INSTITUT INTERNATIONAL DES ASSURANCES

(IIA)

INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

(ISFA)

ANALYSE DES RACHATS INDIVIDUELS EN ASSURANCE

Mémoire présenté et soutenu publiquement en vue de l'obtention du Master en Actuariat

PAR

NOUMSI NIE Raoul Martinien

SOUS L'ENCADREMENT DE :

Dr. Aymric KAMEGA

Actuaire Certifié

Président Directeur Général,

ACAM Vie.

Pr. Louis Aimé FONO

Enseignant-Chercheur en Mathématiques,

Université de Douala.

SEPTEMBRE 2021

Dédicaces

Je dédie ce travail à mes biens aimés parents : Mr et Mme NOUMSI

Savourez en ce travail l'accomplissement de nos efforts communs.

Remerciements

Je remercie en premier toute l'équipe d'Acam vie de m'avoir accueillie au sein de la compagnie dans le cadre de mon apprentissage et de m'avoir proposé ce sujet de mémoire très intéressant. J'aimerais remercier plus particulièrement le Président Directeur Général **Dr. Aymric Kamega** (Actuaire Certifié), pour sa rigueur, sa disponibilité et surtout la justesse des orientations qu'il a bien voulue me suggérer.

Je remercie **Pr Louis Aimé Fono** pour sa disponibilité, ses conseils, son soutien toujours constant face à nos difficultés et sa supervision continue.

Je remercie aussi l'ensemble du corps professoral de l'ISFA et de l'IIA pour la qualité de la formation et pour toutes les connaissances que j'ai acquises tout au long de mon cursus.

Je remercie ma famille : ma mère, qui a toujours ménagé des efforts pour la réussite de mes études, mes frères et sœurs qui m'ont toujours soutenue, conseillée enfin mon épouse et ma fille pour leur accompagnement multiforme.

A tous ceux que je n'ai pas mentionnés dans ce mémoire et qui, de près ou de loin, ont contribué à l'élaboration de ce mémoire.

Voyez en ce mémoire un des fruits de toute l'aide que vous n'avez jamais cessé de m'apporter.

Résumé

Mots clés : comportement de rachat, variables explicatives, rachat structurel, modèle de machine learning.

Au cours de sa vie, un contrat d'épargne peut faire face à de multiples évènements pouvant modifier ses flux financiers futurs probables. C'est le cas notamment lors d'une modification du niveau de la prime ou de la structure de la prime ou lorsqu'il arrive à son échéance ou lors d'un rachat. Ce dernier évènement représente un risque majeur pour l'assureur d'où la nécessité de sa bonne maitrise. C'est dans ce contexte que s'inscrit ce mémoire. L'objectif est d'analyser puis de modéliser les comportements de rachats structurels des assurés d'un portefeuille d'épargne individuelle. L'analyse et l'étude du portefeuille permettent d'abord de détecter les facteurs de risques susceptibles d'impacter la décision de rachat. Une méthodologie de sélection de variables est ensuite appliquée afin de choisir les variables pertinentes qui seront explicatives du comportement de rachat. Ceci a pour finalité l'implémentation de modèles de machine learning (CART et Régression logistique) pour mesurer les effets de chacune de ses variables sur la variable d'intérêt (décision de racheter son contrat ou pas). Les résultats obtenus ne sont pas très satisfaisants liés au nombre et au choix de nos variables ainsi qu'à la dimension de notre portefeuille, les modèles implémentés ont permis néanmoins de dégager certaines caractéristiques qui justifieraient et expliqueraient les comportements de rachat structurel du portefeuille.

Abstract

Keywords: redemption behavior, explanatory variables, structural redemption, machine learning model.

During its life, a savings contract can face multiple events that can modify its probable future financial flows. This is particularly the case when there is a change in the level of the premium or the structure of the premium or when it expires or during a surrender. This last event represents a major risk for the insurer, hence the need for its good control. It is in this context that this brief takes place. The objective is to analyze and then model the repurchase actions of the structures of the policyholders of an individual savings portfolio. The analysis and study of the portfolio first make it possible to detect the risk factors likely to impact the decision to buy back. A variable selection methodology is then applied in order to choose the relevant variables that will be explanatory of the redemption behavior. The purpose of this is to implement machine learning models (Logistic Reggression, CART) to measure the effects of each of its variables on the variable of interest (decision to buy back one's contract or not). The results obtained are not very satisfactory due to the number and choice of our variables as well as to the size of our portfolio, the models implemented have made it possible to retain certain characteristics and trends which justify and explain the results of the portfolio structure.

Table des figures

Listes des tableaux

Liste des abréviations

CIMA : Conférence Interafricaine des Marchés d'Assurances

CART : Classification and Regression Trees

LR : Logistic Regression

Introduction

Le contrat d'assurance vie est un accord entre une compagnie d'assurance qui prend l'engagement irrévocable de verser des prestations au bénéficiaire du contrat en fonction de la réalisation d'évènements aléatoire viagers, en échange de quoi le souscripteur prend l'engagement révocable de verser des cotisations en fonction de la réalisation d'évènements viagers (Milhaud [2011]). Cet accord peut être souscrit à titre individuel (contrat individuel) permettant à l'assuré de constituer une épargne qui sera disponible dans le futur sous forme épargne capitalisée. Ainsi, l'assuré confie son épargne à l'assureur, qui se charge alors de la revaloriser, et promet éventuellement à l'assuré (suivant la nature du contrat) un taux minimal garanti de revalorisation. Cependant, pour des besoins de financement d'un projet personnel ou arbitrage vers un produit plus rentable disponible dans les compagnies concurrentiels, l'assuré peut avoir besoin de l'argent disponible sur son contrat. Face à cette potentielle demande de l'assuré, l'assureur doit satisfaire ses engagements, et ainsi restituer le montant sollicité. Il s'agit d'une opération de rachat.

Le rachat permet à l'assuré de satisfaire son besoin de liquidité en retirant une partie (rachat partiel) ou la totalité de son épargne (rachat total) avant la maturité du contrat prévue à la souscription, ceci moyennant éventuellement des pénalités de rachat. A cet effet, la réglementation des assurances en zone CIMA impose aux compagnies d'assurances d'indiquer sur la police d'assurance notamment pour les contrats qui en comportent, les valeurs de rachat garanties au terme de chacune des 8 premières années au moins, ainsi que, dans le même tableau la somme des primes ou cotisations versées au terme de chacune des mêmes années. Ce rachat est égal à la provision mathématique du contrat diminué éventuellement d'une pénalité qui ne peut dépasser 5% de cette provision mathématique (nulle à l'issue d'une période de 10 ans à compter de la date d'effet du contrat) et lorsque le souscripteur a payé un minimum de prime (15% des primes au cotisations prévues ou au moins 2 primes annuelles, selon l'article 74, 5^{e} alinéa du code CIMA).

Mais, si le rachat est un droit pour l'assuré, il est un risque pour l'assureur dont l'analyse et l'évaluation de son impact sont indispensable pour des enjeux de trésorerie et de compétitivité. C'est donc dans ce contexte que nous nous intéressons dans ce mémoire à la prédiction des comportements de rachat en épargne individuelle. Le but est de mieux cerner les facteurs influençant la décision de rachat des assurés et de prédire cette dernière de la façon la plus précise possible. Pour ce faire, il est nécessaire d'avoir les facteurs explicatifs sur les rachats de manière individuelle. Dès lors, nous pouvons imaginer que les mouvements de rachats peuvent se diviser en deux types de natures bien distinctes : l'une endogène c'est-à-dire liés aux caractéristiques inhérentes aux contrats ou aux caractéristiques individuelles des assurés (rachat structurel) et l'autre exogènes soit en d'autres termes qui ne dépendent ni des produits, ni des assurés mais plutôt lié à l'actualité et au contexte socio-économique (rachat conjoncturel).

Dans ce mémoire, nous nous proposons de modéliser le comportement des rachats suivant les facteurs structurels et les facteurs conjoncturels. Partant de notre problématique et des objectifs, nous avons jugé judicieux de subdiviser notre travail en 3 chapitres :

- Le premier chapitre sera consacré, d'une part, à rappeler les généralités de l'assurance vie ; d'autre part, faire une revue de littérature puis présenter les types de rachats ainsi que la problématique qui en découle ;
- Le deuxième chapitre, quant à lui, traitera de l'analyse descriptive des données ;
- Enfin, le dernier chapitre consistera à décrire les modèles qui seront utiliser pour prédire le comportement de rachat ensuite faire une application des différents modèles sur les données disponibles afin d'obtenir les variables pertinentes.

Au sortir de cette étude, une conclusion générale sera faite dans laquelle l'atteinte ou pas de l'objectif de l'étude sera discutée et s'en suivront des recommandations et une discussion sur les limites de l'étude.

Chapitre 1 : Cadre Conceptuel et Revue de littérature

Ce chapitre présente les différents concepts clés de l'assurance vie nécessaires à la compréhension de l'étude, avec un accent particulier sur la description des rachats en zone CIMA.

1 Généralité sur l'assurance vie

1.1 Présentation générale des contrats

Un contrat d'assurance vie est un engagement réciproque dont le risque, dépend de la durée de la vie humaine. En contrepartie de primes payées par le **preneur d'assurance**, que celui-ci soit le souscripteur (contrat individuel) ou l'adhérent (contrat collectif), la compagnie d'assurance ou **l'assureur** s'engage à verser un capital ou une rente à une ou plusieurs personnes dénommée(s) **bénéficiaire**(s) lorsque le risque survient. A savoir soit le décès de l'**assuré** soit au contraire la survie de ce dernier à un terme donné [7]. Dès lors, un contrat d'assurance vie est caractérisé par les Intervenants aux contrats, sa durée, le type de versement et les garanties.

a- Intervenants aux contrats

Les personnes physiques ou morales qui interviennent au cours de la vie d'un contrat sont :

Assureur : est celui qui porte tous les engagements pris lors de la souscription (valorisation du capital et sa restitution).

Souscripteur : est la personne qui signe la police et paie la prime. Dans le cadre de l'assurance épargne individuelle, le souscripteur est une personne physique agissant à titre individuel pouvant racheter ou transférer le contrat. Toute personne intéressée au contrat (bénéficiaire par exemple) peut se substituer au souscripteur pour payer les primes (article 72) [10].

Assuré : est la personne physique dont le décès ou la survie déclenche le paiement du capital ou de la rente prévue au contrat.

Bénéficiaire : est la personne physique ou morale à qui l'assureur versera le capital ou les rentes en cas de décès de l'assuré, ces montants sont fixés contractuellement lors de la souscription.

b- Durée du contrat

Lors de l'adhésion, le souscripteur a la possibilité de choisir la durée de son contrat et ce, sans aucune contrainte légale. Son choix repose souvent sur ses objectifs personnels mais nous verrons par la suite que les imprévues socio-économiques peuvent amener l'assuré à interrompre son contrat avant l'échéance. Dans le cas d'un contrat d'épargne, il prend normalement fin à la date d'échéance prévue, autrement il peut s'achever lors du décès de l'assurée ou encore suite au rachat total. Si le souscripteur ne rachète pas son contrat et est toujours en vie à la date d'échéance prévue, le contrat est prolongé par tacite reconduction et ce de manière annuelle sauf en cas d'opposition de sa part.

c- Différents types de versement

Le souscripteur a le choix entre plusieurs types de versement des primes relatif à son contrat.

— versement unique : l'assureur ne perçoit qu'une seule prime de la part du souscripteur au moment de la signature du contrat.

— versement périodique : au moment de la souscription du contrat, les deux parties se mettent d'accord sur un calendrier de versement et sur le montant des primes à verser. Ces dernières sont émises par le souscripteur à des échéances déterminées et régulières selon un pas mensuel ou annuel par exemple.

d- Types de garanties

L'assurance vie permet de fructifier l'épargne ou/et de prendre en charge les bénéficiaires de l'assuré après son décès. C'est notamment le cas de l'assurance vie, de l'assurance décès ou de l'assurance mixte.

L'**assurance en cas de vie** prévoit le versement du capital constitué ou d'une rente si l'assuré est toujours en vie au terme du contrat.

L'**assurance en cas de décès** donne lieu au versement d'un capital ou d'une rente à une tierce personne (le bénéficiaire) en cas de décès de l'assuré avant le terme du contrat.

L'**assurance mixte (en cas de vie et de décès)** prévoit le versement d'un capital ou d'une rente, soit à l'assuré, s'il est en vie, soit à un bénéficiaire s'il est décédé.

1.2 Types de contrats

L'assurance vie est un produit de placement puisqu'elle permet de constituer une épargne pour les projets futurs. Sa commercialisation se fait principalement sous deux formes selon que la clientèle est constituée de particuliers ou de personnes morales. Ainsi, elle se regroupe en deux grandes branches à savoir la branche individuelle (dite également la grande branche ou contrats individuels) et la branche collective (dite aussi assurance collective ou branche groupe). Dans la branche individuelle, l'assuré souscrit directement auprès de l'assureur tandis que dans la branche collective, l'assuré qui est un adhérent d'une association est représenté par l'association qui souscrit pour tous ses adhérents. Nous pouvons résumer les différents types de contrats d'assurance vie comme suit :

Assurance vie et épargne

Ce sont des contrats d'assurance en cas de vie comportant des garanties en cas de décès. Ils sont généralement utilisés pour constituer et faire fructifier une épargne, de financer des projets futurs (immobiliers, éducation des enfants, etc.) et/ou optimiser la transmission de son patrimoine à ses proches en cas de décès. Cependant, il ne faut pas confondre contrat d'épargne et contrat de capitalisation en assurance, ce dernier étant un placement de long terme qui ne fait pas intervenir la notion de risque basé sur la durée de vie humaine.

Assurance vie et prévoyance

Ce sont des contrats qui permettent au souscripteur de se protéger contre les risques (décès, invalidité/incapacité) en garantissant le maintien de son niveau de vie. Ces contrats sont souvent qualifiés de contrat à fonds perdus car le souscripteur ne peut en aucun cas récupérer les primes versées. En effet, le contrat de prévoyance prévoit une somme définie en cas de réalisation du risque, qui est indépendante du montant de l'épargne et du temps de cotisation.

Assurance vie et retraite

Ce sont des contrats d'assurance sur la vie à adhésion obligatoire, souscrits par un employeur au bénéfice de ses salariés. Ces contrats permettent au souscripteur de constituer une épargne au cours de sa vie active en vue de préparer sa retraite.

