Mémoire présenté le :

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

Par : QUACH Hong Hanh

Titre : Modélisation de la loi de rachat structurel à l'aide des modèles Machine Learning

Confidentialité : ☒ NON ☐ (Durée : ☐ 1 an ☐ 2 ans)

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

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

. . . . . .

. . . . . .

. . . . . .

Membres présents du jury de l'ISFA

. . . . . .

. . . . . .

. . . . . .

Entreprise : Groupe Pasteur Mutualité
Nom : CAZIER Charles

Signature : DocuSigned by: Charles CAZIER CFCA383243A444E...

Directeur de mémoire en entreprise :
Nom : KABA Fatimata Bintou

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

DocuSigned by: Charles CAZIER CFCA383243A444E...

Signature du candidat

Résumé

Le risque de rachat fait partie des risques majeurs en assurance-vie. C'est l'une des options les plus importantes de l'assurance-vie car elle donne un sentiment d'investissement flexible au produit. Il est l'élément clé de la directive Solvabilité 2 et de la cinquième Etude quantitative (QIS5).

Cette étude porte sur le rachat structruel. Après une bref tour d'horizon de l'environnement réglementaire, fiscal et social de l'assurance vie, nous avons donc proposé de construire pas à pas une loi de rachat adaptée à notre portefeuille.

Le taux de rachat est exposé sous deux types : en nombre et en montant ; et deux aspects : statique et dynamique. Différents modèles sont proposés pour aborder le problème. Ces modèles sont : la régression logistique classique, l'algorithme CART, les forêts aléatoires, l'estimateur de Kaplan-Meier et l'analyse de survie basée sur un modèle d'arbre.

La construction de cette loi fait intervenir une étude approfondie, tant qualitative que quantitative, des causes de rachats. Ces modèles nous permettent d'obtenir la courbe de rachat en fonction de l'ancienneté qui s'adapte au modèle d'inventaire de GPM.

Mots-clés : *rachat structurel*, *régression logistique*, *CART*, *Kaplan-Meier*, *survival tree*, *solvabilité 2*.

Résumé

The surrender risk is one of the major risks in life insurance. It is one of the most important options in life insurance because it gives a flexible investment feeling to the product. It is also the key element of the Solvency 2 Directive and the Fifth Quantitative Study (QIS5).

This study focuses on the surrender structure. After a brief overview of the regulatory, fiscal and social environment for life insurance, we therefore proposed to build a new surrender curve step by step.

The surrender rate is presented in two types : in number and in amount ; and two aspects : static and dynamic. Different models are offered to address the problem. These models are : classical logistic regression, the CART algorithm, random forests, and the Kaplan-Meier estimator and survival analysis based on a tree model.

The construction of this law involves an in-depth study, both qualitative and quantitative, of the causes of surrender. These models allow us to obtain the surrender curve as a function of seniority which adapts to GPM's inventory model.

Key words : *rachat structurel*, *logistique regression*, *CART*, *Kaplan-Meier*, *survival tree*, *solvency 2*.

Remerciement

J'adresse un grand merci à mon tuteur de stage KABA Fatimata Bintou, Chargée d'études actuarielles et ce pour sa disponibilité et ses pertinentes orientations qui m'ont permis de progresser et de réaliser ce travail.

Je témoigne toute ma reconnaissance envers tout le personnel pôle Actuariat de l'entreprise Groupe Pasteur Mutualité.

Je tiens aussi à exprimer ma reconnaissance envers M. CAZIER Charles, Directeur de l'actuariat, pour m'avoir accueilli au sein du Groupe Pasteur Mutualité ainsi que pour toute la bienveillance qu'il a manifesté à mon égard.

Mes remerciements vont également à mes professeurs qui ont su par leur présence, leur compétence, leur persévérance et leur patience nous former puis nous guider et nous livrer les secrets de notre futur métier.

Je remercie aussi Madame Diana DOROBANTU, mon tuteur pédagogique, pour sa disponibilité.

Pour finir, je tiens à exprimer ma profonde gratitude à ma famille qui m'a toujours soutenue ainsi qu'à toutes les personnes qui, de près ou de loin, ont participé à l'élaboration de ce mémoire.

Table des matières

Table des figures

Liste des tableaux

Introduction

Parmi les différents risques auxquels les assureurs-vie sont exposés, le risque de rachat est l'un des plus importants. Il donne au contrat d'assurance des options flexibles qui aident l'assurance-vie à devenir l'un des moyens les plus utilisés pour investir en France.

Il attire beaucoup d'attention dans la modélisation du risque d'assurance-vie car le rachat peut affecter directement la liquidité de l'entreprise. Qu'il s'agisse d'une vague de masse de rachat en rachat dynamique ou de rachat normal provoqué par le comportement de l'assuré, il a un impact important sur la solvabilité de l'assurance. C'est pourquoi, le rachat est un élément clé de l'exigence de capital Solvabilité II (SCR) pour de nombreux assureurs-vie.

De plus, dans le cadre de la mise en œuvre de la réforme prudentielle Solvabilité 2, les calculs de Best Estimate ou bien on dit de projection du passif dépend fortement des hypothèses du modèle. L'hypothèse qui gère les produits épargne est la loi de rachat. En effet, le calcul du SCR s'effectue en calculant les SCR par risques, notamment celui du rachat.

Il rend la recherche d'un modèle de taux de rachat pour un portefeuille épargne indispensable pour les assureurs. Il y a 2 types de rachat défini dans les garanties d'assurance : le rachat total où l'assuré récupère tout son compte de provision mathématique et le rachat partiel où l'assuré récupère une partie de sa provision mathématique. Nous modélisons chaque type de rachat séparément. Il permet de s'adapter au mieux afin que le modèle ALM existant de la compagnie ait une estimation de la prestation globale la plus précise.

Selon les règlementations de Solvabilité II, le rachat est catégorisé en rachat structurel et rachat conjoncturel. Les rachats sont dits structurels si l'analyse des rachats se base sur le comportement des assurés : leur âge, l'ancienneté de leur contrat, le sexe, la catégorie socioprofessionnelle, le réseau de distribution. Tandis que le rachat conjoncturel dépend de la situation économique. Ce document a pour objet de synthétiser et décrire les méthodologies de calculs du taux de rachat structurel des produits d'épargne en euro de Groupe Pasteur Mutualité où le rachat structurel est plus important.

Dans un premier temps, cette étude abordera la problématique en rappelant les différentes définitions liées à l'assurance-vie et au risque de rachat. Nous analysons également le

marché actuel de l'assurance avec un aperçu rapide de l'impact de la pandémie de Covid-19 sur l'assurance-vie. Nous donnons une étude pour connaître la position où le risque de rachat intégrer dans la Solvabilité 2.

Dans un second temps, une analyse descriptive du portefeuille d'épargne en euros de GPM est réalisée à la suite de l'analyse univariée et bivariée. Ensuite, nous nous assurons que nos données ne sont pas constituées de valeurs aberrantes et que les variables ont une forme discrète en utilisant l'algorithme de l'arbre de décision.

Ces résultats sont utilisés dans la section suivante où nous analysons le problème d'un point de vue statique. L'idée principale est de donner un aperçu du fonctionnement du portefeuille, de ses caractéristiques. A l'aide du modèle logistique, de l'arbre de décision et de la forêt aléatoire, nous obtenons certaines caractéristiques importantes du portefeuille et des variables qui semblent les plus discriminantes dans notre modèle.

Dans la section suivante, nous considérons un point de vue dynamique de la structure de rachat telle que nous la modélisons en utilisant le modèle de Kaplan Meier. Ce modèle est comparé avec le modèle existant dans au seins de GPM.

Dans la dernière partie, nous proposons une méthode différente qui combine un modèle d'apprentissage automatique et la théorie des modèle de durée.

Chapitre 1

Contexte de l'étude sur la loi de rachat

Dans cette section, nous introduisons certains des concepts du champ d'assurance qui seront utilisés pour la construction de la loi de rachat dans les sections suivantes. On rappelle les caractéristiques principales d'un contrat d'assurance vie qui en font un véritable produit d'épargne. Nous résumons ensuite l'environnement réglementaire dans lequel ces contrats opèrent ainsi que les risques de rachat en mettant l'accent sur Solvabilité II.

1.1 Les contrats d'assurance vie

1.1.1 Définition d'un contrat d'assurance vie

L'assurance est un contrat, représentée par une police, en vertu duquel une personne physique ou morale reçoit une protection financière ou une indemnisation d'une compagnie d'assurance. La société mutualise le risque des clients pour rendre les paiements plus abordables pour les assurés.

Parties liées dans un contrat d'assurance :

L'assureur : L'assureur est la partie qui s'engage une indemnisation au bénéficiaire du contrat d'assurance en cas de sinistre. En droit des assurances, il peut s'agir d'une société commerciale (SA), d'une société civile (SAM), d'une société européenne, voire d'un intermédiaire d'assurance (agent général d'assurance ou courtiers).

Le souscripteur : dénommé « contractant » ou « preneur d'assurance » en droit communautaire, est la partie qui signe les documents contractuels et qui paie les primes.

L'assuré : la personne physique ou morale sur la tête (en assurance-vie) ou sur les intérêts (assurance dommage) de qui pèse le risque assuré. Il n'est pas nécessairement le souscripteur du contrat.

Le montant que l'adhérent doit payer à la compagnie d'assurance en échange d'une protection s'appelle la prime. En plus des primes, divers frais sont facturés pendant toute la durée du contrat. Il se compose principalement de frais d'entrée (frais de dossier, frais de paiement, etc.) ou de frais de sortie (frais d'arbitrage, d'options, etc.), déterminés préalablement à la signature du contrat.

L'assurance-vie est associée à l'évènement de décès, ce qui signifie que la compagnie d'assurance verse un capital ou une rente lorsqu'une personne décède ou ou en cas de survie. Cela donne à l'assurance-vie un sens de protection et même d'épargne. Soutenir les produits d'assurance-vie nécessite des options complexes pour augmenter la flexibilité du compte d'épargne telles que : l'arbitrage, le rachat et les avantages fiscaux.

Selon la Fédération française de l'assurance (FFA), l'assurance-vie est la principale forme d'épargne en France. En 2020, la valeur des contrats d'assurance-vie dépasse 1 700 milliards d'euros. Fin juillet 2021, les contrats d'assurance-vie s'élevaient à 1 848 milliards d'euros, soit une augmentation de plus de 5% en un an et de plus de 3,47% sur les sept mois du début de l'année (FFA).

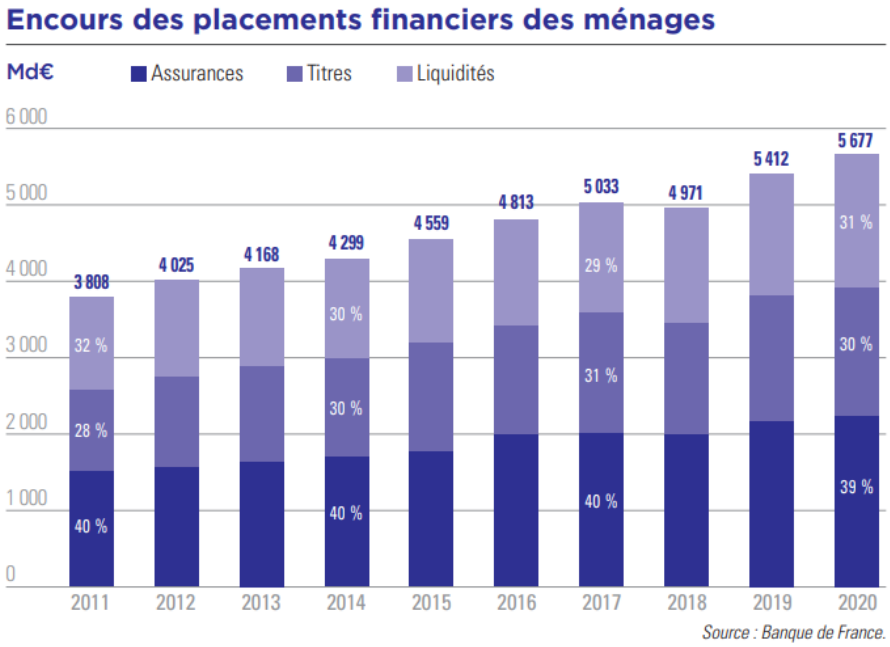


FIGURE 1.1 – Des placements financiers des ménages en 2020

En dehors de l'assurance-vie, il existe d'autres moyens d'allocations de l'épargne : livret A,

épargne logement et paiement de dépôt. Cependant, l'assurance-vie demeure une option concurrentielle. Les contrats d'assurance-vie sont courants parmi les comptes d'épargne puisque 39% des ménages métropolitains en sont propriétaires en 2018. Cependant, les fonds en euro et l'unité de compte se concentrent dans la tranche de revenu élevé.

L'intérêt de souscrire un contrat d'assurance-vie peut se justifier pour trois raisons principales. Premièrement, il offre aux épargnants la possibilité de bénéficier de certains régimes fiscaux. Deuxièmement, l'assurance-vie offre la possibilité de bénéficier d'une épargne en prévision de la retraite qui couvre aussi le risque de décès et de dépendance. Troisièmement, il existe une option qui offre la liberté d'optimiser adéquatement la rentabilité du contrat en arbitrant entre la liquidité d'un produit et son risque.

Bien que le marché de l'assurance-vie s'étende, il a dû faire face à de nombreuses difficultés ces dernières années. Avec les taux bas, l'assurance-vie, notamment le fonds en euros, doit faire face à une difficulté pour respecter le taux minimum garantie. En effet, les taux garantis aux épargnants sur la durée du contrat restent peu élevés pour les supports en euros des contrats individuels et représentent la majeure partie des passifs.

La période de 2019 à 2021 est très particulière alors que le monde traverse le Covid 19. Cette pandémie a touché tous les aspects de la vie humaine, notamment la situation économique. Il nous est difficile de dire quel est l'impact de la pandémie sur le secteur de l'assurance. Tout d'abord, avec les développements compliqués de Covid, la France a traversé cinq vagues de pandémie, il est trop tôt maintenant pour voir son plein impact dans le secteur de l'assurance. Deuxièmement, la restriction de nombreuses activités dans de nombreuses régions du monde nous a mis dans une situation inédite où la question : « Trouverons-nous le nouveau style de vie pour co-vivre avec Covid ». La pandémie pourrait changer la façon dont les gens vivent, mangent et investissent.

Cependant, au moins pour l'instant, nous avons vu ses impacts temporaires. En effet, nous avons observé une baisse du montant de la collecte brute sur l'ensemble de tous supports en 2020. Parallèlement, la collecte nette a été suffisamment dynamique les deux dernières années, notamment les trois premiers trimestres de l'année 2019. Le taux de rendement de l'assurance-vie en 2020 a atteint 1,30% selon les données publiées par la FFA (Fédération française de l'assurance), soit une baisse modérée de 0,16 point par rapport à l'année précédente. Mi-mars 2020, lors du premier confinement, on observe également une forte augmentation des rachats d'assurances.

1.1.2 Des produits d'assurance vie et des types de contrat

L'assurance-vie se divise en 3 types : l'assurance-vie, l'assurance-décès et le contrat mixte vie-décès. Les prestations sont versées de diverses manières qui permettent une flexibilité dans la planification et la distribution financière du ménage, telles que : en capital, en rente viagère, en rente temporaire.

Dans ce mémoire, l'assurance-vie se réfère à l'assurance en cas de vie au sens de l'épargne.

L'assureur peut combiner une ou plusieurs des garanties suivantes pour créer un produit :

- Garantie de taux minimal de revalorisation
- Garantie de participation aux résultats technique et financier
- Garantie de remboursement avant terme de l'épargne constituée (rachat)
- Garantie de remboursement minimum des contrats en UC (plancher)

Sa complexité réside dans sa dépendance vis-à-vis des marchés financiers. Par conséquent, les normes comptables et les cadres réglementaires jouent un rôle important dans la gestion de ces produits.

Il existe deux types de contrats d'assurance-vie : monosupport et multisupport. Quand les contrats proposent un seul support d'investissement, ils sont dits monosupport. Tandis que les contrats multisupports sont constitués d'au moins un fonds en euros et un ou plusieurs supports en unités de compte (obligations, OPCVM, actions, parts de sociétés immobilières, etc.).

Les supports en unités de compte sont structurés de manière similaire à celle des fonds communs de placement, en ce qu'elles regroupent les investissements avec ceux d'autres investisseurs. L'assuré paye la prime en échange de la couverture contre l'évènement de décès et un certain nombre d'unités dans les fonds. La valeur du compte d'épargne dépend des valeurs actuelles de chaque unité. Par conséquent, l'assureur porte le risque.

Au contraire, dans les contrats en euros, les adhérents ont un **Taux garantie minimum**. Réglé selon art. A.132-1 et A.132-3 du Code des assurances, le **Taux garanti minimum** (TMG) est un taux garanti pour une durée n'excédant pas deux ans. Il se compose du taux d'intérêt technique et d'une partie de la participation aux bénéfices. Le TMG est fixé par l'assureur avec un plafond de 75% du taux des emprunts d'État (TME) pendant les huit premières années du contrat et 3,5 % par an ou 60 % du TME par la suite.

Une caractéristique distinctive des fonds en euros est qu'ils fournissent des garanties de capital et toutes les sommes versées ainsi que l'intérêt produit par le TMG ne peuvent baisser. Cette garantie oblige les assureurs à être plus prudent dans la gestion des actifs. Par conséquent, l'assureur préfère des investissements sûrs et amortissables. La majorité des fonds en euros se compose de plus de 80% d'obligations, 10% d'actions et 5-10% d'immobilier.

1.2 L'option de rachat

En assurance-vie, pour donner au client plus de flexibilité dans la gestion de ses biens, l'assureur conçoit le produit avec différentes options sur les mouvements de compte épargne. Des options qui apparaissent très souvent dans les contrats d'assurance-vie sont :

- La clause de participation aux bénéfices
- L'option de rachat : permet à l'adhérent de retirer de l'argent de son compte d'épargne
- L'arbitrage : l'option de changer la répartition du portefeuille d'investissement
- Le versement : permet à l'adhérent de mettre plus d'argent sur son compte d'épargne

L'option de rachat est une des options les plus importantes. En fait, c'est la principale source de prestation de l'assurance-vie puisqu'en 2020, 56% des prestations proviennent de ce motif (FFA). Le Règlement Délégué (UE) 2015/35 Article 1 a défini le « rachat » comme « tout moyen de résilier partiellement ou complètement un contrat ». L'assuré a la permission de racheter à tout moment. Contrairement à un sinistre, l'assuré ne sera pas pénalisé pour cette option. En d'autres termes, cela n'affectera pas leur future possibilité de signer un autre contrat. Cependant, ils pourraient devoir payer une somme d'argent au titre des frais de rachat. La somme payée par la compagnie d'assurance est appelée la valeur de rachat.

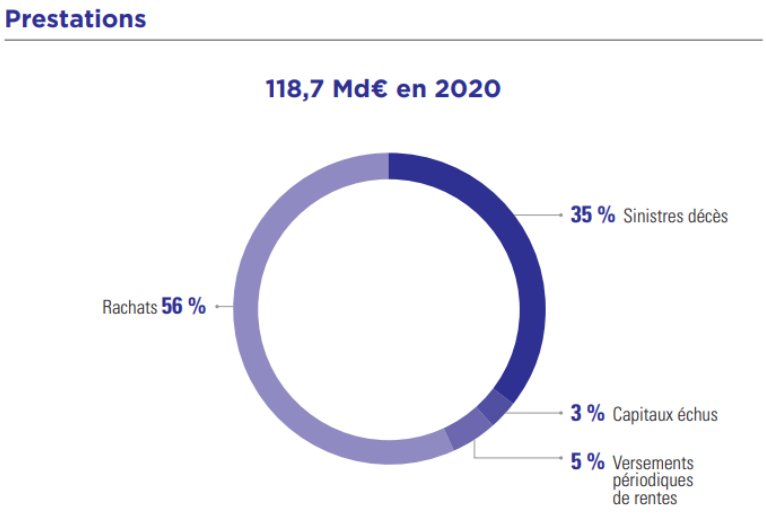


FIGURE 1.2 – Prestations d'assurance vie en 2020

Les adhérents peuvent choisir entre 2 types de rachat : le rachat total et le rachat partiel. Le rachat total est le cas où l'assuré résilie complètement son contrat. Le rachat partiel, en revanche, est le cas où une partie de valeur du compte épargne est demandée volontairement par l'assuré. La valeur de rachat est versée dans un délais de 30 jours après la réception de demande.

Parmi les rachats partiels, dépendant du plan financier des ménages, il en existe deux formes : le ponctuel et le programmé. Le rachat partiel ponctuel répond à un besoin de l'adhérent de récupérer la somme d'argent qui lui est nécessaire. Le rachat partiel programmé peut être mis en place pour bénéficier régulièrement du versement d'une

somme déterminée qui viendra augmenter ses revenus (mensuels ou trimestriels), la somme déterminée étant versée au rythme choisi jusqu'à épuisement du capital acquis.

Une des caractéristiques qui fait du contrat d'assurance-vie un moyen d'investissement intéressant est que la personne qui détient un contrat peut bénéficier d'un avantage fiscal.

- *Fiscalité de rachat*

Selon la Fédération française des assurances, les revenus de l'activité d'assurance sont imposés. En effet, l'assuré peut choisir d'être imposé soit sur le revenu soit sur un Prélèvement Forfaitaire Unique (PFU). En cas d'intégration à la déclaration de revenus, ces intérêts sont soumis en fonction du barème d'imposition. A défaut, le taux du prélèvement forfaitaire est fixé en fonction de la durée du contrat comme :

Pour les versements avant 27/09/2017			**Pour les versements après 27/09/2017**		
>4 ans	Entre 4 ans et 8 ans	< 8 ans	>8 ans	< 8 ans	
				Versements de 150 000 euros maximum	Versements supérieurs à 150 000 euros
35%	15%	7,5%	12,8%	7,5%	7,5% applique à 150 000 et 12,8% à la fraction excédentaire

FIGURE 1.3 – Taux de prélèvement forfaitaire de rachat

Un abattement annuel de 4 600 euros s'applique aux rachats effectués après 8 ans. En général, détenir un contrat pendant plus de 8 ans donne à l'assuré un avantage fiscal plus important. C'est l'un des facteurs importants influençant le comportement des adhérents, en particulier ceux qui possèdent un grand compte d'épargne.

1.3 Environnement règlementaire solvabilité 2

1.3.1 Contexte et définition de la solvabilité

Les objectifs principaux de Solvabilité II sont d'augmenter le niveau de coordination de la surveillance de la solvabilité en Europe, de protéger les assurés, d'introduire des exigences de fonds propres plus sensibles à l'échelle européenne (par rapport aux précédentes exigences minimales de solvabilité) au niveau du risque impliqué et de fournir des incitations adéquates pour une bonne gestion des risques.

L'EIOPA (l'Autorité européenne des assurances et des pensions professionnelles, l'un des principaux organes de surveillance financière de l'UE et issu de l'organisme anciennement

connu sous le nom de CEIOPS) a fourni des conseils techniques et un soutien à la Commission européenne pour l'élaboration des actes délégués (qui fournissent des mises en œuvre des orientations et la directive globale) et était responsable de l'élaboration de certaines des normes techniques et des orientations supplémentaires.

Selon FFA, la solvabilité d'une compagnie d'assurance est sa « capacité à respecter les engagements qu'elle prend envers ses clients ». Jusqu'à la crise de 2007-2010, Solvabilité I était encore en service. Cette crise a été le moteur du durcissement des réglementations en matière de gestion des risques dans les secteurs bancaires, des assurances et de la finance, dont Bâle III pour les banques et la Solvabilité II pour le secteur des assurances. Solvabilité II met l'accent sur le principe de « prudence ».

Par conséquent, la mission du développement de Solvabilité II est d'améliorer Solvabilité I en évaluant et en contrôlant le risque financier systémique. Il est à noter que, comme l'a indiqué la FFSA, le secteur de l'assurance n'a pas été à l'origine de la crise, ni le plus touché. Cependant, sa mission est de stabiliser et d'aider à sortir de la crise. Il est donc très important de s'assurer de la solvabilité de la compagnie d'assurance. La réforme de Solvabilité II est entrée en vigueur le 1er janvier 2016. Les compagnies d'assurances ont eu des années de préparation.

Cette norme se décompose en 3 piliers.

- **Pilier 1 :** aspects quantitatifs, concernant le bilan et la solvabilité
- **Pilier 2 :** aspects qualitatifs, tels que la gouvernance, la gestion des risques au sens large, le processus de supervision solvabilité (ORSA)
- **Pilier 3 :** informations à publier à destination du public et du superviseur.

La provision technique est calculée comme la somme du *Best Estimate* et de la « Marge de risque » qui sont respectivement la meilleure estimation des flux futur et la marge pour prendre en compte les incertitudes des *Best Estimate*.

Le cœur du premier pilier de Solvabilité 2 qui sont définies dans l'article 64 de la directive Solvabilité II est :

- Le SCR (*Solvency Capital Requirement*) : : le minimum de fonds propres que doit posséder la compagnie afin de couvrir avec une probabilité de 99,5% un risque de faillite à horizon un an
- Le MCR (*Minimum Capital Requirement*) : le minimum de fonds propres indispensable pour exercer l'activité d'assurance

Le deuxième pilier porte sur la gestion du contrôle interne des risques ou ERM (Enterprise Risk Management) à travers la mise en place du dispositif « Own Risk and Solvency Assessment (ORSA) ». Le but est également d'harmoniser les différents contrôles des pays européens.

