Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 31 janvier 2025

Par : Adrien PASSUELLO
Titre : Optimisation de la stratégie commerciale des affaires nouvelles en assurance automobile selon la performance individuelle des contrats

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : ADDACTIS France
Signature :

DocuSigned by:
Guillaume ROSOLEK
9D514C92FFE845E...

Directeur de Mémoire en entreprise :
Nom : Simon SAVOYE
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

DocuSigned by:
Simon Savoye
0426841C8C3A4D4...

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

Secrétariat :

Signature du responsable entreprise

Savoye

Bibliothèque :

Signature du candidat

Passuello

Résumé

Le secteur de l'assurance non-vie est marqué par une forte concurrence, nécessitant des stratégies avancées pour ajuster la tarification. Dans ce contexte, il est essentiel pour les assureurs de prédire le comportement futur des assurés afin de maximiser la rentabilité tout en maintenant une base de clients solide. Ce mémoire a pour objectif de proposer une méthode d'optimisation de la stratégie commerciale chez les nouveaux contrats, basée sur une inspiration de la valeur client future (définie comme étant la valeur actuelle nette des profits futurs générés par les assurés), offrant ainsi une approche innovante pour répondre à ces enjeux.

Pour élaborer cette optimisation, plusieurs modèles ont été développés. Le premier est un modèle de durée de vie des contrats, visant à estimer le temps qu'un assuré reste en portefeuille. Le second est un modèle de projection de marges qui estime les profits futurs uniquement en fonction des caractéristiques initiales des assurés. Ces deux modèles sont cruciaux pour le calcul de la valeur client future qui constitue une mesure clé pour évaluer la rentabilité et le potentiel de fidélisation de chaque assuré. En complément, un modèle de transformation a été utilisé pour estimer les probabilités d'acceptation des tarifs proposés aux clients. Cette transformation permet de mieux comprendre comment les assurés réagissent aux ajustements de prix.

La combinaison de ces trois modèles permet de calculer une valeur client potentielle pour chaque individu en fonction du montant proposé lors de sa souscription. L'optimisation tarifaire repose sur cet indicateur qui est recalculé après application de diverses variations à la prime commerciale proposée aux individus. Afin d'identifier les profils à priori fidèles et rentables, une classification des individus a été mise en place. Grâce à cette segmentation, il est possible d'ajuster les tarifs de manière plus précise et équitable, maximisant ainsi la valeur client potentielle globale tout en maintenant un équilibre entre rentabilité et satisfaction des clients.

Ce mémoire établit une méthodologie solide et innovante pour optimiser la stratégie commerciale des affaires nouvelles en assurance automobile, qui pourrait être complexifiée et étendue à divers produits d'assurance pour une meilleure adaptation aux besoins du marché.

Mots-clés : *Optimisation tarifaire*, *Valeur client*, *Assurance non-vie*, *Affaires nouvelles.*

Abstract

The non-life insurance sector is highly competitive, requiring advanced strategies to adjust pricing. In this context, it is essential for insurers to predict the future behavior of policyholders in order to maximize profitability while maintaining a solid customer base. This thesis aims to propose a method for optimizing the strategy for new contracts, inspired by customer lifetime value (defined as the net present value of future profits generated by policyholders), thus offering an innovative approach to addressing these challenges.

Several models have been developed for this optimization. The first is a contract lifetime model, designed to estimate the duration of time a policyholder remains in the portfolio. The second is a margin projection model, which estimates future profits based solely on the initial characteristics of policyholders. These two models are crucial for calculating customer lifetime value, a key measure for evaluating the profitability and retention potential of each policyholder. Additionally, a transformation model was used to estimate the probabilities of acceptance for the prices proposed to the customers. This transformation provides a better understanding of how policyholders respond to price adjustments.

The combination of these three models enables us to calculate a potential customer value for each individual based on the proposed amount at the time of subscription. Pricing optimization relies on this indicator, which is recalculated after applying various adjustments to the commercial premium offered to individuals. To identify potentially loyal and profitable profiles, a classification of individuals was established. This segmentation allows for a more precise and fair adjustment of prices, thereby maximizing overall potential customer value while maintaining a balance between profitability and customer satisfaction.

This thesis establishes a robust methodology for optimizing the commercial strategy of new automobile insurance business, which could be further developed and extended to various insurance products for better adaptation to market needs.

Keywords: *Pricing optimization*, *Customer lifetime value*, *Non-life insurance*, *New business*.

Note de Synthèse

Contexte et motivations

Le secteur de l'assurance non-vie, particulièrement l'assurance automobile, est de plus en plus compétitif. Cette évolution entraîne pour les assureurs un besoin accru de stratégies commerciales avancées, notamment dans l'attraction de nouveaux clients et dans la fidélisation des profils les plus rentables à long terme. Dans un contexte où les bancassureurs, la digitalisation, et la simplification des résiliations facilitent la mobilité des assurés, les compagnies doivent innover pour se distinguer sur un marché saturé. Historiquement, les assureurs ont structuré leurs stratégies tarifaires sur des critères standardisés, mais cette approche présente des limites pour prédire la rentabilité et la fidélité de chaque client.

Dans ce cadre, ce mémoire propose une méthodologie innovante d'optimisation de la stratégie commerciale des affaires nouvelles par une approche dérivée de la notion de "valeur client future". Cette approche permet de mieux comprendre la rentabilité potentielle des nouveaux clients au-delà des seuls critères démographiques, en intégrant la probabilité de rétention et la marge espérée générée pour chaque profil. L'objectif central de ce mémoire est de proposer une méthodologie d'optimisation tarifaire innovante et facilement opérationnelle pour répondre à la problématique suivante : comment optimiser le tarif proposé aux nouveaux assurés lors de la souscription pour maximiser la valeur client potentielle de l'ensemble du portefeuille, tout en maintenant un équilibre entre compétitivité et rentabilité à long terme ?

Valeur client future

En marketing, la valeur client future est généralement définie comme la valeur actuelle nette des profits générés par un client sur toute la durée de sa relation avec l'entreprise. Sa formule théorique est la suivante

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \left(\frac{\left[\text{Prime}_i(t) - \text{Coûts}_i(t)\right] \times \mathbb{P}_i(t)}{(1+r)^t} \right) - \text{CA}_i \tag{1}$$

où
- ▷ $\text{Prime}_i(t)$ est la prime payée par l'assuré i au temps t
- ▷ $\text{Coûts}_i(t)$ représente les coûts de l'assuré i au temps t (sinistres, frais généraux...)
- ▷ $\mathbb{P}_i(t)$ est la probabilité que l'assuré i appartienne encore au portefeuille au temps t
- ▷ r est le taux d'actualisation (ou le coût du capital de la compagnie)
- ▷ $T_h = 10$ est l'horizon de temps choisi pour l'estimation de la valeur client future
- ▷ CA_i représente les coûts d'acquisition du contrat pour l'assuré i

Cependant, dans cette étude portant sur les affaires nouvelles, plusieurs données nécessaires à cette formule théorique ne sont pas disponibles. Notamment, il n'existe pas d'informations sur la sinistralité passée des assurés. Par conséquent, la définition de la valeur client future a été ajustée, en prenant comme estimation du risque la prime pure, c'est-à-dire la partie de la prime correspondant au coût théorique des sinistres.

Ainsi, en notant $\text{PC}_i(t)$ et $\text{PP}_i(t)$ respectivement les primes commerciale et pure de l'assuré i au temps t, la valeur client future est définie de la manière suivante dans ce mémoire

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \frac{\overbrace{\left[\text{PC}_i(t) - \text{PP}_i(t)\right]}^{\text{Marge}_i(t)} \times \mathbb{P}_i(t)}{(1+r)^t} \tag{2}$$

Cette formule est une approximation simplifiée et ajustée pour le contexte spécifique des nouveaux contrats d'assurance, où les coûts d'acquisition sont homogénéisés et peuvent être négligés dans l'optimisation qui sera mise en place. La différence entre la prime commerciale et la prime pure représente ici la marge espérée générée par un assuré, sans prendre en compte les frais et les sinistres individuels. Cette approche permet d'identifier efficacement les profils fidèles et rentables, en offrant une vision simplifiée mais pertinente de la valeur que chaque assuré pourrait apporter au portefeuille.

Ainsi, la valeur client future d'un assuré se construit en fonction de la marge espérée et de la probabilité de rester dans le portefeuille, estimées respectivement via un modèle de projection de marges et un modèle de durée de vie, plus précisément le modèle à risques proportionnels de Cox.

Nota Bene :

> *Bien que le terme de "valeur client future" soit utilisé, la grandeur définie à la formule (2) représente en réalité une "valeur contrat future", car celle-ci s'applique spécifiquement à l'assurance automobile. En effet, le calcul n'intègre pas les informations liées au multi-équipement, c'est-à-dire les autres contrats que le client pourrait détenir auprès de la même compagnie (contrat MRH, Prévoyance, etc.). Cette étude est donc limitée à l'assurance automobile en raison de l'absence d'informations sur le multi-équipement, mais la méthodologie peut être facilement généralisée pour optimiser la stratégie commerciale sur d'autres types de produits d'assurance.*

Enfin, ce mémoire introduit la notion de valeur client potentielle qui tient compte de la probabilité d'acceptation du devis proposé au moment de la souscription. Cette probabilité, appelée probabilité de transformation, est prédite grâce à un modèle de transformation. Lorsque le montant $\text{PC}_i(0) \times (1+v_g)$ est proposé à l'individu i lors de la souscription, où $\text{PC}_i(0)$ représente le montant proposé initialement par l'assureur et v_g une variation entre -30% et $+30\%$, la probabilité de transformation est notée $Tr_i(v_g)$ et la valeur client potentielle de cet individu i est définie par

$$VC_i(v_g) = \sum_{t=0}^{T_h} \frac{\text{Marge}_i(t) \times \mathbb{P}_i(t)}{(1+r)^t} \times Tr_i(v_g) \tag{3}$$

où, pour tout $t \in [\![0, T_h]\!]$, les marges espérées de l'individu i $\text{Marge}_i(t)$ et ses probabilités de survie $\mathbb{P}_i(t)$ dépendent du montant de cotisation proposé $\text{PC}_i(0) \times (1+v_g)$.

Modèles utilisés dans le mémoire

Pour mettre en place, à terme, une optimisation de la stratégie commerciale des affaires nouvelles, plusieurs modèles ont été utilisés afin de calculer les différents éléments de la valeur client potentielle définie dans la formule (3) ci-dessus.

Modèle à risques proportionnels de Cox

Le modèle à risque proportionnel de Cox est utilisé pour estimer les fonctions de survie des contrats, correspondant aux probabilités $\mathbb{P}_i(t)$ qu'un assuré i reste dans le portefeuille au cours du temps t. Ce modèle est bien adapté et permet une estimation fiable de la survie en fonction des caractéristiques initiales des contrats.

Modèle de projection de marges

Le modèle de projection des marges vise à estimer la rentabilité espérée de chaque assuré sur un horizon de 10 ans, en calculant la différence entre les primes commerciales et technique (*i.e.*

primes pures chargées). Dix-huit modèles XGBoost sont développés pour prédire, année par année, les primes en fonction des caractéristiques initiales des assurés et des prédictions des primes de l'année précédente. Chaque modèle XGBoost est donc spécifique à un type de prime (pure ou commerciale) et à une année donnée. Imbriquer les prédictions successivement permet d'obtenir toutes les primes pour chaque individu sur 10 ans uniquement à partir de ses caractéristiques initiales et de ses primes lors de la souscription, ce qui est idéal pour travailler en affaires nouvelles.

Les marges espérées annuelles sont alors calculées avec la formule suivante

$$\text{Marge}_i(t) = \text{PC}_i(t) - \text{PP}_i(t) \times (1 + \text{Taux de chargements}) \tag{4}$$

Nota Bene :

> *La prime pure a été chargée, i.e. transformée en prime technique, pour calculer les marges espérées car les chargements ne représentent pas un gain pour l'assureur mais plutôt un montant destiné à couvrir ses frais.*

On appelle "modèle de projection de marges" ce processus qui calcule les marges dans le temps en s'appuyant sur une série de modèles successifs, où chaque modèle utilise les prédictions du modèle précédent. Les marges espérées projetées par ce modèle sont ensuite intégrées dans le calcul de la valeur client future, apportant une estimation de la rentabilité espérée pour chaque contrat.

Modèle de transformation

Le modèle de transformation, basé sur une régression logistique, est utilisé pour estimer la probabilité d'acceptation du devis proposé lors de la souscription. Il évalue l'influence des variations de la prime commerciale, ainsi que des caractéristiques individuelles des assurés (âge, type de véhicule, etc...), sur la décision des potentiels assurés d'adhérer au contrat.

La probabilité d'acceptation ainsi prédite, appelée probabilité de transformation, est multipliée à la valeur client future pour obtenir la valeur client potentielle de chaque assuré. Ce modèle aide à orienter les stratégies commerciales en anticipant l'impact des variations des tarifs proposés lors de l'adhésion des clients, permettant ainsi d'optimiser la composition du portefeuille d'affaires nouvelles.

Optimisation de la stratégie commerciale

Démarche mise en place

Comme dit plus tôt, l'objectif de ce mémoire est de proposer une méthodologie innovante pour optimiser la stratégie commerciale des affaires nouvelles en assurance automobile. La méthodologie adoptée dans ce mémoire vise à optimiser les primes commerciales proposées lors de la souscription maximisant la valeur client globale. Le première étape consiste à segmenter le portefeuille en clusters de profils similaires en termes de valeur client future calculée pour chaque individu à partir du tarif proposé initialement par l'assureur.

Segmenter un portefeuille en clusters de profils similaires permet d'analyser les comportements et les besoins des clients de manière plus globale. Cette approche aide à identifier des tendances, à anticiper les comportements et à ajuster les stratégies marketing en conséquence. Cela permet aussi de personnaliser les actions en fonction des segments, en créant des offres adaptées qui améliorent la fidélité et la satisfaction. Enfin, puisque l'on cherche à identifier les profils fidèles et rentables, segmenter selon la valeur client future offre un avantage supplémentaire permettant de concentrer les gestes commerciaux sur les clients qui semblent les plus prometteurs à long terme.

Sous les différentes notations ci-dessous, le problème d'optimisation que ce mémoire cherche à résoudre se traduit mathématiquement dans la formule (5)

- ▷ $G \geq 1$ le nombre de clusters (distincts) formant le portefeuille
- ▷ $v_g \in [-30\%, +30\%]$ la variation appliquée au montant de la cotisation pour le cluster $g \in [\![1, G]\!]$

- ▷ n_g le nombre d'individus dans le cluster $g \in [\![1, G]\!]$
- ▷ $\mathbb{P}_i(t)$ la probabilité que l'individu i appartiennent encore au portefeuille au temps t
- ▷ $\text{Marge}_i(t)$ la marge espérée de l'individu i au temps t
- ▷ r le taux d'actualisation constant et égal à 2%
- ▷ $Tr_i(v_g)$ la probabilité de transformation de l'individu i en fonction de la variation v_g
- ▷ $VC_i(v_g)$ la valeur client potentielle de l'individu i à un horizon T_h en fonction de la variation appliquée au montant de la cotisation en entrée

$$\begin{cases} \max\limits_{v_1,\dots,v_G} & VC \text{ potentielle globale} = \sum\limits_{g=1}^{G} VC_g \\ \text{Sous contraintes} \end{cases} \quad \text{avec} \quad VC_g = \sum_{i=1}^{n_g} \overbrace{\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Marge}_i(t)}{(1+r)^t} \times Tr_i(v_g)}^{VC_i(v_g)} \tag{5}$$

Le schéma de la figure 1 récapitule le processus mis en place pour calculer la valeur client potentielle globale que l'on cherche à maximiser.

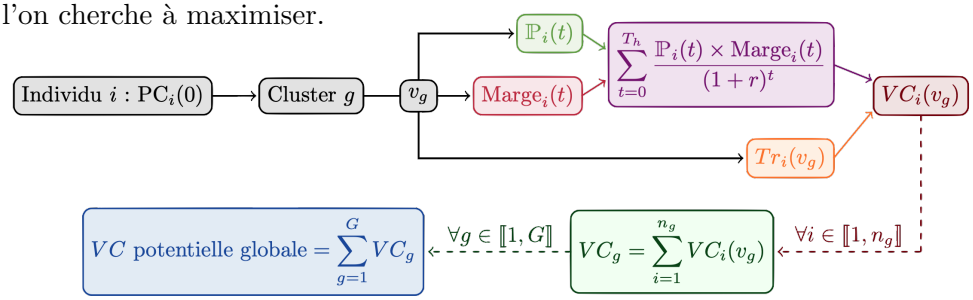


FIGURE 1 : Schéma récapitulatif du processus de calcul de la valeur client potentielle globale

Le processus est donc le suivant : l'assureur a calculé une prime commerciale pour un individu i, notée $\text{PC}_i(0)$. À partir de ce montant et des caractéristiques initiales de son contrat, celui-ci est affecté à un cluster g de profils similaires en termes de valeur client future. Une variation v_g, commune à tous les individus du cluster g, est appliquée au montant de cotisation $\text{PC}_i(0)$ qui devient alors égal à $\text{PC}_i(0) \times (1 + v_g)$. À partir de celui-ci, pour tout temps $t \in [\![0, T_h]\!]$, les marges espérées ($\text{Marge}_i(t)$) et les probabilités de survie du contrat ($\mathbb{P}_i(t)$) sont prédites afin de calculer une nouvelle valeur client future de l'individu i ($\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Marge}_i(t)}{(1+r)^t}$). La probabilité d'acceptation du devis à ce nouveau prix ($Tr_i(v_g)$) est également calculée, puis multipliée à la nouvelle valeur client future de l'individu i pour obtenir sa valeur client potentielle ($VC_i(v_g)$). Cela est effectué pour tous les individus du cluster g, et par somme, la valeur client potentielle du cluster est obtenue (VC_g). Enfin, cela est de nouveau réalisé pour tous les clusters, ce qui donne la valeur client potentielle globale à maximiser.

Résultats de l'optimisation de la stratégie commerciale

Comme dit précédemment, la première étape consiste à segmenter le portefeuille en clusters de profils similaires en termes de valeur client future afin d'identifier les profils à priori fidèles et rentables. La valeur client future de tous les individus du portefeuille d'affaires nouvelles a donc été calculée à partir du montant initialement proposé par l'assureur. Une Classification Ascendante Hiérarchique (C.A.H.) a été utilisée pour créer les 5 clusters de l'étude. Ensuite, les valeurs clients potentielles et les probabilités de transformation moyennes de chaque cluster sont calculées selon toutes les variations possibles (entre -30% et $+30\%$), ces résultats sont répertoriés sous forme de tableaux, dont les échantillons sont disponibles ci-dessous, où une cellule correspond respectivement à la valeur client potentielle et à la transformation moyenne selon le cluster g en colonne et la variation v_g appliquée en index.

v_g \ g	**1**	**2**	**3**	**4**	**5**
-30%	23,7	16,9	7,99	-8,02	-12,6
⋮	⋮	⋮	⋮	⋮	⋮
0%	26,7	32,8	25,0	11,1	-2,4
⋮	⋮	⋮	⋮	⋮	⋮
30%	13,9	27,1	24,5	15,1	0,05

(a) Valeurs clients potentielles (VC_g pour $g \in [\![1,5]\!]$)

v_g \ g	**1**	**2**	**3**	**4**	**5**
-30%	83,8%	83,9%	72,5%	68,0%	33,1%
⋮	⋮	⋮	⋮	⋮	⋮
0%	48,4%	54,5%	50,3%	47,3%	14,8%
⋮	⋮	⋮	⋮	⋮	⋮
30%	22,2%	30,4%	32,1%	30,8%	6,3%

(b) Probabilités de transformation moyennes

TABLE 1 : Échantillons des valeurs clients potentielles et des probabilités de transformation moyennes des clusters selon les variations

Grâce à ces résultats, deux approches d'optimisation ont été réalisées. Tout d'abord une optimisation sans contraintes où il suffit simplement de sommer les valeurs clients potentielles maximales de chaque cluster (*i.e.* les maxima des colonnes du tableau 1a)) puisque, pour rappel, VC potentielle globale $= \sum_{g=1}^{G} VC_g$, où $G = 5$. Bien que cette solution maximise la valeur client potentielle du portefeuille à 101,4M € (contre 93,2M € initialement), celle-ci n'est pas idéale dans la pratique puisqu'elle implique une importante réduction de la proportion de nouveaux clients (passant de 51,32% à 38,72%), directement liée à une augmentation moyenne du montant de la prime commerciale trop élevée (+13,92% en moyenne). La mise en place des contraintes est donc nécessaire.

Les contraintes mises en place ont pour objectif de rester compétitif vis à vis de la concurrence tout en assurant un bon volume du portefeuille et des différents clusters. Les 4 contraintes choisies sont les suivantes :

▷ **Contraintes sur les clusters :**

- (C_1) Appliquer des variations comprises entre -30% et $+30\%$ au montant de la prime commerciale
- (C_2) Ne pas perdre plus de 25 points de transformation dans chaque cluster

▷ **Contraintes globales :**

- (C_3) Ne pas perdre plus de 10 points de transformation dans l'ensemble du portefeuille
- (C_4) Ne pas excéder $+5\%$ de variation moyenne du montant de la prime commerciale

Pour vérifier les contraintes sur les clusters, il suffit d'uniquement garder les cellules des tableaux précédents 1a) et 1b) qui vérifient pour tout $g \in [\![1, G]\!]$ et pour toute variation $v_g \in [-30\%, +30\%]$

(C_1)

$$(C_2) \quad \frac{1}{n_g}\sum_{i=1}^{n_g} Tr_i(v_g) \geq \frac{1}{n_g}\left(\sum_{i=1}^{n_g} Tr_i(0\%)\right) - 25\%$$

Pour vérifier les contraintes globales, une méthodologie combinatoire est mise en place. Il suffit de récupérer toutes les valeurs clients potentielles des clusters et les probabilités de transformation associées pour chaque combinaison de variations possibles sur les clusters (5 clusters et 31 variations possibles $\implies$ 31^5 combinaisons), puis d'uniquement garder les combinaisons $(v_1, ..., v_G) \in [-30\%, +30\%]^G$ qui vérifient

$$(C_3) \quad \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} Tr_i(v_g) \geq \overbrace{\frac{1}{n}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} Tr_i(0\%)\right)}^{=51,32\%} - 10\% \qquad (C_4) \quad \frac{1}{n}\sum_{g=1}^{G} n_g v_g \leq 5\%$$

Ainsi, la valeur client potentielle du portefeuille pour une combinaison correspond donc à la somme des valeurs clients potentielles des clusters associées à cette combinaison. La solution du problème d'optimisation sous contraintes est donc la plus grande valeur client potentielle globale parmi toutes les

combinaisons qui vérifient les contraintes. La combinaison de variations associée est l'allocation optimale recherchée dans cette étude. L'optimisation sous contraintes donne une valeur client potentielle maximale de 99,3M €. Bien que cette valeur soit inférieure à celle obtenue sans contraintes (101,4M €), elle permet de maintenir un taux d'acceptation moyen de 43,77% et de limiter la perte de clients, trouvant ainsi un équilibre entre rentabilité et rétention. Ainsi, la valeur client potentielle a augmenté de plus de 6 millions d'euros par rapport à la situation initiale (93,2M €), simplement en segmentant le portefeuille et en appliquant des contraintes très simples.

Enfin, le tableau 2 ci-dessous présente un récapitulatif des résultats obtenus lors des optimisations de la stratégie commerciale sans et sous contraintes, permettant une comparaison des valeurs clients potentielle globales, des probabilités de transformation moyennes et des variations appliquées pour chaque cluster.

	Valeur client potentielle globale	**Transformation moyenne**	**Variation moyenne**	v_1	v_2	v_3	v_4	v_5
Avant optimisation	93,2 M€	51,32%	0%					
Optimisation sans contrainte	101,4 M€	38,72%	+13,92%	−10%	0%	+12%	+26%	+30%
Optimisation sous contraintes	99,3 M€	43,77%	+4,9%	−16%	0%	+2%	+12%	+18%

TABLE 2 : Récapitulatif des résultats des optimisations sans et sous contraintes

Ces résultats montrent que l'optimisation sans contrainte offre, certes, une meilleure valeur client potentielle globale grâce à des variations tarifaires importantes, allant de -10% pour le cluster 1 à +30% pour le cluster 5. Cependant, ces dernières entraînent un taux de transformation moyen réduit. L'optimisation sous contraintes, bien que générant une valeur client globale légèrement inférieure, est plus équilibrée. Ses variations tarifaires restent modérées, de −16% pour le cluster 1 à +18% pour le cluster 5, permettant de préserver un taux d'acceptation moyen plus élevé. Cette approche est donc préférable, car elle garantit une meilleure attractivité tout en maintenant une amélioration significative de la rentabilité.

Conclusion

Ce mémoire propose une méthodologie innovante pour optimiser la stratégie commerciale des nouveaux contrats d'assurance automobile, en maximisant la valeur client potentielle. L'étude démontre qu'une vision à long terme, combinée à des ajustements tarifaires précis, peut accroître la rentabilité globale du portefeuille tout en maintenant une compétitivité sur un marché exigeant.

Les différentes optimisations menées dans ce mémoire ont permis d'améliorer significativement la stratégie commerciale de l'assureur. Les estimations ont montré qu'une optimisation sans contrainte maximise la valeur client potentielle, mais au prix d'une forte augmentation des primes pour certains segments, entrainant un taux d'acceptation des devis proposés moins élevé. En revanche, l'optimisation sous contraintes équilibre davantage les intérêts de l'assureur, en garantissant un bon taux de nouveaux contrats tout en maximisant les marges annuelles espérées. Une augmentation modérée des primes pour certains clusters a maximisé les marges espérées sans affecter significativement les taux d'acceptation, tandis que des baisses ciblées ont renforcé l'attractivité auprès des profils les plus prometteurs.

L'innovation de cette méthodologie d'optimisation réside dans la flexibilité qu'elle apporte grâce à la segmentation et aux contraintes modulables. En effet, une fois les valeurs client potentielles calculées pour chaque cluster et chaque variation, il devient possible de moduler à l'infini les contraintes de l'optimisation grâce à la méthode combinatoire mise en place. Ainsi, les stratégies tarifaires peuvent être ajustées de manière continue et personnalisée, sans nécessiter de recalculer les valeurs client potentielles. Cette modularité permet ainsi d'adapter rapidement les tarifs aux évolutions du marché et des objectifs commerciaux. Cette approche est également facilement opérationnelle, car elle permet

de mettre en place des ajustements tarifaires précis et modulables de manière réactive, en s'appuyant sur des calculs préétablis. Cela favorise une meilleure adaptation aux contraintes du marché sans augmenter la charge de travail liée au recalcul de la valeur client potentielle pour chaque ajustement tarifaire.

Cependant, certaines limites demeurent, notamment l'absence de données détaillées sur la sinistralité individuelle et le multi-équipement. Ce travail, limité à l'assurance automobile, pourrait être étendu à d'autres produits pour offrir une vision plus complète de la valeur client future. Enfin, bien que l'optimisation tarifaire présentée ici permette de maximiser la valeur client potentielle du portefeuille, elle pourrait soulever des questions d'ordre éthique, notamment en matière de discrimination pour certains profils jugés moins rentables. Une vigilance particulière devra être apportée pour que cette optimisation ne nuise pas aux principes de mutualisation et d'équité actuarielle qui constituent deux des fondements de l'assurance. Ces considérations devront guider l'application pratique de cette méthodologie.

Synthesis note

Context and Motivations

The non-life insurance sector, particularly auto insurance, is becoming increasingly competitive. This evolution creates a heightened need for advanced commercial strategies among insurers, especially in attracting new clients and retaining the most profitable profiles over the long term. In a context where bancassurance, digitalization, and the simplification of contract terminations facilitate policyholder mobility, companies must innovate to stand out in a saturated market. Historically, insurers have structured their pricing strategies around standardized criteria, but this approach has limitations in predicting the profitability and loyalty of each client.

In this context, this thesis proposes an innovative methodology for optimizing the commercial strategy for new business through an approach derived from the concept of "customer lifetime value." This approach enables a deeper understanding of the potential profitability of new clients beyond mere demographic criteria, by integrating retention probability and the expected margin generated for each profile. The central objective of this thesis is to propose an innovative and easily operational pricing optimization methodology to address the following challenge: how to optimize the pricing offered to new policyholders at the time of subscription to maximize the potential customer value of the entire portfolio, while maintaining a balance between competitiveness and long-term profitability?

Customer Lifetime Value

In marketing, customer lifetime value is generally defined as the net present value of the profits generated by a customer over the entire duration of their relationship with the company. Its theoretical formula is as follows

$$\text{Customer Lifetime Value of policyholder } i = \sum_{t=0}^{T_h} \left(\frac{\left[\text{Premium}_i(t) - \text{Costs}_i(t)\right] \times \mathbb{P}_i(t)}{(1+r)^t} \right) - \text{AC}_i \quad (6)$$

where ▷ $\text{Premium}_i(t)$ is the premium paid by policyholder i at time t.

- ▷ $\text{Costs}_i(t)$ represents the costs associated with policyholder i at time t (claims, overhead expenses, etc.).
- ▷ $\mathbb{P}_i(t)$ is the probability that policyholder i remains in the portfolio at time t.
- ▷ r is the discount rate (or the company's cost of capital).
- ▷ $T_h = 10$ is the chosen time horizon for estimating the customer lifetime value.
- ▷ AC_i represents the acquisition costs of the contract for policyholder i.

However, in this study focusing on new business, several data points required for this theoretical formula are not available. Notably, there is no information on the policyholders' past claims experience. Consequently, the definition of customer lifetime value has been adjusted by using the pure premium as an estimate of risk, which represents the portion of the premium corresponding to the theoretical cost of claims.

Thus, denoting by $\mathrm{CP}_i(t)$ and $\mathrm{PP}_i(t)$ the commercial premium and pure premium of policyholder i at time t, respectively, the customer lifetime value is defined in this thesis as follows

$$\text{Customer Lifetime Value of policyholder } i = \sum_{t=0}^{T_h} \frac{\overbrace{\left[\mathrm{CP}_i(t) - \mathrm{PP}_i(t)\right]}^{\mathrm{Margin}_i(t)} \times \mathbb{P}_i(t)}{(1+r)^t} \tag{7}$$

This formula is a simplified and adjusted approximation tailored to the specific context of new insurance contracts, where acquisition costs are homogenized and can be disregarded in the optimization process. The difference between the commercial premium and the pure premium represents the expected margin generated by a policyholder, without accounting for individual expenses or claims. This approach effectively identifies loyal and profitable profiles, providing a simplified yet relevant perspective on the value each policyholder could bring to the portfolio.

Thus, the customer lifetime value of a policyholder is built based on the expected margin and the probability of remaining in the portfolio, estimated respectively through a margin projection model and a survival model, specifically the Cox proportional hazards model.

Nota Bene:

> *Although the term "customer lifetime value" is used, the quantity defined in equation (7) actually represents a "contract lifetime value," as it specifically applies to auto insurance. Indeed, the calculation does not account for information related to multi-product holdings, i.e., other policies the client may have with the same company (home insurance, life insurance, etc.). This study is therefore limited to auto insurance due to the absence of data on multi-product holdings, but the methodology can easily be generalized to optimize the commercial strategy for other types of insurance products.*

Finally, this thesis introduces the concept of potential customer lifetime value, which incorporates the probability of accepting the proposed quote at the time of subscription. This probability, referred to as the transformation probability, is predicted using a transformation model. When the amount $\mathrm{CP}_i(0) \times (1+v_g)$ is offered to policyholder i at subscription, where $\mathrm{CP}_i(0)$ represents the initial amount proposed by the insurer and v_g is a variation between -30% and $+30\%$, the transformation probability is denoted as $Tr_i(v_g)$, and the potential customer lifetime value for policyholder i is defined as

$$PCLV_i(v_g) = \sum_{t=0}^{T_h} \frac{\mathrm{Margin}_i(t) \times \mathbb{P}_i(t)}{(1+r)^t} \times Tr_i(v_g) \tag{8}$$

where, for all $t \in [\![0, T_h]\!]$, the expected margins of i, $\mathrm{Margin}_i(t)$, and their survival probabilities, $\mathbb{P}_i(t)$, depend on the proposed premium amount $\mathrm{CP}_i(0) \times (1+v_g)$.

Models Used in the Thesis

To ultimately implement an optimization of the commercial strategy for new business, several models were employed to calculate the various components of the potential customer lifetime value defined in formula (8) above.

Cox Proportional Hazards Model

The Cox proportional hazards model is used to estimate the survival functions of contracts, corresponding to the probabilities $\mathbb{P}_i(t)$ that a policyholder i remains in the portfolio over time t. This model is well-suited and allows for reliable survival estimation based on the initial characteristics of the contracts.

Margin Projection Model

The margin projection model aims to estimate the expected profitability of each policyholder over a 10-year horizon by calculating the difference between commercial premiums and technical premiums

(*i.e.*, loaded pure premiums). Eighteen XGBoost models are developed to predict, year by year, the premiums based on the initial characteristics of policyholders and the predicted premiums from the previous year. Each XGBoost model is thus specific to a type of premium (pure or commercial) and a given year. By sequentially nesting the predictions, it is possible to obtain all premiums for each over 10 years using only their initial characteristics and premiums at subscription, making it ideal for working with new business.

The expected annual margins are then calculated using the following formula:

$$\text{Margin}_i(t) = \text{CP}_i(t) - \text{PP}_i(t) \times (1 + \text{Loading Rate}) \tag{9}$$

Nota Bene:

> *The pure premium has been loaded, i.e., transformed into a technical premium, to calculate the expected margins, as the loadings do not represent a profit for the insurer but rather an amount intended to cover its expenses.*

The term "margin projection model" refers to this process of calculating margins over time based on a series of successive models, where each model uses the predictions of the previous one. The expected margins projected by this model are then integrated into the calculation of customer lifetime value, providing an estimate of the expected profitability for each contract.

Transformation Model

The transformation model, based on logistic regression, is used to estimate the probability of accepting the proposed quote at the time of subscription. It evaluates the influence of variations in the commercial premium, as well as the individual characteristics of policyholders (age, type of vehicle, etc.), on the decision of potential policyholders to accept the contract.

The predicted probability of acceptance, referred to as the transformation probability, is multiplied by the customer lifetime value to obtain the potential customer lifetime value for each policyholder. This model helps guide commercial strategies by anticipating the impact of premium variations on client enrollment decisions, thereby enabling the optimization of the new business portfolio composition.

Optimization of the Commercial Strategy

Approach

As mentioned earlier, the objective of this thesis is to propose an innovative methodology for optimizing the commercial strategy for new business in auto insurance. The methodology adopted in this thesis aims to optimize the commercial premiums offered at subscription to maximize the overall customer lifetime value. The first step involves segmenting the portfolio into clusters of profiles with similar characteristics in terms of customer lifetime value, calculated for each based on the initial premium proposed by the insurer.

Segmenting a portfolio into clusters of similar profiles allows for a more comprehensive analysis of customer behaviors and needs. This approach helps identify trends, anticipate behaviors, and adjust marketing strategies accordingly. It also enables personalized actions based on the segments, creating tailored offers that enhance loyalty and satisfaction. Finally, since the goal is to identify loyal and profitable profiles, segmenting based on customer lifetime value provides an additional advantage by focusing commercial efforts on customers who appear to be the most promising in the long term.

Using the notations below, the optimization problem addressed in this thesis is mathematically expressed in equation (10):

- $G \geq 1$: the number of (distinct) clusters forming the portfolio
- $v_g \in [-30\%, +30\%]$: the variation applied to the premium amount for cluster $g \in [\![1, G]\!]$

- ▷ n_g: the number of s in cluster $g \in [\![1, G]\!]$
- ▷ $\mathbb{P}_i(t)$: the probability that i remains in the portfolio at time t
- ▷ $\text{Margin}_i(t)$: the expected margin of i at time t
- ▷ r: the constant discount rate, equal to 2%
- ▷ $Tr_i(v_g)$: the transformation probability of i as a function of the variation v_g
- ▷ $PCLV_i(v_g)$: the potential customer lifetime value of i at horizon T_h, as a function of the variation applied to the initial premium amount

$$\begin{cases} \max\limits_{v_1,\dots,v_G} \text{Global Potential } CLV = \sum\limits_{g=1}^{G} PCLV_g \\ \text{Subject to constraints} \end{cases} \quad \text{where} \quad PCLV_g = \sum_{i=1}^{n_g} \overbrace{\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Margin}_i(t)}{(1+r)^t} \times Tr_i(v_g)}^{PCLV_i(v_g)} \tag{10}$$

The diagram in Figure 2 summarizes the process implemented to calculate the global potential customer lifetime value that we aim to maximize.

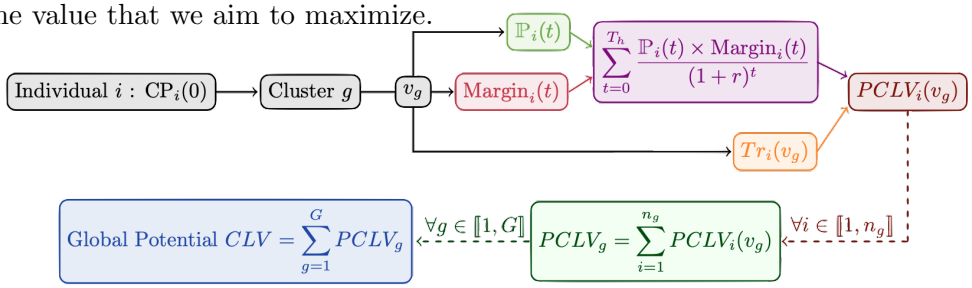


Figure 2: Summary diagram of the process for calculating the global potential customer lifetime value

The process is as follows: the insurer calculates a commercial premium for an individual i, denoted as $\text{CP}_i(0)$. Based on this amount and the initial characteristics of their contract, the individual is assigned to a cluster g of profiles similar in terms of customer lifetime value. A variation v_g, common to all individuals in cluster g, is applied to the premium amount $\text{CP}_i(0)$, which then becomes $\text{CP}_i(0) \times (1 + v_g)$. From this adjusted premium, for all times $t \in [\![0, T_h]\!]$, the expected margins ($\text{Margin}_i(t)$) and contract survival probabilities ($\mathbb{P}_i(t)$) are predicted to calculate a new customer lifetime value for individual i ($\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Margin}_i(t)}{(1+r)^t}$). The probability of accepting the quote at this new price ($Tr_i(v_g)$) is also calculated, and it is multiplied by the new customer lifetime value of individual i to obtain their potential customer lifetime value ($PCLV_i(v_g)$). This process is repeated for all individuals in cluster g, and the potential customer lifetime value of the cluster ($PCLV_g$) is obtained by summing up the individual values. Finally, this is performed for all clusters, resulting in the global potential customer lifetime value, which is the target to be maximized.

Results of the Optimization of the Commercial Strategy

As mentioned earlier, the first step involves segmenting the portfolio into clusters of profiles similar in terms of customer lifetime value to identify profiles that are likely to be loyal and profitable. The customer lifetime value of all individuals in the new business portfolio was calculated based on the premium initially proposed by the insurer. A Hierarchical Clustering method was used to create the 5 clusters analyzed in this study.

Subsequently, the potential customer lifetime values and the average transformation probabilities for each cluster were calculated for all possible variations (ranging from −30% to +30%). These results are

presented in tables, samples of which are provided below, where each cell corresponds to the potential customer lifetime value and the average transformation probability for cluster g (in columns) and the applied variation v_g (in rows).

v_g \ g	**1**	**2**	**3**	**4**	**5**
-30%	23.7	16.9	7.99	-8.02	-12.6
⋮	⋮	⋮	⋮	⋮	⋮
0%	26.7	32.8	25.0	11.1	-2.4
⋮	⋮	⋮	⋮	⋮	⋮
30%	13.9	27.1	24.5	15.1	0.05

(a) Potential Customer Lifetime Values

v_g \ g	**1**	**2**	**3**	**4**	**5**
-30%	83.8%	83.9%	72.5%	68.0%	33.1%
⋮	⋮	⋮	⋮	⋮	⋮
0%	48.4%	54.5%	50.3%	47.3%	14.8%
⋮	⋮	⋮	⋮	⋮	⋮
30%	22.2%	30.4%	32.1%	30.8%	6.3%

(b) Average Transformation Probabilities

Table 3: Samples of Potential Customer Lifetime Values and Average Transformation Probabilities for Clusters Based on Variations

Based on these results, two optimization approaches were carried out. First, an unconstrained optimization, where the potential customer lifetime values are simply maximized by summing the maximum values for each cluster (*i.e.*, the maximum values from the columns of Table 3a), since, as a reminder, global potential $CLV = \sum_{g=1}^{G} PCLV_g$, where $G = 5$. Although this solution maximizes the portfolio's potential customer lifetime value to 101.4M € (compared to 93.2M € initially), it is not ideal in practice as it results in a significant reduction in the proportion of new clients (from 51.32% to 38.72%). This reduction is directly tied to an average increase in the commercial premium that is too high (+13.92% on average). Therefore, the introduction of constraints is necessary.

The constraints implemented aim to remain competitive in the market while ensuring a good portfolio volume and balanced distribution across clusters. The four chosen constraints are as follows:

▷ **Cluster Constraints:**

- (C_1) Apply variations between -30% and $+30\%$ to the commercial premium amount.
- (C_2) Do not lose more than 25 points of transformation in any cluster.

▷ **Global Constraints:**

- (C_3) Do not lose more than 10 points of transformation in the overall portfolio.
- (C_4) Do not exceed a $+5\%$ average variation in the commercial premium amount.

To verify the cluster constraints, it is sufficient to retain only the cells from the previous tables (3a) and (3b) that satisfy, for all $g \in [\![1, G]\!]$ and for all variations $v_g \in [-30\%, +30\%]$, the constraint (C_1), and

$$(C_2) \quad \frac{1}{n_g} \sum_{i=1}^{n_g} Tr_i(v_g) \geq \frac{1}{n_g} \left(\sum_{i=1}^{n_g} Tr_i(0\%) \right) - 25\%$$

To verify the global constraints, a combinatorial methodology is implemented. It involves retrieving all potential customer lifetime values of the clusters and the associated transformation probabilities for each combination of possible variations across the clusters (5 clusters and 31 possible variations $\implies 31^5$ combinations). Then, only the combinations $(v_1, \ldots, v_G) \in [-30\%, +30\%]^G$ that satisfy

$$(C_3) \quad \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} Tr_i(v_g) \geq \overbrace{\frac{1}{n} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} Tr_i(0\%) \right)}^{=51.32\%} - 10\% \qquad (C_4) \quad \frac{1}{n} \sum_{g=1}^{G} n_g v_g \leq 5\%$$

Thus, the potential customer lifetime value of the portfolio for a given combination corresponds to the sum of the potential customer lifetime values of the clusters associated with that combination.

The solution to the constrained optimization problem is therefore the highest global potential customer lifetime value among all combinations that satisfy the constraints. The associated variation combination represents the optimal allocation sought in this study. The constrained optimization yields a maximum potential customer lifetime value of 99.3M €. Although this value is lower than that obtained without constraints (101.4M €), it ensures an average acceptance rate of 43.77% and limits customer loss, thus achieving a balance between profitability and retention. As a result, the potential customer lifetime value increased by more than 6 million euros compared to the initial situation (93.2M €), simply by segmenting the portfolio and applying very straightforward constraints.

Finally, Table 4 below presents a summary of the results obtained from the optimizations of the commercial strategy without and under constraints, allowing for a comparison of the global potential customer lifetime values, average transformation probabilities, and variations applied to each cluster.

	Potential Global Customer Lifetime Value	**Average Transformation**	**Average Variation**	v_1	v_2	v_3	v_4	v_5
Before Optimization	93.2 M€	51.32%	0%					
Optimization without Constraints	101.4 M€	38.72%	+13.92%	−10%	0%	+12%	+26%	+30%
Optimization under Constraints	99.3 M€	43.77%	+4.9%	−16%	0%	+2%	+12%	+18%

Table 4: Summary of Results for Optimizations without and under Constraints

These results show that the unconstrained optimization offers a higher global potential customer lifetime value due to significant pricing variations, ranging from −10% for cluster 1 to +30% for cluster 5. However, these variations lead to a reduced average transformation rate. The constrained optimization, while generating a slightly lower global customer lifetime value, is more balanced. Its pricing variations remain moderate, ranging from −16% for cluster 1 to +18% for cluster 5, allowing for the preservation of a higher average acceptance rate. This approach is therefore preferable, as it ensures greater attractiveness while maintaining a significant improvement in profitability.

Conclusion

This thesis proposes an innovative methodology to optimize the commercial strategy for new auto insurance contracts by maximizing potential customer lifetime value. The study demonstrates that a long-term perspective, combined with precise pricing adjustments, can enhance the overall profitability of the portfolio while maintaining competitiveness in a demanding market.

The various optimizations conducted in this thesis significantly improved the insurer's commercial strategy. The estimates showed that unconstrained optimization maximizes potential customer lifetime value but at the cost of a significant increase in premiums for certain segments, resulting in a lower acceptance rate for the proposed quotes. In contrast, constrained optimization better balances the insurer's interests by ensuring a good rate of new contracts while maximizing expected annual margins. A moderate increase in premiums for some clusters maximized expected margins without significantly affecting acceptance rates, while targeted reductions enhanced attractiveness among the most promising profiles.

The innovation of this optimization methodology lies in the flexibility it provides through segmentation and adjustable constraints. Once the potential customer lifetime values are calculated for each cluster and each variation, it becomes possible to infinitely adjust the optimization constraints using the combinatorial method implemented. Thus, pricing strategies can be continuously and individually adjusted without requiring recalculation of the potential customer lifetime values. This modularity enables a quick adaptation of pricing to market changes and commercial objectives. This approach is also easily operational, as it allows for precise and flexible pricing adjustments to be implemented reactively, relying on pre-established calculations. It facilitates better adaptation to market constraints

without increasing the workload associated with recalculating potential customer lifetime values for each pricing adjustment.

However, certain limitations remain, particularly the absence of detailed data on individual claims history and multi-product holdings. This work, limited to auto insurance, could be extended to other products to provide a more comprehensive view of future customer value. Finally, although the pricing optimization presented here allows for maximizing the potential customer lifetime value of the portfolio, it could raise ethical concerns, particularly regarding discrimination against profiles deemed less profitable. Special attention must be paid to ensure that this optimization does not undermine the principles of risk pooling and actuarial fairness, which are fundamental to insurance. These considerations should guide the practical application of this methodology.

Remerciements

Je tiens à exprimer ma plus profonde gratitude à l'ensemble des personnes qui ont contribué à l'aboutissement de ce mémoire.

Tout d'abord, je tiens à remercier chaleureusement Simon SAVOYE, mon tuteur d'entreprise, pour son accompagnement, sa confiance et tout le savoir qu'il a pu me transmettre durant ces travaux. Ses conseils avisés et son soutien constant ont été essentiels à la bonne conduite de ce mémoire. Au-delà du contexte professionnel, c'est une véritable amitié qui s'est créée.

Je remercie également Guillaume ROSOLEK, Nadhir BABA ARBI, Franck KOFFI et Nabil RACHDI pour leur bienveillance, leur expertise actuarielle, ainsi que le temps qu'ils m'ont accordé pour suivre l'évolution de mes travaux.

Mes remerciements vont aussi à tous mes collègues, Sami BECK, Cédric DENNIEL, Markéta KRÚPOVÁ, Mulah MORIAH et Margaux REGNAULT pour leurs relectures attentives et leurs remarques constructives, qui ont contribué à l'amélioration de ce mémoire. Je tiens à remercier tout particulièrement Margaux, dont les travaux ont servi de point de départ à ce mémoire permettant ainsi d'assurer la continuité de ses recherches.

Je tiens à exprimer ma reconnaissance à mon responsable de Master, Quentin GUIBERT, ainsi qu'à mon tuteur académique, Tachfine EL ALAMI, pour leurs orientations et leur suivi tout au long de ce stage.

Enfin, je souhaite remercier l'Université Paris Dauphine - PSL pour l'excellence de la formation reçue, et ADDACTIS France pour m'avoir accueilli dans le cadre de mon stage, me permettant ainsi de d'approfondir et d'appliquer mes connaissances théoriques dans le monde professionnel.

Table des matières

Introduction

Dans un marché de plus en plus concurrentiel, les assureurs doivent constamment affiner leurs stratégies commerciales pour attirer de nouveaux clients tout en maximisant leur rentabilité à long terme. L'assurance non-vie, et plus particulièrement l'assurance automobile, fait face à des pressions croissantes en matière de compétitivité tarifaire et de fidélisation des clients. Dans ce contexte, l'optimisation de la stratégie commerciale devient cruciale pour garantir la pérennité des portefeuilles de clients.

Historiquement, les assureurs ont longtemps orienté leurs stratégies vers l'acquisition d'affaires nouvelles en s'appuyant sur des barèmes tarifaires standardisés et en segmentant les clients selon des critères démographiques et comportementaux. Cependant, cette approche, bien que pertinente, présente des limites quant à l'anticipation de la rentabilité à long terme d'un client. En effet, certains individus, qui semblent attractifs à première vue, peuvent en réalité s'avérer peu rentables en raison de comportements de résiliation fréquents ou d'une sinistralité élevée.

Ce mémoire s'inscrit dans une démarche d'innovation en proposant une méthode d'optimisation de la stratégie commerciale des affaires nouvelles par une approche basée sur une inspiration d'un indicateur très utilisé en *marketing* : la valeur client future. En s'appuyant sur divers modèles prédictifs, cette méthode vise à identifier les profils de clients à priori susceptibles d'être fidèles et rentables sur le long terme. En conséquence, il devient possible de proposer des tarifs optimisés lors de la souscription, permettant de maximiser la valeur client potentielle du portefeuille tout en renforçant la compétitivité.

L'objectif de cette étude est donc de répondre à la question suivante : comment optimiser le tarif proposé à la souscription d'un contrat d'assurance pour chaque individu, de manière à maximiser la valeur client potentielle du portefeuille d'affaires nouvelles tout en maintenant un équilibre entre compétitivité et rentabilité à long terme ?

Pour répondre à cette problématique, des techniques avancées de modélisation sont appliquées sur des données provenant d'un portefeuille de nouveaux contrats d'assurance automobile. Ces modèles permettent de prédire la durée de vie des contrats, d'estimer les marges annuelles générées et de calculer les probabilités d'acceptation des différentes offres tarifaires. En combinant ces éléments, la valeur client du portefeuille peut être calculée et maximisée, selon différentes stratégies, en ajustant le montant proposé aux potentiels nouveaux assurés.

Chapitre 1

Contexte et motivation de l'étude

1.1 Cadre de l'assurance non-vie

L'assurance est un service reposant sur deux concepts qui remonteraient à des siècles : la mutualisation et le transfert de risque. Établie en France par un cadre juridique dès le XVII$^{\text{ème}}$ siècle, l'assurance a été renforcée par les progrès mathématiques dans le domaine des probabilités. Avec les évolutions économiques et la complexification des échanges, elle est devenue indispensable. Aujourd'hui, l'assurance constitue un pilier fondamental de notre structure sociale et économique.

Juridiquement, un contrat, de quelque nature qu'il soit, est défini comme un accord de volontés créant un lien de droit, qui permet à une personne (le créancier) d'exiger d'une autre (le débiteur) qu'elle réalise une prestation. Dans le cadre de l'assurance, il est question de contrat aléatoire, *i.e.* un contrat dans lequel l'étendue des contreparties réciproques n'est pas connue dès la conclusion du contrat. En effet, l'assureur s'engage à indemniser l'assuré de certains risques ou sinistres éventuels, moyennant une prime ou cotisation. Les engagements des deux parties sont formalisés par un contrat d'assurance stipulant le risque couvert, la prestation en cas de survenance de ce dernier et la prime que l'assuré verse à l'assureur pour supporter ce transfert de risque. La principale caractéristique de l'assurance, qui la différencie des autres secteurs, est l'inversion du cycle de production : l'assureur fixe et perçoit des primes avant même de connaître le montant des sinistres et des prestations qu'il devra verser. Cela exige notamment d'utiliser des techniques statistiques et probabilistes avancées. Le calcul mathématique des primes doit équilibrer deux aspects : la mutualisation et la segmentation. La mutualisation, qui consiste à regrouper un large nombre d'assurés, permet de diminuer la volatilité du risque tandis que la segmentation vise à créer des sous-groupes homogènes en termes de risque, ce qui demande une analyse plus détaillée des assurés.