Assurance vie et emprunt

Dans ce contrat, l'assureur se substitue à l'emprunteur assuré si ce dernier vient à décéder avant le terme du prêt. L'assureur verse en une seule fois à l'établissement prêteur une somme égale au capital restant dû à la date du décès, majoré des intérêts courus depuis la dernière échéance de remboursement.

Assurance vie et Indemnité de fin de carrière

Il s'agit d'un contrat d'assurance vie par lequel l'assureur s'engage, en contrepartie des primes reçues, à verser aux membres du personnel de la Contractante, au moment de leur départ à la retraite, une indemnité déterminée conformément aux dispositions du Code du Travail ou de la Convention Collective ou d'un Accord d'Entreprise particulier.

2 Présentation du rachat

En zone CIMA, les caractéristiques de certains contrats relatifs à l'assurance de personnes et aux opérations de capitalisation prévoit une valeur de rachat à tout moment. L'article 74 du Code CIMA permet d'interrompre son contrat avant le terme initialement prévu et d'obtenir de l'assureur le versement de la provision mathématique constituée à la date dudit rachat. Le Code CIMA prévoit aussi la possibilité d'appliquer dans certains cas des pénalités en cas de rachat avant l'échéance. Pour les contrats d'épargne, l'assuré peut choisir de racheter son contrat. Dans ce cas, il récupère la provision mathématique de son contrat, qui représente la somme que l'assureur doit mettre en réserve pour faire face aux engagements futurs pris à l'égard de l'assuré. Cette provision est égale aux sommes versées par l'assuré nettes de frais, et des intérêts acquis à la date donnée.

Le rachat peut être total (l'assuré récupère la totalité de sa provision mathématique) ou partiel (récupère qu'une partie de sa provision mathématique, et son contrat reste en portefeuille). La motivation du rachat relève d'un comportement d'urgence liée à un besoin soudain et urgent de l'épargne constituée. Autrement dit, il s'agit d'une part, des évènements couteux et prévus qui se sont produit dans la vie d'un assuré (financement d'un projet : mariage, naissance, biens immobiliers, etc.). D'autre part, il s'agit des évènements soudains et imprévus liées à la réputation, la qualité des offres avec la concurrence, les stratégies de vente de la compagnie ou encore une crise socioéconomique (pandémie, guerre civile, etc) [2].

Des lors, le comportement de rachat des assurés joue un rôle essentiel dans le résultat d'une ligne de produit et sa compréhension permet à l'assureur d'anticiper des possibles vagues de rachats/non-rachat qui peuvent lui être fortement préjudiciables. Ainsi, il existe bien des facteurs pouvant affecter les comportements de rachat. Globalement ces facteurs de risque peuvent se résumer en deux grandes catégories, les effets structurels et les effets conjoncturels.

2.1 Description du Rachat structurel

Généralement, les assurés dans la zone CIMA souscrivent au contrat d'épargne dans le but de constituer des fonds permettant de résoudre un projet personnel à une période bien choisie de leur vie. Face à cet objectif, l'assuré effectue des cotisations périodiques ou unique auprès de l'assureur, lui en retour doit revaloriser ses primes et constituer des provisions. A une date quelconque de la vie du contrat ce projet peut connaitre quelques modifications obligeant l'assuré a retiré en partie ou en totalité l'épargne constituée.

Par exemple, un assuré souhaite financer un bien immobilier décide de souscrire à un contrat d'épargne sur plusieurs années, au cours de la vie de ce contrat une opportunité de voyage d'étude se présente pour son enfant, bien que cette opportunité d'étude ne soit pas l'objectif principal de ce contrat, il sera aussi l'un des projets futurs de l'assuré puisqu'il a pour devoir d'offrir une meilleure éducation scolaire à ses enfants. Compte tenu du fait que ce contrat d'épargne ne soit pas destiné à financer les frais de voyage et de scolarité, mais dans le souci de garantir un meilleur avenir pour son enfant, l'assuré sera contraint de racheter en partie ou en totalité son épargne constituée pour réaliser ce projet prévu de sa vie ou encore peut racheter le contrat parce que le projet qui a motivé la souscription s'est réalisé avant l'échéance. Ceci correspond à un désir ou un besoin de liquidité immédiat de la part de l'assuré qui rachète pour des raisons personnelles généralement inconnues de l'assureur.

Dans ce cas, le retrait n'est pas dû à un problème de sécurité ou de rentabilité des fonds mais plutôt un besoin urgent de cash nécessaire pour résoudre des problèmes prévus dans sa vie. Ces phénomènes de sortie anticipée en zone CIMA trouvent sa cause dans la maladie, les projets ou d'autres raisons. Ainsi une bonne connaissance des causes des sorties anticipées pourrait donc permettre à l'assureur de construire des produits permettant de mieux contrôler les rachats, par des pénalités de rachat, ou des conditions

qui peuvent inciter l'assuré à conserver son contrat ou encore à réduire sa fréquence de rachat partiel. Partant d'un portefeuille d'expérience, l'assureur pourra alors se protéger contre ce risque structurel basé sur les désirs personnels de l'assuré, que l'on peut analyser en étudiant ses caractéristiques personnelles, ainsi que les caractéristiques du contrat.

2.2 Description du Rachat conjoncturel

Les principales raisons de détention d'un contrat d'assurance-vie épargne sont la sécurité des fonds et le rendement. Lorsqu'elles ne sont pas respectées font partie des effets conjoncturels qui peuvent motiver un assuré à retirer une partie ou la totalité de l'épargne constituée avant l'échéance du contrat. Deux conclusions peuvent être tirées de ce constat en zone CIMA.

D'une part, la relation entre l'assureur et l'assuré via les intermédiaires d'assurance ou les outils à disposition de l'assuré (les services en ligne par exemple) a un impact sur l'attachement de l'assuré. Puisque certains intermédiaires profitent de l'inattention des assurés pour expliquer le contrat à leur guise dans le seul but de bénéficier des commissions au détriment de la bonne information sur les différentes garanties. Par ignorance, l'assuré signe le contrat sans prendre le temps de le lire attentivement, des années après, l'assuré constate que sa part de plus-value est assez faible par rapport à ses attentes. En effet, le taux de revalorisation communiquer de manière orale par l'intermédiaire est différent de ce qui est réellement renseigné dans le contrat. Cette sous information ou mal information est souvent à l'origine des rachats anticipés.

D'autre part, l'image de l'assureur peut être menacé sur les réseaux sociaux ou médias pour non-respect du délai de prestation ou encore le directeur général de la compagnie fait l'objet d'une poursuite judiciaire pour blanchiment. Par conséquent, l'image de la compagnie est salie alors un assuré prudent peut racheter son contrat de peur de ne plus pouvoir récupérer son épargne constituée. Ainsi, la place de la sécurité des fonds laisse à penser que tant que la garantie de récupérer la totalité de ses avoirs n'est pas compromise, certains clients ne rachèteront pas. De là, il existe une réelle incertitude et imprécision concernant la connaissance du risque de rachat conjoncturel puisqu'il est directement lié au raisonnement humain. Celui-ci n'est pas toujours précis mais approximatif.

On pourrait aussi prendre en compte le fait que certains assurés quittent le portefeuille d'une compagnie pour une autre soit parce qu'ils ont un proche dans cette compagnie ou font confiance aveuglement à un intermédiaire d'assurance qui le fait passer d'une compagnie à une autre plus pour son intérêt (commission) que pour la garantie de ses fonds.

Il va de soi que les évènements imprévus tels que les stratégies de vente et la réputation de la compagnie peuvent contraindre l'assuré à racheter son contrat pour des raisons de sécurité ou de rentabilité de ses fonds et non plus pour un évènement prévu nécessitant un besoin immédiat de liquidité.

3 Revue de la littérature et présentation de la problématique

Le comportement d'un assuré est a priori quelque chose de subjectif, délicat à mesurer, surtout lorsque s'y mêle des aspects sociologiques (sexe, âge, niveau de vie) ou conjoncturels (baisse du taux d'intérêt, ...), ou même les deux : réaction des assurés, supposés rationnels, suite à la variation d'un agrégat macroéconomique. Les facteurs explicatifs du rachat peuvent donc être divers et variés. On peut distinguer deux natures d'études sur les facteurs explicatifs du rachat. D'une part, celles qui s'intéressent aux comportements des assurés et d'autre part, celles qui tentent d'expliquer les taux de rachat de manière plus agrégée. La revue de littérature utilisée dans toute la suite est inspirée des mémoires de (Goune [2016], D Médée& G Gwenaëlle [2015] et N RAKAH [2012]) où nous nous attarderons uniquement sur le comportement des assurés.

3.1 Comportement des assurés face aux rachats

Dans cette sous-section, nous nous intéressons aux travaux empiriques réalisées sur les variables explicatives de la survenance du rachat chez les assurés disposant d'un contrat d'assurance-vie.

Dans son étude Xavier Milhaud (2011), constate l'existence d'un clivage dans la littérature entre rachat conjoncturel et structurel. Il propose alors un modèle GLM avec des variables explicatives caractéristiques de l'assuré et d'autres traduisant le contexte économique. Dans un premier temps, l'auteur ne prend pas en compte le contexte économique et se concentre sur la modélisation du rachat structurel. Pour ce faire, il utilise des modèles de segmentation CART et de régression logistique puis étudie l'impact de

variables telles que l'âge, le sexe de l'assuré ou le type de contrat, le montant et la fréquence des primes, la richesse de l'assuré et l'ancienneté du contrat. Il en ressort que le sexe et la prime de risque de l'assuré ne sont pas des variables significatives. L'ancienneté du contrat, le type de contrat et l'option de participation aux bénéfices ont, au contraire, un impact important sur le rachat. Cette segmentation permet de dresser quelques types de profils risqués pour le rachat : les personnes jeunes, les assurés qui versent une prime périodiquement (particulièrement les versements annuels et bimensuels), les assurés les moins fortunés, les contrats avec une clause de participation aux bénéfices. Dans un second temps Xavier Milhaud [4] prend en compte, via une modélisation mélange, l'environnement économique et la forte corrélation qui existe alors entre les comportements des assurés en période de crise. Le modèle développé donne de bons résultats sur certaines lignes de produits et tend à prouver qu'il n'est pas nécessaire de prendre en compte trop de facteurs de risque : l'ancienneté du contrat, le contexte économique et un troisième facteur discriminant suffisent dans la majorité des cas.

Eling et Kiesenbauer publient en 2011 un article sur le sujet, en se servant d'un portefeuille allemand d'historique 2000-2010. A partir d'une régression de Poisson, d'un modèle binomial et d'une régression binomiale négative, les auteurs introduisent comme potentiels variables explicatives de décision de rachat : l'âge de l'assuré, son sexe, l'ancienneté du contrat et le type de produit. Ils trouvent à l'issue de leurs travaux que toutes ces variables permettent d'expliquer les décisions de rachat.

Cerchiara et *al.* mènent en 2009 en Italie fait une étude similaire sur un portefeuille d'épargne comprenant des données d'une fenêtre d'observation plus grande : 1991-2007. Dans leur modèle de régression de Poisson, ils postulent que l'âge de l'assuré à la souscription, son ancienneté dans le portefeuille, l'année de rachat et le type de contrat influencent les décisions de rachat. Ils concluent finalement que toutes les co-variables introduites jouent un rôle dans les décisions de rachat.

Le même sujet intéresse en 1986 Renshaw et Haberman qui réalisent leur étude à partir des données d'historique 1 an (1976) de sept compagnies d'assurance-vie. Ils utilisent un modèle de régression logistique et un modèle binomial avec pour variables explicatives : l'âge de l'assuré à la souscription et au rachat, son sexe, le type de contrat, la compagnie considérée et l'ancienneté dans le contrat. Leurs travaux aboutissent à la conclusion que

l'âge à la souscription, l'ancienneté dans le contrat, la compagnie et le type de contrat sont des facteurs explicatifs de décision de rachat. Ils notent qu'une interaction entre le type de contrat et l'ancienneté de l'assuré dans le portefeuille se révèle également significative.

Pour clore cette sous-section, nous résumons dans le tableau ci-après les variables explicatives retenues dans les études antérieures. Les facteurs les plus influents (d'après les auteurs).

Auteur et année de publication	**Variables explicatives**
Milhaud X. (2011)	-âge -sexe -ancienneté - fréquence de la prime -prime de risque -prime d'épargne -clause de participation au bénéfice -indicateur de richesse de l'assuré
Eling et Kiesenbauer (2011)	-âge - sexe -ancienneté - type de produit
Cerchiara et *al* (2009)	-âge -ancienneté -année de rachat -type de contrat
Renshaw et Haberman (2007)	-âge -ancienneté -compagnie -sexe -type de contrat

Tableau 1 : Récapitulatif des variables explicatives du Rachat

3.2 Présentation de la problématique

Nous nous proposons d'améliorer la connaissance du risque de rachat en zone CIMA, sous un angle différent de beaucoup de professionnels qui abordent le problème des rachats comme une variable observée, et à la relier avec des déterminants, soit endogènes (caractéristiques de l'assuré, du contrat d'assurance), soit exogènes à l'assureur (conjoncture macroéconomique, dépendant de la différence entre un taux benchmark venant du marché et le taux servi par le contrat d'assurance). Cette approche est légitime dans la mesure où l'assuré est rationnel, cultivé, a une bonne connaissance et peut investir sur les marchés financiers. A notre sens, dans la zone CIMA, les rachats structurels gardent presque les mêmes facteurs explicatifs (caractéristiques de l'assuré, du contrat d'assurance, etc) que celle rencontré dans la littérature. Par contre pour le rachat conjoncturel les assurés dans la zone ne s'intéressent pas aux marchés financiers dont il serait aberrant de vouloir expliquer ce type de rachat via les différents facteurs énumérés par certains auteurs. Dès lors, les facteurs de modélisation des rachats conjoncturels diffèrent selon l'environnement, en zone CIMA se sont les évènements soudain et imprévus (l'image de la compagnie, le type d'offre etc) qui pourraient déclencher ce type de rachat. Dans la suite de notre étude nous comptons identifier les différents facteurs favorisant le rachat et d'en ressortir les profils décrivant les comportements des assurés en zone CIMA. Il s'articule ainsi autour de la recherche des causes de rachat et de résiliation : quels facteurs peuvent influencer le comportement de l'assuré, et de quelle manière ? Notre modélisation, qui sera abordée à partir du chapitre 2 et 3, est une modélisation théorique et pratique, là où beaucoup d'autres démarches (machine learning) se fondent sur de l'analyse de données, et relève l'ensemble des variables pertinentes pour expliquer chaque type de rachat.