L'objectif principal du troisième pilier est de définir les informations accessibles au grand

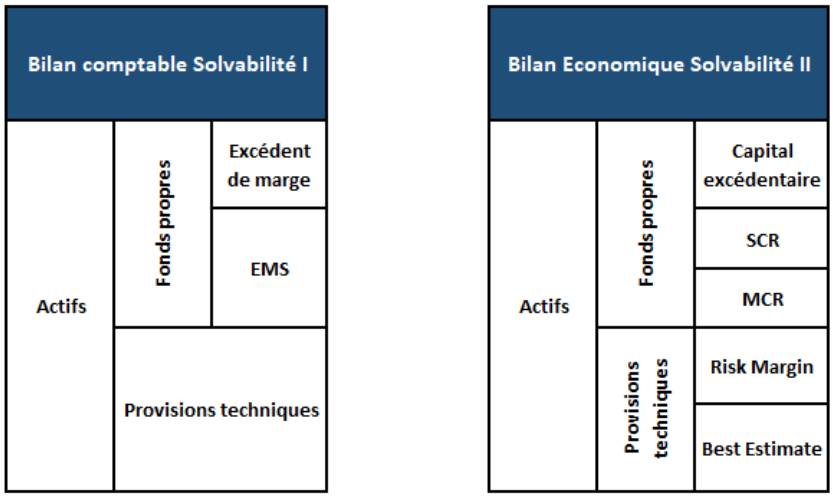


FIGURE 1.4 – Comparaison du bilan de solvabilité

public et accessibles aux autorités de contrôle, afin de pouvoir exercer leur pouvoir de surveillance. Ces informations, tant quantitatives que qualitatives, devraient être présentées annuellement et, pour certaines, trimestriellement. Ils visent à préciser le contenu attendu du rapport de solvabilité et la situation financière (« SFCR ») et le rapport régulier au superviseur (« RSR »).

Ces informations couvrent les éléments suivants :

- Performance financière
- Profils de risques, données et hypothèses sur lesquelles ils sont basés
- Mesures d'incertitudes, incluant la mesure d'adéquation des estimations antérieures et la sensibilité des résultats à la volatilité du marché...

La norme est exprimée au travers des Etudes Quantitatives d'Impact (Quantitatives Impact Studies, QIS). Le dernier QIS est le cinquième qui fait publié en mars 2011 par l'EIOPA. Le QIS fournit des informations sur l'impact quantitatif des réformes. Il réglemente spécifiquement les exigences techniques, les exigences de capital de solvabilité (SCR) et la classification des fonds propres.

1.3.2 Formule standard

Solvabilité II oblige les compagnies d'assurance à calculer le SCR pour s'assurer de leur capacité à payer leurs obligations à l'avenir. C'est aussi la centrale des premières pilier dont on a parlé. Les calculs sont effectués à l'aide de formules standard, de modèles internes ou de modèles internes partiels. Le SCR et le MCR de GPM sont calculés sur la base de la formule standard.

Le SCR se définit dans l'Article 104 de la Directive Solvabilité II par la formule :

$$SCR = BSCR - Adj + SCR_{op}$$

où

- BSCR (Basis Solvency Capital Requirement) est le montant de fonds propres requis de base
- Adj : Ajustement - effet d'absorption
- SCR_{op} : Le capital règlementaire requis pour couvrir le risque opérationnel

Nous présentons le terme NAV (Net asset value) qui correspond à la différence entre la valeur marchande de l'actif et le montant de *Best Estimate*. La cartographie des risques est fixée dans la formule standard. Le BSCR se décompose en 6 sous-modules. Les sous-modules, selon l'article 105 et le QIS5, sont suivants :

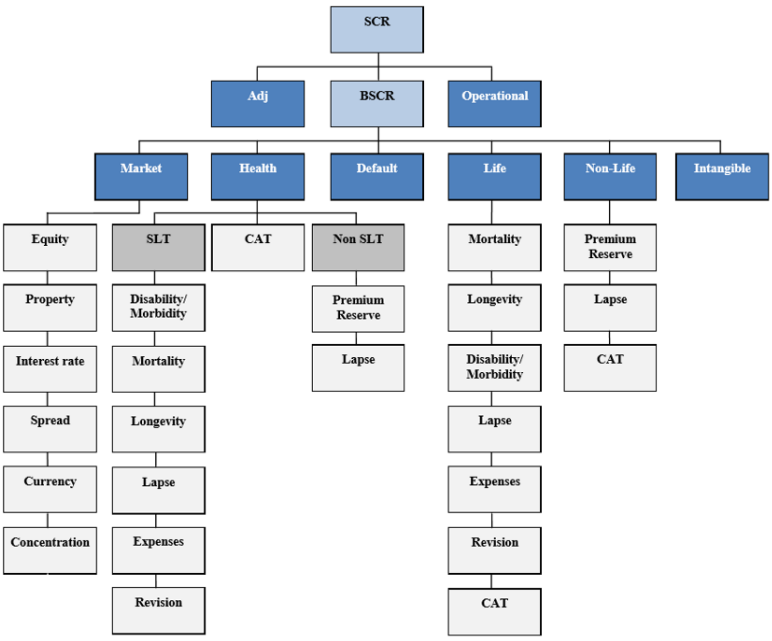


FIGURE 1.5 – Cartographie des risques

- Risque de marché : des risques liés aux fluctuations de la valeur des différents instruments financiers liés aux actifs et passifs de l'entreprise.
- Risque de santé : le risque découlant de l'assurance santé et de réassurance qu'elle soit modélisée comme une assurance vie ou non-vie
- Risque de défaut de contrepartie : il reflet les pertes potentielles causées par des défauts de qualité de crédit imprévus

- Risque vie : le risque découlant du processus de gestion de l'activité de l'assurance vie
- Risque non-vie : le risque découlant du processus de gestion de l'activité de l'assurance non-vie
- Risque des actifs incorporels : y compris les risques de marché (perte des prix dans le marché des actifs, manque de liquidité...) et les risques internes.

Le BSCR est calculé selon la méthode « bottom-up ». En d'autres termes, nous calculons le SCR de chaque sous-module, puis les ajoutons pour obtenir le BSCR. Il est défini dans la formule suivante :

$$BSCR = \sqrt{\sum_i \sum_j Corr_{i,j} \times SCR_i \times SCR_j} + SCR_{\text{intangibles}}$$

On a

$$SCR_{intangibles} = 0,8 \times IA$$

où IA = valeur des actifs incorporels

$Corr_{i,j}$ correspondent à les corrélations entre les sous-modules qui sont déterminés en QIS5 dans la matrice de corrélation suivante :

j i	Market	Default	Life	Health	Non-life
Market	1				
Default	0.25	1			
Life	0.25	0.25	1		
Health	0.25	0.25	0.25	1	
Non-life	0.25	0.5	0	0	1

FIGURE 1.6 – Matrix de corrélation des sous-modules

Dans le calcul, les SCR_i et SCR_j sont remplacés par SCR_{mkt}, SCR_{life}, $SCR_{non-life}$, SCR_{health}, $SCR_{default}$. L'idée de couvrir un évènement bicentenaire se transmet en appliquant le choc instantané sur chaque facteur de risque. Le SCR va se calculer avec la variation ΔNAV qui est la différence de NAV en scénario choc et en hypothèse centrale.

1.3.3 L'impact de Solvabilité sur l'inventaire de GPM

Ici, nous reprendrons l'activité de valorisation de GPM en 2020 au fur et à mesure de l'impact de ce cadre sur la technique de provision et le bilan de l'entreprise. L'inventaire du Groupe Pasteur Mutualité est réalisé sous réglementation S1 et S2. Ce tableau ci-dessous nous montre les provisions techniques calculées avec la norme S1 :

31/12/2020 (en €)	PSAP 31/12/2020	PM 31/12/2020	Total	Ligne d'activité
Altiscore Epargne €	775 199	487 187 929	487 963 128	Assurance avec PB
Altiscore Epargne uc	0	82 257 263	82 257 263	Assurance indexée et un uc
Altiscore Rentes viagères	7 947	57 676 602	57 684 549	Assurance avec PB
Repag	0	230 510 629	230 510 629	Assurance avec PB
AGMF Epargne	1 331	32 437 668	32 438 999	Assurance avec PB
Dexia	32 664	1 877 978	1 910 643	Assurance santé
VE KO	1 451	477 211	478 662	Assurance avec PB
Parmateam	209 989	277 998	487 987	Autre assurance vie
GAV	201 110	0	201 110	Autre assurance vie
CIR SENOIS	125 442	0	125 442	Assurance de protection du revenu
Assor	84 866	0	84 866	Assurance de frais médicaux
Prévoyance forfaitaire	7 019	0	7 019	Assurance de protection du revenu
ADOHA	274 802	527 911,73	802 714	Assurance de protection du revenu
Hospi	1 333	0,00	1 333	Assurance de frais médicaux
Total provisions hors PPE	1 723 154	893 231 190	894 954 344	
PPE		36 971 514	36 971 514	
Total provisions techniques	**1 723 154**	**930 202 704**	**931 925 858**	

FIGURE 1.7 – Provisions technique Solvabilité I

Les deux dispositions concernées en S1 sont les Provisions Mathématiques (PM) et les Provision pour les sinistres à payer (PSAP). En effet, le changement de Solvabilité I à Solvabilité II a eu un grand impact sur la provision technique de l'entreprise. On peut le voir dans le tableau suivant des Provision Technique calculées avec les deux cadres réglementaires segmentés par différents LOB (Line of Bussiness) de GPM :

31/12/2020 (en €)	PT S1 brut réassurance	PT S2 brut réassurance	Best estimate (BE)	BE cédé	Marge pour risque
Lob 1 : frais médicaux	86 199	187 336	183 623	0	3 712
Lob 2 : protection du revenu	584 772	652 896	639 957	50 951	12 939
Lob 29 : santé SLT	1 910 643	1 984 231	1 944 909	1 361 003	39 322
Lob 30 : vie avec PB	809 075 967	987 148 401	966 974 371	0	20 174 030
Lob 31 : unités de compte	82 257 263	80 525 143	78 929 351	0	1 595 793
Lob 32 : autre assurance vie	818 062	855 340	837 860	289 088	17 480
Lob 33: Rentes issues de contrats d'assurance non-vie et liées aux obligations d'assurance maladie	221 438	244 326	239 484	23 944	4 842
Total	**894 954 344**	**1 071 597 673**	**1 049 749 554**	**1 724 987**	**21 848 119**

FIGURE 1.8 – Provisions technique Solvabilité II

1.4 Le risque de rachat structurel

Selon Solvabilité II, la compagnie d'assurance s'expose aux risques suivants :

- Les risques de marché
- Les risques de crédit
- Les risques vie
- Les risques IARD
- Les risques opérationnels

Parmi ce risque, le rachat est un risque très important que les compagnies d'assurance doivent surveiller. En effet, les rachats massifs causent des problèmes de liquidités et forcent la vente d'actifs.

De nombreuses recherches ont été menées pour étudier et modéliser ce risque car il peut affecter grandement la santé financière de la compagnie d'assurance, notamment, Albizzati & Geman (1994) ; Bacinello (2003) ; Grosen & Jørgensen (2000). Surtout dans le contexte de Solvabilité II, lorsque les réglementations concentraient à la « prudence », les auteurs comme Burkhart (2018) disposent également d'études précieuses pour approfondir cette question.

Il pourrait également entraîner une perte de bénéfices futurs potentiels (Kuo et al. (2003)). Plus précisément, les rachats précoces pourraient entraîner des pertes substantielles si l'assureur n'est pas en mesure de récupérer les coûts d'acquisition.

En outre, l'option de rachat peut renforcer la sélection défavorable, surtout lorsque les adhérents peuvent racheter avec de faibles frais. Les rachats diminuent également l'efficacité de la mutualisation des risques. Enfin, des taux de rachat élevés peuvent avoir un effet négatif sur la réputation de l'assureur, ce qui peut entraîner le rachat d'un plus grand nombre d'assurés et nuire à de nouvelles affaires.

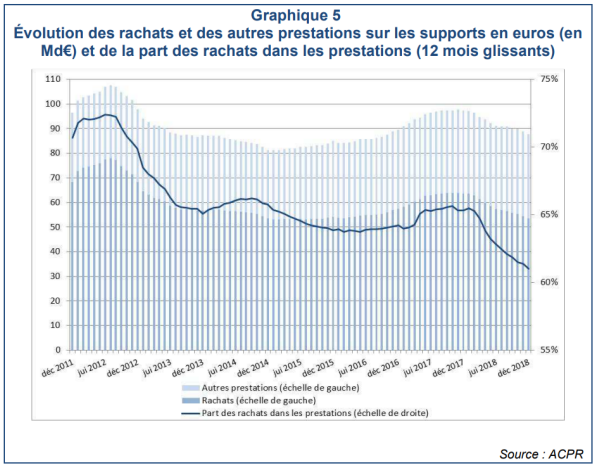


FIGURE 1.9 – Évolution des rachats et des autres prestations sur les supports en euros et de la part des rachats dans les prestations

Selon ACPR, les rachats fait une partie majeure dans les prestations payées par les assureurs sur les support en euro. Depuis 2011, cette part est en décroissant mais elle reste encore à 61% en 2018.

Pour GPM, le risque de rachat a un impact important sur le résultat du produit en euro. En effet, l'année 2020 marque une réduction du portefeuille due au rachat à hauteur de 42 809 k€ sur Alticore Epargne, entraînant ainsi une réduction des engagements. Le REPAG étant un produit en run-off, 8% des rentes ont été liquidées suite au décès du bénéficiaire générant une reprise de provisions mathématiques de 22 346 k€ (Source : Rapport sur la solvabilité et la condition financière, GPM 2021)

Dans ce contexte, une bonne connaissance du comportement des adhérents est donc cruciale afin de permettre une meilleure anticipation des taux de rachat et ainsi de assurer une bonne gestion d'actif-passif.

Résultat de souscription (en k€)	31/12/2021	31/12/2020	Variation en %
Primes brutes	9 440	13 041	-28 %
Prestations brutes	77 594	68 472	13 %
Variation de provisions	-63 015	-41 310	-53 %
Résultat technique brute	**-5 140**	**-14 121**	**64 %**
Primes cédées	0	0	0 %
Prestations cédées	0	0	0 %
Variation de provisions cédées	0	0	0 %
Résultat technique cédé	**0**	**0**	**0 %**
Primes nettes	9 440	13 041	-28 %
Prestations nettes	77 594	68 472	13 %
Variation provisions nettes	-63 015	-41 310	-53 %
Résultat technique net	**-5 140**	**-14 121**	**64 %**
Frais administratifs	1 329	1 417	-6 %
Frais de gestion des investissements	1 813	1 198	51 %
Frais de gestion des sinistres	643	714	-10 %
Frais d'acquisition	956	1 040	-8 %
Frais généraux	0	0	0 %
Total frais	**4 742**	**4 369**	**9 %**
Résultat technique net y compris frais	**-9 882**	**-18 490**	**47 %**

FIGURE 1.10 – Résultat technique épargne fonds euros

L'article 79 de Directive Solvabilité II souligne que les hypothèse doivent être « réaliste et fondée sur des informations actuelles crédibles ». Ces hypothèses, y compris les risques de rachat, doivent prendre en compte toutes les données internes et les données externes afin de refléter au mieux les propriétés du portefeuille. Le calcul de la probabilité de rachat des adhérents affecte le stress test du calcul de SCR, l'étude d'ALM et la comptabilité de l'entreprise selon Milhaud et al. (2010). En d'autre terme, sous la carde de Solvabilité 2, la mission de comprendre le risque de rachat devient cruciale pour les assureurs.

Dans l'assurance-vie, le risque de rachat est mentionné dans la sous-module de « *Lapse* » ou « cession » de la formule standard. Ce sous-module comprend différents scénarios de stress qui représentent des chocs sur le taux de rachat, tels que décrits à l'article 95 du règlement autorisé (UE) 2015/35 et QIS5 partie 7.

En effet, le risque de rachat représente le risque de perte ou de variation du passif du fait des variations du taux d'exercice des options du contrat d'assurance. Cela comprend

les options de résiliation, de réduction, de limitation ou de suspension totale ou partielle de la couverture d'assurance ainsi que les options permettant la mise en place totale ou partielle, le renouvellement, l'augmentation, l'extension ou la reprise de la couverture d'assurance.

Dans la partie suivante, nous étudierons plus en détail l'agrégation du sous-module « vie » et la spécification de l'exigence de fonds propres pour le risque de rachat. Nous détaillerons également les deux catégories de risque de rachat et un rapide résumé de la façon dont les études sont réalisées.

1.4.1 Sous-module risque de souscription vie

Le sous-module risque de souscription d'assurance vie couvre l'ensemble des risques liés aux risques couverts et aux processus suivis dans la conduite des affaires. Il se décompose encore en 7 sous-modules : *risque de mortalité*, *risque de longévité*, *risque d'invalidité/incapacité*, *risque de rachat*, *risque des frais*, *risque de révision* et *risque de catastrophe*.

En général, le SCR se calcule par :

$$\begin{aligned} SCR = \Delta NAV &= NAV_{avantchoc} - NAV_{choc} \\ &= (A_{\text{avant choc}} - BE_{\text{avant choc}}) - (A_{choc} - BE_{choc}) \end{aligned}$$

où

NAV(Net asset value) : valeur marché de fond propres de l'assurance

A : valeur marché de l'actif

BE : Best Estimate

Dans le cas du risque de souscription vie, un choc n'affectera pas les actifs de l'entreprise. Autrement dit, $A_{\text{avant choc}} = A_{choc}$. Donc, $SCR = BE_{choc} - BE_{\text{avant choc}}$

Le SCR_{life} se définit comme :

$$SCR = \sqrt{\sum_{r,c} CorrLife_{r,c} \times Life_r \times Life_c}$$

$Life_{r,c}$ sont respectivement le SCR des 7 facteurs des risques. $CorrLife_{r,c}$ correspond à la corrélation entre deux facteurs de risque. Voici la matrice corrélation des risques :

Il faut noter que le risque de rachat est corrélé aux risques de longévité, de frais et de catastrophe. Le capital exigé pour ce risque se définit comme :

	Mortality	Longevity	Disability	Lapse	Expenses	Revision	CAT
Mortality	1						
Longevity	-0.25	1					
Disability	0.25	0	1				
Lapse	0	0.25	0	1			
Expenses	0.25	0.25	0.5	0.5	1		
Revision	0	0.25	0	0	0.5	1	
CAT	0.25	0	0.25	0.25	0.25	0	1

FIGURE 1.11 – Matrix de corrélation de sous-module vie

$$Life_{lapse} = max(Lapse_{down}; Lapse_{up}; Lapse_{mass})$$

Où $Lapse_{down}$, $Lapse_{up}$ et $Lapse_{mass}$ correspondent à la ΔNAV dans le cas du taux de rachat structurel choqué et dans le cas de scénarios de rachat de masse (ou dit rachat conjoncturel). Donnons maintenant la définition d'un rachat structurel et d'un rachat conjoncturel.

Le taux de rachat structurel représente le comportement de rachat basé sur les caractéristiques des assurés : leurs âges, l'ancienneté de leur contrat, le sexe, le type de leur contrat, etc. Dans un environnement d'économique « normal », selon d'anciennes recherches, la décision de rachat est traitée comme le risque de décès, c'est-à-dire qu'elle est considérée comme une cause « exogène » de résiliation du contrat. Une telle décision peut être motivée par plusieurs raisons « personnelles » hors du contrôle de la compagnie d'assurance. Les statistiques suffisantes sur les rachats permettent d'estimer les taux de rachat attendus (Bacinello (2005))

Le taux de rachat conjoncturel, au contraire, est le résultat d'un événement majeur sur la situation économique. En effet, l'effet de rachat conjoncturel peut être plus grave que le rachat total car il implique une vague potentielle de rachats. Il est également plus difficile à modéliser. On peut citer ici quelques variables qui pourraient avoir un impact sur le taux de rachat conjoncturel :

- Environnement économique et financier : état des marchés financiers, inflation, croissance, chômage
- Environnement concurrentiel : taux concurrentiel, lancement de nouveaux produits, etc.
- Évolution de la législation, fiscale entre autre

1.4.2 Le choix de modélisation de la loi de rachat structurel

Chez GPM, le risque de rachat est le quatrième risque le plus élevé impactant le SCR. Le SCR et le MCR sont calculés en utilisant la formule standard. Les deux années 2019 et 2020 montrent le montant de rachat stable.

En €	2020	2019	Variation
SCR souscription vie	26 785 895	15 918 013	+68%
Longévité	21 802 459	12 562 328	+74%
Mortalité	157 964	159 972	-1%
Rachat	365 733	372 986	-2%
Frais	4 325 238	2 735 351	+58%
Catastrophe	134 501	87 376	+54%

FIGURE 1.12 – SCR souscription vie

Bien que la modélisation du risque de rachat structurel et conjoncturel soit aussi importante pour les assureurs dans le contexte de Solvabilité 2, dans le cadre de mon stage au GPM, la modélisation de la loi de rachat structurel est mise en priorité comme ma mission principale. Elle est bâtie en prenant en compte la faisabilité en raison de la contrainte de durée du stage, la disponibilité des données et le besoin urgent de l'entreprise. Par conséquent, dans cette étude, nous nous concentrons sur le rachat structurel dans le cas de l'assurance-vie épargne/retraite en appliquant les approches statique et dynamique avec des modèles de Machine Learning.

Ici, on peut noter que la prise en compte des facteurs et des modèles conjoncturel dans le futur ne devrait pas impacter la pertinence notre modèles. L'idée de tenir compte à la fois du rachat structurel et du rachat conjoncturel a été avancée dans la litterature, mais n'a pas encore été résolue de façon systématique. Il reste encore un sujet intéressant. Une des solutions qui a été poursuivie est de construire un modèle qui n'a pas de contrainte dans les types de variables comme dans l'étude de JAMAL (2017). De cette manière, on peut inclure la variable économique qui est une composante temporelle comme dans l'etude de Loisel et al. (2021).

Tout au long de cette étude, ***on fera référence au rachat structurel***. Nous étudierons différents aspects de l'évènement de rachat correspondant aux différents types de variables cibles. Nous examinerons la survenance et la non-survenance de rachat grâce à la modélisation d'une variable binaire au chapitre 3. Ici, des modèles de *Machine Learning* sont appliqués en tant que méthode principale.

Ensuite, nous regardons les rachats au sens du temps de l'événement dans chapitre 4. Nous tentons de modéliser la durée du maintien en portefeuille de l'adhérent jusqu'à ce qu'il décide de racheter. Cette approche est pertinente pour l'avis de l'ACPR. En effet, selon ACPR, les rachats structurels, estimés à partir de l'historique des taux de rachat observés, seraient modélisés en fonction de l'ancienneté du contrat mais impliquent parfois

aussi l'âge à la souscription. Dans le présent ce chapitre, on s'interroge également sur l'homogénéité du portefeuille. Le modèle du machine learning apportera une solution simple aux questions.

Et au final, nous proposons une nouvelle approche qui intègre le Machine Learning dans le modèle de survie.

1.4.3 Définition de taux de rachat

Dans un premier temps, nous nous intéressons au montant que la compagnie d'assurance doit payer sans une distinction entre rachat partiel et rachat total. Nous définirons respectivement le taux de rachat personnel par montant et par nombre pour un horizon d'un an comme :

$$TR_{personnel;montant} = \frac{\text{Montant de rachat (soit partiel soit total)}}{\text{Provision Mathématique d'ouverture}}$$

$$TR_{personnel;nombre} = \frac{\text{Nb de personne rachat}}{\text{Nb de personne dans portefeuille}}$$

Ces indicateurs nous donnent un aperçu du montant total et de la fréquence des obligations que l'assureur doit respecter chaque année. Cependant, il est également très important de distinguer entre le rachat total et partiel car la nature du risque est différente due au comportement de l'adhérent.

De plus, l'input nécessaire pour le modèle d'inventaire de GPM est la table de rachat par ancienneté et par type de rachat. Donc, de la même manière, nous construisons le taux de rachat partiel et total en montant et en nombre. Nous avons deux variables qui doivent être définies : le montant de rachat et la PM. Le calcul du « Montant de rachat » se fait en fin d'année. Il s'agit des rachats d'une année calendaire. La PM est calculée en début d'année.

Chapitre 2

Études préliminaires du portefeuille

Pour aborder le sujet principal de ce mémoire, il est d'abord nécessaire de présenter le cadre dans lequel s'effectue notre travail. Dans cette section, nous présenterons brièvement les activités de GPM pour comprendre la nature particulière des profils clients chez GPM ainsi que les deux produits de fonds en euros concerné. Ensuite, nous analyserons les données d'un point de vue statistique et préparerons les données pour l'étape de construction du modèle dans la partie III de l'étude.

2.1 Description du portefeuille en euro de GPM

Le Groupe Pasteur Mutualité (GPM) est un groupe mutualiste d'assurances destiné aux professionnels de la santé. Avec près de 130 000 adhérents et un chiffre d'affaires de 217 millions d'euros en 2016, GPM est l'une des premières mutuelles dans le secteur de la santé en France.