L'assurance se divise en deux grandes catégories : l'assurance de personnes, dont les indemnisations sont en général forfaitaires (compensation financière dont le montant est fixé à l'avance, indépendamment du préjudice réel subi), et l'assurance de biens et responsabilité, communément appélée assurance non-vie. L'assurance de personnes comprend diverses branches comme l'assurance vie et l'assurance maladie. En revanche, l'assurance de biens et responsabilité englobe des couvertures telles que l'assurance automobile, habitation et responsabilité civile. Elle permet de recevoir une indemnisation en cas de survenance d'un évènement aléatoire appelé sinistre. La figure 1.1 illustre la répartition des cotisations perçues par les assureurs en France en 2022 d'après une étude statistique de France Assureurs (2023). Il apparaît clairement que l'assurance de biens et responsabilité constitue un peu plus d'un quart de l'assurance en France et que l'assurance automobile, sur laquelle se portera les travaux de ce mémoire, ne représente que 10,4% de l'ensemble des cotisations en 2022.

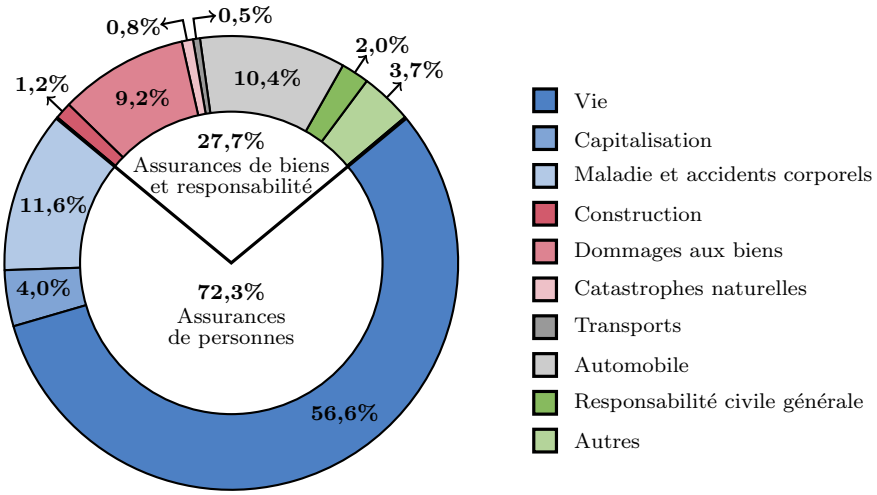


FIGURE 1.1 : Parts de cotisations en fonction des différentes branches d'assurance en France en 2022 (FRANCE ASSUREURS (2023))

L'assurance automobile (hors flotte) est tout de même la branche majoritaire de l'assurance de biens et responsabilité avec 54,3% des cotisations des particuliers en 2022, suivie de l'assurance des biens des particuliers à 31% (FRANCE ASSUREURS (2023)).

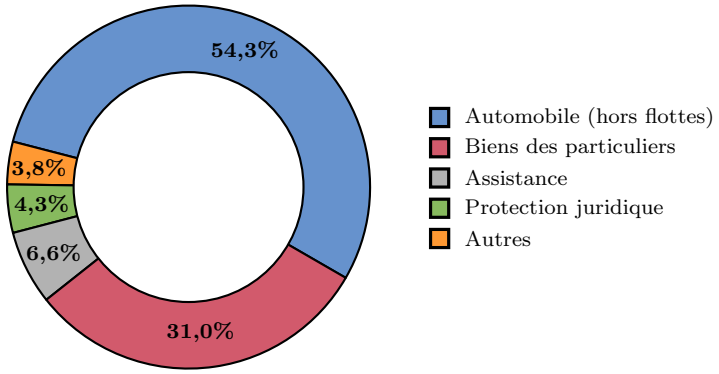


FIGURE 1.2 : Parts de cotisations des particuliers en assurance de biens et de responsabilité en 2022 (FRANCE ASSUREURS (2023))

La branche de l'assurance automobile, de même que les autres branches de l'assurance, est en constante évolution, entrainant une concurrence de plus en plus accrue. Cette dernière peut être expliquée par 3 phénomènes majeurs.

Premièrement, l'équilibre précédemment établi vient être perturbé par l'arrivée de nouveaux acteurs sur le marché tels que les bancassureurs. En effet, depuis les années 1990, les bancassureurs captent de plus en plus de parts de marché, passant d'environ 10% des cotisations en assurance de biens et responsabilité en 2010, à près de 16% douze ans plus tard, comme le montre la figure 1.3 :

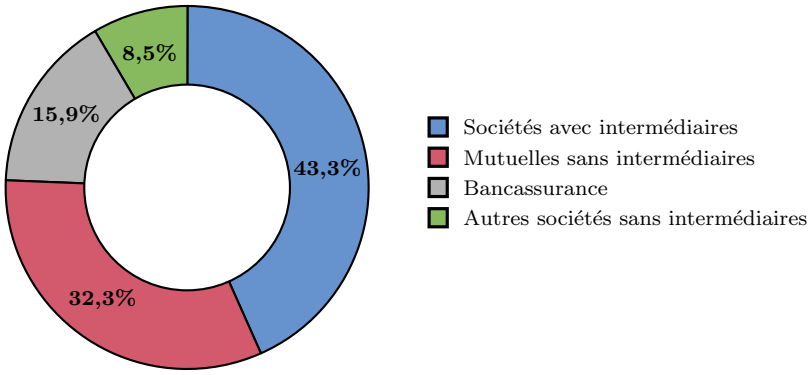


Figure 1.3 : Parts de cotisations des par mode de distribution pour l'assurance automobile en 2022 (France Assureurs (2023))

Deuxièmement, la digitalisation de la société impacte énormément le secteur de l'assurance. En plus de l'augmentation du nombre d'offres dématérialisées, les clients bénéficient désormais d'un accès à l'information de plus en plus simple et rapide. Par exemple, les sites de comparateurs d'assurance permettent aux assurés potentiels de confronter les différentes offres et de choisir la plus attractive. Les assureurs doivent s'adapter en conséquence à cette nouvelle transparence et proposer des prix d'entrée encore plus compétitifs.

Enfin, l'allègement des procédures de résiliation, et les assouplissements législatifs de manière générale, contribuent à cette accroissement de la concurrence entre les assureurs. Depuis 2008, la loi Châtel oblige les assureurs de notifier les assurés du renouvellement tacite de leur contrat dans les semaines qui précèdent la date limite de résiliation. À celle-ci s'est ajoutée en 2015 la loi Hamon qui permet aux assurés de résilier leur contrat d'assurance n'importe quand après au moins une année complète d'assurance. Enfin, la loi n°2022-1158 du 16 août 2022 portant sur des mesures d'urgence pour la protection du pouvoir d'achat simplifie les modalités de résiliation en ligne.

Dans ce contexte de concurrence, il devient crucial pour les différents acteurs de fidéliser les meilleurs assurés en termes de risque, de longévité et de rentabilité dans le portefeuille et d'attirer les meilleurs clients potentiels. La prédiction du comportement des assurés s'avère être un atout considérable pour le bon fonctionnement économique des assureurs. Ce mémoire s'inscrit dans cette perspective, avec pour objectif d'identifier les profils rentables à acquérir et optimiser le montant de cotisation à proposer à ces profils lors de leur souscription d'un contrat d'assurance automobile. À cette fin, une variante d'un indicateur largement utilisé dans le domaine du *marketing* et de la gestion d'entreprise sera employée : la valeur client future.

1.2 Valeur client future : définition

Le principal intérêt de la valeur client future réside dans sa capacité à aider les entreprises à segmenter leurs clients, allouer les ressources de manière plus efficace et formuler des stratégies centrées sur la fidélisation des clients rentables. Par exemple, les entreprises peuvent utiliser la valeur client pour identifier les clients les plus lucratifs et leur offrir des services personnalisés ou des offres exclusives, augmentant ainsi leur satisfaction et leur fidélité. De plus, la valeur client est devenue cruciale pour évaluer la performance des stratégies de *marketing* relationnel. En se concentrant sur la valeur à long terme plutôt que sur les transactions individuelles, les compagnies peuvent mieux comprendre et anticiper les comportements futurs des clients, optimisant ainsi leurs efforts commerciaux et maximisant leur retour sur investissement.

La valeur client future est généralement définie comme la valeur actuelle nette des profits futurs générés par un assuré. En d'autres termes, il s'agit de l'ensemble des profits générés par un assuré tout au long de sa relation avec la compagnie d'assurance (KUMAR, RAMANI et BOHLING (2004)). Bien que son approche soit similaire à celle des flux de trésorerie actualisés (*discounted cash flows*) utilisés en finance, 2 différences sont à noter. Tout d'abord, la valeur client future est définie et calculée pour chaque individu ce qui permet de différencier les profils de risque les plus rentables. Enfin, contrairement au domaine financier, la valeur client incorpore une probabilité pour les assurés de quitter le portefeuille à l'avenir pour se tourner vers des concurrents.

Différentes approches pour évaluer la valeur client peuvent être trouvées dans la littérature (BERGER et NASR (1998), DWYER (1997)). Toutefois, l'attention sera portée sur la définition donnée par GUPTA, LEHMANN et STUART (2004), ainsi que REINARTZ et KUMAR (2003). Selon eux, la valeur client future de l'assuré i, pour $i \in [\![1, n]\!]$, avec n le nombre d'individus du portefeuille, est définie comme

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \left(\frac{\left[\text{Prime}_i(t) - \text{Coûts}_i(t)\right] \times \mathbb{P}_i(t)}{(1+r)^t} \right) - \text{CA}_i \tag{1.1}$$

où
- ▷ $\text{Prime}_i(t)$ est la prime payée par l'assuré i au temps t
- ▷ $\text{Coûts}_i(t)$ représente les coûts de l'assuré i au temps t (sinistres, frais généraux...)
- ▷ $\mathbb{P}_i(t)$ est la probabilité que l'assuré i appartienne encore au portefeuille au temps t
- ▷ r est le taux d'actualisation (ou le coût du capital de la compagnie)
- ▷ T_h est l'horizon de temps choisi pour l'estimation de la valeur client
- ▷ CA_i représente les coûts d'acquisition du contrat pour l'assuré i

Cependant, pour notre étude, une autre définition, fortement inspirée de cette valeur client future, sera utilisée. Comme les travaux se portent sur les affaires nouvelles, les informations sur la sinistralité des individus ne sont pas disponibles. La prime pure constitue donc la meilleure estimation mathématique du risque. Egalement, faute d'accès aux informations sur les coûts d'acquisition, l'hypothèse forte d'une homogénéité de ces coûts entre les individus est posée. Ceux-ci représentent ainsi juste une "translation" de la valeur client future et peuvent donc être retirés car ils ont un impact négligeable dans l'optimisation qui sera mise en place.

En notant $\text{PC}_i(t)$, $\text{PP}_i(t)$ respectivement les primes commerciale et pure de l'assuré i au temps t, $\mathbb{P}_i(t)$ la probabilité que l'assuré i soit encore en portefeuille au temps t, r le taux d'actualisation et T_h l'horizon de temps choisi, la valeur client future est définie de la manière suivante dans ce mémoire

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \frac{\overbrace{\left[\text{PC}_i(t) - \text{PP}_i(t)\right]}^{\text{Marge}_i(t)} \times \mathbb{P}_i(t)}{(1+r)^t} \tag{1.2}$$

Cette définition inspirée de la valeur client future (1.1) doit être considérée avec prudence et n'est en théorie que spécifiquement applicable à notre étude. En effet, celle-ci tient compte d'une sinistralité lissée sur l'ensemble du portefeuille via la prime pure et non d'une sinistralité individuelle. De plus, la prime commerciale reflète certes les chargements, mais également la stratégie mise en place par la compagnie d'assurance pour chaque profil, qui peut alors choisir de marger fortement sur un segment précis de la population, tout en étant très compétitif sur un autre segment pour conserver un équilibre de compétitivité. C'est pourquoi, soustraire la prime pure à la prime commerciale, permet non pas de refléter la marge générée par un assuré mais plutôt sa marge espérée (ou anticipée). Cela permet tout de même d'identifier quels sont les segments de profils ciblés par la compagnie.

Ainsi, deux quantités à modéliser émergent de cette formule (1.2) : la probabilité de rester en portefeuille (probabilité de rétention) et la marge espérée. Elles seront respectivement estimées grâce à un modèle de durée de vie (plus précisément un modèle à risques proportionnels de Cox) et un modèle de projection de marges (via XGBoost), développés dans les chapitres 2 et 3.

Pour la suite des travaux, l'expression "valeur client future" sera utilisée, par simplification, pour se référer à cette variante de la notion de valeur client future définie à l'équation (1.2).

1.3 Optimisation de la stratégie commerciale : présentation de la démarche

Toujours dans un but de fidélisation et de compétitivité sur le marché de l'assurance, une optimisation de la stratégie commerciale de l'assureur sera développée. Le principe de celle-ci est de trouver le montant de cotisation optimal à proposer aux individus lors de leur souscription à un contrat d'assurance automobile afin de maximiser la valeur client du portefeuille. En théorie, une optimisation tête par tête pourrait être appliquée. Cependant, cela ne permettrait pas de segmenter le portefeuille en différents profils similaires, ce qui est essentiel pour mieux cibler les profils rentables à fidéliser.

L'objectif est donc de segmenter le portefeuille en différents clusters de profils similaires auxquelles des variations seront appliquées à la cotisation annuelle proposée à l'entrée dans le portefeuille. En conséquence, certains clients accepteront d'entrer dans le portefeuille à ce nouveau prix, tandis que d'autres ne le feront pas (cf. Chapitre 4). Une valeur client, non plus future mais désormais potentielle, pourra alors être calculée pour chaque cluster, puis au niveau global, permettant de comparer les différentes stratégies commerciales afin de maximiser cette valeur client potentielle globale.

Mathématiquement, cette optimisation peut être exprimée de la manière suivante, soient

- $G \geq 1$ le nombre de clusters (distincts) formant le portefeuille
- $v_g \in [-30\%, +30\%]$ la variation appliquée au montant de la cotisation pour le cluster $g \in [\![1, G]\!]$
- n_g le nombre d'individus dans le cluster $g \in [\![1, G]\!]$
- $\mathbb{P}_i(t)$ la probabilité que l'individu i appartiennent encore au portefeuille au temps t
- $\text{Marge}_i(t)$ la marge espérée de l'individu i au temps t
- r le taux d'actualisation constant et égal à 2%
- $Tr_i(v_g)$ la probabilité de transformation de l'individu i en fonction de la variation v_g
- $VC_i(v_g)$ la valeur client potentielle de l'individu i à un horizon T_h en fonction de la variation appliquée au montant de la cotisation en entrée

Ainsi, en notant VC_g la valeur client potentielle du g$^{\text{ème}}$ cluster, l'objectif est de résoudre le problème d'optimisation sous contraintes suivant

$$\begin{cases} \max\limits_{v_1,\dots,v_G} \; VC \text{ potentielle globale} = \sum\limits_{g=1}^{G} VC_g \\ \\ \text{Sous contraintes} \end{cases} \tag{1.3}$$

$$\text{avec} \quad VC_g = \sum_{i=1}^{n_g} \underbrace{\sum_{t=0}^{T_h} \frac{\overbrace{\mathbb{P}_i(t)}^{\text{Chapitre 2}} \times \overbrace{\text{Marge}_i(t)}^{\text{Chapitre 3}}}{(1+r)^t} \times \overbrace{Tr_i(v_g)}^{\text{Chapitre 4}}}_{VC_i(v_g)}$$

Chaque élément de la valeur client potentielle du $g^{\text{ème}}$ cluster VC_g seront développés dans les chapitres 2, 3 et 4.

Les contraintes permettent de fixer des seuils pour rester compétitif. Elle peuvent être globales comme spécifiques aux clusters. Par exemple, une restriction peut être imposée pour limiter la hausse du montant de la prime commerciale à 5% en moyenne sur tout le portefeuille. Pour cela, la variation appliquée peut être augmentée dans certains clusters et réduite dans d'autres afin de parvenir à un équilibre. Un autre exemple de contraintes croisées entre les clusters et le portefeuille serait de ne pas excéder les 25% de pertes de clients par clusters tout en n'excédant pas les 10% de pertes dans l'ensemble du portefeuille.

La sélection des clusters et les différentes contraintes seront développées dans le chapitre 5.

1.4 Description des données

Comme mentionné dans la section 1.1, notre étude se concentre sur l'assurance automobile. Il est important de différencier deux catégories dans ce domaine : l'assurance des flottes de véhicules et celle des particuliers. L'assurance des flottes, généralement souscrite par des entreprises, couvre plusieurs véhicules sous un même contrat. Dans ce cas, l'entreprise, en tant que personne morale, souscrit le contrat qui protège l'ensemble de ses véhicules. À contrario, un contrat souscrit par un particulier, en tant que personne physique, ne couvre qu'un seul véhicule. Ces deux types d'assurance diffèrent par leur nature et leurs spécificités. Les tarifs appliqués et le comportement des clients varient énormément, rendant nécessaire une analyse séparée et spécifique pour ces deux types d'assurance automobile. Ce mémoire est focalisé sur les contrats des particuliers uniquement.

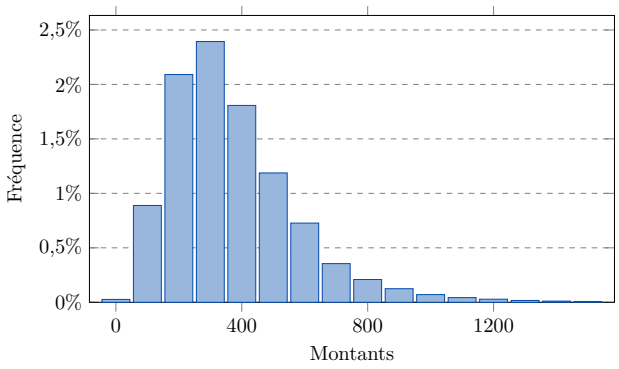
Les travaux ont été réalisés sur une base de données d'un organisme d'assurance utilisée par le cabinet à des fins de recherche et développement. Il s'agit d'une base de contrats d'assurance automobile pour particuliers comprenant des informations sur les assurés, leur comportement (par exemple le coefficient Bonus-Malus), leurs véhicules et leurs montants de cotisation annuelle. Cette base de contrats contenant les informations classiques a été complétée, d'une part du véhiculier de l'association SRA (Sécurité Réparation Automobile) qui renseigne plusieurs caractéristiques techniques précises sur les véhicules assurés, et d'autre part d'un zonier donnant des informations inhérentes au lieu de résidence des assurés. Le portefeuille mis à disposition contient environ 1,4 millions de contrats répartis sur plus de 5,2 millions de lignes sur une période de 13 ans : de 2009 à 2021. Une base d'affaires nouvelles, contenant environ 450 000 contrats, a également été extraite de cette base, et servira de base d'étude pour la suite des travaux. Cependant, aucune information sur la sinistralité passée des individus n'est disponible.

Ces données ont déjà été nettoyées et retraitées dans de précédents travaux de R&D (REGNAULT (2023)). De plus, dans la base des affaires nouvelles, des données de durée de vie des contrats et des indicateurs de résiliation et censure à droite (cf. sous-section 2.1.2) sont également présents. Ces différentes informations seront expliquées plus tard et permettront de modéliser la durée de vie des contrats via un modèle à risques proportionnels de Cox (cf. Chapitre 2).

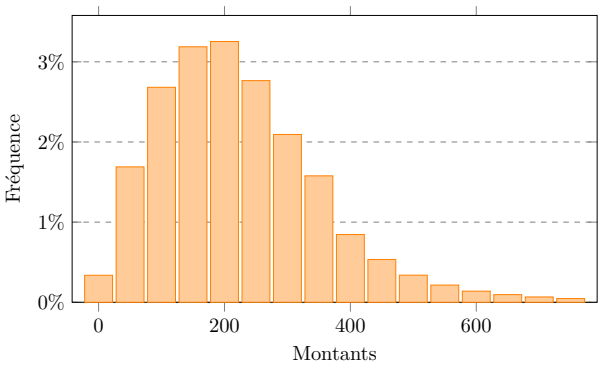
1.4.1 Analyse statistique du portefeuille

La base complète permet de suivre avec précision la structure du portefeuille dans le temps et ainsi de comprendre le comportement du porteur de risque, *i.e.* l'assureur, en matière de stratégie. Cette base est initialement composée de plus 5,2 millions de lignes et 167 variables. Pour chaque contrat, une ligne par année d'exercice est disponible, contenant les primes pures et commerciales d'une année

complète d'assurance ainsi que toutes les informations relatives à l'assuré. Cependant comme dit plus tôt, aucune base sur la sinistralité passée n'est disponible. La prime pure constitue donc la meilleure estimation mathématique du risque porté par les assurés. Voici ci-dessous les distributions des primes commerciales et pures dans l'ensemble du portefeuille représentées sous forme d'histogrammes de densité :



(a) Histogramme de densité de la prime commerciale

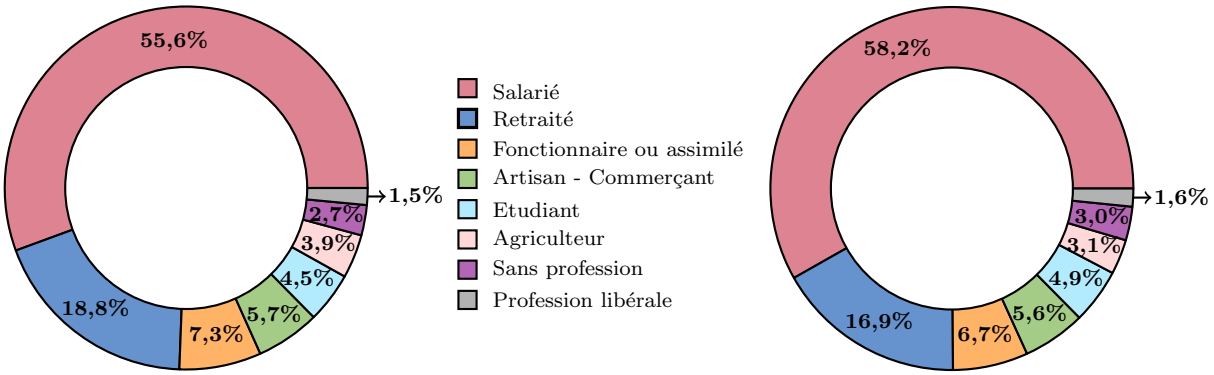


(b) Histogramme de densité de la prime pure

Figure 1.4 : Histogrammes de densité des primes commerciales et pures dans l'ensemble du portefeuille

Ces deux graphes permettent de visualiser les différences de distribution entre les montants payés par les assurés et leur coûts théoriques.

La répartition de certaines variables importantes des contrats est maintenant examinée dans l'ensemble du portefeuille mais aussi dans le sous-ensemble utilisé pour l'étude, c'est-à-dire la base des affaires nouvelles. L'analyse commence par les différentes catégories socio-professionnelles présentes dans le portefeuille.



(a) Ensemble du portefeuille

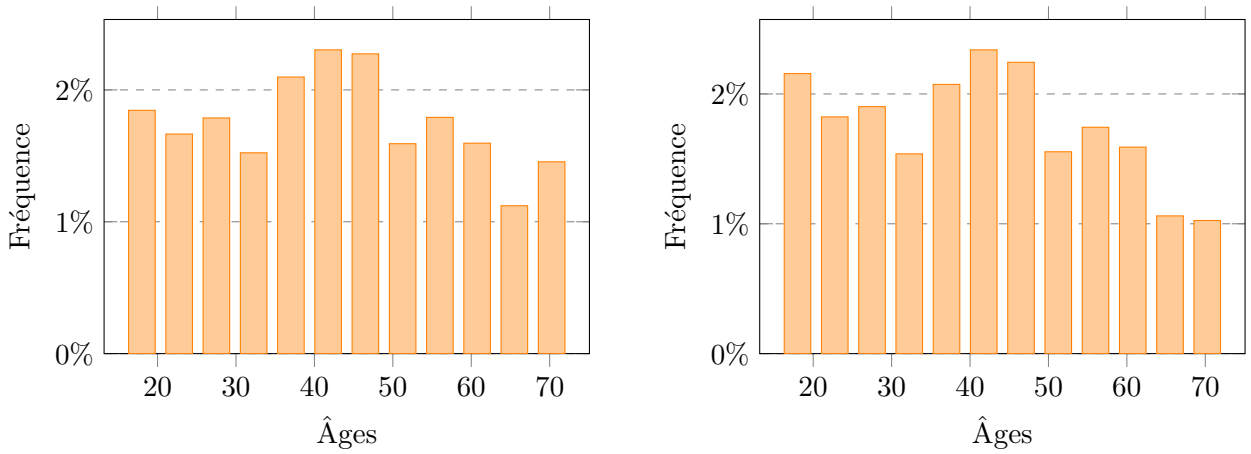
(b) Base des affaires nouvelles

Figure 1.5 : Répartition des différentes catégories socio-professionnelles dans l'ensemble du portefeuille et dans la base des affaires nouvelles

La figure 1.5 représente la répartition des catégories socio-professionnelles dans le portefeuille et dans la base des affaires nouvelles. Il est clair que la base des affaires nouvelles possède une répartition similaire au portefeuille complet. Plus de la moitié des individus sont des salariés, suivi de plus de 15% de retraités. Le reste du portefeuille est composé de fonctionnaires, d'artisans-commerçants, d'étudiants, d'agriculteurs, de personnes exerçant une profession libérale et des personnes sans activité. Dans l'ensemble, chaque catégorie est suffisamment bien représentée pour pouvoir être analysée en

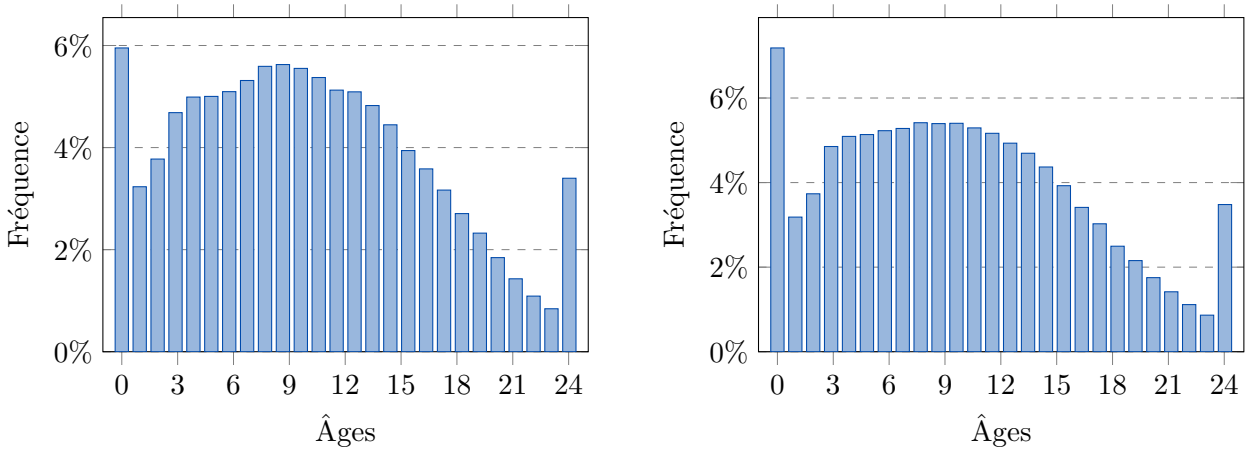
détails pendant l'étude.

De la même façon, les figures 1.6 et 1.7 représentent respectivement les histogrammes des âges des conducteurs et des anciennetés des véhicules assurés selon la base. Dans l'ensemble, le portefeuille et la base d'étude présentent une répartition homogène des âges des assurés. Il en est de même pour les véhicules bien qu'une décroissance soit visible à partir de 10 ans d'ancienneté. Les âges des assurés ont été bornés à 75 ans tandis que ceux des véhicules à 25 ans. Ceci explique la dernière barre des histogrammes de la figure 1.7. Enfin, on peut également noter qu'énormément de véhicules assurés sont neufs, avec près de 6% dans le portefeuille et 7% dans la base des affaires nouvelles.



(a) Ensemble du portefeuille (b) Base des affaires nouvelles

FIGURE 1.6 : Histogrammes des âges des conducteurs dans l'ensemble du portefeuille et dans la base des affaires nouvelles

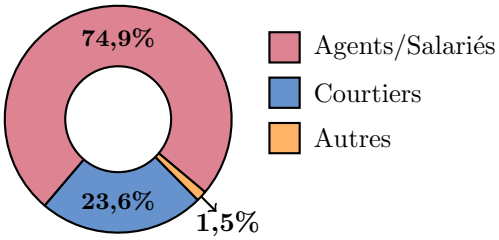


(a) Ensemble du portefeuille (b) Base des affaires nouvelles

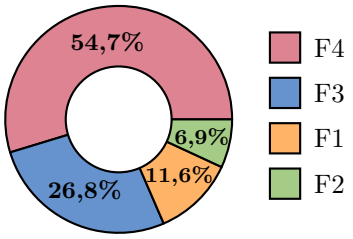
FIGURE 1.7 : Histogrammes des anciennetés des véhicules dans l'ensemble du portefeuille et dans la base des affaires nouvelles

Enfin, les répartitions dans le portefeuille et dans la base des affaires nouvelles de 4 variables sont représentées dans les figures 1.8 et 1.9. Dans la suite du mémoire, les impacts de ces variables sur les modèles utilisés seront étudiées. Il est clair que les répartitions entre les deux bases sont assez homogènes. Les variables représentées sont les différents réseaux de distribution, les formules pouvant être souscrites, le fractionnement de paiement et enfin le nombre d'autres contrats différents détenus

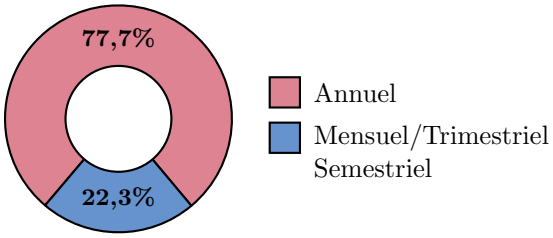
par les assurés. Les réseaux de distribution sont les moyens utilisés par les individus pour souscrire leur contrat d'assurance automobile. Ils sont répartis entre les agents généraux et salariés de la compagnie d'assurance, des courtiers d'assurance et d'autres moyens (partenaires ou internet). Les formules vont de la plus complète, *i.e.* la couverture la plus étendue, F4 à la moins complète F1. Le fractionnement de paiement est regroupé en 2 modalités, soit un seul paiement annuel de la cotisation, soit plusieurs paiements par an (tous les 1, 3 ou 6 mois). Pour finir, un seul indicateur de multi-équipement, *i.e.* le fait de posséder plusieurs contrats d'assurance auprès de ce même assureur, est disponible dans ces bases de données. Il s'agit du nombre d'autres contrats (en plus de celui étudié) dont disposent les assurés de ce portefeuille. Ainsi, les clients en affaires nouvelles n'ayant aucun autre contrat sont considérés comme étant en "conquête" pour l'assureur.



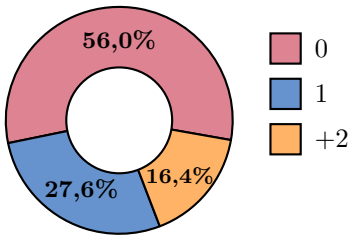
(a) Parts des différents réseaux de distribution



(b) Parts des différentes formules souscrites



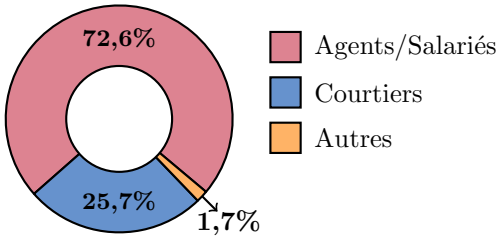
(c) Parts des différents fractionnements de paiement



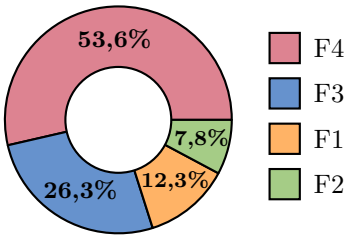
(d) Parts du nombre d'autres contrats par assuré

Figure 1.8 : Répartition des modalités de plusieurs variables de l'ensemble du portefeuille

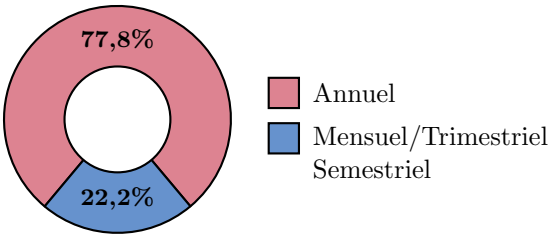
Maintenant que le contexte et les différents concepts ont été introduits, il est convient d'aborder la modélisation des probabilités de rétention dans le portefeuille au fil du temps, en utilisant des modèles de durée de vie.



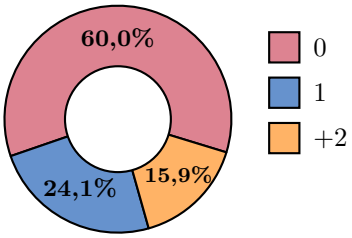
(a) Parts des différents réseaux de distribution



(b) Parts des différentes formules souscrites



(c) Parts des différents fractionnements de paiement



(d) Parts du nombre d'autres contrats par assuré

FIGURE 1.9 : Répartition des modalités de plusieurs variables de la base des affaires nouvelles

Chapitre 2

Modèles de durée de vie

Ce chapitre explore les modèles de durée de vie, et plus particulièrement le modèle à risques proportionnels de Cox, afin de modéliser la probabilité que l'assuré i appartienne encore au portefeuille au temps t, noté $\mathbb{P}_i(t)$ dans la formule suivante de la valeur client potentielle du $g^{\text{ème}}$ cluster

$$VC_g = \sum_{i=1}^{n_g} \sum_{t=0}^{T_h} \frac{\boxed{\mathbb{P}_i(t)} \times \text{Marge}_i(t)}{(1+r)^t} \times Tr_i(v_g)$$

Les modèles de durée de vie consistent en la modélisation et l'analyse d'une variable aléatoire T à valeurs dans $\mathbb{R}_+$ qui représente le temps passé dans un certain état. Dans notre étude, T représente la durée de vie des contrats, c'est-à-dire le temps que chaque individu passe dans le portefeuille entre la souscription et sa résiliation ou fin d'observation.

Il est important de noter que la principale particularité des modèles de durée est de modéliser des variables aléatoires positives, la loi normale ne peut donc pas être considérée comme la loi de référence. De plus, la fonction de répartition ne sera pas la représentation de la loi utilisée pour interpréter les résultats. Pour cela, la fonction de survie et le taux de hasard (ou taux de risque instantané) définis ci-dessous seront utilisés comme références.

Enfin, la modélisation de la durée de vie est sujette à 2 phénomènes de données incomplètes : la censure et la troncature. Un modèle est dit censuré lorsque le résultat n'est observé que partiellement pour certains individus et un modèle est tronqué lorsque la variable aléatoire n'est observable que sur une sous-partie de $\mathbb{R}_+$.

2.1 Définitions et notations

2.1.1 Notations des modèles de durée

Considérons une variable aléatoire T à valeurs dans $\mathbb{R}_+$ et notons $F(t) = \mathbb{P}(T \leq t)$ pour tout $t \geq 0$ sa fonction de répartition.

Fonction de survie

La fonction de survie de T est définie comme le complémentaire de la fonction de répartition F, pour tout $t \geq 0$

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t)$$

Il est également possible d'introduire la densité associée à la durée de vie

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{d}{dt} F(t) = -\frac{d}{dt} S(t)$$

Or par définition $\int_{\mathbb{R}_+} f(u)\ du = 1$ ainsi, la fonction de survie peut aussi être exprimée directement à partir de la densité de probabilité

$$S(t) = 1 - \int_0^t f(u)\ du = \int_t^{+\infty} f(u)\ du$$

Dans le cadre de la modélisation de la durée de vie des contrats, pour une date t fixée, $S(t)$ représente la probabilité que le contrat soit encore dans le portefeuille à la date t.

Taux de hasard

Le taux de hasard, aussi appelé taux de risque instantané ou encore force de mortalité dans le cas d'études démographiques, mesure la propension instantanée de survenue d'un évèvement (résiliation dans l'étude) à un moment donné, conditionnellement au fait que cet évènement n'est pas encore survenu à ce moment-là.

Dans le cas d'une variable aléatoire T continue, le taux de hasard est défini par

$$\begin{aligned} h : \mathbb{R}_+ &\longrightarrow \mathbb{R}_+ \\ t &\longmapsto \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T > t)}{\Delta t} \end{aligned}$$

Dans notre étude, la valeur de $h(t)$ peut être interprétée comme le risque qu'un individu résilie son contrat dans un petit intervalle de temps suivant t, sachant qu'il n'a pas résilié avant t. Une valeur élevée de $h(t)$ signifie un risque élevé de résiliation immédiate, tandis qu'une valeur faible indique un risque faible.

Le taux de hasard cumulé peut également être introduit comme étant l'intégrale du taux de hasard

$$H(t) = \int_0^t h(u)\ du$$

Or par définition de la densité f, pour tout $t \geq 0$, la relation suivante s'applique

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\Delta t} = \mathbb{P}(T > t) \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T > t)}{\Delta t} = S(t)h(t)$$

D'où

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} S(t) \times \frac{1}{S(t)} = -\frac{d}{dt} \ln S(t)$$

L'intégration entre 0 et t conduit donc à l'expression suivante

$$S(t) = \exp\left(- \int_0^t h(u)\ du \right) = \exp\big(- H(t) \big)$$

De la même façon que la densité f et la fonction de répartition F, la fonction de survie et le taux de hasard définissent la loi de T de manière unique.

2.1.2 Censure et troncature

Censures

Soient T et C les durées de vie dépendant d'un même vecteur de caractéristiques.

T est dite censurée par C lorsque T n'est pas directement observé mais la variable aléatoire $Y = \inf\{T, C\}$ dans le cas de la censure à droite ou $Y = \max\{T, C\}$ dans celui de la censure à gauche.

En d'autres termes, si, pour la censure à droite (respectivement à gauche), $T \geq C$ (respectivement $T < C$), alors la variable C sera observée à la place de T. C est appelée variable de censure.

L'observation de Y est de plus complétée par celle d'une indicatrice $\delta = \mathbb{1}_{\{T<C\}}$ (resp. $\delta = \mathbb{1}_{\{T\geq C\}}$) qui indique la nature de la variable observée : résiliation ou censure.

La censure à droite est omniprésente dans les études longitudinales. Il faut distinguer :

- ▷ La censure fixe a lieu lorsque l'évènement de censure est déterministe. La valeur de C est ainsi connue à priori pour toutes les trajectoires. Les trajectoires sont soumises à une censure fixe du fait de l'arrêté des données.
- ▷ La censure aléatoire a lieu lorsque l'évènement de censure est lui aussi de nature aléatoire. Cela signifie que la valeur de C ne sera connue que pour les trajectoires effectivement censurées. La date de résiliation n'est pas connue à priori et ne sera connue que si la résiliation est survenue avant l'arrêté des données.

Dans le cas de la censure aléatoire, il est nécessaire de faire l'hypothèse de censure non-informative qui stipule l'indépendance entre les variables T et C. Les censures fixe et aléatoire se traiteront alors de la même façon.

La censure à gauche suppose la présence d'individus pour lesquels la résiliation est survenue avant le début de la période d'observation. Elle demeure néanmoins anecdotique.

Troncatures

Soient T et Q les durées de vies dépendant d'un même vecteur de caractéristiques.

T est dite tronquée par Q lorsque la variable T n'est pas observée si $T < Q$ dans le cas de la troncature à gauche ou si $T \geq Q$ dans le cas de la troncature à droite.

La troncature à gauche est présente dans les études de durée de vie d'un contrat dès lors que l'observation des trajectoires ne démarre pas systématiquement dès le début du contrat des individus. En effet, si Q correspond à l'âge (hypothétique) de souscription au contrat et T à l'âge de résiliation, alors les individus pour lesquels $T < Q$ sont ceux dont les contrats sont résiliés avant même d'avoir été comptabilisés dans l'étude. Ainsi, les contrats résiliés prématurément seront sous-représentés dans le portefeuille. La durée de vie moyenne des contrats au sein de celui-ci sera donc différente de celle de la population de contrats originale.

La troncature à droite survient, par exemple, lorsque l'on dispose, pour un type de contrat, uniquement de la base de données contenant les informations sur les contrats déjà résiliés. Ces données ne permettent pas alors d'observer les contrats dont l'âge T de résiliation est supérieur à l'âge Q atteint à la date d'arrêté des données. Cela entraîne une représentation incomplète de la durabilité réelle des contrats.

Dans le cadre du phénomène de troncature, l'information selon laquelle certaines trajectoires ne sont pas du tout observées doit logiquement être prise en compte par le biais des trajectoires observées.

Illustration de la censure et de la troncature

Dans le cadre classique des études longitudinales de durée de vie, les phénomènes de censure à droite et de troncature à gauche sont principalement observés. Pour illustrer et mieux comprendre leur impact en pratique, il est possible de s'appuyer sur une représentation des trajectoires :

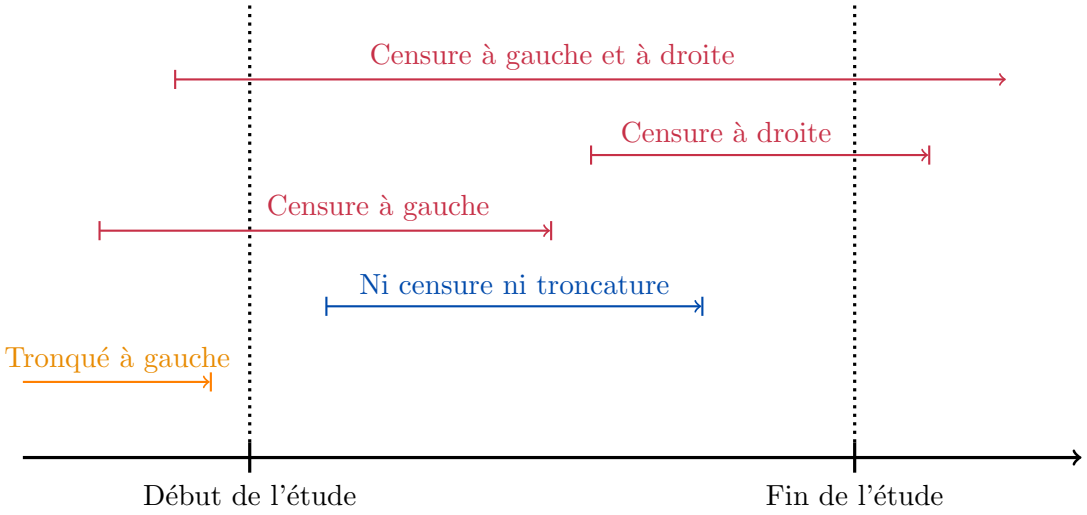


FIGURE 2.1 : Exemples illustratifs des phénomènes de censure et troncature

2.2 Estimateur non paramétrique de Kaplan-Meier

2.2.1 Définition et construction de l'estimateur

L'estimateur non paramétrique de Kaplan-Meier, introduit par KAPLAN et MEIER (1958), est une méthode statistique utilisée pour estimer la fonction de survie. Cette méthode tient compte des évèvements censurés. Montrons comment le construire.

En se plaçant dans le cas d'une censure à droite, soit $(T_1, ..., T_n) \overset{\text{i.i.d.}}{\sim} T$, où T désigne la durée de vie du contrat, et $(C_1, ..., C_n) \overset{\text{i.i.d.}}{\sim} C$, où C est la variable de censure, avec T et C indépendantes, ainsi

$$\begin{cases} Y_i = \inf\{T_i, C_i\} \\ \delta_i = \mathbb{1}_{\{T_i \leq C_i\}} \end{cases}$$

où

- ▷ Y_i est la durée observée du contrat i
- ▷ T_i est la durée de vie du contrat i
- ▷ C_i est la durée avant la censure
- ▷ δ_i est l'indicatrice de censure : si $\delta_i = 1$, alors $T_i \leq C_i$ donc tout le contrat est observé, sinon $T_i > C_i$ et le contrat n'est donc pas terminé à la date d'extraction de la base

Ainsi, le but est d'estimer la loi de T à partir de $(Y_1, ..., Y_n)$ et $(\delta_1, ..., \delta_n)$.

Pour cela, montrons qu'il est possible d'estimer le taux de hasard sans introduire de biais. Notons K la fonction de répartition de Y et K_1 la fonction de répartition (et k_1 la densité) des observations non censurées définie par

$$K_1(t) = \mathbb{P}(Y \leq t, \delta = 1) = \int_0^t \big(1 - G(u)\big) f(u)\, du \tag{2.1}$$

où G est la fonction de répartition de C et f la densité de T.

Sous l'hypothèse de censure indépendante, il s'ensuit

$$\begin{aligned}
1-K(t) &= \mathbb{P}(Y \geq t) \\
&= \mathbb{P}(\inf\{T, C\} \geq t) \\
&= \mathbb{P}(T \geq t, C \geq t) \\
&= \mathbb{P}(T \geq t) \times \mathbb{P}(C \geq t) \text{ par indépendance entre } T \text{ et } C \\
1-K(t) &= S(t)\big(1-G(t)\big) \qquad (2.2)
\end{aligned}$$

Or en dérivant l'équation (2.1), il en découle

$$\begin{aligned}
& k_1(t) = \big(1-G(t)\big)f(t) \\
\iff & \frac{k_1(t)}{1-K(t)} = \frac{\big(1-G(t)\big)f(t)}{1-K(t)} \\
\iff & \frac{k_1(t)}{1-K(t)} = \frac{f(t)}{S(t)} \text{ d'après (2.2)} \\
\iff & h(t) = \frac{k_1(t)}{1-K(t)} = \frac{\lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y < t+\Delta t, \delta = 1)}{\Delta t}}{\mathbb{P}(Y \geq t)} \\
\iff & \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t+\Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq Y < t+\Delta t, \delta = 1 | Y \geq t)}{\Delta t} \qquad (2.3)
\end{aligned}$$

En ordonnant les individus dans l'ordre croissant des temps observés $(Y_1, ..., Y_n)$, nous avons $(Y_{(1)} < ... < Y_{(\ell)})$, $\ell \leq n$. Le taux de hasard au temps $Y_{(i)}$ est estimé par $\dfrac{d_i}{r_i}$ où d_i et r_i* représentent respectivement le nombre de résiliations observées et le nombre d'individus à risque au temps $Y_{(i)}$.

Ainsi, $\dfrac{d_i}{r_i}$ est un estimateur de $\mathbb{P}(t \leq Y < t+\Delta t, \delta = 1 | Y \geq t)\big|_{t=T_{(i)}}$ dans l'équation (2.3).

L'estimateur de Kaplan-Meier repose sur l'hypothèse que rester dans le portefeuille après un temps t équivaut à être dans le portefeuille juste avant le temps t et à ne pas résilier au temps t. Ainsi, avec l'échéancier $0 = t_{(0)} < t_{(1)} < ... < t_{(n)}$, pour $i \in [\![1, n]\!]$ il apparaît que

$$\begin{aligned}
S(t_{(i)}) &= \mathbb{P}(T > t_{(i)}) \\
&= \mathbb{P}(T > t_{(i)} | T > t_{(i-1)}) \times \mathbb{P}(T > t_{(i-1)}) \\
S(t_{(i)}) &= p_i \times S(t_{(i-1)}) \qquad (2.4)
\end{aligned}$$

et $p_i = 1 - \mathbb{P}(T \leq t_{(i)} | T > t_{(i-1)})$ est estimé par

$$\widehat{p}_i = 1 - \frac{d_i}{r_i} \text{ où } \begin{cases} d_i \text{ représente le nombre de résiliations observées au temps } t_{(i)} \\ r_i \text{ représente le nombre d'individus à risque au temps } t_{(i)} \end{cases} \qquad (2.5)$$

D'où par récurrence directe, l'expression de l'estimateur non paramétrique de Kaplan-Meier est définie par la fonction en escalier suivante

$$\widehat{S}^{KM}(t) = \prod_{t_j < t} \widehat{p}_j = \prod_{t_j < t} \left(1 - \frac{d_j}{r_j}\right) \qquad (2.6)$$

*Il est possible d'écrire $r_i = \sum_{j=1}^{n} \mathbb{1}_{\{Y_{(i)} \leq Y_j\}}$

Nota Bene :
> *L'estimateur de Kaplan-Meier possède plusieurs propriétés developpées en annexe A.1.*

2.3 Modèle à risques proportionnels de Cox

Le modèle à risques proportionnels de Cox, introduit par COX (1972), est une méthode de régression utilisée dans l'analyse de survie. Le terme "risques proportionnels" signifie que le modèle suppose que les ratios des risques (ou des taux) de l'événement entre différentes valeurs des covariables sont constants dans le temps. En d'autres termes, le rapport des taux de survenue de l'événement pour deux individus reste constant quelque soit le moment regardé dans le temps. De plus, cette méthode tient compte des évènements censurés.

2.3.1 Forme du modèle

Le modèle à risques proportionnels de Cox se définit avec le taux de hasard suivant

$$h(t|X) = h_0(t)\exp(X\beta) \tag{2.7}$$

où

- ▷ X est le vecteur des covariables (caractéristiques des assurés à leur souscription)
- ▷ β est le vecteur de paramètres à estimer
- ▷ $h_0(\cdot)$ est le taux de hasard de base. Il est inconnu et indépendant de X. Il peut être interprété comme le taux de hasard d'un individu référent dont les caractéristiques sont toutes nulles (*i.e.* $X = (0\ ...\ 0)^\top$).

Nota Bene :
> $\exp(X\beta)$ *est indépendant du temps, ce terme représente le risque relatif associé à l'individu de caractéristiques* X*.*

Soit T la variable aléatoire qui modélise la durée de vie d'un contrat, de fonction de répartition $F(t) = \mathbb{P}(T \leq t) = 1 - S(t)$ (où S est la fonction de survie), et de densité f.

Or pour tout vecteur de covariables X, il est vérifié que

$$h(t|X) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \leq t)}{\Delta t} = \frac{f(t|X)}{S(t|X)} = \frac{1}{S(t|X)} \times \frac{-\partial}{\partial t} S(t|X) = -\frac{\partial}{\partial t} \ln\big(S(t|X)\big)$$

Donc en intégrant entre 0 et t, il en découle

$$-\ln\big(S(t|X)\big) = \int_0^t h(u|X)\ du \iff S(t|X) = \exp\left(-\int_0^t h(u|X)\ du\right)$$

et

$$\int_0^t h(u|X)\ du = \exp(X\beta)\int_0^t h_0(u)\ du$$

$$\iff \exp\left(-\int_0^t h(u|X)\ du\right) = \exp\left(-\exp(X\beta)\underbrace{\int_0^t h_0(u)\ du}_{=:H_0(t)}\right)$$

$$\iff S(t|X) = S_0(t)^{\exp(X\beta)} \iff S(t|X) = \exp\left(-H_0(t)\exp(X\beta)\right) \tag{2.8}$$

2.3.2 Estimation des paramètres et de la fonction de survie

Pour estimer les paramètres et la fonction de survie, il y a 3 étapes :

(1) Estimation de β via la maximisation d'une vraisemblance partielle

(2) Estimation du risque de base cumulé $H_0(t) := \int_0^t h_0(u)\, du$

(3) Estimation de la fonction de survie

(1) Vraisemblance partielle de Cox

Plaçons-nous dans le cas d'une censure à droite et posons donc $(T_1, ..., T_n) \overset{\text{i.i.d.}}{\sim} T$, où T est la durée de vie du contrat, et $(C_1, ..., C_n) \overset{\text{i.i.d.}}{\sim} C$, où C est la variable de censure, avec T et C indépendantes, ainsi

$$\begin{cases} Y_i = \inf\{T_i, C_i\} \\ \delta_i = \mathbb{1}_{\{T_i \leq C_i\}} \end{cases}$$

où

- ▷ Y_i est la durée observée du contrat i
- ▷ T_i est la durée de vie du contrat i
- ▷ C_i est la durée avant la censure
- ▷ δ_i est l'indicatrice de censure : si $\delta_i = 1$, alors $T_i \leq C_i$ donc tout le contrat est observé, sinon $T_i > C_i$ et le contrat n'est donc pas terminé à la date d'extraction de la base

$\widehat{\beta}$ s'obtient par maximisation de la vraisemblance partielle suivante

$$\mathcal{L}(x, \beta) = \prod_{Y_i, \delta_i = 1} \frac{\exp\left(x_i^\top \beta\right)}{\sum_{Y_i \leq Y_j} \exp\left(x_j^\top \beta\right)} \tag{2.9}$$

<u>Nota Bene</u> :

> *La log-vraisemblance partielle s'écrit donc* $\ell(x, \beta) = \sum_{Y_i, \delta_i = 1} x_i^\top \beta - \sum_{Y_i, \delta_i = 1} \ln \left[\sum_{Y_i \leq Y_j} \exp\left(x_j^\top \beta\right) \right]$.