Chapitre 2 : Analyse descriptive des données

L'objectif de ce chapitre est de réaliser les statistiques descriptives de notre portefeuille afin d'avoir une première intuition sur les facteurs explicatifs de la décision de rachat que nous pourrons retenir comme variables pertinentes pour la suite dans le cadre de la modélisation.

1 Présentation des données

1.1 Description des données

Les fichiers de données mis à notre disposition proviennent du portefeuille des contrats d'épargne individuelle d'une compagnie d'assurance. Ils sont stockés sous format Excel contenant un grand nombre de variables. Dans ce portefeuille le contrat le plus ancien a été souscrit le 1 er Novembre 2016 et le plus récent le 31 juillet 2021. Nous avons reçu deux fichiers l'un nommé « Composition » qui contient 2 174 contrats, chacune de ses lignes possèdent les informations nécessaires à la souscription telles que le numéro de police, le type de produit, la date effet, la date de naissance, la prime, etc. L'autre fichier nommé « prestation » est constitué 866 contrats ayant connu soit un rachat partiel, soit un rachat total (ou encore échu et réglé). Il contient sur les lignes la date de rachat, le numéro de police, le type de prestation, etc. Certaines de ses variables permettent de tenir compte de l'influence de la conjoncture, d'autres étant des variables structurelles propre à l'assuré. Après fusion des deux fichiers et au vu de l'objectif de l'étude, il convient de faire un choix quant aux variables à conserver et à étudier ou encore créer des variables qui permettront de mieux comprendre la décision de rachat.

1.2 Retraitement des bases

Avant d'entamer la description proprement dite des variables, il a fallu au préalable retraiter la base de données afin de ne disposer que des variables qui seront par la suite utiles dans la modélisation des comportements de rachats. Les retraitements suivants ont été effectués :

— Suppression des types de produits n'ayant pas connu de rachats pendant la période de l'étude ;

— Analyse et retraitement des données aberrantes ;

— Création des variables Age de l'assuré à la souscription, Ancienneté du contrat lors du rachat, Nombre de rachat partiel et Statut (Rachat total/ Non Rachat total) ;
— Exclusion de tous les contrats dont l'ancienneté est strictement inférieure à 1 an (durée minimale pour prétendre faire un rachat dans cette compagnie).

Au sortir de cette phase d'apurement de données, nous nous retrouvons avec une base de 1 409 contrats allant de 2016 à 2020 constituée : de variables propres à l'assuré (âge à la souscription, genre), de variables liées au contrat (ancienneté, prime de base, type de produit, fractionnement, nombre de rachat partiel, statut) et de variables conjoncturelles (réseau de distribution et saisonnalité).

2 Analyse descriptive univariée

2.1 Analyse univariée des caractéristiques personnelles de l'assuré

Sur la fenêtre d'étude (2016-2020), le portefeuille étudié est composé en majorité de femmes (57 %) et l'âge moyen des assurés à la souscription se situe à 38 ans, avec un coefficient de variation (rapport entre l'écart-type et la moyenne) de 0.25 (<1), ce qui dénote d'une population homogène en termes de générations.

2.2 Analyse univariée des variables contractuelles

Au sein du portefeuille étudié, une grande majorité (plus de 85 %) des assurés paye des primes mensuelles, il nous est donc difficile de détecter l'influence de la variable fractionnement de prime, c'est pourquoi elle n'est pas retenue dans la suite. Les variables contractuelles à étudier sont au nombre de quatre.

• **Code Produit** : Variable permettant de donner la proportion de souscription par produits commercialisés

Code Produit	Proportions (%)
100	30.21 %
400	3.33 %
410	1.63 %
420	0.78 %
430	0.16 %
500	27.36 %
600	36.53 %
Total	100%

Tableau 2 : Répartition des contrats par type de produit (prime unique/périodique)

Au regard du tableau 2, nous remarquons que les produits 400, 410, 420, 430 sont faiblement représentés. Pour une meilleure analyse, nous allons retenir uniquement les produits homogènes en termes de poids. Ainsi, la variable **CodeProduit** aura désormais 3 modalités (100, 500, 600) résumer dans le tableau 3.

Code Produit	Proportions(%)
100	32,10%
500	29,08%
600	38,82%

Tableau 3 : Répartition des contrats par type de produit (prime périodique)

Partant de la restriction faite sur le type de produit, le portefeuille étudié est désormais constitué de 1 324 contrats. La suite de notre étude portera sur ce portefeuille.

- **Ancienneté :** est une variable calculée pour les contrats en cours comme la différence entre la date d'arrêté des données de l'étude et la date d'effet des contrats. Pour les contrats qui ont été résiliés ou rachetés au cours de l'année nous faisons plutôt la différence entre la date de situation observée et la date d'effet. Faisons une analyse de la moyenne d'ancienneté par contrat suivant l'année de souscription de 2016 à 2020.

Année	Ancienneté Moyenne	Ecart-type	Coefficient de variation
2016	2,46	1,33	0,54
2017	3,10	1,08	0,35
2018	1,88	0,85	0,45
2019	1,66	0,53	0,32
2020	1,23	0,20	0,16

Tableau 4 : Ancienneté moyenne dans le contrat suivant l'année

On constate que l'ancienneté moyenne se situe autour de 2018 environ 2 ans, un écart-type sensiblement égale à 1 an, soit un coefficient de variation de 0.45.

- **Prime de base :** Variable décrivant la prime mensuelle que l'assuré est prêt à épargné au cours de la durée de son contrat. Le tableau 5 donne la prime de base moyenne des assurés, pour les années 2016 à 2020.

Année	Prime de base Moyenne	Ecart-type	Coefficient de variation
2 016	36 111	45 192	1,25
2 017	28 975	45 805	1,58
2 018	29 975	43 214	1,44
2 019	31 186	51 529	1,65
2 020	34 570	60 170	1,74

Tableau 5 : Prime de base moyenne par année de souscription.

Les coefficients de variation élevés (supérieure à 1) traduisent la variabilité des primes, qui peuvent alors significativement différer d'un assuré à l'autre, puisqu'il ne dispose pas les mêmes caractéristiques contractuelles (niveau de prime de base, type de produit, etc.). En observant ces primes suivant le statut, la majorité des contrats qui sont rachetés correspondent aux primes de base dont le montant est de 10 000 ou 20 000. Nous pouvons dont regrouper la variable **Prime de base** en 3 groupes ($[5000,\ 15\,000[$; $[15\,000,\ 30\,000[$; $[30\,000,\ INF[$).

➢ **Variable décrivant le rachat du contrat :**

• **NbreRachatPartiel :** Variable décrivant le nombre de rachat partiel sur un contrat pendant la période de l'étude.

Nombre de Rachat Partiel	0	1	2	3	4	5	Total
Effectifs	1041	225	48	6	3	1	1324
Proportions(%)	78,63%	16,99%	3,63%	0,45%	0,23%	0,08%	100%

Tableau 6 : Répartition des contrats par nombre de rachat partiel

Au regard du Tableau 6, environ 17 % des assurés de ce portefeuille ont effectués au moins un rachat partiel pendant la période d'étude, et près de 79 % n'ont effectués aucun rachat. Comme certaines modalités sont faiblement représentés, nous allons regrouper les modalités 3, 4, 5 en seule modalité noté 3 de proportion 0.76 %.

• **Statut :** Etat du contrat à la date de l'étude (1=rachat total/0= pas de rachat total).

Statut	Nbre de Rachat	Proportion(%)
0	966	72,96%
1	358	27,04%
Total	1 324	100%

Tableau 7 : Répartition des contrats suivant le statut.

Le Tableau 7 présente la distribution de la variable « **Statut** » nous remarquons que plus du quart des contrats du portefeuille sont racheter par anticipation (avant échéance).

3 Etude de l'effet de chaque covariable sur le rachat Total

3.1 Calcul des taux de rachat

Nous détaillons dans cette section le calcul du taux de rachat. Il existe plusieurs façons de mesurer un taux de rachat. Le calcul du taux de rachat peut être abordé suivant deux grandes approches. La première consiste à évaluer le taux de rachat en montant de provision mathématique, la deuxième quant à-elle repose sur un calcul en nombre de contrats. En effet, le rachat total d'un contrat entraîne par définition une baisse du nombre de polices et par conséquent une diminution de la provision mathématique strictement égale à la valeur du contrat qui est sorti du portefeuille de l'assureur. Ainsi, le taux de rachat total peut donc être calculé en montant ou en nombre. Dans le cadre de notre étude le calcul du taux de rachat se fera en nombre pour chaque mois de l'étude (exemple Annexe A). La formule est la suivante :

$$Taux\ de\ rachat\ du\ mois\ (i) = \frac{nombre\ de\ contrats\ rachetés\ au\ mois\ (i)}{nombre\ de\ contrats\ exposés\ au\ rachat\ au\ mois\ (i)}$$

3.2 Effet des variables qualitatives sur le rachat total

- **Influence de la variable Ancienneté sur le taux de rachat total**

Partant de ce qui précède sur l'analyse univarié de l'ancienneté (écart-type étant égale à 1). Nous allons faire un regroupement de cette variable en amplitude 1. Représentons une distribution du taux de rachat total en fonction de l'ancienneté.

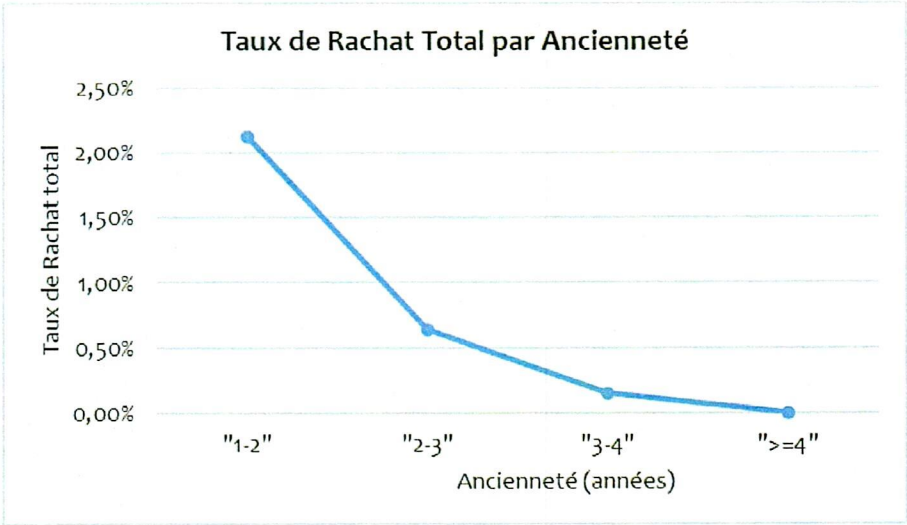

Figure 1 : Taux de Rachat total suivant l'ancienneté

Au regard de la figure 1, nous observons un pic entre 1 et 2 ans d'ancienneté. De plus les taux ont tendance à décroitre en fonction de l'ancienneté. Ce pic est lié au faite que l'assureur mise sur les contrats à court terme dans cette compagnie.

- **Influence du Type de produit sur le taux de rachat total**

Partant du regroupement proposer précédemment sur la variable **TypeProduit.** Nous allons faire une distribution du taux de rachat total par type de produit.

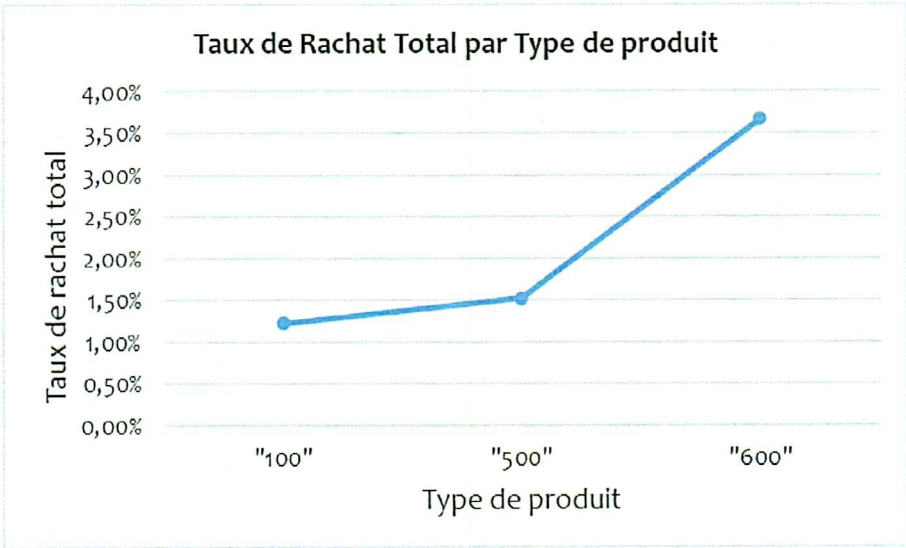

Figure 2 : Taux de rachat total par type de produits

Au regard de la figure 2, nous constatons que le produit de code 600 est le plus souscrit et aussi celui qui est le plus racheté avant échéance. Ceci peut se justifier par le fait qu'il s'agit d'un produit qui a un taux revalorisation mais ayant une pénalité de rachat nulle.

➢ **Taux de Rachat total suivant le Genre**

Partant de l'analyse univariée de la variable Genre nous avons dans le portefeuille plus de femme que d'homme. Représentons le taux de rachat total moyen par genre.