Le Groupe Pasteur Mutualité est issu de l'Association Générale des Médecins de France (AGMF), fondée en 1858 pour assurer la santé et le bien-être des médecins. En effet, GPM a été créé par les médecins et pour les médecins. Aujourd'hui, le groupe a élargi son portefeuille de produits pour inclure des contrats d'épargne, de retraite et d'assurance contre l'incendie, les accidents et autres risques (P&C), ainsi que le profil de ses adhérents à l'ensemble des professionnels de santé.

Le Groupe Pasteur Mutualité (GPM) est le premier groupe mutualiste français de premier plan géré par des professionnels de la santé, proposant des solutions conçues par des professionnels de la santé. Leur participation et leur rôle actif garantissent des produits qui répondent parfaitement aux attentes de tous les adhérents. Dans les adhérents de

GPM on peut compter :

- Médecins libéraux
- Les praticiens hospitaliers
- Pharmaciens
- Chirurgiens-Dentistes, Infirmiers
- Sages-femmes
- Vétérinaires
- Masseurs-kinésithérapeutes
- Professions libérales
- Autres professions de santé

GPM Assurances (GPMA) est une des trois entités de GPM, créée en 1997 et régie par le Code des Assurances, propose des contrats d'assurance vie. GPMA gère les contrats de type épargne et retraite.

GPMA commercialise des contrats d'épargne en euros (Comptes, Bons, PEP, Altiscore Multisupports et Retraites, Plans d'Epargne PEP). AGMF Epargne est un produit d'assurance vie mono-support en euros. Il est fermé à la commercialisation au 01/01/1994. Cependant, GPMA gère toujours ce produit pour ses prestations.

2.2 Analyse descriptive des données

2.2.1 Analyse univariée

Les données de GPMA comprennent des comptes épargne proviennent de deux produits : AGMF Epargne et Altiscore dans un horizon d'observation de 2013 à 2019. Le choix de la période de référence est crucial dans la mesure où il doit nous permettre d'obtenir suffisamment de données pour atténuer l'effet de la censure et de la troncature sur notre estimation.

A noter que nous avons accès aux données de 2020. Cependant, 2020 est une année de volatile due à l'émergence de la pandémie de Covid-19. Cela affecte notre construction de modèles. C'est pourquoi nous retirons 2020 de l'étude et que notre fin d'observation est 2019. Comme nous nous concentrons davantage sur le rachat structurel et le rachat dynamique réglé par Solvabilité II, la recherche d'une sur-augmentation du taux de rachat n'entrera pas dans le cadre de cette étude.

Nous travaillons avec deux bases de données : l'information d'adhérent et l'information des transactions. En effet, les informations des adhérents sont utilisées pour analyser le comportement des clients et ensuite le montant des rachats au niveau des adhérents. Dans cette étude, c'est principalement dans le but de comprendre les caractéristiques des adhérents et la tendance du rachat. L'ensemble des données d'informations de chaque transaction est utilisé pour modéliser le taux de rachat en utilisant un modèle de durée, dont nous parlerons dans les prochaines parties.

Du point de vue de GPM, nous nous intéressons à l'évolution du taux de rachat en général tant pour AGMF Epargne que pour Altiscore. L'analyse bivariée et univariée se fera donc, avec chaqu'un des deux produits pour analyser les comportements. À partir de maintenant, nous allons analyser la base de données agrégée des deux produits distinguable par la variable PRODUIT.

Nom de variable	Description	Type
NOADH	ID d'adhérent	
TAUX	Taux technique applique dans l'année d'observation	Continue $(0;1)$
PM	Provision mathématique (valeur d'épargne) acquise aux 31/12/N-1	Continue
PB	Participation bénéfice acquis aux 31/12/N-1	Continue
ANCIEN	Ancienneté de contrat à 2019 ou la date de rachat	Continue
AGE	Age de contrat à 2019 ou la date de rachat	Continue
PRODUIT	« agmf » ou « altiscore »	
RACH_COUNT	Rachat ou pas	Catégorielle
MT_RACHAT	Montant de rachat	Continue
RACHTOT	Montant de rachat total	Continue
RACHPART	Montant de rachat partiel	Continue
TAUXRACH	Taux rachat personnel	Continue $(0;1)$
MOTI	Motif de sortie le portefeuille	Catégorielle

Table 2.1 – Descriptions des variables

Malgré la complexité, les données sont de bonne qualité. Les données sont également assez récentes et contiennent des informations riches. Il existe plusieurs points de contrôle des données que nous devons vérifier avant de les utiliser pour garantir la qualité des données : incohérence, incomplétude, exactitude, précision et manquant. Les données de GPM ne contiennent aucune valeur manquante. Grâce à certaines vérifications croisées entre les fichiers de données, les données sont exactes.

Cependant, il contient encore quelques incohérences qui découle de la nature du produit. Il seront expliquées dans la partie suivant de traitement des valeurs aberrantes. De plus, avec le volume de données, la mise en forme des données est compliquée à réaliser par Excel. Pour expliquer plus loin, chaque adhérent a de nombreuses transactions différentes à partir de son compte. Cela rend difficile le suivi du compte à chaque instant de calcul.

Notre base de données des adhérents comprend 11 138 observations et 9 variables qui conduit 348 112 différents transactions pendant 7 ans. Chaque ligne correspond aux informations d'une personne à la date d'observation du (31/12/2019). Le tableau 2.1 permet de garder une trace de la description des noms de variables utilisées dans le modèle, du contenu et de la format des données.

Notons que l'année d'observation est définie par l'année de de la sortie du portefeuille. Donc, il s'agit soit de l'année de rachat total, soit de l'année de décès, soit de la dernière année de survenance du portefeuille (2019). Le portefeuille d'étude est composé d'une majorité de produit d'Altiscore (60,97%) avec l'âge moyen de 64,5 ans. La plupart du portefeuille n'est pas racheté (68,46%) et 8,04% du portefeuille est sorti pour la raison hors rachat (décès).

Dans la partie suivante, nous étudions plus en détail les critères statistiques de chaque variable. Cette étape nous donne une idée plus précise des variables du modèle. Nous faisons aussi un contrôle de la qualité des données car nous détectons des anomalies.

	PM	PB	MT_RACHAT	TAUX	AGE	ANCIEN	RACHTOT	RACHPART	TAUXRACH
Min	0	0	0	0	2.85	0.67	0	0	0
Median	11 632	253.34	0	0	70.87	19.50	0	0	0
Mean	37 051	845.64	5263	0.13	69.54	19.25	4657	630	0.26
Max	2 692 096	57 497.07	925 889	2.00	108.24	35.30	925 889	284 100	46.39

Table 2.2 – Descriptive analyse

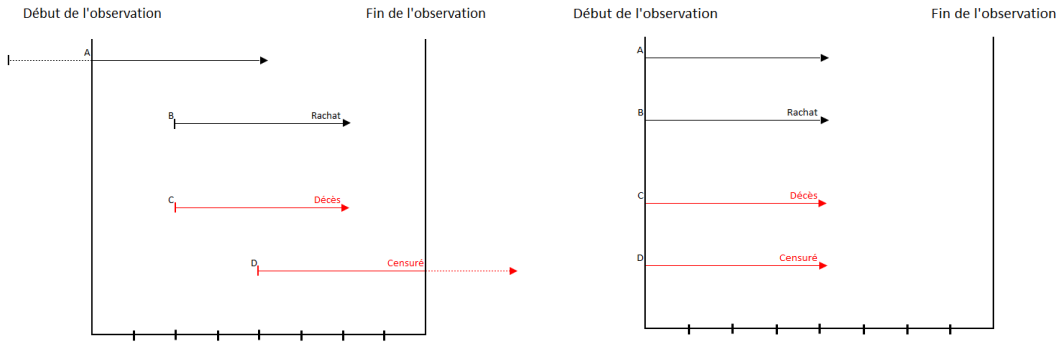
Notez que nos données sont censurées à droite et tronquées à gauche. La censure signifie que les événements peuvent être détectés, mais les valeurs (mesures) ne sont pas complètement connues. Supposons que l'événement de durée de vie qui dans notre cas ont l'ancienneté de contrat représenté par X, alors la valeur exacte de X est inobservable mais on peut mesurer $T = min(X, C)$. Cela se produit lorsque notre horizon d'observation se termine en 2019, il y a des contrats qui ne se sont pas rachetés à cette date.

La troncature signifie qu'un objet ne peut être détecté que si sa valeur est supérieure à un certain nombre. Une caractéristique de la métrique de survie est que nous ne pouvons l'observer qu'à partir du temps 0. En d'autres termes, X n'est observable que sur un sous-ensemble de $[0, \infty)$.

Dans notre période d'observation, nous n'observons souvent qu'une partie de la vie des observations. L'adhérent A n'est observé qu'après que son contrat ait été en portefeuille pendant quelques années. La date de l'événement de l'adhérent B est respectée. Ce sont deux cas où nos données ne sont pas censurées et parmi eux, A est troncature.

Les cas C et D sont ce que nous appelons des données censurés (censures). L'adhérent C ne rachète pas son contrat. Il sort du portefeuille par un autre événement qui est le décès. L'adhérent D n'a pas racheté son contrat dans le délai d'observation.

Le début des contrats de B, C, D commencent après un certain temps dans la période d'observation. Dans l'analyse de survie, nous avons pu standardiser la date de début de contrat comme sur la figure 2.1.b avec la date de censure est désormais variable.



(a) Vue dans calendrier (b) Vue dans analyse de survie

FIGURE 2.1 – Période d'observation

Caractéristique de l'âge et ancienneté du portefeuille

Les questions que l'on se pose sont celles-ci : « quelle est la caractéristique de la variable âge de ce portefeuille ? » et « allons-nous observer des différences d'âge par rapport à la décision de rachat ? ». La prochaine analyse répondra à ces deux questions.

Ici, l'âge maximum de ce portefeuille est de 108,24 ans. C'est un point intéressant car on voit que le portefeuille de GPM est particulier qui effectivement, a un impact sur la modélisation et montre qu'un tableau d'expérience est nécessaire. Ils ont des espérances de vie supérieures à la population générale. Le graphique 2.2 montre la répartition de l'âge des adhérents.

La médiane est de 69 ans, supérieure à l'âge de la retraite de 62 ans en France. En détaillant, 50% du portefeuille a plus de 69 ans, ce qui signifie que la majeure partie du portefeuille d'investissement vont à la retraite. Par conséquent, nous nous attendons plus au comportement de rachat que de versement.

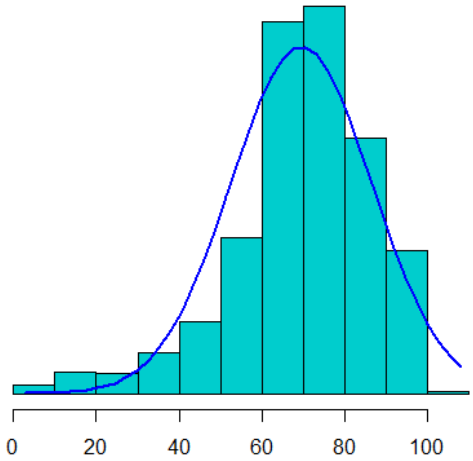


FIGURE 2.2 – Histogramme des fréquences d'âge des adhérents

Ensuite, nous voulons savoir s'il y a une différence d'âge moyen dans la décision de rachat. À l'aide de la figure de boxplot ci-dessous, nous avons comparé l'âge moyen des deux groupes : rachetés et non-rachetés. Ici, 1 représente le groupe d'adhérents qui ont rachetés et 0 représente le contraire.

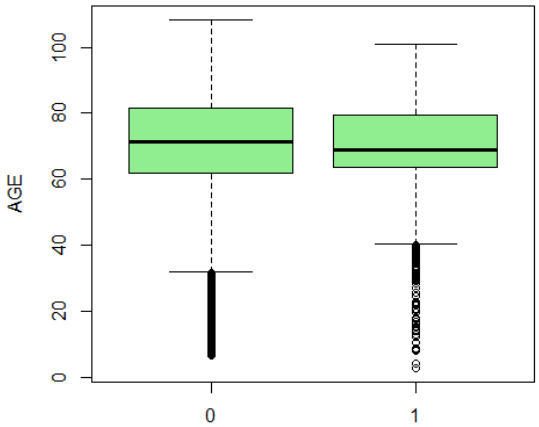


FIGURE 2.3 – Box-plots de deux groupes Non-rachat et Rachat

Nous remarquons qu'il n'y a pas de différence d'âge moyen de l'assuré. En effet, l'âge moyen du groupe de personnes ayant racheté est légèrement inférieur à celui de l'autre groupe. Cependant, cette différence est mineure. Les boîtes sont de même hauteur et relativement courtes, concentrées entre 60 et 80 ans.

Les observations de moins de 40 ans sont considérées comme des valeurs aberrantes. La moustache du groupe non-rachat est plus longue. Cela signifie que l'âge varie davantage au sein de ce groupe. En revanche, l'âge est plus concentré dans le groupe de personnes

qui ont racheté.

Ensuite, nous examinons la distribution de la variable ancienneté dans le portefeuille. Cela nous aiderait à voir comment cette variable se comporte car c'est l'une des variables les plus utilisées dans la modélisation du taux de rachat structurel.

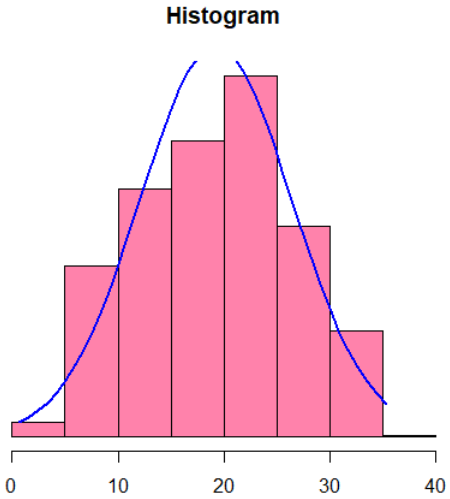


FIGURE 2.4 – Histogramme des fréquences d'ancienneté des adhérents

Dans le graphique 2.4, notre objectif est d'étudier la durée du contrat. On trouve que nombre de contrat dans la période de 0 à 5 ans d'ancienneté est plus bas et ce nombre augmente rapidement dans les années suivantes, notamment après la dixième année. Les adhérents dans ce portefeuille ont une tendance à conserver leur contrat très longtemps. Il est vu comme un signe du fait que ce portefeuille n'est pas sensible au changement du marché et du taux. Les adhérents conservent les contrats plutôt pour le but de protéger que d'investir. En d'autres termes, pour ce portefeuille particulier du GPM, l'élément de rachat structurel est plus important.

Segmentation du portefeuille par type de produit

Les contrats AGMF occupe 38,2% des observations. Nous proposons une hypothèse selon laquelle nous n'avons pas de différence de Montant de rachat (MT_RACHAT) entre ces deux groupes de produits. Ici, nous voulons voir si la variable PRODUIT qui divise le portefeuille est statistiquement significative. Pour tester cette hypothèse, nous utilisons le test ANOVA.

L'ANOVA permet de voir si une variable numérique a des valeurs différentes en fonction de plusieurs groupes. C'est une généralisation du test de Student permettant de comparer plus de deux groupes. Dans le cas d'un test ANOVA, même si le test est significatif, on ne sait pas en revanche quelles catégories sont concernées. Il faudra donc faire un test de

Student pour comparer les groupes 2 par 2.

Dans ce cas, nous n'avons que 2 groupes donc cela ne posera pas de problème. L'une des hypothèses de l'ANOVA est que la variable cible doit avoir une distribution normale. Comme nous l'avons vu précédemment dans le tableau 2.2, la médiane du montant de rachat est égale à 0 qui est supérieur à la moyenne (égale à 5263). Le montant de rachat est donc, jusqu'ici asymétrique. Pour cette raison, le test ANOVA ne fonctionne pas dans cette situation.

Nous proposons un test alternatif d'ANOVA : le test de Kruskal-Wallis. Le test de Kruskal-Wallis par rang est une alternative non paramétrique au test ANOVA, qui étend le test de Wilcoxon à deux échantillons dans la situation où il y a plus de deux groupes.

Kruskal-Wallis rank sum test	
data : MT_RACHAT by PRODUIT	
Chi-squared	212.11
df	1
P-value	<0.0001

Table 2.3 – Résultat du test de Kruskal-Wallis

Une p-value petite signifie qu'il y a une différence entre la moyenne du Montant de rachat des deux groupes (altiscore et agmf). Autrement dit, on doit distinguer les deux produits. Ceci est important pour notre conclusion sur l'utilisation du taux brut général des deux produits. À partir du tableau 2.3 ci-dessous, nous concluons qu'il existe une différence entre les montants de rachat moyens de chaque groupe. En d'autres termes, les types de comptes d'épargne que les adhérents possèdent ont un impact sur la décision de rachat. Cela vient du fait que le taux technique et les frais de sortie de ces produits sont différents.

Caractéristique du comportement de sortie du contrat des adhérents

La figure 2.5 montre les différents types de sortie du portefeuille. 1, 2 et 3 représentent respectivement la sortie de rachat partiel, de rachat total et de décès. Parmi eux, seuls 18,73% ont quitté le portefeuille pour cause de rachat, 7,9% ont quitté le portefeuille pour cause de décès, les 73,35% restants sont toujours dans le portefeuille avec un contrat actif. Notons qu'une personne qui rachète partiellement sont contrat ne sera pas considérée comme sortie le portefeuille.

On voit bien ici le problème de la censure. On observe le portefeuille depuis 7 ans, pourtant la durée de vie d'un contrat d'assurance-vie est beaucoup plus longue. C'est pourquoi, à la fin de notre observation, il y a beaucoup de contrats qui sont toujours dans le portefeuille.

Le nombre de rachats est très faible par rapport à l'observation totale ce qui conduit à des

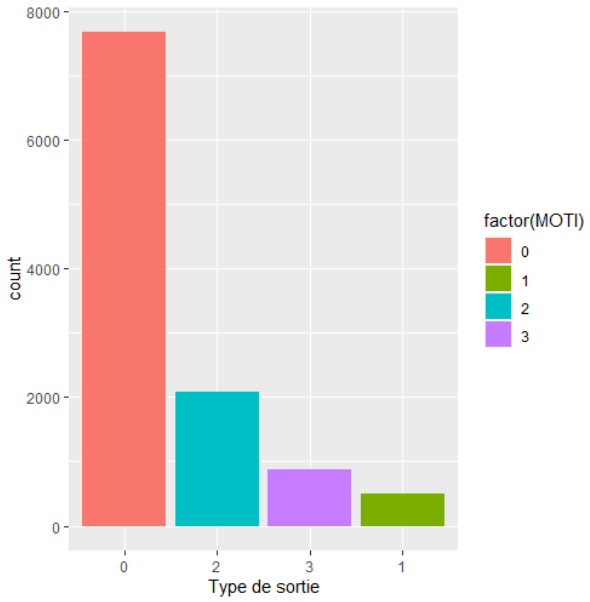


FIGURE 2.5 – Types de sortie

données déséquilibrées (difficulté de déterminé distribution et modélisation). Donc, il est très important de faire une étude de validation de la segmentation des bases de données *training* et *testing*.

Analyse de taux de rachat personnel

La variable TAUXRACH ne sera pas utilisée dans le modèle mais dans cette étude nous effectuons toujours une analyse dans le but de détecter les anomalies et les valeurs aberrantes. Nous rappelons que nous avons défini la variable TAUXRACH comme suit :

$$TR_{personnel;montant} = \frac{\text{Montant de rachat (soit partiel soit total)}}{\text{Provision Mathematique ouverture}}$$

$$TR_{personnel;nombre} = \frac{\text{Nb de personne rachat}}{\text{Nb de personne dans portefeuille}}$$

TAUXRACH est compris entre 0 et 1. Cependant, en réalité, ce n'est pas toujours le cas. Les raisons de l'augmentation ou de la diminution des comptes d'épargne au cours de l'année comprennent : les dépenses, la participation aux bénéfices, les nouveaux versements, etc. Pour cette raison, le montant qu'une personne peut réellement racheter peut parfois être beaucoup plus élevé que la PM du début d'année. Pour cette raison, nous choisissons un seuil admissible pour TAUXRACH i.e nous avons décidé que TAUXRACH $> Var_{99,5}\%$ est une valeur aberrante.

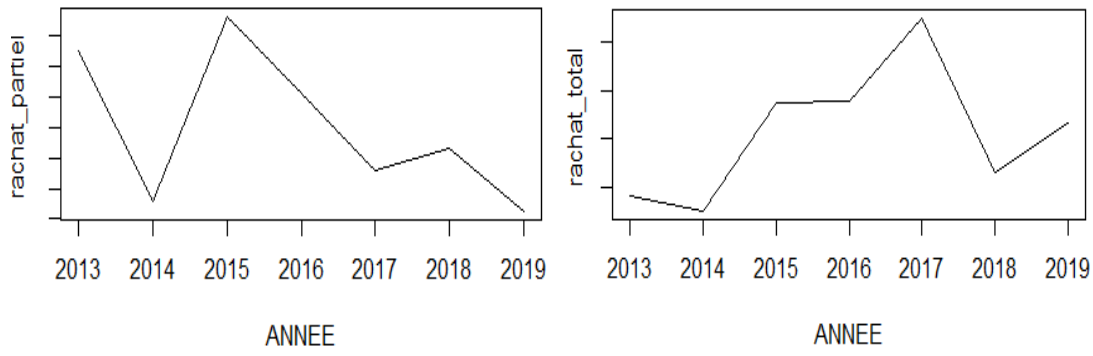
2.2.2 Analyse bivariée

Dans une régression, la multi-colinéarité est un problème qui survient lorsque certaines variables de prévision du modèle mesurent le même phénomène. Une multi-colinéarité prononcée s'avère problématique, car elle peut augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter. Par conséquent, pour les modèles de régression (ols, ridge, lasso, glm), il est très important de se renseigner sur la multi-colinéarité. En revanche, il est moins important pour les modèles basés sur des arbres (arbres de decision, random forest).

Dans cette section, nous étudions la corrélation inter-variables : en savoir plus sur les interactions entre les variables pour clarifier le comportement des clients, ajouter des indicatives pour le modèle de régression. De plus, on répond la question : quelle variable peut être utilisée pour modéliser le rachat ?

Ici, en utilisant une méthode assez naïve pour estimer le taux brut, nous construisons un taux forfaitaire par année. Il présente une tendance et des premières idées des taux de rachat. En moyenne, le taux de rachat partiel et total sont respectivement de 2,8% et 3,1% par an. Ses écart-type sont de 0,25% et 1,38%. On constate une baisse du taux de rachat en 2014.

Il est clair que taux de rachat total et partiel sont peu corrélés car une personne qui rachète totalement son contrat ne peut pas faire partiellement. Tandis que le taux de rachat total dépend davantage de la situation économique, le taux de rachat partiel dépend davantage de la situation personnelle.



(a) Taux de rachat partiel (b) Taux de rachat total

FIGURE 2.6 – Taux de rachat en montant par année

Le graphique ci-dessus montre qu'il existe des différences de taux de rachat total en nombre en fonction de l'ancienneté. Concrètement, à partir de dix ans d'ancienneté, cette proportion tend à augmenter rapidement. Cela peut s'expliquer par l'impact des avantages fiscaux. Le taux de rachat commence à diminuer après 25 ans à cause de la taux de

mortalité.

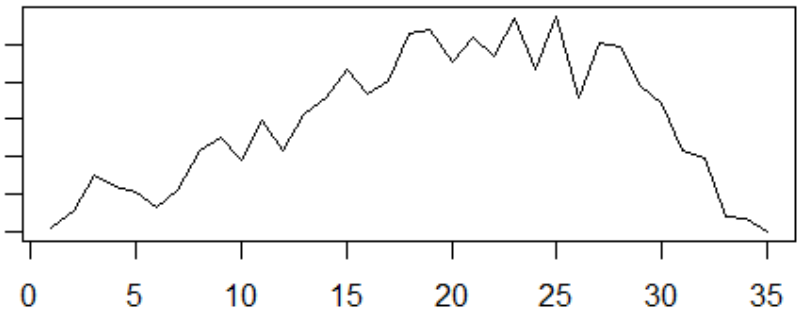


FIGURE 2.7 – Taux de rachat en nombre en fonction de l'ancienneté

Ensuite, nous nous intéressons à l'interaction entre les variables. Pour répondre à cette question, nous observons la matrice de corrélation.

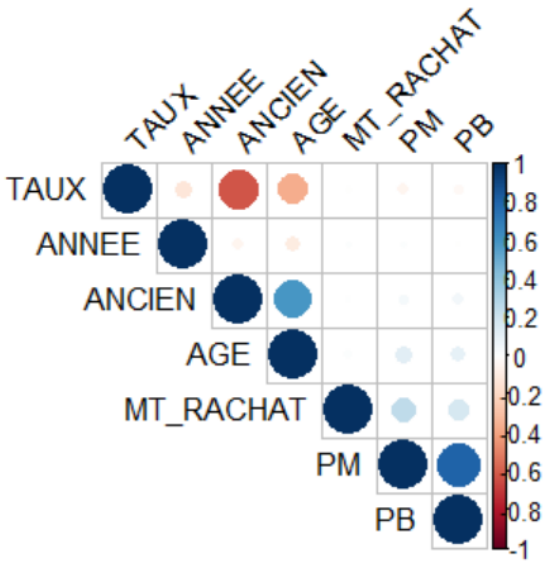


FIGURE 2.8 – Matrice de corrélation

A partir de la matrice de corrélation, on remarque une corrélation entre la variable PB et PM à cause de la nature de création ces variables. C'est aussi le cas de l'ancienneté et du taux technique parce que même si les adhérents vont différents taux techniques au début du contrat, ce taux a tendance de diminuer dans les années futurs et finalement atteindre 0.