Cette vraisemblance partielle ne dépend par du taux de hasard de base $h_0(\cdot)$, une estimation de β peut être réalisée sans définir $h_0(\cdot)$.

Ce n'est pas une vraisemblance mais elle se comporte de la même manière, elle dispose donc de propriétés asymptotiques similaires.

Le vecteur $\widehat{\beta}$ est ainsi estimé tel que $\dfrac{\partial \mathcal{L}}{\partial \beta}\left(x, \widehat{\beta}\right) = 0$. L'algorithme de *Newton-Raphson* défini ci-dessous permet d'obtenir une approximation de la solution.

<u>Nota Bene</u> :

> *Pour plus de simplicité, la log-vraisemblance partielle définie précédemment peut être appliquée à cet algorithme. La* $k^{\text{ème}}$ *composante de cette log-vraisemblance partielle est égale à*
>
> $$\sum_{Y_i, \delta_i = 1} x_{i,k} - \sum_{Y_i, \delta_i = 1} \frac{\sum_{Y_i \leq Y_j} x_{j,k} \exp\left(x_j^\top \beta\right)}{\sum_{Y_i \leq Y_j} \exp\left(x_j^\top \beta\right)}$$

Algorithme 1 : Algorithme de Newton-Raphson

Input :

- $\diamond$ $\dfrac{\partial \mathcal{L}}{\partial \beta} : \mathbb{R}^n \to \mathbb{R}^n$ le gradient de $\mathcal{L}$
- $\diamond$ $\dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ la matrice Hessienne de $\mathcal{L}$
- $\diamond$ $\beta_0 \in \mathbb{R}^n$ une première approximation de $\widehat{\beta}$
- $\diamond$ $\varepsilon > 0$ une précision
- $\diamond$ $k = 0$

Tant que $\left\| \dfrac{\partial \mathcal{L}}{\partial \beta}(\beta_k) \right\| > \varepsilon$, **calculer :**

- $\diamond$ $\beta_{k+1} = \beta_k - \left(\dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top}(\beta_k) \right)^{-1} \dfrac{\partial \mathcal{L}}{\partial \beta}(\beta_k)$
- $\diamond$ $k = k + 1$

Output : $\widehat{\beta} = \beta_k$

(2) Risque de base cumulé

Le risque de base cumulé $H_0(t) = \int_0^t h_0(u)\, du$ est estimé à l'aide de l'estimateur de Breslow (1972)

$$\widehat{H}_0(t) = \sum_{Y_i \leq t} \frac{\delta_i}{\sum_{Y_i \leq Y_j} \exp\left(x_j^\top \widehat{\beta}\right)} \tag{2.10}$$

(3) Fonction de survie estimée

En combinant les deux résultats précédents, voici une expression de l'estimateur de la fonction de survie de Cox

$$\widehat{S}^{Cox}(t|X) = \exp\left(-\widehat{H}_0(t) \exp\left(X\widehat{\beta}\right)\right) \tag{2.11}$$

2.3.3 Interprétation

Un avantage non négligeable du modèle à risques proportionnels de Cox est son interprétabilité des coefficients. En effet, l'analyse des ratios de hasard *(hazard ratios)*, qui mesurent le risque de survenance de l'évènement (ici résiliation du contrat) pour un individu par rapport à un autre, offre une compréhension immédiate de l'effet d'une modalité sur la durée de vie.

Soit une $X_{i,j}$ la valeur de la $j^{\text{ème}}$ covariable de l'individu i ($i \in [\![1, n]\!]$ et $j \in [\![1, p]\!]$). L'objectif est de mesurer l'effet de $X_{i,j}$ sur le risque instantané toutes choses égales par ailleurs. Pour cela, cette variable est supposée continue et le ratio de hasard est étudié quand cette variable augmente d'une unité

$$\begin{aligned} HR(X_{i,j}, X_{i,j}+1) &= \frac{h(t|X_{i,1}, ..., X_{i,j}, ..., X_{i,p})}{h(t|X_{i,1}, ..., X_{i,j}+1, ..., X_{i,p})} \\ &= \frac{h_0(t)\exp(X_{i,1}\beta_1 + ... + X_{i,j}\beta_j + ... + X_{i,p}\beta_p)}{h_0(t)\exp(X_{i,1}\beta_1 + ... + (X_{i,j}+1)\beta_j + ... + X_{i,p}\beta_p)} \\ HR(X_{i,j}, X_{i,j}+1) &= \exp(-\beta_j) \end{aligned}$$

Ainsi, 3 cas se présentent lors de l'étude des risques relatifs :

- ▷ Si $\beta_j < 0$, alors $\exp(-\beta_j) > 1$ et donc le risque que l'évènement se produise diminue (resp. augmente) quand $X_{i,j}$ augmente (resp. diminue).
- ▷ Si $\beta_j > 0$, alors $\exp(-\beta_j) < 1$ et donc le risque que l'évènement se produise diminue (resp. augmente) quand $X_{i,j}$ diminue (resp. augmente).
- ▷ Si $\beta_j = 0$, alors $\exp(-\beta_j) = 1$ et donc la variable $X_{i,j}$ n'a pas d'impact sur le risque instantané.

Les coefficients estimés de $\widehat{\beta} = \left(\widehat{\beta}_1, ..., \widehat{\beta}_p\right)$ peuvent donc être interprétés en sortie de modèle.

2.3.4 Hypothèses du modèle à risques proportionnels de Cox

Ce modèle repose sur des hypothèses qu'il faut vérifier : la log-linéarité et la proportionnalité.

Log-linéarité

Rappelons l'expression de la fonction de hasard dans le modèle à risques proportionnels de Cox

$$h(t|X) = h_0(t)\exp(X\beta)$$

D'où par application de la fonction logarithme népérien, l'expression devient

$$X\beta = \ln\left(\frac{h(t|X)}{h_0(t)}\right)$$

Ceci implique une linéarité des variables explicatives continues sur le taux de hasard. Cette hypothèse donne, par exemple pour l'âge des assurés, que le risque relatif entre deux assurés de 20 et 30 ans est le même que le risque relatif entre deux assurés de 40 et 50 ans. En effet, en prenant deux individus A et B de caractéristiques respectives X_A et X_B, la relation de linéarité du risque relatif est bien illustrée dans l'équation suivante

$$\begin{aligned} HR\left(X_A, X_B\right) &= \frac{h\big(t|X_A\big)}{h\big(t|X_B\big)} \\ &= \frac{h_0(t)\exp\left(X_A\beta\right)}{h_0(t)\exp\left(X_B\beta\right)} \\ HR\left(X_A, X_B\right) &= \exp\big((X_A - X_B)\beta\big) \end{aligned} \tag{2.12}$$

Pour vérifier cette hypothèse, il convient de faire une analyse des résidus de martingale, et sinon déterminer la fonctionnelle la plus adaptée pour satisfaire la log-linéarité.

Les modèles de survie subissent de la censure, ce qui ne permet pas de donner une expression des résidus contrairement à un modèle linéaire généralisé (dont les résidus correspondent à la différence entre les prédictions et les observations). C'est pourquoi Barlow et Prentice (1988) ont exposé une approche martingale pour obtenir les résidus des modèles présentant de la censure, reposant sur la théorie des processus de comptage. Deux ans plus tard, Therneau, Grambsch et Fleming (1990) ont présenté des approches graphiques pour vérifier l'hypothèse de log-linéarité du modèle à risques proportionnels de Cox en se basant sur les résidus de martingale. La présentation suivante reprend donc les notations des deux articles mentionnés précédemment. Soient :

- ▷ $N_i(t) = \mathbb{1}_{\{T_i \leq t\}}$ le processus de comptage égal à 1 à partir de l'instant où l'individu i sort du portefeuille
- ▷ $\lambda_i(t)$ la fonction d'intensité du processus N
- ▷ $\Lambda_i(t) = \int_0^t \lambda_i(u)\ du$ le processus (croissant) d'intensité cumulée, aussi appelé compensateur
- ▷ $Y_i(t) = 1 - N_i(t) = \mathbb{1}_{\{T_i > t\}}$ le processus indiquant si l'individu i est encore à risque au temps t

Dans le cadre du modèle à risques proportionnels de Cox, le processus Λ s'expriment à partir du processus Y et de la fonction de hasard h. En notant h_i la fonction de hasard définie par le modèle à risques proportionnels de Cox pour l'individu i de caractéristiques X_i, Λ_i s'écrit

$$\begin{aligned}\Lambda_i(t) &= \int_0^t Y_i(u)h_i(u)\ du \\ &= \int_0^t \mathbb{1}_{\{T_i>u\}}h_0(u)\exp(X_i\beta)\ du \\ \Lambda_i(t) &= \int_0^{t\wedge T_i} \exp(X_i\beta)\ dH_0(u) \end{aligned} \tag{2.13}$$

La décomposition de Doob-Meyer donne que la différence entre le processus de comptage N et le processus d'intensité cumulée Λ est une martingale (localement de carrée intégrable) dont l'expression est donc la suivante

$$M_i(t) = N_i(t) - \int_0^{t\wedge T_i} \exp(X_i\beta)\ dH_0(u) \tag{2.14}$$

Donc l'estimateur de la martingale M_i, noté $\widehat{M_i} := \widehat{M_i}(+\infty)$, pour chaque individu est égal à la différence entre le nombre de résiliations observées et le nombre de résiliations attendues par le modèle à risques proportionnels de Cox

$$\widehat{M_i} = \delta_i - \widehat{H}_0(T_i)\exp\left(X_i\widehat{\beta}\right) \tag{2.15}$$

Intuition :

> *Sous les notations du modèle à risques proportionnels de Cox, nous avons* $\mathbb{E}[H(Y)] = \mathbb{P}(T \leq C)$. *Ainsi, la définition des résidus de martingale découle de ce résultat en remarquant que pour tout individu* $i \in [\![1, n]\!]$, *nous avons* $\mathbb{E}[\delta_i - H(Y_i)|X_i] = 0$. *Il suffit donc de remplacer* H *par son estimateur défini à l'équation (2.10)grâce à l'estimateur de Breslow* (1972).

Pour valider l'hypothèse de log-linéarité, il suffit de représenter les résidus de martingale $(\widehat{M_i})_{i\in[\![1,n]\!]}$ en fonction du temps (ou une transformation du temps) pour chaque covariable numérique. Il est également possible de représenter sur le même graphique une courbe représentant l'évolution moyenne des résidus en fonction du temps. Cette courbe donne la tendance générale. Toute différence par rapport à une droite horizontale d'équation $y = 0$ représente une déviation par rapport à l'hypothèse de log-linéarité.

Voici un exemple de représentation, en figure 2.2, avec l'âge des assurés comme covariable :

Proportionnalité

Le modèle de Cox étant également connu sous le nom de *modèle à risques proportionnels de Cox* doit satisfaire une proportionnalité entre les taux de hasard.

Pour valider cette hypothèse, deux approches sont proposées : une approche graphique pour les variables qualitatives ou discrètes, et une approche par l'analyse des résidus de Schoenfeld standardisés pour les variables continues.

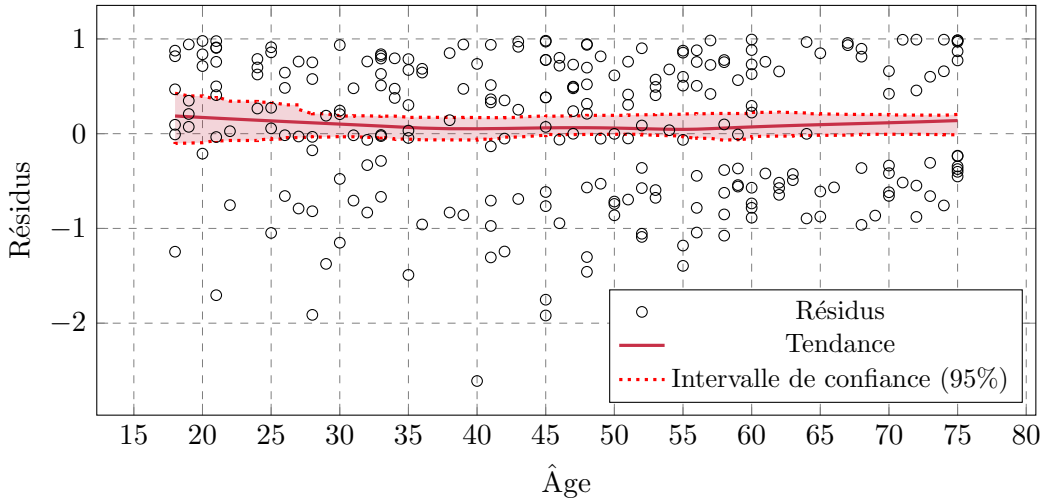


Figure 2.2 : Exemple de représentation graphique des résidus de martingale en fonction de l'âge des assurés

▷ **Méthode 1 : Représentations graphiques**

Par application de la fonction continue sur $x \in \mathbb{R}_+^* \mapsto \ln[-\ln(x)]$, la relation suivante est obtenue

$$S(t|X) = S_0(t)^{\exp(X\beta)} \iff \ln[-\ln(S(t|X))] = \ln[-\ln(S_0(t))] + X\beta$$

Soient deux individus A et B de caractéristiques différentes X_A et X_B, alors

$$\begin{cases} \ln[-\ln(S(t|X_A))] = \ln[-\ln(S_0(t))] + X_A\beta \\ \ln[-\ln(S(t|X_B))] = \ln[-\ln(S_0(t))] + X_B\beta \end{cases}$$

$$\implies \ln[-\ln(S(t|X_A))] = \underbrace{\ln[-\ln(S(t|X_B))]}_{\text{facteur de translation}} + \underbrace{(X_A - X_B)\beta}_{\text{relation linéaire}}$$

Donc l'hypothèse de proportionnalité est vérifiée si les courbes de $t \mapsto \ln[-\ln(S(t|X_A))]$ et $t \mapsto \ln[-\ln(S(t|X_B))]$, représentées dans le même plan en fonction du temps t, sont parallèles.

Un exemple de représentation est disponible dans la figure 2.3 ci-dessous :

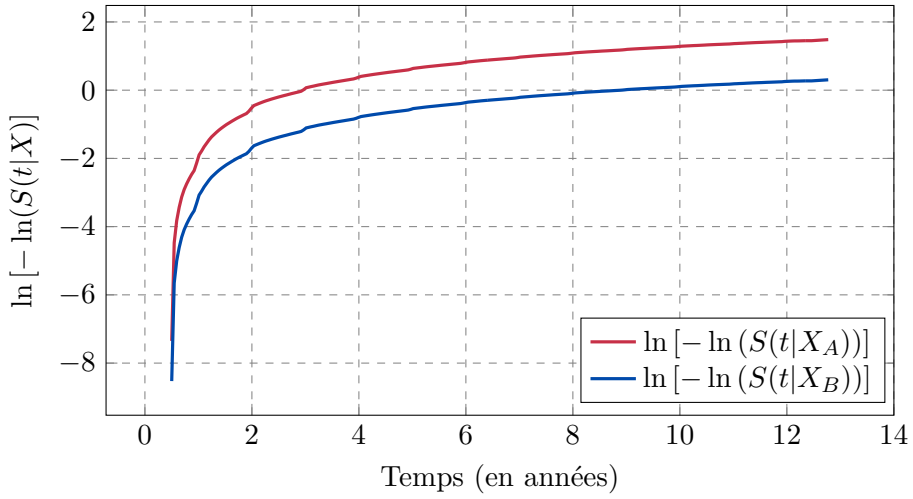


FIGURE 2.3 : Exemple de représentation de courbes d'équation $t \mapsto \ln[-\ln(S(t|X))]$

⚠ Cela ne fonctionne pas avec les variables continues qu'il faut alors segmenter en plusieurs classes. Néanmoins, cette discrétisation génère de la perte d'informations, il vaut donc mieux privilégier la seconde méthode pour ce type de variables.

▷ **Méthode 2 : Analyse des résidus standardisés de Schoenfeld**

Cette seconde méthode est préférable pour les variables explicatives continues.

Les résidus de Schoenfeld, introduits par SCHOENFELD (1982), concernent les covariables et non le taux de hasard. Il y en a autant que de covariables et ils s'appliquent uniquement pour les individus non censurés (*i.e.* $\{i \in [\![1, n]\!] \mid Y_i = T_i\}$).

Un résidu lié à une covariable représente l'écart entre la valeur prise par cette covariable pour un individu au moment de la survenance de l'évènement (ici résiliation) et la moyenne de cette covariable parmi tous les individus exposés au risque à ce même moment. Les résidus peuvent être standardisés à l'aide de la covariance. THERNEAU et GRAMBSCH (1994) ont montré que les résidus standardisés de Schoenfeld doivent vérifier, sous l'hypothèse de proportionnalité, l'approximation suivante

$$\mathbb{E}[s_{k,j}] + \widehat{\beta}_j \approx \beta_j(t_k) \tag{2.16}$$

où $s_{k,j}$ représente le résidu de Schoenfeld standardisé au temps t_k de la $j^{\text{ème}}$ covariable et $\widehat{\beta}_j$ l'estimation du coefficient β_j obtenu par le modèle à risques proportionnels de Cox.

Les résidus ne sont pas une mesure de l'ajustement du modèle aux données, ils s'interprètent comme une mesure de la différence de profils entre un individu qui subit l'évènement (*i.e.* qui résilie son contrat) et l'ensemble des individus exposés à ce risque.

Comme pour l'hypothèse de log-linéarité, pour valider l'hypothèse de proportionnalité, il faut représenter les résidus en fonction du temps (ou une transformation du temps) pour chaque covariable continue. Il est également possible de représenter sur le même graphique une courbe représentant l'évolution moyenne des résidus en fonction du temps. Cette courbe donne la tendance générale. Toute différencepar rapport à une droite horizontale d'équation $y = 0$ représente unedéviation par rapport à l'hypothèse de proportionnalité.

Voici un exemple de représentation dans la figure 2.4 :

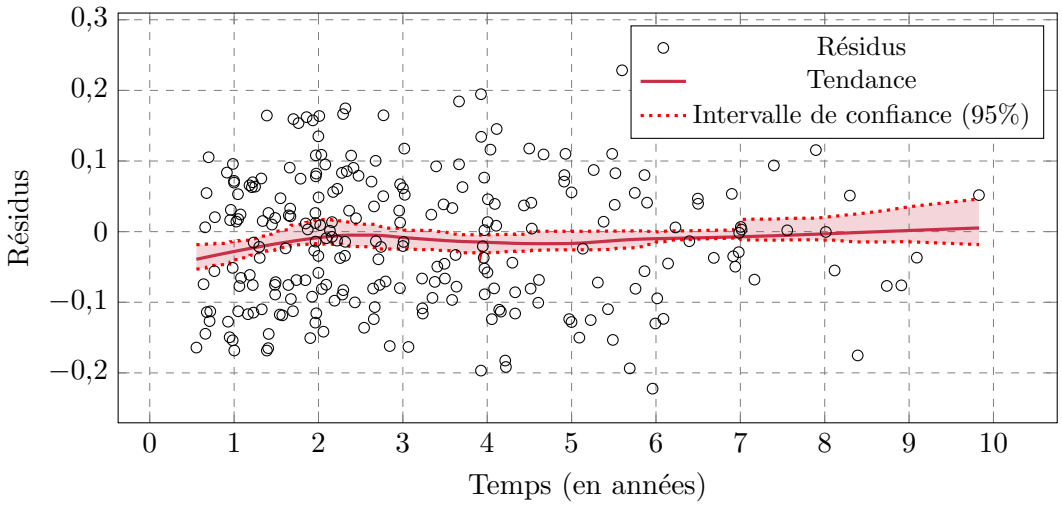


Figure 2.4 : Exemple de représentation graphique des résidus de Schoenfeld en fonction du temps pour une covariable continue

2.3.5 Validité et performance du modèle à risques proportionnels de Cox

Pour valider le modèle à risques proportionnels de Cox, plusieurs tests existent comme le test de Wald, le test du *log-rank* ou encore le test du rapport de vraisemblances. La théorie de ces derniers est développée en annexe A.2.

De plus, la métrique qui évalue la performance du modèle est appelée la concordance, une mesure connue depuis longtemps en statistiques principalement popularisée par le biostatisticien Harrell (1982). Celle-ci est la métrique d'évaluation la plus utilisée pour mesurer la qualité d'un modèle à risques proportionnels de Cox comme développé dans l'article de Therneau et Watson (2017).

Deux couples (x_i, y_i) et (x_j, y_j), pour $i \neq j$, sont dits concordants si x et y suivent le même ordre, c'est-à-dire ($x_i < x_j$ et $y_i < y_j$) ou ($x_i > x_j$ et $y_i > y_j$). Ainsi, la concordance est la proportion de couples concordants parmi tous les couples possibles d'un échantillon. Mathématiquement, cela se traduit pour $x = (x_1, ..., x_n) \in \mathbb{R}^n$ et $y = (y_1, ..., y_n) \in \mathbb{R}^n$ par

$$c(x, y) = \frac{\sum_{i \neq j} \mathbb{1}_{\{x_i < x_j,\ y_i < y_j\}} + \mathbb{1}_{\{x_i > x_j,\ y_i > y_j\}}}{(n-1)^2} \tag{2.17}$$

Cependant en analyse de survie, les couples ne peuvent pas être ordonnés à cause de la censure. En effet, en considérant n individus caractérisés par la durée de vie de leur contrat T et leur indicateur de censure δ (comme définis dans la section 2.1.2), certains couples de durée de vie ne peuvent être comparés comme par exemple une observation censurée à 5 ans et une résiliation à 6 ans. Dans ce cas, Harrell, Califf, Pryor, Lee et Rosati (1982) proposent de modifier la formule (2.17) pour ne considérer que des couples comparables. Alors, pour tout $(i, j) \in [\![1, n]\!]$, $i \neq j$, l'indicateur $K(i, j)$ est défini de la façon suivante

$$K(i,j) = \mathbb{1}_{\{T_i < T_j,\ \delta_i = 1\}} + \mathbb{1}_{\{T_i = T_j,\ \delta_i = 1,\ \delta_j = 0\}} = \mathbb{1}_{\{\delta_i = 1\}} \left(\mathbb{1}_{\{T_i < T_j\}} + \mathbb{1}_{\{T_i = T_j,\ \delta_j = 0\}} \right)$$

Nota Bene :

> *L'indicatrice $\mathbb{1}_{\{T_i = T_j,\ \delta_i = 1,\ \delta_j = 0\}}$ suit la convention que si l'individu j subit une censure au même moment que l'individu i résilie son contrat, alors nécessairement le contrat de l'individu j a une durée de vie supérieure à celle de l'individu i.*

Ils définissent ainsi la concordance comme la part de tous les couples ordonnés, c'est-à-dire tous les individus $i \neq j$ tels que $K(i,j) = 1$ ou $K(j,i) = 1$, pour lesquels le score de risque x prédit correctement cet ordre. Si τ représente la limite supérieure du temps (par exemple 9 si l'étude dure 10 ans), alors la concordance peut être définie par

$$C = \frac{\sum_{i \neq j} \mathbb{1}_{\{T_i < \tau\}} K(i,j) \left(\mathbb{1}_{\{x_i > x_j\}} + \frac{\mathbb{1}_{\{x_i = x_j\}}}{2} \right)}{\sum_{i \neq j} \mathbb{1}_{\{T_i < \tau\}} K(i,j)} \tag{2.18}$$

Dans le cadre d'un modèle à risques proportionnels de Cox, le score de risque x correspond au prédicteur linéaire $X\beta$ prédit par le modèle, *i.e.* $X\widehat{\beta}$. Ainsi, le modèle est considéré comme cohérent si un contrat d'un individu possède une durée de vie plus longue qu'un second, alors le risque qu'il supporte, représenté par son prédicteur linéaire, est inférieur à celui du contrat du deuxième individu.

Enfin, C varient entre 0 à 1, indiquant un score de risque parfaitement discordant à concordant, et une valeur de 0,5 signifie que le score de risque est indépendant des temps d'événements, suggérant des prédictions purement aléatoires du modèle.

2.4 Extensions du modèle à risques proportionnels de Cox

Pour modéliser la durée de vie d'un contrat d'assurance non-vie, plusieurs variantes du modèle à risques proportionnels de Cox peuvent être particulièrement adaptées en fonction des spécificités des données et des objectifs de l'analyse. Voici quelques exemples de modèles de Cox qui pourraient être pertinents :

- ▷ **Modèle de Cox stratifié :**

  Si les contrats varient significativement en fonction de certaines variables catégorielles (comme le type de contrat, la région géographique, ou le réseau de distribution), utiliser un modèle de Cox stratifié peut aider à ajuster les estimations en permettant à chaque strate de posséder son propre risque de base. Cela permet de mieux contrôler les facteurs de variation qui ne sont pas directement mesurés par les covariables incluses dans le modèle.

- ▷ **Modèle de Cox à *time-varying covariates* :**

  Dans le cas où les covariables, telles que le montant de la prime, le niveau de couverture ou des indicateurs économiques externes, changent avec le temps, il peut être judicieux d'utiliser un modèle de Cox avec des covariables dépendantes du temps pour capturer l'effet dynamique de ces covariables sur le risque de résiliation.

- ▷ **Modèles de risques concurrents :**

  Si plusieurs types d'événements de sortie sont possibles tels que la résiliation du contrat, le non-renouvellement, ou d'autres formes de censure, un modèle de risques concurrents peut être approprié pour analyser ces risques en compétition (cf. le modèle de Fine & Gray).

- ▷ **Modèle de Cox à risques non proportionnels :**

  Si l'hypothèse de risques proportionnels n'est pas vérifiée, ce modèle peut permettre de modéliser cette variabilité. Cela peut être pertinent pour des contrats où l'effet des facteurs de risque peut s'accroître ou décroître avec l'ancienneté du contrat.

Le choix du modèle dépend de la structure des données et des hypothèses retenues. Une analyse exploratoire des données et des tests préliminaires pour vérifier les hypothèses des modèles peuvent également aider à orienter le choix du modèle le plus adapté. Cependant pour la suite des travaux, uniquement le modèle de Cox stratifié sera développé.

2.4.1 Modèle de Cox stratifié

Ce modèle permet de gérer des données où l'hypothèse des risques proportionnels n'est pas respectée pour toutes les covariables sur l'ensemble des données. Cette méthode est particulièrement utile pour analyser des données qui incluent des sous-groupes ou des strates ayant des risques de base différents.

Dans le modèle de Cox stratifié, chaque strate a sa propre fonction de risque de base $h_0(\cdot)$, mais l'effet des covariables est supposé être le même dans toutes les strates. Cela signifie que le modèle permet à chaque strate d'avoir son propre profil de risque inhérent, tout en estimant des effets de covariables qui sont communs à toutes les strates.

Il est aussi possible de stratifier en fonction de plusieurs variables simultanément.

Supposons qu'une variable catégorielle dispose de K modalités, ainsi le modèle relatif à la $k^{\text{ème}}$ strate ($k \in [\![1, K]\!]$) s'écrit

$$h_k(t|X) = h_{0,k}(t) \exp(X\beta) \tag{2.19}$$

Pour rappel, dans le modèle classique, la vraisemblance partielle est donnée par

$$\mathcal{L}(x, \beta) = \prod_{Y_i, \delta_i = 1} \frac{\exp\left(x_i^\top \beta\right)}{\sum_{Y_i \leq Y_j} \exp\left(x_j^\top \beta\right)} \tag{2.20}$$

Ce produit peut être très simplement réorganisé par strate, ce qui donne une vraisemblance partielle par strate. Il est ensuite possible de calculer les vraisemblances partielles pour chaque strate, notées $\mathcal{L}_k(x, \beta)$, pour tout $k \in [\![1, K]\!]$. La vraisemblance partielle du modèle de Cox stratifié est alors égale au produit des $\mathcal{L}_k(x, \beta)$.

<u>Nota Bene</u> :

> ⚠ *Le modèle de Cox stratifié ne revient pas à calculer des modèles séparément pour chaque niveau de la variable de stratification !*

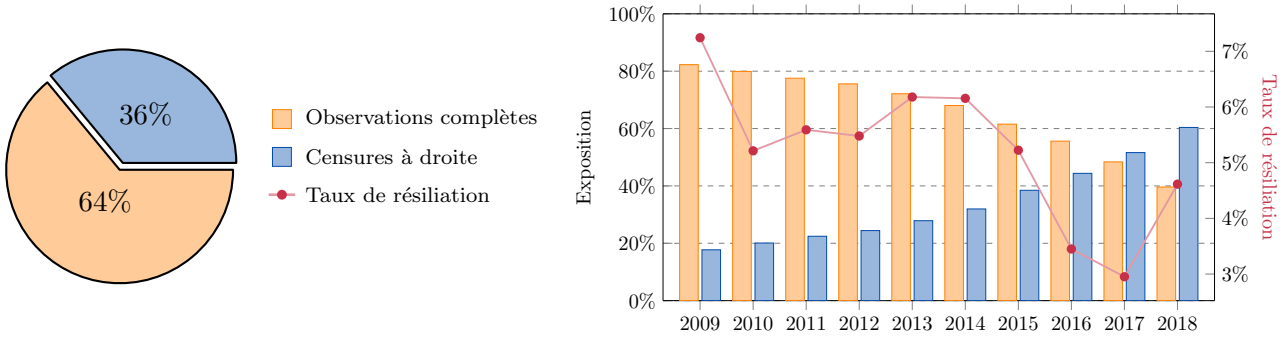
Ce modèle part de l'hypothèse de non-interaction entre la variable de stratification et les variables explicatives. Pour tester cette hypothèse, plûtot que de calculer des modèles séparément pour chaque strate, il est possible de partir du modèle stratifié et de rajouter des termes d'interaction entre la variable de stratification et chaque variable explicative. Un test du rapport de vraisemblances est ensuite effectué entre le modèle stratifié de départ et le modèle stratifié avec interactions.

2.5 Traitement des données et application au portefeuille

2.5.1 Traitement des données

La base de données utilisée pour modéliser la durée de vie des contrats est la base des affaires nouvelles. Ces données ont déjà été nettoyées et retraitées dans de précédents travaux de R&D (REGNAULT (2023)). De plus, des données de durée de vie des contrats et des indicateurs de résiliation et censure à droite (cf. 2.1.2) sont également présents.

Une analyse a été effectuée pour déterminer les états des données à la fin de l'étude. Lors de l'extraction de la base, deux cas de figure sont envisageables : soit le contrat a été résilié avant cette date, soit il est encore valide. Si la date de fin du dernier statut d'un contrat précède la date d'extraction, ce contrat est considéré comme terminé et entièrement observé. Dans le cas contraire, le contrat est toujours actif et est donc considéré comme censuré à droite. La figure 2.5(a) montre la répartition des contrats complètement observés (64%) et des contrats censurés à droite (36%) dans l'ensemble des données de l'étude :



(a) Parts des deux états possibles à la fin de l'étude

(b) Histogrammes des états en fonction de l'année de début d'observation

FIGURE 2.5 : Parts des états possibles à la fin de l'étude et histogrammes des états en fonction de l'année de début d'observation

La figure 2.5(b) permet de voir très clairement que le nombre de contrats censurés est croissant avec l'année de souscription. Cela est une conséquence directe du fait que plus un contrat est souscrit récemment, plus sa probabilité d'être encore en portefeuille est grande. Concernant le taux de résiliation, celui-ci oscille entre 7% et 3% pendant la période d'étude mais montre tout de même une tendance décroissante.

Ces différentes informations vont permettre de modéliser la durée de vie des contrats via un modèle à risques proportionnels de Cox (cf. section 2.3). En effet, bien que la théorie de Kaplan-Meier ait été étudiée pour une compréhension complète des modèles de durée de vie et de l'analyse de survie, l'utilisation du modèle à risques proportionnels de Cox dans l'application est justifiée par sa capacité à fournir une analyse plus robuste et détaillée des effets des covariables sur la survie tout en prenant en compte directement les observations censurées à droite, ce qui est essentiel pour répondre aux questions spécifiques de l'étude.

Enfin, chaque variable catégorielle a subi un encodage *One-Hot* avec suppression de la modalité la plus représentée. L'encodage *One-Hot* est une méthode de transformation des variables catégorielles en un format binaire où chaque modalité est représentée par un vecteur dont une seule entrée est à 1 et toutes les autres sont à 0. Il est également possible de supprimer une des colonnes sans perte d'information en raison de la redondance implicite dans la représentation. En effet, en supprimant

une colonne, il est toujours possible de déduire sa valeur en utilisant les valeurs des autres colonnes restantes. Pour l'étude, il a été choisi de supprimer la colonne la plus représentée. Par exemple, la variable catégorielle `RESEAU_DIST`, qui représente le réseau de distribution de chaque contrat, possède 3 modalités : "Agents/Salariés", "Courtiers" et "Autres", où "Agents/Salariés" est la modalité la plus représentée. L'encodage *One-Hot* de cette variable est donc la suivante :

RESEAU_DIST	RESEAU_DIST_Courtiers	RESEAU_DIST_Autres
Agents/Salariés	0	0
Courtiers	1	0
Autres	0	1

Table 2.1 : Exemple d'encodage *One-Hot* pour la variable `RESEAU_DIST`

2.5.2 Modélisation de la durée de vie

Comme expliqué dans la section 2.3, le modèle à risques proportionnels de Cox se comporte de manière assez similaire à une régression linéaire. Pour obtenir le modèle optimal, les variables les plus significatives ayant une faible volatilité ont été sélectionnées. Un modèle à 41 variables (19 avant l'encodage) est alors obtenu, où chaque modalité (sauf la plus représentée) des variables catégorielles forment une variable en raison de l'encodage *One-Hot*. Les variables retenus pour le modèle sont listées ci-dessous, ainsi que leurs modalités (dans l'ordre décroissant d'exposition) :

- ▷ Âge du conducteur (`AGE_COND` $\in [18; 75]$)
- ▷ Ancienneté du véhicule (`ANCIENNETE_VEHICULE` $\in [0; 25]$)
- ▷ Montant de la prime commerciale de la première année d'assurance (`PRIME_COMM` $\in$ [32,95 ; 7196,40])
- ▷ Nombre de contrats différents (`NBR_CTR_DIFF` $\in [\![0; 5]\!]$)
- ▷ Nombre de garanties (`NBR_GARANTIES` $\in \{2; 3; 6; 7\}$)
- ▷ Fractionnement de paiement (`FRAC_PAIEMENT` : "Annuel" ou "Mensuel/Trimestriel/ Semestriel")
- ▷ Réseau de distribution (`RESEAU_DIST` : "Agents/Salariés", "Courtiers" ou "Autres")
- ▷ Coefficient Bonus-Malus (`COEF_BONUS_MALUS` $\in [0,5; 1,95]$)
- ▷ Niveau de franchise dommage (`NIVEAU_FRANCHISE_DMG` : "Sans Franchise", "Standard", "Faible", "Elevé", "Sans objet" ou "Doublé")
- ▷ Niveau de franchise bris de glace (`NIVEAU_FRANCHISE_BDG` : "Sans Franchise", "Standard", "Sans objet" ou "Elevé")
- ▷ Catégorie socio-professionnelle (`CSP` : "Salarié", "Artisan - Commerçant", "Fonctionnaire ou assimilé", "Agriculteur", "Retraité", "Profession libérale", "Sans profession" ou "Etudiant")
- ▷ Indicateur de résiliation du précédent assureur (`RESIL_PRECEDENT_ASS` $\in \{0; 1\}$)
- ▷ Type de conducteur secondaire (`TYPE_COND_SCD` : "NR", "Enfant" ou "Conjoint")
- ▷ Contrat d'assurance habitation résidence principale (`OPT_CTR_HAB` $\in \{0; 1\}$)
- ▷ Second contrat d'assurance automobile (`OPT_SCD_CTR_AUTO` $\in \{0; 1\}$)
- ▷ Type d'habitation (`TYPE_HAB` : "Mais-Prop", "Autre-NR", "Mais-Loc", "App-Loc", "App-Prop", "Autre-Prop", "Autre-Loc", "Mais-NR" ou "App-NR")
- ▷ Indicateur de plusieurs options (`OPT_MOB`, `OPT_INDEM`, `OPT_CONT` $\in \{0; 1\}$)

Pour visualiser l'impact de chaque variable du modèle, le graphique des *Log-Hazard ratios* est représenté. Il s'agit, comme leur nom l'indique, des logarithmes des ratios de hasard définis à la sous-section 2.3.3. Mathématiquement, il s'agit donc des coefficients $-\beta_j$, pour $j \in [\![1, p]\!]$. Ce graphique permet de visualiser l'impact logarithmique des différentes covariables sur le taux de risque instantané. Une valeur positive indique une augmentation du risque de résiliation avec la variable, tandis qu'une valeur négative indique une diminution du risque. Ainsi, sur la figure 2.6, plus le *Log-Hazard ratio* d'une variable est négatif (resp. positif), plus cette variable contribue positivement (resp. négativement) à la survie. Par exemple, la variable qui a le plus d'impact positif sur la survie est l'indicateur de contrat d'assurance habitation résidence principale. À l'inverse, la variable qui a le plus d'impact négatif sur la survie est la variable dont le type d'habitation est "Autre-Loc" (location d'un logement qui n'est ni une maison, ni un appartement).

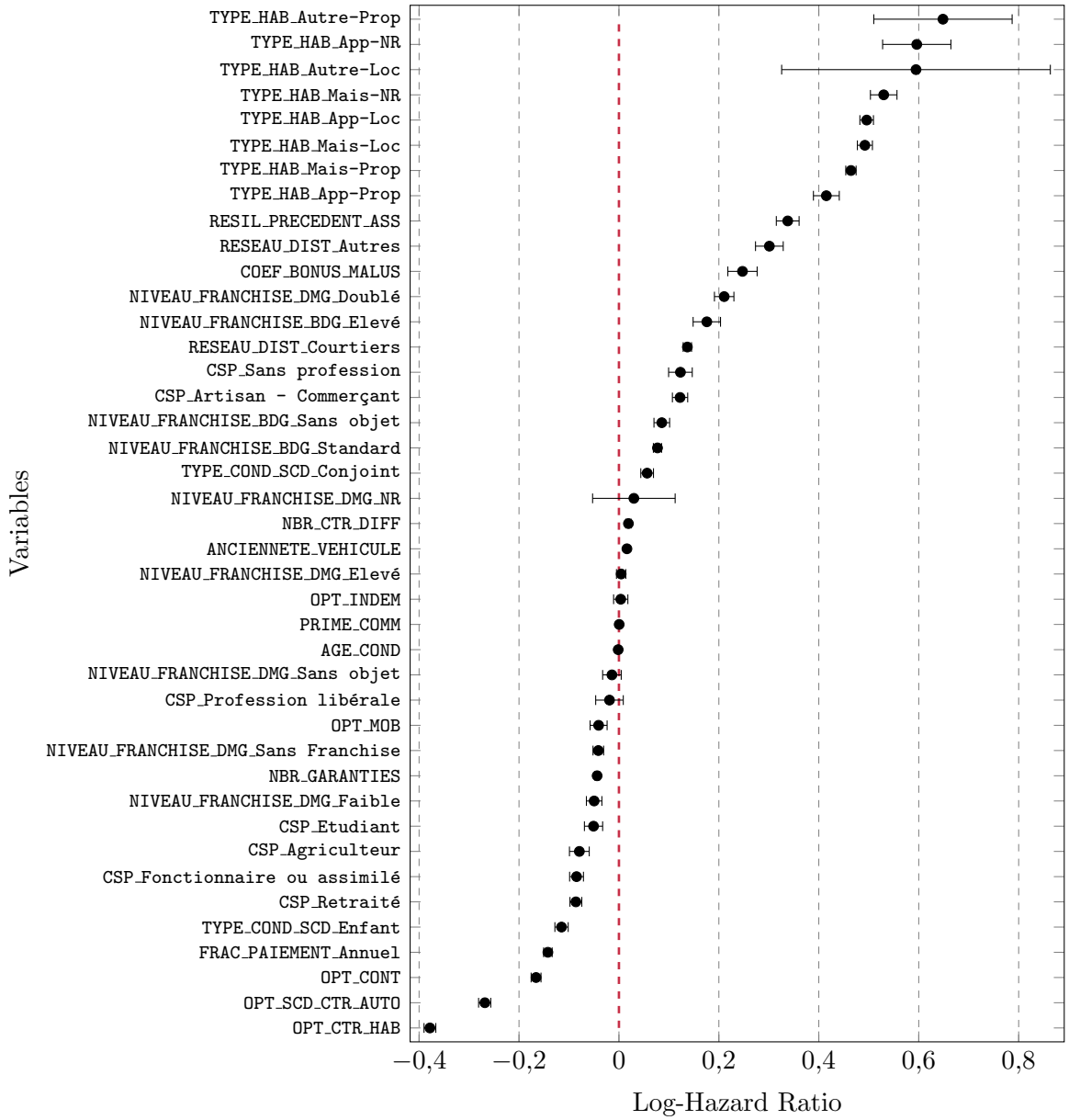


FIGURE 2.6 : Log-Hazard ratios avec intervalles de confiance à 95% du modèle optimal

Le graphique des *Log-Hazard ratios* montre aussi que la variable `RESIL_PRECEDENT_ASS`, qui indique si le contrat a été résilié par son précédent assureur, est significativement associée à un risque de

résiliation accru avec un rapport de risque (*Hazard ratio*) de 1,4 (*Log-Hazard ratio* = 0,33), ce qui signifie qu'avoir été résilié par son précedent assureur multiplie le risque de nouvelle résiliation par 1,4. À contrario, la variable `NBR_GARANTIES`, par exemple, qui représente le nombre de garanties dans le contrat, semble avoir un effet protecteur avec un rapport de risque égal à 0,95 ce qui signifie que pour chaque garantie en plus dans le contrat, le risque de résiliation est multiplié par 0,95. Ces résultats sont cohérents avec la réalité.

Enfin, le modèle présente une concordance égale à 0,6198 ce qui indique une bonne capacité pour prédire la survie des individus et pour différencier les individus selon celle-ci.

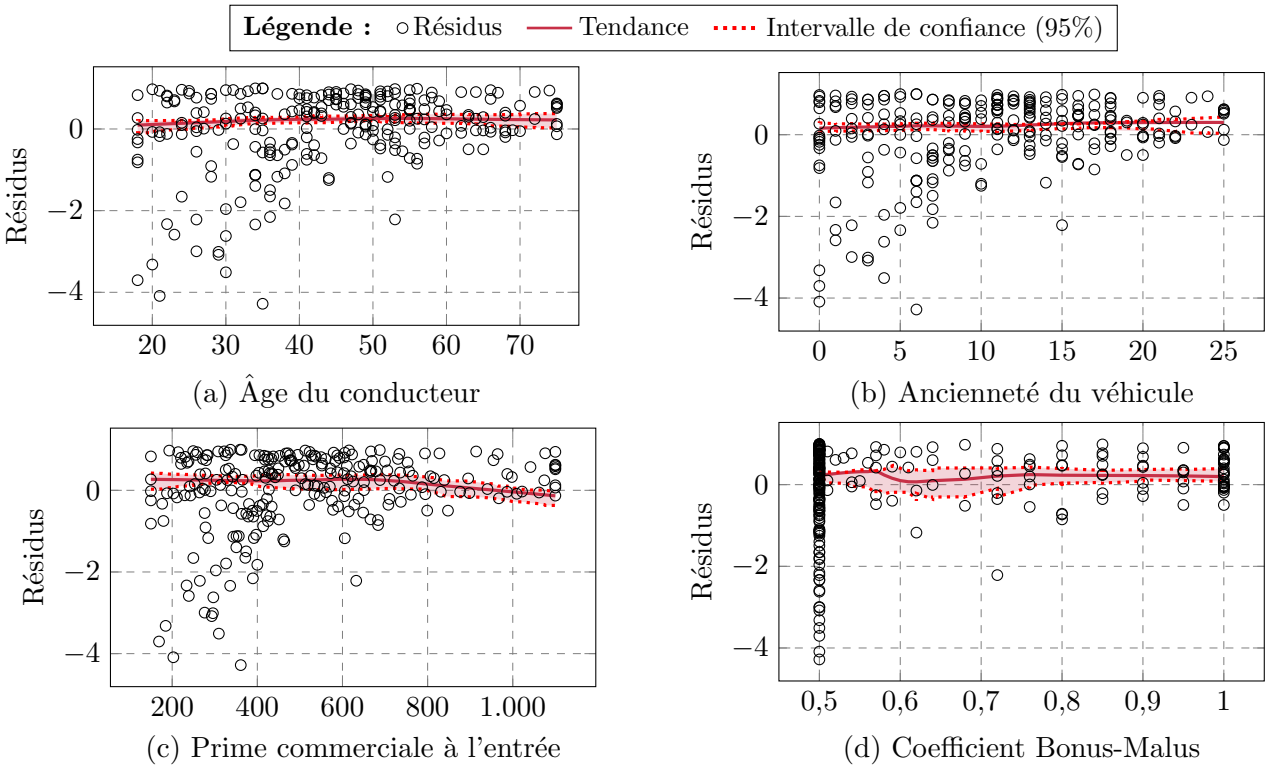
2.5.3 Vérification des hypothèses du modèle

Désormais que le modèle a été créé, les hypothèses développées dans la sous-section 2.3.4 doivent être vérifiées.

Hypothèse de log-linéarité

Commençons par l'hypothèse de log-linéarité. Pour rappel, pour valider cette hypothèse il faut représenter les résidus de martingale $(\widehat{M}_i)_{i\in[\![1,n]\!]}$ en fonction du temps pour chaque covariable numérique incluse dans le modèle puis, il faut étudier la tendance générale de ces résidus. Toute différence par rapport à la droite horizontale d'équation $y = 0$ représente une déviation par rapport à l'hypothèse de log-linéarité.

Dans notre modèle, les variables numériques (continues) sont l'âge du conducteur et du véhicule, la prime commerciale proposée à l'entrée en portefeuille et le coefficient Bonus-Malus. Voici, dans la figure 2.7, des échantillons des résidus de martingale (et leur tendance) pour chacune de ces variables :



(a) Âge du conducteur

(b) Ancienneté du véhicule

(c) Prime commerciale à l'entrée

(d) Coefficient Bonus-Malus

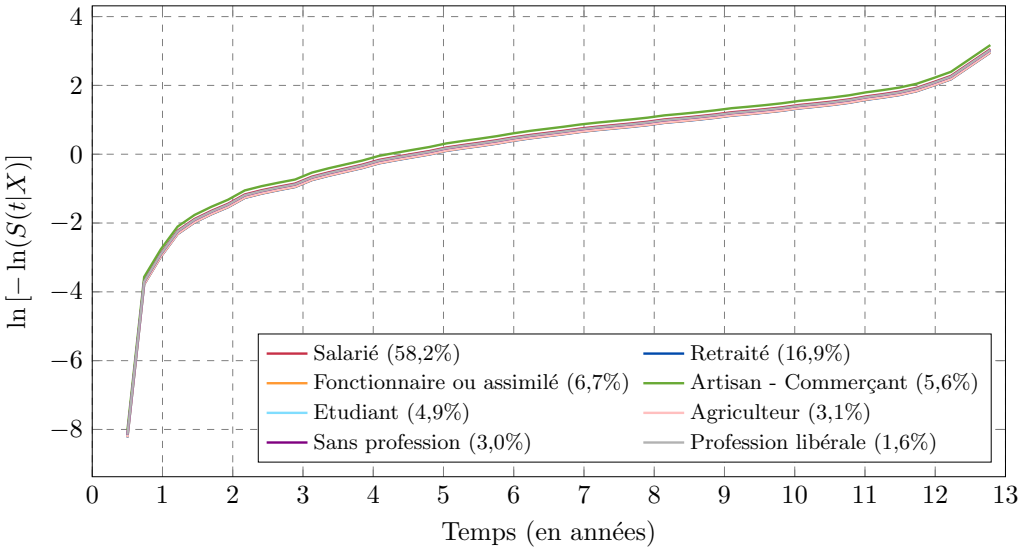
Figure 2.7 : Echantillon des résidus de martingale en fonction des variables numériques (continues) du modèle à risques proportionnels de Cox

Sur chacun des graphes, malgré de légères fluctuations, la tendance est toujours bien centrée autour de 0. Il est alors possible de conclure que l'hypothèse de log-linéarité semble être vérifiée pour le modèle construit.

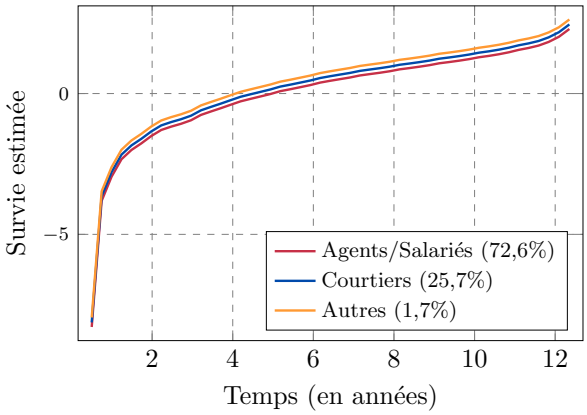
Hypothèse de proportionnalité

Passons maintenant à l'hypothèse de proportionnalité. Pour rappel, deux approches sont proposées pour valider cette hypothèse : une approche pour les variables qualitatives ou discrètes et une approche par l'analyse des résidus de Schoenfeld pour les variables continues.

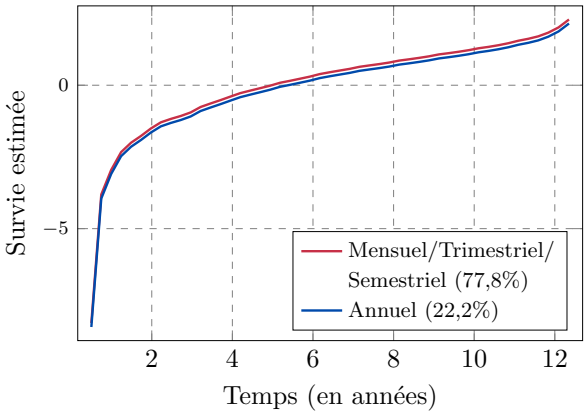
Commençons par la première méthode, les courbes $t \longmapsto \ln\left[-\ln(S(t|X=x))\right]$ sont représentées pour chaque modalité x de la variable qualitative X dans un même graphique. Pour valider l'hypothèse de proportionnalité des variables catégorielles, il suffit alors de vérifier que les courbes sont parallèles (au sens qu'elles ne se croisent pas et qu'elles restent à une même distance au cours du temps). Voici par exemple de tels graphiques pour la catégorie socio-professionnelle, le réseau de distribution et le fractionnement de paiement :



(a) Selon la catégorie socio-professionnelle



(b) Selon le réseau de distribution



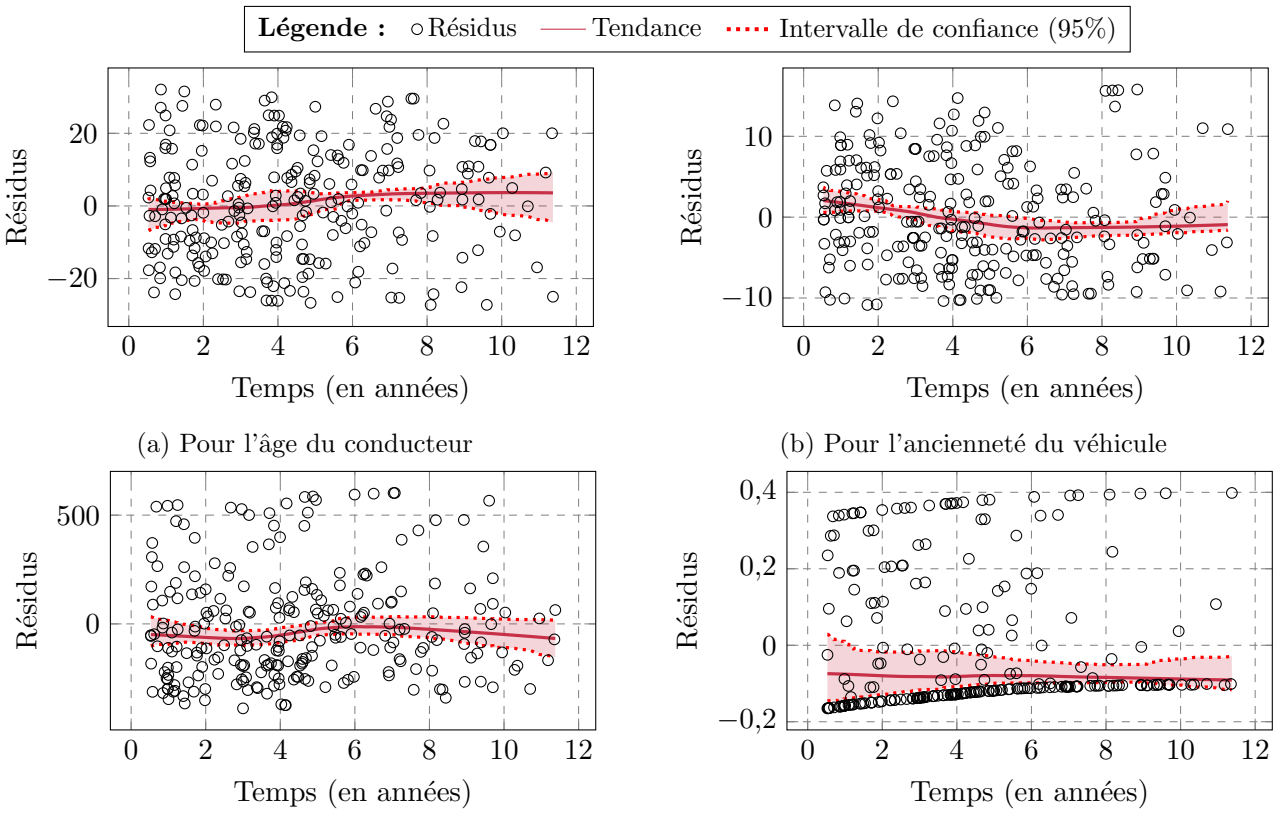
(c) Selon le fractionnement de paiement

FIGURE 2.8 : Courbes d'équation $\ln[-\ln(S(t|X))]$ selon 3 variables catégorielles

D'après la figure 2.8, pour ces 3 variables, l'hypothèse de proportionnalité est vérifiée puisque, pour chacune d'elles, les courbes de chaque modalité ne se croisent pas et gardent la même distance au cours du temps.