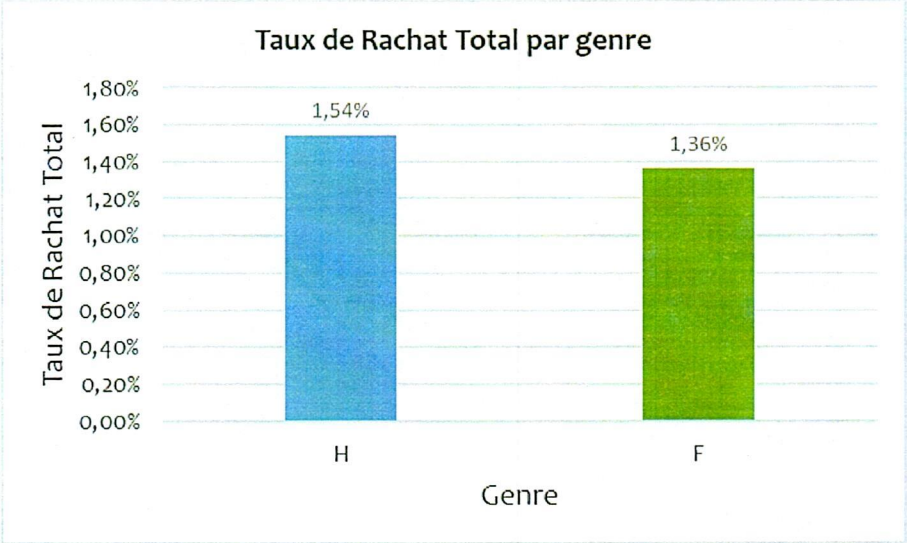

Figure 3 : Taux de rachat total suivant le genre

Au regard de la figure 3, les hommes font plus de rachat anticipé que les femmes. Ceci peut avoir une explication si on prend l'exemple d'un ménage dans la zone, l'homme a plus un besoin récurent de liquidité que la femme vue qu'il a l'obligation de prendre en charge les besoins familiaux prévus ou imprévus tels que le logement, la santé, l'éducation, les dépenses alimentaires et biens d'autres.

➢ **Comparaison Taux de Rachat Total et Taux de Rachat Partiel**

En rappel, le rachat total d'un contrat entraîne par définition une baisse du nombre de polices et un retrait total de la provision mathématique. Par contre le rachat partiel consiste à réduire sa provision mathématique. Nous allons représenter les courbes décrivant le taux de rachat total et celui du rachat partiel

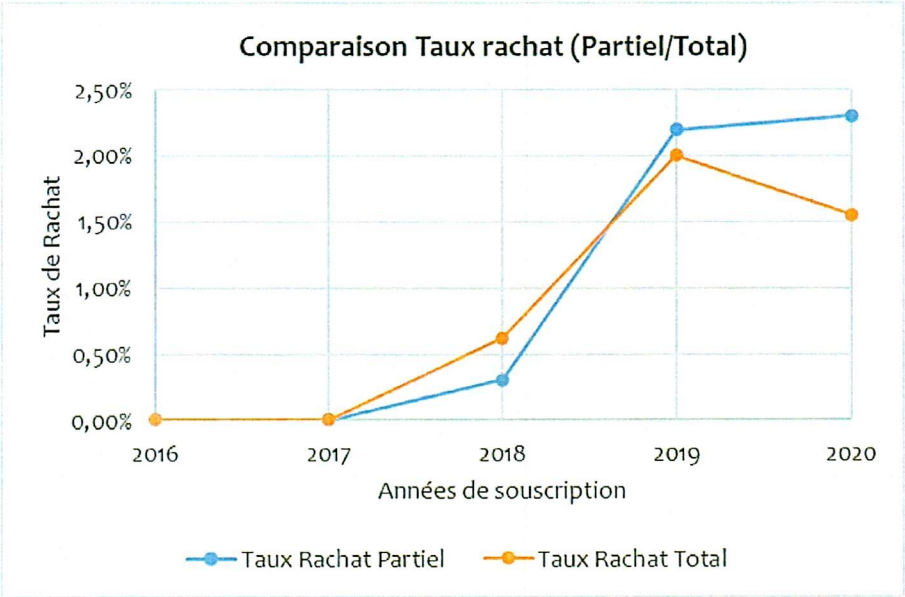

Figure 4 : Comparaison des taux de rachats (partiel/total)

Nous observons sur la figure 4 un gap élevé en 2020 entre le rachat total et le rachat partiel. Ceci peut avoir tout son sens puisqu'en 2020, le pays abritant cette compagnie a connu la pandémie du COVID qui a entrainé le confinement. Elle a obligé les assurés a retiré une partie de leur épargne pour subvenir à leurs besoins quotidiens. Les assurés ne pouvaient pas racheter totalement leur contrat vu que l'objectif pour lequel il constituait l'épargne n'étais pas encore atteint.

3.3 Effet des variables quantitatives sur le rachat total

- Influence de la variable **Age à la souscription** sur le **Statut**

Particulièrement pour le cas de l'étude sur l'impact que l'âge à la souscription peut avoir sur la décision de rachat. Nous allons déterminer le taux de rachat total qui est considéré comme la probabilité qu'un assuré rachète ou non son contrat au cours d'une année. Une manière d'estimer cette probabilité est donnée dans la démarche de construction de l'estimateur de Kaplan-Meier ou celle de Hoem qui eut s'intéresse à la probabilité de survie [8]. Dans notre cas la survie sera ramenée aux faites de racheter ou non son contrat. Nous représentons ci-après les taux de rachat total obtenus par l'approche Kaplan-Meier et Hoem suivant l'âge à la souscription

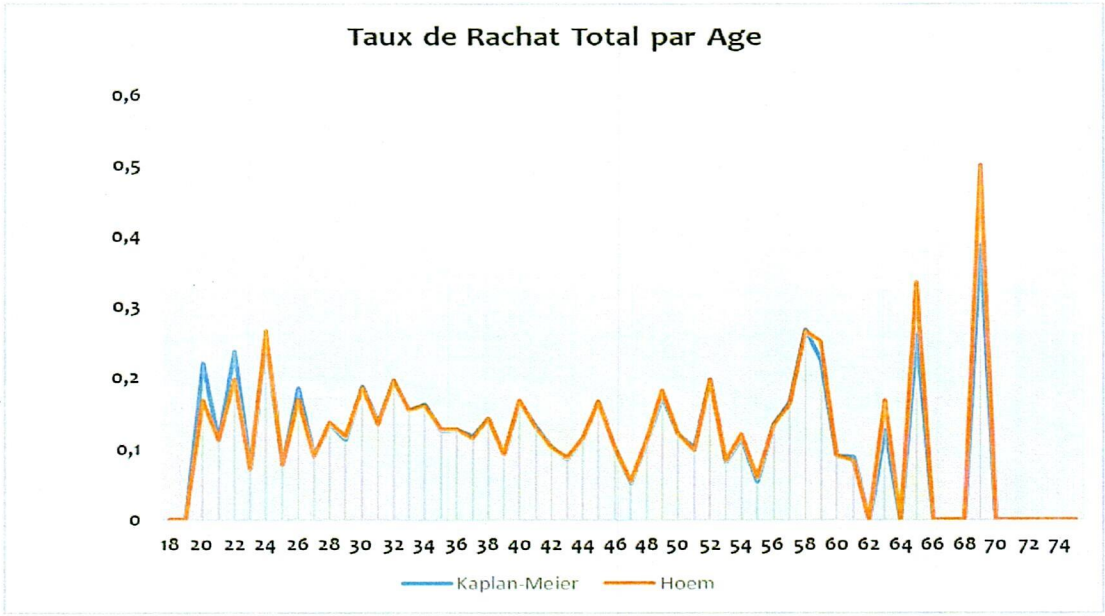

Figure 5 : Taux de Rachat Total par Age

Nous observons trois phases sur la figure 5 qui se matérialisent par un train linéaire croissant de taux à partir de 19 ans jusqu'à 32 ans, ensuite il diminue progressivement entre les âges 32 et 55 ans enfin même si nous observons une légère croissance autour de 58 ans nous avons en majorité des pics atypiques au-delà de 55 ans puisqu'il n'y a pas assez de données à ces âges donc les taux sont volatiles. Ceci s'explique par le fait que les assurés de la première tranche ont des besoins quotidiens de cash pour s'intégrer dans la société. Par contre ceux de la deuxième tranche sont en phase de réalisation dont ont un besoin de liquidité pour financer l'achat d'un bien immobilier ou la concrétisation d'autres projets par conséquent le rachat ne peut pas être régulier. La dernière tranche, l'assuré peut être entrain de constituer un patrimoine pour des bénéficiaires. Un regroupement par tranche d'âge est donné comme suit : $[18,\ 32[$; $[32,\ 55[$ et $[55,\ INF[$.

- **Distribution du taux de rachat suivant la Prime de base**

Partant de ce qui précède sur la variable **Prime de base** le coefficient de variation est supérieure à 1 et la moyenne des primes dans le portefeuille est environ 31 000. Ceci prouve que le portefeuille des primes n'est pas homogène et qu'il pourrait être délimité en classes. Voici une représentation du taux de rachat total suivant le regroupement fait sur la prime.

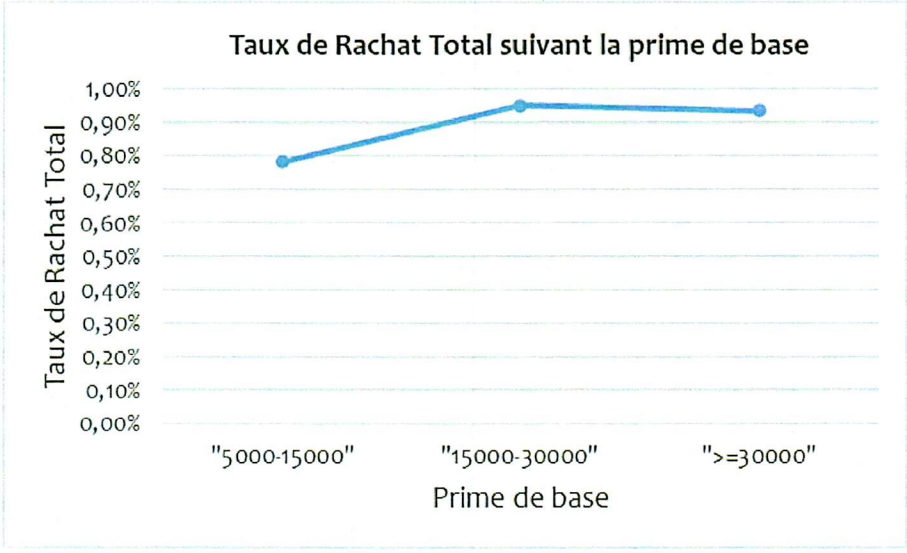

Figure 6 : Taux de Rachat Total par prime

La figure 6, nous permet de voir qu'il semble exister une tendance linéaire entre la prime de base et le rachat total. Donc, la variable prime de base pourra être conservée telle quelle dans la modélisation.

3.4 Conclusion sur l'analyse préliminaire des variables à l'étude

Parvenu au terme de ce chapitre dont le but principal était de rechercher les potentielles variables explicatives du rachat d'un point de vue descriptif (avant la mise en œuvre des modèles de machine learning), nous notons tout d'abord avoir affaire à un portefeuille où le taux moyen annuel de survenance du rachat total avoisine 1.5%. Comme variable identitaire de l'assuré (le genre et l'âge) ont peu laissé entrevoir les signes d'un facteur discriminant. Dans le registre des variables contractuelles, nous avons noté que les variables ci-dessous pourraient être discriminantes :

- Ancienneté au contrat ;
- Prime de base ;
- Nombre de rachats partiels
- Type de produits

Pour compléter ces caractéristiques personnelles des assurés ainsi que celle des variables contractuelles, nous mobiliserons d'autres variables explicatives afin d'analyser l'influence de l'effet conjoncturel sur les décisions de rachat. Les variables conjoncturelles que nous retenons sont :

Saisonnalité : est une variable temporelle correspondant à la date de sortie pour les rachats et les autres sorties, pour les contrats en cours elle a été fixée à 31 Juillet 2021. En effet, nous l'avons introduit comme variable parce que l'assuré, selon certaines périodes de l'année, peut décider de racheter son contrat ou pas. Nous avons délimité cette variable en quatre classes : **Fête de fin année** (Novembre, Décembre, Janvier), **Période courante 1** (Février, Mars, Avril), **Période courante 2** (Mai, Juin, Juillet), **Rentrée Scolaire** (Août, septembre, Octobre).

Réseau de distribution : est une variable correspondant à la commercialisation des produits ayant pour modalités : **Bureau direct, Conseiller vie, Personnel vie, Apporteur divers, Courtier.**

Chapitre 3 : Modélisation des Comportements de Rachat

L'objectif de ce chapitre est de proposer des méthodes permettant de modéliser le comportement du rachat structurel des assurés partant des informations dont nous disposons sur le portefeuille d'épargne. Modéliser ce comportement lors du rachat revient à l'expliquer par des facteurs personnels et/ou contractuels de l'assuré. La modélisation comportementale apparait comme étant un processus qui permet d'étudier le lien entre une variable à expliquer (Rachat ou non de son contrat au cours de l'année) et un ensemble d'autres variables qui seront explicatives.

1 Modélisation Théorique

1.1 Motivation

Le deuxième chapitre et certains travaux qui ont été effectués sur le même sujet nous ont permis d'avoir des intuitions sur les facteurs de risque susceptible d'expliquer les comportements de rachat. De là, nous avons reconstitué un nouveau portefeuille d'épargne à partir de certaines variables qui peuvent expliquer le rachat.

Nous disposons ainsi d'un portefeuille de 1409 contrats d'épargne de 2016 à 2020 ayant 7 variables. Parmi ces dernières, nous avons celles qui concernent l'assuré (âge à la souscription, genre) ensuite celle propre au contrat (ancienneté, prime, type de produit) enfin une variable décrivant le rachat (nombre de rachats partiels, statut).

Au départ, nous voulions modéliser les rachats totaux ainsi que les rachats partiels mais nous nous sommes contentés finalement de ne considérer que l'absence ou non de Rachat total comme étant la variable d'intérêt et pour prendre en compte les rachats partiels, la variable **Nombre de Rachats Partiels** a été créé.

La variable Statut (Rachat total/ Non Rachat total) représente la variable à expliquer. C'est une variable qualitative dichotomique. Si on la note Y, on a $Y \in \{0, 1\}$

$$Y_i = \begin{cases} 1 & si\ Rachat\ au\ cours\ de\ l'année \\ 0 & sinon \end{cases}$$

Les autres variables sont des variables explicatives que nous avons choisi de manière intuitive ou par le biais de la revue de la littérature. Pour une meilleure modélisation, Il faudra effectuer des tests statistiques afin de vérifier statistiquement l'existence d'un lien entre ces variables et celle que l'on doit expliquer.

Un premier modèle intuitif qui semble être adapter pour ce type de modélisation est le cas particulier de la régression logistique des Modèles Linéaires Généralisés. Bien qu'une autre approche par modèle mélange a été développée très récemment dans la littérature par Xavier Milhaud.

1.2 Tests Statistiques

Comme les variables explicatives qui ont été retenu au chapitre précédent se sont fait de manière intuitive. Pour appliquer le modèle choisi à nos données nous devons au préalable vérifier statistiquement l'existence d'un lien entre chacune de ses variables explicatives et la variable à expliquer « **Statut** ».