Nous rappelons d'une erreur fréquente est de confondre la multi-colinéarité et la corrélation. Deux variables sont dites colinéaires si elles mesurent la « même chose » tandis que

corrélation représente la relation linéaire des variables. Nous utilisons la corrélation comme un signe pour décider s'il faut faire une analyse plus approfondie de la multi colinéarité. Il existe différentes mesures de la multi-colinéarité, notamment l'approche classique : de la VIF (variance inflation factor (VIF)).

2.3 Détection et traitement des valeurs aberrante et anomalie

Comme indiqué dans la partie précédente, nous choisissons un seuil pour TAUXRACH car nous avons décidé que TAUXRACH $Var_{99,5}\%$ est une valeur aberrante. Plus précisément avec cette étape, 56 observations sont supprimées.

Il existe plusieurs approches pour déterminer l'anomalie. En général, ils suivent la même idée : définir les observations « normales » et filtrer tout le reste. Ils sont construits pour déterminer la normalité donc, ça cause le résultat de la détection d'anomalies contient trop de faux positifs ou peut détecter trop peu d'anomalies. Dans ce mémoire, nous présentons une méthode de détection d'anomalies (voir Liu et al. (2008)) qui s'appelle « Isolation Forest » qui peut effectivement résoudre ce problématique.

Cette méthode est basée sur le principe de l'arbre de décision. De sorte que dans cette partie nous allons donner une idée générale de son algorithme sans entrer dans le détail de la manière de construire un arbre de décision. Le terme concerné par l'arbre sera expliqué plus en détail dans la partie III.

Le cœur de l'algorithme est « d'isoler » les anomalies en créant des arbres de décision sur des attributs aléatoires. Le partitionnement aléatoire produit des chemins significativement plus courts pour les anomalies car :

- Moins d'instances (d'anomalies) entraînent des partitions plus petites
- Les valeurs d'attribut distinguables sont plus susceptibles d'être séparées

Par exemple, une observation de rachat avec une valeur PM inférieure à 10 euros peut être considéré comme aberrant. Il est plus facile de détecter de telles observations, car nous voyons rarement des observations a une valeur de PM inférieure à 10 euros. En revanche, lorsqu'une observation est considérée comme « normale », il existe de nombreux points similaires autour d'elle, ce qui la rend plus difficile à « isoler ». Par conséquent, lorsqu'une forêt d'arbres aléatoires produit collectivement des longueurs de chemin plus courtes pour certains points particuliers, il est fort probable qu'il s'agisse d'anomalies.

Définissons $X = x_1, ..., x_n$ comme l'ensemble d'observation, l'algorithme pour construire une « *Isolation Tree* » se définit en sélectionnant au hasard un attribut q avec une valeur de fractionnée p jusqu'à ce que soit :

- le nœud n'a qu'une seule instance, ou
- toutes les données du nœud ont les mêmes valeurs

Donc, chaque nœud dans l'arbre est soit un nœud externe sans enfant, soit un nœud interne avec un « test » et exactement deux nœuds filles. Lorsque l'iTree est complètement développé, chaque point de X est isolé sur l'un des nœuds externes. Intuitivement, les points anormaux sont ceux avec la plus petite longueur de chemin dans l'arbre, où la longueur de chemin $h(x_i)$ au point x_i se défini par le nombre d'étapes depuis la racine pour atteindre le nœud externe.

Après avoir détecté les valeurs aberrantes à l'aide de l'algorithme Isolation Forest, nous supprimons toutes les valeurs aberrantes pour évaluer les performances améliorées du système en termes de précision de classification. Avec nos données, il y a 476 observations qui sont considérée comme aberrantes. Par conséquent, nous choisissons de le supprimer.

2.4 Transformées de discrétisation

Certains algorithmes de classification ne traitent que des attributs nominaux et ne peuvent pas gérer les mesures sur une échelle numérique. En outre, les performances de nombreux algorithmes d'apprentissage automatique se dégradent pour les variables qui ont des distributions de probabilité qui ne sont pas standards.

Certaines variables d'entrée peuvent avoir une distribution très asymétrique, telle qu'une distribution exponentielle. Les valeurs aberrantes font que la distribution est très répandue, ce qui est le cas de la variable PM dans notre étude. Ces problèmes peuvent rendre un jeu de données difficiles à modéliser avec une gamme de modèles d'apprentissage automatique.

En tant que tel, il est souvent souhaitable de transformer chaque variable d'entrée pour avoir une distribution de probabilité standard. Une approche consiste à transformer la variable numérique pour avoir une distribution de probabilité discrète où chaque valeur numérique se voit attribuer une étiquette et les étiquettes ont une relation ordonnée (ordinale).

Différentes méthodes pour regrouper les valeurs en k groupes discrets peuvent être utilisées :

- Uniforme : chaque groupe a la même largeur dans la plage de valeurs possibles pour la variable.
- Quantile : chaque groupe a le même nombre de valeurs, réparties en centiles.
- « Clustered » : les clusters sont identifiés et des exemples sont attribués à chaque groupe.

Une transformée de discrétisation K-means tentera d'ajuster k clusters pour chaque va-

riable d'entrée, puis affectera chaque observation à un cluster. À moins que la distribution empirique de la variable ne soit complexe, le nombre de grappes sera probablement petit, par exemple 3 à 5.

Chapitre 3

Étude statiques avec des méthodes de Machine Learning

Les objectifs de cette partie sont triples : compréhension du portefeuille, choix des variables dominant et approche de l'idée d'apprentissage automatique par l'introduction de l'utilisation des différents modèles. Nous aimerions avoir une idée des déclencheurs possibles de la décision rachat en l'expliquant en fonction de variables. C'est ce qu'on appelle « modèle statique » qui se réfère à une détection de risque à une certaine date. Dans notre cas, cette date est fixée au 31/12/2019. Cela nous aide à mettre en évidence les facteurs discriminants dans la décision de rachat

3.1 Méthodologies des modèles

3.1.1 Régression logistique

La régression logistique est un algorithme de classification supervisée. Il s'agit d'un cas particulier du modèle linéaire généralisé (GLM) ou de variables cibles, utilisant la modalité [0,1]. L'objectif principal est de calculer la probabilité de rachat au niveau du contrat. La régression logistique est une alternative à l'analyse discriminante linéaire de la méthode de Fisher de 1936.

Les GLMs sont apparus pour la première fois dans Nelder & Wedderburn (1972). Ils ont été appliqués à de nombreux problèmes et sont couramment utilisés dans le domaine des

statistiques et en actuariat Charpentier & Denuit (2005). Le but du modèle GLM est de trouver la relation entre la variable objectif Y et la variable indépendante X où X est un vecteur de dimensions k. On a $X = (X_1, X_2, ..., X_k)$. Cette relation est exprimée par la fonction de lien h. Donc, un modèle de GLM est défini par :

$$h((\mathbb{E}[Y|X = x] = x\beta)$$

$\beta = (\beta_0, \beta_1, ...\beta k)$ est la vecteur de coefficient de la régression. Dans le cas de la régression logistique, la fonction de lien correspond à la fonction logit qui est defini par :

$$logit(p) = Log(\frac{p}{1-p})$$

Nous disposons d'un échantillon Ω de taille n. La valeur prise par Y pour un individu ω est notée $Y(\omega)$ et $X = (X_1(\omega), X_2(\omega), ..., X_k(\omega))$. Car la variable Y n'accepte que 2 valeurs 0 ou 1, la probabilité que Y prenne la valeur 1 se défini par $\mathbb{P}(Y(\omega) = 1|X(\omega)) = p(\omega)$. Le logit d'un individuel est :

$$ln(\frac{p(\omega)}{1-p(\omega)}) = \beta_(0) + \beta_1 X_1(\omega) + ...$$

La ratio $\frac{p}{1-p} = \frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=0|X)}$ représente le "odd". En autre terme, si un individuel présent un odd de 4, ça signifie qu'il a 4 fois plus de chances d'être égale à 1 que d'être égale à 0.

L'une des propriétés importantes de la fonction logit est la transformation d'une valeur défini sur l'intervalle $[0, 1]$ en une valeur sur l'intervalle $(-\infty, +\infty)$. Le odd-ratio est très important pour les modèles de classification. Nous étudions sa signification à travers l'exemple suivant.

En supposant que la probabilité de rachat est de 0,9 étant donné les conditions sur les variables indépendantes ou en d'autres termes, $\mathbb{P}(Y = 1|X) = 0,9$, alors la probabilité de non rachat est de 0,1 et l'odd-ratio est égal à 9. Cela signifie qu'un adhérent a à profil fixée (âge, année de contrat, PM et PB) a 9 fois plus de chance de racheter son contrat que de rester dans le portefeuille.

- Estimation des paramètres par la maximisation de la vraisemblance

L'objectif est d'estimer le vecteur de coefficients β. Contrairement à la régression linéaire normale, nous ne pouvons pas utiliser l'estimation des moindres carrés ordinaire puisque Y n'est pas normalement distribué. Nous résolvons ce problème en appliquant la méthode du maximum de vraisemblance. Ici, toutes les observations de l'échantillon sont indépendamment supposées être des distributions de Bernoulli pour calculer la fonction de vraisemblance. Donc :

$$\mathbb{P}[Y(\omega)|X(\omega)] = p(\omega)^{y(\omega)} \times (1 - p(\omega))^{1-y(\omega)}$$

La vraisemblance s'écrit :

$$L = \prod_{\omega} p(\omega)^{y(\omega)} \times (1 - p(\omega))^{1-y(\omega)}$$

La log-vraisemblance est donc :

$$L = \sum_{\omega} y(\omega) ln(p(\omega)) + (1 - y(\omega)) ln(1 - p(\omega))$$

Nous cherchons β tel que : $\dfrac{\partial ln(L)}{\partial \hat{\beta}}$. Cette log-vraisemblance est une fonction convexe donc il existe une solution unique. Cependant, il n'existe pas de solution analytique. Nous pouvons résoudre ce problème d'optimisation en utilisant algorithme de Newton-Raphson. Dans cette étude, nous n'expliquons pas le détail de cet algorithme. Dans la section suivante, nous étudierons comment on évalue le modèle.

- Goodness of fit

La question est maintenant de savoir « le modèle est-il considéré comme bon ? » En d'autres termes, nous voulons savoir si l'estimation est proche de la réalité. Pour la régression linéaire (OLS), nous utilisons R^2 où R^2 est une mesure statistique qui détermine la proportion de variance de la variable dépendante qui peut être expliquée par la variable indépendante.

Par exemple, un R^2 de 60% révèle que 60% des données correspondent au modèle de régression. En général, un R^2 plus élevé indique un meilleur ajustement pour le modèle.

Rappelons le modèle de régression linéaire. R^2 est calculé en utilisant les résidus. Les résidus de la régression logistique au contraire, sont tous infinis. Par conséquent, nous ne pouvons pas calculer le R^2. Il existe de nombreuses études utilisant différents mesures comme pseudo R^2.

Mittlböck & Schemper (1996) ont étudié 12 mesures différentes. La méthode couramment utilisée et rapportée dans le logiciel statistique R est celle proposée par McFadden et al. (1973). Nous présenterons ce pseudo R^2. Soit L_0 la valeur de la fonction de vraisemblance pour un modèle sans prédicteur, et soit L_M la vraisemblance du modèle estimé. Le pseudo R^2 de McFadden est défini comme :

$$R^2_{McF} = 1 - \frac{ln(L_M)}{ln(L_0)}$$

Contrairement à la régression linéaire, le pseudo R^2 est celui qui ne montre pas de diminution proportionnelle de l'erreur. La régression linéaire suppose l'homoscédasticité, que la variance d'erreur est la même. Dans la régression logistique, chaque valeur du score de prédiction a une valeur différente avec une diminution proportionnelle de l'erreur. Par conséquent, il est faux de considérer R^2 comme une diminution proportionnelle de l'erreur au sens universel de la régression logistique.

- Tests de significative des coefficients

Pour voir si la variable a réellement un impact sur le modèle, nous allons faire un test significativité du coefficient. Le premier test que nous ferons est basé sur la question : si chaque coefficient est significatif. Notre hypothèse sera donc :

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Ce test est appelé test de Wald. La statistique de Wald est $Z = \dfrac{\hat{\beta}_i}{se(\hat{\beta}_i)}$. Pour déterminer la p-valeur de ce test, nous utilisons une distribution normale. Et donc, nous construisons l'intervalle de confiance du coefficient comme : $\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} se(\hat{\beta}_i)$

- Tests de signicativité globale du modèle

Pour tester si au moins une variable a un impact sur le modèle, nous utilisons l'hypothèse :

$$H_0 : \beta_1 = \beta_2 = ...\beta_J = 0$$

$$H_1 : \text{un des coefficients au moins est non nul}$$

Nous nous intéressons à la statistique $G = -2log(L_0) - (-2log(L))$ où L_0 est la vraisemblance d'un modèle constant (en utilisant la moyenne) et L est la vraisemblance d'un modèle avec l'ensemble des variables. Nous trouvons une p-value en localisant G dans une χ^2-distribution avec k degrés de liberté. Un modèle peut être considéré comme globalement significatif si la probabilité critique (la p-value) est inférieure au niveau de signification défini.

3.1.2 Abre de CART

Le principe d'un Arbre de CART est de construire les sous-groupes les plus « homogènes » du point de vue de la variable à prédire. Dans la méthode de CART, nous rencontrons les termes suivants :

- Racine : représente le premier niveau de l'arbre. Il comprend la distribution de variable à prédire.
- Variable de segmentation : la première variable utilisée pour construire l'arbre

– « Feuille de l'arbre » ou nœud externe : le dernier nœud de l'arbre où aucune branche ne sort

Si Y est une variable catégorielle, l'arbre s'appellera « Arbre de classification ». Au contraire, si Y est numérique, l'arbre s'appellera « Arbre de régression »

Il existe différentes façons de construire un arbre de décision : ID3, C4.5, *Chi-square automatic interaction detection* (CHAID), etc. L'algorithme CART Breiman et al. (1984) est le plus utilisé. En particulier, pour modélisation des taux de rachat, l'algorithme CART ainsi que le modèle d'arbre de décision sont très couramment utilisés, notamment dans les études de Milhaud et al. (2011). Le terme CART signifient *Classification And Regression Tree* (CART) est utilisé pour désigner à la fois l'arbre de classification et de régression. Les arbres utilisés pour la régression et les arbres utilisés pour la classification présentent certaines similitudes, mais aussi certaines différences, telles que l'indice de diversité. Dans ce mémoire, comme Y est le choix de rachat ou non, nous allons détailler de l'arbre de classification.

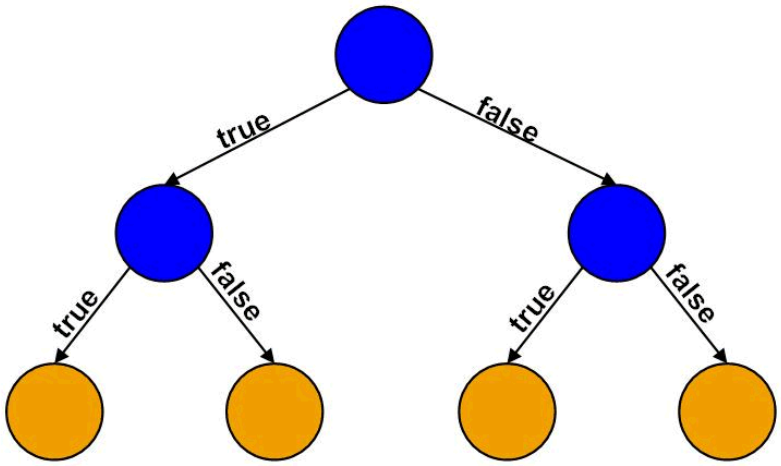


FIGURE 3.1 – Illustration de l'algorithme de l'arbre de décision

Ensuite, nous étudierons en détail l'algorithme pour construire un arbre. Les questions sont maintenant les suivantes :

– Comment décider quelle variable à prendre au niveau du nœud ?

– Comment décider le seuil de segmentation du nœud ?

– Quel est le critère pour décider la taille adéquate de l'arbre ?

– Comment mesurer la précision du fractionnement de l'arbre ?

Pour repondre à ces questions, on présente l'indice de diversité de Gini.

- L'indice de diversité de Gini

Le but de l'arbre est de créer des feuilles pures. En d'autres termes, toutes les observations de ce nœud adoptent un seul mode Y. Le choix de la variable divisée dépend de la capacité à diviser les données en groupes les plus purs. Pour mesurer la pureté, chaque méthode de construction d'arbre utilise un indice différent, et l'arbre CART utilise l'indice de Gini.

Supposons que X prenne k classes et Y prenne m modalités. Donc, p_{mk} représente la proportion de d'observation qui prennent la valeur de m-ème dans la classe k-ème. L'indice de Gini est défini comme suivant :

$$G = 1 - \sum_{k=1}^{K} p_{mk}^2$$

Lorsque tous les éléments de l'ensemble sont dans la même classe de la variable cible, celle-ci atteint sa valeur minimale (zéro). L'algorithme calcule le coefficient de Gini de chaque variable explicative et sélectionne la variable avec le plus petit coefficient de Gini comme variable de segmentation.

Si X est une variable continue, comme les variables de PM d'ouverture, de taux technique de notre étude sont des variables continues, l'algorithme utilisé est le suivant :

- Trier les données X du plus grand au plus petit
- Calculer la valeur moyenne de X pour chaque valeur adjacente
- Calculer l'indice de Gini pour chaque poids moyen et choisir le poids moyen avec l'indice de Gini le plus bas au seuil de segmentation de la variable

• L'élagage d'arbre

Les arbres CART peuvent créer des arbres trop complexes, et ces arbres ne peuvent pas être bien généralisés à partir des données d'apprentissage. C'est ce qu'on appelle le « sur-apprentissage », ce qui signifie que les données expliquent un très bon ensemble de dataset « *training* », mais ne peuvent pas prédire de nouvelles données. Des mécanismes tels que l'élagage sont nécessaires pour éviter ce problème. Un arbre plus petit avec moins de divisions peut entraîner moins de variance et une meilleure interprétation en échange d'un léger biais.

Dans CART, le critère utilisé pour élaguer l'arbre est le « *cost complexity* ». Il se définit par l'erreur de classification d'un arbre plus un facteur de pénalité pour la taille de l'arbre. Cet algorithme est construit en utilisant le principe de la validation k-fold. Le facteur de pénalité est basé sur un paramètre, appelons-le α, qui est la pénalité par nœud. Le critère de complexité pour cet arbre est donc :

$$ERR(T) + \alpha \times L(T)$$

où $ERR(T)$ est la fraction des observations de données de validation qui sont mal classées par l'arbre T et $L(T)$ est le nombre de feuille de l'arbre T.

Lorsque $\alpha = 0$, l'arbre s'appelle l'arbre maximal. $T_1, T_2, ...$ respectivement sont les sous-arbres. Au fur et à mesure que nous augmentons α, avec chaque valeur α, nous obtenons un sous-arbre qui minimise le score de l'arbre qui est égal à la somme mentionnée ci-dessus. Dès qu'on trouve le sous-arbre correspond à sa valeur α, nous effectuons une validation croisée k fois et calculons $ERR(T)$ pour chaque « *testing set* ». L'alpha et l'arbre qui donnent le plus faible $ERR(T)$ en moyenne seront choisis comme arbre optimal.

L'un des avantages de l'utilisation de l'arbre de décision est que la connaissance de la relation entre les variables descriptives. Contrairement à la méthode de régression, y compris la Régression Logistique, la relation n'est pas nécessairement linéaire. L'arbre de décision peut détecter un pattern complexe de données. Cependant, il reste de nombreux défauts. C'est un algorithme glouton car il résout le problème localement. Il peut être aussi peu robustes. Notons que les arbres élagués fonctionnent toujours mieux sur l'ensemble de validation,mais pas nécessairement sur l'ensemble de test (en fait, ils ont également des performances égales ou pires sur l'ensemble de « training »).

3.1.3 L'algorithme de Random Forest

L'arbre de décision rencontre avec un problème d' « *overfitting* » et l'ignorance des variables en cas de petite taille d'échantillon. L'idée de la forêt aléatoire qui a été proposé par Ho (1998), et développé par Breiman (2001) pour devenir une méthode robuste de partition récursive particulièrement bien adaptée aux problèmes de petite taille d'échantillon. Les forêts aléatoires surpassent généralement les arbres de décision, mais leur précision est inférieure à celle des arbres améliorés par gradient.

Les forêts aléatoires sacrifient l'interprétabilité, mais améliorent généralement de manière significative les performances du modèle final. La forêt aléatoire est l'un des algorithmes d'apprentissage automatique les plus utilisés pour la classification. Il peut également être utilisé pour les modèles de régression (c'est-à-dire les variables cibles continues), mais il fonctionne principalement avec les modèles de classification (c'est-à-dire les variables cibles catégorielles).

En général, la forêt aléatoire est un modèle composé de nombreux arbres de décision. Plutôt que de simplement calculer la moyenne de la prédiction des arbres, ce modèle utilise deux concepts clés qui lui donnent le nom de « aléatoire » : échantillonnage aléatoire des points de données d'apprentissage lors de la création d'arbres et de sous-ensembles aléatoires de variables pris en compte lors de la division des nœuds.

Chaque arbre est cultivé comme suit :

Etape 1 : Sélection aléatoire des échantillons. Chaque arbre est formé sur environ 2/3 des données d'apprentissage. Les échantillons sont tirés avec remplacement, connu sous le nom de *bootstrap*. Cela signifie que certains échantillons seront utilisés plusieurs fois dans un même arbre. Les prédictions sont faites en faisant la moyenne des prédictions de chaque arbre de décision. Cette étape d'apprentissage sur différents sous-ensembles des données,

puis de calculer de la moyenne des prédictions, est connue sous le nom de *bagging*.

Etape 2 : Sélection aléatoire des variables. Certaines variables prédictives (par exemple, m) sont sélectionnées au hasard parmi toutes les variables prédictives. Les nœuds sont partitionnés en utilisant une affectation optimale pour ces m variables. Fondamentalement, m est la racine carrée du nombre total de tous les prédicteurs de la classification. Pour la régression, m est le nombre de tous les prédicateurs divisé par 3.

Etape 3 : Pour chaque arbre, en utilisant les données restantes (36,8%), calculez le taux d'erreur de classification - *out of bag* (OOB). Agréger l'erreur de toutes les arbes pour déterminer le taux d'erreur OOB global pour la classification. Si nous cultivons 200 arbres, un record sera en moyenne OOB pour environ 0,37 * 200 = 74 arbres.

Etape 4 : Chaque arbre donne une classification sur les données restantes (OOB), et nous disons que l'arbre «vote» pour cette classe. La forêt choisit la classification ayant le plus de voix sur tous les arbres de la forêt. Pour une variable dépendante binaire, le vote sera OUI ou NON, comptez les votes OUI. Il s'agit du score RF et le pourcentage de votes OUI reçus est la probabilité prévue. Dans le cas de la régression, il s'agit de la moyenne de la variable dépendante.

«Aléatoire» se réfère principalement à deux processus : des observations aléatoires pour faire croître chaque arbre et des variables aléatoires sélectionnées pour le fractionnement à chaque nœud. Le taux d'erreur de la forêt dépend de deux choses :

- La corrélation entre deux arbres quelconques de la forêt. L'augmentation de la corrélation augmente le taux d'erreur de la forêt.
- La force de chaque arbre individuel de la forêt. Un arbre avec un faible taux d'erreur est un classificateur puissant. Augmenter la force des arbres individuels diminue le taux d'erreur globalement.

La réduction de mtry (nombre de variables aléatoires utilisées dans chaque arbre) réduit à la fois la corrélation et la force. Quelque part entre les deux se trouve une gamme "optimale" de mtry. Une autre variable important dans l'algorithme de forêt aléatoire est le nombre d'arbres utilisés dans la forêt ou "ntree".

La forêts aléatoires est un algorithme puissant mais elle n'est pas bonnes pour généraliser des cas avec des données nouvelles. Par exemple, si je vous dis qu'une glace coûte 1$, 2 glaces coûtent 2$ et 3 glaces coûtent 3$, combien coûtent 10 glaces ? Une régression linéaire peut facilement comprendre cela, tandis qu'une forêt aléatoire n'a aucun moyen de trouver la réponse.

Les forêts aléatoires sont biaisées en faveur de la variable catégorielle ayant plusieurs niveaux (catégories). C'est parce que la sélection des caractéristiques basée sur la réduction des impuretés est biaisée en faveur de la préférence des variables avec plus de catégories, de sorte que la sélection des variables (importance) n'est pas précise pour ce type de données.