Pour les variables numériques continues, cette première méthode ne peut être appliquée que si ces variables sont segmentées en différentes catégories, ce qui génère une perte d'informations. Il est donc préférable de calculer et d'analyser les résidus de Schoenfeld. Comme pour l'hypothèse de log-linéarité, les résidus de Schoenfleld sont représentés en fonction de chaque covariable continue incluse dans le modèle puis leur tendance générale est étudiée. Toute différence par rapport à la droite horizontale d'équation $y = 0$ représente une déviation par rapport à l'hypothèse de proportionnalité.

Comme dit précédemment, les variables continues du modèle construit sont l'âge du conducteur et du véhicule, la prime commerciale proposée à l'entrée en portefeuille et le coefficient Bonus-Malus. Voici, dans la figure 2.9, des échantillons des résidus de Schoenfeld (et leur tendance) pour chacune de ces variables explicatives :



(a) Pour l'âge du conducteur

(b) Pour l'ancienneté du véhicule

(c) Pour la prime commerciale à l'entrée

(d) Pour le coefficient Bonus-Malus

Figure 2.9 : Echantillon des résidus de Schoenfeld en fonction du temps pour les variables continues du modèle à risques proportionnels de Cox

Comme pour les résidus de martingale, malgré des fluctuations, la tendance est toujours bien centrée autour de 0 sur chaque graphe. Notons tout de même que pour le coefficient Bonus-Malus, la tendance est légèrement inférieure à 0 mais reste autour de -0,07, ce qui est très proche de 0.

À la lumière des résultats des deux méthodes, il est possible de conclure que l'hypothèse de proportionnalité semble être vérifiée pour le modèle construit.

Tests statistiques

Pour appuyer les résultats précédents, les tests de Wald, du *log-rank* et du rapport de vraisemblances, dont les théories sont développées en annexe A.2, ont été réalisés.

TEST DE WALD :

Le tableau des résultats du test de Wald pour chaque variable est disponible en annexe A.3. Pour résumer, presque toutes les variables sont très significatives puisqu'elles ont une p-value inférieures à 0,05. Seules la catégorie socio-professionnelle "Profession libérale" et le niveau de franchise bris de glace "Sans objet" ne respectent pas seuil de 5% mais ces covariables restent quand même assez significative avec des p-values respectivement égales à 0,138 et 0,062. Il en résulte que toutes ces variables semblent être importantes pour le modèle. De même, au niveau global, le test de Wald possède une statistique de test égale à 51867,20, sa p-value est donc bien inférieure à 0,05.

TEST DU LOG-RANK :

Dans le tableau 2.2 se trouvent les résultats des tests du *log-rank* sur chaque variable catégorielle du modèle construit. Pour rappel, ce test est utilisé pour vérifier si les groupes ont des risques proportionnels constants dans le temps. Les résultats sont très significatifs au seuil de 0,05 pour chaque variable. Cela confirme donc une fois de plus que l'hypothèse de proportionnalité semble être vérifiée pour ce modèle.

Covariables	**Degrés de liberté**	**Statistique de test**	**p-value**	**Significativité**
FRAC_PAIEMENT	1	3761,01	$< 0,05$	$\star\star\star$
CSP	8	5094,03	$< 0,05$	$\star\star\star$
RESEAU_DIST	2	1343,34	$< 0,05$	$\star\star\star$
TYPE_COND_SCD	2	1036,23	$< 0,05$	$\star\star\star$
TYP_HAB	8	9168,48	$< 0,05$	$\star\star\star$
NIVEAU_FRANCHISE_DMG	5	8669,98	$< 0,05$	$\star\star\star$
NIVEAU_FRANCHISE_BDG	3	9174,57	$< 0,05$	$\star\star\star$
NBR_CTR_DIFF	2	966,47	$< 0,05$	$\star\star\star$
NBR_GARANTIES	3	7823,32	$< 0,05$	$\star\star\star$
OPT_MOB	2	1036,23	$< 0,05$	$\star\star\star$
OPT_INDEM	1	483,06	$< 0,05$	$\star\star\star$
OPT_CONT	1	3942,47	$< 0,05$	$\star\star\star$
OPT_CTR_HAB	1	882,44	$< 0,05$	$\star\star\star$
OPT_SCD_CTR_AUTO	1	673,97	$< 0,05$	$\star\star\star$
RESIL_PRECEDENT_ASS	1	1484,52	$< 0,05$	$\star\star\star$

TABLE 2.2 : Résultats des test du *log-rank* pour les variables catégorielles du modèle à risques proportionnels de Cox

TEST DU RAPPORT DE VRAISEMBLANCES :

Pour davantage valider la qualité du modèle créé, un test du rapport de vraisemblances a été réalisé. Celui-ci compare le modèle complet avec un modèle nul sans covariables. La statistique du test est égale à 47260,24 impliquant une p-value largement inférieure à 0,05. Ce résultat suggère que les covariables incluses dans le modèle apportent une contribution significative à la prédiction de la survie. Autrement dit, le modèle construit précédemment explique significativement mieux les données de survie que le modèle nul.

2.5.4 Résultats du modèle

Maintenant que les hypothèses sous-jacentes au modèle ont bien été vérifiées, les fonctions de survie prédites par le modèle vont pouvoir être étudier. La fonction de survie d'un individu i représente la probabilité d'être encore dans le portefeuille à tout temps $t \in [0, T_h]$. Il s'agit donc du terme $\mathbb{P}_i(t)$, pour $t \in [0, T_h]$, dans la définition de la valeur client future choisie pour notre étude (cf. formule (1.2)).

Afin d'étudier l'impact de certaines variables sur la durée de vie des contrats, regardons les fonctions de survie moyennes estimées selon ces variables. Commençons par la catégorie socio-professionnelle des assurés dans la figure 2.10 suivante :

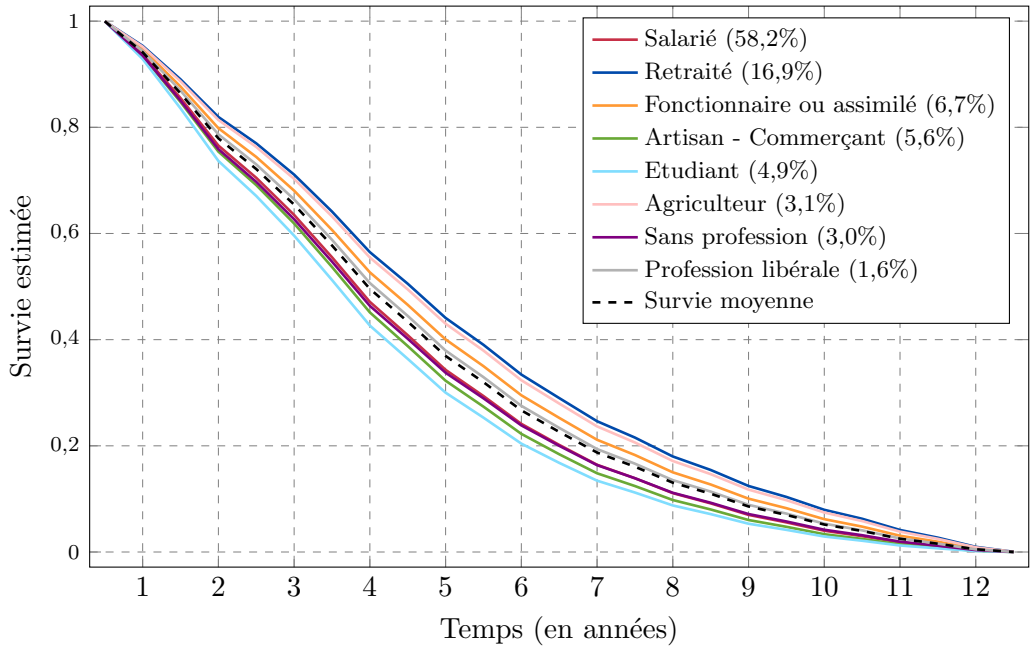


Figure 2.10 : Fonctions de survie moyennes estimées par le modèle à risques proportionnels de Cox selon les catégories socio-professionnelles

D'après ces fonctions de survie moyennes estimées par le modèle, les retraités et les agriculteurs sont les profils les plus fidèles et les plus à même de rester longtemps dans le portefeuille. À contrario, les étudiants et les artisans - commerçants sont les profils ayant le moins de chance de rester dans le portefeuille. Par exemple, en moyenne, les retraités ont 36% de chances d'être encore dans le portefeuille au bout de 6 ans contre seulement 20% pour les étudiants. Ces observations peuvent s'expliquer par des facteurs sociaux et comportementaux associés à chaque groupe.

Retraités :

- Les retraités peuvent disposer d'une situation financière plus stable grâce à leurs pensions de retraite et à leur habitude de vie plus économe, ce qui peut leur permettre de maintenir plus facilement leurs contrats d'assurance.
- En général, les personnes retraités ne se déplacent pas aussi fréquemment que les travailleurs actifs. Par conséquent, cela réduit le risque de sinistres ce qui peut conduire à une relation plus longue et stable avec leur assureur.
- Les retraités peuvent être plus averses au risque et préférer maintenir une couverture d'assurance plus complète pour se protéger contre d'éventuels imprévus.

Agriculteurs :

- ▷ Les agriculteurs vivent souvent en milieu rural, où le trafic est moins dense et les risques d'accidents sont moindres par rapport aux zones urbaines. Cela peut contribuer à une expérience de conduite plus sûre et à une relation plus longue avec l'assureur.
- ▷ Les véhicules agricoles sont essentiels pour leur activité professionnelle, ce qui rend les agriculteurs plus enclins à maintenir leurs polices d'assurance pour éviter toute interruption de leur activité.

Artisans - Commerçants :

- ▷ Les artisans et commerçants utilisent très souvent leurs véhicules pour des raisons professionnelles, ce qui augmente les risques d'accident et donc de demandes d'indemnisations, pouvant conduire à des résiliations en cas de mauvaise gestion de l'assureur ou à des primes plus élevées.
- ▷ Leurs revenus peuvent être plus volatils, en fonction des lois du marché ou de la saison, ce qui pourrait affecter leur capacité à payer régulièrement des primes dont le montant peut augmenter dans le temps.
- ▷ Afin de réduire les coûts, ils pourraient opter pour des polices moins coûteuses ou avec une couverture moindre, augmentant le risque de changer fréquemment d'assureur.

Étudiants :

- ▷ Les étudiants disposent souvent de ressources financières limitées, ce qui peut les amener à choisir des polices d'assurance moins coûteuses.
- ▷ Étant des conducteurs novices, ils peuvent être plus sujets aux accidents, entraînant des primes plus élevées et une possible résiliation du contrat.
- ▷ Les étudiants sont davantages sujets à des changements de vie à venir, ce qui peut augmenter la possibilité de résiliation.
- ▷ Il s'agit peut-être de leur premier assureur, une relation de confiance ne s'est pas encore installée à l'inverse de personnes retraitées qui peuvent être assurées depuis des années chez cette compagnie.

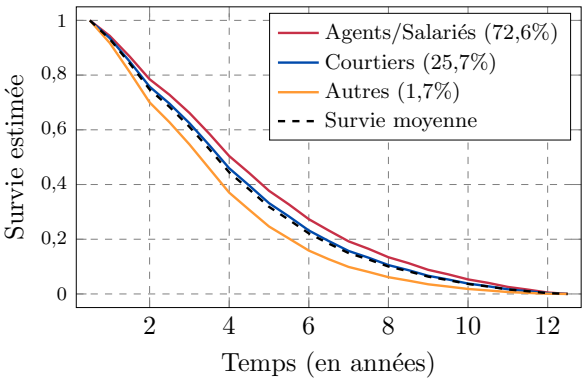
L'impact des modalités d'autres variables sur la durée de vie des contrats est également étudié grâce à la figure 2.11. Tout d'abord, le réseau de distribution joue un rôle important dans la survie des contrats. En effet, sur la figure 2.11(a) il apparaît que souscrire par un autre moyen qu'un agent général, comme par exemple auprès d'un courtier ou par *internet*, diminue fortement la probabilité de survie car la relation avec le client est souvent moins personnelle et plus transactionnelle. Au contraire, les agents généraux développent des relations de confiance plus étroites et offrent un service personnalisé, ce qui favorise une fidélité accrue et donc une rétention plus longue.

Ensuite, il est également intéressant de noter qu'un fractionnement de paiement annuel est meilleur en terme de survie qu'un fractionnement mensuel, trimestriel ou semestriel. Cela vient du fait que l'assureur peut offrir une réduction ou une prime moins élevée pour un paiement en une seule fois, réduisant ainsi les coûts administratifs et le risque de non-paiement. Les assurés ont également moins de démarches à faire au cours de l'année, ce qui diminue leur risque de résiliation. Cela témoigne également d'une meilleure santé financière des assurés qui sont alors moins sensibles au prix.

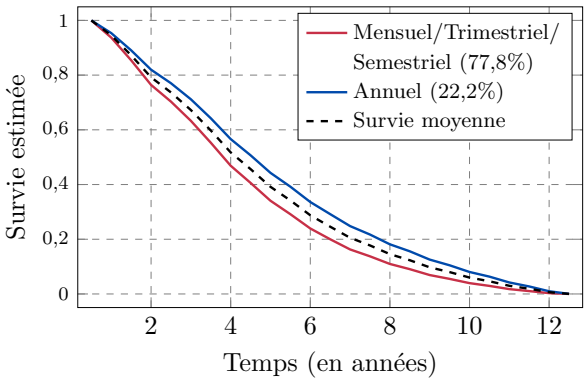
De plus, les clients ayant une couverture plus étendue perçoivent une plus grande protection contre divers risques, ce qui les incite à maintenir leur contrat sur le long terme. Ils peuvent être plus investis dans leur relation avec l'assureur réduisant leur propension à chercher des alternatives et augmentant leur fidélité.

Enfin, lorsqu'un client possède au moins un autre contrat que celui de son assurance automobile, cela

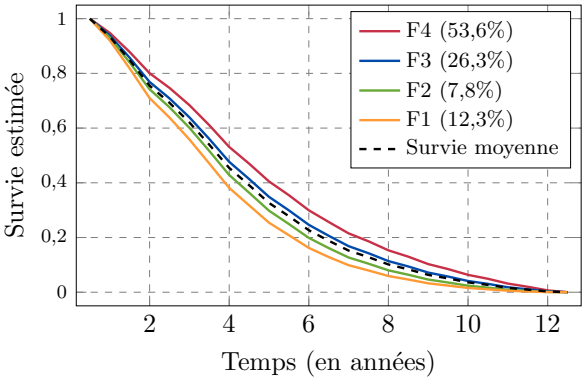
signifie qu'il est déjà fidélisé, celui-ci est satisfait du service fourni par son assureur et il cherchera donc moins à résilier qu'un autre nouveau client comme le montre la figure 2.11(d).



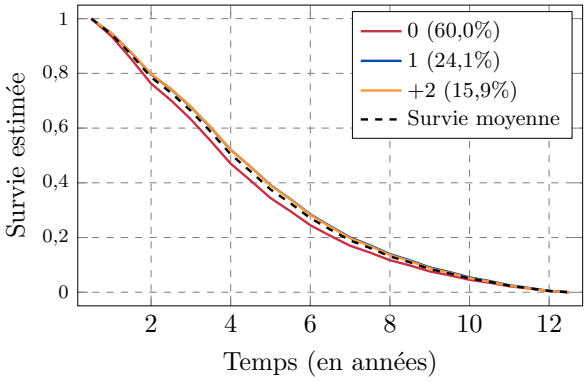
(a) Selon le réseau de distribution



(b) Selon le fractionnement de paiement



(c) Selon la formule souscrite



(d) Selon le nombre de contrats différents

Figure 2.11 : Fonctions de survie moyennes estimées par le modèle à risques proportionnels de Cox selon 4 variables

2.5.5 Modélisation du modèle stratifié

Un modèle de Cox stratifié, dont la théorie a été introduite à la section 2.4, a également été créé. Afin d'étudier plus précisément l'impact du réseau de distribution (variable `RESEAU_DIST`) sur la durée de vie des contrats, cette variable a été employée comme variable de stratification. En effet, le canal de distribution joue un rôle essentiel sur la durée de vie des contrats comme vu précédemment. De plus, le réseau de distribution peut témoigner d'une stratégie commerciale de l'assureur, qui peut être incité à fidéliser plus particulièrement les assurés ayant souscrit via des agents généraux plutôt qu'un courtier ou que par *internet*. Ainsi, pouvoir estimer un risque de base pour chaque canal peut permettre de mieux estimer la durée de vie des contrats.

Ce modèle part de l'hypothèse de non-interaction entre la variable de stratification et les variables explicatives. Pour tester cette hypothèse, plûtot que de calculer des modèles séparément pour chaque strate, il est possible de partir du modèle stratifié et de rajouter des termes d'interaction entre la variable de stratification et chaque variable explicative (numérique). Un test du rapport de vraisemblances est ensuite effectué entre le modèle stratifié de départ et le modèle stratifié avec interactions. Pour rappel, l'hypothèse nulle du test du rapport de vraisemblances dans notre cas est que le modèle

stratifié avec interactions est meilleur statistiquement que le modèle stratifié de départ. Il faut donc rejeter cette hypothèse nulle pour que l'hypothèse de non-interaction soit validée. Voici les résultats :

Variables	Rapport de vraisemblances	p-value	Conclusion	Interactions avec "Autres"	Interactions avec "Courtiers"
AGE_COND	29,5	$< 0{,}05$	Rejet de H0	0,0002	-0,0010
ANCIENNETE_VEHICULE	3,8	0,15	Non rejet de H0	0,0037	-0,0002
PRIME_COMM	82,8	$< 0{,}05$	Rejet de H0	0,0002	0,0001
COEF_BONUS_MALUS	20,4	$< 0{,}05$	Rejet de H0	0,0916	0,0652

TABLE 2.3 : Résultats du test de rapport de vraisemblances entre le modèle stratifié et le modèle stratifié avec interactions

Les deux dernières colonnes du tableau 2.3 représentent les valeurs des coefficients β pour les interactions entre chaque variable numérique et les modalités "Autres" et "Courtiers" de la variable de stratification RESEAU_DIST. Cependant, ici l'hypothèse nulle $\mathcal{H}_0$ n'est pas rejetée pour la variable ANCIENNETE_VEHICULE. Or, les valeurs des coefficients β pour les interactions associées à celle-ci avec la variable de stratification RESEAU_DIST sont très proches de 0, ce qui témoigne tout de même d'une faible importance de ces interactions dans le modèle. Il est donc raisonnable de supposer que l'hypothèse semble être validée pour toutes les variables numériques du modèle.

Cependant, pour la suite de l'étude, le modèle à risques proportionnels de Cox développé à la sous-section 2.5.2 sera gardé. En effet, la concordance entre les deux modèles sont très proches ($\approx$ 0,62 pour les deux modèles) et l'intérêt principal du modèle stratifié est de gérer les données où l'hypothèse des risques proportionnels n'est pas respectée, ce qui n'est pas le cas ici puisque l'hypothèse de proportionnalité est bien vérifiée d'après la sous-section 2.5.3.

2.6 Classification Ascendante Hiérarchique

La Classification Ascendante Hiérarchique (C.A.H.) est une méthode de classification non supervisée visant à regrouper des individus selon un critère de similarité de sorte que les classes créées soient composées d'observations homogènes et qu'elles soient le plus distinctes possible les unes des autres. Le principe est donc de créer, à chaque étape, une partition réunissant les deux éléments les plus proches au sens d'un indice d'agrégation qui mesure la dissimilarité entre ces deux éléments. Lors de l'initialisation, chaque observation forme une classe. À chaque itération, les observations sont regroupées selon l'indice d'agrégation jusqu'à ce que l'ensemble des individus ne forme qu'une seule et unique classe.

Pour appliquer une C.A.H., il faut au préalable choisir un critère de similarité. Un indice d'agrégation δ est une règle de calcul qui estime la dissimilarité entre deux classes A et B déjà existantes en tenant compte des dissimilarités entre les éléments de A et ceux de B. Quelques indices classiques sont listés en annexe A.4. Pour cette étude, l'indice de Ward est préféré car il permet de ne pas isoler les observations atypiques. Définissons le plus en détails :

$$\delta_{\text{Ward}}(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B) \tag{2.21}$$

où n_A et n_B sont les effectifs des classes A et B, g_A et g_B sont les centres de gravité de A et B, et d est la distance euclidienne.

De plus la décomposition de l'inertie (décomposition de Huygens) joue un rôle primordial dans la compréhension de la C.A.H. avec l'indice de Ward. En effet, l'inertie inter-classe, qui mesure la dispersion entre les classes, est à maximiser alors que l'inertie intra-classe, qui mesure la variabilité

des éléments au sein d'une classe, est à minimiser. L'algorithme cherche donc à regrouper les classes tout en minimisant la perte de l'inertie inter-classe et en minimisant le gain de l'inertie intra-classe. Cependant ces deux actions sont équivalentes d'après la décomposition de Huygens*.
Voici donc comment fonctionne l'algorithme de la Classification Ascendante Hiérarchique :

Algorithme 2 : Classification Ascendante Hiérarchique avec l'indice de Ward

Initialisation : chaque individu forme une classe

Jusqu'à ce que *l'ensemble des individus appartiennent à une seule et même classe,* **faire :**

- ◇ Calcul de la matrice de Ward à partir de l'équation (2.21) : $\mathbb{W} = \left(\delta_{\text{Ward}}(E_i, E_j)\right)_{i,j\in[\![1,n]\!]}$
- ◇ Réunion des deux groupes les plus proches au sens de l'indice de Ward $\longrightarrow$ cela garantit la minimisation de la perte de l'inertie inter-classe, et donc celle du gain de l'inertie intra-classe
- ◇ Calcul du centre de gravité de la nouvelle classe

Enfin, les agrégations successives peuvent être représentées graphiquement dans un dendrogramme. L'analyse de ce dernier, couplée à la courbe de l'inertie en fonction du nombre de classes, permettent de choisir le nombre de classes idéal.

Un exemple numérique de classification ascendante hiérarchique avec l'indice d'aggrégation de Ward est disponible en annexe A.5.

2.7 Classes de durée de vie des contrats

Dans une optique d'analyse plus approfondie de la survie des contrats du portefeuille d'affaires nouvelles, des classes de durée de vie des contrats ont été créées. Pour cela, une méthodologie précise a été mise en place. La première étape a consisté à récupérer les prédicteurs linéaires $X\beta$ du modèle pour chaque individu. L'objectif est de découper l'histogramme de la figure 2.12 en plusieurs classes en sachant que plus le prédicteur linéaire d'un individu est faible, plus cet individu possède une probabilité de survie élevée.

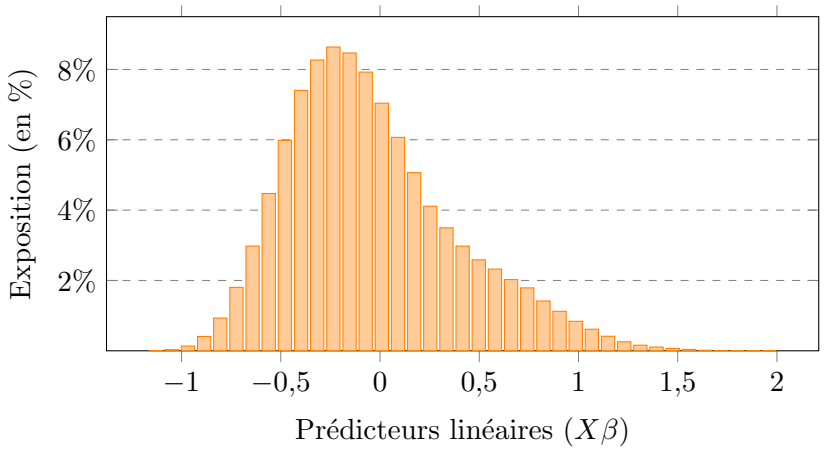


Figure 2.12 : Histogramme des prédicteurs linéaires $X\beta$ issus du modèle à risques proportionnels de Cox

*Soient $E = \{x_i : i \in [\![1,n]\!]\}$ l'ensemble des observations, g son centre de gravité et $\left(E_j\right)_{j\in[\![1,K]\!]}$ les K classes disjointes deux à deux telles que $E = \bigcup_{j=1}^{K} E_j$ (avec $n_j = \text{Card}(E_j)$ et $g_j = \text{c.d.g}(E_j)$), alors la décomposition de Huygens permet d'exprimer l'inertie totale de E comme la somme des inerties inter-classe et intra-classe :
$\mathcal{I}_{\text{Totale}} = \mathcal{I}_{\text{Inter}} + \mathcal{I}_{\text{Intra}} \iff \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) = \frac{1}{n}\sum_{j=1}^{K} n_j d^2(g_j, g) + \frac{1}{n}\sum_{j=1}^{K}\sum_{e\in E_j} d^2(e, g)$

En raison du volume important de données, 20 C.A.H. ont été réalisées sur des sous-ensembles de 50 000 données chacun, selon le principe du *Bagging* (cf. sous-section 3.1.1), plutôt qu'une seule classification ascendante hiérarchique. Répéter ce processus permet d'assurer la stabilité de la classification. Avant de procéder aux C.A.H., le nombre de classes à créer doit être défini. Pour cela, il faut observer la courbe de l'inertie en fonction du nombre de classes. Les courbes d'inertie de chaque C.A.H. sont représentées dans la figure 2.13 ainsi que la courbe moyenne (en noire). Le nombre de classes choisi est celui à partir duquel la baisse d'inertie n'est plus significative. D'après la courbe d'inertie moyenne, il a été décidé de créer 5 classes de durée de vie des contrats puisqu'au delà de 5, le gain en performance est minime.

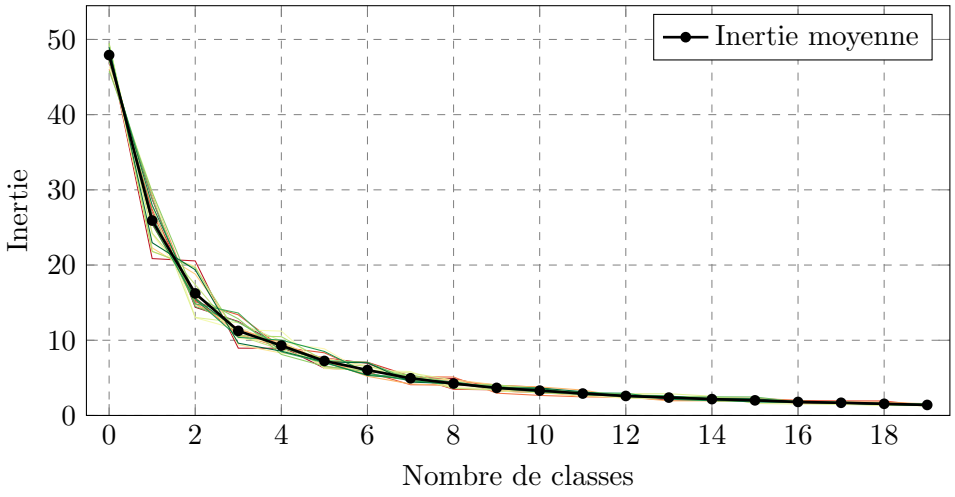


FIGURE 2.13 : Courbes d'inertie en fonction du nombre de classes pour les 20 C.A.H. et la courbe d'inertie moyenne

Les valeurs maximales et minimales des prédicteurs linéaires pour chaque classe créée par chacune des C.A.H. sont récupérées ainsi que leurs moyennes dans la figure 2.14. Les bornes moyennes vont servir de bornes pour les classes de durée de vie des contrats :

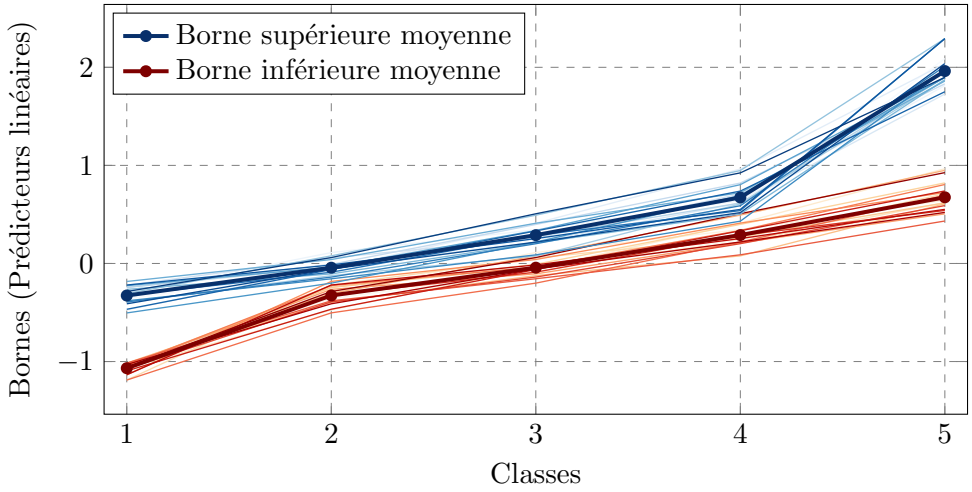


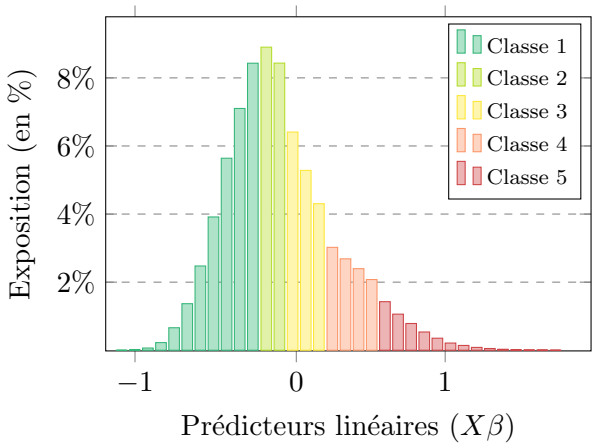
FIGURE 2.14 : Bornes supérieures et inférieures par classes pour les 20 C.A.H. et les bornes moyennes

Les valeurs des bornes pour chaque classe sont répertoriées dans le tableau 2.4 ci-dessous :

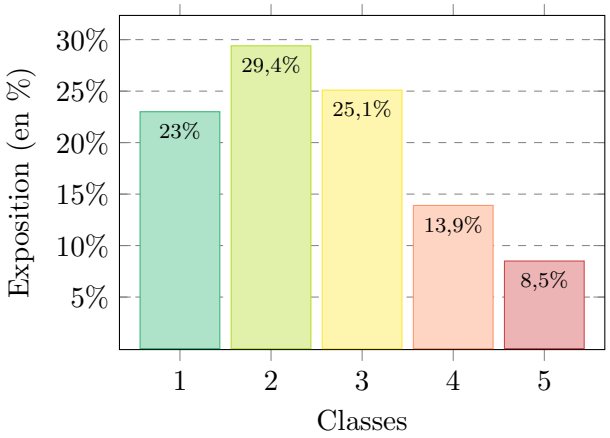
Classes	**1**	**2**	**3**	**4**	**5**
Bornes	$]-\infty\ ;\ -0,37[$	$[-0,37\ ;\ -0,06[$	$[-0,06\ ;\ 0,3[$	$[0,3\ ;\ 0,72[$	$]0,72\ ;\ +\infty[$

Table 2.4 : Valeurs des bornes pour chaque classe de durée de vie des contrats

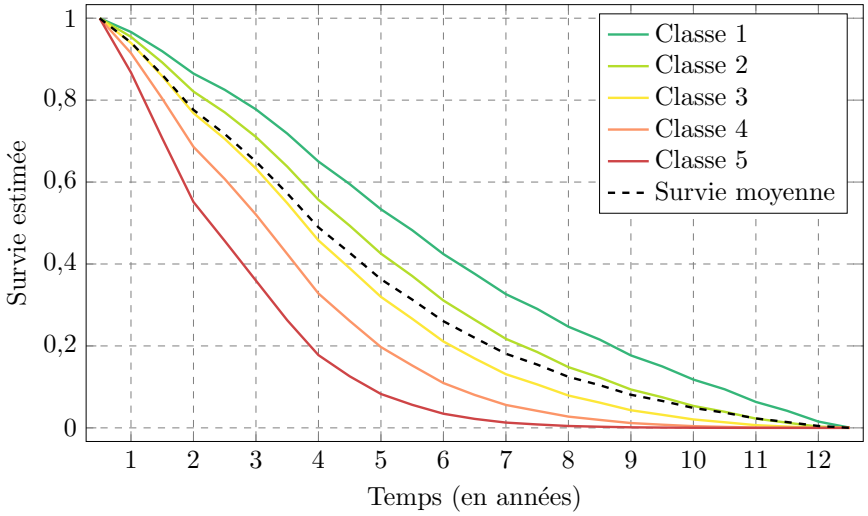
Les prédicteurs linéaires peuvent finalement être segmentés en 5 groupes, comme illustré dans la figure 2.15(a), ce qui donne 5 classes de durée de vie, dont la répartition est représentée dans la figure 2.15(b) ci-dessous. Plus la classe est petite (*i.e.* proche de 1), plus la probabilité de rester en contrat plus longtemps est élevée, comme l'indiquent les fonctions de survie moyennes estimées par le modèle à risques proportionnels de Cox, présentées dans la figure 2.15(c). Il convient également de noter que les classes 1 et 2 présentent une survie moyenne supérieure à celle de l'ensemble du portefeuille.



(a) Histogramme des prédicteurs linéaires $X\beta$



(b) Histogramme des classes de durée de vie



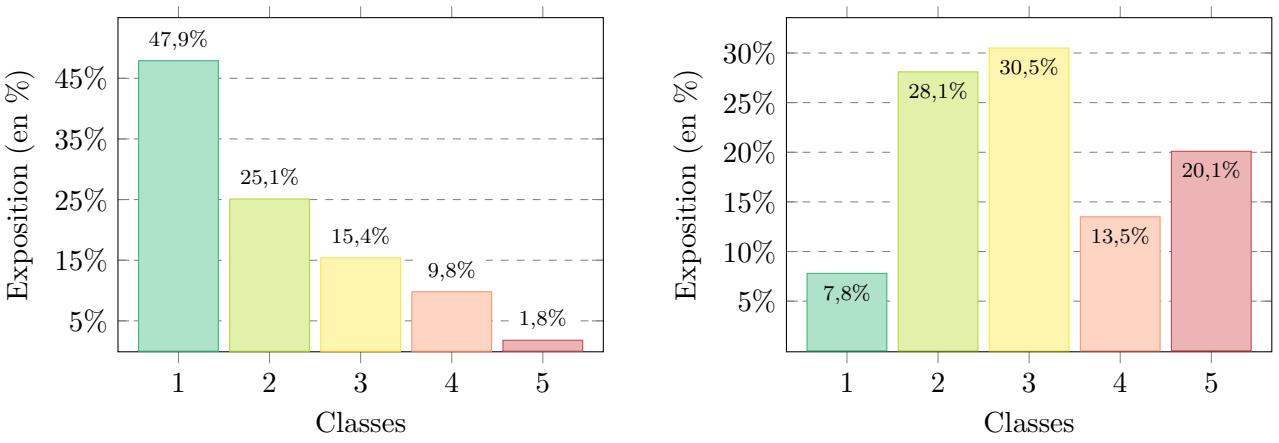
(c) Fonctions de survie moyennes estimées selon les classes

Figure 2.15 : Répartition des prédicteurs linéaires et des classes de durée de vie, et fonctions de survie moyennes estimées selon les classes

Comme dans la sous-section 2.5.4 où l'impact des modalités de plusieurs variables sur la survie estimée par le modèle à risques proportionnels de Cox est étudié, regardons comment se répartissent les modalités de ces variables dans les classes de durée de vie des contrats.

Concentrons-nous d'abord sur deux catégories socio-professionnelles : les retraités et les étudiants. Il a

été vu que ces deux groupes sont ceux qui ont respectivement le plus et le moins de chances de rester dans le portefeuille au cours du temps. Ce résultat est clairement vérifié par la figure 2.16. En effet, presque 50% des retraités se trouvent dans la classe 1 et moins de 2% dans la classe 5 alors que les étudiants sont représentés à 7,8% dans la meilleure classe et 20,1% dans la classe la plus mauvaise. Il est cependant important de noter que les étudiants ne constituent pas nécessairement des profils défavorables, car près de 50% d'entre eux se répartissent entre les classes 2 et 3.



(a) Pour les retraités (16,9%)

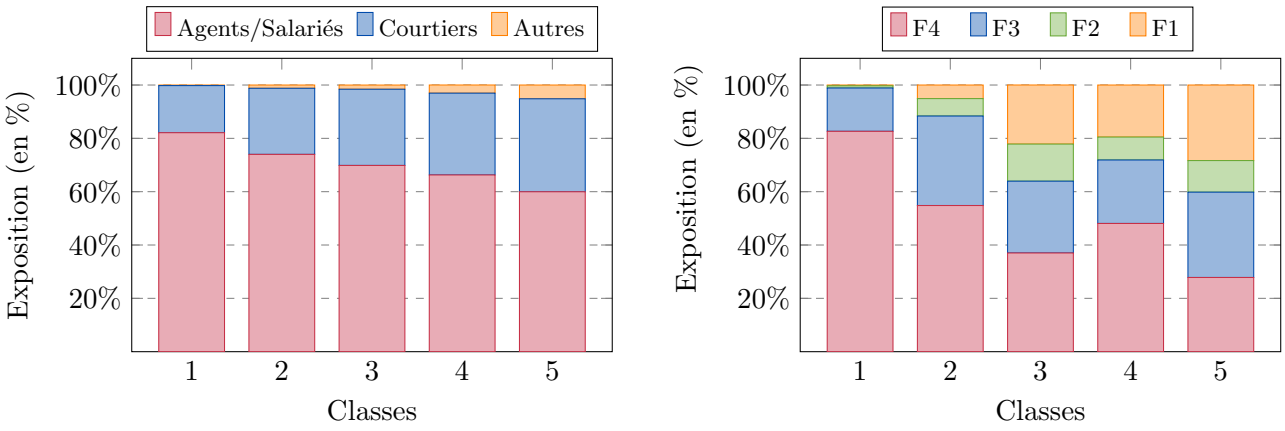
(b) Pour les étudiants (4,9%)

FIGURE 2.16 : Répartition de 2 catégories socio-professionnelles dans les classes de durée de vie

Nota Bene :

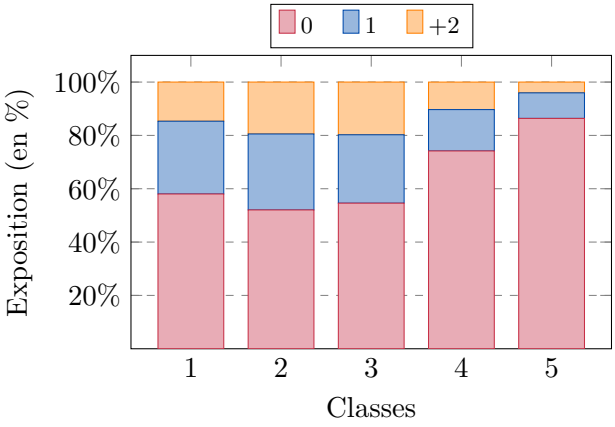
> *Les répartitions de chaque catégorie socio-professionnelle dans les classes de durée de vie sont disponibles en annexe A.6.*

En ce qui concerne le réseau de distribution, la formule souscrite et le nombre de contrats différents détenus par les assurés, les résultats de la sous-section 2.5.4 sont encore vérifiés dans la figure 2.17. Plus de 80% des souscriptions réalisées auprès d'agents généraux sont regroupées dans la classe 1, tandis que ce chiffre tombe à 60% dans la dernière classe. De même, une formule complète est un indicateur de fidélité, puisque plus de 99% des souscriptions en formules F3 ou F4 se trouvent dans la première classe, contre à peine 60% dans la cinquième classe. Enfin, la figure 2.17(c) montre que le multi-équipement, *i.e.* le fait de posséder d'autres contrats d'assurance auprès du même assureur, joue un rôle significatif sur la durée de vie des contrats. Par exemple, à peine plus de 15% d'assurés possédant au moins un autre contrat sont présents dans la classe 5 contre plus de 40% dans les 3 premières classes.



(a) Selon le réseau de distribution

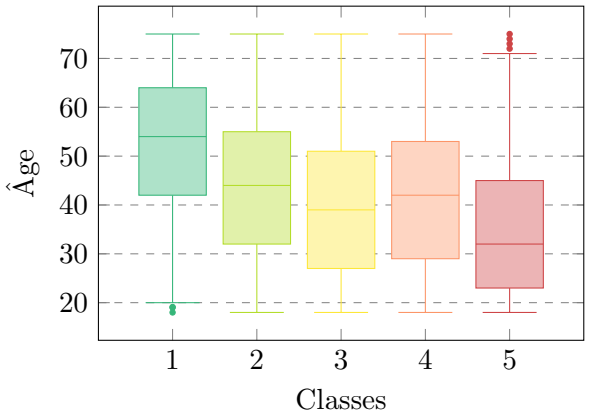
(b) Selon la formule souscrite



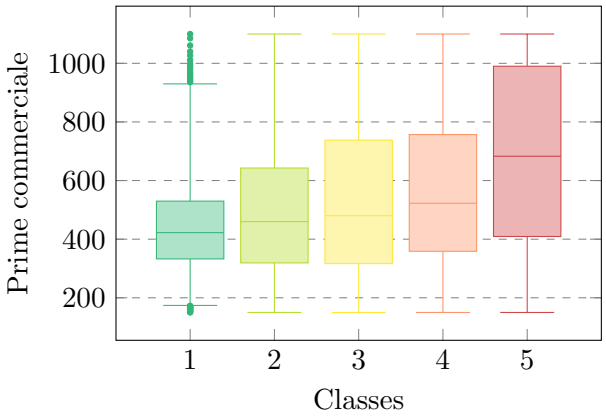
(c) Selon le nombre de contrats différents

Figure 2.17 : Répartition des modalités de plusieurs variables dans les classes de durée de vie

Concernant les variables numériques continues, des *boxplots* (ou boîtes à moustache) ont été représentés dans la figure 2.18 pour l'âge du conducteur et le montant de la prime commerciale à l'entrée dans le portefeuille. Pour commencer, l'âge du conducteur est décroissant avec les classes. Ce résultat était attendu puisque l'âge est fortement corrélé avec les catégories socio-professionnelles. Les personnes plus âgées, comme les retraités, sont en majorité dans la classe 1 avec un âge médian de 54 ans, alors que les jeunes sont énormément dans la classe 5 (âge médian égal à 32 ans). À l'inverse, le montant de cotisation à l'entrée dans le portefeuille tend à augmenter pour les classes de moindre qualité, passant d'un tarif médian d'environ 420€ dans la classe 1 à près de 700€ dans la classe 5. L'explication rationnelle de ce phénomène est qu'un montant de cotisation plus bas réduit la probabilité que l'assuré quitte le portefeuille pour trouver un contrat plus avantageux ailleurs.



(a) Boxplots de l'âge du conducteur par classes

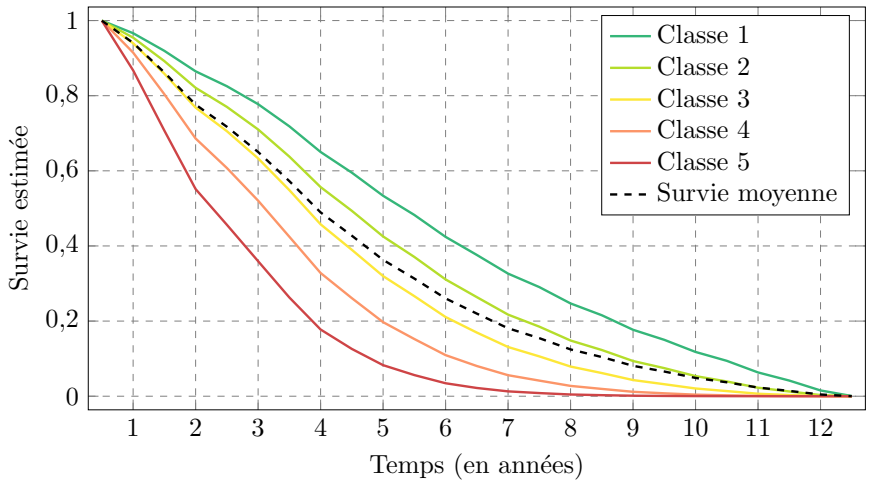


(b) Selon la prime commerciale à l'entrée

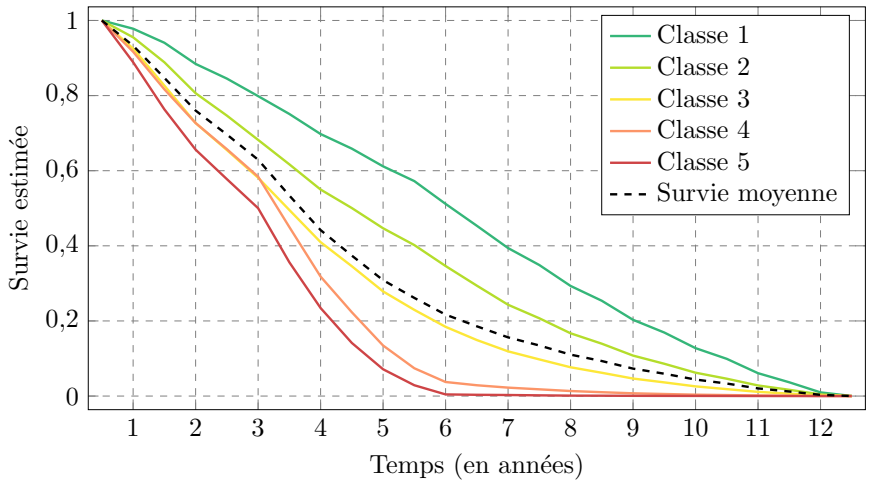
Figure 2.18 : Boxplots par classes selon l'âge du conducteur et la prime commerciale à l'entrée dans le portefeuille

2.8 Comparaison avec l'estimateur de Kaplan-Meier

Comme expliqué à la section 2.5, la durée de vie des contrats pour cette étude a été modélisée par un modèle à risques proportionnels de Cox et non via l'estimateur de Kaplan-Meier. Ce choix a été justifié par la capacité du modèle à risques proportionnels de Cox à fournir une analyse plus robuste et détaillée des effets des covariables sur la survie, ce qui est essentiel pour répondre aux questions spécifiques de l'étude. Pour appuyer ces arguments, les fonctions de survies moyennes estimées par les deux méthodes selon les classes ont été tracées dans la figure 2.19 ci-dessous.



(a) Fonctions de survie moyennes estimées par le modèle à risques proportionnels de Cox selon les classes



(b) Fonctions de survie moyennes estimées par l'estimateur de Kaplan-Meier selon les classes

FIGURE 2.19 : Fonctions de survie moyennes estimées par le modèle à risques proportionnels de Cox et l'estimateur de Kaplan-Meier selon les classes

En se basant sur ces fonctions de survie moyennes, le modèle à risques proportionnels de Cox parvient à capturer de manière plus nuancée les différences de survie entre les différentes classes de durée de vie des contrats, fournissant ainsi une estimation plus fidèle. En revanche, les fonctions de survie estimées par Kaplan-Meier vont certes dans le même sens mais ont tendance à être plus aplaties,

conduisant à une sous-estimation notable de la survie, en particulier pour les classes 4 et 5 où plusieurs individus survivent en pratique plus de 6 ans, ce qui ne semble pas être le cas à la vue de la courbe moyenne de la classe 5. Cela souligne la supériorité du modèle à risques proportionnels de Cox pour évaluer de manière robuste et plus précise les probabilités de survie des contrats du portefeuille étudié dans le calcul de la valeur client future.

2.9 Synthèse sur la durée de vie des contrats

Ainsi, bien que divers modèles de durée de vie des contrats aient été construits, la probabilité qu'un individu i appartiennent encore au portefeuille à tous les temps t, pour $t \in [\![0, T_h]\!]$, notée $\mathbb{P}_i(t)$ dans la formule de la valeur client future utilisée pour notre étude, sera modélisée grâce à la fonction de survie estimée par le modèle à risques proportionnels de Cox. Le choix de ce modèle par rapport à l'estimateur de Kaplan-Meier ou le modèle de Cox stratifié est justifié par sa facilité d'implémentation, ses bonnes performances, sa grande interprétabilité et sa capacité à considérer la censure à droite.

L'étude des fonctions de survie moyennes estimées par le modèle retenu selon les modalités de plusieurs variables, ainsi que la création de classes de durée de vie des contrats, créées uniquement dans un but d'analyse plus approfondie du portefeuille, ont permises de rendre compte de l'impact de certaines caractéristiques des contrats sur leur durée de vie. Par exemple, un contrat, ayant la couverture la plus étendue (formule F4), souscrit via un agent par un retraité déjà assuré pour un autre contrat chez cet assureur aura en moyenne une plus grande durée de vie estimée qu'un nouveau client étudiant ayant souscrit pour un contrat avec une formule simple (F1) via *internet* ou via un courtier d'assurance.

Maintenant que la probabilité de rétention $\mathbb{P}_i(t)$ a été modélisé, il convient de projeter les marges espérées individuelles à chaque année d'assurance pour les individus du portefeuille d'affaires nouvelles, notées $\text{Marge}_i(t)$ pour tout $i \in [\![1, n]\!]$ et $t \in [\![0, T_h]\!]$ dans la formule de la valeur client future suivante

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \frac{\text{Marge}_i(t) \times \mathbb{P}_i(t)}{(1+r)^t} \tag{2.22}$$

Chapitre 3

Modèle de projection de marges

Ce chapitre, explique comment ont été projetées les marges espérées (ou anticipées) individuelles notées $\text{Marge}_i(t)$ dans la formule suivante de la valeur client potentielle du $g^{\text{ème}}$ cluster

$$VC_g = \sum_{i=1}^{n_g} \sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \boxed{\text{Marge}_i(t)}}{(1+r)^t} \times Tr_i(v_g)$$

3.1 Présentation du modèle XGBoost

Pour rappel, en l'absence d'informations sur les coûts annuels des contrats, il est impossible de calculer les marges annuelles pour chaque contrat. Il convient donc de calculer les marges espérées (ou anticipées) grâce aux primes pures et commerciales. Puisque ces deux types de primes sont disponibles pour tous les individus du portefeuille complet pendant toute leur période d'observation, celles-ci peuvent être projetées sur un certain horizon de temps, noté T_h, et les marges espérées individuelles peuvent être calculées pour chaque année. Pour projeter ces primes, une méthode de *Machine Learning* très puissante est utilisée : l'*eXtrem Gradient Boosting* communément appelé XGBoost.