1.2.1 Test du chi 2

Le test du chi 2 permet de tester l'indépendance de deux variables qualitatives à l'aide d'un tableau de contingence. Le tableau de contingence associé aux variables sous l'hypothèse d'indépendance est obtenu en effectuant le produit des fréquences marginales. Il faut ensuite comparer la distribution empirique obtenue avec le tableau de contingence et la distribution théorique calculée en évaluant la distance entre les deux distributions bidimensionnelles désignées sous les notations :

$$\{f_{ij}: 1 \leq i \leq I ; 1 \leq j \leq J\} \; et \; \{f_{i.} \times f_{.j}: 1 \leq i \leq I ; 1 \leq j \leq J\}$$

La distance est mesurée via la quantité :

$$\chi^2_{(I-1)(J-1)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(ni,j - n..f_{i.} \times f_{.j})^2}{n..f_{i.} \times f_{.j}}$$

est appelée distance du chi 2 à $ddl = (I-1)(J-1)$ degrés de liberté.

Le test d'indépendance du chi-deux permet de prendre une décision quant à l'hypothèse d'indépendance. Cela revient à tester l'hypothèse suivante :

$$H_0 : f_{ij} = f_{i.} \times f_{.j}$$

$$H_1 : f_{ij} \neq f_{i.} \times f_{.j}$$

Après avoir sélectionné les variables statistiquement liées à la variable à expliquer, il convient de détecter parmi celles-ci, les variables fortement inters corrélés. En effet, une forte corrélation entre les variables explicatives sélectionnées peut nuire à la qualité de la régression par la suite. Comme précédemment, le test du chi2 peut nous permettre d'établir l'existence d'un effet entre nos variables explicatives qualitatives si on les croise dans un tableau de contingence. Toutefois, il faudrait fixer une limite au-delà de laquelle

on peut considérer que les variables sont fortement associées. Pour éviter de faire ce choix non judicieux parfois à bien des égards, nous privilégions le test du V de Cramer. Ce dernier permet de comparer l'intensité du lien ou le degré de relation entre deux variables qualitatives.

1.2.2 Test du V de Cramer

Soient A et B deux variables qualitatives à p et q modalités respectivement, n le nombre d'individus sur lesquels A et B ont été observés. La table de contingence est un tableau croisé de q colonnes et p lignes. Si on note n_{ij} le nombre d'individus possédant à la fois la modalité i de la variable A et la modalité j de la variable B. Alors, le V de Cramer est la racine carrée du χ^2 divisé par le χ^2max. Le χ^2max théorique correspond à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1 dont la formule est la suivante [3] :

$$V = \sqrt{\frac{\chi^2}{n \times \min(p-1, q-1)}}$$

1.3 Approche par Régression Logistique

Dans un portefeuille d'assurés donné Ω, on peut observer sur les assurés un certain nombre d'informations (caractéristiques du produit et/ou de l'assuré) représentés par des variables aléatoires $X_1, \ldots, X_p$. Cependant sur les assurés de Ω, il y a une information principale qui nous intéresse dans le cadre de notre étude les assurés qui rachètent ou non leurs contrats, cette information est représentée par la variable Y. Pour reconnaitre parmi les caractéristiques de l'assuré ou/et du contrat les informations qui motivent les assurés a racheté leur contrat. Nous allons nous servir de la régression logistique dont l'objectif est triple : calculer des probabilités de rachat au niveau du contrat, sélectionner les variables les plus significatives pour le déclenchement du rachat et définir une classification des modalités des variables sélectionnées.

1.3.1 Formalisation mathématique

Soit Ω un échantillon d'assurés de taille n, $X = (X_1, \ldots, X_p)$ un ensemble de p variables explicatives et Y est la variable à expliquer binaire (1 : rachat ou 0 : non rachat). Pour un assuré ω de Ω on a $P[Y(\omega) = k/X(\omega)]$ la probabilité conditionnelle qu'un assuré

ω rachète ou non son contrat est vérifié par : $Y = k, k \in \{0; 1\}$ sachant qu'on a observé un certain nombre d'informations $X_1(\omega), \ldots, X_p(\omega)$ sur lui.

- **Hypothèse de base du modèle :**

$$\exists\, \vec{\beta} = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1} \;/\; \forall\, X(\omega) \in \mathbb{R}^p,$$

$$logit(P[Y(\omega) = k \;/\; X(\omega)]) = \beta_0 + \beta_1 X_1(\omega) + \ldots + \beta_p X_p(\omega)$$

Où la fonction logit est une bijection strictement croissante de $]0; 1[$ vers $\mathbb{R}$ définie par : $\forall \in\,]0; 1[\,, logit(u) = \ln\left(\frac{u}{1-u}\right)$.

Sa bijection réciproque est la fonction logistique définie par :

$$logit^{-1} : \mathbb{R} \to]0; 1[$$

$$u \to logit^{-1}(s) = \frac{1}{1+e^{-s}}$$

- **Déduction de $\boldsymbol{P[Y(\omega) = 1 \,/X(\omega)]}$ et $\boldsymbol{P[Y(\omega) = 0 \,/X(\omega)]}$**

Comme $P[Y(\omega) = k \;/\; X(\omega)] = logit^{-1}(\beta_0 + \beta_1 X_1(\omega) + \ldots + \beta_p X_p(\omega))$

$$\boldsymbol{P[Y(\omega) = 1 \,/X(\omega)]} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1(\omega)+ \ldots+\beta_p X_p(\omega))}}$$

$$\boldsymbol{P[Y(\omega) = 0 \,/X(\omega)]} = \frac{1}{1+e^{(\beta_0+\beta_1 X_1(\omega)+ \ldots+\beta_p X_p(\omega))}}$$

2.3.2 Estimation des paramètres du modèle

A partir des observations $(X(\omega);\, Y(\omega)) \in \mathbb{R}^{p+1} \times \{0; 1\},\ \omega = 1, \ldots, n$; on a la fonction de vraisemblance conditionnelle suivante :

- **Fonction de vraisemblance du couple**

$$V_{Y(\omega)/X(\omega)}(\vec{\beta}) = \prod_{\omega=1}^{n} P[Y(\omega) \;/\; X(\omega)]$$

Comme $[Y(\omega) \in \{0; 1\}/\; X(\omega)]$, suit la loi de Bernoulli de paramètre $P[Y(\omega) = 1/\, X(\omega)]$ on a : $V_{Y(\omega)/X(\omega)}(\vec{\beta}) = \prod_{\omega=1}^{n}(P[Y(\omega) = 1/\, X(\omega)])^{y(\omega)}(1 - P[Y(\omega) = 1/\, X(\omega)])^{1-y(\omega)}$

$$V_{Y(\omega)/X(\omega)}(\vec{\beta}) = \prod_{\omega=1}^{n} \left(\frac{P[Y(\omega)=1/\, X(\omega)]}{1-P[Y(\omega)=1/\, X(\omega)]}\right)^{y(\omega)} P[Y(\omega) = 0/\, X(\omega)]$$

- **Estimation de $\widehat{\boldsymbol{\beta}}$**

En maximisant le logarithme népérien de la fonction de vraisemblance V, l'estimateur $\hat{\beta}$ est définit par :

$$\hat{\beta} = argmax_{\vec{\beta} \in \mathbb{R}^{p+1}}\, ln\; V_{Y(\omega)/X(\omega)}(\vec{\beta})$$

où $F(\overrightarrow{\beta}) = ln\ (V_{Y(\omega)/X(\omega)}(\vec{\beta})) = \sum_{\omega=1}^{n} Y(\omega)\ \langle \vec{\beta}, X(\omega)\rangle - ln(1 + e^{\langle \vec{\beta}, X(\omega)\rangle})$ avec $F: \mathbb{R}^{p+1} \to \mathbb{R}$.

Pour estimer $\vec{\beta}$, nous allons utiliser la méthode de Newton-Raphson afin d'optimiser la log-vraisemblance à partir des données et obtenir $\hat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p})$.

Maximiser F revient à résoudre l'équation $\vec{\nabla} F(\hat{\beta}) = \overrightarrow{0}$. Une procédure itérative pour le faire est l'algorithme de Newton-Raphson qui est décrit comme suit :

1. On se donne une valeur initiale $\widehat{\beta_0} \in \mathbb{R}^{p+1}$;
2. On définit la $(k+1)^{ème}$ valeur approchée $\hat{\beta}^{k+1}$ à partir de la $k^{ème}$ $\hat{\beta}^{k}$ par :

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \left[\nabla^2 F(\hat{\beta}_k)\right]^{-1} \vec{\nabla} F(\hat{\beta}_k)$$

où $\nabla^2 F(\hat{\beta}_k)$ est la matrice Hessienne de F en $\hat{\beta}$;

3. On répète la deuxième étape jusqu'à obtenir

$\| \hat{\beta}_k - \hat{\beta}_{k-1} \| \leq \varepsilon \| \hat{\beta}_k \|$ avec ε fixé (par exemple $\varepsilon = 10^{-3}$) ;

4. Ainsi, on pourra prendre $\hat{\beta} = \hat{\beta}_k$ obtenu au dernier rang k.

Après avoir obtenu les estimations des coefficients de régression $\hat{\beta}$ nous pouvons en déduire l'estimation de la probabilité individuelle de rachat de chaque assuré.

$$\hat{p} = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k),$$

Où les $\hat{\beta}_i$ sont les coefficients de régression estimés. Donc chaque assuré a sa propre probabilité de rachat, dépendant de ses caractéristiques personnelles et contractuelles. [11]

1.3.3 Test de significativité des paramètres du modèle

Les paramètres $\beta_1, \dots, \beta_p$ permet de trouver les principaux déclencheurs de rachat. On peut tester la significativité de chacun d'eux à l'aide de la statistique de Wald qui est égale au rapport entre une estimation d'un paramètre et son écart-type estimé. La procédure revient à tester la nullité des coefficients, elle est décrite comme suit :

- **Hypothèse du test**

Pour $j \in \{1, \dots, p\}$ fixé, où p est le nombre de covariables dans le modèle, soit β_j le paramètre correspondant à la $j^{ème}$ covariable du modèle. Les hypothèses à tester sont :

Hypothèse nulle (H_0) : $\beta_j = 0$;

Hypothèse alternative (H_1) : $\beta_j \neq 0$.

➢ **Statistique du test**

$$q_j = \frac{\sqrt{n}\ \hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

Sous (H_0) $\quad q_j \rightarrow_L N(0,1)$

➢ **Résultat du test**

Soit un seuil α fixé,

Si $p - value < \alpha$ alors l'hypothèse (H_0) est rejetée, donc le paramètre β_j est significativement non nul. Sinon, la covariable X_j peut être retirée du modèle.

1.3.4 Mesure de l'effet d'une variable explicative X_j sur la prédiction Y

Pour mesurer l'effet d'une variable explicative sur la variable d'intérêt, nous avons recours à l'Odds-ratio qui est une mesure de quantification du risque. L'Odds-ratio mesure l'évolution du rapport des probabilités d'apparition de l'évènement $Y = 1$ contre $Y = 0$, lorsque X_j passe de x_j à x_{j+1}. Il sert aussi à mesurer le contraste entre les effets d'une variable qualitative.

Considérons l'étude du comportement de rachat d'un contrat (variable à expliquer en 0/1) en fonction de l'ancienneté. Nous voulons évaluer la différence en termes de probabilité de rachat avec un changement de caractéristiques entre deux individus. Soient p_1 la probabilité de racheter totalement son contrat si l'assuré a passé trois ans dans le portefeuille et p_0 la probabilité de racheter si l'assuré a passé deux ans dans le portefeuille.

L'Odds-ratio (OR) se calcule selon la formule [9] suivante :

$$OR = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_0}{(1-p_0)}} = \frac{e^{3\beta_1}}{e^{2\beta_1}} = e^{\beta_1}$$

$$\begin{cases} OR\ (Y(\omega) = 1/\ X(\omega)) > 1, & \textit{indique une influence positive de } X_j \textit{ sur } Y \\ OR\ (Y(\omega) = 1/\ X(\omega)) < 1, & \textit{indique une influence négative de } X_j \textit{ sur } Y \\ OR\ (Y(\omega) = 1/\ X(\omega)) = 1, & \textit{le rachat ou non rachat ont les meme chances de se realisé} \end{cases}$$

1.3.5 Sélection et validation du modèle

Le principe de sélection du modèle consiste à choisir en présence de M1, M2, …, Mn modèles, un seul, le « meilleur » parmi eux par rapport à notre échantillon ou à sélectionner

automatiquement le modèle adéquat. La validation constitue, quant à elle la seconde étape pour vérifier statistiquement si le modèle retenu est le « bon ».

a- Sélection d'un modèle

Un meilleur modèle est toujours évalué par rapport à un critère donné. Parmi les nombreux critères, ceux de l'AIC et BIC sont les plus utilisés. L'idée est de comparer deux modèles donnés en utilisant leur vraisemblance. En réalité, plus la vraisemblance est grande, plus la log-vraisemblance est grande, meilleur est le modèle en termes d'ajustement. En se basant sur cette philosophie, on serait amené à choisir le modèle qui maximise la vraisemblance à savoir le modèle saturé, mais ce dernier sur-ajuste les données car il est sur-paramétré. C'est pourquoi il est nécessaire pour choisir un modèle parcimonieux qui pénalise la vraisemblance par une fonction du nombre de paramètres. Pour exemple l'AIC et BIC sont des fonctions du nombre de paramètres [3] définis comme suit :

- L'AIC (Akaike Information Criterion) est défini pour un modèle logistique M à p paramètres par :

$$AIC(M) = -2\,F(\hat{\beta}_n) + 2p$$

Avec $\hat{\beta}_n$ l'EMV (estimateur du Maximum de Vraisemblance) des paramètres du modèle.

- Le BIC (Bayesian Information Criterion) du modèle M est quant à lui défini par :

$$BIC(M) = -2\,F(\hat{\beta}_n) + p \log n$$

Pour chaque modèle, ces deux critères sont calculés et le modèle qui sera retenu sera celui qui minimise l'AIC ou le BIC.

