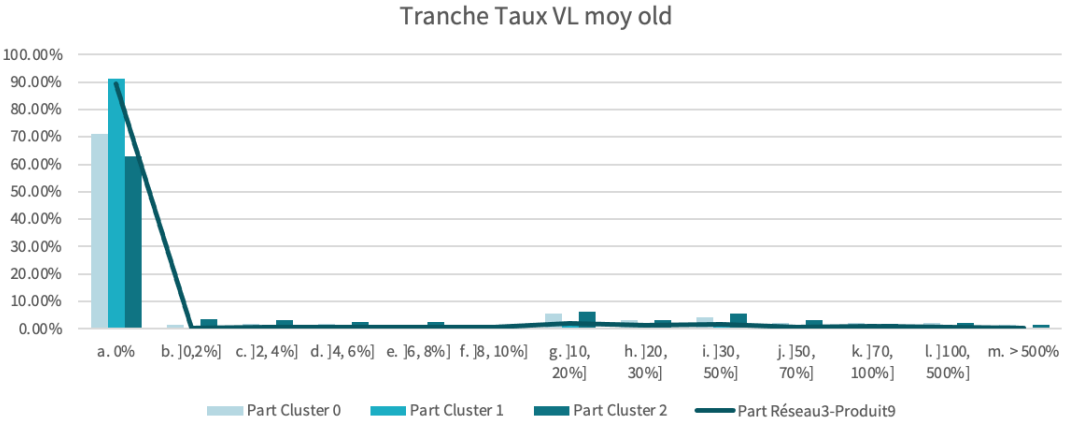

FIGURE 48 – Distribution des taux moyens historiques de VL selon les *clusters* du Réseau3-Produit9

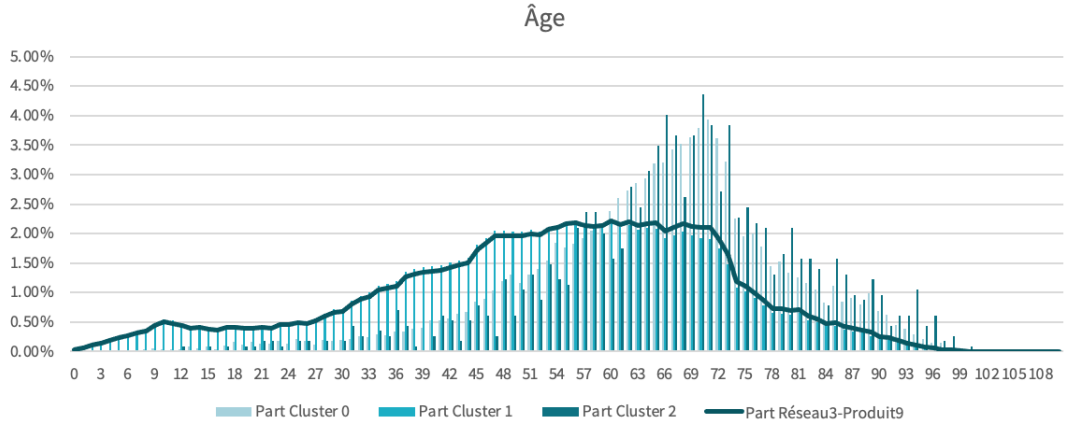

FIGURE 49 – Distribution des âges selon les *clusters* du Réseau3-Produit9

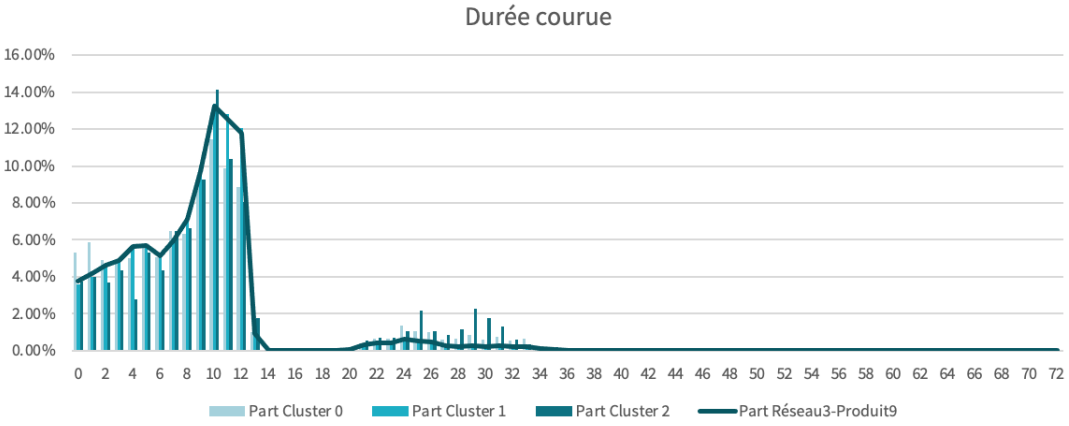

FIGURE 50 – Distribution des durées courues selon les *clusters* du Réseau3-Produit9