3.1.1 *Boosting* vs. *Bagging*

Avant de définir le modèle XGBoost, il convient de s'intéresser à l'apprentissage d'ensemble ou *ensemble learning*. Il s'agit d'une technique en *machine learning* où plusieurs modèles (souvent appelés modèles de base ou apprenants) sont combinés pour améliorer la performance prédictive globale par rapport à un modèle unique. Il existe plusieurs types de méthodes d'apprentissage d'ensemble, parmi lesquelles figurent le *Bagging* (contraction de *Bootstrap Aggregating*) et le *Boosting*.

Le *Bagging* améliore la précision des modèles d'apprentissage en combinant les prédictions de plusieurs modèles de base, souvent des arbres de décision, entraînés en parallèle sur des sous-ensembles de données indépendants. Voici comment se déroule le processus de *Bagging* :

1. *Bootstrap Sampling* : création de plusieurs sous-échantillons de l'ensemble de données d'entraînement en utilisant la méthode de *bootstrap* (échantillonnage avec remplacement).
2. *Training* : entraînement d'un modèle de base sur chaque sous-échantillon.
3. *Aggregation* : combinaison des prédictions de tous les modèles de base pour obtenir une prédiction finale. Pour les problèmes de classification, cela peut être fait par vote majoritaire, et pour les problèmes de régression, par moyenne.

Le *Boosting* quant à lui améliore la performance prédictive en entraînant séquentiellement des modèles faibles, chaque nouveau modèle corrigeant les erreurs des précédents pour réduire le biais et accroître la précision globale. Voici comment il procède :

1. *Sequential Training* : entraînement des modèles de base séquentiellement, chaque nouveau modèle étant formé pour corriger les erreurs des modèles précédents.
2. *Weight Adjustment* : les observations mal prédites par les modèles précédents reçoivent un poids plus élevé pour que le modèle suivant se concentre sur ces erreurs.
3. *Combinaison* : les prédictions finales sont obtenues par une combinaison pondérée des prédictions de tous les modèles.

La figure 3.1 illustre bien ces deux processus et la différence entre ces deux types de méthodes d'apprentissage d'ensemble :

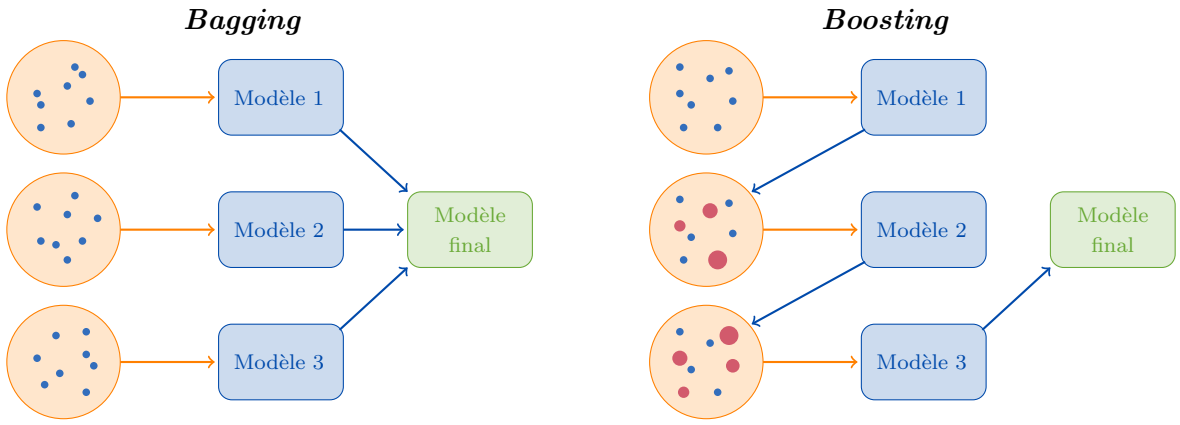


FIGURE 3.1 : Schéma explicatif du *Bagging* et du *Boosting*

Ces deux méthodes d'apprentissage d'ensemble présentent des avantages et des inconvénients dont les principaux sont listés dans le tableau 3.1 ci-dessous :

	Bagging	***Boosting***
Avantages	◇ Réduction de la variance du modèle ◇ Moins sensible au sur-apprentissage par rapport à un modèle unique	◇ Réduction du biais du modèle ◇ Transforme des modèles faibles en modèles forts et performants
Inconvénients	◇ Les modèles de base doivent être indépendants les uns des autres ◇ Approche peu efficace avec des modèles biaisés	◇ Susceptible au sur-apprentissage si les modèles de base sont complexes. ◇ Plus complexe et gourmand en temps de calcul

TABLE 3.1 : Avantages et inconvénients entre le *Bagging* et le *Boosting*

Pour la suite des travaux, la méthode du *Boosting* a été privilégiée car celle-ci est particulièrement efficace pour améliorer la performance des modèles faibles en réduisant à la fois le biais et la variance. Son processus séquentiel permet une adaptation dynamique aux erreurs des modèles précédents, ce qui conduit à une amélioration continue du modèle final. Bien que le *Boosting* soit plus susceptible au sur-apprentissage, le modèle XGBoost qui sera utilisé inclut des régularisations et des mécanismes de contrôle du sur-apprentissage, ce qui rend la modélisation plus robuste et efficace. Enfin, pour cette étude, il convient de projeter les marges espérées sur une horizon de temps important. Les résultats

seront alors de plus en plus volatils. L'utilisation de modèles XGBoost permet d'assurer de bonnes performances malgré cette perte de précision au cours du temps.

3.1.2 Définition du modèle

Le modèle *eXtrem Gradient Boosting*, ou XGBoost, a été développé par Chen et Guestrin (2016). Comme son nom l'indique, cet algorithme repose sur le *Gradient Boosting*, une technique d'apprentissage supervisé pour la classification et la régression. XGBoost est donc une implémentation avancée de cette méthode qui construit un modèle prédictif sous la forme d'une ensemble de modèles faibles, généralement des arbres de décision. Voici une approche simplifiée du modèle :

1. *Initialisation* : le processus débute par une prédiction initiale, en général la moyenne des valeurs cibles pour un problème de régression ou le mode pour un problème de classification.
2. *Ajout des arbres* : à chaque étape de *boosting*, un nouvel arbre de décision est ajouté pour corriger les erreurs des prédictions précédentes. Chaque arbre est ajusté pour prédire les résidus de la somme des arbres précédents.
   - ◇ Fonction de Perte : l'algorithme minimise une fonction de perte $l(y, \hat{y})$, où y est la valeur réelle et $\hat{y}$ est la prédiction.
   - ◇ Gradient : le gradient de la fonction de perte est calculé pour chaque point de données, indiquant la direction et l'ampleur de l'erreur.
   - ◇ Poids des Arbres : les arbres sont pondérés en fonction de leur capacité à corriger les erreurs des prédictions précédentes.
3. *Régularisation* : pour éviter le sur-apprentissage, XGBoost utilise des techniques de régularisation (Ridge ou Lasso).
4. *Prédiction Finale* : la prédiction finale est obtenue en additionnant les prédictions pondérées de tous les arbres.

Ainsi, XGBoost combine plusieurs arbres de décision simples qui sont entraînés ensemble, ce qui permet de réduire le biais du modèle. La méthode de *boosting* utilisée par cet algorithme entraîne les prédicteurs de manière séquentielle et adaptative, en accordant plus de poids aux observations mal prédites par le modèle précédent.

Traduisons cela mathématiquement : soit $X \in \mathcal{M}_{n,p}(\mathbb{R})$ la matrice des n observations des p variables explicatives (caractéristiques) et $y \in \mathbb{R}^n$ le vecteur des n observations de la variable d'intérêt. Un ensemble de K arbres de décision prédit l'observation y_i, pour $i \in [\![1, n]\!]$ de la façon suivante

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), \;\; f_k \in \mathcal{F} \tag{3.1}$$

où $\mathcal{F}$ est l'espace des arbres de décision et, pour tout $k \in [\![1, K]\!]$, f_k retourne le poids associé à la feuille que l'arbre k assigne aux caractéristiques X_i.

L'objectif est de minimiser la fonction de perte suivante

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.2}$$

où

- ▷ l est une fonction de perte convexe et différentiable (par exemple l'erreur quadratique pour la régression)

- ▷ $\hat{y}_i$ est la prédiction de l'observation y_i
- ▷ K est le nombre d'arbres créés par le modèle (hyperparamètre)
- ▷ f_k représente le $k^{\text{ème}}$ arbre du modèle
- ▷ Ω est un terme de régularisation qui contrôle la complexité du modèle pour éviter le sur-apprentissage

Chaque arbre f_k est construit pour minimiser l'erreur résiduelle des prédictions précédentes. En notant η le taux d'apprentissage (*learning rate*) qui contrôle la contribution de chaque arbre et $f_k(X_i)$ la prédiction de l'arbre k pour l'observation i, il est possible d'écrire

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + \eta f_k(X_i)$$

Le terme de régularisation Ω est défini comme suit

$$\Omega(f) = \gamma T + \frac{\lambda}{2}\|w\|^2$$

avec T le nombre de feuille de l'arbre f, w le vecteur des poids associés aux feuilles de f, γ et λ sont des hyperparamètres du modèle qui contrôlent la régularisation., et enfin $\|\cdot\|$ est la norme ℓ_1 (Lasso) ou ℓ_2 (Ridge)*.

Voici un schéma qui représente le fonctionnement du XGBoost :

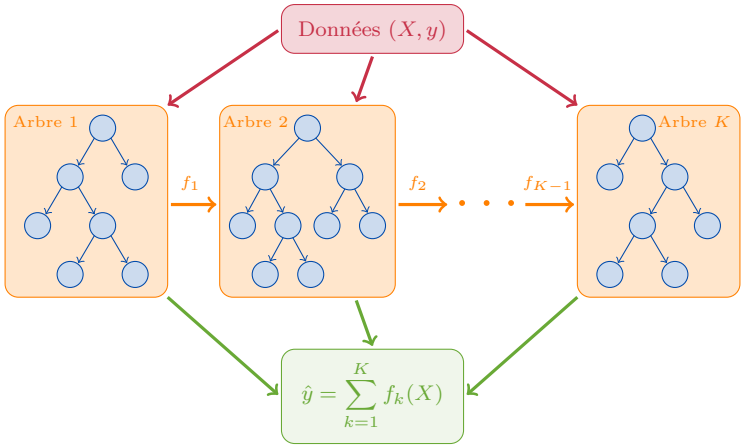


FIGURE 3.2 : Schéma explicatif du modèle XGBoost

S'agissant d'un modèle très sophistiqué, il apparaît très clairement dans la définition ci-dessus qu'un grand nombre d'hyperparamètres peuvent affecter la performance. Il faut distinguer deux familles d'hyperparamètres : les paramètres propres au *boosting* comme la profondeur maximale de l'arbre ou le taux d'apprentissage η, et les paramètres d'apprentissage tels que la fonction de perte et la métrique d'évaluation.

3.1.3 Interprétabilité, fonction de perte et métriques d'évaluation

Interprétabilité

Bien que XGBoost soit souvent considéré comme un modèle "boîte noire" en raison de sa complexité et de son opacité, plusieurs techniques permettent d'expliquer et d'interpréter ses prédictions.

*Norme ℓ_1 : $\|\cdot\|_1 : x \in \mathbb{R}^d \longmapsto \sqrt{\sum_{i=1}^{d} |x_i|} \in \mathbb{R}_+$ Norme ℓ_2 (euclidienne) : $\|\cdot\|_2 : x \in \mathbb{R}^d \longmapsto \sqrt{\sum_{i=1}^{d} |x_i|^2} \in \mathbb{R}_+$

L'interprétabilité est cruciale pour comprendre les décisions du modèle dans sa construction. Parmi ces techniques figurent, par exemple, les graphiques d'importance des variable permettant de comprendre quelles variables ont le plus d'impact sur les prédictions. Un autre outil de visualisation très utilisé est l'analyse des valeurs de Shapley (ou *shap values*), introduits par Shapley (1953), permettant d'observer l'effet marginal de chaque caractéristique ou leurs dépendances entres elles. Ces *shap values* permettent ainsi une interprétation locale de modèles complexes comme le XGBoost.

Fonction de perte et métriques d'évaluation

Dans XGBoost, comme dans tout algorithme d'apprentissage supervisé, il est crucial de définir une métrique et une fonction de perte appropriées pour évaluer les performances du modèle et orienter l'optimisation pendant l'entraînement. La fonction de perte est utilisée pour ajuster les prédictions du modèle, tandis que la métrique est utilisée pour évaluer la qualité des prédictions. Ces deux éléments dépendent souvent du type de problème à résoudre et des objectifs spécifiques de l'étude. Dans notre contexte, l'objectif est de projeter des primes ayant des valeurs continues, ce qui en fait un problème de régression. Penchons-nous sur la fonction de perte et les métriques d'évaluation choisies.

La fonction de perte mesure la différence entre les prédictions du modèle $\hat{y}$ et les valeurs réelles des données y. Elle guide l'optimisation du modèle en minimisant cette différence. La fonction de perte choisie est la plus classique* des problèmes de régression : l'erreur quadratique moyenne (MSE : *Mean Squared Error*) définie comme suit

$$l(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.3}$$

Les métriques d'évaluation sont des mesures quantitatives utilisées pour évaluer la qualité des prédictions d'un modèle pendant son entraînement et sa validation. Pour cette étude, 4 métriques différentes sont calculées :

- ▷ **Racine carrée de l'erreur quadratique moyenne (RMSE)** :
  Appelée RMSE pour *Root Mean Squared Error*, il s'agit tout simplement de la racine carrée de la MSE définie ci-dessus
  $$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$
  Celle-ci punit sévèrement les erreurs importantes, ce qui peut être utile si de grandes erreurs doivent être évitées.
- ▷ **Erreur Absolue Moyenne (MAE)** :
  Appelée MAE pour *Mean Absolute Error*, il s'agit de la moyenne des valeurs absolues des erreurs comme son nom l'indique
  $$\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$
  La MAE donne une mesure plus robuste des erreurs car elle n'amplifie pas les grandes erreurs autant que la RMSE.
- ▷ **Coefficient de détermination (Score R^2) :**

*En raison de ses propriétés différentiables facilitant l'optimisation.

Le coefficient de détermination mesure la proportion de la variance des observations y qui est expliquée par les prédictions $\hat{y}$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{où } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ est la moyenne des observations}$$

Un R^2 égal à 1 indique un modèle parfait, alors qu'un R^2 nul indique que le modèle ne fait pas mieux qu'une prédiction moyenne constante.

▷ **Coefficient de Gini normalisé :**

Le coefficient (ou indice) de Gini a été introduit par GINI (1912) et est dérivé de la courbe de LORENZ (1905). En économie, cette dernière est une représentation graphique de la distribution cumulée des revenus ou des richesses. Elle trace la proportion cumulée du revenu total de la population (en ordonnées) en fonction de la proportion cumulée de la population (en abscisses). La courbe de Lorenz peut être généralisée mathématiquement comme une représentation graphique (un *P-P plot*) comparant la distribution d'une variable à une distribution uniforme hypothétique de cette variable.

Pour calculer le coefficient de Gini, les observations y sont ordonnées en fonction des valeurs croissantes des prédictions $\hat{y}$. Soit $\big(\sigma_{\hat{y}}(1), ..., \sigma_{\hat{y}}(n)\big)$ les indices de $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)$ tels que $\hat{y}_{\sigma_{\hat{y}}(1)} < ... < \hat{y}_{\sigma_{\hat{y}}(n)}$, les observations $(y_1, ..., y_n)$ sont alors triées selon cet ordre : $(y_{\sigma_{\hat{y}}(n)}, ..., y_{\sigma_{\hat{y}}(n)})$. La formule du coefficient de Gini est la suivante

$$G(y, \hat{y}) = \frac{2\sum_{i=1}^{n} i y_{\sigma_{\hat{y}}(i)}}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n} \tag{3.4}$$

Cependant, pour notre étude, le coefficient de Gini normalisé (ou ajusté) a été préféré. Il s'agit d'une version modifiée du coefficient de Gini qui vise à améliorer sa comparabilité et son interprétabilité, notamment en tenant compte de certaines de ses limites, comme les différences de taille de population ou les biais potentiels. Il est défini comme le rapport entre le Gini où les observations sont triées dans l'ordre croissant des valeurs des prédictions et le Gini où les observations sont triées dans leur ordre croissant. En d'autres termes, il peut être défini à partir de l'équation (3.4) comme

$$G_{\text{Norm}}(y, \hat{y}) = \frac{G(y, \hat{y})}{G(y, y)} \tag{3.5}$$

Comme pour le R^2, un coefficient de Gini normalisé égal à 1 indique un modèle parfait et de moins en moins précis qu'il se rapproche de 0.

3.1.4 Pourquoi utiliser le XGBoost ?

Pour le modèle de projection de marges, le choix de XGBoost est motivé par des raisons de performance prédictive, bien que ce modèle soit plus complexe et moins interprétable, étant souvent qualifié de "boîte noire". En effet, des modèles plus simples comme les modèles linéaires généralisés (cf. sous-section 4.1.2) avaient déjà été testés lors de travaux antérieurs (REGNAULT (2023)), mais XGBoost s'est avéré plus performant pour ce type de projection de marges. Grâce à son approche de boosting séquentiel, il permet une correction itérative des erreurs, maximisant ainsi la précision des projections. De plus, le temps de calcul s'est avéré être court, rendant ce choix tout à fait acceptable et compatible avec les contraintes opérationnelles. Ce modèle se montre donc plus adapté aux exigences de précision et de robustesse nécessaires pour cette étude.

3.2 Traitement des données et méthodologie

Maintenant que la théorie a été expliquée, l'application au portefeuille peut être abordée. Avant de projeter les marges espérées, il est nécessaire d'expliquer comment la base de données a été retraitée.

3.2.1 Traitement des données

Afin de projeter les marges espérées jusqu'à un certain horizon T_h, le portefeuille complet a dû être transformé en base image afin d'être capable de suivre un contrat et ses primes dans le temps selon les caractéristiques initiales de l'assuré. Par exemple les tableaux 3.2 et 3.3 représentent des échantillons respectivement du portefeuille complet et de la base image :

Contrat	**Année d'observation**	**Prime commerciale**	**Prime pure**	$\cdots$
5	2009	163,32	83,83	$\cdots$
5	2010	132,36	80,91	$\cdots$
6	2009	206,68	61,32	$\cdots$
6	2010	198,02	60,45	$\cdots$
8	2009	630,84	524,25	$\cdots$
8	2010	541,92	448,88	$\cdots$
8	2011	486,76	340,00	$\cdots$
8	2012	465,64	320,58	$\cdots$
8	2013	437,72	290,78	$\cdots$
8	2014	493,08	281,90	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 3.2 : Échantillon du portefeuille complet

Contrat	**Année d'entrée**	**PC(0)**	**PC(1)**	**PC(2)**	**PP(0)**	**PP(1)**	**PP(2)**	$\cdots$
5	2009	163,32	132,36	-	83,83	80,91	-	$\cdots$
6	2009	206,68	198,02	-	61,32	60,45	-	$\cdots$
8	2009	630,84	541,92	486,76	524,25	448,88	340,00	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 3.3 : Échantillon de la base image

Dans ces exemples, les contrats 5 et 6 ont été dans le portefeuille pendant 2 ans et le contrat 8 pendant 6 ans. Les caractéristiques initiales des contrats et les primes lors de ces années d'assurance sont récupérées et ne forment qu'une seule ligne dans la base image.

De plus, le changement de situation, également appelé avenant, est un facteur important à prendre en compte dans la vie d'un contrat d'assurance. Un changement de véhicule, de domicile ou de situation maritale constituent un motif courant pour que l'assuré résilie son contrat. Les données disponibles ne renseignent pas le motif de résiliation, observer ces changements peut donc permettre une meilleur compréhension pour fidéliser les clients.

Les avenants qui peuvent être considérés dans l'étude et leurs nombres dans le portefeuille sont les suivants :

- ▷ Changement de véhicule : 214850

- ▷ Changement de formule (F1, F2, F3 et F4) : 115661
- ▷ Changement du nombre de garanties (7 garanties au total) : 109560
- ▷ Changement du nombre de contrats différents détenus par l'assuré : 267194
- ▷ Changement du niveau de franchise dommage et bris de glace : 192437 et 152620
- ▷ Changement d'adresse : 141828

Cela a permis de créer une base image avec prise en compte de ces avenants : lorsqu'un avenant a lieu, une nouvelle ligne est créée comme s'il s'agissait d'un nouveau contrat. Voici ci-dessous, dans le tableau 3.4, un échantillon de cette base image :

Contrat	**Année d'entrée**	**Identifiant SRA**	**Formule**	**PC(0)**	**PC(1)**	**PC(2)**	···
5	2009	CI04166	**F3**	163,32	-	-	···
5	2010	CI04166	**F2**	132,36	-	-	···
6	2009	NA09054	F1	206,68	198,02	-	···
8	2009	**BM59689**	**F4**	630,84	-	-	···
8	2010	**RE32005**	**F3**	541,92	486,76	465,64	···

TABLE 3.4 : Échantillon de la base image illustrant les avenants

Cet exemple montre que le contrat 5 a changé de formule en 2010, adoptant une formule moins complète, ce qui diminue directement le montant de sa prime commerciale. Cependant, il ne sera pas observé plus longtemps. Le contrat 6 est quant à lui observé pendant 2 ans sans qu'aucun changement n'ait lieu. Enfin le contrat 8 change de véhicule et de formule en 2010 après une année d'assurance. Ces changements diminue sa prime commerciale, il restera en portefeuille encore plusieurs années.

Ces avenants expliquent les changements drastiques de tarifs d'une année à l'autre. L'avantage certain de prendre en considération ces avenants est d'obtenir des résultats plus interprétables en observant des primes à priori croissantes dans le temps. Cependant, cela ne reflète pas bien l'évolution des primes puisqu'en pratique un changement n'entraîne pas forcément une résiliation comme supposé ici en créant un nouveau contrat de manière implicite (nouvelle ligne dans la base). Pour l'étude, les avenants ont d'abord été pris en compte afin d'évaluer comment les primes projetées doivent se comporter et si les résultats sont biens interprétables, puis la projection des primes a été réalisée sur la base image sans prise en compte des avenants. Les résultats sans la prise en considération des avenants a donné des résultats similaires à ceux avec leur prise en compte. C'est pourquoi le choix de la base image sans prise en compte des avenants, *i.e.* semblable au tableau 3.3, pour la suite de l'étude a été fait. Celle-ci servira de base d'apprentissage pour comprendre le comportement des assurés et apprendre à projeter les primes pures et commerciales selon les caractéristiques individuelles à l'entrée en portefeuille et le montant payé lors de la première année d'assurance afin de prédire les primes futures des affaires nouvelles.

Une amélioration possible pour considérer les changements de situation des assurés serait de créer un modèle de changement d'état via des chaînes de Markov. Ce modèle donnerait pour chaque individu à chaque instant la probabilité que les différents avenants s'opèrent. Ces probabilités vont ensuite servir à pondérer les primes considérants les avenants possibles pour obtenir une prime moyenne.

3.2.2 Méthodologie de la projection des marges espérées

Pour projeter les marges espérées individuelles dans le temps, une méthodologie précise a été mise en place. Tout d'abord, le choix d'un horizon de temps T_h a dû être fait. Puisque l'objectif final est de

réaliser une optimisation de la stratégie commerciale sur une base d'affaires nouvelles, la proportion de contrats en affaires nouvelles qui reste dans le portefeuille au fur et à mesure des années a été regardé. Cette proportion est de près de 5% pour $t = 9$ et 2,3% pour $t = 10$ comme le montre la figure 3.3, ce qui n'est pas négligeable. Cependant l'horizon choisi sera de 10 ans, soit $T_h = 9$, pour la suite de l'étude puisque les résultats au bout de 11 ans sont trop volatils et ne présentent pas de bons effets.

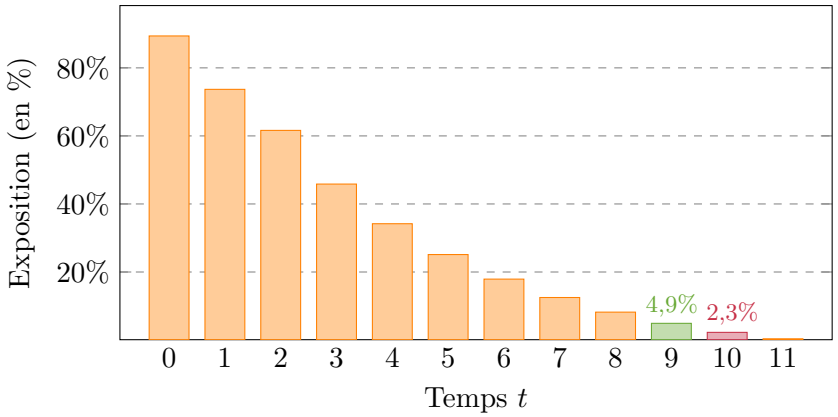


Figure 3.3 : Distribution de la proportion de contrats selon leur durée de vie pour le portefeuille d'affaires nouvelles

La seconde étape a été de projeter les primes pures et commerciales par année pour chaque individu. Pour cela, un modèle XGBoost a été créé pour chaque type de prime et pour chaque année. Par exemple, pour prédire les primes pures et commerciales de l'année $n+1$, 2 modèles (un pour chaque type de prime) sont entrainés sur un échantillon de données où les primes de l'année $n+1$ sont connues, puis les primes de l'année $n+1$ sont prédites sur l'ensemble du portefeuille. Les variables explicatives des modèles sont les informations initiales des assurés. Parmi celles-ci, la prime observée est remplacée par la prédiction des primes de l'année précédente (sauf pour prédire les primes de la $2^{\text{ème}}$ année d'assurance où ce sont bien les primes observées de la $1^{\text{ère}}$ année qui sont utilisées). Ainsi, dans cet exemple, pour projeter la prime pure (respectivement commerciale) de l'année $n+1$ d'assurance, une des variables explicatives du modèle est la prédiction de la prime pure (respectivement commerciale) de la $n^{\text{ème}}$ année. Imbriquer les prédictions successivement va certes rendre les résultats plus volatils mais cela permet d'obtenir toutes les primes pour chaque individu sur toute la période uniquement à partir de ses caractéristiques initiales et de ses primes lors de la souscription, ce qui sera l'idéal lorsque les prédictions seront appliquées sur le portefeuille d'affaires nouvelles.

Ainsi, un total de 18 modèles XGBoost ont été créés pour projeter les primes pures et commerciales pendant 10 ans. De plus, chaque modèle, pour chaque type de prime et pour chaque année, a subi une validation croisée afin d'obtenir les hyperparamètres offrant les meilleures performances pour les différentes métriques d'évaluation définies à la sous-section 3.1.3.

Enfin, les marges espérées individuelles par année ont été calculées selon la formule suivante

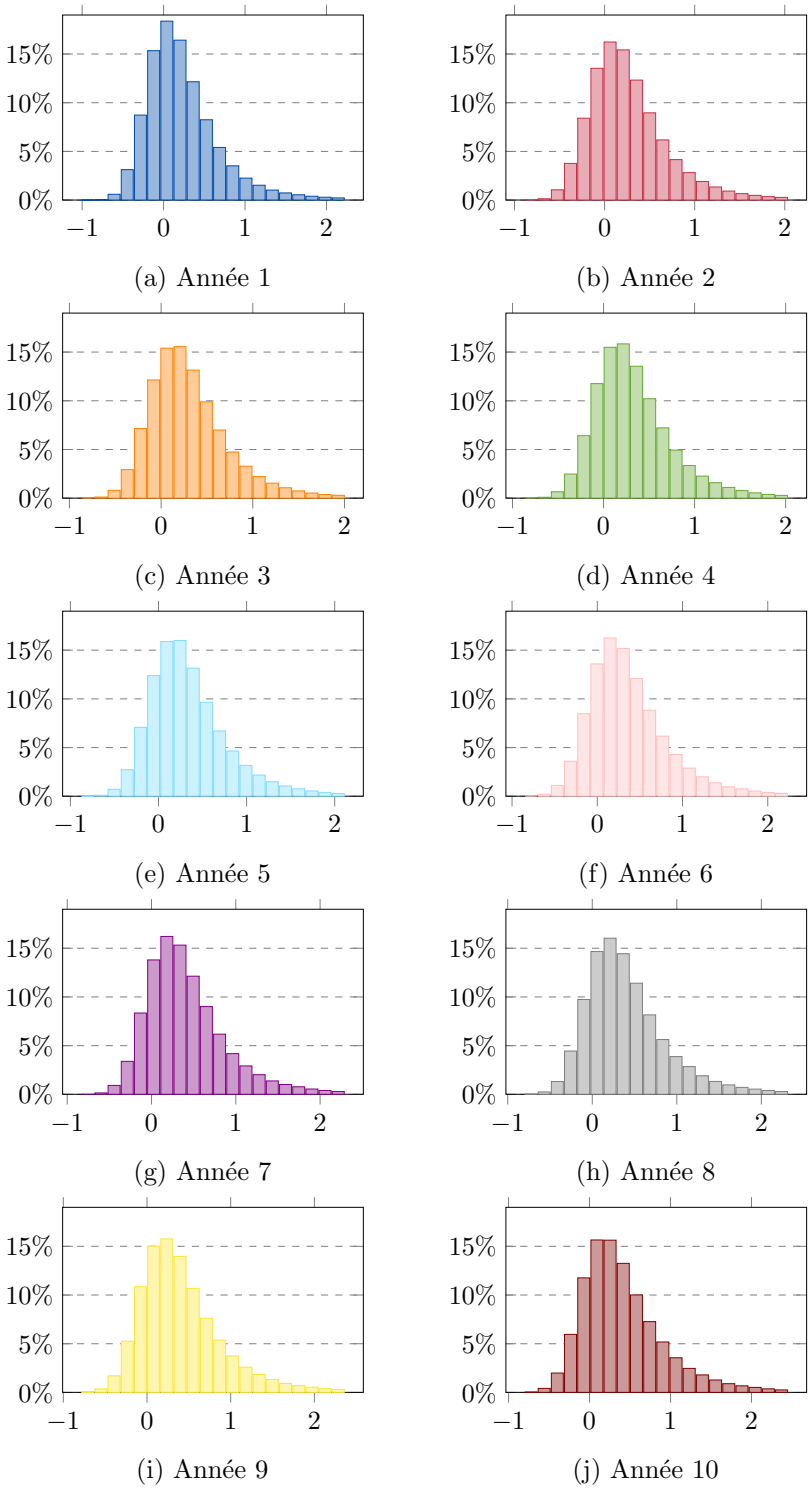
$$\text{Marge}_i(t) = \text{PC}_i(t) - \text{PP}_i(t) \times (1 + \text{Taux de chargements}) \tag{3.6}$$

Étant donné que l'analyse est effectuée du point de vue de l'assureur, la prime pure a été chargée pour calculer la marge espérée car les chargements ne représentent pas un gain pour l'assureur mais plutôt un montant destiné à couvrir ses frais. Ici, les chargements dépendent de la formule souscrite et sont données par l'assureur.

Cependant, il est important de vérifier que les marges espérées relatives, définies dans la formule (3.7) ci-dessous, sont biens aléatoires et reflètent la stratégie mise en place par l'organisme d'assurance pour chaque profil.

$$\text{Marge relative}_i(t) = \frac{\text{PC}_i(t) - \text{PP}_i(t) \times (1 + \text{Taux de chargements})}{\text{PP}_i(t) \times (1 + \text{Taux de chargements})} \tag{3.7}$$

Pour cela, les histogrammes des marges espérées relatives observées sont représentés pour chaque année d'assurance dans la figure 3.4 ainsi que la moyenne et l'écart-type des marges espérées relatives observées au cours du temps dans la figure 3.5 :



(a) Année 1 (b) Année 2

(c) Année 3 (d) Année 4

(e) Année 5 (f) Année 6

(g) Année 7 (h) Année 8

(i) Année 9 (j) Année 10

FIGURE 3.4 : Histogrammes des marges espérées relatives observées selon les années

Tout d'abord, il est clair que les marges espérées relatives observées sont toujours aléatoires comme en témoigne les 10 histogrammes de la figure 3.4. Ensuite, d'après la figure 3.5, la marge espérée relative moyenne est croissante au cours du temps, ce qui montre que l'écart entre les primes pures et commerciales n'est pas constant mais augmente dans le temps. Enfin, l'écart-type est certes décroissant les premières années mais devient croissant après 4 années.

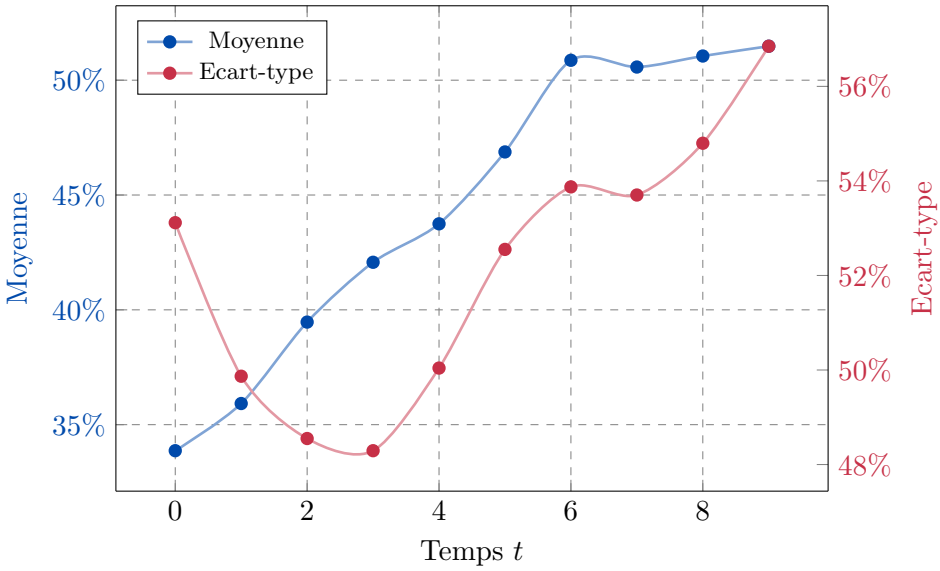


Figure 3.5 : Moyenne et écart-type des marges espérées relatives observées au cours du temps

Nota Bene :

> *Les marges présentées dans cette étude sont des marges espérées, définies comme l'écart entre la prime commerciale et la prime technique, cette dernière étant calculée à partir de la prime pure ajustée d'un taux de chargement. Contrairement au ratio combiné (CoR) du marché, qui intègre* ***ex post*** *l'ensemble des coûts réels (sinistralité observée, frais de gestion, acquisition, réassurance, etc.) et s'établissait à 98,7% en 2023 selon* France Assureurs *(2024), ces marges sont construites sur des hypothèses* ***ex ante*** *et ne tiennent pas compte des ajustements a posteriori liés à la survenue effective des sinistres. Ainsi, des marges élevées peuvent être observées dans ce cadre théorique, sans pour autant signifier que ces niveaux de rentabilité se matérialisent en pratique.*

Le schéma de la figure 3.6 synthétise la méthodologie de projection pour un individu i de caractéristiques initiales X_i. À partir de ses caractéristiques et de ses primes pure et commerciale à l'entrée en portefeuille, 18 modèles sont successivement créés pour prédire les primes pures et commerciales de chaque année jusqu'à 10 ans ($t = T_h = 9$), respectivement notées $\widehat{\mathrm{PP}}_i(\cdot)$ et $\widehat{\mathrm{PC}}_i(\cdot)$. Chaque année, les marges espérées sont alors calculées grâce à la formule (3.6).

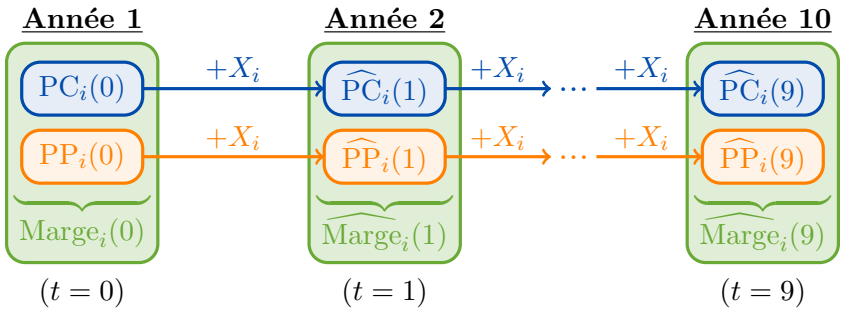


FIGURE 3.6 : Schéma récapitulatif de la méthode de projection des marges espérées

Comme dit plus tôt, l'intérêt principal de prédire les primes successivement en utilisant les prédictions des primes précédentes est d'obtenir une projection des primes pour tous les individus sur chaque période uniquement en ne connaissant que le montant de leurs primes à l'entrée dans le portefeuille. Par exemple, après projection des primes, l'échantillon de la base image du tableau 3.3 devient :

Contrat	**Année d'entrée**	**PC(0)**	**PP(0)**	**Marge(0)**	$\widehat{\textbf{PC}}(1)$	$\widehat{\textbf{PP}}(1)$	$\widehat{\textbf{Marge}}(1)$	$\widehat{\textbf{PC}}(2)$	$\widehat{\textbf{PP}}(2)$	$\widehat{\textbf{Marge}}(2)$	$\cdots$
5	2009	163,32	83,83	**52,31€**	166,23	85,12	**49,45€**	181,23	87,05	**61,80€**	$\cdots$
6	2009	206,68	61,32	**119,67€**	204,76	64,18	**113,69€**	205,03	68,34	**108,06€**	$\cdots$
8	2009	630,84	524,25	**-43,87€**	554,11	460,26	**-38,24€**	504,91	345,85	**59,80€**	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

TABLE 3.5 : Échantillon de la base image après projection des marges espérées

Le tableau 3.5 rend compte des très bonnes performances des modèles de projection des primes qui semblent réussir à bien capter les tendances générales. Les contrats 5 et 6 sont rentables dès le début, mais il est intéressant de noter que le contrat 8 est déficitaire pour l'assureur pendant les deux premières années. Ces marges espérées négatives au début de l'observation du contrat s'expliquent par les valeurs de primes commerciales proches des primes pures, qui une fois chargées seront supérieures. La prime pure prédite diminue de près de 115€ entre la deuxième et la troisième année d'assurance, rendant le contrat enfin rentable, car l'assuré devient moins risqué. En effet, celui-ci avait 18 ans au début du contrat et venait tout juste d'obtenir son permis. Au bout de 3 ans, il a acquis assez d'expérience et présente moins de risque pour l'assureur. Ainsi, le déficit initial s'explique par la volonté de l'organisme d'assurance d'attirer ce jeune conducteur en lui proposant une offre très attractive afin d'assurer une bonne mutualisation de son portefeuille. Ce résultat montre que les modèles réussissent bien à capter les changements des caractéristiques des assurés au cours du temps en gardant uniquement leurs informations à l'entrée dans le portefeuille.

Dans la suite, on appellera "modèle de projection de marges" ce processus qui calcule les marges espérées dans le temps en s'appuyant sur une série de modèles successifs, où chaque modèle utilise les prédictions du modèle précédent.

3.3 Résultats sur l'échantillon test

Certes, l'entraînement à la prédiction des primes et à la projection des marges espérées a été réalisé sur l'ensemble de la base de données afin d'apprendre sur un nombre maximum d'individus mais ces prédictions ont également été réalisées, au préalable, sur des échantillons d'entrainement et de test afin de rendre compte des performances des modèles et de l'absence de sur-apprentissage. Les résultats seront concentrés uniquement sur la $2^{\text{ème}}$ année d'assurance ($t = 1$), étant la première année de projection, et la $10^{\text{ème}}$ et dernière année de projection ($t = T_h = 9$). Pour ces deux périodes, deux graphiques ont été représentés : le graphique de l'importance des variables selon le poids qu'elles apportent aux modèles et les graphiques de marges espérées moyennes observées et prédites (sur l'échantillon de test) en fonction de 4 des variables les plus importantes.

Le graphique d'importance des variables est basé sur la fréquence d'utilisation de chaque variable dans les arbres de décision. Plus précisément, il s'agit du nombre de fois qu'une variable est utilisée pour diviser les données dans tous les arbres du modèle, ce qui contribue à augmenter son poids. Ces graphiques sont la réunion des importances des variables des modèles XGBoost pour les deux types de primes et les valeurs correspondent à une pondération des importances de chaque modèle. Ici, seulement les 10 variables les plus importantes seront affichées.

Les graphiques de marges espérées moyennes observées et prédites en fonction des variables permettent de voir si les modèles ont bien prédit les marges espérées (en moyenne). Les marges espérées moyennes observées et prédites en fonction d'une variable y sont présentées, ainsi que l'exposition de cette variable selon ses modalités. Les variables continues (ayant trop de modalités différentes) ont été regroupées en 8 segments distincts via une méthode de quantiles. Les valeurs de l'axe des abscisses ont dû être supprimées afin d'anonymiser le portefeuille.

$2^{\text{ème}}$ année d'assurance ($t = 1$)

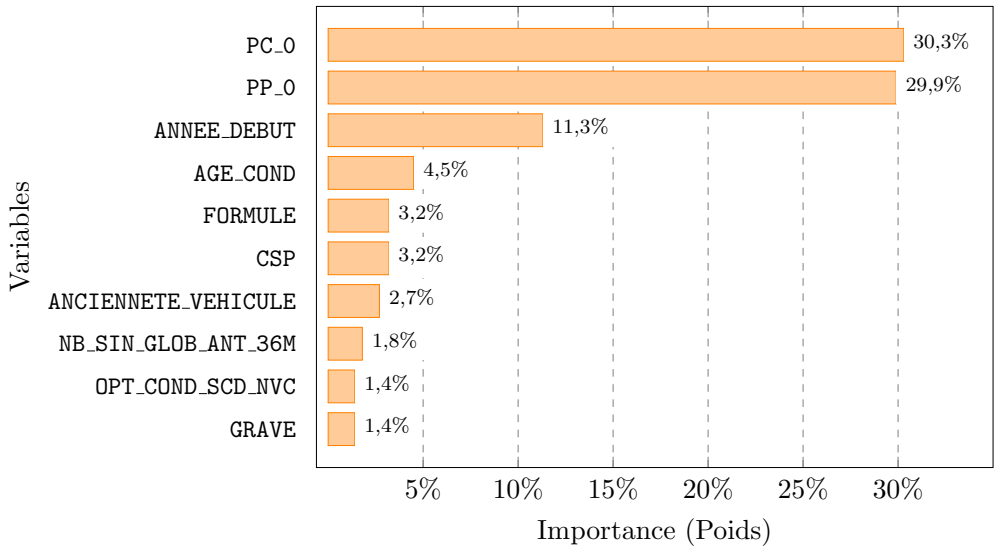
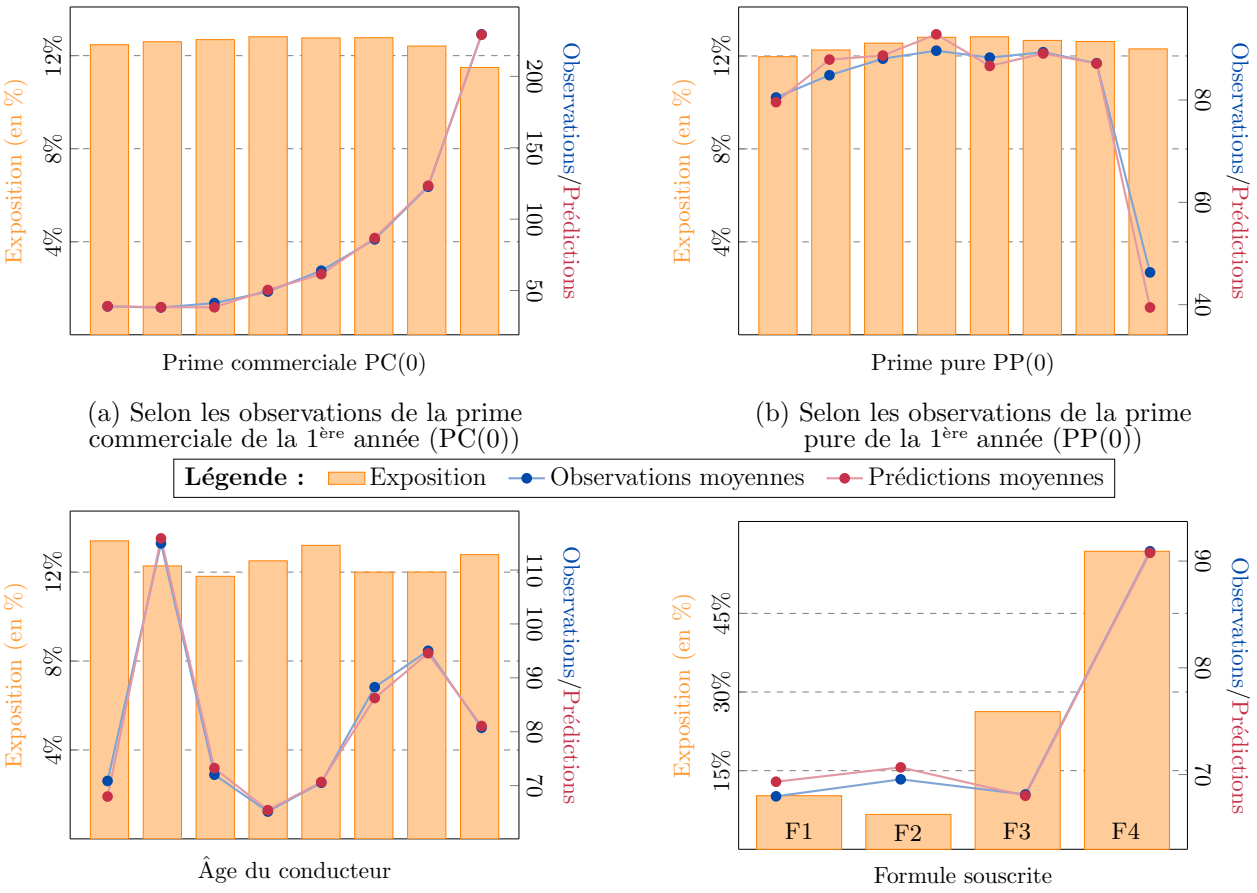


FIGURE 3.7 : Importance des variables de la $2^{\text{ème}}$ année d'assurance ($t = 1$)

Comme le montre la figure 3.7, les deux variables les plus importantes (plus de 60% de poids) du modèle de projection de marges sont les primes pures et commerciales des individus à leur entrée en portefeuille. Ce résultat était forcément attendu puisque les marges espérées de la deuxième année d'assurance dépendent fortement du tarif payé et du risque théorique de l'année précédente. Parmi les

autres variables, figurent également l'année d'entrée en portefeuille, qui capte la tendance inflationniste, et surtout des variables qui témoignent d'une stratégie commerciale de l'organisme d'assurance comme l'âge de l'assuré, sa catégorie socio-professionnelle ou encore la formule souscrite.

Les marges espérées de la 2ème année d'assurance semblent avoir été très bien prédites comme l'illustre la figure 3.8, les prédictions moyennes sont très proches des observations moyennes.



(a) Selon les observations de la prime commerciale de la 1ère année (PC(0))

(b) Selon les observations de la prime pure de la 1ère année (PP(0))

(c) Selon l'âge du conducteur

(d) Selon la formule souscrite

FIGURE 3.8 : Observations et prédictions moyennes des marges espérées de la 2ème année d'assurance ($t = 1$) selon 4 des variables les plus importantes

Certains résultats intéressants témoignant de la stratégie de l'assureur sont visibles sur ces graphiques. Par exemple, la figure 3.8(c) montre que l'assureur réalise très peu de marges sur les jeunes conducteurs (première barre), ce qui indique une volonté d'en attirer le plus possible dans son portefeuille. Pourtant, une marge espérée moyenne très élevée est observée pour le deuxième segment d'âge et qui peut s'expliquer par un effet de sélection naturelle combiné à une revalorisation tarifaire après la période jeune conducteur. En effet, le premier segment correspond aux jeunes conducteurs, considérés comme plus risqués mais attirés dans le portefeuille grâce à des offres tarifaires attractives, ce qui limite la marge espérée initiale. Après quelques années, ces conducteurs subissent une hausse tarifaire significative afin d'ajuster leur prime à leur sinistralité réelle. Toutefois, seuls certains d'entre eux acceptent ces revalorisations et restent assurés. Ce phénomène introduit un biais de sélection : la sinistralité effective de ce groupe devient inférieure aux attentes initiales, alors que les primes commerciales restent élevées, ce qui gonfle mécaniquement la marge espérée. De même, la figure 3.8(a), représentant les marges espérées moyennes selon la prime commerciale payée à l'entrée dans le portefeuille montre que plus un individu paie cher son assurance, plus l'assureur margera sur son contrat

en moyenne. Ce résultat est fortement corrélé avec la formule souscrite, visible dans la figure 3.8(d), puisqu'une couverture plus ou moins étendue influence directement le montant de la prime commerciale, et donc la marge espérée. Enfin, il est également important de noter que, d'après la figure 3.8(b), plus la prime pure est élevée (et donc plus l'individu est risqué), moins l'assureur dégagera de marges de ces contrats. Encore une fois, cela peut expliquer la volonté de l'assureur de ne pas trop pénaliser les individus risqués (comme des jeunes conducteurs) en restant compétitifs sur leur segments.

Par ailleurs, les valeurs des métriques entre les échantillons d'entraînement et de test sont présentes dans le tableau 3.6 pour cette première année de projection. Leur analyse détaillée révèle que les résultats sont très bons mais cela est en grande partie dû à la forte performance du XGBoost et non au sur-apprentissage.

Echantillon	RMSE	MAE	R^2	Gini normalisé
Entraînement	68,77	34,66	86,3%	95,2%
Test	69,17	34,75	86,5%	95,3%

Table 3.6 : Métriques d'évaluation selon l'échantillon pour la projection des marges espérées de la $2^{\text{ème}}$ année d'assurance ($t = 2$)

Enfin, la figure 3.9 représentant un échantillon des marges espérées prédites en fonction des marges espérées observées sur l'échantillon de test confirme bien ces bons résultats, malgré quelques *outliers*, puisque la tendance linéaire suit bien une droite d'équation $y = x$.

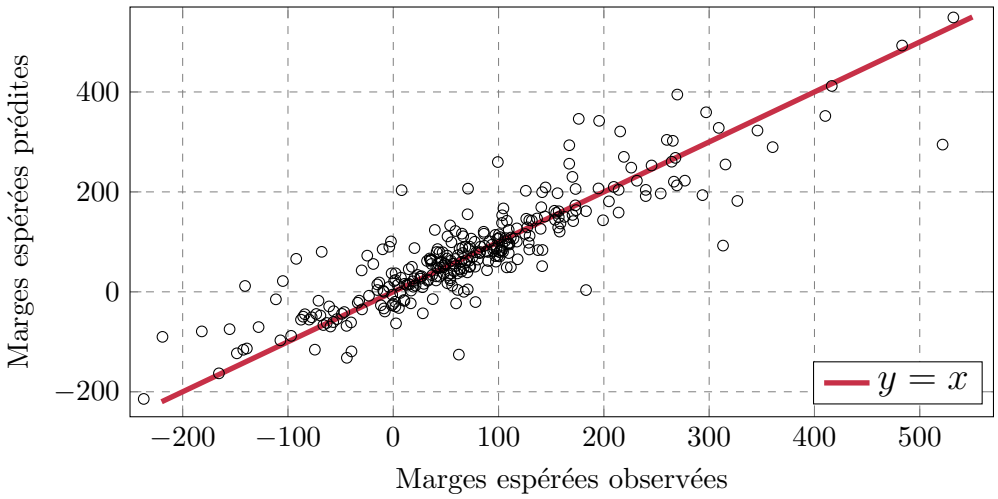


Figure 3.9 : Marges espérées prédites en fonction des marges espérées observées de la $2^{\text{ème}}$ année d'assurance ($t = 1$)

$10^{\text{ème}}$ année d'assurance (t = 9)

La figure 3.10 donne des résultats similaires à ceux de la $2^{\text{ème}}$ année d'assurance. En effet, les prédictions des primes pures et commerciales de la $9^{\text{ème}}$ année d'assurance (*i.e.* $\widehat{\text{PP}}(8)$ et $\widehat{\text{PC}}(8)$) apparaissent dans les variables les plus importantes à plus de 40% pour projeter les marges espérées de la $10^{\text{ème}}$ année d'assurance ($t = 9$). L'année d'entrée dans le portefeuille explique toujours la tendance inflationniste. Les mêmes variables reflétant la stratégie commerciale de l'assureur, telles que l'âge du conducteur, la catégorie socio-professionnelle, la formule et le coefficient Bonus-Malus, sont également présentes.