Par ailleurs, un autre critère est la capacité de prédiction du modèle. L'idée est de chercher à comparer les pouvoirs de prédiction des différents modèles logistiques et de choisir celui qui prédit le mieux. Soit un modèle M_β fournissant une estimation $\hat{p}_\beta(x) = p_{\hat{\beta}}$, une règle de prévision $\hat{g}_\beta$ évidente à partir de cette estimation est :

$$\hat{g}_\beta(x) = \begin{cases} 1 & si\ \hat{p}_\beta(x) \geq s \\ 0 & sinon \end{cases}$$

Avec $s \in [0,\ 1]$ un seuil fixé par l'utilisateur et qui prend par défaut la valeur 0.5 dans la plupart des logiciels statistiques. Il existe plusieurs critères pour mesurer la performance d'une règle de prévision $\hat{g}$. Le plus classique est d'estimer les probabilités d'erreur.

Soit un échantillon $\Omega = (X_1, Y_1), \ldots, (X_n, Y_n)$ et $\hat{g} : \mathbb{R}^{n+1} \rightarrow \{0, 1\}$ une règle de prévision construite à partir de cet échantillon, la probabilité d'erreur de $\hat{g}$ est définie par :

$$L(\hat{g}) = \mathbb{P}(\hat{g}(X) \neq Y/\Omega)$$

Etant donné N règles $\hat{g}_n \ n = 1, \ldots, N$, l'approche consiste à :

- Estimer les probabilités d'erreur de toutes les règles candidates à l'aide de l'échantillon.
- Choisir la règle qui possède la plus petite estimation.

La difficulté réside dans le choix d'un bon estimateur pour $L(\hat{g})$. Intuitivement, une première idée consisterait à utiliser :

$$\hat{L}(\hat{g}) = \frac{1}{n} \sum_{i=1}^{n} 1_{\hat{g}(X_i) \neq Y_i}$$

Mais ceci n'est pas un bon estimateur de $L(\hat{g})$; en effet, la loi des grands nombres ne peut pas s'appliquer car les variables $1_{\hat{g}(X_i) \neq Y_i}$ ne sont pas indépendantes. De plus, le principal problème est que l'échantillon Ω est utilisé deux fois pour calculer $\hat{g}$ puis pour estimer $L(\hat{g})$. [3]

Pour pallier à ce problème, les praticiens utilisent la procédure d'apprentissage/validation qui consiste à découper l'échantillon Ω en deux :
-Un échantillon d'apprentissage $\Omega_l = (X_i, Y_i), i \in J_m$ de taille m $(avec\ m + l = n)$, utilisé pour estimer les probabilités d'erreur de chaque règle $L(\hat{g})$ par :

$$L_n(\hat{g}) = \frac{1}{n} \sum_{i \in J_m} 1_{\hat{g}(X_i) \neq Y_i}$$

avec $J_l \cup J_m = \{1, \ldots, n\}$ et $J_l \cap J_m = \emptyset$.
Il faut noter que $L_n(\hat{g})$ est un estimateur sans biais de $L(\hat{g})$.

Une autre approche pour sélectionner un modèle est la sélection automatique. Autant pour les deux méthodes précédentes, on est en présence de modèles déjà construits et il fallait en choisir un, cette approche de sélection consiste à chercher parmi les variables explicatives $X_1, X_2, \ldots, X_p$, celles qui expliquent le mieux Y pour un critère donné (AIC ou BIC par exemple). Toutefois, sélectionner le meilleur modèle selon le critère qu'on s'est fixé sera coûteux en temps de calcul. C'est pourquoi il convient d'utiliser les méthodes de recherche pas à pas.

Il existe trois grandes méthodes de recherche pas à pas : la méthode ascendante (Forward selection), la méthode descendante (Backward selection) et la méthode progressive (Stepwise selection) [9].

— Méthode Backward : elle part du modèle initial avec toutes les variables disponibles. Au fur et à mesure une variable explicative est retirée. On commence par retirer la variable non significative dont la p-value est la plus élevé. Mais s'il s'agit de la modalité d'une variable, elle est retirée puis regrouper avec la modalité de référence. On répète cet algorithme jusqu' à l'obtention d'un modèle constitué uniquement des variables significatives marquant ainsi l'arrêt.

— Méthode Forward : c'est le contraire de la méthode Backward. A chaque pas, une variable explicative est ajoutée au modèle.

— Méthode Stepwise : elle combine les deux précédentes. C'est la même méthode que celle ascendante mais des variables déjà introduites peuvent être retirées. En effet, au fur et à mesure qu'on introduit de nouvelles variables, il se peut que certaines variables déjà introduites ne soient plus significatives.

b- Validation d'un modèle

Après avoir sélectionné ou choisi un modèle, il est nécessaire de mener une étude pour le valider ou l'affiner en mesurant sa qualité d'ajustement globale par rapport aux données observées. Plusieurs tests ont été proposés dans la littérature.

- **Le test d'adéquation de la déviance**

L'idée ici est de se baser sur la vraisemblance. Plus celle-ci est proche de 1, plus le modèle est proche des données. La déviance d'un modèle M étant définie comme suit :

$$D_M = 2(F_{sat} - F_n(\hat{\beta}_n))$$

avec $\hat{\beta}_n$ l'EMV des paramètres et F_n la log-vraisemblance du modèle.

Elle nous permet de comparer la vraisemblance du modèle avec celle du modèle saturé qui est parfait en termes d'adéquation aux données. Ainsi une déviance faible implique une bonne adéquation. Le test d'adéquation de la déviance pose les deux hypothèses suivantes :

H_0 : Le modèle est adéquat, les données sont bien générées selon par exemple notre modèle logistique contre l'hypothèse.

H_1 : Il ne l'est pas.

En présence de données répétées, la déviance suit une loi du χ^2_{T-P} sous H_0 lorsque $n_t \to \infty,\ t = 1, \dots, T$.

On rejettera ainsi H_0 si $D_{M,obs}$ est plus grande que le quantile d'ordre $1 - \alpha$ de la loi χ^2_{T-P}.

- **Le test d'adéquation de Pearson**

Avec les mêmes hypothèses que le test de déviance, il s'applique toujours dans le cadre des données répétées. La statistique du test est :

$$P = \sum_{t=1}^{T} \frac{(y_t - n_t p_{\hat{\beta}_n}(X_t))^2}{n_t p_{\hat{\beta}_n}(X_t)(1 - p_{\hat{\beta}_n}(X_t))}$$

P suit une loi du χ^2_{T-P} sous H_0 lorsque $n_t \to \infty,\ t = 1, \dots, T$.

H_0 sera rejetée si P_{obs} est plus grande que le quantile d'ordre 1 - α de la loi χ^2_{T-P}.

Remarquons que ces deux tests d'adéquation sont asymptotiques et utilisables uniquement dans le cas de données répétées.

- **Le test de Hosmer-Lemeshow**

En présence de données individuelles, c'est le test de Hosmer-Lemeshow qui est utilisé. Il consiste à évaluer la concordance entre les valeurs prédites et observées des observations regroupées en quantiles, typiquement des déciles. Il dépend du nombre de groupes fixés a priori, et il est peu puissant en cas de mauvaise spécification. Plusieurs approches ont été proposées pour construire ce test. Nous présentons ici la démarche proposée par Laurent Rouvière [1]. En présence de données individuelles $(X_1, Y_1), \dots, (X_n, Y_n)$, la statistique du test se construit comme suit :

i) Les probabilités estimées $P_{\hat{\beta}_n}(x_i)$ sont classées par ordre croissant.

ii) Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand).

iii) Si on note :

- m_k les effectifs du groupe k
- o_k le nombre de succès $(Y = 1)$ observé dans le groupe k
- μ_k la moyenne des $\hat{P}_{\beta}(x_i)$ dans le groupe k.

Alors la statistique est définie par :

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k \mu_k)^2}{m_k \mu_k (1 - \mu_k)}$$

C^2 suit approximativement sous H_0 un χ^2_{K-2}.

1.4 Approche par la méthode de CART

Nous présentons dans cette section comment construire l'arbre à partir des données d'un portefeuille d'assurés. Pour cela, nous allons le faire suivant une chronologie de l'algorithme de CART.

1.4.1 Principes généraux pour la construction de l'arbre

- **Notation :**

$P_{k/N}$: Poids d'une classe $(Y = k)$ dans un nœud N de £ est la proportion des individus de N qui sont dans la classe $(Y = k)$, on notera $P_{k/N} = \frac{card\ (N \cap (Y=k))}{card\ A}$;

P_N : Poids d'un nœud N dans £ est la proportion des assurés d'un portefeuille A qui sont dans le nœud N, on notera $P_N = \frac{card\ N}{card\ A}$;

$i(N)$: indice d'impureté du nœud N dans £ ;

$I(£)$: Indice global d'impureté de l'arbre £.

- **Ensemble de départ :**

A : Portefeuille des assurés qui ont racheté ou non leurs contrats (racine de l'arbre £).

- **Divisions la racine A en deux sous-ensembles disjoints :**

Nous commençons par diviser la racine A en deux sous-ensembles disjoints appelés nœuds, notés N_1 et N_2. Chaque noeud est ensuite divisé de la même manière (s'il contient au moins deux éléments). Au final, nous obtenons une partition de A en plusieurs sous-ensemble appelés nœuds terminaux ou feuilles.

1.4.2 Mesurons la qualité de la division d'un nœud N en N_1 et N_2 grâce à un critère d'impureté :

$i(N) = \psi(P_{1/N}, \dots, P_{G/N})$ où ψ est la fonction d'impureté.

Par conséquent, **l'impureté globale de l'arbre** £ est donnée par :

$I(£) = \sum_{N \in £^*} P_N\, i(N)$ avec $£^*$(ensemble des nœuds terminaux de £).

La pureté de l'arbre est censée augmenter à chaque division puisque plus on descend en profondeur dans l'arbre £, moins les nœuds doivent tendre à être impurs en ce qui est des valeurs de Y en leur sein. Pour y arriver, on examine le problème concret suivant :

Relation entre $I(£)$ et $I(£')$ lorsque $£'$ a été obtenu à partir de £ par division d'un nœud terminal N de £ en deux fils N_1 et N_2 est donné par :

$$\Delta I(£) = I(£) - I(£') = P_N \Delta i(N/d)$$

$$\text{Où } d = (N_1, N_2) \text{ et } \Delta i(N/d) = i(N) - P_{N_1/N} i(N_1) - P_{N_2/N} i(N_2)$$

$$P_{N_s/N} = \frac{card\ N_s}{card\ N}, \quad s = 1,\ 2.$$

Ainsi, il faut choisir la division $d = (N_1, N_2)$ à opérer au nœud N de telle sorte que $\Delta I(£)$ soit supérieure à 0 et le plus grand possible c'est-à-dire $\Delta i(N/d) \geq 0$.

1.4.3 Critères pratique de division d'un nœud :

Plus on descend dans l'arbre, plus les individus doivent tendre à se ressembler du point de vue de leurs valeurs pour les variables X_j. De telle sorte que ceux qui se retrouvent dans un même nœud terminal N tendent à se ressembler significativement (tout en se distinguant de ceux qui ne sont pas dans N) du point de vue de leurs valeurs pour X_j. Ainsi, il faut choisir X_j^* une des variables de X_j et diviser le nœud N à base des valeurs de X_j^* sur les individus de N, telle sorte que ceux envoyés dans le même fils se ressemble plus entre eux en termes des valeurs de X_j^* que ceux envoyés dans l'autre fils. Mais en essayant de garanti un $\Delta i(N/d)$ aussi grand que possible.

Schéma algorithmique pour la division d'un nœud

1 Pour j = 1 à p faire

1.1 Identifier l'ensemble $D_j(N)$ des divisions possibles du nœud Nqui peuvent être décider uniquement à base des valeurs de la seule variable X_j^* sur les individus de N.

1.2 Pour $d \in D_j(N)$, calculer $\Delta i(N/d)$

1.3 Adopter comme meilleur division possible du nœud N à base des seules valeurs de la variable X_j^* dans N, la division $d_j(N) \in D_j(N)$ et vérifiant :

$$d_j(N) = argmax\ \Delta i(N/d)$$

Au sortie de cette boucle « pour » on a identifié p divisions potentiel du nœud N

$$d_1, \ldots, d_p$$

2 Adopter comme division finale à effectuer du nœud N, la division $d_{j^*}(N)$ avec $j^* \epsilon \{1, \ldots, p\}$ tel que $j^* = argmax\ \Delta i\big(N/\ d_j(N)\big)$

a- **Division possible d'un nœud $N \in £$ à base d'une variable X_j pour j $\epsilon\{1, \dots, p\}$**

- **Cas ou X_j est binaire dans N :**

Les divisions possibles de N à base des valeurs (0 et 1) de X_j sur les assurés de N.

$$d_{j^*}(N) = (N_1, N_2) \quad \text{tel que} \begin{cases} N_1 = \left\{i \in N / x_j^{(i)} = 0\right\} \\ N_2 = \left\{i \in N / x_j^{(i)} = 1\right\} \end{cases}$$

- **Cas ou X_j est qualitative ordinale ou quantitative discret dans N :**

Les divisions possibles de Nà base des modalités ordonnée $m_1 < m_2 < \cdots < m_q$de X_j sur les assurés de N.

$$d_{j^*}(N) = (N_1, N_2) \quad \text{tel que} \begin{cases} N_1 = \left\{i \in N / x_j^{(i)} \leq m_l\right\} \\ N_2 = \left\{i \in N / x_j^{(i)} \geq m_{l+1}\right\} \end{cases}$$

- **Cas ou X_j est quantitative continue dans N :**

On peut au préalable effectuer un tri dans l'ordre croissant de l'échantillon de ses valeurs avant de commencer la construction de l'arbre. Cependant, parfois pour accélérer on pourra préférer discrétiser les valeurs de $X_j$dans le nœud N, par exemple par le min, les quantiles et le max des $x_j^{(i)}$, pour $i \epsilon N$. Ainsi, on peut transformer $X_j$dans Npar une variable qualitative ordinale de modalités ordonnées. On applique la division précédente sur ce nœud.

- **Cas ou X_j est qualitative nominale dans N à $q \geq 3$ modalités :**

On fixe l'une des modalités comme modalité de référence

$$d_{j^*}(N) = (N_1, N_2) \quad \text{tel que} \begin{cases} N_1 = \left\{i \in N / x_j^{(i)} = m_q\right\} \\ N_2 = \left\{i \in N / x_j^{(i)} \in \left\{m_1, \dots, m_{q-1}\right\}\right\} \end{cases}$$

1.4.4 Critères d'arrêts naturels :

L'étape suivante consiste à définir quand arrêter les divisions, ce qui relève du choix de l'utilisateur. Certaines règles d'arrêt sont naturelles tandis que d'autres sont purement arbitraires : les divisions s'arrêtent évidemment lorsque les observations des variables explicatives dans une classe donnée sont identiques ; soit définir un nombre minimal d'observations dans un nœud (plus ce nombre est petit et plus le nombre de feuilles sera grand) ; soit choisir un seuil $\beta \in [0; 1]$ de décroissance minimum de l'impureté permettant d'arrêter la division lorsque $\Delta i(N/d) \leq \beta\, i(N)$.