3.2 Évaluation de performance

Maintenant, nous voulons voir si la valeur observée de Y est proche des prédictions $\hat{Y}$. Dans la section suivante, nous présenterons quelques techniques appliquées aux modèles de classification supervisée.

3.2.1 La matrice de confusion

La matrice de confusion effectue des mesures pour le problème de classification de l'apprentissage automatique où la sortie peut être de deux classes ou plus. Elle est souvent utilisé pour décrire les performances d'un modèle de classification (ou "classificateur") sur un ensemble de données de test dont les vraies valeurs sont connues. La matrice de confusion elle-même est relativement simple à comprendre.

Il s'agit d'un tableau avec 4 combinaisons différentes de valeurs prédites et réelles. Nous démontrons un cas où la variable cible n'a que deux classes ici dans la figure 3.2. C'est également le cas de notre étude puisque nous essayons de prédire si un adhérent rachète ou non son contrat d'ici la fin de la période observable.

		Valeur actuelle	
		Positif (1)	Négatif (0)
Valeur prédite	Positif (1)	a	b
	Négatif (0)	c	d

FIGURE 3.2 – Illustration de la matrice de confusion

Nous nous concentrons sur les ratios suivants :

- a sont les vrais positifs. Les observations sont correctement classées positives
- c sont les faux positifs. Les observations sont incorrectement classées positives (erreur de type 1)
- de la même manière, b et d respectivement sont les faux négatifs (erreur de type 2) et vrais négatifs

- La *sensibilité* (en anglais *recall*). Le ratio représente de toutes les classes positives, le taux de bonne prédiction

$$S_e = \text{Sensibilité} = Recall = \frac{a}{a+b}$$

- La *spécificité* indique la proportion de négatifs détectés

$$S_p = \text{Spécificité} = \frac{d}{c+d}$$

- La *précision* représente parmi toutes les classes que nous avons prédites comme positives, combien sont réellement positives.

$$precision = \frac{a}{a+c}$$

- La *taux de succès*(en anglais *accuracy*) représente parmi toutes les classes, combien d'entre eux nous avons prédit correctement.

$$accuracy = \frac{a+d}{n}$$

- La *F-measure*. Il est difficile de comparer deux modèles avec une faible précision et une sensibilité élevé. Donc, pour les rendre comparables, nous utilisons la F-measure. Il permet de mesurer la sensibilité et la précision en même temps. Il utilise la moyenne harmonique à la place de la moyenne arithmétique en punissant davantage les valeurs extrêmes.

$$F - measure = \frac{2 \times S_e \times precision}{S_e + precision}$$

3.2.2 La courbe ROC - AUC

Lorsque nous devons vérifier ou visualiser les performances du problème de classification multi-classes, nous utilisons la courbe AUC (Area Under The Curve) ROC (Receiver Operating Characteristics). C'est l'une des mesures d'évaluation les plus importantes pour vérifier les performances de tout modèle de classification.

La courbe ROC met en relation le taux de vrais positifs (la sensibilité) et le taux de faux positifs (1 - spécificité) dans un graphique. Dans le processus de prédiction de Y, nous comparerons la probabilité que Y soit égal à 1 avec le seuil s. Normalement ce seuil est égal à 0,5 . La courbe ROC varie s entre 0 et 1. Pour chaque s, on calcule le taux de vrais positifs et le taux de faux positifs.

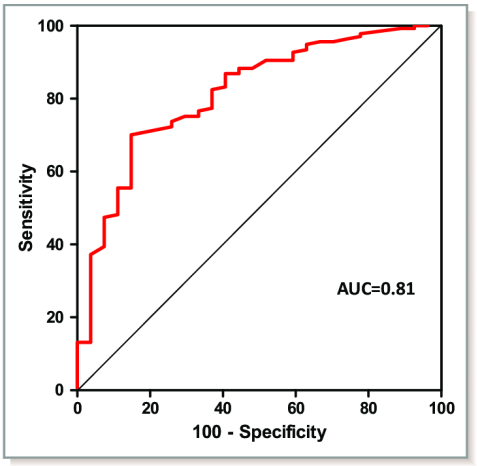


FIGURE 3.3 – Illustration du courbe

AUC est la zone en dessous de la courbe ROC. Lorsque l'AUC est de 0,7, cela signifie qu'il y a 70 % de chances que le modèle soit capable de faire la distinction entre la classe positive et la classe négative. Lorsque l'AUC est d'environ 0, le modèle renvoie en fait les classes. Cela signifie que le modèle prédit une classe négative en tant que classe positive et vice-versa. L'AUC permet de comparer rapidement la capacité de prédire des différents modèles.

La question est maintenant « quelle doit être l'AUC ? ». Il est facile de comparer les modèles pour voir quels modèles sont les meilleurs en utilisant l'AUC. Mais quand il s'agit d'identifier un seuil qui représente un « bon » modèle, c'est compliqué. Cela est dû au fait que chaque domaine et chaque étude accepteront un niveau différent d'AUC. Dans cette étude, nous utilisons la référence du cours de Data Science de l'Université Lyon 2 :

Valeur de l'AUC	Commentaire
$AUC = 0.5$	Pas de discrimination.
$0.7 \leq AUC < 0.8$	Discrimination acceptable
$0.8 \leq AUC < 0.9$	Discrimination excellente
$AUC \geq 0.9$	Discrimination exceptionnelle

FIGURE 3.4 – Interprétation des valeurs du AUC

Ici, un modèle avec un AUC de 70% est acceptable.

3.3 Résultats des différents modèles présentés

3.3.1 Logistique régression

Tout d'abord, nous construirons un modèle logistique de base sur les variables dont nous disposons : provision mathématiques, participation aux bénéfices, type de produits (altiscore ou agmf) et taux technique.

Paramètres	**Estimation**	**Erreur type**	**Khi 2 de Wald**	$Pr > Khi2$
(Intercept)	0.725	0.2311	3.137	0.001 71
AGE	−0.010 92	0.002 571	−4.247	<0001
ANCIEN	−0.032 36	0.006 108	−5.298	<0001
PB	0.001 067	0.000 161	6.631	<0001
PM	−0.000 032	0.000 004	−8.081	<000
PRODUITalt	−1.303	0.094 320	−13.815	<0001
TAUX	−1.176	0.1723	−6.828	<0001

Table 3.1 – Modèle logistique sans segmentation

Sur la base du modèle ci-dessus, nous pouvons voir que toutes les valeurs p sont inférieures à 1%. On peut donc rejeter l'hypothèse que le coefficient d'individu est nul. En d'autres termes, toutes les variables ont un effet sur le modèle.

On peut remarquer que la probabilité de racheter à la fin de la période d'observation son contrat AGMF est supérieure à Altiscore. Le coefficient de la variable PRODUIT est de -1,303. La variable TAUX a également un impact important sur les chances de rachat. On aussi observe que les valeurs des coefficients sont très petites, notamment les deux variables PM et PB.

Nous interprétons le résultat comme suit : si la valeur de $PM_{ouverture}$ de l'adhérent augmente de 1 euro tandis que d'autres facteurs restent inchangés, alors le $log(odd)$ augmente de 0,000032 fois. Autrement dit, la odd-ratio augmente $e^{0,000032} = 1,000032 \approx 1$. Cela montre que la segmentation est nécessaire.

De plus, nous mesurons également l'efficacité du modèle à travers la courbe ROC-AUC ci-dessous. Avec un AUC = 69%, le niveau de prédiction du modèle ne répond pas aux exigences discriminantes acceptables.

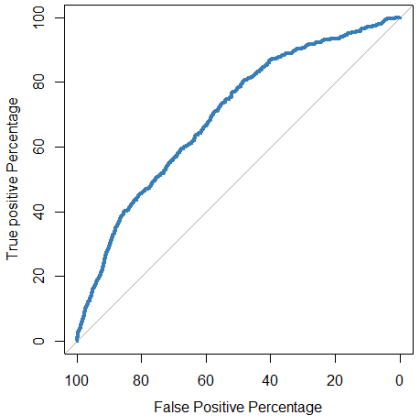


FIGURE 3.5 – Courbe de ROC-AUC modèle logistique sans segmentation

Comme mentionné dans l'idée générale du traitement des données dans la partie 2, la transformation de discrétisation des données peut être effectuée avec 3 méthodes : uniforme, quantile et *cluster*. Nous essayons chacune de ces méthodes pour trouver la meilleure méthode de transformation de nos données.

Dans cette étude, la méthode qui s'est avérée créer la meilleure discrétisation est le *clustering* en tant qu'application simple de l'arbre de décision. En effet, nous construisons un arbre (régression ou classification) pour chaque variable. Dans ces arbres, les variables cibles et explicatives sont les mêmes. Il aide à créer un *cluster* de la nature de chaque variable en maximisant les dispersions interclasses.

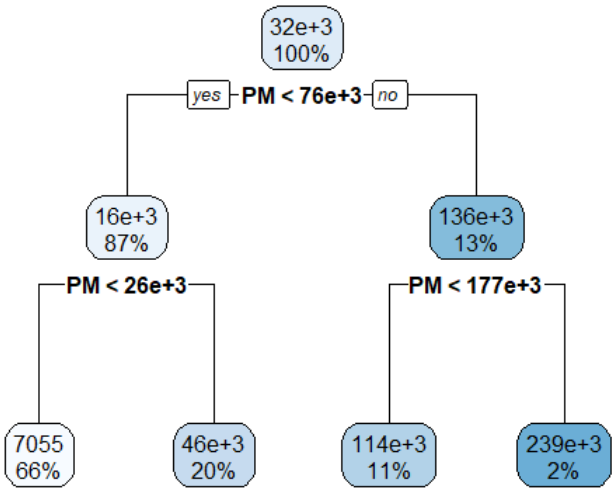


FIGURE 3.6 – Segmentation variable PM

Cette méthode nous donne un grand avantage par rapport à d'autres méthodes de *clustering* car elle est très simple et efficace. De plus, cela nous donne la possibilité d'interpréter (par rapport à k-means). Les résultats du processus de discrétisation sont présentés dans le tableau suivant :

Variable	Classe	Tranche
AGE	Classe 1	Moins de 42 ans
	Classe 2	De 42 à 62 ans
	Classe 3	De 62 à 70 ans
	Classe 4	Plus de 70 ans
PM	Classe 1	Moins de 26 404,61 euro
	Classe 2	De 26 404,61 à 75 929,65 euro
	Classe 3	De 75 929,65 à 176 732,6 euro
	Classe 3	Plus de 176 732,6 euro
PB	Classe 1	Moins de 1 602 euro
	Classe 1	De 1 602 à 3 904 euro
	Classe 1	Plus de 3 904 euro
TAUX	Classe 1	Taux égale à 0%
	Classe 2	plus de 0%

TABLE 3.2 – Segmentation des variables

Maintenant que nos données sont discrétisées, nous construisons un nouveau modèle avec ces variables. Le tableau 3.3 fournit des informations d'estimation des coefficients pour le nouveau modèle.

Dans ce nouveau modèle, presque toutes les variables sauf l'intercept et la tranche_AGE_class4 sont significatives au seuil de 1%. Cette tranche représente le groupe de personne de plus de 70 ans. Cela signifie que nous ne pouvons pas rejeter l'hypothèse selon laquelle le coefficient de ces variables est égal à zéro.

Cela conduit au fait que nous ne pouvons pas conclure les différences entre le *odd ratio* d' adhérent dans la tranche d'âge 1 (inférieur à 42 ans) et la tranche d'âge 4 (au-dessus de 72 ans). On ne peut conclure que lorsqu'une adhérent qui est dans la tranche d'âge, PM, PB de classe 1 ; taux égal à 0% et 0 année d'ancienneté, alors son *odd ratio* égal à $e^{0,062245}$ $(e^{Intercept})$.

Cependant, ce modèle a la capacité de prédire même lorsque les coefficients ne sont pas significatifs. Une vue de la puissance prédictive est faite à travers la courbe ROC, comme c'était le cas pour les modèles précédents. Ce nouveau modèle nous donne une meilleure

Paramètres	**Estimation**	**Erreur type**	**Khi 2 de Wald**	$Pr > Khi2$
(Intercept)	0.061 245	0.235 996	0.260	0.795 238
tranche_AGEclass 2	0.574 670	0.191 674	2.998	0.002 716
tranche_AGEclass 3	0.699 220	0.195 908	3.569	<0001
tranche_AGEclass 4	−0.158 458	0.198 126	−0.800	0.423 836
ANCIEN	−0.043 697	0.006 023	−7.255	<0001
tranche_PBclass 2	0.951 096	0.242 587	3.921	<0001
tranche_PBclass 3	2.613 176	0.487 587	5.359	<0001
tranche_PMclass 2	−0.266 475	0.097 635	−2.729	0.006 347
tranche_PMclass 3	−1.591 492	0.276 122	−5.764	<0001
tranche_PMclass 4	−4.054 326	0.638 347	−6.351	<0001
PRODUITalt	−1.456 748	0.093 236	−15.624	<0001
TAUX	−0.784 867	0.152 547	−5.145	<0001

TABLE 3.3 – Modèle logistique avec segmentation

valeur d'AUC qui est de 70,43% par rapport au modèle sans segmentation. Le modèle a été amélioré mais l'AUC n'est toujours pas entre 0,8 et 0,9. Donc, la question qui se pose est : « Et si la relation entre le *odd ratio* et les variables n'était pas expliquée par une fonction linéaire ? ». Ce problème est résolu avec un modèle basé sur un arbre dans la partie suivante.

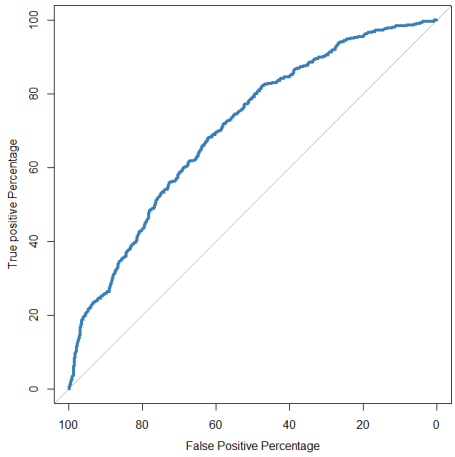


FIGURE 3.7 – Coubre ROC-AUC nouvelle modele

3.3.2 Abre CART - Random Forest

L'arbre de décision est dit plus « visualisable », cependant, en réalité, dans certains cas, la structure de l'arbre est également si compliquée que nous ne pouvons pas le suivre en détail. Cela se produit lorsque nous avons trop de variables différentes dans notre modèle ou que la variable cible est continue, ce qui conduit à de nombreux seuils de séparation différents. En effet, dans notre cas, l'arbre est très complexe. Nous ne pouvons pas observer chaque feuille même après l'élagage. Par conséquent, nous construisons l'arbre et la forêt aléatoire en même temps puis comparons ces deux modèles.

Rappelons que l'arbre de décision est construit en deux étapes : construction d'arbres et élagage d'arbres. L'élagage des arbres se fait en minimisant le «*cost complexity*» qui est noté dans R avec le «*complexity parameter*».

Après avoir fait pousser un maximum d'arbre, on essaie de tailler l'arbre pour éviter le sur-apprentissage. Ici, le graphique a deux dimensions, l'axe des x est la taille de l'arbre et l'axe des y représente l'erreur du modèle. Dans l'axe ci-dessous, nous pouvons voir que la complexité du coût correspond au nombre d'arbres. La fonction de R choisira alors la taille de l'arbre correspond au cp minimum.

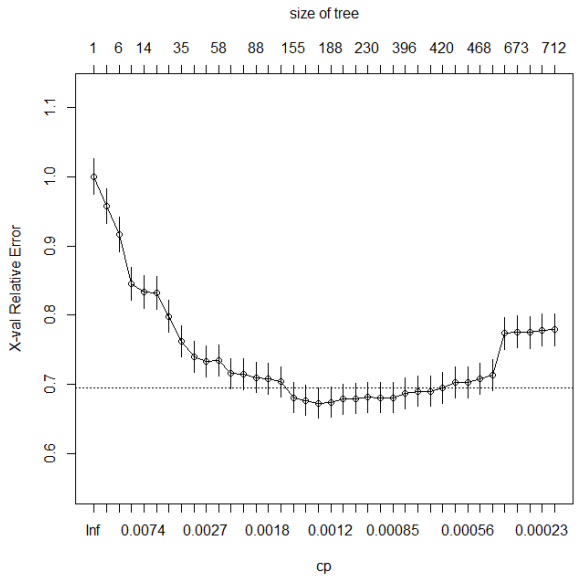


FIGURE 3.8 – CP en foncion de error

Après l'élagage, notre modèle donne un AUC de 80%. L'arbre reste très complexe à visualiser. Nous allons donc essayer d'interpréter certaines de ses informations. L'un d'eux est la table de l'importance des variables (table 3.4). L'âge est la plus importante parmi les variables et la Taux technique est la moins importante.

Il est intéressant de noter que les variables PB, PM et ANCIEN sont les 3 prochaines variables importantes. Les importances de ces variables sont très proches. Cependant, elle

sont complètement différents des deux dernières variables : PRODUIT et TAUX

	AGE	PB	PM	ANCIEN	PRODUIT	TAUX
Variable importance	973.107 88	644.317 03	576.823 31	484.950 51	87.808 56	26.263 84

Table 3.4 – Table de l'importance des variables

Puis, nous allons construire un modèle de forêt aléatoire. Pour éviter le sur-apprentissage, la forêt aléatoire nécessite également une étape d'élagage. On va d'abord parler de la matrice de confusion. Il s'agit d'un tableau qui affiche les données observées en lignes et les données prédites par l'algorithme en colonnes. La somme de la diagonal de 6256 est le nombre d'observations bien classées.

L'erreur de *out-of-bag* est de 12,84%. C'est l'un des indicateurs que l'on mentionne fréquemment lorsque l'on parle de modèle forêt aléatoire. Cela signifie que le modèle a 12,84% de chance de prédire mal. Si la personne ne rachète pas mais que le modèle prédit qu'il l'a fait, cela signifie que nous surestimons le risque réel et coûte plus chère en SCR pour l'entreprise. En revanche, si la personne a racheté mais que le modèle dit qu'il ne l'a pas fait, c'est plus dangereux où la sous-estimation du risque peut entraîner une perte potentielle de liquidité. Il n'y a que 0,8% de chance que nous ayons cette erreur ce qui équivaut à 7% de l'erreur.

OOB estimate of error rate : 12.84%		
	0	1
0	5862	58
1	864	394

Table 3.5 – La matrice de confusion

Cependant, lorsque vous travaillez sur l'ensemble de test pour voir la puissance de prédiction, l'AUC de la forêt aléatoire est de 89,85%. Cela prouve que la forêt aléatoire est un modèle très robuste pour prédire la classification.

Pour examiner l'importance d'une variable, nous avons 2 méthodes : l'impureté de Gini et la précision de diminution.

L'« impureté de Gini » ou la « *mean decrease impurity* » est définie comme la diminution totale de l'impureté du nœud en moyenne sur tous les arbres. « *Mean decrease accuracy* », d'autre part, mesurez la diminution de la précision sur les données OOB lorsque vous permutez de manière aléatoire les valeurs de cette fonctionnalité. Si la diminution est faible, alors la caractéristique n'est pas importante, et vice-versa.

L'impureté de Gini pourrait être assez trompeuse lorsqu'elle réduit l'importance de la variable PRODUIT. En effet selon l'indice, elle a un discriminant fort dans le modèle

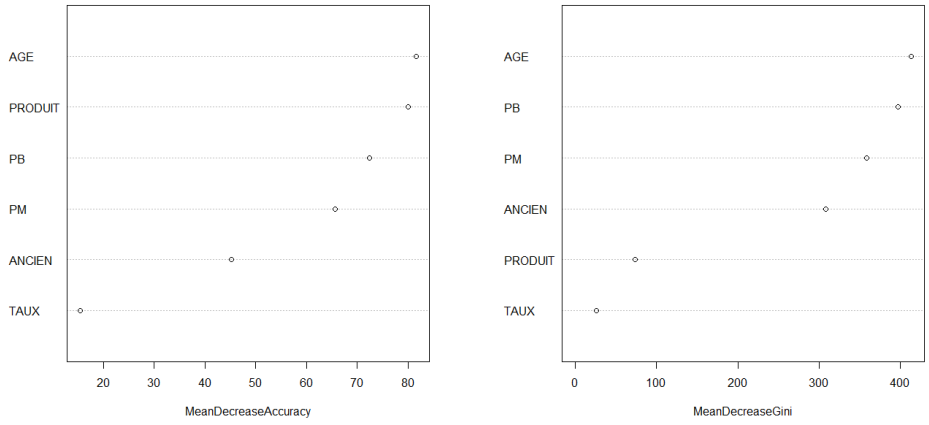


FIGURE 3.9 – L'importance des variables

mais réellement, elle n'apparaît qu'une seule fois dans le premier nœud. Avec ce résultat, il pourrait sembler que la variable TAUX ait un petit impact sur notre modèle.

Ensuite, nous nous intéressons à une intrigue très intéressante : partialPlot. Il nous donne l'effet marginal d'une variable sur le résultat d'un modèle d'apprentissage automatique (Friedman (2001)). Il peut montrer si la relation entre la cible et une caractéristique est linéaire, monotone ou plus complexe. Pour une classification, x-axis est le *odd-ratio* qui dans cette étude est le *odd-ratio* du rachat.

Dans le graphique de « *Patial Dependence on Age* », le nombre de personnes ayant rachetées leur contrats augmente rapidement de l'année 5 à l'année 8. Cela s'explique par l'avantage fiscal. Cependant, il passe de 8 à 10. Selon la loi du comportement, si l'adhérent ne bénéficie pas de l'avantage fiscal, il conservera alors l'investissement pour des objectifs financiers à long terme. Ce pourcentage a également tendance à augmenter progressivement jusqu'à l'âge de 25 ans.

Le graphique de dépendance partielle selon l'âge montre que l'odds ratio est assez stable de 0 à 70, et tend à augmenter rapidement au cours des années suivantes. Ceci est cohérent avec les caractères du produit d'épargne/retraite. Les adhérents achètent plus à l'approche de l'âge de la retraite. En revanche, l'odd-ratio de rachat est plus faible pour les $PM_{ouverture}$ faibles. Ceci est raisonnable car il s'agit d'un produit d'épargne et les adhérents ont tendance à attendre que la valeur accumulée soit remboursée. Ce qui est intéressant, cependant, c'est que après 10 000 euro, l'augmentation de la PM n'implique pas une augmentation du *odd-ratio*. Les adhérents avec un PB élevé, ce qui signifie que leurs épargnes sont rentables, sont plus susceptibles de conserver le contrat.

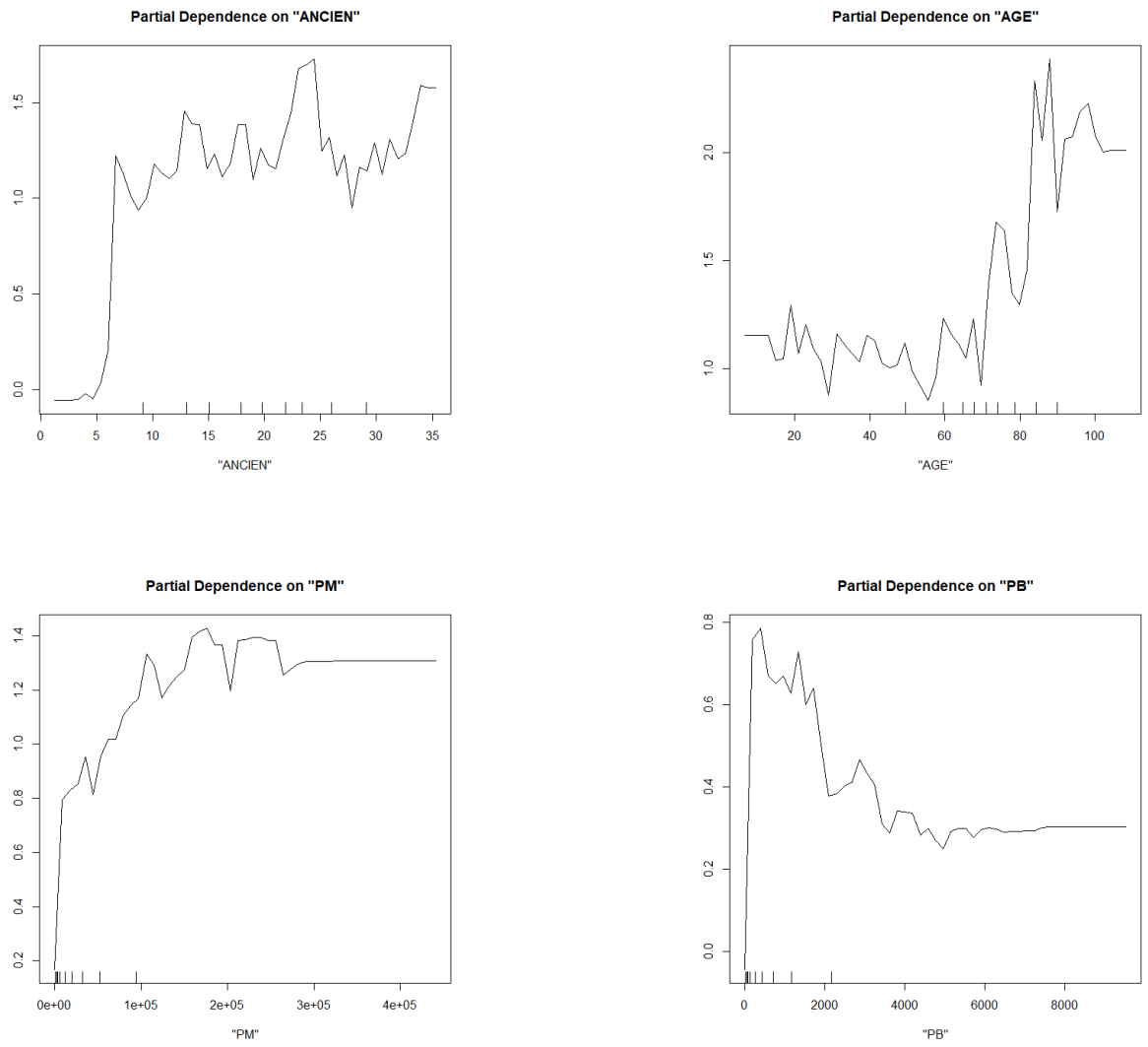


FIGURE 3.10 – Partial Dépendance des variables

3.4 Conclusion et critique

L'étude statique ci-dessus donne un aperçu du comportement de rachat des adhérents. En effet, à travers la construction et la comparaison de modèles, nous pouvons voir que la relation entre l'odd-ratio et les variables indépendantes est une relation non linéaire. L'utilisation d'un modèle d'arbre peut améliorer considérablement les résultats de prédiction de notre modèle.