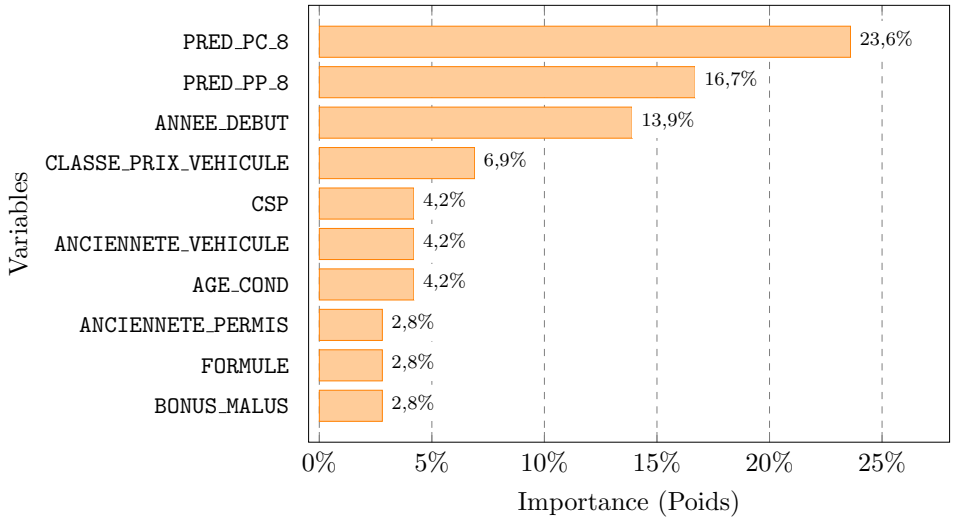
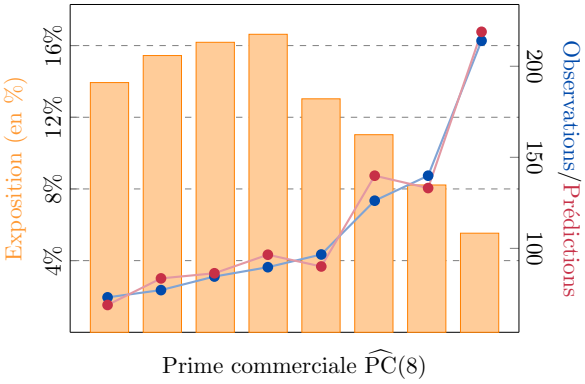
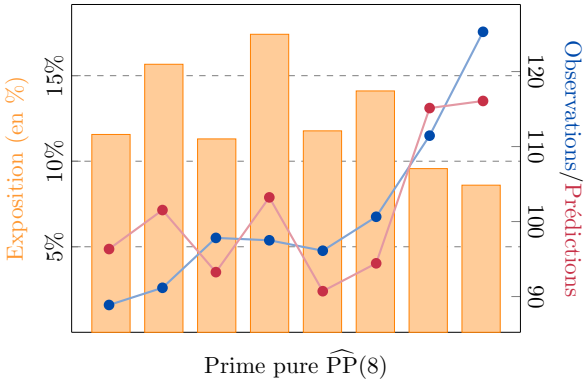


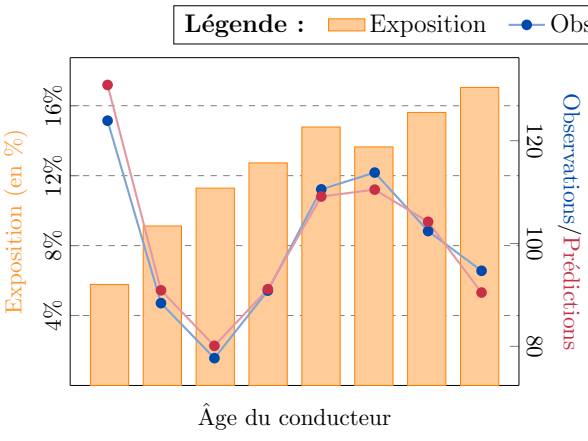
FIGURE 3.10 : Importance des variables de la 10ème année ($t = 9$)



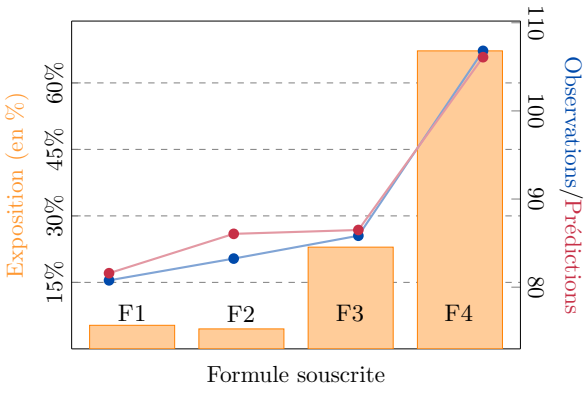
(a) Selon les prédictions de la prime commerciale de la 9ème année $\left(\widehat{\text{PC}}(8)\right)$



(b) Selon les prédictions de la prime pure de la 9ème année $\left(\widehat{\text{PP}}(8)\right)$



(c) Selon l'âge du conducteur



(d) Selon la formule souscrite

FIGURE 3.11 : Observations et prédictions moyennes des marges espérées de la 10ème année d'assurance ($t = T_h = 9$) selon 4 des variables les plus importantes

Par ailleurs, la figure 3.11 présente des prédictions qui en moyenne, correspondent moins bien aux observations (en particulier les marges espérées moyennes selon la prime pure de la 9ème année d'assurance présente dans la figure 3.11(b)). Ces résultats s'expliquent simplement par le fait que ces prédictions ont été faites à partir d'une succession de prédictions et entrainées sur un échantillon plus petit. Les projections sont donc de plus en plus volatiles dans le temps. Les figures 3.11(a) et 3.11(d) montrent des interprétations sur les marges espérées assez similaires à celles de la 2ème année d'assurance. Cependant, il est très intéressant de noter que dans la figure 3.11(c), représentant les observations et les prédictions moyennes des marges espérées selon l'âge du conducteur à l'entrée dans le portefeuille, la proportion de jeunes conducteurs a fortement diminué. Cela est lié au fait que certains de ces assurés ont quitté le portefeuille pendant les 9 dernières années. Les assurés restants ne sont plus jeunes conducteurs et sont donc moins risqués. Leur primes pures ont alors diminué au cours du temps, creusant un écart important avec leur primes commerciales, ce qui implique de fortes marges pour l'assureur avec une observation moyenne d'environ 125€ et une prédiction moyenne d'environ 130€ pour ces profils. Concernant les primes pures, la figure 3.11(b) donne des résultats inverses à ceux de la 2ème année d'assurance puisque cette fois, l'assureur réalise une marge espérée plus importante sur les profils ayant un risque théorique important.

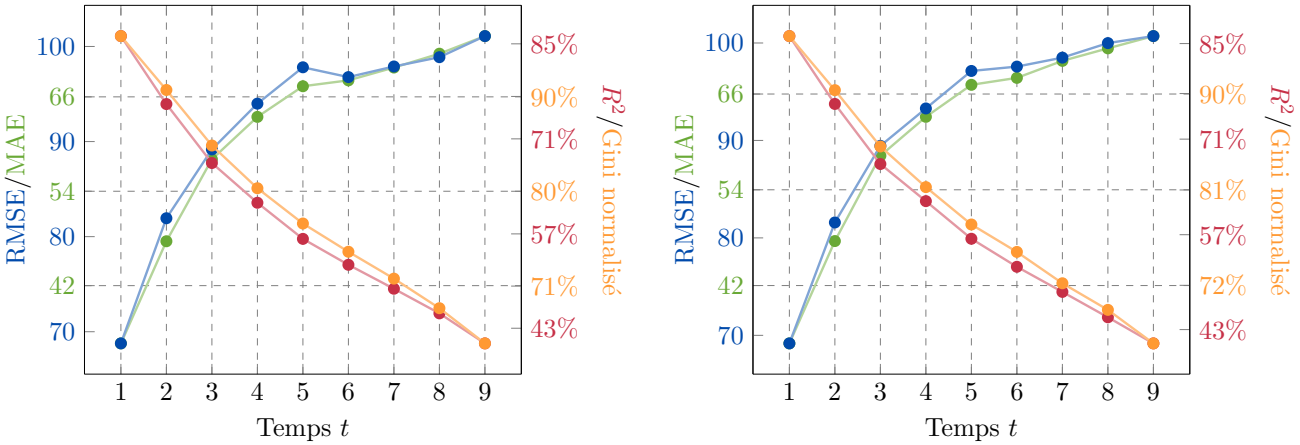
Malgré cette perte de précision, les valeurs des métriques pour cette dernière année de projection, répertoriées dans le tableau 3.7, sont tout de même très bonnes et témoignent d'une bonne performance :

Echantillon	**RMSE**	**MAE**	$\mathbf{R^2}$	**Gini normalisé**
Entraînement	101,09	73,73	40,7%	66,1%
Test	100,71	73,26	40,9%	66,2%

Table 3.7 : Métriques d'évaluation selon l'échantillon pour la projection des marges espérées de la 10ème année d'assurance ($t = T_h = 9$)

Métriques

Concernant les métriques d'évaluation, les performances sont très bonnes bien que les prédictions soient de moins en moins précises au fil du temps. Comme dit précédemment pour expliquer les résultats du modèle pour la 10ème année d'assurance, cette perte de précision est logique puisque les modèles ont été entrainés sur des échantillons de plus en plus petits et avec comme variables explicatives les prédictions des primes précédentes et non les primes réellement observées. La figure 3.12 représentent les évolutions des valeurs des métriques au cours du temps pour les échantillons d'entraînement et de test. Nous obtenons des valeurs de métriques cohérentes, *i.e.* évoluant dans le sens logique, avec la RMSE et la MAE qui augmentent avec le temps à l'inverse du coefficient de détermination R^2 et du Gini normalisé qui diminuent au fur et à mesure que le temps passe. De plus ces valeurs sont très proches entre les échantillons d'entraînement et de test, ce qui tend à confirmer l'absence de sur-apprentissage.



(a) Echantillon d'entraînement

(b) Echantillon de test

FIGURE 3.12 : Valeurs des métriques d'évaluation sur les échantillons d'entraînement et de test au cours du temps

3.4 Synthèse et premier calcul de valeur client future

Maintenant que l'apprentissage pour projeter les marges espérées a été réalisé et validé sur l'ensemble du portefeuille, il est possible d'appliquer la projection des marges espérées individuelles sur toute la période d'étude, *i.e.* $T_h = 9$, pour le portefeuille d'affaires nouvelles où l'on dispose uniquement des caractéristiques initiales ainsi que les primes pures et commerciales lors de la souscription de ces contrats. La figure 3.13 représente par exemple l'évolution des marges espérées au cours du temps selon les différentes catégories socio-professionnelles :

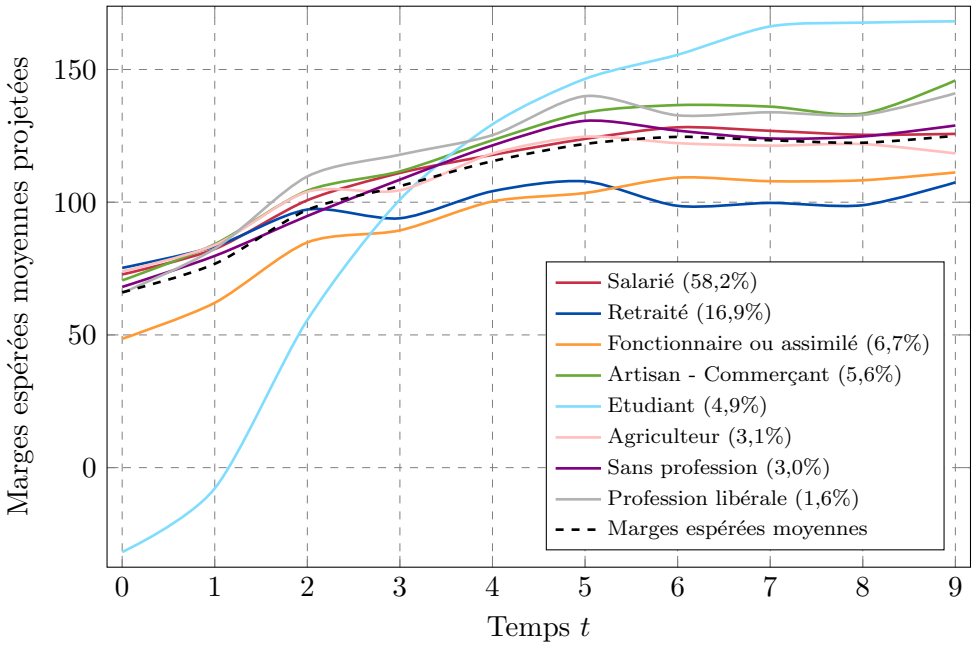


FIGURE 3.13 : Evolution des marges espérées moyennes projetées au cours du temps selon la catégorie socio-professionelle

Les observations faites précédemment concernant les jeunes conducteurs, et donc les étudiants,

sont bien vérifiées. En effet, ceux-ci possèdent en moyenne des marges espérées négatives pendant leurs deux premières années d'assurance, ce qui témoigne d'une volonté de l'assureur de leur proposer un tarif très attractif à la souscription. Au bout de 10 ans, ceux-ci sont les profils les plus rentables pour l'assureur qui aura alors augmenter l'écart entre la prime commerciale et la prime technique (prime pure chargée) progressivement au cours du temps jusqu'à atteindre un équilibre à partir de $t = 7$. Voici comment s'explique cette marge espérée moyenne croissante dans le temps. Lors de leur souscription, les primes pures et commerciales des étudiants sont très proches et assez élevées en raison du risque important qu'ils portent. L'assureur ne peut que très peu marger, voire pas du tout puisque les jeunes sont très sensibles au prix. Cependant, au cours du temps, le risque porté par un jeune conducteur, et donc sa prime pure, baisse car celui-ci a acquis de l'expérience (sauf en cas de sinistres). La prime commerciale quant à elle n'évolue pas énormément en raison des revalorisations annuelles. Ce mécanisme explique donc la hausse des marges espérées réalisées dans le temps chez les étudiants. Par ailleurs, cela peut expliquer la faible durée de vie des contrats de ces individus qui sont incités à changer d'assureur au bout de quelques années puisque leur tarif est bien plus élevé que s'ils souscrivaient chez un concurrent en affaire nouvelle.

Les autres catégories socio-professionnelles ont augmenté progressivement leurs marges espérées au cours du temps, malgré quelques fluctuations, mais de manière bien moins importante que les étudiants. Les retraités sont au final les individus sur lesquels l'assureur réalise le moins de marges espérées. Cependant, cela ne veut pas dire qu'ils seront mauvais en termes de valeur client future. En effet, en combinant ces marges espérées projetées à la survie estimée par le modèle à risque proportionnels de Cox du chapitre prédédent, la valeur client future peut être calculée pour tous les individus du portefeuille d'affaires nouvelles. La formule utilisée spécifiquement dans notre étude pour calculer la valeur client future d'un individu est rappelé dans l'équation (3.8) ci-dessous

$$\text{Valeur client future de l'assuré } i = \sum_{t=0}^{T_h} \frac{\text{Marge}_i(t) \times \mathbb{P}_i(t)}{(1+r)^t} \tag{3.8}$$

Ainsi, l'évolution (cumulative) de la valeur client future moyenne au cours du temps selon les catégories socio-professionnelles est représentée dans la figure 3.14. Il est clair que la probabilité de survie des contrats jouent un rôle important dans le calcul de la valeur client future puisqu'à l'inverse des marges espérées projetées, les retraités sont les meilleurs profils selon cet indicateur et les étudiants sont les plus mauvais. Ces derniers vont certes rapporter le plus après 10 ans dans le portefeuille mais uniquement s'il reste aussi longtemps en contrat.

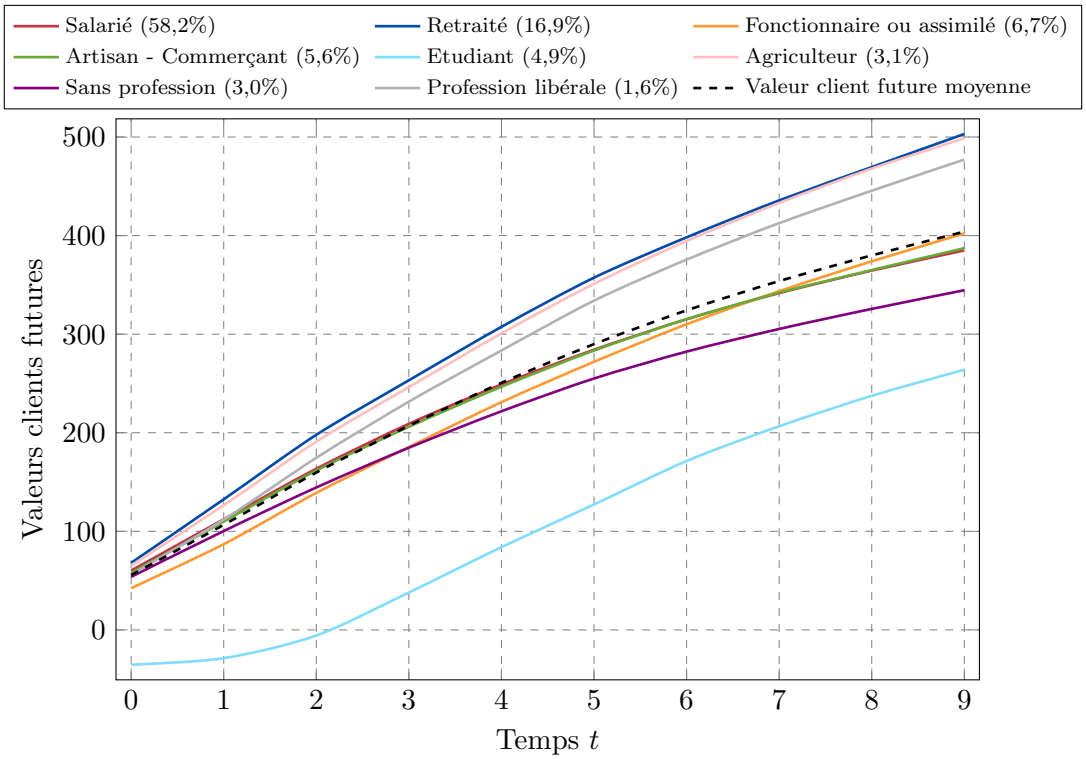


FIGURE 3.14 : Evolution de la valeur client future moyenne au cours du temps selon la catégorie socio-professionnelle

Enfin, la valeur client future retenue correspond à la valeur à l'horizon d'étude choisi $t = T_h = 9$. Le choix de cet horizon est donc très important pour évaluer la fidélité et la rentabilité des assurés puisqu'un horizon plus ou moins long donne des résultats bien différents. Par exemple, si l'étude s'arrêtait à $t = 3$, *i.e.* après 4 ans d'assurance, alors les fonctionnaires ou assimilés serait l'un des plus mauvais profils en termes de valeur client future. Cependant, ceux-ci vont accumuler de plus en plus de valeur au cours du temps jusqu'à devenir, au bout de 10 ans, la $4^{\text{ème}}$ meilleure catégorie socio-profesionnelle en moyenne, derrière les retraités, les agriculteurs et les personnes pratiquant une profession libérale, et ont une valeur client future moyenne très proche de celle du portefeuille entier.

La valeur client future a donc été calculée à partir du montant de cotisation proposé par l'assureur à chacun des assurés lors de leur souscription. Il convient maintenant de vérifier si ces assurés acceptent de payer ce montant ou non. Pour modéliser ces probabilités d'acceptation des devis, un modèle de transformation sera implémenté.

Chapitre 4

Modèle de transformation

Ce chapitre est consacré à la modélisation des probabilités de transformation, notées $Tr_i(\cdot)$ dans la formule suivante de la valeur client potentielle du $g^{\text{ème}}$ cluster ‘

$$VC_g = \sum_{i=1}^{n_g} \sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Marge}_i(t)}{(1+r)^t} \times \boxed{Tr_i(v_g)}$$

Lors de la création d'un contrat, l'assureur propose un devis à son client. Ce dernier a alors la possibilité d'accepter ce devis ou non si le montant lui convient. Afin d'estimer les probabilités que les contrats soient acceptés par les assurés, un modèle de transformation est implémenté via un modèle linéaire généralisé bien précis : la régression logistique.

4.1 Théorie des modèles linéaires généralisés

Les modèles linéaires généralisés (*Generalized Linear Models*) ont été introduits par Nelder et Wedderburn (1972) et complétés par Nelder et McCullagh (1989). Ces modèles sont communément utilisés en assurance non-vie et permettent de modéliser des variables de fréquence, de sévérité, ou encore des probabilités dans notre cas. Dans un premier temps, le cadre du modèle linéaire gaussien est posé, puis les modèles linéaires généralisés seront expliqués pour enfin présenter la régression logistique.

4.1.1 Modèle linéaire gaussien

Le modèle linéaire gaussien, ou modèle de régression linéaire classique, est un outil statistique fondamental utilisé pour examiner la relation entre une variable réponse (ou dépendante) et une ou plusieurs variables explicatives. Ce modèle repose sur l'hypothèse que les résidus suivent une distribution normale.

La formule générale du modèle linéaire gaussien peut être exprimée sous la forme matricielle suivante

$$Y = X\beta + \varepsilon \tag{4.1}$$

- $Y \in \mathbb{R}^n$: vecteur des n observations de la variable réponse
- $X \in \mathcal{M}_{n,p+1}(\mathbb{R})$: matrice des n observations des p variables explicatives, appelée matrice *design*, dont la première colonne (*intercept*) est remplie de 1 et est de rang plein ($\text{rg}(X) = p + 1 < n$).

- ▷ $\beta \in \mathbb{R}^{p+1}$: vecteur des coefficients à estimer
- ▷ $\varepsilon \in \mathbb{R}^n$: vecteur des résidus qui sont supposés centrés, non corrélés, homoscédastiques* et gaussiens, *i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ où I_n est la matrice identité de taille n et l'écart-type $\sigma > 0$ est constant.

Sous cette forme, les coefficients $\beta = (\beta_0, ..., \beta_p)^\top$ sont généralement estimés par la méthode des moindres carrés ordinaires (*Ordinary Least Squares*) qui consiste à minimiser le terme $\|Y - X\beta\|^2$ (où $\|\cdot\|$ est la norme euclidienne). . Il peut être démontré† très facilement que le vecteur de coefficients ainsi estimé $\widehat{\beta}^{\text{MC}}$ dispose d'une expression explicite : $\widehat{\beta}^{\text{MC}} = \left(X^\top X\right)^{-1} X^\top Y$.

Le modèle linéaire gaussien est un outil puissant et largement utilisé en statistique pour sa simplicité et sa capacité à expliquer la relation linéaire entre les variables. Cependant, dans le cadre de modélisation de la probabilité d'acceptation d'un devis, il doit être généralisé pour gérer les distributions non normales comme la distribution de Bernoulli. Cette dernière est plus appropriée pour modéliser des variables binaires telles que la conversion ou non d'un devis, permettant ainsi de mieux appréhender la nature du phénomène étudié.

4.1.2 Modèle linéaire généralisé

Les modèles linéaires généralisés sont une extension des modèles de régression linéaire classiques qui permettent de modéliser une variable réponse en fonction d'une ou plusieurs variables explicatives lorsque les hypothèses de normalité des résidus et de linéarité ne sont pas respectées. Ils sont particulièrement utiles lorsque la variable réponse suit une distribution différente de la loi normale. Ces modèles reposent sur 3 composantes : la distribution de la variable réponse (composante aléatoire), le prédicteur linéaire (composante déterministe) et la fonction de lien.

La variable réponse Y doit appartenir à la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$. Une variable aléatoire Y possède une densité de probabilité, par rapport à une mesure dominante ν, notée $f_{\theta,\phi}$ appartenant à la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$ si $f_{\theta,\phi}$ s'écrit pour tout $y \in \mathbb{R}$

$$f_{\theta,\phi}(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{4.2}$$

où $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont des fonctions connues et dérivables telles que :

- ▷ $b(\cdot)$ est 3 fois dérivable
- ▷ $b'(\cdot)$ est inversible, *i.e.* $(b')^{-1}(\cdot)$ existe.
- ▷ $\theta \in \Theta \subseteq \mathbb{R}$ est le paramètre naturel et $\phi \in \mathcal{B} \subseteq \mathbb{R}_+^*$ est le paramètre de dispersion.

Si Y admet une densité de probabilité appartenant à la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$ alors

$$\mathbb{E}_\theta[Y] = b'(\theta) \text{ et } \mathbb{V}\text{ar}_\theta(Y) = a(\phi)b''(\theta)$$

Le prédicteur linéaire correspond au même $X\beta$ que dans le cadre gaussien, où X est la matrice *design* et β le vecteur de coefficients à estimer.

Enfin, la fonction de lien $g(\cdot)$ est une fonction monotone, différentiable et inversible qui établit la relation suivante entre le prédicteur linéaire et la variable réponse Y

$$g\left(\mathbb{E}[Y|X]\right) = X\beta \iff \mathbb{E}[Y|X] = g^{-1}(X\beta) \tag{4.3}$$

*Les résidus $(\varepsilon_1, ..., \varepsilon_n)$ ont une variance (niveau de bruit) constante $\sigma^2 > 0$.

† $\min\limits_{\beta\in\mathbb{R}^{p+1}} \|Y - X\beta\|^2 \iff \min\limits_{\beta\in\mathbb{R}^{p+1}} \left[Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta\right] \rightsquigarrow \Delta_\beta\left[\|Y - X\beta\|^2\right] = 0_{\mathbb{R}^{p+1}}$

$\iff -2X^\top Y + 2X^\top X\beta = 0_{\mathbb{R}^{p+1}} \iff \widehat{\beta}^{\text{MC}=\left(X^\top X\right)^{-1}X^\top Y}$ avec $X^\top X$ inversible car X est de rang plein.

En théorie, le lien canonique associé à la distribution de la variable réponse est choisi en raison de sa simplicité mathématique, de ses propriétés théoriques avantageuses et de son interprétabilité.
Celui-ci est défini comme suit : soit Y une variable aléatoire qui admet une densité appartenant à la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$, telle que $\mathbb{E}_\theta[Y] = b'(\theta) = \mu$ alors la fonction $g(\mu) = (b')^{-1}(\mu)$ est appelée fonction de lien canonique.

Voici un tableau récapitulant les liens canoniques selon certaines distributions usuelles de la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$:

Distribution	**Support**	**Nom du lien**	**Fonction de lien**
Normale	$\mathbb{R}$	Identité	$g(\mu) = \mu$
Exponentielle	$\mathbb{R}_+^*$	Réciproque	$g(\mu) = -\dfrac{1}{\mu}$
Gamma			
Poisson	$\mathbb{N}$	Log	$\ln(\mu)$
Bernoulli	$\{0,1\}$	Logit	$\ln\left(\dfrac{\mu}{1-\mu}\right)$
Binomiale	$[\![0, N]\!]$		$\ln\left(\dfrac{\mu}{N-\mu}\right)$

Table 4.1 : Tableau des lois usuelles de la famille exponentielle naturelle $\mathcal{F}_\theta^{\text{Nat}}$ et leur lien canonique associé

L'estimation des coefficients $\beta = (\beta_0, ..., \beta_p)^\top$ et du paramètre ϕ s'effectue par la méthode du maximum de vraisemblance. Soit $Y = (Y_1, ..., Y_n)$ où les Y_i sont indépendants et $y = (y_1, ..., y_n)$ les observations de Y, la vraisemblance s'écrit

$$\mathcal{L}(y|\beta, \phi) = \prod_{i=1}^{n} f_{\theta,\phi}(y_i) \tag{4.4}$$

D'où la log-vraisemblance

$$\ell(y|\beta, \phi) = \sum_{i=1}^{n} \ln\left(f_{\theta,\phi}(y_i)\right) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi) \tag{4.5}$$

Les estimateurs $\widehat{\beta}^{\text{EMV}}$ et $\widehat{\phi}^{\text{EMV}}$ sont ainsi obtenus par résolution du système d'équations différentielles suivants

$$\begin{cases} \dfrac{\partial}{\partial \beta_j}\ell(y|\beta, \phi) = 0 \text{ pour tout } j \in [\![0, p]\!] \\ \dfrac{\partial}{\partial \phi}\ell(y|\beta, \phi) = 0 \end{cases} \tag{4.6}$$

En théorie, les équations de log-vraisemblance pour estimer les β_j pour $j \in [\![0, p]\!]$ sont égales à

$$\frac{\partial}{\partial \beta_j}\ell(y|\beta, \phi) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}\text{ar}(Y_i)} h'(x_i^\top \beta) x_{i,j} = 0 \text{ où } h(\cdot) := g^{-1}(\cdot)$$

Elles sont donc simplifiées lorsque le lien canonique est utilisé, pour tout $j \in [\![0, p]\!]$

$$\frac{\partial}{\partial \beta_0}\ell(y|\beta, \phi) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} x_{i,j} = 0$$

En général les équations de log-vraisemblance sont transcendantes, c'est-à-dire qu'il n'est en général pas possible de donner une expression analytique des estimateurs du maximum de vraisemblance $\widehat{\beta}^{\text{EMV}}$ et $\widehat{\phi}^{\text{EMV}}$ en les résolvant. Il est alors possible de les déterminer par des procédures d'optimisation comme l'algorithme de Newton-Raphson (cf. Algorithme 1) ou l'algorithme IRLS (*Iterative Reweighted Least Squares*) décrit par GREEN (1984) :

Algorithme 3 : Algorithme IRLS

Input :

- $\diamond$ $\dfrac{\partial \mathcal{L}}{\partial \beta} : \mathbb{R}^n \to \mathbb{R}^n$ le gradient de $\mathcal{L}$
- $\diamond$ $\dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ la matrice Hessienne de $\mathcal{L}$
- $\diamond$ $\beta_0 \in \mathbb{R}^n$ une première approximation de $\widehat{\beta}$
- $\diamond$ $\varepsilon > 0$ une précision
- $\diamond$ $k = 0$

Tant que $\left\| \dfrac{\partial \mathcal{L}}{\partial \beta}(\beta_k) \right\| > \varepsilon$, **calculer :**

- $\diamond$ $\beta_{k+1} = \beta_k - \left(\mathbb{E}\left[\dfrac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top}(\beta_k) \right] \right)^{-1} \dfrac{\partial \mathcal{L}}{\partial \beta}(\beta_k)$
- $\diamond$ $k = k + 1$

Output : $\widehat{\beta} = \beta_k$

Enfin, la significativité des variables explicatives est vérifiée grâce à un test de Wald comme dans le modèle à risques proportionnels de Cox (cf. A.2).

4.2 Régression logistique

La régression logistique a été largement développée par BERKSON (1944) - (1951) et améliorée par la suite par COX (1958). Cette méthode est largement employée pour modéliser des probabilités, en particulier lorsque le modèle doit être à la fois facilement interprétable et offrir de bonnes performances. Ainsi, modéliser la probabilité de transformation des contrats, *i.e.* la probabilité que les individus acceptent les devis proposés par l'assureur, convient parfaitement à l'utilisation d'un tel modèle.

4.2.1 Définition

La régression logistique est un modèle linéaire généralisé où $Y|x$ suit une loi de Bernoulli de paramètre $p(x)$ et la fonction de lien associée est la fonction logit définie par

$$g : x \in]0,1[\longmapsto \log\left(\frac{x}{1-x}\right) \in \mathbb{R} \tag{4.7}$$

Cette fonction est symétrique en $x = 0,5$ sur son support $]0,1[$ comme le montre la figure 4.1(a). Cette fonction est donc bien adaptée pour prédire un état binaire d'acceptation ou non du devis.

Ainsi par application de la formule (4.3) et comme $Y|X$, avec $X = (x_1, ..., x_n) \in \mathcal{M}_{n,p+1}(\mathbb{R})$ la

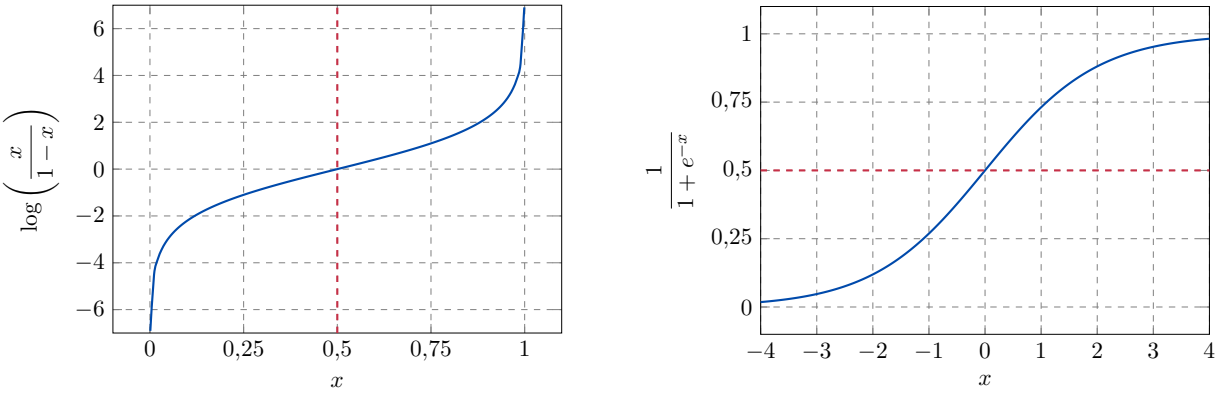
matrice *design*, suit une loi de Bernoulli* de paramètre $p(X) = \big(p_1(x_1), ..., p_n(x_n)\big)$, il en découle que

$$\left.\begin{array}{l}\mathbb{E}[Y|X] = g^{-1}(X\beta)\\ \mathbb{E}[Y|X] = p(X)\end{array}\right\} \implies g^{-1}(X\beta) = p(X) \tag{4.8}$$

où g^{-1} est la fonction sigmoïde définie par

$$g^{-1} : x \in \mathbb{R} \longmapsto \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \in]0,1[\tag{4.9}$$

Cette fonction inverse permet de transformer une valeur réelle en probabilité dans l'intervalle]0,1[comme le montre la figure 4.1(b). Ainsi en l'appliquant à la combinaison linéaire $X\beta = \beta_0 + x_1\beta_1 + ... + x_p\beta_p$, $p(X)$ représente bien un vecteur de probabilités, avec β estimé par la méthode du maximum de vraisemblance comme dans le cas général.



(a) Fonction de lien logit

(b) Fonction sigmoide

Figure 4.1 : Représentations des fonctions logit et sigmoïde

4.2.2 Interprétation

La régression logistique se distingue principalement par la facilité d'interprétation de ses coefficients. Les *odds ratios* offrent un moyen direct et simple d'interprétation pour analyser les résultats. Ils permettent de quantifier l'effet d'une variable continue ou de comparer les effets d'une variable qualitative sur la variable dépendante à partir de la formule de Bayes.

Soient deux individus A et B de caractéristiques respectives X_A et X_B. Dans le cadre de la régression logistique, l'*odds ratio* entre ces deux individus s'exprime de la manière suivante

$$\begin{aligned} OR(X_A, X_B) &= \frac{\mathbb{P}(Y=1|X=X_A)/\mathbb{P}(Y=0|X=X_A)}{\mathbb{P}(Y=1|X=X_B)/\mathbb{P}(Y=0|X=X_B)} \\ &= \frac{p(X_A)\big(1-p(X_B)\big)}{p(X_B)\big(1-p(X_A)\big)} \\ &= \frac{\dfrac{1}{1+e^{-X_A\beta}} \times \dfrac{e^{-X_B\beta}}{1+e^{-X_B\beta}}}{\dfrac{1}{1+e^{-X_B\beta}} \times \dfrac{e^{-X_A\beta}}{1+e^{-X_A\beta}}} \\ OR(X_A, X_B) &= \exp\big(\beta(X_A - X_B)\big) \end{aligned} \tag{4.10}$$

*__Rappel :__ Si $N \sim \mathcal{B}er(q)$ alors pour $k \in \{0,1\}$, $\mathbb{P}(N=k) = q^k(1-q)^{1-k}$, *i.e* $\mathbb{P}(N=k) = \left\{\begin{array}{l} q \text{ si } k=1 \\ 1-q \text{ si } k=0\end{array}\right.$ et $\mathbb{E}[N] = q$.

Ainsi, trois cas sont possible :

- ▷ Si $OR(X_A, X_B) = 1$, alors les probabilités de transformation sont identiques pour les individus ayant les caractéristiques X_A et X_B. Autrement dit, la présence des caractéristiques des individus A ou B n'a pas d'impact sur la probabilité que l'événement se produise.
- ▷ Si $OR(X_A, X_B) > 1$, alors les caractéristiques de l'individu A augmentent la probabilité de transformation par rapport à celles de l'individu B. En d'autres termes, A présente plus de chances d'accepter le devis que B, ce qui indique que les caractéristiques X_A exercent une influence plus importante sur la probabilité de transformation.
- ▷ Si $OR(X_A, X_B) < 1$, alors la situation est inversée par rapport au point précédent : les caractéristiques de l'individu A réduisent la probabilité de transformation par rapport à celles de l'individu B. Autrement dit, les caractéristiques X_A sont associées à un risque d'accepter le devis plus faible que les caractéristiques X_B.

Dans le cadre de la régression logistique, cette propriété des *odds ratios* permet une interprétation directe des coefficients β_j, pour $j \in [\![0, p]\!]$. Un coefficient positif signifie que l'augmentation d'une variable continue, ou une modalité spécifique d'une variable catégorielle, influence négativement la transformation, diminuant ainsi les chances de conversion des devis.

4.2.3 Métriques usuelles

Afin d'étudier la performance d'un modèle de régression logistique, plusieurs métriques sont utilisées :

- ▷ ***Accuracy* (Exactitude) :** il s'agit de la proportion de prédictions correctes sur l'ensemble des prédictions. Elle est calculée de la manière suivante :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

  - ◇ TP = *True Positives* (vrais positifs)
  - ◇ TN = *True Negatives* (vrais négatifs)
  - ◇ FP = *False Positives* (faux positifs)
  - ◇ FN = *False Negatives* (faux négatifs)

- ▷ **Précision :** il s'agit de la proportion de vraies prédictions positives parmi toutes les prédictions positives. Elle est définie par

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

- ▷ **Sensibilité :** il s'agit de la proportion de vrais positifs détectés parmi tous les positifs. Elle se calcule comme suit

$$\text{Sensibilité} = \frac{TP}{TP + FN}$$

- ▷ **F1-Score :** il s'agit de la moyenne harmonique entre la précision et la sensibilité. Elle équilibre les deux métriques précédentes

$$\text{F1-Score} = 2 \times \frac{\text{Précision} \times \text{Sensibilité}}{\text{Précision} + \text{Sensibilité}}$$

- ▷ **Matrice de confusion :** il s'agit du tableau qui présente la répartition des vraies et fausses prédictions du modèle. Elle est représentée dans le tableau 4.2 ci-dessous. Chaque cellule contient le nombre de prédictions positives ou négatives selon leur observations réelles. Cette matrice permet ainsi de détecter rapidement si le modèle classe les individus correctement.

Observations / Prédictions	**Positif (1)**	**Négatif (0)**
Positif (1)	TP	FP
Négatif (0)	FN	TN

Table 4.2 : Matrice de confusion

Cependant, ces métriques nécessitent de définir un seuil de Bayes, compris entre 0 et 1, qui détermine à partir de quelle valeur de probabilité prédite une observation est classée comme positive ou négative. Ce seuil, censé déterminer une frontière claire entre la conversion et la non conversion des devis, est difficile à établir pour notre étude et sera abordé dans la sous-section 4.3.1.

Il convient donc de définir d'autres métriques qui seront plus adaptées pour étudier la performance du modèle de transformation.

▷ **Courbe ROC (*Receiver Operating Characteristic*) et AUC (*Area Under the Curve*) :**

La courbe ROC, développée lors de la Seconde Guerre Mondiale pour détecter les objets ennemis sur les champs de bataille, est un graphe permettant de mesurer la performance d'un modèle de classification binaire à différents seuils de décision. Elle est tracée en représentant le taux de vrais positifs TPR, égal à la sensibilité, par rapport au taux de faux positifs FPR égal à 1 - Spécificité où la Spécificité est définie par

$$\text{Spécificité} = \frac{FP}{FP + TN}$$

La courbe ROC est tracée en variant le seuil de Bayes : en diminuant le seuil, plus d'observations sont prédites comme positives, ce qui augmente à la fois le TPR et le FPR. Idéalement, il faut maximiser le TPR tout en minimisant le FPR.

L'AUC est quant à elle une mesure numérique qui quantifie la performance d'un modèle. Elle représente la surface sous la courbe ROC, et sa valeur varie entre 0 et 1 :

- ◇ Si AUC = 1 alors le modèle est parfait, il prédit correctement toutes les observations.
- ◇ Si AUC = 0,5 alors le modèle ne fait pas mieux qu'une classification aléatoire.
- ◇ Si AUC < 0,5 alors le modèle est pire qu'une classification aléatoire (ce qui suggère un problème dans le modèle)

Ainsi, l'AUC donne en une seule valeur une indication globale de la performance du modèle et ce que la courbe ROC permet de visualiser. Un modèle avec une AUC élevée a une bonne capacité de prédiction, ce qui signifie qu'il est capable de bien séparer les observations positives des négatives sur une large gamme de seuils de Bayes.

Voici ci-dessous dans la figure 4.2 deux exemples de représentation d'une courbe ROC. La droite $y = x$ correspondant à un modèle complètement aléatoire est tracée en tirets rouges. Plus la courbe ROC se rapproche du point de coordonnées (0,1), meilleur est le modèle. Ainsi, dans cet exemple, la courbe orange représente un modèle plus performant que celui associé à la courbe bleu.

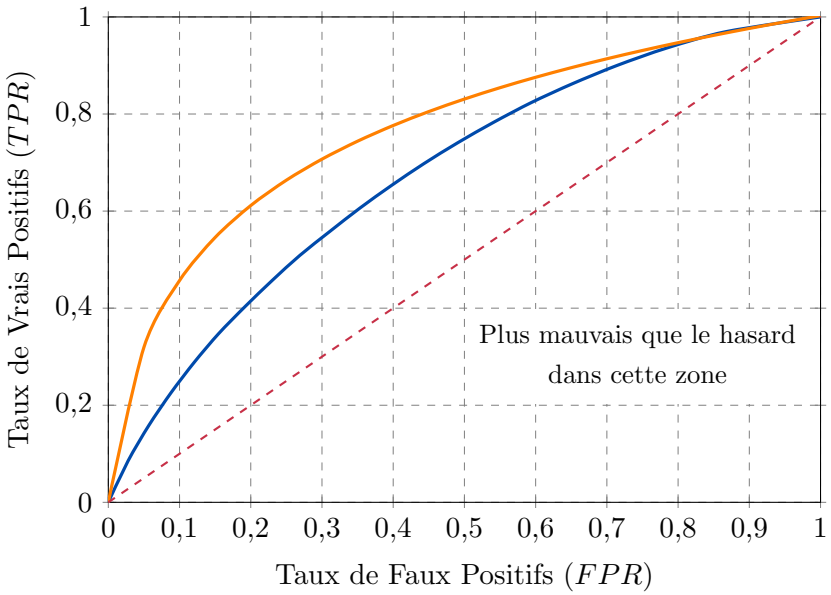


FIGURE 4.2 : Exemple de courbes ROC

▷ **Coefficient de Gini :** initialement défini dans la sous-section 3.1.3, il peut être exprimé différemment dans le cadre de problème de classification binaire comme ici. Celui-ci s'exprime en fonction de l'AUC et vaut

$$\text{Gini} = 2 \times \text{AUC} - 1$$

Ainsi, si l'AUC est de 0,5 (ce qui correspond à un modèle aléatoire), le coefficient de Gini sera de 0. À l'inverse, si l'AUC est de 1 (modèle parfait), le coefficient de Gini sera également de 1, ce qui reflète une parfaite discrimination entre les classes.

▷ ***Log-Loss* (Erreur logarithmique) :**

Le *log-loss* est une métrique de perte que le modèle doit chercher à minimiser. Il mesure la performance en termes de probabilités prédites. Plus la probabilité prédite pour la vraie classe est élevée, plus le *log-loss* sera faible. Pour y le vecteur des observations et $\hat{y}$ le vecteur des prédictions, il est défini par

$$\text{Log-loss}(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{4.11}$$

4.3 Construction du modèle et application au portefeuille

Le problème majeur rencontré lors de l'étude est l'absence d'informations sur l'acceptation des devis des individus. Ceci est normal puisque le portefeuille d'affaires nouvelles est exclusivement composé de devis convertis. Or, dans l'optique d'une optimisation de la stratégie commerciale, il sera nécessaire de modéliser la sensibilité des individus de ce portefeuille au prix lorsque celui-ci varie. La solution trouvée a donc été de construire et d'entraîner un modèle de transformation sur une autre base de données similaire du même organisme d'assurance disposant de cette information (devis converti ou non).

Ce modèle a été implémenté via une régression logistique. Les variables retenues dans le modèle sont, bien évidemment, le montant de cotisation proposé aux assurés, certaines caractéristiques des assurés

(âge, coefficient Bonus-Malus,...) et des informations sur le véhicule assuré (ancienneté, classe de prix,...).

Un *mapping* (processus consistant à établir une correspondance entre les colonnes de deux bases de données) a donc été réalisé entre cette base de devis (base d'apprentissage) et la base de l'étude (affaires nouvelles) afin de pouvoir prédire les probabilités de transformation pour chaque contrat du portefeuille d'affaires nouvelles.

4.3.1 *Strong classifier* vs. *Weak classifier*

Lors de l'implémentation du modèle, une question importante s'est posée : les prédictions doivent-elles être sous forme de probabilité (*Weak classifier*) ou sous forme binaire (*Strong classifier*)* ? En effet, l'indicateur de conversion du devis que l'on cherche à prédire est binaire (0 ou 1). Cependant, la régression logistique prédit naturellement la probabilité d'appartenance à une classe donnée, *i.e.* la probabilité qu'un client accepte un devis dans notre cas. En revanche, en appliquant un seuil de Bayes, le modèle donne une classification binaire où une probabilité supérieure (respectivement inférieure) à ce seuil est interprétée comme une conversion (respectivement non conversion) du devis proposé.

Puisque pour cette étude les résultats doivent davantage être interprétés sous forme de risque, utiliser les prédictions sous forme de probabilité semble plus adapté que de prendre une décision catégorique pour chaque individu via une classification binaire. De plus, la figure 4.3, représentant les densités de probabilités de transformation prédites selon l'observation par la régression logistique, confirme cette idée puisque la visualisation de ces densités ne permet pas de définir de seuil de Bayes efficace. En effet, la densité des devis convertis présente une allure de "double cloche". La première cloche étant trop confondue avec la densité des devis non convertis, il est difficile de décider d'un seuil à partir duquel les prédictions peuvent être binarisées. Malgré tout, le modèle parvient à distinguer, parmi les fortes probabilités, une majorité de devis convertis, et parmi les faibles probabilités, une majorité de devis non convertis. Également, les deux distributions ne sont pas identiques, ce qui peut constituer une preuve de différence entre les individus qui acceptent le devis et ceux qui ne l'acceptent pas.

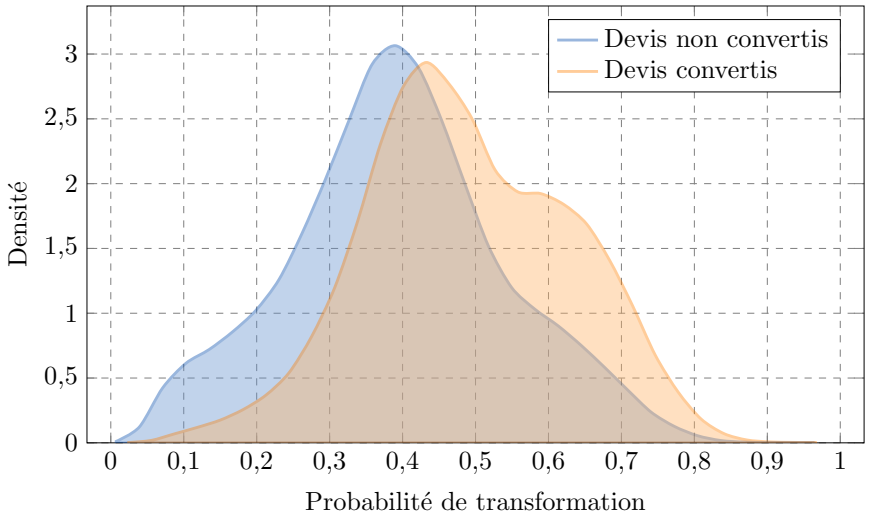


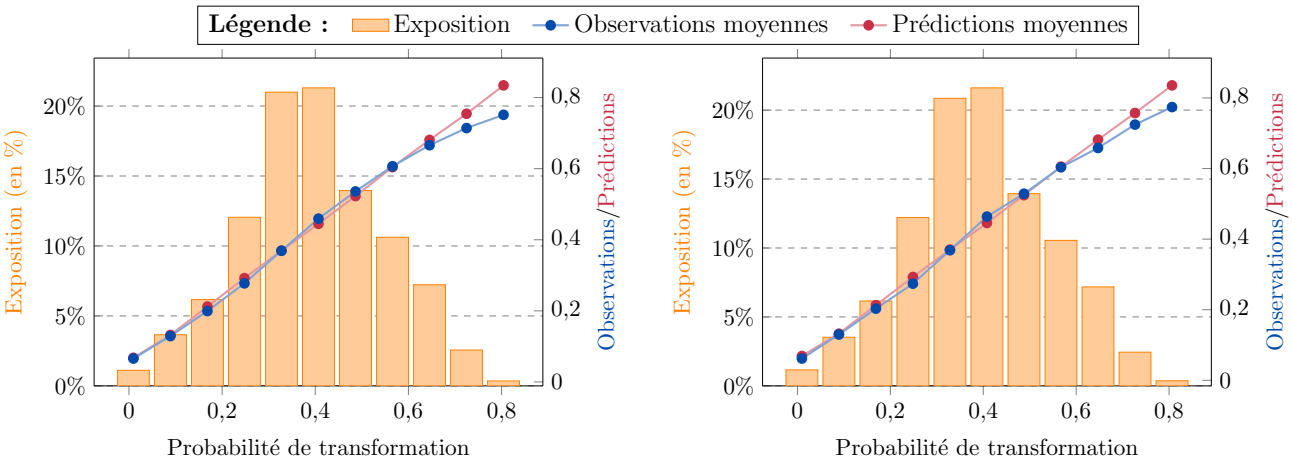
Figure 4.3 : Densités de probabilités de transformation prédites selon l'observation

Le choix d'utiliser des probabilités en sortie de modèle a donc été retenu pour la suite de l'étude.

*Ces termes sont plus généralement utilisés dans la méthode du *Boosting*, mais dans le cadre de la régression logistique, ils peuvent être considérés.