En conclusion il n'y a pas de règle d'arrêt dans les CART : la construction de l'arbre £ s'arrête lorsqu'on ne peut plus diviser aucun de ses nœuds terminaux ou feuilles. Ainsi, les variables sélectionnées pertinentes pour la prédiction sont celles qui ont servi à diviser au moins un nœud et les autres X_j ont donc été éliminées comme étant non pertinentes pour prédire Y.

2 Modélisation Pratique

2.1 Sélection des variables explicatives

Nous avons retenu à partir des statistiques descriptives du chapitre 2 et en s'inspirant des travaux qui ont été réalisés sur le même sujet les covariables suivantes :

- Age à la souscription ;
- Genre ;
- Ancienneté au contrat ;
- Prime de base ;
- Nombre de rachats partiels ;
- Type de produits.

Cette étape consiste à vérifier statistiquement l'existence d'un lien entre chacune de ces variables et la variable statut (Rachat total/Non Rachat total) qu'on veut expliquer. Au préalable nous avons catégorisé toutes les variables explicatives retenues de manière intuitive lors de l'analyse descriptive avant de réaliser le test d'indépendance du chi 2 entre chacune d'elle avec la variable d'intérêt. Les p-values obtenues lors de ce test nous permet de conclure sur le rejeter ou non de l'hypothèse nulle Ho d'indépendance de chacune des

variables explicatives avec la variable d'intérêt (c'est-à-dire p-value < 5%). Dans notre étude le test de chi 2 de la variable d'intérêt **Statut** (Rachat total/Non Rachat total) avec les variables explicatives sont résumés dans le tableau 8.

Variables	P-value
Ancienneté	$2.442\ e^{-08}$
Agesouscription	0.1 945
Prime	0.1 802
Sexe	0.0 707
TypeProduit	0.0 625
nbrepartiel	$7.514\ e^{-07}$

Tableau 8 : Récapitulatif des p-values du test de chi 2.

Au sortie de ce test du chi 2, les variables ayant un lien significatif avec la variable d'intérêt Statut sont : l'**Ancienneté** et le **nombre de rachat partiel**.

Par la suite, nous avons étudié le degré de corrélation entre l'Ancienneté et le nombre de rachat partiel afin d'éviter les problèmes de multi colinéarité dans la régression. Pour cela, nous avons fait le test V de Cramer, nous avons obtenu des coefficients très proches de 0 (tous inférieurs à 0.5). D'où, nous en déduisons que nos covariables ne sont pas dépendantes.

En somme, les variables **Ancienneté** et **nombre de rachat partiel** sont les variables explicatives qui seront intégrer dans la régression logistique.

2.2 Mise en pratique de la régression logistique

Pour cette phase de modélisation, nous avons divisé notre échantillon en deux : un échantillon d'apprentissage (70%) pour la phase d'apprentissage et un échantillon test (30%) pour la phase de test du modèle obtenu lors de l'apprentissage. Ces échantillons ont été obtenus par tirage aléatoire simple sans remise en s'assurant qu'ils aient chacun la même structure de donnée que celle de la base de données mère. Ainsi, notre échantillon d'apprentissage contient 926 contrats dont 252 contrats qui ont été rachetés, soit une proportion de rachat d'environ 27% comparable avec la proportion de la base de données

globale (27%). L'échantillon test contient quant à lui 398 contrats dont 106 contrats rachetés, soit une proportion d'environ 27% également.

a- Implémentation du modèle et Résultats

Les variables explicatives retenues après l'analyse descriptive sont toutes des variables plusieurs modalités. Notre premier réflexe a été de construire un modèle logistique complet avec toutes nos covariables à l'aide de la fonction glm de R. Le but est de pouvoir détecter les modalités non significatives qui peuvent découler d'une mauvaise répartition de notre portefeuille (Confère Annexe B).

Au regard des p-value de ce modèle, nous constatons que les variables explicatives qui n'avaient pas de lien avec la variable d'intérêt statuts lors du test de chi 2 sont non significatives. Ainsi nous allons refaire le modèle en prenant compte uniquement les variables explicatives retenues lors du test du chi 2 (Confère Annexe C).

Ce modèle permet de voir que certaines modalités d'une même variable ne sont pas significatives d'où la nécessité de s'interroger sur leur éventuel regroupement avec d'autres modalités. Nous allons regrouper la modalité non significative dont la p-value est la plus élevé avec la modalité de référence ce qui donne une covariable avec un nombre de modalités réduit puis on refait la modélisation. Nous allons répéter ce processus et s'arrêter uniquement lorsque toutes modalités des variables seront significatives (Voir Annexe D).

Notons que l'AIC obtenu avec ce modèle est inférieur à celui du modèle initial contenant toutes les variables.

Par ailleurs, afin d'interpréter ces résultats, nous avons analysé les odds-ratio de chacun de nos facteurs de risque présent dans le modèle final.

	Intercept	nbrerachatpartiel (1)	ancienneté (2-3)	ancienneté (4-3)
OR	0.4902	0,4193	0.7212	0.3540

Tableau 9 : Tableau récapitulatif des Odds-Ratio

Nous constatons que l'odds ratio de toutes les variables retenues dans notre modèle sont inférieure à 1 alors toute modification d'une d'entre elle en hausse aura une influence négative sur la probabilité de rachat.

b- Validation du modèle

Ayant à notre disposition des données individuelles, nous avons effectué le test d'Hosmer-Lemeshow pour évaluer la pertinence du modèle logistique retenu. Sous R, nous avons utilisé la fonction hoslem.test de la librairie (ResourceSelection). La p-valeur obtenue ($< 2.2\ e^{-16}$) n'étant pas supérieure au seuil $\alpha = 0.05$, on admet alors que le modèle n'est pas bien adapté aux données. Toutefois, nous avons aussi évalué le pouvoir prédictif du modèle. Pour cela, nous avons repris exactement tous les regroupements que nous avons effectués sur les données apprentissage lors de la construction du modèle sur les données test afin d'obtenir le modèle prédiction. Pour évaluer la qualité de prédiction nous avons utilisé la matrice de confusion évaluant le taux de bien et mal classés sur l'échantillon test. Afin de pouvoir analyser les résultats fournis par la matrice de confusion, nous avons calculé certains indicateurs :

- Le taux de succès correspondant à la probabilité de bon classement du modèle ;
- La sensibilité ou encore Taux de Vrais Positifs qui indique la capacité du modèle à retrouver les positifs (rachats) ;
- La précision qui indique la proportion de vrais positifs parmi les individus qui ont été classés positifs ;
- La spécificité, qui à l'inverse de la sensibilité, indique la proportion de négatifs détectés ;
- Le taux de faux positifs qui correspond à la proportion de négatifs qui ont été classés positifs.

Taux de succès	Sensibilité	Précision	Spécificité	Taux de Faux positifs
73,62%	11,96%	40,00%	97,97%	2,03%

Généralement un « bon » modèle doit avoir un taux d'erreur et un taux de faux positif proches de 0 et une sensibilité et une spécificité proches de 1. Nos résultats bien qu'ils soient représentatifs et satisfaisants, ne nous permettent pas d'attester que notre modèle est bon. En effet, un modèle peut être jugé comme adéquat que si on le compare à d'autres modèles. Nous allons implémenter le modèle de CART pour pouvoir faire une comparaison.

Néanmoins, la méthode de régression logistique nous a permis de sélectionner les variables les plus pertinentes pour expliquer le comportement du rachat.

2.3 Mise en pratique de la méthode de CART

Dans la classe des arbres de décisions, nous avons implémenté l'algorithme de CART à partir de la Library **Scikit-learn** (Decision Trees) de Python. Pour le faire, nous avons appliqué sur nos données le schéma typique de l'analyse prédictive en scindant aléatoirement nos données en échantillon d'apprentissage et de validation dont les tailles respectives sont 926 et 398 assurés. Dans le but de développer le modèle sur le premier et évaluer les performances sur le second à travers la confrontation des classes observées et prédites.

a- Implémentation du modèle et Résultat

Nous construisons d'abord un arbre T_{max} sur les données d'apprentissage sans définir un critère d'arrêt optimal ensuite à partir de l'échantillon de validation nous calculons l'erreur de prédiction de l'arbre maximal T_{max} qui est de 29,40% correspondant aux termes non diagonaux de la matrice de confusion. Nous remarquons que cet arbre a trop de feuille et admet une représentation complexe en taille finale et à un mauvais taux de classification.

Taux de succès	Sensibilité	Précision	Spécificité	Taux de Faux positifs
70,60%	10,68%	36,54%	88,81%	11,19%

Au vue de ses différentes incongruités observées sur l'arbre T_{max}, nous nous sommes attelés à l'élaguer cet arbre pour obtenir un arbre plus performant (T_{elague}). Comme nous l'avons vu plus haut dans la partie théorique, le critère d'arrêt de division d'un arbre peut se faire en définissant un nombre minimal d'observation dans un nœud (plus ce nombre est petit et plus le nombre de feuilles sera grand). Ainsi dans notre étude, un sommet de l'arbre n'est pas segmenté s'il est composé de moins de 200 individus et une segmentation est validée si et seulement si les feuilles générées comportent tous au moins 65 observations. Nous obtenons ainsi l'arbre élagué représenter comme suit :

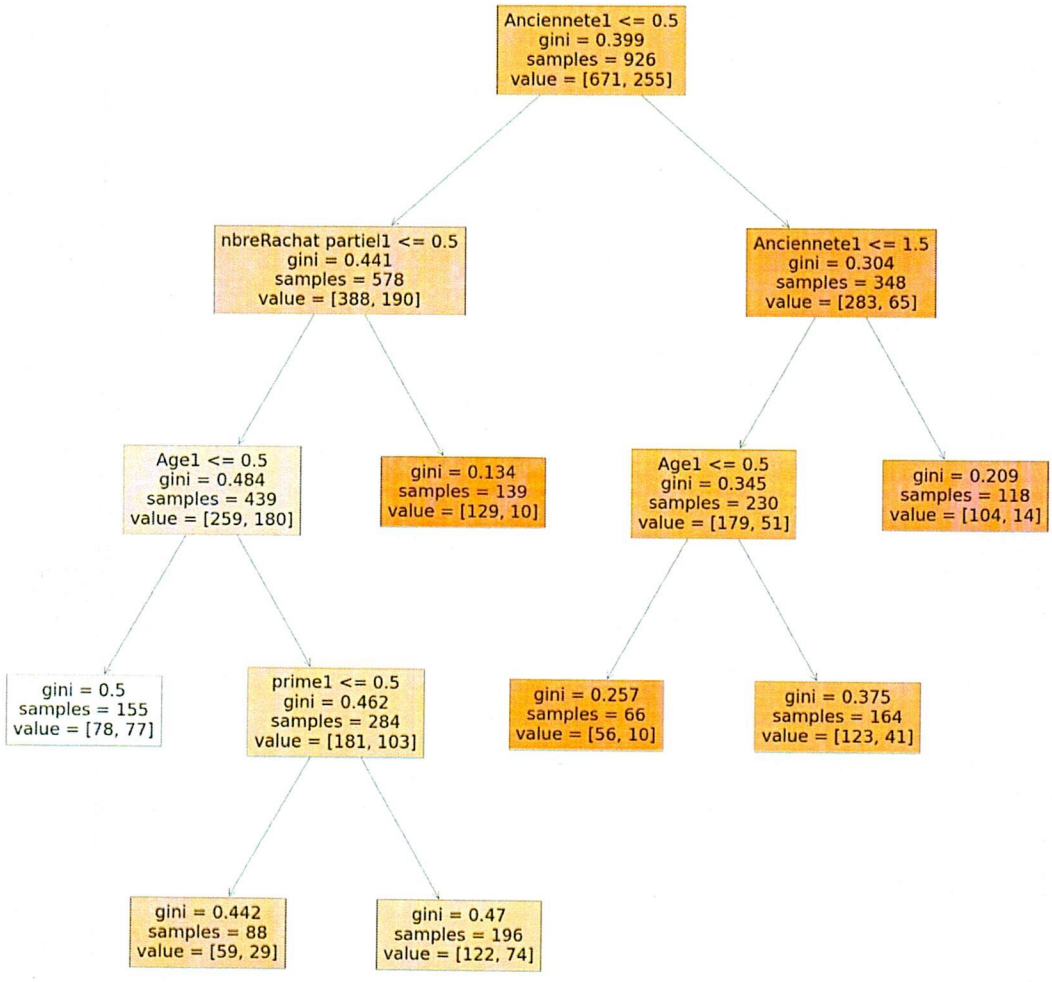

Figure 7 : Arbre de décision élagué

Une lecture de l'arbre, si l'ancienneté $\leq 0{,}5$ alors l'interprétation est la suivante la branche de gauche est le Non Rachat Total et celle de droite est le Rachat Total. Le reste de l'arbre se lit de la même manière.

Les variables sélectionnées dans la construction de l'arbre sont l'ancienneté, le nombre de rachat partiel, l'âge à la souscription et la prime de base. Nous remarquons que le genre et le type de produit n'apparaissent pas dans cet arbre parce que leurs effets ne semblent pas être significatifs.