Entre les deux modèles dit *tree-based*, le modèle de forêt aléatoire a des résultats robustes pour conclure sur la possibilité de rachat. Son pouvoir de prédiction (AUC) dans cette

base de données est de 89,95%. La variable âge s'avère être systématiquement l'une des variables discriminantes. Par ailleurs, les variables qui représente la richesse de l'adhérent i.e la PM et la PB jouent également un rôle très important. La décision de rachat dépend également de l'ancienneté du contrat et du type de contrat que détiennent les clients.

Avec la dépendance partielle des variables, nous avons analysé davantage dans le modèle Random Forest. Il permet de visualiser l'interaction entre chaque variable et la variable cible. Il nous donne une explication précieuse du comportement de l'adhérent, ce qui est utile pour la prochaine partie de l'étude.

Cependant, il a encore des faiblesses. L'un d'elle est le fait que nous avons ignoré l'effet de censure des données qui est l'une des problématiques centrales de cette base de données. Bien qu'il soit très clair que le pouvoir de prédiction de la forêt aléatoire est fort, nous ne pouvons modéliser que l'effet du rachat total mais pas du rachat partiel. Notez que la variable cible pour le rachat partiel est « combien d'argent une personne rachète ? » mais pas « cette personne rachète-t-elle ? ». Dans la partie suivante, nous présentons le modèle qui peut corriger ces faiblesses.

Chapitre 4

Approche de modèle d'analyse de survie

Cette section étudie les taux de rachat en termes de modèles de durée. L'objectif principal de cette section est de surmonter les faiblesses des études statiques de la section précédente. En effet, le modèle de survie prend en compte la problématique des données censurées, en complément des recherches sur le rachat partiel et le taux de rachat selon le montant. On introduit les notion générale de modèle de durée. Ensuite, nous étudierons le modèle de Kaplan Meier et ses étapes de mise en œuvre et ses résultats. En fin, nous étudierons une manière simple d'intégrer des variables explicatives dans le modèle.

4.1 Contexte littérature et critique

Le modèle de survie est un modèle applicable dans de nombreux domaines de la vie. En parlant de survie, nous pouvons immédiatement penser à des événements tels que la mort, la maladie ou l'invalidité. En effet, le modèle de survie concerne le moment auquel une personne, une chose ou un phénomène peut rencontrer un « événement ». C'est pour cette raison que l'analyse de survie a tant d'applications en biologie, en science des matériaux, etc. En particulier, l'analyse de survie a de grandes applications en assurance.

Cette application ne s'arrête pas seulement aux études d'espérance de vie ou de la durée jusqu'à l'incapacité/invalidité. Il existe aussi des études de faillite avec les modèles de durée et dans notre étude celui des taux de rachat. Il existe de nombreux travaux de recherche sur le taux de rachat qui utilisent cette méthode.

Nous notons que ces formes de rachat peuvent affecter énormément la nature du risque. En effet, le rachat total implique que le contrat prendra fin immédiatement après l'événement. Cela lui donne un caractère de sortie de contrat. Par conséquent, il peut être modélisé à l'aide de méthodes d'analyse de survie.

Au contraire, une personne qui rachat partiellement, ponctuellement ou de façon programmée son contrat n'implique pas une sortie de contrat. Pour cette raison, une simple analyse de survie ne suffit pas. Cependant, si nous pensons que chaque euro est un « comme une personne », dans ce cas, chaque euro rendu est une sortie définitive assimilable à une personne. Il y a encore beaucoup de critiques dans cette approche car quand une personne rachète son contrat, la quantité de rachat est vue comme un vague de personnes qui décède en même temps pour une même raison. De plus, il est très compliqué de surveiller « chaque euro » car cela signifie un énorme ensemble de données avec des centaines de millions de lignes.

De plus si un client choisit une option de rachat programmée, l'obligation de paiement est maintenant certain. Le modèle deviendra désormais un modèle pour voir quand la PM sera payé totalement. Cela implique une complexité très élevée en raison du fait que ce paiement dépend non seulement de la longévité du risque mais aussi du risque financier à travers les bénéfices de participation. Cependant, pour simplifier le modèle, nous utiliserons toujours les méthodes d'analyse de survie et de rachat partiel étudié en montant.

Une difficulté particulière liée à l'analyse de survie provient principalement du fait que seuls certains individus ont vécu l'événement, et le temps de survie d'un sous-ensemble de ce groupe d'étude est inconnu. Ce phénomène a déjà été mentionné au chapitre 2 où on analyse les caractéristiques des données.

La censure peut se produire des manières suivantes :

(a) L'adhérent n'a pas (encore) racheté son contrat à la fin de l'étude.

(b) L'adhérent décède ce qui rend l'observation impossible.

Ces temps de survie interrompus sous-estiment le temps réel jusqu'à l'événement. En visualisant le processus de survie de l'individu sur une chronologie, ces incidents (supposés se produire) ont dépassé la fin de la période de suivi. Cette situation est considéré une censure à droite. La plupart des données de survie sont des observations qui sont censurées à droite. Il existe des données de censure par intervalle et de censure à gauche et des méthodes de modélisation de celles-ci (Hosmer Jr & Lemeshow (1999)). Cependant, compte tenu de nos recherches, nous ne considérons que les données qui sont censurées à droite.

4.2 Construction de taux brut

Il existe plusieurs méthodes de construction de taux brut dans l'analyse de survie ou dans notre étude, de loi de rachat. Grâce à un graphique du cours de « *Machine Learning for Survival Analysis* » de l'université du Michigan, nous montre un bref résume des différentes approches qui peuvent être utilisées pour résoudre des problèmes d'analyse de survie. Suivant le modèle statistique, nous avons trois approches : l'approche paramétrique, semi-paramétrique et non-paramétrique. Ce sont les trois approches les plus utilisées dans les modèles actuariels. En plus de cela, nous avons l'approche Machine Learning qui est très intéressante et nous en apprendrons davantage au Chapitre 5.

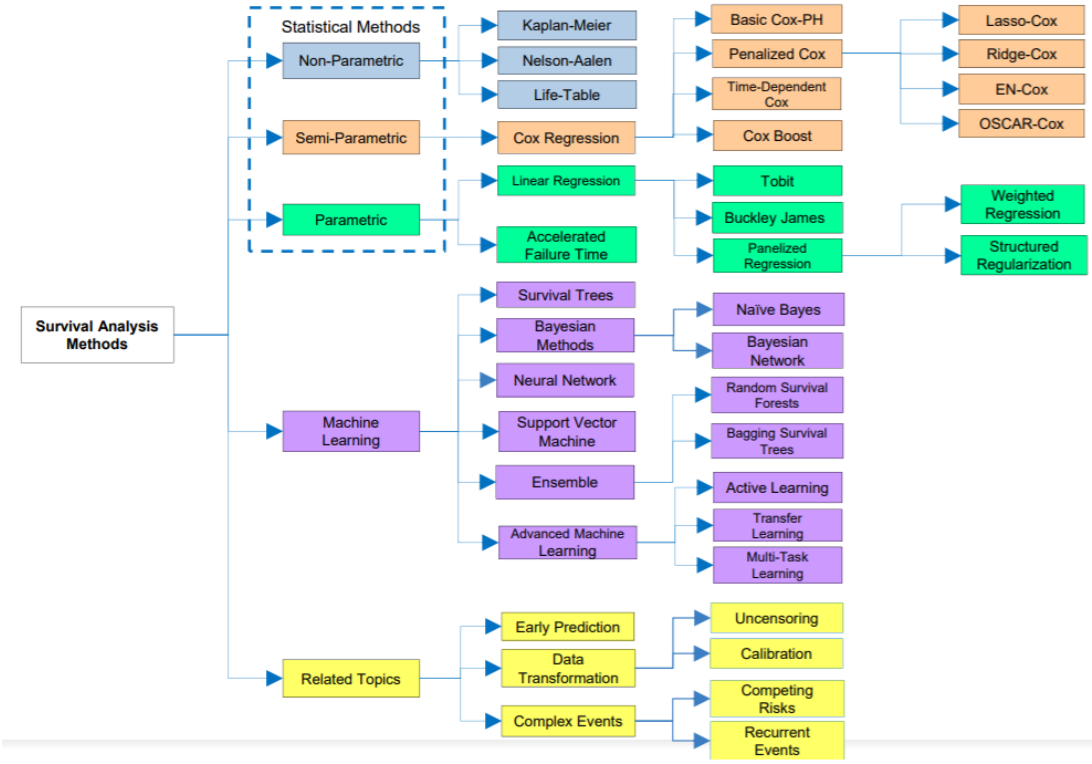


FIGURE 4.1 – L'approche d'analyse de survie

Les modèles paramétriques ou semi-paramétriques supposent une distribution paramétrique de la fonction de hasard. Il s'agit d'estimer les paramétriques du modèles avec le maximum de vraisemblance. Les modèles non paramétriques, au contraire, ne nécessitent aucune hypothèse sous forme de fonction de risque. Cela nous donne une liberté de choix lorsque nous estimons le taux de rachat, en échange de la complexité de la technique statistique. Dans cette étude, nous nous concentrerons sur l'approche non paramétrique.

Tout d'abord, rappelons quelques notations concernant l'approche non-paramétrique :

– X_i : le temps jusqu'à évènement de i-ème observation (dans notre cas c'est le temps

jusqu'à rachat)

- C_i : horizon de censure à droite de la i-ème observation
- D_i : variable qui indique la censure de la i-ème observation et qui prend deux valeurs 0 ou 1. Donc $D_i = 1$ si $X_i \leq C_i$ et $D_i = 0$ si non.
- $N_i(t)$: variable qui indique la i-ème observation et qui prend deux valeurs 0 ou 1 où $N_i(t) = 1_{\{X_i \leq t\}}$.

Soit X une variable définie sur $[0, +\infty]$ et soit $F(x) = P(X \leq x)$ sa fonction de répartition. La fonction de survie est donc : $S(x) = 1 - F(x)$ et la densité :

$$f(x) = \frac{dF(t)}{dt} = \lim_{h \to 0} \frac{P(x \leq X \leq x + h)}{h}$$

La fonction de survie conditionnelle donne la probabilité de survie dans $u + x$ d'une personne qui a déjà survécu jusqu'à u. Elle est définie comme :

$$S_u(x) = P(X > u + x | X > u) = \frac{P(X > x + u)}{P(X > u)} = \frac{S(u + x)}{S(u)}$$

La fonction de hasard $h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{dlnS(x)}{dx}$ représente la fréquence de défaillance par unité de temps. La fonction de hasard cumulé est donc $H(x) = \int_0^x h(s)ds$. La relation entre la fonction de survie et sa fonction de hasard est :

$$S(x) = exp(-\int_0^x h(s)ds) = exp(-H(x))$$

Dans le cas où la variable X prend des valeur discrètes entières, la fonction de survie s'écrit : $S(k) = \sum_{m \geq k+1} p_m$ avec $p_k = P(X = k)$. La fonction de hasard $h(k) = P(X = k | X > k - 1)$. La relation entre la fonction de survie et la fonction de hasard est : $1 - h(k) = \dfrac{S(k)}{S(k-1)}$. Et on trouve $S(k) = \prod_{m=1}^{k}(1 - h(m))$.

Comme nous ne pouvons pas observer X en raison du problème des données censurées et nous ne pouvons observer que T où $T = X \wedge C$. Donc, nous en đéuíón que notre observation se porte uniquement sur la variable $N^1(t) = 1_{\{X \leq t; D=1\}}$.

4.2.1 Estimation de Nelson-Allen

Selon Nelson (1972), la fonction de hasard cumulé peut s'écrire :

$$\hat{H}_{NA}(t) = \int_0^t \frac{d\bar{N}^1(u)}{\bar{R}(u)}$$

où

- $\bar{R}(t)$: nombre d'observation à risque juste avant t
- $\bar{N}^1(t) = \sum_{i=1}^{n} N_i^1(t)$: nombre d'observations événement avant t

On peut ainsi réécrire l'équation ci-dessus sous la forme plus intuitive suivante :

$$\hat{H}_{NA}(t) = \sum_{T_i} \frac{d_i}{r_i}$$

avec d_i, r_i respectivement le nombre de personnes qui rachètent en T_i et nombre de personne à risque juste avant T_i.

Notons que $\mathbb{E}[\hat{H}_{NA}(t)] \leq \hat{H}_{NA}(t)$ ce qui implique que : l'estimateur de Nelson-Aalen a bien tendance à sous-estimer la fonction de hasard cumulée du modèle. Cette propriété est prouvée en utilisant la forme fonction du hasard cumulé (pas la forme intuitive) car cette fonction est continue à droite.

En exploitant la relation $S(t) = e^{-\hat{H}_{NA}(t)}$, on peut ainsi proposer comme estimateur de la fonction de survie de Harrington et Flemming :

$$\widehat{S_{HF}}(t) = e^{-\hat{H}_{NA}}(t)$$

Donc,

$$\hat{q}_x = 1 - \hat{p}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)} = 1 - exp(-\sum_{x<T_i<x+1} \frac{d_i}{r_i})$$

4.2.2 Estimation de Kaplan-Meier

La méthode de Kaplan-Meier a été introduite en 1958 par Edward L. KAPLAN et Paul MEIER dans un article du journal de l' « *American Statistical Association* » intitulé « *Nonparametric Estimation from Incomplete Observations* ».

L'estimateur de Kaplan Meier est défini par :

$$\widehat{S_{KM}}(t) = \prod_{T_i} (1 - \frac{d_i}{r_i})$$

$$\hat{q}_x = 1 - \hat{p}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)} = 1 - \prod_{x<T_i<x+1} (1 - \frac{d_i}{r_i})$$

Les estimateurs de Kaplan Meier et Nelson Aalen sont asymptotiquement équivalents. Une caractéristique clé de l'estimateur de Nelson Aalen est qu'il est supérieur à l'estimateur Kaplan Meier. En expression mathématique, on a $\widehat{S_{KM}}(t) \leq \widehat{S_{NA}}(t)$.

Lorsque les taux de sortie sont décroissants, l'estimateur de Kaplan-Meier est plus pertinent que celui de Nelson-Aalen. Ainsi, l'estimateur de Kaplan-Meier est adapté à notre étude dans la mesure où les taux de rachat sont décroissants.

Interprétation avec le modèle de la loi de rachat en montant :

$$\text{Taux de rachat d'ancienneté x} = 1 - \prod_{x<T_i<x+1} (1 - \frac{\text{Montant de rachat à moment i}}{PM_i})$$

Ce taux représente le taux de rachat dans l'année d'une personne d'ancienneté x où la PM_i est le montant de provision mathématique juste avant T_i. Il est calculé par :

$$PM_i = \text{PM au début d'année} + \sum \text{les transactions jusqu'à } T_i$$

Notons que ces transactions peuvent être positives ou négatives. Un des inconvénients de modèle de Kaplan Meier, c'est qu'il faut suivre la PM dans un fréquence plus élèves. Suivre la PM en temps réel est trop compliqué car les données de GPMA sont très volumineuses. Cette réalité conduit au fait que nous devons baser notre modèle sur une estimation de calcul pour la PM. En effet, le moment où les adhérents souhaitent remettre son épargne, ils ne connaissaient pas le montant des participation bénéfices qu'il percevra en fin d'année ou le taux applicable à son compte.

Normalement, cela peut être résolu en devisant l'intervalle d'un an en parties d_t égale à un mois. Ainsi, notre travail consiste à calculer la PM en début de mois et le montant de rachat chaque mois. A noter que le mois de l'année et le mois équivalent à ancienneté ne correspondent pas. Nous nous intéressons au mois selon l'ancienneté. Dans cette étude, en respectant les principes de l'estimateur de Kaplan Meier, nous suivrons chaque client au moment de son rachat (sans aucun arrondi au mois). Il conduit à une complexité de calculer le PM à chaque date de rachat due au volume des données. Cependant,en utilisant les fonctions du package **data.table** pour la manipulation des données dans R au lieu de la structure par **data.frame**, cela devient possible.

4.3 Lissage de taux brut

Les taux de rachat bruts sont généralement sensibles aux fluctuations d'échantillonnage donc, ils peuvent représente des irrégularités. Le processus d'estimation des taux bruts considère généralement les anciennetés indépendamment, et donc ne considère pas les interactions qui existent Planchet & Thérond (2006). Il est nécessaire de lisser ces taux afin d'obtenir une courbe de rachat qui progresse graduellement avec l'ancienneté.

La méthode de lissage a deux restrictions. Nous devons s'assurer que les taux estimés sont proches des taux initiaux (la précision). Nous souhaitons aussi que les taux ajustés soient le plus régulièrement possible (la régularité).

Il existe différentes techniques qui peuvent faire aussi l'objet d'une utilisation conjointe :

- Les lissages paramétriques : Makeham
- Les lissages non paramétriques : Whittaker-Henderson
- Les modèles relationnels : Brass

Les modélisations paramétriques ou relationnelles sont à privilégier dans le cas des petits échantillons. Il est facile d'interpoler l'estimation des taux de rachat à des ancienneté sans observation et d'extrapoler ceux à des anciennetés plus élevées.

Dans cette étude, nous choisissons la méthode lissages non paramétriques. Le principe consiste à substituer aux données brutes des valeurs lissées, en supposant que le rachat du groupe étudié est plutôt régulier, sans qu'aucune loi sous-jacente n'intervienne. Les méthodes les plus courantes sont celles des moyennes mobiles et de Whittaker-Henderson. Cette méthode trouve son origine dans les études de Whittaker (1922), et des contributions à la théorie ont été apportées par Henderson (1924), et d'autres.

En ce qui concerne le lissage, nous avons souvent besoin d'utiliser certaines définitions dans les séries temporelles comme :

- La différence avant : $\Delta u(x) = u(x+1) - u(x)$
- La différence avant : $\bigtriangledown u(x) = u(x) - u(x-1)$

où u est une serie. Ensuite nous avons :

$$\begin{aligned}\Delta^2 u(x) &= \Delta(\Delta u(x)) \\ &= \Delta(u(x+1) - u(x)) \\ &= \Delta u(x+1) - \Delta u(x) \\ &= u(x+2) - 2u(x+1) + u(x)\end{aligned}$$

En général,

$$\Delta^n u(x) = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} u(x+j)$$

- Lissage par le modèle Makeham

Le lissage avec le modèle de Makeham est un lissage paramétrique où on exprime de taux de hasard par la formule :

$$\mu_x = \alpha + \beta \times c^x$$

avec $\alpha > 0, \beta > 0, c > 1$. L'interprétation de cette formule est la décomposition du taux de rachat en :

– Un élément indépendant de l'ancienneté du contrat.

– Un élément croissant exponentiellement avec l'ancienneté du contrat

Le taux de rachat lissé est donc :

$$\begin{aligned} q_x = 1 - exp(-\int_x^{x+1} \mu_s ds) &= 1 - exp(-\int_x^{x+1} (\alpha + \beta \times c^s) ds) \\ &= 1 - exp(-\alpha - \frac{\beta c^x (c-1)}{ln(c)}) \end{aligned}$$

- Lissage avec la méthode Whittaker-Henderson

L'idée de base de cette méthode est de minimiser la somme des deux critères : la fidélité et la régularité. Le critère de fidélité se définit par :

$$F = \sum_{i=1}^{p} w_i (q_i - \hat{q}_i)^2$$

où $w_i > 0$ est le poids du taux de rachat de la i-ème ancienneté. z étant un paramètre du modèle, le critère de régularité s'écrit :

$$S = \sum_{i=1}^{p-z} (\Delta^2 q_i)^2$$

Le modèle de Whittaker-Henderson cherche l'estimation $\hat{q}_i$ en minimisant la somme :

$$M = F + h \times S$$

où h est un autre paramètre du modèle. Plus h petit plus la courbe des taux lissés sera proche des taux empiriques. Nous résolvons cela en trouvant $\hat{q}_i$ de sorte que $\frac{\partial M}{\partial q_i} = 0$. Exprimé sous forme matricielle, on a :

$$q = \begin{pmatrix} q_1 \\ \vdots \\ q_p \end{pmatrix} \qquad \text{et} \qquad \hat{q} = \begin{pmatrix} \hat{q}_1 \\ \vdots \\ \hat{q}_p \end{pmatrix}$$

Ensuite, la matrice de poids s'écrit :

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_p \end{pmatrix}$$

On obtient :

$$F = (q - \hat{q})^T W (q - \hat{q})$$

Et :

$$S = (\Delta^2 q)^T \Delta^2 q$$

avec $\Delta^2 q = K_i q$ où K_z la matrice de taille (p-z,p). Donc, on a :

$$M = (q - \hat{q})^T W (q - \hat{q}) + h q^T K_z^T K_z q$$

En résolvant $\dfrac{\partial M}{\partial q_i} = 0$, le taux lissé se calcule par :

$$q_{lisse} = (W + h K_z^T K_z)^{-1} W \hat{q}$$

Le choix des paramètres du modèle (h,z) est basé sur la perception humaine. Nous avons deux critères de sélection jusqu'à obtenir la courbe la plus appropriée. L'une est une forme de courbe qui maintient une forme propre aux taux bruts, et l'autre est basée sur des critères de validité statistique découlant de la taux lissée.

4.4 Validation de la loi de rachat

L'objectif est de s'assurer du respect de certaines règles de cohérence. Plusieurs éléments sont à vérifier : croissance des taux de rachat avec l'ancienneté, correspondance avec les données initiales, respect des connaissances a priori du rachat. Différents critères et tests sont susceptibles de mesurer la qualité de la modélisation.

Les tests suivants utilisée sont : le Test du Khi-deux ; Standardized mortality ratio (SMR) ; Test des changements de signe. Une fois l'ajustement (ou le lissage) effectué, le test du Khi-2 permet de vérifier la qualité globale des taux révisés en s'assurant qu'ils ne sont pas « trop loin » des taux estimés. L'hypothèse est :

$$H_0 : \hat{q} = q_i$$

$$H_0 : \hat{q} \neq q_i$$

On calcule la statistique :

$$Z = \sum_{i=1}^{p} n_i \frac{(\hat{q}_i - q_i)^2}{q_i(1 - q_i)}$$

Le SMR est défini comme le rapport du nombre de rachat observé au nombre de rachat prédits dans une population de référence, avec l'objectif de décider si la mortalité du

groupe observé est identique à celle du groupe de référence ; on a ainsi :

$$SMR = \frac{D}{E} = \frac{\sum_{i=1}^{p} D_i}{\sum_{i=1}^{p} n_i q_i}$$

Dans cette expression, E est une constante et D est une variable aléatoire binomiale que l'on peut approcher par une loi de Poisson.

4.5 Analyse des résultats

Dans cette partie, nous présenterons les résultats de 3 parties : construction de taux brute, lissage, extrapolation et validation. Ici, pour l'estimation des taux bruts suite à la présentation dans le partie précédant des différents avantages et inconvénients des modèles existants, nous avons choisi le modèle Kaplan Meier.