4.3.2 Résultats du modèle

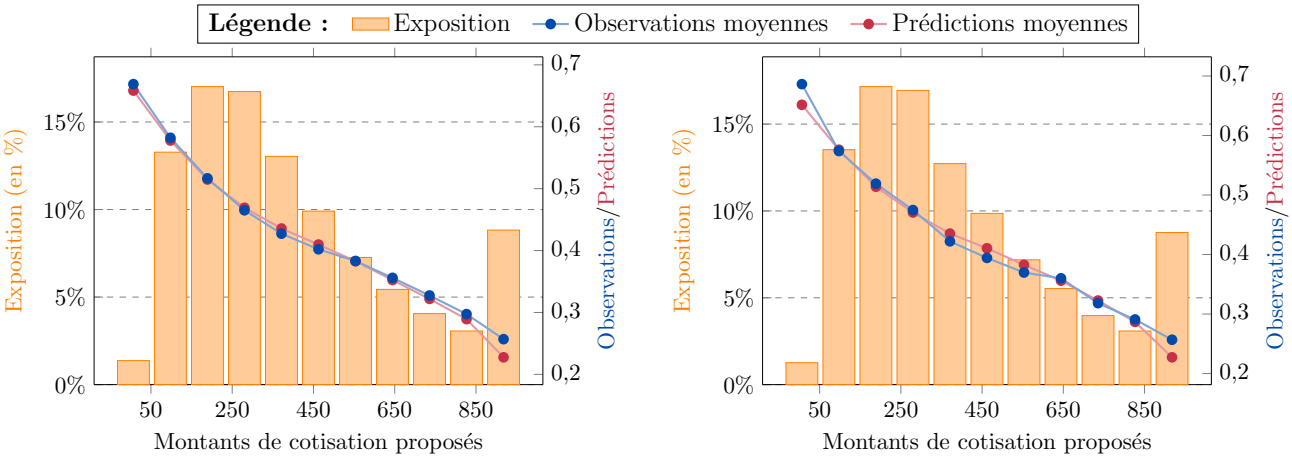
Afin de vérifier que le modèle de transformation prédit correctement la variable d'intérêt, les observations moyennes (moyennes des 0 et 1) et les probabilités de transformation moyennes prédites ont été représentées dans la figure 4.4 sur les échantillons d'entraînement (80%) et de test (20%) de la base de devis. Le modèle semble bien prédire la conversion des devis en moyenne et le modèle ne fait pas de sur-apprentissage car les résultats sont assez similaires sur les deux échantillons.



(a) Échantillon d'entrainement (b) Échantillon de test

FIGURE 4.4 : Observations et prédictions moyennes des probabilités de transformation selon l'échantillon de la base de devis

En regardant la figure 4.5 qui représente les observations et prédictions moyennes de la transformation (conversion) en fonction du montant de cotisation proposé selon les échantillons de la base de devis, le constat est le même puisque le modèle semble très bien calibré et ne fait pas de sur-apprentissage. En outre, un résultat logique émane de cette figure : la probabilité de transformation est, en moyenne, décroissante quand le montant proposé augmente.



(a) Échantillon d'entraînement (b) Échantillon de test

FIGURE 4.5 : Observations et prédictions moyennes de la transformation en fonction du montant de cotisation proposé selon l'échantillon de la base devis

Afin d'évaluer davantage la qualité du modèle, plusieurs métriques définies à la sous-section 4.2.3

ont été calculées sur les échantillons d'entraînement et de test. La matrice de confusion n'a certes pas pu être calculée puisque les prédictions n'ont pas été binarisées. Les courbes ROC pour chaque échantillon ont été représentées dans la figure 4.6, et les valeurs de l'AUC, du *log-loss* et du coefficient de Gini sont répertoriées dans le tableau 4.3. Les courbes ROC et les valeurs des métriques des deux échantillons sont très proches ce qui conforte l'absence de sur-apprentisage du modèle. Les valeurs de l'AUC et du *log-loss*, respectivement autour de 0,67 et 0,63, indiquent une capacité discriminante acceptable malgré des erreurs de prédictions. Le coefficient de Gini est quant à lui cohérent avec l'AUC.

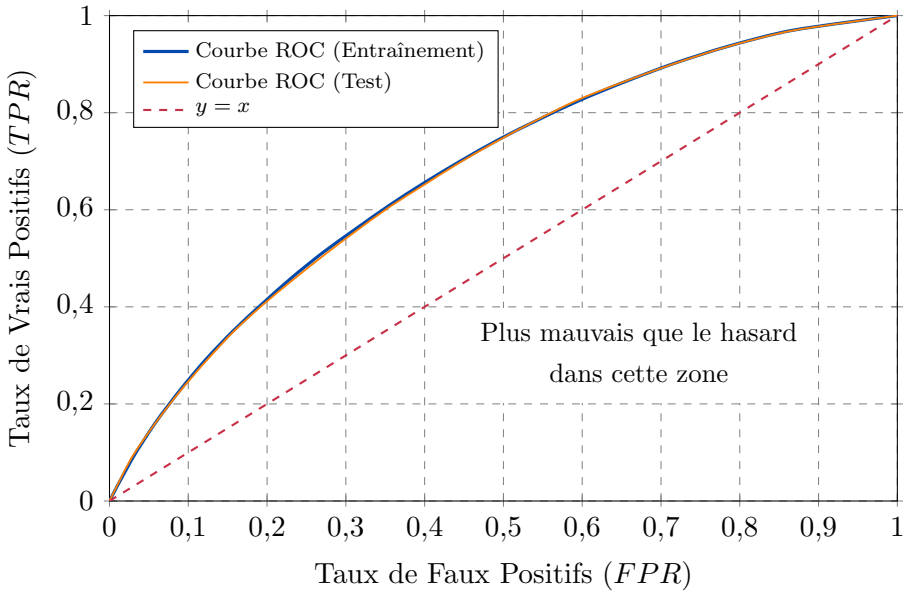
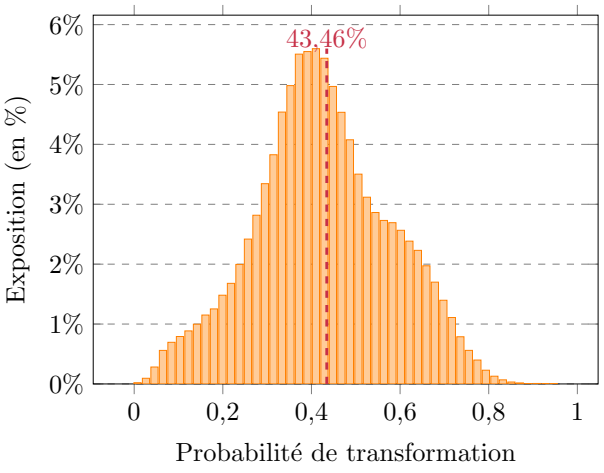


Figure 4.6 : Courbes ROC pour le modèle de transformation

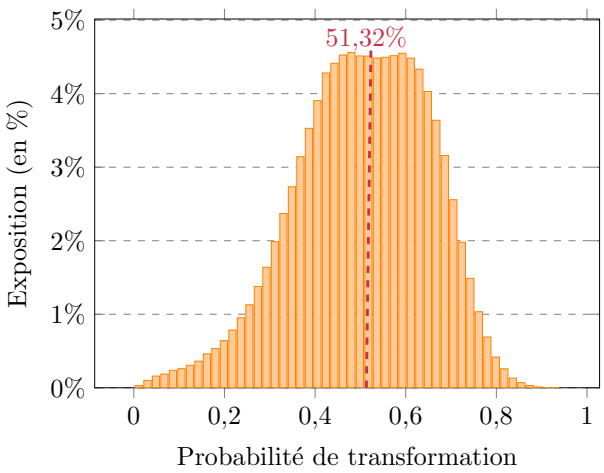
Échantillon	**AUC**	***Log-loss***	**Coefficient de Gini**
Entraînement	0,6788	0,6341	0,3577
Test	0,6774	0,6345	0,3548

Table 4.3 : Valeurs des métriques d'évaluation du modèle de transformation

Désormais que le modèle a été appris sur la base de devis, le *mapping* avec le portefeuille d'affaires nouvelles peut être fait et cela permet de prédire des probabilités de transformation pour chaque contrat de la base d'étude. Les histogrammes des probabilités de transformation prédites par le modèle sont représentés dans la figure 4.7 pour la base de devis et la base d'étude. La transformation moyenne dans la base de devis, d'environ 43%, est inférieure à celle de la base d'étude ($\simeq$ 51%). Cet écart de près de 8% s'explique par le fait que les devis du portefeuille d'affaires nouvelles ont, en effet, été convertis dans la pratique, la sensibilité au prix des individus est donc assez bien captée.



(a) Base de devis (base d'apprentissage)



(b) Base d'étude (affaires nouvelles)

FIGURE 4.7 : Histogramme des probabilités de transformation prédites selon les bases

4.3.3 Limites du modèle de transformation

Dans le cadre de cette étude, certaines simplifications ont été nécessaires en raison des données disponibles, ce qui entraîne des limites par rapport à un contexte réel.

Dans un cadre idéal, l'évaluation de la transformation devrait inclure des données présentant des prix différents aux même individus et leur réaction (acceptation ou non) face à ces devis. Or, les prix observés étant statiques (toutes choses égales par ailleurs), il n'est pas possible d'estimer une élasticité au prix. En d'autres termes, le modèle actuel ne capture pas pleinement les comportements des assurés en réponse aux fluctuations tarifaires. De plus, le modèle suppose implicitement que l'acceptation d'un devis dépend uniquement du prix et des caractéristiques initiales du client, alors qu'en réalité, les décisions d'achat sont influencées par des facteurs exogènes comme le contexte économique, la perception du rapport qualité-prix, la notoriété de la compagnie ou l'expérience précédente avec l'assureur. Enfin, le fait de ne pas disposer d'informations sur les concurrents et leurs offres limite l'aptitude du modèle à répliquer les comportements réels d'achat.

Afin de pallier ces limites, des axes d'amélioration sont possibles. En particulier, l'introduction d'une dimension dynamique permettant de mesurer l'élasticité au prix pourrait enrichir le modèle. Cela pourrait se faire par l'intégration de données historiques sur les variations tarifaires et les réponses des clients. En cas d'absence d'informations sur différents devis proposés aux clients, une solution serait d'implémenter une modèle PSM (*Propensity Score Matching*, modèle qui a pour vocation de simuler un *AB testing*) afin d'agrandir artificiellement le portefeuille avec plusieurs devis proposés par contrat. Le principe est le suivant : le portefeuille est scindé en k groupes selon leur propension à convertir les devis (par exemple les classes de transformation de la section suivante 4.4). Pour chaque contrat de chaque groupe, le modèle recherche n profils similaires (appelés jumeaux) dans les autres groupes et attribue au contrat d'origine les prix proposés à ces jumeaux. Le même contrat, avec les mêmes caractéristiques, apparaît désormais $n \times (k-1)+1$ fois dans le portefeuille avec $n \times (k-1)+1$ montants différents permettant ainsi d'explorer plusieurs scénarios tarifaires. Enfin, le modèle de transformation est appliqué à ce portefeuille élargi et donne diverses probabilités par contrat, permettant une meilleure évaluation de la sensibilité au prix des potentiels clients.

Néanmoins, un tel modèle nécessite de créer un modèle pour détecter les jumeaux entre les groupes et implique surtout une multiplication très importante de la taille de la base. Pour simplifier le processus et limiter la multiplication de la taille du portefeuille par 3, il peut être choisi de ne rechercher, pour

un contrat, qu'un seul jumeau dans les deux groupes les plus proches (en termes de transformation) de celui du contrat.

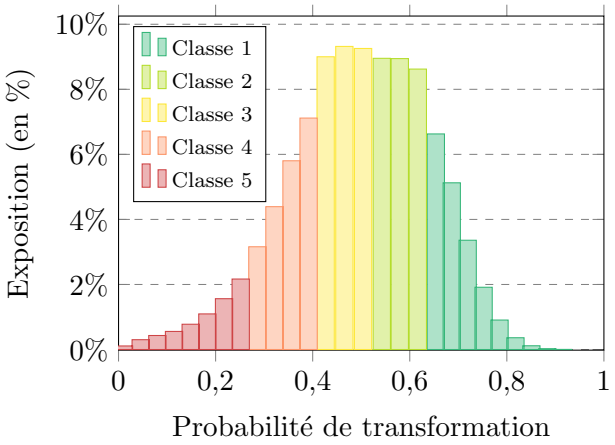
De plus, l'ajout de variables exogènes, telles que l'évolution des tarifs des concurrents (via un *benchmark*) ou des indicateurs macro-économiques, permettrait d'améliorer la précision du modèle.

Malgré ces limites, ce modèle de transformation constitue une première approche pour estimer la probabilité d'acceptation des devis et est adapté aux données disponibles pour la suite de cette étude.

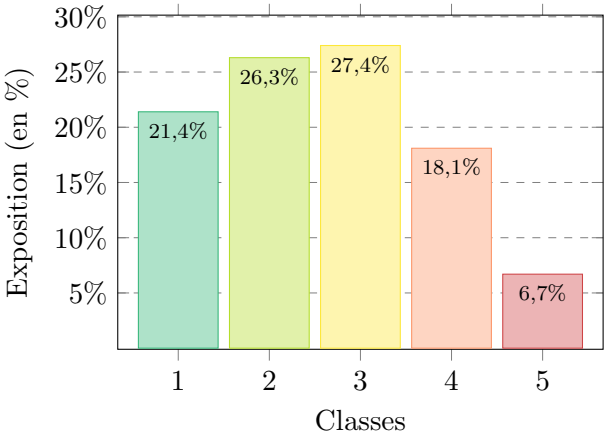
4.4 Classes de transformation

De la même manière que pour les classes de durée de vie des contrats créées à la section 2.7, différents groupes, appelés classes de transformation, ont été créés afin d'analyser plus en détails quels sont les profils ayant le plus de chances d'accepter les devis proposés par l'assureur dans le portefeuille des affaires nouvelles.

Pour cela, la même méthodologie est mise en place : 20 classifications ascendantes hiérarchiques ont été réalisées sur des sous-ensembles de 50 000 données (en raison du volume trop important du portefeuille) afin de segmenter l'histogramme de la figure 4.7(b) en différentes classes. La courbe d'inertie moyenne des C.A.H. en fonction du nombre de groupes montre que le gain d'inertie n'est plus significatif entre 5 et 6 groupes. Il a donc été décidé de créer 5 classes de transformation. Leur répartition est représentée dans la figure 4.8(b) et la figure 4.8(a) représente l'histogramme des probabilités de transformation prédites segmenté selon les classes grâce aux bornes moyennes répertoriées dans le tableau 4.4.



(a) Histogramme des probabilités de transformation prédites



(b) Histogramme des classes de transformation

Figure 4.8 : Répartition des probabilités de transformation prédites et des classes induites

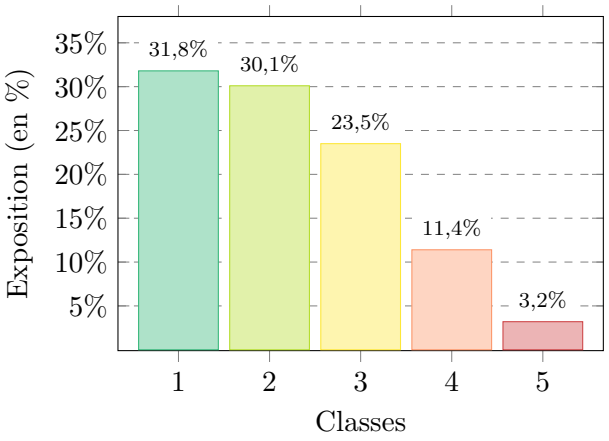
Classes	**5**	**4**	**3**	**2**	**1**
Bornes	$[0;\ 0,28[$	$[0,28\ ;\ 0,41[$	$[0,41\ ;\ 0,53[$	$[0,53\ ;\ 0,64[$	$]0,64\ ;1]$

Table 4.4 : Valeurs des bornes pour chaque classe de transformation

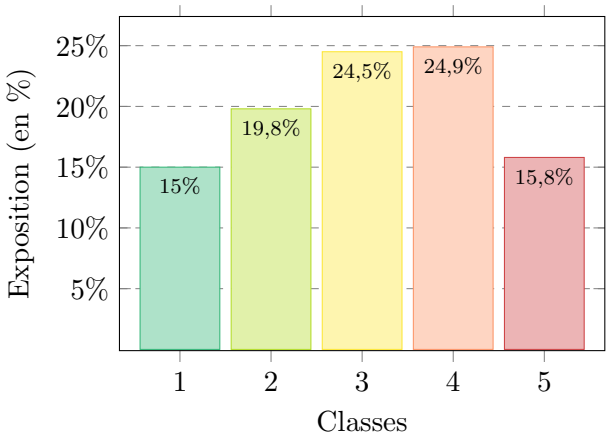
Comme pour les classes de durée de vie des contrats, la classe 1 représente la meilleure classe en terme de transformation, *i.e.* la classe des individus qui ont le plus de chance d'accepter le devis proposé, et la classe 5 la moins bonne classe.

L'impact de plusieurs variables sur la probabilité de transformation estimée par le modèle de transformation est étudié en regardant comment se répartissent les modalités de ces variables dans les classes de transformation.

Tout d'abord, pour les catégories socio-professionnelles, les répartitions entre les classes pour les agriculteurs et les personnes ayant une profession libérale sont représentées dans la figure 4.9. D'après l'histogramme 4.9(a), les agriculteurs ont une très forte probabilité d'acceptation du devis proposé par l'assureur, puisqu'ils sont énormément présents dans les meilleures classes. À l'inverse, les personnes exerçant une profession libérale sont biens répartis dans toutes les classes, donc en particulier dans la classe 5 où plus de 15% d'entre eux sont représentés, contre à peine plus de 3% chez les agriculteurs.



(a) Pour les agriculteurs (3,1%)



(b) Pour les personnes ayant une profession libérale (1,6%)

FIGURE 4.9 : Répartition de 2 catégories socio-professionnelles dans les classes de transformation

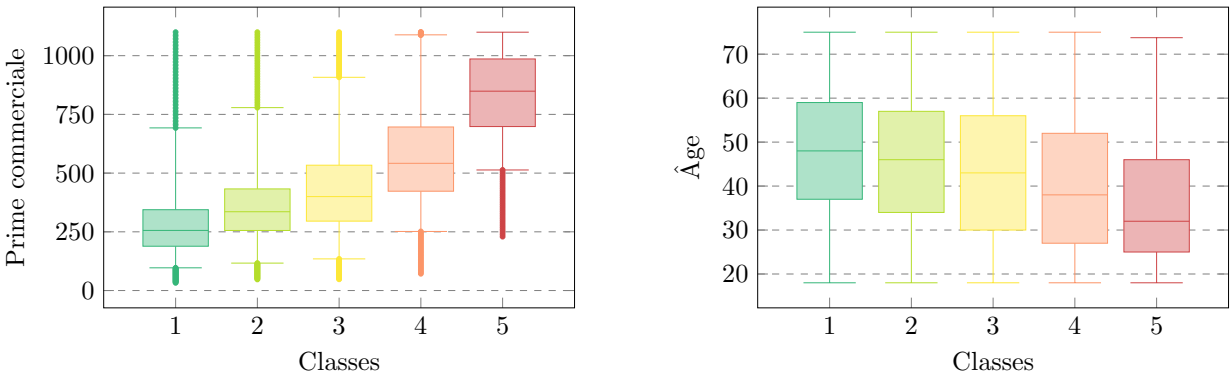
Nota Bene :

> *Les répartitions de chaque catégorie socio-professionnelle dans les classes de transformation sont disponibles en annexe A.7.*

Ces différences peuvent s'expliquer par des comportements distincts face aux offres d'assurance et par des besoins spécifiques. Les agriculteurs peuvent, par exemple, être confrontés à des besoins d'assurance plus immédiats pour protéger leurs outils de travail. Il peuvent alors tendre à accepter plus facilement les propositions d'assurance, d'où leur répartition plus élevée dans les classes 1 et 2. En revanche, les personnes ayant une profession libérale, du fait d'une plus grande flexibilité financière et d'attentes plus élevées, peuvent être plus sélectifs et peuvent être à la recherche d'offres plus personnalisées ou compétitives, ce qui explique leur présence plus marquée dans les classes 4 et 5, associées à une faible probabilité d'acceptation. De plus, il est clair que ce résultat est fortement corrélé au montant proposé par l'assureur. En effet, le montant moyen proposé aux agriculteurs est d'environ 360€ contre environ 530€ pour les professions libérales.

Le montant de cotisation proposé lors du devis semble donc avoir un impact sur ces deux catégories socio-professionnelles mais cela semble également être le cas sur l'ensemble du portefeuille. D'après la figure 4.10(a), représentant le *boxplot* (boîte à moustache) du montant du devis par classe, plus le montant de cotisation proposé est élevé, moins les individus ont de chances d'accepter le devis. Par exemple, le montant que premier quartile, *i.e.* quantile 25%, de la classe 5 (700€) est supérieur au quantile 99% de la classe 1 (690€).
De même, l'âge du conducteur a une impact significatif sur la répartition dans les classes de transformation puisque celui-ci décroit lorsque l'on est dans une moins bonne classe. Ce résultat est bien évidemment corrélé au montant de cotisation puisque les individus les plus jeunes sont les conducteurs les moins expérimentés, donc les plus risqués, et ainsi les individus possédant le montant de cotisation

le plus élevé avec, par exemple, un montant moyen d'environ 640€ chez les étudiants.

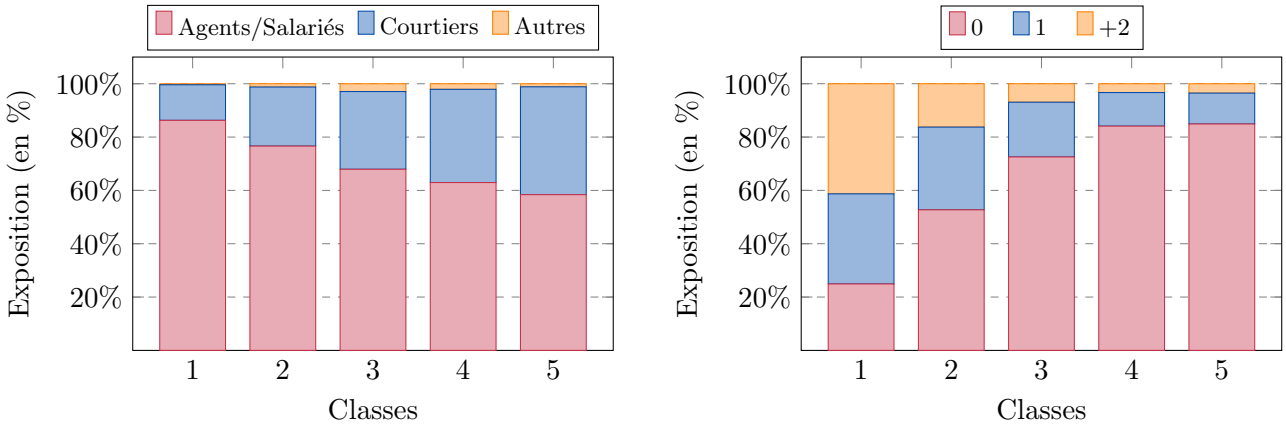


(a) Selon le montant de cotisation proposé

(b) Selon l'âge du conducteur

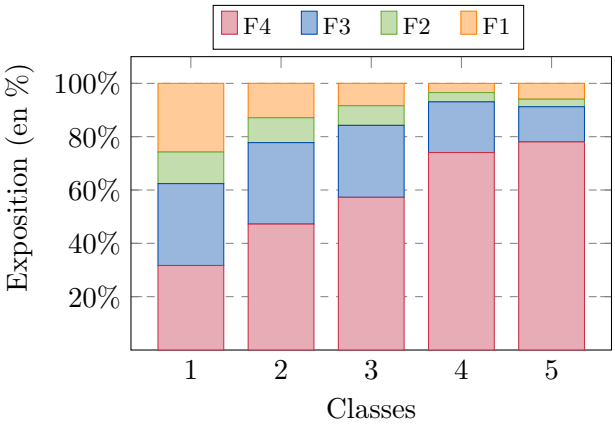
Figure 4.10 : Boxplots par classes de transformation selon le montant de cotisation proposé par l'assureur et l'âge du conducteur

Enfin, comme lors de l'analyse des classes de durée de vie des contrats, l'impact des différents réseaux de distribution, des formules proposées et du nombre d'autres contrats détenus par les assurés est examiné dans la figure 4.11. Tout d'abord, une relation de confiance semble être installée entre les agents généraux et les assurés puisque plus de 80% des devis ayant le plus de chances d'être acceptés sont proposés par des agents généraux et cette proportion baisse progressivement jusqu'à un peu moins de 60% dans la classe 5. De même, les clients qui possèdent déjà au moins un autre contrat chez cet assureur sont plus enclins à accepter le devis proposé avec près de 80% d'entre eux dans la classe 1 contre à peine plus de 15% dans la classe 5. Cela peut être lié à des offres commerciales proposées par l'assureur à ses clients les plus fidèles.



(a) Selon le réseau de distribution

(b) Selon le nombre de contrats différents



(c) Selon la formule proposée

FIGURE 4.11 : Répartition des modalités de plusieurs variables dans les classes de transformation

Pour finir, la formule proposée semble avoir un impact bien différent à ce qui pourrait être attendu. En effet, plus la couverture est étendue, moins les individus semblent vouloir accepter le montant proposé par l'assureur. Les formules plus complètes peuvent être perçues comme trop coûteuses ou incluant des garanties superflues, décourageant les individus qui privilégient des offres plus abordables et adaptées à leurs besoins réels.

4.5 Synthèse sur la transformation

Pour conclure sur ce chapitre, malgré l'absence d'informations sur l'acceptation des devis des individus du portefeuille d'affaires nouvelles, un modèle de transformation a été créé via une autre base de données du même assureur. Les probabilités de transformation des devis du portefeuille d'affaires nouvelles ont alors pu être prédites.

Ainsi, pour l'optimisation de la stratégie commerciale développée dans le prochain chapitre, les probabilités de transformation seront de nouveau prédites pour chaque individu i pour différentes variations du montant de cotisation. Il est ainsi possible de recalculé la valeur client future en partant du montant de la prime commerciale proposé $\mathrm{PC}_i(0)$ ayant subie une variation v_g, *i.e.* $\mathrm{PC}_i(0)\times(1+v_g)$, qui une fois multiplié par la probabilité d'acceptation du devis à ce prix ($Tr_i(v_g)$), donne la valeur client potentielle de l'individu i notée $VC_i(v_g)$ et définie par

$$VC_i(v_g) = \sum_{t=0}^{T_h} \frac{\mathrm{Marge}_i(t) \times \mathbb{P}_i(t)}{(1+r)^t} \times Tr_i(v_g) \tag{4.12}$$

où, pour tout $t \in [\![0, T_h]\!]$, les marges espérées de l'individu i $\mathrm{Marge}_i(t)$ et ses probabilités de survie $\mathbb{P}_i(t)$ dépendent du nouveau montant de cotisation proposé $\mathrm{PC}_i(0) \times (1+v_g)$.

Chapitre 5

Optimisation de la stratégie commerciale

Pour rappel, l'objectif de ce mémoire est de trouver l'allocation optimale de la prime commerciale proposée à chaque individu lors de la souscription d'un contrat d'assurance automobile qui maximise la valeur client potentielle du portefeuille. Pour cela, une méthodologie complexe d'optimisation de la stratégie commerciale, faisant appel à tous les modèles précédemment introduits, sera mise en place.

5.1 Premier résultat d'optimisation

Avant de mettre en place une optimisation complexe de la stratégie commerciale, il est possible d'effectuer une optimisation simple et directe pour maximiser la valeur client potentielle du portefeuille en appliquant une même variation v au montant de la cotisation proposée à l'entrée dans le portefeuille, notée $\mathrm{PC}_i(0)$ pour tout individu $i \in [\![1, n]\!]$.

Pour cela, il suffit de calculer la valeur client potentielle globale pour plusieurs variations de la cotisation et de regarder quelle est celle qui maximise cette valeur client potentielle. Ainsi, il suffit d'appliquer la formule (5.1) pour toute variation $v \in [-30\%, +30\%]$ du montant de la prime commerciale proposée à l'entrée dans le portefeuille.

$$\text{Valeur client potentielle globale} = \sum_{i=1}^{n} \overbrace{\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \mathrm{Marge}_i(t)}{(1+r)^t}}^{\text{Valeur client future de l'individu } i} \times Tr_i(v) \tag{5.1}$$

où pour tout $i \in [\![1, n]\!]$ et $t \in [\![0, T_h]\!]$, $\mathrm{Marge}_i(t)$ et $\mathbb{P}_i(t)$ dépendent du nouveau montant de cotisation proposé $\mathrm{PC}_i(0) \times (1 + v)$.

L'optimum correspond donc à la valeur client potentielle globale maximale et l'allocation optimale correspond à la variation associée.

Cependant, il est clair que modifier le montant de la prime commerciale proposée aux clients a une incidence directe sur la probabilité de transformation des devis. En effet, augmenter (respectivement diminuer) ce montant diminue (respectivement augmente) la probabilité de transformation. Ce problème d'optimisation consiste donc à trouver un équilibre entre l'indicateur de rentabilité et de fidélité, représenté par la valeur client future, et le taux d'acquisition de nouveaux contrats, représenté par la probabilité de transformation.

La formule (5.1) est donc appliquée pour toute variation $v \in [-30\%, +30\%]$ (à pas de 1%). Chaque point de la figure 5.1 représente une valeur client potentielle globale en fonction de la variation de prix associée (en abscisse) et de la transformation moyenne du portefeuille (en couleur). Ainsi, le point ayant pour abscisse 0% (entouré en bleu) correspond à la valeur client potentielle du portefeuille d'affaires nouvelles avec le prix initialement proposé par l'assureur (puisqu'aucune variation n'est appliquée). Cette valeur client s'élève à environ 93,2 millions d'euros pour une probabilité de transformation moyenne de 51,32% comme vu dans le chapitre précédent. L'optimum est égal à 94,4 millions d'euros et est atteint pour une variation du prix pour l'ensemble du portefeuille de 8%, induisant une transformation moyenne de 39,52%.

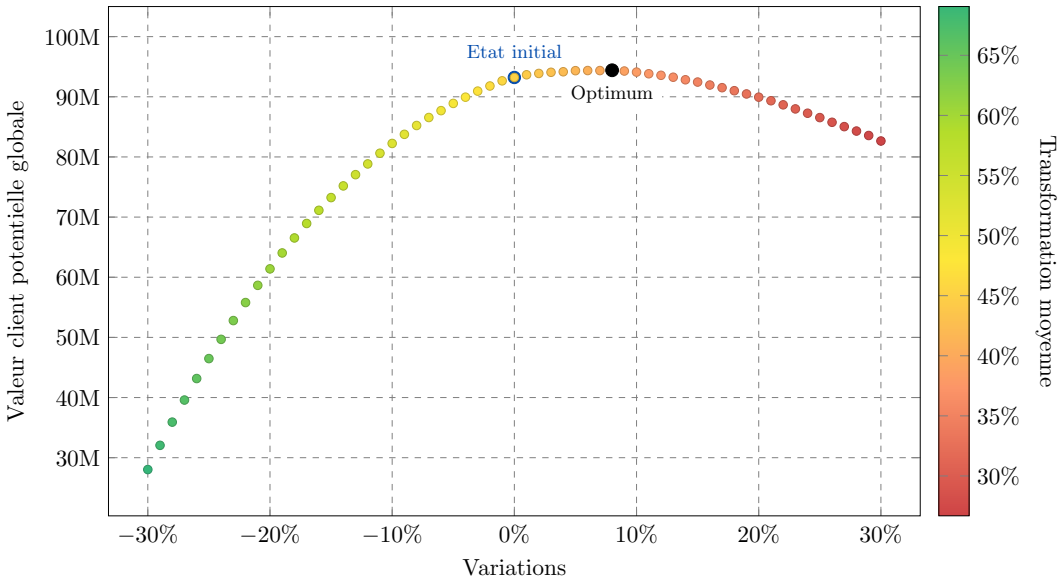


FIGURE 5.1 : Résultat de l'optimisation simple et directe

Nota Bene :

> *Les variations de prix étudiées s'étendent de −30% à +30%, un domaine relativement large qui peut ne pas refléter fidèlement la réalité du marché. Toutefois, ce choix s'inscrit dans un cadre théorique, où l'objectif est d'identifier un optimum à l'intérieur de cet intervalle, plutôt qu'une solution située au bord. Cette approche permet d'éviter des résultats biaisés par une contrainte trop restrictive sur l'amplitude des variations de prix. De plus, le modèle de transformation utilisé influence fortement l'optimisation mise en place. Étant donné que ce modèle présente certaines lacunes comme précisé à la sous-section 4.3.3, il peut impacter fortement la structure concave recherchée dans le problème d'optimisation, rendant ainsi nécessaire une amplitude plus large des variations appliquées pour garantir une analyse pertinente. Enfin, cette approche permettra de mieux observer l'impact des contraintes dans l'optimisation. Sans contrainte, les résultats ont tendance à se situer aux extrêmes, ce qui ne correspond pas toujours à un véritable optimum. L'ajout de contraintes, même simples, permet d'obtenir des évolutions de prix plus réalistes et mieux alignées avec les standards du marché. Dans une application réelle, ces contraintes seraient plus nombreuses et précises, mais leur effet resterait le même : guider naturellement l'optimisation vers des solutions cohérentes et applicables.*

Cette méthode d'optimisation s'avère être peu efficace dans la pratique puisqu'elle implique un changement de prix commun à tous les individus sans prendre en compte leurs caractéristiques. En effet, certains profils verront une augmentation de leur prix qu'ils ne devraient pas avoir, tandis que d'autres subiront une hausse moins importante. De plus, augmenter le prix de 8% a une incidence

considérable sur la probabilité de transformation moyenne du portefeuille puisque celle-ci diminue de presque 12%, pour un gain de seulement 1,2 millions d'euros de valeur client potentielle globale.

Il convient donc d'appliquer une autre méthode d'optimisation. Celle-ci repose sur la segmentation du portefeuille en différents clusters de profils similaires en termes de fidélité et de rentabilité, où les contrats subiront une variation du montant de la cotisation propre à leur cluster. Il conviendra toutefois de nuancer les résultats qui seront présentés dans la suite, notamment en raison du modèle de transformation, qui, comme précisé à la sous-section 4.3.3, présente certaines limites et ne permet pas d'estimer pleinement l'élasticité des clients aux variations tarifaires. Ce modèle ne prend pas en compte la concurrence et fait abstraction de la stratégie commerciale globale de l'assureur, qui pourrait influencer la fixation des prix au-delà des seules considérations de transformation et de rentabilité. La stratégie commerciale pourra d'ailleurs s'exprimer directement ou par un *proxy* via des contraintes supplémentaires, permettant d'intégrer des objectifs comme la part de marché ou la fidélisation dans l'optimisation.

5.2 Rappel de la démarche

L'objectif de ce mémoire est de trouver l'allocation optimale de la prime commerciale proposée aux individus souhaitant souscrire un contrat d'assurance automobile. Cette allocation est dite optimale dans le sens où elle maximise la valeur client potentielle du portefeuille d'affaires nouvelles sous différentes contraintes.

Pour rappel, sous les différentes notations ci-dessous, ce problème d'optimisation se traduit mathématiquement par la formule (5.2)

- $G \geq 1$ le nombre de clusters (distincts) formant le portefeuille
- $v_g \in [-30\%, +30\%]$ la variation appliquée au montant de la cotisation pour le cluster $g \in [\![1, G]\!]$
- n_g le nombre d'individus dans le cluster $g \in [\![1, G]\!]$
- $\mathbb{P}_i(t)$ la probabilité que l'individu i appartiennent encore au portefeuille au temps t
- $\text{Marge}_i(t)$ la marge espérée de l'individu i au temps t
- r le taux d'actualisation constant et égal à 2%
- $Tr_i(v_g)$ la probabilité de transformation de l'individu i en fonction de la variation v_g
- $VC_i(v_g)$ la valeur client potentielle de l'individu i à un horizon T_h en fonction de la variation appliquée au montant de la cotisation en entrée

$$\begin{cases} \max\limits_{v_1,\dots,v_G} \quad VC \text{ potentielle globale} = \sum\limits_{g=1}^{G} VC_g \\ \text{Sous contraintes} \end{cases} \tag{5.2}$$

$$\text{avec} \quad VC_g = \sum_{i=1}^{n_g} \overbrace{\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \text{Marge}_i(t)}{(1+r)^t} \times Tr_i(v_g)}^{VC_i(v_g)}$$

Le schéma de la figure 5.2 récapitule le processus mis en place pour calculer la valeur client potentielle globale que l'on cherche à maximiser :

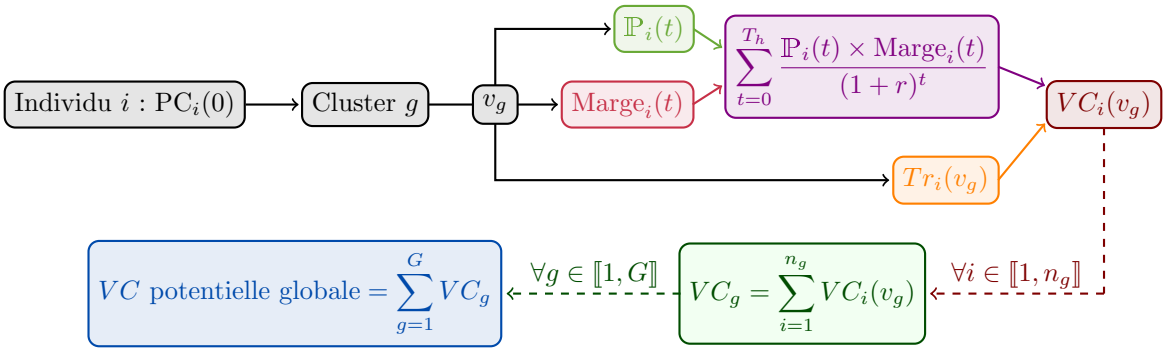


FIGURE 5.2 : Schéma récapitulatif du processus de calcul de la valeur client potentielle globale

Le processus est donc le suivant : l'assureur a calculé une prime commerciale pour un individu i, notée $\mathrm{PC}_i(0)$. À partir de ce montant et des caractéristiques initiales de son contrat, celui-ci est affecté à un cluster g de profils similaires (le choix des clusters sera développé dans la section suivante 5.3). Une variation v_g, commune à tous les individus du cluster g, est appliquée au montant de cotisation $\mathrm{PC}_i(0)$ qui devient alors égal à $\mathrm{PC}_i(0) \times (1 + v_g)$. À partir de celui-ci, pour tout temps $t \in [\![0, T_h]\!]$, les marges espérées ($\mathrm{Marge}_i(t)$) et les probabilités de survie du contrat ($\mathbb{P}_i(t)$) sont prédites afin de calculer une nouvelle valeur client future de l'individu i ($\sum_{t=0}^{T_h} \frac{\mathbb{P}_i(t) \times \mathrm{Marge}_i(t)}{(1+r)^t}$). La probabilité d'acceptation du devis à ce nouveau prix ($Tr_i(v_g)$) est également calculée, puis multipliée à la nouvelle valeur client future de l'individu i pour obtenir sa valeur client potentielle ($VC_i(v_g)$). Cela est effectué pour tous les individus du cluster g, et par somme, la valeur client potentielle du cluster est obtenue (VC_g). Enfin, cela est de nouveau réalisé pour tous les clusters, ce qui donne la valeur client potentielle globale à maximiser.

5.3 Choix et analyse des clusters

Avant d'expliciter le choix fait pour segmenter le portefeuille en différents clusters, il faut souligner l'importance de cette segmentation pour notre étude. En effet, une optimisation "tête par tête" aurait pu être mise en place. Cependant, celle-ci ne permet ni une bonne mutualisation des risques, ni d'identifier les profils similaires, et donc de mieux cibler les profils rentables à fidéliser. De plus, une telle optimisation nécessite de créer un problème de Lagrange, dont la résolution numérique peut être coûteuse en temps de calcul.

Segmenter un portefeuille en clusters de profils similaires permet d'analyser les comportements et les besoins des clients de manière plus globale. Cette approche aide à identifier des tendances, à anticiper les comportements et à ajuster les stratégies marketing en conséquence. Cela permet aussi de personnaliser les actions en fonction des segments, en créant des offres adaptées qui améliorent la fidélité et la satisfaction. Enfin, puisque l'on cherche à identifier les profils fidèles et rentables, segmenter selon la valeur client future offre un avantage supplémentaire permettant de concentrer les gestes commerciaux sur les clients qui semblent les plus prometteurs à long terme.

Ainsi, la valeur client future de tous les individus du portefeuille d'affaires nouvelles est calculée à partir du montant de cotisation $\mathrm{PC}_i(0)$, pour tout $i \in [\![1, n]\!]$, calculé initialement par l'assureur comme dans la section 3.4. La figure 5.3 représente la répartition des valeurs clients futures des individus du portefeuille.

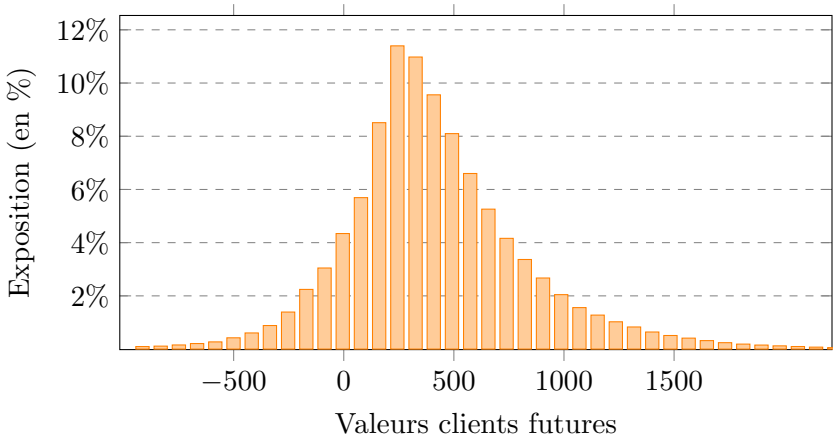
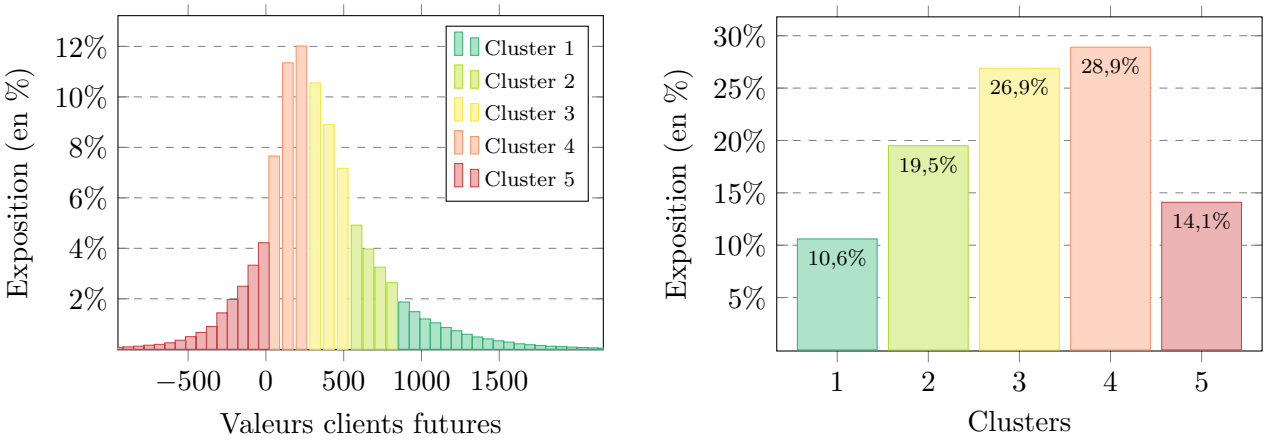


Figure 5.3 : Histogramme des valeurs clients futures à horizon 10 ans ($T_h = 9$)

Comme pour les classes de durée de vie des contrats et les classes de transformation, respectivement aux sections 2.7 et 4.4, cet histogramme est segmenté via la même méthode de classification ascendante hiérarchique. D'après la courbe d'inertie moyenne des 20 C.A.H. en fonction du nombre de groupes, le gain d'inertie ne semble plus être significatif entre 5 et 6 groupes. 5 clusters ont ainsi été créés et leur répartition est représentée dans la figure 5.4(b). La figure 5.4(a) représente l'histogramme des valeurs clients futures segmenté selon les clusters grâce aux bornes moyennes répertoriées dans le tableau 5.1. Il est clair que les individus du cluster 1 sont les plus fidèles et rentables puisqu'ils ont les valeurs clients futures les plus élevées du portefeuille. À l'inverse, plus de 85% des individus du cluster 5 ont des valeurs clients futures négatives, ce qui montre qu'ils sont soit déficitaires pour l'assureur, soit très peu rentables puisque la valeur client future maximale est d'environ 30€ dans ce cluster.



(a) Histogramme des valeurs clients futures

(b) Histogramme des clusters

Figure 5.4 : Répartitions des valeurs clients futures à horizon 10 ans ($T_h = 9$) et des clusters induits

Clusters	**5**	**4**	**3**	**2**	**1**
Bornes	$]-\infty;\ 32[$	$[32\ ;\ 296[$	$[296\ ;\ 552[$	$[552\ ;\ 906[$	$[906\ ;+\infty[$

Table 5.1 : Valeurs des bornes pour chaque cluster

L'évolution (cumulative) des valeurs clients futures moyennes selon les clusters a été tracée dans la figure 5.5 ci-dessous. La valeur client future utilisée pour segmenter le portefeuille correspond à celle au temps $t = T_h = 9$. À cette date, la valeur client future moyenne du cluster 1 est supérieure à 1200€ alors

que celle du cluster 5 est proche de -200€ et elle est négative pendant toute la période d'étude. Ce graphique reflète une fois encore les écarts importants de rentabilité au sein du portefeuille, ce qui démontre que la segmentation semble être primordiale pour mieux allouer les montants de cotisations proposés aux assurés lors de leur souscription.

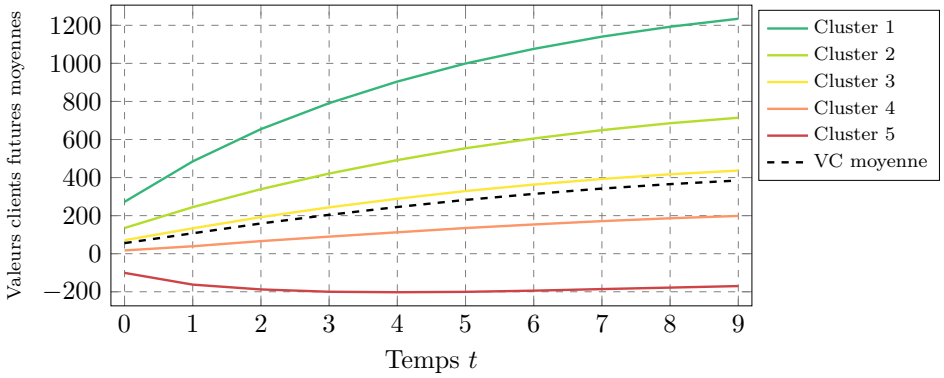
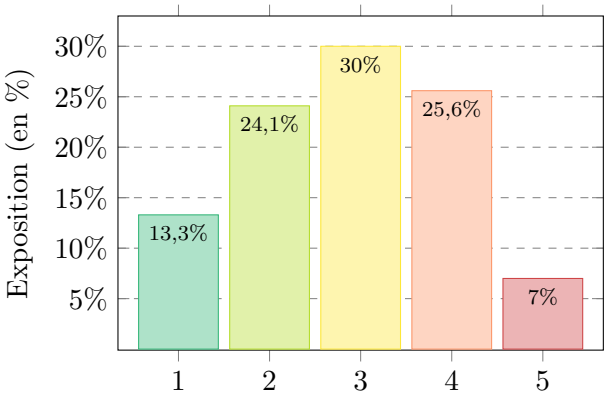


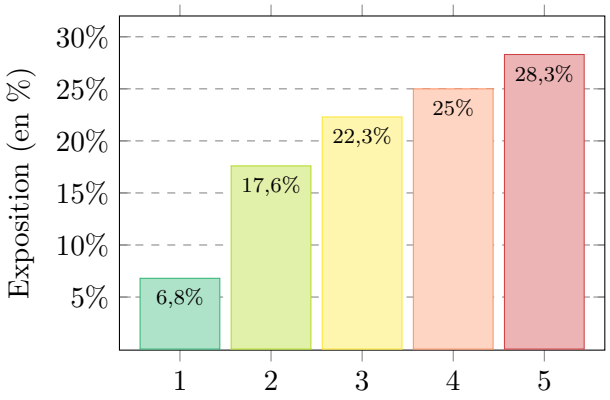
FIGURE 5.5 : Evolution de la valeur client future moyenne au cours du temps selon les différents clusters

Ensuite, en regardant comment se répartissent les modalités de certaines variables au sein des clusters, il est possible d'étudier leur impact sur la valeur client future, et donc d'avoir une idée des profils qui semblent, à priori, les plus fidèles et rentables.

Comme vu dans la figure 3.14, représentant l'évolution de la valeur client future moyenne au cours du temps selon la catégorie socio-professionnelle, les retraités sont les meilleurs profils en termes de valeur client future à horizon $T_h = 9$, avec une moyenne d'environ 500€, à l'inverse des étudiants qui sont les plus mauvais avec une valeur client future moyenne d'à peine plus de 250€ au bout de 10 ans. La répartition de ces deux catégories socio-professionnelles au sein des clusters est représentée dans la figure 5.6. Les retraités semblent avoir une répartition assez normale dans les clusters, avec très peu d'exposition dans le cluster 5, ce qui confirme le fait que la plupart des retraités sont fidèles et rentables. À contrario, les étudiants sont sous-représentés dans le cluster 1 et sont de plus en plus présents que le cluster est mauvais en termes de valeur client future, jusqu'à atteindre plus de 28% d'exposition dans le cluster 5.



(a) Pour les retraités (16,9%)



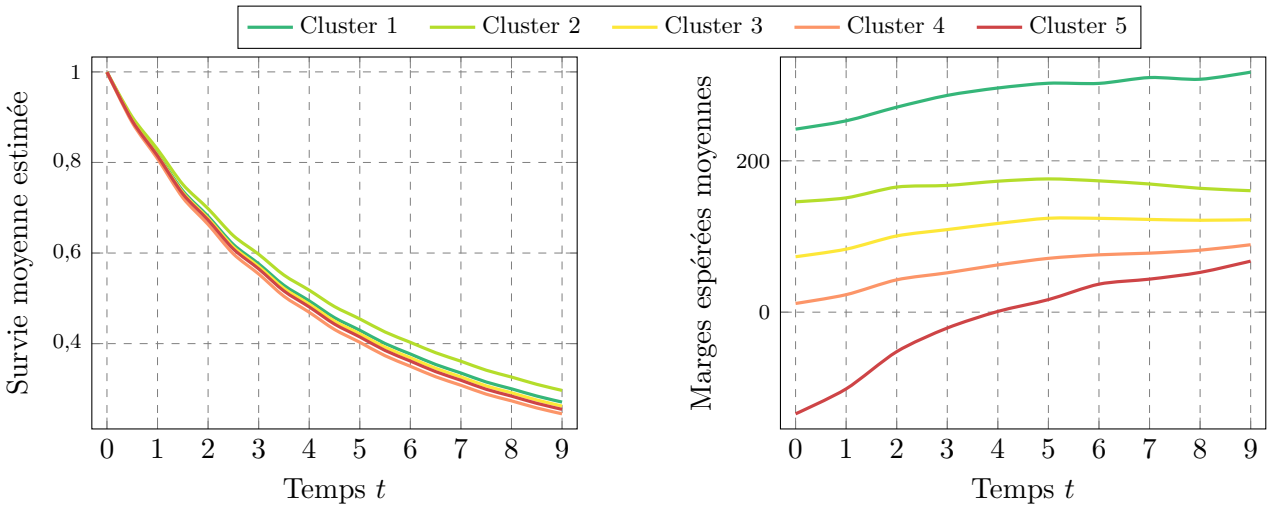
(b) Pour les étudiants (4,9%)

FIGURE 5.6 : Répartition de 2 catégories socio-professionnelles dans les clusters

Nota Bene :

> *Les répartitions de chaque catégorie socio-professionnelle dans les clusters sont disponibles en annexe A.8.*

Pour finir sur l'analyse de ces clusters, les deux principaux éléments définissant la valeur client future dans ce mémoire, *i.e.* la probabilité de survie et les marges espérées individuelles, ont été représentés en moyenne dans la figure 5.7 selon les clusters. Ces deux figures permettent de comprendre que la valeur client future dépend bien à la fois de la survie des contrats et des marges espérées réalisées sur ces derniers. Ainsi, d'après la figure 5.7(a), le cluster 1 et le cluster 5 ne sont pas les groupes ayant la survie la plus élevée ou la plus basse. C'est la combinaison de la survie avec les marges espérées qui permet d'identifier les meilleurs profils en termes de valeur client future. En effet, la survie permet d'identifier les individus les plus fidèles mais les plus rentables sont caractérisés par les marges espérées. Puisque celles du cluster 5 sont en moyenne négatives les 5 premières années d'assurance ($t = 4$) et peinent à atteindre les 60€ en moyenne au bout de 10 ans, il est clair que les individus les moins rentables sont présents dans ce cluster. À l'inverse, les marges espérées annuelles moyennes du cluster 1 augmentent progressivement au cours du temps, passant d'environ 240€ à plus de 300€ en 10 ans. La valeur client future est donc bien plus impactée par les marges espérées individuelles que les probabilités de survie puisque l'écart entre les marges espérées moyennes des différents clusters est beaucoup plus important que l'écart entre les fonctions de survie moyennes.