- **Evaluation du test :**

Pour évaluer les performances prédictives de l'arbre, nous appliquons l'arbre T_{elague} sur l'échantillon test composé de 398 observations. Nous obtenons à partir de la

fonction (**arbreFirst.predict)** de Python une prédiction 33 rachats sur les 398 contrats. Nous confrontons les classes observées et prédites via la matrice de confusion.

		predicted	
		Y=0	Y=1
observed	Y=0	273	22
	Y=1	92	11

Tableau 10 : Matrice de confusion de l'arbre élagué

Afin de pouvoir analyser les résultats fournis par la matrice de confusion, nous avons calculé certains indicateurs résumés dans le tableau suivant :

Taux de succès	Sensibilité	Précision	Spécificité	Taux de Faux positifs
71,36%	10,68%	33,34%	92 ,54%	7,46%

L'arbre (Figure 7) paraît surdimensionné. Nous avions remarqué notamment que plusieurs feuilles issues du même sommet père portaient des conclusions identiques (voir couleur identique sur les nœuds terminaux issu du même père). Alors, nous allons introduire un nouveau paramètre pour réduire la taille de l'arbre. Nous spécifions (**max_leaf_nodes** = 3) c.-à-d. nous souhaitons obtenir un arbre qui produit 3 règles au maximum (en priorité les segmentations qui maximisent les contributions). Nous obtenons l'arbre suivant :

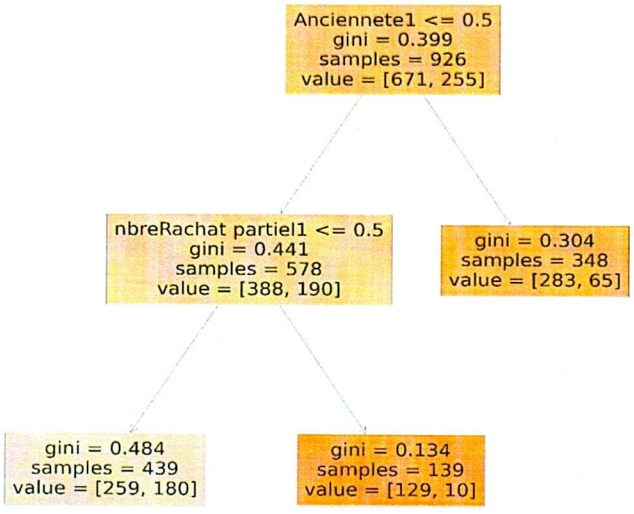

Figure 8 : Arbre de décision élagué final

L'arbre est fortement simplifié, tout en maintenant ses qualités prédictives puisque nous obtenons exactement la même matrice de confusion que sur l'échantillon test, et par conséquent des valeurs identiques des indicateurs de performances. D'où, les variables significatives présentent sont **l'ancienneté** et le **nbre de rachat partiel.**

2.4 Conclusion

Nous allons comparer ses deux modèles de segmentation suivant le critère de performance. Il s'agit de la sensibilité et de la spécificité. La sensibilité est définie comme le nombre de success sur le nombre de contrats rachetés observés, et la spécificité est le nombre de correct rejections sur le nombre de contrats non-rachetés observés. Le tableau résume les critères de performance des différentes méthodes de classification.

	CART ($T_{elagué}$)	LR
S_e	10,68%	11,96%
S_p	92,84%	97,97%
$(1 - S_e)$	89,32%	88,04%

L'analyse du comportement des rachats par la régression logistique présentent moins de misses et plus de correct rejection, les résultats étant comparables et les erreurs presque équilibrées entre les deux méthodes.

D'un point de vue technique, nous avons vu que le processus d'analyse du comportement des assurés lors du rachat peut se réaliser soit par l'emploi du modèle logistique soit par l'emploi de la méthode de CART, les résultats étant en adéquation. Des profils type de risque se dégagent plus facilement à partir des statistiques descriptives ou des CART, tandis que le modèle de régression logistique donne accès à des indicateurs intéressants tels que le odds-ratio. Les deux modèles apportent des résultats complémentaires et font intervenir des hypothèses bien différentes, mais servent globalement une même cause puisque le compromis entre sensibilité et spécificité est meilleur avec CART mais le nombre de misses est un peu plus élevé, ce qui nous conduirait ici à choisir le modèle de régression logistique pour plus de prudence. Ce modèle a permis néanmoins de dégager certaines caractéristiques telles que l'ancienneté, le nombre de rachat partiel qui justifieraient et expliqueraient les comportements de rachat structurel du portefeuille.

Conclusion

L'objectif de ce mémoire a été d'analyser, de comprendre, d'expliquer et de prédire les comportements de rachat du portefeuille d'épargne individuelle de la compagnie, ce qui passe d'abord par l'identification des facteurs de risque (structurels et conjoncturels) susceptibles d'influencer la décision de rachat et qui se poursuit par la mesure de leurs effets.

Des statistiques descriptives ont permis d'identifier les facteurs de risque à prendre en compte dans la modélisation. Pour cette dernière, intuitivement, l'approche probabiliste a été privilégié en faisant recours aux techniques de machine Learning. Notre variable d'intérêt étant dichotomique, c'est une régression logistique qui a été utilisée. Un volet entier a été dédié à la sélection des variables à introduire dans le modèle pour réduire la dimension des variables explicatives. Il découle de l'analyse et de la modélisation que certaines variables à effets structurels comme l'Ancienneté, le nombre de rachats partiels influencent la décision de rachat des assurés. Pour valider ce modèle de régression logistique, nous avons au préalable implémenter l'algorithme de CART sur nos données puis une comparaison entre les deux modèles a été fait en se basant sur des critères de performance tels que la sensibilité et la spécificité. Au sortie de cette comparaison le modèle de régression logistique a le taux erreur le moins élevé par conséquent il est le plus prudent.

Comme tout travail scientifique, cette étude ne saurait être parfaite et présente ainsi quelques limites. L'une des principales concerne la non prise en compte de l'effet conjoncturels de nos données. En outre, notre étude aurait été complète si nous avions traité les deux types de rachats surtout que les comportements de rachat structurels et le rachat conjoncturels ne sauraient être identiques. En perspective, nous proposons ainsi d'étendre cette étude au cas du rachat conjoncturel. Bien qu'en zone CIMA, la prise en compte des variables conjoncturelles est problématique au sens où elles sont difficilement modélisables. Puisque nous nous sommes confrontés à la difficulté de modéliser un comportement humain qui relève parfois de facteurs cachés tels que la réputation de la compagnie et les relations d'amitié ou de famille. Mais nous avons proposé à la fin de notre analyse descriptive deux variables permettant de mesurer certains effets conjoncturels (la

saisonnalité et le réseau de distribution). Nous pouvons aussi citer la non prise en compte de certaines variables explicatives susceptibles de déterminer la décision de rachat. Nous pensons par exemple à la catégorie socioprofessionnelle et les raisons du rachat que nous avons exclu vu le très haut de données manquantes et incohérentes. Faisant suite aux résultats obtenus et aux difficultés rencontrées (la taille des données insuffisantes et la période observation courte) pour la mise en œuvre de cette étude, nous suggérons au service client de renforcer la proximité avec les clients, toujours avoir à dialoguer avec les assurés lors du rachat pour tirer la bonne information du motif (comme c'est le cas dans les banques lors de la demande de crédit). Ceci pourra permettre d'affiner l'étude en segmentant les données en profils de risque.

Bibliographie

[1] **ROUVIERE LAURENT (2015)** Régression Logistique avec R, Université Rennes 2, UFR Sciences Sociales.

[2] **MILHAUD XAVIER** (2011) Segmentation et modélisation des comportements de rachat en Assurance Vie.

[3] **RAKOTOMALA RICCO** (2017) Pratique de la Régression Logistique, Université Lumière Lyon2.

[4] **MILHAUD XAVIER** (2012) Mélanges de GLMs et nombre de composantes : application au risque de rachat en Assurance Vie, Université Claude Bernard Lyon 1.

[5] **RAKAH NAFOULAH** (2012) Modélisation des rachats dans les contrats d'épargne, Centre des Etudes Actuarielles.

[6] **LADIAS NICOLAS** (2012) Analyse des causes de rachats sur des contrats d'assurance vie, Mémoire d'actuaire, Institut de Science Financière et d'Assurances.

[7] **DJAGANA OUATTARA** (2020) Cours d'assurance des personnes en Master d'actuariat, Institut International des assurances Cameroun.

[8] **AYMRIC KAMEGA** (2021) Cours de modèle de durées en Master actuariat, Institut International des assurances Cameroun.

[9] **NDONG NGUEMA** (2016) Cours de durées censurées, Master de statistique Appliquée, Ecole Polytechnique, Yaoundé.

[10] **CIMA** (2019) Traité instituant une organisation intégrée de l'industrie des assurances dans les Etats Africains. Paris, France.

[11] **Raoul NOUMSI** (2019) Analyse du cout des sinistres en Assurance Automobile. Ouvrage, Edition Universitaire Européenne.

Annexe

Annexe A : Calcul du taux de Rachat par Mois (un cas)

Taux de Rachat Total par Mois suivant les Types de Produits

1*/ Taux de rachat par produit

	Rachat éffectué par mois 2020												
Type de produit	janv-20	févr-20	mars-20	avr-20	mai-20	juin-20	juil-20	août-20	sept-20	oct-20	nov-20	déc-20	Total
100	2	5	2	1	1	2	9	0	3	1	1	6	33
500	7	5	2	0	0	4	3	5	1	2	2	1	32
600	4	17	8	5	4	7	12	5	5	8	6	16	97
Total	13	27	12	6	5	13	24	10	9	11	9	23	162

	Possibilité de rachat par mois en 2020												
Type de produit	janv-20	févr-20	mars-20	avr-20	mai-20	juin-20	juil-20	août-20	sept-20	oct-20	nov-20	déc-20	Total
100	256	281	298	305	310	313	332	345	353	361	364	374	3892
500	279	294	305	317	322	328	329	336	345	349	354	357	3915
600	46	91	127	171	197	209	237	256	281	310	346	374	2645
Total	581	666	730	793	829	850	898	937	979	1020	1064	1105	10452

2020	janv-20	févr-20	mars-20	avr-20	mai-20	juin-20	juil-20	août-20	sept-20	oct-20	nov-20	déc-20	Total
Taux de rachat	2,24%	4,05%	1,64%	0,76%	0,60%	1,53%	2,67%	1,07%	0,92%	1,08%	0,85%	2,08%	19,49%

	Rachat éffectué par mois												
Type de produit	janv 2019	févr 2019	mars 2019	avr 2019	mai 2019	juin 2019	juil 2019	août 2019	sept 2019	oct 2019	nov 2019	déc 2019	Total
100	0	4	5	1	7	5	7	1	7	5	7	5	54
500	0	1	0	0	2	2	3	13	12	1	4	6	44
600	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	0	5	5	1	9	7	10	14	19	6	11	11	98

	Possibilité de rachat par mois												
Type de produit	janv 2019	févr 2019	mars 2019	avr 2019	mai 2019	juin 2019	juil 2019	août 2019	sept 2019	oct 2019	nov 2019	déc 2019	Total
100	167	181	188	197	198	208	216	219	227	231	242	255	2529
500	51	75	101	120	126	149	170	192	220	237	260	276	1977
600	0	0	0	0	0	0	1	1	1	1	6	33	43
Total	218	256	289	317	324	357	387	412	448	469	508	564	4549

2019	janv 2019	févr 2019	mars 2019	avr 2019	mai 2019	juin 2019	juil 2019	août 2019	sept 2019	oct 2019	nov 2019	déc 2019	Total
Taux de rachat	0,00%	1,95%	1,73%	0,32%	2,78%	1,96%	2,58%	3,40%	4,24%	1,28%	2,17%	1,95%	24,36%

Annexe B : Résultat du modèle régression logistique de toutes les variables

glm (formula = base.appren$rachattotal ~ base.appren$prime + base.appren$anciennete +
base.appren$nbrepartiel + base.appren$Age + base.appren$sexe +
base.appren$produit, family = binomial(link = logit))

Deviance Résiduels :

Min	1Q	Median	3Q	Max
-1.1180	-0.8784	-0.6406	1.3167	2.2197

Coefficients :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.55325	0.23223	-2.382	0.01720 *
base.appren$prime(1.5e+04,3e+04]	0.19224	0.17717	1.085	0.27790
base.appren$prime(3e+04,Inf]	0.37268	0.19760	1.886	0.05929 .
base.appren$anciennete(2,3]	-0.41318	0.19295	-2.141	0.03225 *
base.appren$anciennete(3,4]	-1.11854	0.35745	-3.129	0.00175 **
base.appren$anciennete(4,Inf]	-15.98751	388.26282	-0.041	0.96715
base.appren$nbrepartiel1	-0.97157	0.23751	-4.091	4.3e-05***
base.appren$nbrepartiel2	-0.44784	0.44826	-0.999	0.31777
base.appren$nbrepartiel3	0.22216	0.87347	0.254	0.79923
base.appren$Age(32,55]	-0.07419	0.16590	-0.447	0.65475
base.appren$Age(55,Inf]	-0.19792	0.40079	-0.494	0.62143
base.appren$sexeM	-0.15512	0.15493	-1.001	0.31672
base.appren$produit500	-0.20713	0.21610	-0.958	0.33782
base.appren$produit600	0.03932	0.20102	0.196	0.84494

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1084.1 on 925 degrees of freedom

Residual deviance: 1019.0 on 912 degrees of freedom

AIC: 1047

Number of Fisher Scoring iterations: 15

Annexe C : Résultat de la régression logistique final sur les données apprentissage

glm(formula = base.appren$rachattotal ~ base.appren$nbrepartiel + base.appren$anciennete, family = binomial(link = logit))

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0169	-0.9410	-0.6144	1.4338	2.3636

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.58513	0.09699	-6.033	1.61e-09 ***
base.appren$nbrepartiel1	-0.98649	0.23594	-4.181	2.90e-05 ***
base.appren$nbrepartiel2	-0.47889	0.43937	-1.090	0.275744
base.appren$nbrepartiel3	0.19516	0.85793	0.227	0.820054
base.appren$anciennete(2,3]	-0.40352	0.18409	-2.192	0.028384 *
base.appren$anciennete(3,4]	-1.15856	0.34097	-3.398	0.000679 ***
base.appren$anciennete(4,Inf]	-15.98094	389.25757	-0.041	0.967252

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1084.1 on 925 degrees of freedom

Residual deviance: 1015.7 on 919 degrees of freedom

AIC: 1039.7

Number of Fisher Scoring iterations: 15

Annexe D : Résultat de la régression logistique final sur les données Test

```
glm(formula = base.appren$rachattotal ~ base.appren$nbrepartiel +
    base.appren$anciennete, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8933  -0.8933  -0.7782   1.4911   2.3197

Coefficients:
                                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                     -0.71281   0.09459    -7.536   4.84e-14 ***
base.appren$nbrepartiel1        -0.86914   0.23474    -3.703   0.000213 ***
base.appren$anciennete(2,3]     -0.32677   0.17926    -1.823   0.042323 **
base.appren$anciennete(3,4]     -1.03838   0.33935    -3.060   0.002214 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1084.1  on 925  degrees of freedom
Residual deviance: 1011.9  on 922  degrees of freedom
AIC : 1026.9
Number of Fisher Scoring iterations : 4
```