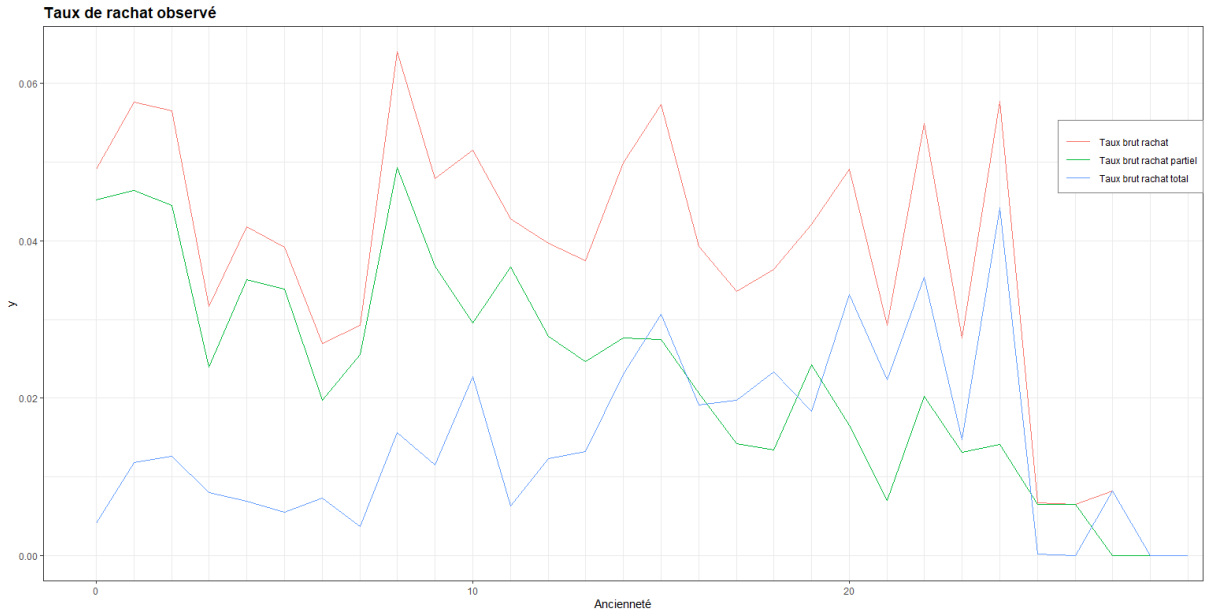


FIGURE 4.2 – Taux brut estimé par Kaplan Meier

Dans le graphique ci-dessus, si nous nous concentrons sur la partie des 10 premières années, nous pouvons facilement voir une tendance en forme de W avec deux pics en troisième année et huitième année pour le taux rachat partiel et le taux de rachat en général. Nous constatons également une tendance très intéressante : le taux de rachat en général est plus affecté par le taux de rachat partiel au cours des 10 premières années. En revanche, il est plus impacté par le taux de rachat total à partir de 10 ans.

Avec ce que nous avons observé, le taux de rachat global par l'estimation de Kaplan Meier est assez fiable dans les premières années, mais il devient moins fiable pour les 25 ans et

plus car les données disponibles sont rares. De plus, il y a des adhérents qui sont toujours dans le portefeuille (problème de censure). Donc, on ne sait pas l'ancienneté maximal que le portefeuille peut atteindre.

Le taux de rachat partiel tend à baisser à partir de la dixième année. Ceci s'explique par 2 raisons, soit la valeur de la PM a été remboursée au client, soit l'adhérent décide après un certain temps de racheter le total pour recevoir la PM restant. En revanche, le taux de rachat total a tendance à augmenter.

Ici, en utilisant le lissage non-paramétrique, on construit trois courbes de taux de rachat avec des paramètres différents. λ est le paramètre de lissage. Une valeur λ plus faible mettra davantage l'accent sur la reproduction de l'observation aussi bonne que possible au prix d'une moins grande régularité. À son tour, une valeur plus élevée de λ forcera le résultat lissé à être aussi lisse que possible avec un écart éventuellement plus important par rapport aux données d'entrée. d est l'ordre des différences.

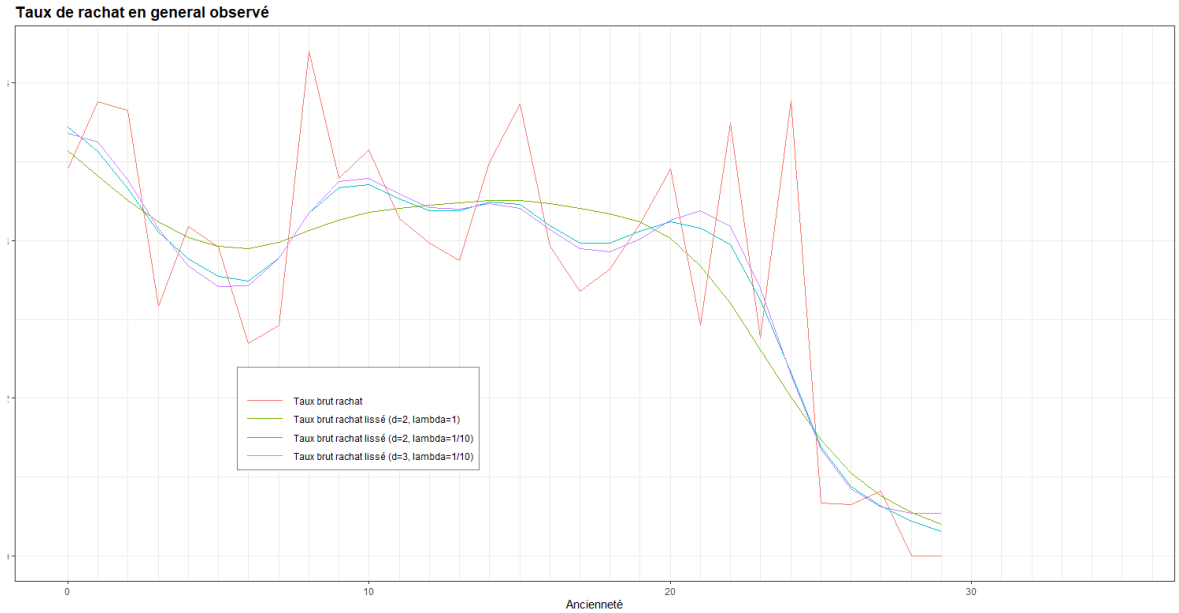


FIGURE 4.3 – Taux de rachat en général lissé

En comparaison les 3 courbes : bleue, verte et violette, on voit que la verte est la plus lissée. Cependant, la ligne verte perd de nombreuses caractéristiques de la ligne de taux brut (rouge). Concrètement, la ligne rouge représente une augmentation du taux de rachat en l'an 8 tandis que la courbe vert sous-estime cette tendance. Les lignes bleues et violettes respectent la plupart des propriétés des taux bruts.

Nous exécutons d'abord le test de la table lissé. Le test du Khi-deux permet de savoir s'il y a correspondance entre la théorie et la répartition observée. Le test du Khi-deux permet donc de voir si un échantillon est conforme à la théorie ou s'il en diffère significativement. Rappelons que H_0 : l'observation ne diffère pas de la théorie. Une p-value $> 10\%$ signifie

qu'on ne peut pas rejeter H_0. En d'autre terme, si la p-value $> 10\%$, on accepte le courbe lissée.

Test de Khi^2 de Pearson	
X^2	754
df	728
p-value	0.2448

TABLE 4.1 – Résultat de Khi test

Pour GPM, il est important de faire la distinction entre le rachat total et partiel. Par conséquent, dans la partie suivante, nous utiliserons la même technique pour obtenir le taux de rachat de la courbe pour chaque type de rachat.

Dans cette partie, nous voulons comparer la courbe construite et la courbe précédemment utilisée par GPM. Cela nous donne un simple test de sensibilité du modèle. En effet, la courbe de rachat de GPM utilisé en 2019 est construite à partir d'une estimation par la moyenne des données historiques. L'avantage de cette méthode est sa simplicité. Cependant, il reste beaucoup d'inconvénients. L'un d'eux est le fait qu'ils ignorent que plus le portefeuille vieillit, plus il a de chances de rachat. De plus, cette approche n'est pas appropriée pour des données censurées.

Dans la figure ci-dessous, nous comparons la courbe de rachat total de GPM en 2019 et la courbe obtenue par Kaplan-Mier.

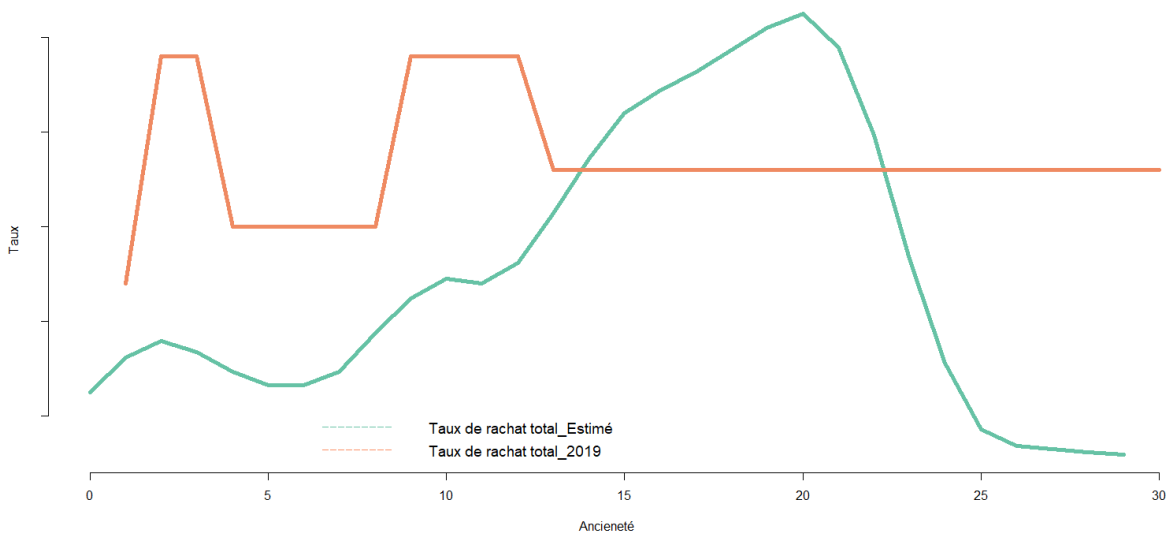


FIGURE 4.4 – Comparaison des taux de rachat total

Une caractéristique de la courbe de rachat de 2019 (celle en orange) est qu'elle crée une

forme en M pour la période de 0 à 12 ans d'ancienneté. Cette tendance est conservée dans la courbe de rachat estimée par Kaplan-Meier (celle en vert). Cependant, dans cette période, la courbe orange est toujours au dessus du vert ce qui signifie que le taux de rachat 2019 a surestimé le risque. Le taux de rachat total augmente rapidement après 10 ans d'ancienneté et commence à baisser en 20 ans d'ancienneté. Cette tendance n'est pas reflétée par l'ancienne courbe car nous avons le même niveau de rachat de 14 ans à 30 ans d'ancienneté.

4.6 Clustering avec un algorithme non-supervisé

Dans l'étude de la méthode Kaplan Meier, nous avons fait l'hypothèse que la population est homogène. L'hétérogénéité est la conséquence d'un mélange de sous-groupes caractérisées chacune par des variables observables. Si la population est en fait hétérogène, il faut tenir compte des caractéristiques de chaque sous-groupe. Planchet & Thérond (2006)

Une des façons de décider d'un sous-groupe est basée sur l'expérience. Par exemple, dans la construction de table de mortalité utilisant également la méthode de Kaplan Meier, nous divisons la population en deux groupes d'hommes et de femmes. Ceci est basé sur des études démographiques d'espérance de la vie. C'est logique et naturel. Cependant, pour le modèle de loi de rachat, le regroupement est basé sur des variables explicatives : l'âge d'adhérents, la durée du contrat, etc. La plupart de ces variables sont numériques. La décision de sous-groupe devient plus compliquée. Dans son article, Planchet & Thérond (2006) introduit également des méthodes pour intégrer des variables explicatives. Nous pouvons en citer quelques-uns : modèle additif d'Aalen et modèle de Lin et Ying.

Dans cette étude, nous utiliserons la méthode de *clustering* pour déterminer la division de chaque sous-groupe. Après avoir créé différents sous-groupes, nous utiliserons le test du log-rank pour vérifier la différence entre les courbes des taux de rachat pour chaque groupe.

Le « *Clustering* » (classification automatique) consiste à regrouper un ensemble d'objets de telle sorte que les objets du même groupe soient le plus similaires (dans un certain sens) les uns aux autres qu'à ceux des autres groupes (*clusters*). Il s'agit d'une tâche important dans l'exploration de données et d'une technique courante d'analyse de données statistiques, utilisée dans de nombreux domaines, notamment la reconnaissance de formes, l'analyse d'images, la recherche d'informations, la bio-informatique, la compression de données, l'infographie et l'apprentissage automatique.

Webster (Merriam-Webster Online Dictionary, 2008) définit « *clustering* » comme « une technique de classification statistique pour découvrir si les individus d'une population appartiennent à différents groupes en effectuant des comparaisons quantitatives à caractères multiples. En autre terme, l'objectif est d'identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement des clients) On veut que :

- Les individus dans un même groupe se ressemblent le plus possible
- Les individus dans des groupes différents se démarquent le plus possible

Mais quelle est la notion de similitude ? En fait, il est parfois compliqué de le déterminer. Alors que les humains peuvent facilement détecter un « *cluster* » dans deux et peut-être trois dimensions, nous avons besoin d'algorithmes automatiques pour les données de haute dimension. C'est ce défi ainsi que le nombre inconnu de *clusters* pour les données qui ont abouti à des milliers d'algorithmes de *clustering* qui ont été publiés et qui continuent d'apparaître.

4.6.1 L'évaluation de la tendance au clustering

Avant d'appliquer une méthode de *clustering* sur des données, une question simple est de savoir si les ensembles de données contiennent des *clusters* significatifs (c'est-à-dire des structures non aléatoires). En d'autre terme, nous voulons voir si nos données sont vraiment hétérogènes. Si oui, combien de sous-groupe y a-t-il ?. Ce processus est défini comme l'évaluation de la tendance au *clustering* ou la faisabilité de l'analyse de *clustering*.

Cependant, un gros problème dans l'analyse de *cluster*, est que les méthodes de *clustering* renverront des *clusters* même si les données ne contiennent aucun *cluster*. De plus, le *cluster* dépend beaucoup des données existantes mais échoue facilement à se séparer de la même manière ou de la même logique avec les nouvelles données.

Dans cette partie, nous présentons une méthode pour détecter l'hétérogénéité de la population : critère statistique d'Hopkins selon l'étude de Lawson & Jurs (1990) :

$$H_0 : \text{Pas de } \textit{clustering} \text{ significatif}$$

$$H_1 : \text{L'ensemble de données contient un } \textit{clustering} \text{ significatif}$$

Soit D l'ensemble de données. La statistique d'Hopkins est calculée en suivant ces étapes :

- Prendre un échantillon au hasard n points de D $p_1, p_2, \ldots, p_n$
- Calculer la distance $x_i = distance(p_i, p_j)$ où p_j le plus proche voisin de p_i
- Simuler un ensemble de donnée de n points de $q_1, q_2, \ldots, q_n$ qui suit une distribution uniforme
- Calculer la distance $y_i = distance(q_i, q_j)$ où q_j le plus proche voisin de q_i

L'hypothèse est redéfinie comme :

$$H_0 : \text{D suit une loi uniforme}$$

$$H_1 : \text{D ne suit pas une loi uniforme}$$

La statistique d'Hopkins est :

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Si D a une distribution d'uniforme, $\sum_{i=1}^{n} y_i$ et $\sum_{i=1}^{n} x_i$ seront très poches. Donc, H sera d'environ 0,5. Au contraire, s'il y a un *cluster*, la distance des points de D sera plus basse que l'ensemble des données simulées, donc H augmentera. Une valeur de H supérieur à 0,75 est considérée comme une tendance de *clustering*.

4.6.2 La méthode K-Means

La méthode des K-Means (méthode des centres mobiles) est une technique de classification automatique. Elle vise à produire un regroupement de manière que les individus du même groupe soient semblables et les individus de groupes différents soient dissemblables.

L'algorithme des K-means a été publié pour la première fois en 1955. Malgré le fait que la méthode a été proposé il y a plus de 50 ans et que des milliers d'algorithmes de *clustering* ont été publiés depuis lors, la K-means est encore largement utilisé.

La méthode dite des k-means (k-moyennes) introduite par MacQueen et al. (1967) commence effectivement par un tirage pseudo-aléatoire de centres ponctuels. Cependant la règle de calcul des nouveaux centres n'est pas la même. On n'attend pas d'avoir procédé à la réaffectation de tous les individus pour modifier la position des centres : chaque réaffectation d'individus entraîne une modification de la position du centre correspondant. En une seule itération, cette procédure peut ainsi donner une partition de bonne qualité. Mais celle-ci dépendra de l'ordre des individus sur le fichier, ce qui n'est pas le cas pour la technique exposée précédemment 1.

- Méthodologies et paramètres de modèle

Soit $X = \{x_i\}$ avec $i = 1, ..., n$ l'ensemble des n points de d-dimension être regroupés en un ensemble de K *clusters*, $C = \{c_k, k = 1, ..., K\}$. L'algorithme K-means trouve une partition telle que l'erreur quadratique entre la moyenne empirique d'un *cluster* et les points du *cluster* est minimisée. Soit μ_k l'espérance de *cluster* c_k. L'erreur quadratique entre μ_k et c_k est défini par :

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2$$

Le but de K-means est de minimiser la somme de l'erreur quadratique sur tous les K clusters :

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2$$

Les principales étapes de l'algorithme sont (Jain & Dubes (1988)) :

- Sélection au hasard de K point de c_k
- Création une nouvelle partition en affectant chaque modèle à son centre de *cluster* le plus proche. La définition de la distance diffère entre les études mais il s'agit souvent de la distance Euclidienne
- Calcul du nouveau c_k
- Répétitions les étapes 2 et 3 jusqu'à ce qu'aucun point ne change son *cluster*

Il y a 3 paramètres dans le modèle de K-means : la valeur de K, le choix initial des centres de classes et la définition de la distance. Entre eux, le choix de la valeur de K est le plus compliqué (Jain (2010)). La méthode utilisée pour déterminer K sera discutée dans la partie suivante de ce mémoire où nous appliquons le modèle K-means à des données réelles.

4.6.3 Application

- Évaluation de la tendance de *clustering*

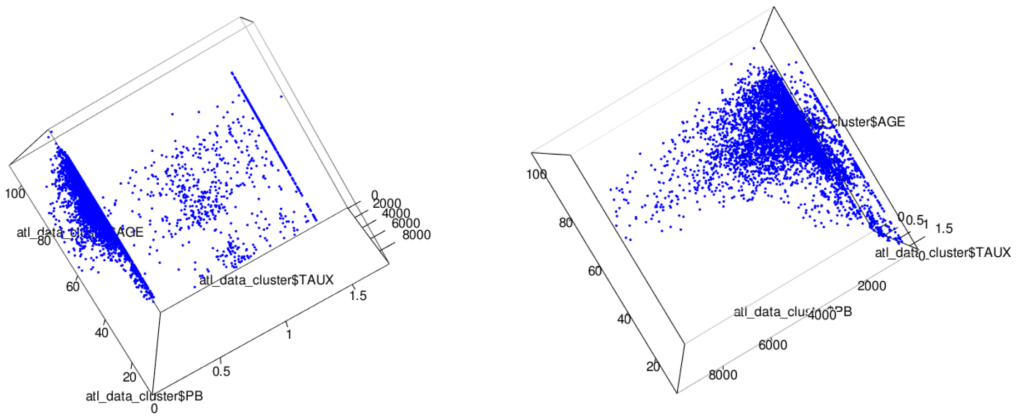


FIGURE 4.5 – Vision 3D de la tendance de la clustering

Nous avons pris les statistiques d'Hopkins sur l'ensemble de données GPMA. Grâce à un échantillonnage aléatoire, nous prélevons un échantillon de 3 000 observations. Les résultats montrent que la valeur des statistiques d'Hopkins est 0.0098. Notez que la fonction utilisée dans R est la fonction **hopkins** dans le package **clustertend**. Pour cette fonction, le résultat est 1-H. Pour cette raison, la valeur des statistiques d'Hopkins est inférieure à 0,5, nous avons donc une indication de *clustering* dans notre ensemble de données.

- Choix du paramètre K

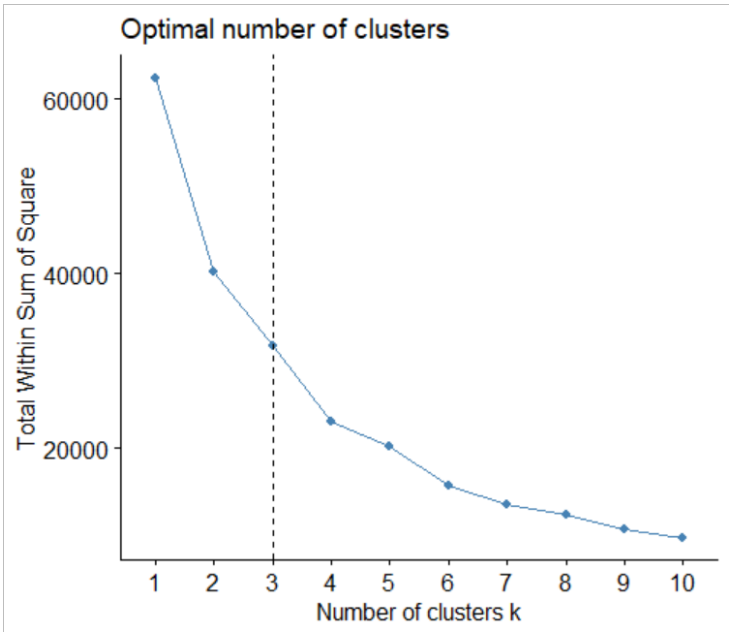


FIGURE 4.6 – Choix d'optimum nombre de clusters

Comme dans le graphique, nous pouvons voir que le total dans la somme des carrés diminue rapidement de K = 2 à K =4 puis ralentit pour les k supérieur à ce seuil. Nous appliquons ensuite la méthode des k moyennes pour chaque valeur de K. Pour voir les résultats plus clairement, nous nous appuyons sur le graphique suivant :

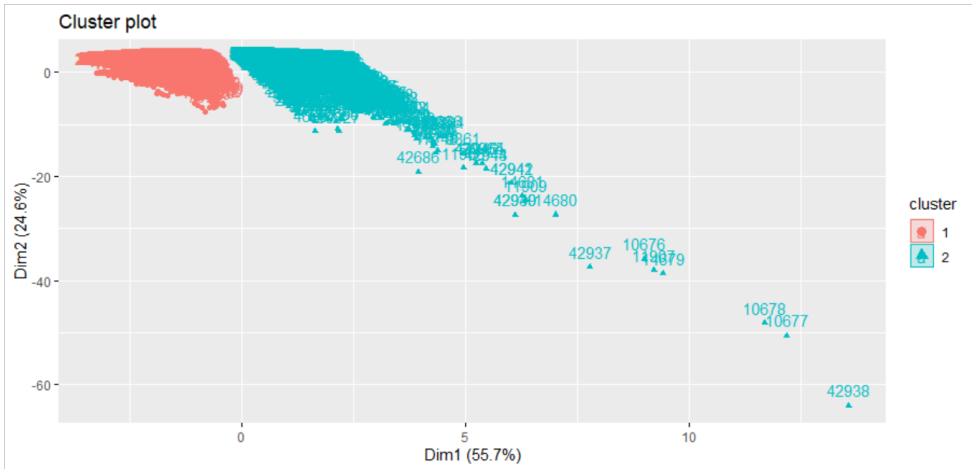


FIGURE 4.7 – Plot de cluster avec K = 2

Les valeurs aberrantes affectent gravement le résultat du modèle. Deux regroupements (K = 2) sont raisonnables dans ce cas, ce qui simplifie également l'application au modèle de Kaplan Meier.

Ensuite, nous construirons les courbes de taux de rachat pour chaque regroupement

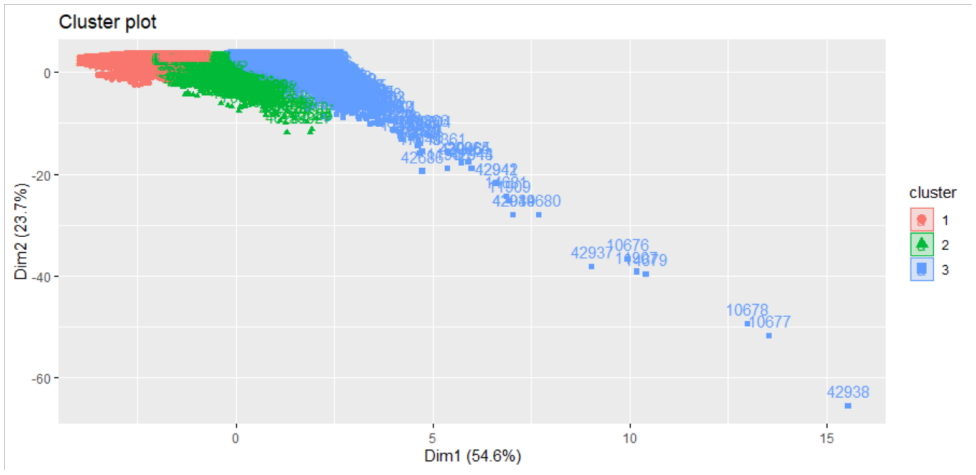


FIGURE 4.8 – Plot de cluster avec K = 3

de groupes homogènes. Pour finir, nous utiliserons critère log-rank pour comparer la différence entre ces taux de rachats et leur dissemblance.

4.6.4 Avantages et inconvénients

K-means est un algorithme de *clustering* assez simple dans l'interprétation des résultat par rapport à d'autre méthode de *clustering*. Il produit des résultats de *clustering* rapide. Grâce à cette fonctionnalité, nous pouvons l'utiliser pour l'appliquer au modèle Kaplan Meier.

K-means est une méthode très efficace pour le *clustering*. Cependant, ce modèle a aussi ses faiblesses. Le temps de traitement des données avec cet algorithme est encore très long car nous devons faire le calcul à chaque observation plusieurs fois. Pour nos ensembles de données volumineux, cela peut être un défi. L'algorithme trouve seulement l'optimal local, même si dans le cas où on peut facilement trouver l'optimum global en observant le graphique. Toujours, K-means converge à l'optimum global (Meilă (2006)).

Dans le même temps, la sélection des premiers points de l'algorithme et l'ordre des points peut également modifier les résultats. Pour résoudre ce problème, nous allons faire des K-means plusieurs fois et faire attention au traitement des valeur aberrants qui fausserait les centres des clusters. Cette méthode ne convient que lorsque la valeur K est petite.

Un autre inconvénient de ce modèle est qu'il ne peut pas garantir l'efficacité du *clustering* lorsqu'il est utilisé comme pré-traitement pour l'estimation de Kaplan meier car il n'est pas conçu pour une analyse de survie configurée. Cependant, avec ce modèle simple de *clustering*, nous avons réfléchi à l'idée de diviser le portefeuille en différents sous-groupes. Dans la partie suivante, nous examinerons un modèle conçu pour résoudre le problème de l'analyse de survie avec la même idée de segmenter le portefeuille.

Chapitre 5

Tree-based survival modeles

Pour prendre en compte différentes variables explicatives dans les modèles de survie, on peut proposer différentes méthodes. Un modèles d'analyse de survie courant est le modèle de Cox (Cox (1972)), qui modélise le taux de risque d'un événement avec une combinaison linéaire d'effets de covariables. Bien que ce modèle soit largement utilisé et facilement interprétable, sa nature paramétrique le rend inefficace d'identifier les effets non linéaires ou les interactions entre les covariables (Bou-Hamad et al. (2011)). Comme nous l'avons appris dans la partie précédente, la relation entre la variable cible et les variables indépendantes est non linéaire. Les modèles qui sont proposés pour remplacer la régression de Cox sont : le « modèle du Cox dépendant du temps » et l'apprentissage automatique. Dans cette section, nous étudierons un modèle de machine learning intéressant pour traiter le problème d'analyse de survie : arbre de survie et forêt aléatoire de survie.

5.1 Méthodologies

5.1.1 Le modèle de Cox

Dans cette section, nous entrerons brièvement dans les détails du modèle de régression de Cox et pourquoi ce modèle est généralement mentionné dans l'étude pour l'analyse de survie.

Le modèle de régression proposé par Cox est un modèle de régression multiple pour l'analyse des données de survie censurées. Il est utilisé pour étudier le modèle de nombreuses variables avec la fonction de risque. Le modèle est défini avec la fonction de hasard et les paramètres suivants :

$$\lambda(t, z) = \lambda_0(t)exp(b_1 z_1 + ... + b_i z_i + ... + b_p z_p)$$

Pour un individu ayant un ensemble de variables z où $z = (z_1, z_2, ...)$, la fonction de hasard au temps t est définie comme $\lambda(t, z)$. L'objectif est d'estimer les coefficients $b = (b_1, b_2, ...)$. Cette fonction dépend du risque de base $\lambda_0(t)$.

Le modèle de régression de Cox est un modèle semi-paramétrique. Nous ne voulons pas estimer le λ_0. C'est, en fait, le même pour tous les individus. Nous sommes plus préoccupés par le rapport des risques instantanés de décès pour deux individus exposés à des facteurs de risques différents. C'est aussi un modèle à risque proportionnel car le rapport est donc indépendant du temps.

Comme mentionné, le modèle est basé sur l'hypothèse forte qui doit être vérifiée avant d'estimer les coefficients. Avec chaque covariable, nous testons si son effet est dépendant du temps ou non. Le test habituellement utilisé est le test des résidus de Schoenfeld.

Grâce à la méthode du maximum de vraisemblance, nous pourrons modéliser les différents coefficients β_k avec $k = 1, 2, ...$ Dans cette section, la régression de Cox n'est pas le cœur de notre étude, nous n'irons donc pas plus loin, les détails du modèle seront présentés en annexe.

Les méthodes paramétriques et semi-paramétriques (notamment, le modèle de risques proportionnels de Cox) qui permettent d'associer le temps de survie aux variables explicatives. Cependant, de tels modèles nécessitent de pré-spécifier la forme fonctionnelle des variables explicatives, y compris leurs interactions.

De plus, la puissance d'une analyse de Cox dépend largement du nombre d'événements dans l'ensemble de données (mentionné par Christensen (1987)). Harrell (1983) recommande que le nombre de variables prédictives examinées ne dépasse pas environ 5 à 10% du nombre d'événements, pour garantir l'exactitude du modèle. Le choix des variables est également très important dans ce modèle.

5.1.2 Survival tree

Les arbres de survie offrent une approche relativement flexible lorsque la forme des effets des variables explicatives est inconnue et ont également une plus grande capacité à détecter automatiquement les interactions en fonction des données observées. Ils sont l'alternative non paramétrique des modèles (semi-) paramétriques et présentent également l'avantage d'une interprétation plus facile.

La plupart des techniques de construction d'arbres de survie actuellement disponibles ne sont pas basées sur un test formel de signification et peuvent donc être sujettes à des découvertes fallacieuses sur l'effet des variables explicatives. Dans cet article, nous avons proposé un algorithme de partitionnement récursif pour construire un arbre de survie qui sélectionne la variable de division via un test statistique formel. Contrairement

aux algorithmes existants qui se concentrent uniquement sur l'hétérogénéité d'événement, l'algorithme proposé fournit un cadre pour identifier les sous-groupes en fonction de l'hétérogénéité du temps des événements et / ou de la censure.

Les techniques de partitionnement récursif (également appelées arbres) sont une alternative populaire aux modèles paramétriques. Lorsqu'ils sont appliqués aux données de survie, les algorithmes d'arbre de survie partitionnent l'espace des covariables en régions de plus en plus petit (nœuds) contenant des observations avec des résultats de survie homogènes. La distribution de survie dans les partitions finales (feuilles) peut être analysée à l'aide de diverses techniques statistiques telles que les estimations de la courbe de Kaplan-Meier (Kaplan & Meier (1958)).

Il existe deux règles partagées qui peuvent être utilisées pour faire pousser des arbres de survie. Dans ce modèle, la réponse ou le résultat y associé à l'individu i est une paire de valeurs spécifiant un temps de survie non négatif et censurant les informations. Notons la réponse pour i par $Y_i = (T_i, \delta_i)$. On dit que l'individu est censuré à droite au temps T_i si $\delta_i = 0$ et est décédé au temps T_i si $\delta_i = 1$. En cas de décès, T_i sera désigné comme un événement et le décès comme un événement. Un individu i qui est censuré à droite à T_i signifie simplement que l'individu est connu pour avoir été vivant à T_i, mais l'heure exacte du décès est inconnue.

Comme mentionné précédemment, l'une des tâches les plus importantes lors de la construction d'un arbre de décision est de déterminer une règle de division. Dans cette étude, nous introduisons la méthode des différences et des articles autour de cette problématique. Et après cela nous décrirons en détail la règle de fractionnement « log-likelihood » qui est construite en R.

L'idée d'un arbre de survie est introduite par Gordon & Olshen (1985). Il vise à rechercher l'homogénéité de chaque nœud par la métrique de Wasserstein. Bien que ce critère ne soit pas couramment utilisé, il mentionne également la possibilité d'utiliser des statistiques log-rank et une statistique de rapport de vraisemblance paramétrique qui, plus tard, est devenue le fondement de nombreuses recherches. Le test du log-rank pour la division des arbres de survie est un concept bien établi Segal (1988). Il s'est avéré robuste dans les contextes de risque proportionnel et non proportionnel LeBlanc & Crowley (1993). Hothorn et al. (2006) ont implémenté un arbre de survie sans biais en utilisant le test du log-rank comme méthode de fractionnement.

En se basant sur le principe du risque proportionnel, LeBlanc & Crowley (1992) proposent un critère de fractionnement basé sur une mesure de déviance de nœud entre un modèle saturé log – vraisemblance et un log – vraisemblance maximisé. Avec cette méthode, la log – vraisemblance est approximée en remplaçant la fonction de risque cumulatif de base par l'estimateur de Nelson – Aalen. Cette méthode est implémentée dans le package rpart de R (Therneau et al. (2010)).

Cette règle de fractionnement est dérivée d'un modèle à risques proportionnels qui suppose que la distribution de survie sous-jacente pour chaque observation est donnée par :

$$\mathbb{P}(S_i \leq t) = 1 - \exp^{(-\theta_i \Lambda(t))}$$

où $\Lambda(t)$ est la fonction de risque cumulé de base et les coefficients θ_i sont les ajustements au risque cumulé de base pour chaque observation. Dans un modèle d'arbre de survie, nous remplaçons $\Lambda(t)$ avec une estimation empirique de la probabilité cumulée de décès à chacun des moments d'observation qui est l'estimateur de Nelson-Aalen.

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i}$$

L'objectif d'un modèle d'arbre de survie est d'optimiser les coefficients de risque θ_i en maximisant la vraisemblance de la feuille.

5.1.3 Survival random forest

Notez que cette règle de fractionnement est différente de celle de l'arbre de survie. Cependant, l'estimateur du nœud terminal est gardé comme l'estimateur de Kaplan-Meier. Dans cette section, nous avons introduit les règles de fractionnement du log-rank et du log-rank score qui sont bien pris en charge par le package randomForestSRC. Nous rappelons que $d_{i,j}$ et $r_{i,j}$ respectivement sont le nombre de décès et d'individus à risque à t_i dans nœud j.

– Critères de Log-rank

Le test du log-rank pour un fractionnement à la valeur c pour le prédicteur x est le suivant :

$$L(x,c) = \frac{\sum_{i=1}^{N}(d_{i,1} - r_{i,1}\frac{d_i}{r_i})}{\sqrt{\sum_{i=1}^{N}\frac{r_{i,1}}{r_i}(1-\frac{r_{i,1}}{r_i})(\frac{r_i - d_i}{r_i - 1})d_i}}$$

La valeur $|L(x,c)|$ est la mesure de la séparation des nœuds. Une grande valeur de $L(x,c)$ signifie que les différences entre deux groupes sont importantes et donc, la répartition est meilleure. Par conséquent, le meilleur fractionnement est déterminé en trouvant le prédicteur x^* et la valeur de fractionnement c^* tels qu'ils maximisent $|L(x,c)|$.

– Critères du Log-rank score

Supposons que le prédicteur x a été ordonné de telle sorte que $x_1 \leq x_2 \leq ... \leq x_n$. On calcule les « rangs » pour chaque temps de survie T_l :

$$a_l = \delta_l - \sum_{k=1}^{\Gamma_l} \frac{\delta_k}{n - \Gamma_k + 1}$$

où Γ_k . Le test de log-rank est défini par :

$$S(x,c) = \frac{\sum_{x_k \leq c}(a_j - n_l \bar{a})}{\sqrt{n_l(1-\frac{n_l}{n})s_a^2}}$$

où $\bar{a}$ et s_a^2 sont respectivement le moyenne de l'échantillon et la variance de l'échantillon de a_j. La meilleure division correspond aux x et c qui maximisent $S(x,c)$.

– Estimations des nœuds terminaux

A chaque nœud terminal, nous déterminerons ensuite une estimation de Kaplan Meier. Comme nous avons discuté de cette estimation dans les parties précédentes, nous n'aurons ici qu'un aperçu très rapide de la façon dont l'OOB est calculé pour chaque individu i :

$$S(x,c) = \frac{\sum_{x_k \leq c}(a_j - n_l \bar{a})}{\sqrt{n_l(1-\frac{n_l}{n})s_a^2}}$$

5.2 Résultat

En utilisant l'arbre de survie, le portefeuille est divisé en 4 groupes en raison de 2 variables : l'âge de l'adhérent et son participation aux bénéfices. A noter que nos arbres ont élagage afin d'éviter le « *over-fitting* ».

Le premier groupe représente les personnes dont l'âge est supérieur à 69 ans et qui ont perçu plus de 0,005 euro de participation aux bénéfices en 2019. Il s'agit du groupe d'adhérents qui a déjà pris sa retraite. C'est pourquoi leur fonction de survie dans 20-ème année d'ancienneté est plus élevée par rapport aux autres groupes. En d'autres termes, ils sont moins susceptibles d'être rachetés mais plus susceptibles d'être censurés (pour cause de décès).

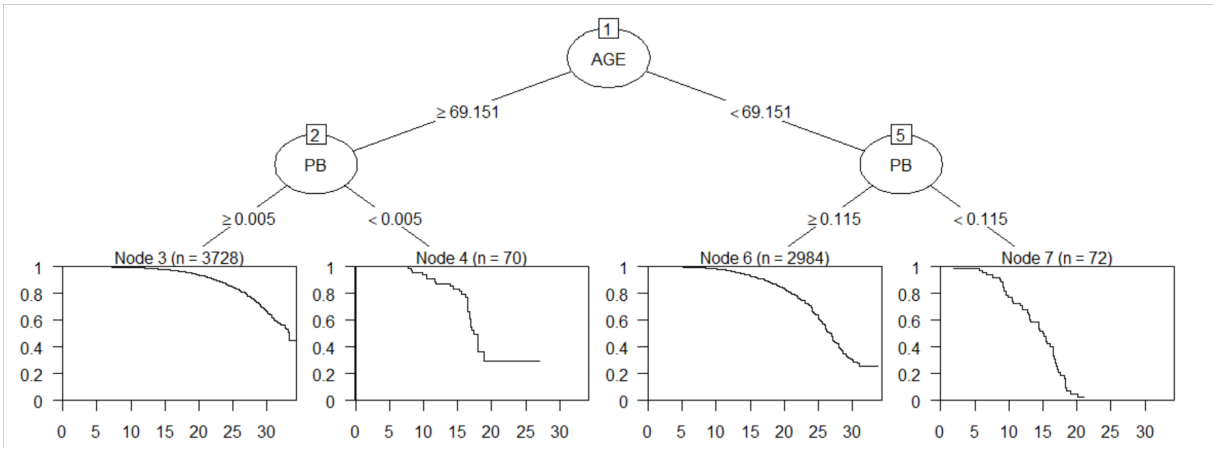


FIGURE 5.1 – Résultat de Survival tree

Notez que la séparation de la variable PB ici avec des seuils de 0,005 et 0,115 ne représente que le fait que le compte d'épargne reçoit toujours de la PB ou non. Pour prouver ce fait, nous examinons la distribution de la PB. Ici, on le voit, 20% du portefeuille a une PB inférieure à 1 euro. Mais parmi les personnes qui ont de la PB de plus d'1 euro, leur répartition est diverse gravement.

Le nœud 6 qui se compose d'adhérents de moins de 69 ans et ayant un PB supérieur à 0,115 semble être le groupe typique où la fonction de survie se comporte normalement. La fonction de survie diminue rapidement après la 25ème année de contrat. Il y a 2984 observations qui appartiennent à ce groupe. Les nœuds 4 et 7 ont une queue distribution bizarre qui se termine avec 20 ans d'ancienneté.

Sample size	4 402
Number of deaths	524
Number of tree	100
Forest terminal node	15
(OOB) CRPS	0.05105744
(OOB) Requested performance error	0.13957391

TABLE 5.1 – Résultat de Survival Random Forest

Ici pour la forêt aléatoire, sa plus grande faiblesse est le fait qu'elle ne nous permet pas d'exporter une courbe de rachat implémentable. Dans cette section, la capacité de mise en œuvre et le pouvoir de prédiction sont tout aussi importants, par conséquent, le modèle de forêt aléatoire ne sera pas utilisé.

En conclusion, cette partie nous a donné une idée nouvelle pour traiter le problème de survie. L'arbre de survie nous laisse plus de choix pour définir la distribution de la variable. Cela nous donne également plus de liberté dans le choix de la relation entre les variables indépendantes et la variable cible.

Le modèle a montré beaucoup d'avantages cependant, il reste des faiblesse. Premièrement, le résultat final peut être beaucoup trop complexe pour être intégré dans modèle d'ALM de l'entreprise. Bien qu'après le processus d'élagage, le nombre de nœud final ne soit que de 4, cela pose un problème à l'entreprise où nous ne pouvons pas accepter que 1 ou 2 courbes.

Deuxièmement, le modèle est faible lorsqu'il a de nouvelles données qui arrivent. Nous l'avons mentionné dans la partie sur les modèle utilisant les arbres. En général, si la valeur d'une variable indépendante n'a pas existé dans le modèle, le modèle risque de ne pas la catégoriser.

Conclusion

L'étude réalisée s'inscrit dans le cadre de la maîtrise du risque de rachat. Dans ce but, nous avons effectué des analyses pour une meilleure connaissance de notre portefeuille et du risque de rachat.

Dans la première section, nous avons fait une synthèse des différentes terminologies concernant l'assurance-vie et le risque de rachat. Nous avons récapitulé ensuite l'environnement solvabilité 2 pour savoir où le rachat intervient dans les fonds propres de l'entreprise. Cela prouve l'importance de ce risque pour la compagnie d'assurance.

Une remarque à travers cette analyse qualitative est que l'option de rachat entraîne un avantage fiscal pour les adhérents qui conservent leur compte jusqu'à la 8-ème année de contrat. Ces caractéristiques affectent le comportement des adhérents.

La deuxième partie a contribué à une meilleure compréhension du portefeuille d'un point de vue statistique qui est une étape fondamentale avant la construction de modèles quantitatifs. Les études descriptives univariées et bivariées ont montré quelques caractéristiques importantes du portefeuille.

L'espérance de vie du portefeuille est beaucoup plus élevée que l'espérance normale. Ils semblent garder leur compte d'épargne très longtemps et s'avèrent peu sensibles aux marché financiers. Les variables qui pourraient influer sur la décision de rachat sont : l'âge de l'adhérent, la provision mathématique en début d'année et le montant de la participation au bénéfice de l'année précédente. L'ensemble de données a également besoin d'un traitement des valeurs aberrantes et de la discrétisation.

La troisième partie est celle où nous avons commencé à construire le modèle. Notre premier point de vue a été d'utiliser le modèle statique. Cela nous a permis de mieux comprendre notre portefeuille. Nous avons proposé 3 modèles : la régression logistique, les arbres de décision et des forêts aléatoire. La forêt aléatoire est le modèle où les résultats les plus robustes. L'AUC est de 89,95%. Cela pourrait signifier que si nous devons prédire les risques de rachat d'une personne l'année prochaine, le modèle est raisonnable pour cette utilisation. Cependant, il faut aller au-delà de ce modèle car le modèle ALM de GPM est construit sur la courbe de rachat en fonction d'ancienneté des adhérents. C'est ce que nous faisons dans la quatrième partie.

Dans cette partie, nous avons commencé par donner la compréhension nécessaire à l'analyse de survie en récapitulant les 2 estimateurs les plus utilisés : l'estimateur de Kaplan Meier et de Nelson-Aalen. Nous avons ensuite construit la courbe de rachat Kaplan Meier par ancienneté et l'avons comparée à la courbe actuellement utilisée par GPM. Cette courbe s'avère acceptable pour remplacer l'ancienne courbe de GPM. Nous voulions aller un peu plus loin dans l'analyse car la question de savoir si : « l'hypothèse d'homogénéité est satisfaite dans notre portefeuille » se pose. Nous avons prouvé ensuite que le portefeuille est hétérogène à travers l'analyse et le test de Kmeans.

Dans la dernière partie, nous avons donné une méthode alternative pour résoudre la question autre que le modèle de Cox. Nous avons introduit un modèle d'arbre et de forêt aléatoire appliqué à l'analyse de survie. Ces modèles satisfont aux contraintes d'homogénéité et aux caractéristiques de l'analyse de survie. Cependant, comme le modèle est nouveau, il est difficile de le maîtriser complètement pour être utilisé dans la réalité. Pourtant, cela donne la nouvelle idée d'un développement ultérieur à l'avenir.

Pour conclure, l'analyse de survie du modèle univariée s'est avérée robuste et pertinente pour une utilisation par GPM. Nous avons développé des modèles pour estimer le risque de rachat aux des analyses utilisant des modèles dépendants de plusieurs variables. Cependant, pour garder le modèle simple et l'intégrer facilement dans le modèle ALM de compagnie, nous utilisons le modèle univariée.

Bibliographie

Albizzati, M.-O. & Geman, H. (1994), 'Interest rate risk management and valuation of the surrender option in life insurance policies', *Journal of Risk and Insurance* pp. 616–637.

Bacinello, A. R. (2003), 'Pricing guaranteed life insurance participating policies with annual premiums and surrender option', *North American Actuarial Journal* **7**(3), 1–17.

Bacinello, A. R. (2005), 'Endogenous model of surrender conditions in equity-linked life insurance', *Insurance : Mathematics and Economics* **37**(2), 270–296.

Bou-Hamad, I., Larocque, D. & Ben-Ameur, H. (2011), 'A review of survival trees', *Statistics surveys* **5**, 44–71.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), 'Classification and regression trees–crc press', *Boca Raton, Florida* .

Burkhart, T. (2018), 'Surrender risk in the context of the quantitative assessment of participating life insurance contracts under solvency ii', *Risks* **6**(3), 66.

Charpentier, A. & Denuit, M. (2005), 'Mathématiques de l'assurance non-vie', *Economica, Paris* .

Christensen, E. (1987), 'Multivariate survival analysis using cox's regression model', *Hepatology* **7**(6), 1346–1358.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society : Series B (Methodological)* **34**(2), 187–202.

Friedman, J. H. (2001), 'Greedy function approximation : a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Gordon, L. & Olshen, R. A. (1985), 'Tree-structured survival analysis.', *Cancer treatment reports* **69**(10), 1065–1069.

Grosen, A. & Jørgensen, P. L. (2000), 'Fair valuation of life insurance liabilities : the impact of interest rate guarantees, surrender options, and bonus policies', *Insurance : Mathematics and Economics* **26**(1), 37–57.

Harrell, F. E. (1983), 'The phglm procedure', *Supplemental Library User's Guide* pp. 267–294.

Henderson, R. (1924), 'A new method of graduation', *Transactions of the Actuarial Society of America* **25**, 29–40.

Ho, T. K. (1998), 'The random subspace method for constructing decision forests', *IEEE transactions on pattern analysis and machine intelligence* **20**(8), 832–844.

Hosmer Jr, D. W. & Lemeshow, S. (1999), 'Applied survival analysis : regression modelling of time to event data (1999)', *Eur Orthodontic Soc* pp. 561–2.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. J. (2006), 'Survival ensembles', *Biostatistics* **7**(3), 355–373.

Jain, A. K. (2010), 'Data clustering : 50 years beyond k-means', *Pattern recognition letters* **31**(8), 651–666.

Jain, A. K. & Dubes, R. C. (1988), *Algorithms for clustering data*, Prentice-Hall, Inc.

JAMAL, S. (2017), 'Lapse risk modeling with machine learning techniques : an application to structurel lapse drivers'.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association* **53**(282), 457–481.

Kuo, W., Tsai, C. & Chen, W.-K. (2003), 'An empirical study on the lapse rate : The cointegration approach', *Journal of Risk and Insurance* **70**(3), 489–508.

Lawson, R. G. & Jurs, P. C. (1990), 'New index for clustering tendency and its application to chemical problems', *Journal of chemical information and computer sciences* **30**(1), 36–41.

LeBlanc, M. & Crowley, J. (1992), 'Relative risk trees for censored survival data', *Biometrics* pp. 411–425.

LeBlanc, M. & Crowley, J. (1993), 'Survival trees by goodness of split', *Journal of the American Statistical Association* **88**(422), 457–467.

Liu, F. T., Ting, K. M. & Zhou, Z.-H. (2008), Isolation forest, *in* '2008 eighth ieee international conference on data mining', IEEE, pp. 413–422.

Loisel, S., Piette, P. & Tsai, C.-H. J. (2021), 'Applying economic measures to lapse risk management with machine learning approaches', *ASTIN Bulletin : The Journal of the IAA* **51**(3), 839–871.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297.

McFadden, D. et al. (1973), 'Conditional logit analysis of qualitative choice behavior'.

Meilă, M. (2006), The uniqueness of a good optimum for k-means, *in* 'Proceedings of the 23rd international conference on Machine learning', pp. 625–632.

Milhaud, X., Gonon, M.-P. & Loisel, S. (2010), 'Les comportements de rachat en assurance vie en régime de croisière et en période de crise', *Risques : les cahiers de l'assurance* **83**, 76–81.

Milhaud, X., Loisel, S. & Maume-Deschamps, V. (2011), 'Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ?', *Bulletin Français d'Actuariat* **11**(22), 5–48.

Mittlböck, M. & Schemper, M. (1996), 'Explained variation for logistic regression', *Statistics in medicine* **15**(19), 1987–1997.

Nelder, J. A. & Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society : Series A (General)* **135**(3), 370–384.

Planchet, F. & Thérond, P. (2006), 'Modèles de durée', *Economica* .

Segal, M. R. (1988), 'Regression trees for censored data', *Biometrics* pp. 35–47.

Therneau, T. M., Atkinson, B. & Ripley, M. B. (2010), 'The rpart package', *R Foundation for Statistical Computing : Oxford, UK* .

Whittaker, E. T. (1922), 'On a new method of graduation', *Proceedings of the Edinburgh Mathematical Society* **41**, 63–75.