(a) Fonctions de survie moyennes estimées selon les différents clusters

(b) Evolution moyenne des marges espérées dans le temps selon les différents clusters

Figure 5.7 : Fonctions de survie moyennes estimées et évolution moyenne des marges espérées dans le temps selon les différents clusters

Maintenant que les clusters de profils similaires en termes de fidélité et rentabilité ont été créés, la résolution du problème d'optimisation défini à la section précédente 5.2 peut être réalisée, dans un premier temps sans contrainte.

5.4 Optimisation sans contrainte

Avant de résoudre le problème d'optimisation en tenant compte de diverses contraintes, il est possible d'optimiser la valeur client potentielle du portefeuille sans en imposer. En effet, le processus décrit dans la section 5.2 est appliqué pour toute variation du montant de cotisation initial comprise entre -30% et +30% (à pas de 2% pour des raisons de puissance de la machine utilisée). Ainsi, une valeur client potentielle et la probabilité de transformation moyenne associée sont calculées pour chaque cluster et pour chacune des variations possibles. Ces informations sont stockées sous forme de tableaux, dont les échantillons sont disponibles dans les tableaux 5.2 et 5.3 ci-dessous.

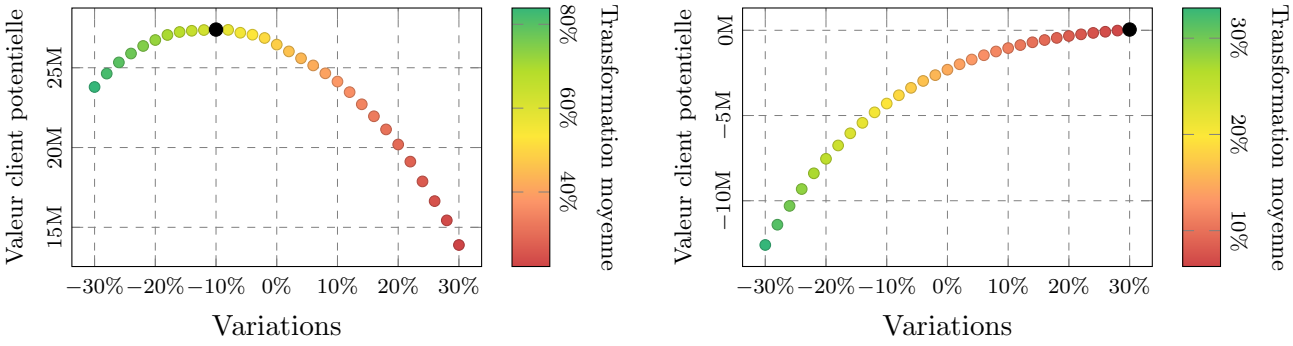
Variations \ Clusters	1	2	3	4	5
-30%	23,7	16,9	7,99	-8,02	-12,6
-28%	24,6	18,8	9,83	-5,94	-11,4
-26%	25,3	20,5	11,5	-3,97	-10,3
⋮	⋮	⋮	⋮	⋮	⋮
0%	26,7	32,8	25,0	11,1	-2,4
⋮	⋮	⋮	⋮	⋮	⋮
26%	16,6	28,3	25,0	15,3	-0,09
28%	15,4	27,7	24,8	15,2	-0,02
30%	13,9	27,1	24,5	15,1	0,05

TABLE 5.2 : Echantillon des valeurs clients potentielles des clusters (VC_g pour $g \in [\![1, 5]\!]$) selon chaque variation (en millions d'euros)

Variations \ Clusters	1	2	3	4	5
-30%	83,8%	83,9%	72,5%	68,0%	33,1%
-28%	81,6%	81,8%	70,8%	66,6%	31,5%
-26%	79,4%	79,8%	69,2%	65,1%	29,8%
⋮	⋮	⋮	⋮	⋮	⋮
0%	48,4%	54,5%	50,3%	47,3%	14,8%
⋮	⋮	⋮	⋮	⋮	⋮
26%	24,9%	33,0%	34,2%	32,7%	7,0%
28%	23,6%	31,7%	33,2%	31,8%	6,6%
30%	22,2%	30,4%	32,1%	30,8%	6,3%

TABLE 5.3 : Echantillon des probabilités de transformation moyenne des clusters selon chaque variation

Ainsi, pour trouver la solution du problème d'optimisation sans contrainte, il suffit de sommer les valeurs clients potentielles maximales de chaque cluster, disponibles dans le tableau 5.2, puisque VC potentielle globale $= \sum_{g=1}^{G} VC_g$ où $G = 5$. La figure 5.8 représente les valeurs clients potentielles des clusters 1 et 5 en fonction de la variation appliqué au montant de cotisation des individus des clusters (en abscisse), et en fonction de la probabilité de transformation moyenne (en couleur) issue du tableau 5.3. Les points noirs • correspondent aux optima globaux pour ces deux clusters.



(a) Cluster 1

(b) Cluster 5

Figure 5.8 : Résultats de l'optimisation sans contrainte des clusters 1 et 5

Nota Bene :

> *Ces même graphiques de résultats de l'optimisation sans contrainte sont disponibles pour tous les clusters en annexe A.9.*

Les résultats sont répertoriés dans le tableau 5.4. Les deux premières colonnes correspondent à la valeur client potentielle et à la tranformation moyenne si aucune variation n'est appliquée au montant de cotisation proposé aux individus à l'entrée dans le portefeuille. La troisième colonne représente l'allocation optimale des variations recherchée dans notre étude. Enfin, les deux dernières colonnes correspondent à la valeur client potentielle et la transformation moyenne calculées suite à l'application des variations optimales.

Les individus du cluster 1, étant les profils les plus fidèles et rentables en termes de valeur client future, bénéficient d'une réduction de 10% de leur prime commerciale. Ceci implique nécessairement une hausse de plus de 10% de la probabilité de transformation moyenne, permettant ainsi l'intégration d'un plus grand nombre de profils similaires dans le portefeuille. Le cluster 2 ne subit aucun changement, puis le montant de la prime commerciale augmente progressivement entre les clusters 3, 4 et 5.

	Valeur client potentielle initiale	**Transformation moyenne initiale**	**Variation de la prime commerciale**	**Valeur client potentielle**	**Transformation moyenne**
Cluster 1	26,7M €	48,4%	-10%	27,4M €	59,9%
Cluster 2	32,8M €	54,5%	0%	32,8M €	54,5%
Cluster 3	25,0M €	50,3%	+12%	25,9M €	42,4%
Cluster 4	11,1M €	47,3%	+26%	15,3M €	32,7%
Cluster 5	-2,4M €	14,8%	+30%	45 224 €	6,3%
Portefeuille	**93,2M €**	**51,32%**	**+13,92%**	**101,4M €**	**38,72%**

Table 5.4 : Tableau de résultats de l'optimisation sans contrainte

Cependant, comme vu avec la première optimisation simple et directe à la section 5.1, augmenter le prix globalement impacte fortement la transformation moyenne du portefeuille. Cela est encore une fois le cas ici puisque cette dernière a baissé de presque 13% en augmentant en moyenne le montant de la prime commerciale de près de 14%. Cette solution maximise la valeur client potentielle du portefeuille à 101,4 millions d'euros, ce qui représente une hausse de plus de 8 millions par rapport à l'état initial, et surtout une hausse de 7 millions par rapport à la solution d'une augmentation de 8% de la prime commerciale de tous les individus. La variation de cette nouvelle solution est nettement

plus significative, pourtant la transformation ne diminue que de façon marginale (moins de 1%) par rapport à la solution précédente. Cela montre l'impact d'une segmentation du portefeuille, qui permet de mieux cibler les profils à attirer et ceux sur lesquels augmenter la prime commerciale, et donc à fortiori la marge espérée, malgré leur potentielle non acceptation.

Bien que cette solution maximise la valeur client potentielle du portefeuille, celle-ci n'est pas idéale dans la pratique puisqu'elle implique une importante réduction du volume du portefeuille, directement liée à une augmentation moyenne du montant de la prime commerciale trop élevée. La mise en place de contraintes de compétitivité est donc nécessaire.

5.5 Optimisation sous contraintes

Avant de passer aux résultats, il convient d'expliciter le choix des contraintes et de les définir mathématiquement.

5.5.1 Choix des contraintes

Les contraintes mises en place ont pour objectif de rester compétitif vis à vis de la concurrence tout en assurant un bon volume du portefeuille et des différents clusters. En effet, 4 contraintes ont été choisies : 2 sur les clusters et 2 au niveau global.

- ▷ **Contraintes sur les clusters :**
  - (C_1) Appliquer des variations comprises entre -30% et +30% au montant de la prime commerciale
  - (C_2) Ne pas perdre plus de 25 points de transformation dans chaque cluster
- ▷ **Contraintes globales :**
  - (C_3) Ne pas perdre plus de 10 points de transformation dans l'ensemble du portefeuille
  - (C_4) Ne pas excéder +5% de variation moyenne du montant de la prime commerciale

Les contraintes sur les clusters se traduisent simplement en : pour tout cluster $g \in [\![1, G]\!]$, l'application d'une variation v_g comprise entre -30% et +30% à la prime commerciale des individus ne doit pas faire baisser la probabilité de transformation moyenne de plus de 25% par rapport à la transformation moyenne initiale. Ainsi, il suffit de garder uniquement les cellules des tableaux précédents 5.2 et 5.3 qui vérifient pour tout $g \in [\![1, G]\!]$ et pour toute variation $v_g \in [-30\%, +30\%]$ (à pas de 2%) (C_1)

$$(C_2) \quad \frac{1}{n_g} \sum_{i=1}^{n_g} Tr_i(v_g) \geq \frac{1}{n_g} \left(\sum_{i=1}^{n_g} Tr_i(0\%) \right) - 25\%$$

Par exemple, après application de ces contraintes, les clusters 1 et 2 ne pourront respectivement pas subir une variation de la prime commerciale supérieure à +22% et +26%, puisqu'à partir de celles-ci les probabilités de transformation moyenne ont diminué de plus de 25 points de transformation par rapport à l'état initial.

Les contraintes globales se traduisent par : la probabilité de transformation moyenne du portefeuille doit être supérieure à la transformation moyenne initiale (51,32%) moins 10% et la variation moyenne appliquée doit être inférieure à 5%. Pour vérifier ces contraintes, une méthodologie combinatoire est mise en place. Il suffit de récupérer toutes les valeurs clients potentielles des clusters et les probabilités de transformation associées pour chaque combinaison de variations possibles sur les

clusters (5 clusters et 31 variations possibles $\implies 31^5$ combinaisons), puis d'uniquement garder les combinaisons $(v_1, ..., v_G) \in [-30\%, +30\%]^G$ qui vérifient

$$(C_3)\ \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} Tr_i(v_g) \geq \underbrace{\frac{1}{n}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} Tr_i(0\%)\right)}_{=51,32\%} - 10\% \qquad (C_4)\ \frac{1}{n}\sum_{g=1}^{G} n_g v_g \leq 5\%$$

Les contraintes mises en place pour cette étude sont assez simples, mais un avantage certain de notre méthodologie combinatoire est qu'elles sont facilement modulables. En effet, les contraintes sur les clusters s'appliquent simplement sur les tableaux 5.2 et 5.3, et une fois toutes les combinaisons calculées, des contraintes globales différentes de celles définies précédemment peuvent être appliquées très aisément.

5.5.2 Résultats de l'optimisation sous contraintes

Ainsi, la valeur client potentielle du portefeuille pour une combinaison correspond donc à la somme des valeurs clients potentielles des clusters associées à cette combinaison. La solution du problème d'optimisation sous contraintes est donc la plus grande valeur client potentielle globale parmi toutes les combinaisons qui vérifient les contraintes. La combinaison de variations associée est l'allocation optimale recherchée dans cette étude.

Cette solution optimale est représentée dans le tableau 5.5, ainsi que dans la figure 5.9 où chaque point correspond à la valeur client potentielle du cluster en abscisse et dont la couleur renvoie à la probabilité de transformation moyenne de celui-ci. Ces probabilités et les variations appliquées sont annotées au-dessus de ces points.

	Valeur client potentielle initiale	**Transformation moyenne initiale**	**Variation de la prime commerciale**	**Valeur client potentielle**	**Transformation moyenne**
Cluster 1	26,7M €	48,4%	-16%	27,2M €	67,2%
Cluster 2	32,8M €	54,5%	0%	32,8M €	54,5%
Cluster 3	25,0M €	50,3%	+2%	25,3M €	48,9%
Cluster 4	11,1M €	47,3%	+12%	14,4M €	40,2%
Cluster 5	-2,4M €	14,8%	+18%	-445 967 €	8,8%
Portefeuille	**93,2M €**	**51,32%**	**+4,9%**	**99,3M €**	**43,77%**

Table 5.5 : Tableau de résultats de l'optimisation sous contraintes

La solution vérifie bien les contraintes globales car la probabilité de transformation moyenne est égale à 43,77% qui est bien supérieure à $51,32\% - 10\% = 41,32\%$, et la variation moyenne est égale à 4,9%. Un équilibre entre les variations appliquées aux différents clusters a donc été trouvé par rapport à la solution sans contrainte, pour une perte minime de valeur client potentielle globale d'environ 2 millions d'euros. Le montant de cotisation proposé aux individus du cluster 1 lors de leur souscription doit diminuer de 16% afin d'attirer et de fidéliser un maximum de ces profils fidèles et rentables à long terme. Le cluster 2, comme pour la solution sans contrainte, ne subit aucune variation, et enfin les autres clusters voient leur montants proposés augmenter progressivement mais de manière moins abrupte que dans la solution précédente.

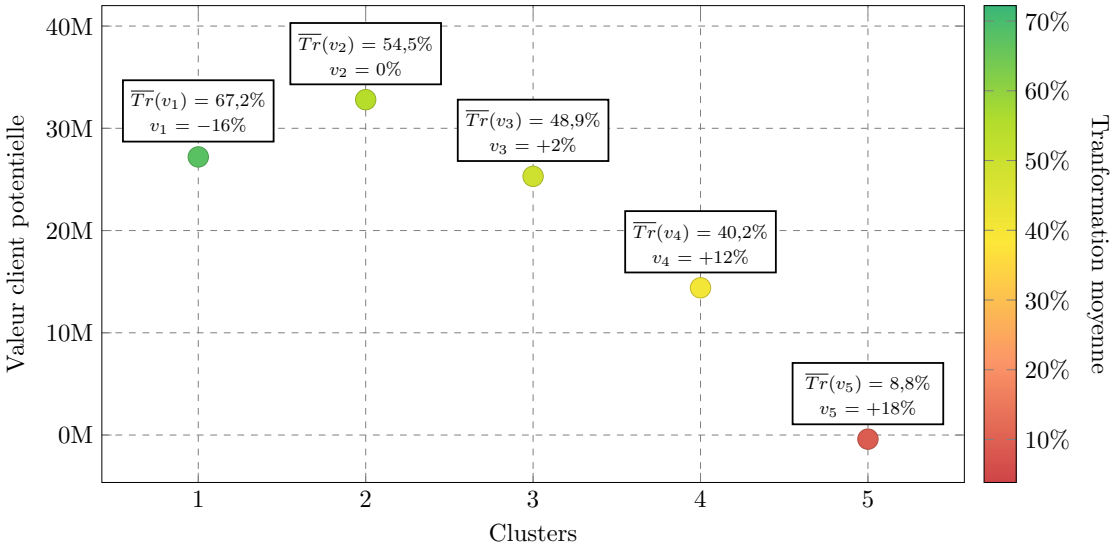


FIGURE 5.9 : Résultats de l'optimisation sous contraintes

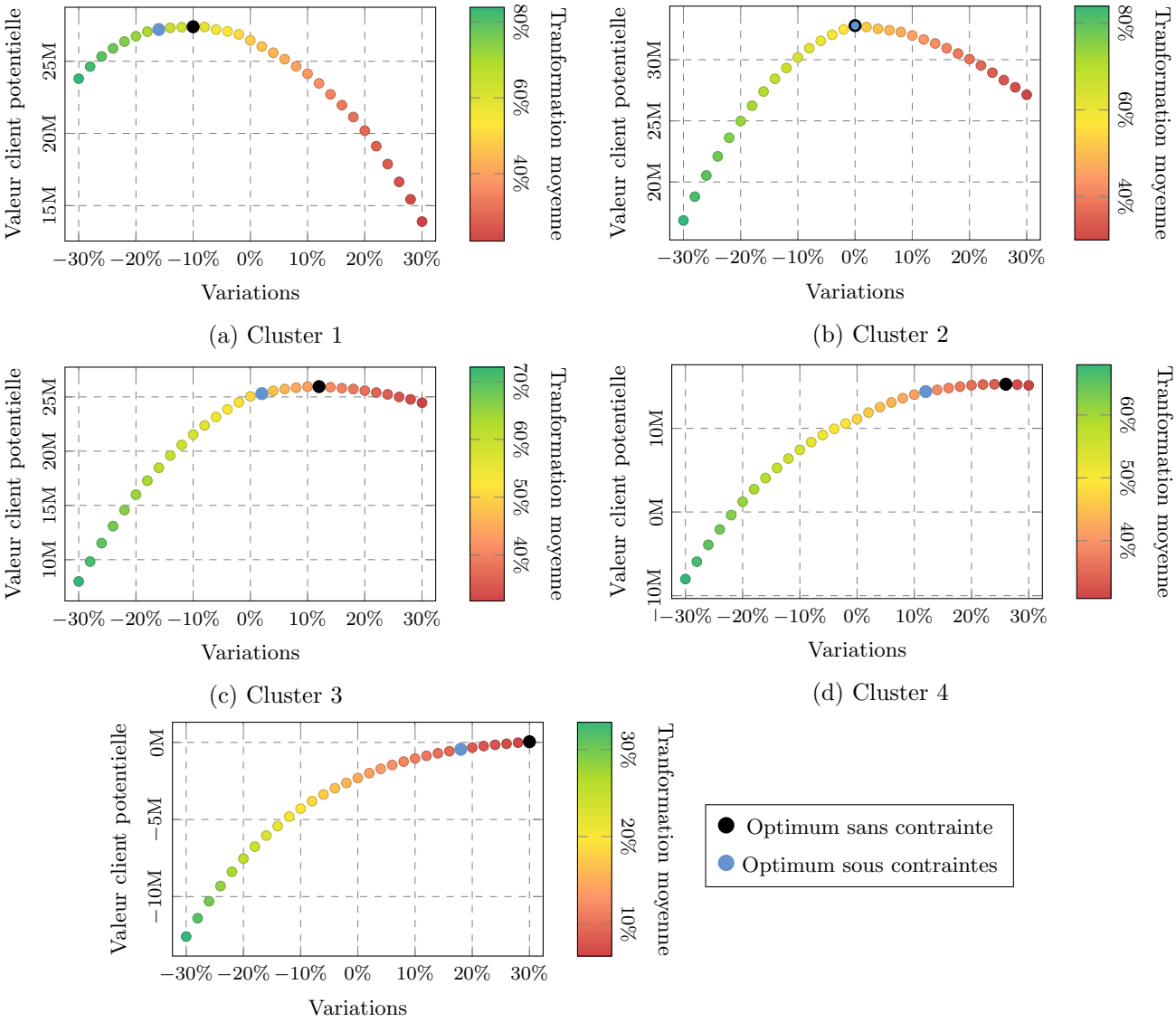
Ainsi, la valeur client potentielle a augmenté de plus de 6 millions d'euros par rapport à la situation initiale, simplement en segmentant le portefeuille et en appliquant des contraintes très simples.

5.5.3 Comparaison des résultats

Il est possible de comparer plus en détails les résultats des optimisations sans et sous contraintes. Pour cela, les graphiques représentant les valeurs clients potentielles en fonction de la variation appliquée et de la probabilité de transformation moyenne sont affichés pour chaque cluster dans la figure 5.10. Les points noirs • et les points bleus • correspondent respectivement aux optima sans et sous contraintes.

Ces graphiques permettent de bien observer les différences entre les deux solutions trouvées. Un décalage des optima s'opèrent vers la gauche, excepté le cluster 2 qui reste au même point, lors de l'application des contraintes afin de trouver un équilibre pour les satisfaire.

Enfin, les sommes des primes commerciales et des marges espérées collectées par l'assureur la première et la dixième année d'assurance sont répertoriées dans le tableau 5.6 avant optimisation et après l'application des 3 méthodes d'optimisation (simple, sans et sous contraintes). Ces grandeurs permettent de rendre compte directement de l'effet des différentes stratégies commerciales sur le chiffre d'affaire et sur les gains générés à court et long termes pour l'assureur.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

Figure 5.10 : Résultats des optimisations sans et sous contraintes de la stratégie commerciale pour chaque cluster

	Somme des primes à $t = 0$	Somme des primes à $t = T_h = 9$	Somme des marges espérées à $t = 0$	Somme des marges espérées à $t = T_h = 9$	Valeur client potentielle globale	Transformation moyenne
Avant optimisation	199,0M €	191,1M €	29,7M €	56,6M €	93,2M €	51,32%
Après optimisation simple	214,9M €	193,6M €	45,6M €	60,2M €	94,4M €	39,52%
Après optimisation sans contrainte	220,6M €	195,5M €	51,3M €	61,1M €	101,4M €	38,72%
Après optimisation sous contraintes	204,3M €	191,9M €	35,1M €	57,5M €	99,3M €	43,77%

Table 5.6 : Tableau comparatif des différentes stratégies commerciales

Dans tous les cas, optimiser la stratégie commerciale en faisant varier le montant de cotisation proposé aux individus lors de leur souscription a un impact positif sur les grandeurs du tableau 5.6

(sauf la transformation moyenne). Cependant, l'assureur doit trouver un bon équilibre entre maximiser ses profits immédiats et sa rentabilité à long terme tout en minimisant sa perte de nouveaux clients. La meilleure stratégie commerciale semble toutefois bien être celle de l'optimisation sous contraintes qui trouve cet équilibre en essayant d'attirer des profils potentiellement fidèles et rentables tout en augmentant le tarif de ceux qui le sont moins.

5.5.4 Innovation de la méthodologie d'optimisation

L'innovation de cette méthodologie d'optimisation réside dans la flexibilité qu'elle apporte grâce à la segmentation et aux contraintes modulables. En effet, une fois ces valeurs client potentielles calculées pour chaque cluster et chaque variation (cf. tableau 5.2), il devient possible de moduler à l'infini les contraintes de l'optimisation. Pour vérifier les contraintes, une méthode combinatoire est mise en place. Cette méthode consiste à parcourir toutes les combinaisons possibles de variations tarifaires pour chaque cluster, en récupérant les valeurs client potentielles et les probabilités de transformation associées. Dans cette étude, avec 5 clusters et 31 variations possibles par cluster (de -30% à $+30\%$), cela représente 31^5 combinaisons. La méthode permet alors de sélectionner uniquement les combinaisons $(v_1, \dots, v_G) \in [-30\%; +30\%]^G$ qui respectent l'ensemble des contraintes imposées, comme les seuils de rentabilité et de rétention.

Ainsi, en utilisant cette méthode combinatoire, les stratégies tarifaires peuvent être ajustées de manière continue et personnalisée, sans nécessiter de recalculer les valeurs client potentielles. Cette modularité permet ainsi d'adapter rapidement les tarifs aux évolutions du marché et des objectifs commerciaux. Cette approche est également facilement opérationnelle, car elle permet de mettre en place des ajustements tarifaires précis et modulables de manière réactive, en s'appuyant sur des calculs préétablis. Cela favorise une meilleure adaptation aux contraintes du marché sans augmenter la charge de travail liée au recalcul de la valeur client potentielle pour chaque ajustement tarifaire.

L'optimisation présentée précédemment peut être affinée en intégrant des contraintes supplémentaires permettant d'assurer un équilibre entre rentabilité, compétitivité et cohérence commerciale. Voici des exemples de contraintes pouvant être rajoutées afin que la stratégie tarifaire optimale reste réaliste et alignée avec les objectifs commerciaux de l'assureur :

- ▷ *Stabilité tarifaire :* Limiter les variations excessives des primes d'une période à l'autre pour éviter de perturber les clients et assurer une cohérence commerciale.
- ▷ *Rentabilité minimale par segment :* Garantir qu'aucun segment de clientèle ne devienne non rentable en imposant une marge minimale par groupe.
- ▷ *Cohérence tarifaire :* Restreindre les écarts de primes au sein d'un même segment afin d'éviter des différences trop marquées entre profils similaires.
- ▷ *Gagner des parts de marché minimale :* Augmenter le volume de souscriptions sur l'ensemble du portefeuille.
- ▷ *Couverture des profils à risque :* Préserver une diversification du portefeuille en imposant un minimum de souscriptions pour les profils considérés comme plus risqués.

Les résultats obtenus sur les clusters de clients peuvent également être exploités au-delà des ajustements tarifaires pour définir des actions différenciées selon les segments. Plutôt que d'appliquer un simple mouvement tarifaire à l'entrée, ces résultats peuvent orienter des adaptations de produits et des stratégies commerciales ciblées. Par exemple, les profils considérés comme plus risqués pourraient se voir proposer des garanties plus limitées avec des franchises plus élevées, tandis que les segments jugés plus intéressants en termes de fidélisation et de rentabilité pourraient faire l'objet d'efforts commerciaux renforcés sur le multi-équipement.

5.5.5 *Reverse engineering*

Maintenant que l'optimisation de la stratégie commerciale a été réalisée. Il serait intéressant de mettre en place une méthode de *reverse engineering* pour identifier les principaux facteurs expliquant la segmentation en clusters du portefeuille étudié. En effet, le *reverse engineering* est une méthode consistant à "décomposer" un modèle pour examiner son fonctionnement et analyser sa structure ainsi que les variables sous-jacentes qui le régissent. Une potentielle approche serait donc de construire un modèle linéaire généralisé, en particulier une régression logistique multinomiale, où la variable à expliquer est le cluster. Puisque les clusters prennent des valeurs entières allant de 1 à 5, une telle régression permettrait d'identifier les variables explicatives influençant l'appartenance des individus à un cluster spécifique.

Une fois le modèle optimal créé, la prédition de la classification des individus dans les clusters peut être réalisée sur les données. Les variations spécifiques aux clusters, issues des résultats de l'optimisation sous contraintes, peuvent être appliquées aux montants de cotisation des individus initialement calculés par l'assureur. Ces nouveaux montants pourraient alors être intégrés en tant qu'*offset* dans le modèle de tarification de l'assureur, permettant ainsi d'ajuster au mieux les prix proposés pour se rapprocher des prix optimisés tout en gardant la structure de modélisation existante de grille tarifaire.

5.6 Axes d'amélioration et limites du mémoire

Bien que ce mémoire soit relativement complet, il présente certaines limites et possède de nombreux axes d'amélioration pouvant donner des résultats plus précis. Voici une liste non exhaustive des aspects pouvant être améliorés :

▷ **Valeur client future :**

Tout d'abord, dans ce mémoire, un abus de langage a été fait concernant la valeur client future. Comme précisé dans le chapitre 1, une inspiration de la valeur client future a été définie et utilisée par manque d'informations sur la sinistralité des individus (étant donné que l'étude porte sur les affaires nouvelles) et sur les coûts d'acquisition.

De plus, puisque l'étude a uniquement été appliquée sur un portefeuille d'assurance automobile, il s'agit en réalité d'une valeur contrat. En effet, la valeur client future est un indicateur très complexe qui fait appel au multi-équipement. En d'autres termes, un client peut posséder plusieurs contrats différents (automobile, MRH, prévoyance, etc.) et sa valeur client future est définie comme la somme de ses valeurs contrats futures, pondérées par des probabilités de souscription à de nouveaux produits.

Par manque d'informations sur ce multi-équipement, l'étude a été restreinte à un seul contrat. Elle peut cependant facilement être généralisée à d'autres types de produits d'assurance, couplés à des modèles de souscription.

▷ **Modèle à risques proportionnels de Cox :**

Le principal problème de ce modèle de durée de vie est qu'il ne prend pas en compte l'évolution du montant de la prime commerciale dans le temps. Une amélioration possible serait d'implémenter une des extensions de ce modèle : le modèle de Cox à *time varying covariates*, où la prime commerciale est la variable qui varie dans le temps. Bien que l'on dispose de ces primes commerciales projetées sur 10 ans grâce au modèle de projection de marges, ce type de modèle est toutefois assez compliqué à créer et est excessivement lourd en temps de calcul.

Un approche simpliste serait d'ajouter les primes commerciales prédites pour chaque période parmi

les variables explicatives du modèle à risques proportionnels de Cox créé au chapitre 2.

▷ **Modèle de projection de marges :**

Comme expliqué à la sous-section 3.2.1, lors du traitement des données et de la transformation du portefeuille en base image, les avenants n'ont pas été considérés, *i.e.* les caractéristiques initiales de chaque contrat sont gardées pour toute la période de projection. Pour pallier à cela, des chaînes de Markov peuvent être utilisées.

En effet, la propriété de Markov est la suivante : l'état du temps t dépend uniquement de l'état au temps $t-1$. Ainsi, il est possible de calculer des probabilités de changements d'états (véhicule, formule, adresse, etc.) à chaque période sur plusieurs segments du portefeuille (la segmentation permet d'avoir des probabilités différentes selon les profils). Une matrice de transition peut alors être créée sur chaque segment et sert à calculer les primes annuelles des individus par pondération des primes calculées en considérant les différents changements de situation.

Cette solution est cependant très coûteuse en temps de calcul puisqu'elle implique le calcul d'une matrice de transition sur chaque segment du portefeuille, puis la prédiction de k primes commerciales et k primes pures par individu et par période si l'on considère k changements d'états possibles.

▷ **Modèle de transformation :**

Concernant le modèle de transformation, celui-ci peut être grandement amélioré puisqu'il ne capte pas assez bien la sensibilité au prix des individus, comme précisé à la sous-section 4.3.3. En effet, pour chaque contrat, un seul montant et donc une seule probabilité d'acceptation est disponible. Il est donc compliqué d'obtenir une bonne sensibilité pour les individus.

La meilleure solution (en l'absence d'informations sur différents devis proposés aux clients) serait de créer un modèle PSM (*Propensity Score Matching*) afin d'agrandir artificiellement le portefeuille avec plusieurs devis par contrat.

Le principe est le suivant : le portefeuille est scindé en k groupes plus ou moins bons en terme de conversion des devis (par exemple les classes de transformation de la section 4.4). Pour chaque contrat de chaque groupe, le modèle recherche n profils similaires (appelés jumeaux) dans les autres groupes et récupère le montant de cotisation proposé à ces jumeaux. Le même contrat, avec les mêmes caractéristiques, apparaît désormais $n \times (k-1)+1$ fois dans le portefeuille avec $n \times (k-1)+1$ montants différents. Enfin, le modèle de transformation est appliqué à ce portefeuille élargi et donne diverses probabilités par contrat, permettant une meilleure évaluation de la sensibilité au prix des potentiels clients.

Néanmoins, un tel modèle nécessite de créer un modèle pour détecter les jumeaux entre les groupes et implique surtout une multiplication très importante de la taille de la base. Pour simplifier le processus et limiter la multiplication de la taille du portefeuille par 3, il peut être choisi de ne rechercher, pour un contrat, qu'un seul jumeau dans les deux groupes les plus proches (en termes de transformation) de celui du contrat.

▷ **Limites éthiques et d'équité actuarielle :**

Enfin, les résultats de ce mémoire ne doivent pas inciter les assureurs à discriminer les profils les moins fidèles et rentables. En effet, via la segmentation mise en place dans ce dernier chapitre, les mauvais profils en termes de valeur client future subissent une augmentation importante de leur montant de cotisation à l'entrée dans le portefeuille. Cette hausse peut être interprétée comme une discrimination de ces individus et peut mener à une mauvaise mutualisation du portefeuille si l'assureur propose des tarifs excessifs à ces individus.

Le principe de l'équité actuarielle requiert que les primes payées par les assurés soient proportionnelles au risque qu'ils représentent pour l'assureur, sans qu'une discrimination excessive soit exercée envers certains profils.

Il est donc important de mettre en perspective les impacts sociaux que peuvent avoir une telle optimisation de la stratégie commerciale. La flexibilité des contraintes via la méthode combinatoire utilisée dans l'étude permet cela. En effet, comme expliqué à la sous-section 5.5.1, les contraintes sont extrêmement modulables. Il est possible de les rendre davantage strictes en limitant par exemple les variations ou la probabilité transformation moyenne spécifiquement à chaque cluster, ou encore en assurant un seuil minimum de profils particuliers à attirer (étudiants, personnes ayant souscrit par l'intermédiaire d'un courtier d'assurance, etc.).

Conclusion

L'objectif de ce mémoire était de créer une méthodologie innovante pour identifier les profils fidèles et rentables à acquérir et pour optimiser la stratégie commerciale des nouveaux contrats en maximisant la valeur client potentielle du portefeuille étudié. Cette étude a permis de démontré l'importance d'adopter une vision à long terme pour maximiser non seulement les profits immédiats, mais aussi la rentabilité potentielle des assurés sur toute la durée de leur relation avec l'assureur.

Les différentes étapes de la démarche, reposant sur des techniques de modélisation avancées, ont permis d'identifier les éléments contribuant à l'optimisation du tarif à la souscription. Le modèle à risques proportionnels de Cox a servi à estimer la durée de vie des contrats. Parallèlement, un modèle de projection de marges complexe a permis d'estimer la rentabilité future pour chaque contrat, en intégrant uniquement les informations initiales et le montant proposé à la souscription. Ces deux modèles ont ensuite été combinés à un modèle de transformation afin de rendre compte de la sensibilité des individus aux différentes offres tarifaires étudiées lors de la recherche de l'optimum.

L'un des principaux apports de cette étude est la démonstration qu'il est possible d'optimiser la stratégie commerciale dès l'entrée des clients dans le portefeuille. En segmentant ce dernier en clusters de profils similaires en termes de valeur client future, puis, en appliquant des ajustements de primes différenciés selon les clusters, la valeur client potentielle globale a pu être maximisée tout en maintenant des tarifs compétitifs, grâce à des contraintes prédéfinies, pour attirer les nouveaux souscripteurs. Cette approche permet de répondre à la fois aux exigences de rentabilité de l'assureur et à celles de compétitivité dans un marché fortement concurrentiel.

Les différentes optimisations menées dans ce mémoire ont permis d'améliorer significativement la stratégie commerciale de l'assureur. Pour chaque méthode, en ajustant le montant de la cotisation à l'entrée, des améliorations significatives en termes de rentabilité à long terme du portefeuille ont été constatées. Les estimations ont montré qu'une optimisation sans contrainte maximise la valeur client potentielle, mais au prix d'une forte augmentation des primes pour certains segments, entrainant un taux d'acceptation des devis proposés moins élevé. En revanche, l'optimisation sous contraintes équilibre davantage les intérêts de l'assureur, en garantissant des probabilités de transformation compétitives tout en maximisant les marges annuelles dégagées. Les résultats finaux montrent que chaque cluster a bénéficié d'une modification spécifique du tarif. Une augmentation modérée des primes dans certains clusters a permis d'accroître les marges sans affecter de manière significative les taux de conversion des devis. À l'inverse, des ajustements à la baisse chez les profils les plus fidèles et rentables ont renforcé l'attractivité des offres, favorisant ainsi une probabilité d'acceptation plus élevée. Cette stratégie a permis de maximiser la valeur client potentielle tout en limitant les pertes de prospects.

Plusieurs défis ont émergé tout au long de ce travail de recherche. Premièrement, l'absence de données sur la sinistralité individuelle des nouveaux assurés a nécessité l'utilisation de la prime pure comme approximation des risques, ce qui pourrait être affiné avec l'accès à des données plus détaillées. Deuxièmement, chaque modèle utilisé pourrait être considérablement amélioré si des ressources de

calcul plus puissantes étaient disponibles. Par ailleurs, un abus de langage a été utilisé tout au long de l'étude en faisant référence à la "valeur client", alors qu'il s'agit en réalité d'une "valeur contrat". Cette distinction est importante, car l'analyse s'est limitée aux contrats d'assurance automobile. Une complexification de cette approche, en l'appliquant à plusieurs types de produits d'assurance, permettrait une vision plus complète de la relation client sur l'ensemble des produits proposés par l'assureur. Enfin, bien que l'optimisation tarifaire présentée ici permette de maximiser la valeur client potentielle du portefeuille, elle pourrait soulever des questions d'ordre éthique, notamment en matière de discrimination pour certains profils jugés moins rentables. Une vigilance particulière devra être apportée pour que cette optimisation ne nuise pas aux principes de mutualisation et d'équité actuarielle qui constituent deux des fondements de l'assurance.

Bibliographie

Andersen P. K. & Gill R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* Vol. 10 (No. 4) : p. 1100-1120.

Andersen P. K., Borgan Ø., Gill R. D. & Keiding N. (1996). Statistical Models Based on Counting Processes. Springer Series in Statistics. Springer New York.

Barlow W. E. & Prentice R. L. (1988). Residuals for Relative Risk Regression. *Biometrika* Vol. 75 (No. 1) : p. 65-74.

Bendle N. T., Farris P. W., Pfeifer P. E. & Reibstein D. J. (2010). Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Pearson Education, Inc. Upper Saddle River, New Jersey.

Berger P. D. & Nasr N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* Vol. 12 (No. 1) : p. 17-30.

Berkson J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* Vol. 39 (No. 227) : p. 357-365.

Berkson J. (1951). Why I Prefer Logits to Probits. *Biometrics* Vol. 7 (No. 5) : p. 327-339.

Blattberg R. C. & Deighton J. A. (1996). Manage marketing by the customer equity test. *Harvard Business Review* Vol. 74 (No. 4) : p. 136-144.

Bouaziz O. (2016–2017). *Analyse de survie : Méthodes non paramétriques*. Support de cours. Université Paris Descartes. URL : https://helios2.mi.parisdescartes.fr/~obouaziz/KMSurv.pdf.

Breslow N. E. (1972). Contribution to the discussion on the paper by D. R. Cox, Regression models and Life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 34 (No. 2) : p. 216-217.

Breslow N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics* Vol. 30 (No. 1) : p. 89-99.

Breslow N. E. (1975). Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review* Vol. 43 (No. 1) : p. 45-57.

Breslow N. E. & Crowley J. J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics* Vol. 2 (No. 3) : p. 437-453.

Brookmeyer R. & Crowley J. J. (1982). A Confidence Interval for the Median Survival Time. *Biometrics* Vol. 38 (No. 1) : p. 29-41.

Chen T. & Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. *Conference: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* : p. 785-794.

Cox D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 20 (No. 2) : p. 215-242.

Cox D. R. (1972). Regression models and Life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 34 (No. 2) : p. 187-220.

Cox D. R. (1975). Partial Likelihood. *Biometrika* Vol. 62 (No. 2) : p. 269-276.

Damien Hennom (2016). *Création d'un indicateur de valeur client en assurance non-vie*. Mémoire d'actuariat. Paris : ENSAE.

Dwyer F. R. (1997). Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing* Vol. 11 (No. 4) : p. 6-13.

France Assureurs (2023). Les données clés de l'assurance française en 2022 : p. 1-110.

France Assureurs (2024). L'assurance automobile des particuliers en 2024 : p. 1-4.

Gardes L. (2023). *Chapitre 3 - Test de comparaison de fonctions de survie*. Cours Analyse de survie, 14 à 18. Gardes (2023). Institut de Recherche Mathématique Avancée (IRMA), Strasbourg. url : https://irma.math.unistra.fr/~gardes/Resume_Cours_Survie.pdf.

Gill R. D. (1980). Censoring and Stochastic Integrals. Mathematical Centre Tracts 124. Sponsored by the Netherlands Organization for Advancement of Pure Research, Amsterdam.

Gill R. D. (1992). *Lectures on Survival Analysis*. Ecole d'Eté de Probabilités de Saint Flour XXII, ed. P. Bernard. Springer Lecture Notes in Mathematics. url : https://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/c2503b7c346eda1cc1257f130064ce83/$FILE/stflour0.pdf.

Gini, C. (1912). Variabilità e Mutuabilità. Republié dans Memorie di metodologia statistica en 1955. Ed. E. Pizetti & T. Salvemini.

Goodman L. A. & Kruskal W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* Vol. 49 (No. 268) : p. 732-764.

Green P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternative. *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 46 (No. 2) : p. 149-192.

Greenwood M. (1926). The Natural Duration of Cancer. *Reports on Public Health and Medical Subjects* (No. 33) : p. 1-26.

Gupta S. & Zeithaml V. A. (2006). Customer Metrics and Their Impact on Financial Performance. *Marketing Science* Vol. 25 (No. 6) : p. 718-739.

Gupta S., Lehmann D. R. & Stuart J. A. (2004). Valuing Customers. *Journal of Marketing Research* Vol. 41 (No. 1) : p. 7-18.

Harrell F. E., Califf R. M., Pryor D. B., Lee K. L. & Rosati R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* Vol. 247 (No. 18) : p. 2543-2546.

Kaplan E. L. & Meier Paul (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* Vol. 53 (No. 282) : p. 457-481.

Kumar V., Ramani G. & Bohling T. (2004). Customer lifetime value approaches and best practice applications. *Journal of Interactive Marketing* Vol. 18 (No. 3) : p. 60-72.

Lin D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis* Vol. 13 : p. 471-480.

Lorenz M. O. (1905). Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* Vol. 9 (No. 70) : p. 209-219.

Margaux Regnault (2023). *Modélisation des comportements de fidélité des assurés : sensibilité au prix et applications à l'optimisation tarifaire sur un portefeuille auto*. Mémoire d'actuariat. Paris : ENSAE.

McCullagh P. & Nelder J. A. (1989). Generalized Linear Models. Second Edition. Statistics and Applied Probability (No. 37). Boca Raton : Chapman et Hall.

Nelder J. A. & Wedderburn R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)* Vol. 135 (No. 3) : p. 370-384.

Reinartz W. J. & Kumar V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. *Journal of Marketing* Vol. 64 (No. 4) : p. 17-35.

Reinartz W. J. & Kumar V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing* Vol. 67 (No. 1) : p. 77-99.

Schoenfeld D. (1982). Partial Residuals for The Proportional Hazards Regression Model. *Biometrika* Vol. 69 (No. 1) : p. 239-241.

Shapley L. S. (1953). A Value for n-Person Games. *Contributions to the Theory of Games II* : sous la dir. d'Harold W. Kuhn & Albert W. Tucker, p. 307-317.

Shaw R. (1993). Making Database Marketing Work. *Journal of Information Technology* Vol. 8 (No. 2) : p. 110-117.

Stute W. (1994). The Bias of Kaplan-Meier Integrals. *Scandinavian Journal of Statistics* Vol. 21 (No. 4) : p. 475-484.

Stute W. (1995). The Statistical Analysis of Kaplan-Meier Integrals. *Lecture Notes-Monograph Series* Vol. 27 : p. 231-254.

Therneau T. M. & Grambsch P. M (1994). Proportional hazard tests and diagnostics based on weighted residuals. *Biometrika* Vol. 81 (No. 3) : p. 515-526.

Therneau T. M. & Grambsch P. M. (2000). Modeling Survival Data: Extending the Cox Model. Statistic for Biology and Health. Springer New York.

Therneau T. M. & Watson D. A. (2017). The concordance statistic and the Cox model. *Technical Report - Department of Health Sciences Research, Mayo Clinic Rochester, Minnesota* Vol. 85 : p. 515-526.

Therneau T. M., Grambsch P. M. & Fleming T. R. (1990). Martingale-based residuals for survival models. *Biometrika* Vol. 77 (No. 1) : p. 147-160.

Venkatesan R. & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Marketing* Vol. 68 : p. 106-125.

Annexes

A.1 Propriétés de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier construit à la section 2.2 possède plusieurs propriétés intéréssantes.

Absence de censure :

Si l'échantillon étudié ne présente aucune censure, alors l'estimateur de Kaplan-Meier $\widehat{S}^{KM}$ coïncide parfaitement avec la fonction de survie empirique $\widehat{S}_n$

$$\widehat{S}_n = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)} \leq t\}} \tag{3}$$

où $\left(T_{(1)}, ..., T_{(n)}\right)$ est la statistique d'ordre* du n-échantillon $(T_1, ..., T_n)$.

Comportement asymtotique :

Si pour tout $t \in \mathbb{R}_+$ $S(t) > 0$, alors sous les mêmes notations que dans la section 2.1, l'inégalité suivante est trouvée

$$0 \leq \mathbb{E}\left[\widehat{S}^{KM}(t) - S(t)\right] \leq F(t)K(t)^n \tag{4}$$

Ainsi, l'estimateur de Kaplan-Meier est biaisé mais asymtotiquement sans biais si $K(t) \neq 1$. En effet, d'après l'ouvrage de Gill (1980) *Censoring and Stochastic Integrals*, en posant $\tau_K = \inf\left\{t \in \mathbb{R}_+ | \; K(t) = 1\right\}$, alors la convergence en probabilité suivante est obtenue

$$\sup_{0 \leq t \leq \tau_K} \left|\widehat{S}^{KM}(t) - S(t)\right| \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \tag{5}$$

De même, l'article *Cox's Regression Model for Counting Processes : A Large Sample Study* de Andersen et Gill (1982) donne la convergence en loi suivante, pour tout $t \leq \tau < \tau_K$

$$\sqrt{n}\left(\widehat{S}^{KM}(t) - S(t)\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathcal{V}^2(t)\right) \tag{6}$$

où $\mathcal{V}^2(t) = \left(S(t)\right)^2 \displaystyle\int_0^t \frac{f(u)}{\left(S(u)\right)^2\left(1 - G(u)\right)}\, du = \left(S(t)\right)^2 \int_0^t \frac{h(u)}{1 - K(u)}\, du$

Nota Bene :

> *L'estimateur de Kaplan-Meier présente des problèmes de convergence dans la queue de distribution à cause de la censure. En effet, $\widehat{S}^{KM}(t)$ ne peut pas être consistant pour $t > \tau$ car il n'y a pas d'observations au delà de τ_K par définition de K.*

**i.e.* échantillon classé dans l'ordre croissant

L'estimateur (consistant) de Greenwood (1926) permet d'estimer la variance asymtotique $\mathcal{V}^2$ et est défini, pour tout $t < \tau_K$, par

$$\widehat{\mathcal{V}}^2(t) = \left(\widehat{S}^{KM}(t)\right)^2 \sum_{t_j<t} \frac{d_j}{r_j(r_j - d_j)} \tag{7}$$

Il est donc possible de construire l'intervalle de confiance asymptotique de la fonction de survie $S(t)$ au niveau $1-\alpha$, $\alpha \in]0,1[$

$$IC^A_\alpha = \left[\widehat{S}^{KM}(t) - q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}} \frac{\widehat{\mathcal{V}}(t)}{\sqrt{n}} \;;\; \widehat{S}^{KM}(t) + q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}} \frac{\widehat{\mathcal{V}}(t)}{\sqrt{n}}\right] \tag{8}$$

avec $q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\dfrac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0,1)$.

A.2 Théories des tests de Wald, du *log-rank* et du rapport de vraisemblances

Test de Wald

Le test de Wald est utilisé pour évaluer l'importance des coefficients dans le modèle de Cox. Il teste l'hypothèse nulle selon laquelle un ou plusieurs coefficients dans un modèle de régression sont égaux à zéro. La formule pour la statistique de Wald et la p-value associée dépend du contexte spécifique, mais dans le cas d'un modèle à risques proportionnels de Cox ou d'un modèle linéaire généralisé, voici comment est défini ce test.

Soit $\beta = (\beta_1, ..., \beta_p)^\top$ le vecteur des coefficients de la régression et $\widehat{\beta} = \left(\widehat{\beta}_1, ..., \widehat{\beta}_p\right)^\top$ le vecteur de leurs estimations.

Pour un coefficient β_k, $k = 1, ..., p$, il faut tester :

$$\mathcal{H}_0 : \beta_k = 0 \qquad \text{contre} \qquad \mathcal{H}_1 : \beta_k \neq 0$$

Pour cela, la statistique de test se calcule avec la formule suivante

$$W_k = \frac{\widehat{\beta}_k^2}{\mathbb{V}\left(\widehat{\beta}_k\right)} \tag{9}$$

Celle-ci suit asymptotiquement une loi du Khi-Deux à 1 degré de liberté. Par conséquent, l'hypothèse nulle est rejetée avec un risque d'erreur $\alpha \in]0, 1[$ si $W_k > q_{1-\alpha}^{\chi_1^2}$, avec $q_{1-\alpha}^{\chi_1^2}$ le quantile d'ordre $1-\alpha$ de la loi du Khi-Deux à 1 degré de liberté.

La p-value est donc donnée par

$$p\text{-value} = \mathbb{P}\left(W_k > \chi_1^2\right) = 1 - F_{\chi_1^2}(W_k) \tag{10}$$

où $F_{\chi_1^2}$ est la fonction de répartition d'une loi du Khi-Deux à 1 degré de liberté.

De manière analogique, il est également possible de tester la nullité de plusieurs coefficients. Il faut alors tester les hypothèses suivantes pour $r \in [\![1, p]\!]$:

$$\mathcal{H}_0 : \beta_r = ... = \beta_p = 0 \qquad \text{contre} \qquad \mathcal{H}_1 : \exists k \in [\![r, p]\!] \text{ tel que } \beta_k \neq 0 \ ^*$$

La statistique de test est définie par

$$W = \left(\widehat{\beta}_r, ..., \widehat{\beta}_p\right) \left[\mathbb{V}\begin{pmatrix}\widehat{\beta}_r \\ \vdots \\ \widehat{\beta}_p\end{pmatrix}\right]^{-1} \begin{pmatrix}\widehat{\beta}_r \\ \vdots \\ \widehat{\beta}_p\end{pmatrix} \tag{11}$$

Celle-ci suit asymptotiquement une loi du Khi-Deux à $p - r - 1$ degrés de liberté. Par conséquent, l'hypothèse nulle est rejetée avec un risque d'erreur $\alpha \in]0, 1[$ si $W > q_{1-\alpha}^{\chi^2_{p-r-1}}$, avec $q_{1-\alpha}^{\chi^2_{p-r-1}}$ le quantile d'ordre $1-\alpha$ de la loi du Khi-Deux à $p - r - 1$ degrés de liberté.

La p-value est donc donnée par

$$p\text{-value} = \mathbb{P}\left(W > \chi^2_{p-r-1}\right) = 1 - F_{\chi^2_{p-r-1}}(W) \tag{12}$$

*L'hypothèse alternative $\mathcal{H}_1$ peut se réécrire : au moins un coefficient β_k est non nul, pour $k \in [\![r, p]\!]$.

où $F_{\chi^2_{p-r-1}}$ est la fonction de répartition d'une loi du Khi-Deux à $p-r-1$ degrés de liberté.

Test du *log-rank*

Le test du *log-rank*, aussi connu sous le nom de test de Mantel-Cox, compare les courbes de survie entre deux ou plusieurs groupes. Ce test est utilisé pour vérifier si les groupes ont des risques proportionnels constants dans le temps. Celui-ci n'est valide que sous l'hypothèse de censure à droite.

Le test du *log-rank* est cohérent avec le modèle à risques propotionnels de Cox car il évalue la même hypothèse nulle d'égalité des taux de hasard entre les groupes, ce qui implique que les ratios de hasard sont constants dans le temps, une des principales hypothèses de ce modèle de durée.

La présentation suivante est inspirée du chapitre *Test de comparaison de fonctions de survie* du support de cours de LAURENT GARDES (2023) sur l'*Analyse de survie*, pages 14 à 18.

COMPARAISON DE 2 GROUPES :

Plaçons-nous dans le cadre de 2 groupes indépendants d'individus et notons n_1 et n_2 les nombres d'individus dans les groupes 1 et 2. Les groupes étant indépendants, les individus d'un groupe ne peuvent être présents dans l'autre et alors $n = n_1 + n_2$ représente le nombre d'individus dans la fusion des groupes. Pour les deux groupes, les échantillons suivants sont observés :

$$\begin{cases} \left(\mathbf{Y}^{(1)}, \mathbf{\Delta}^{(1)}\right) := \left\{\left(Y_1^{(1)}, \delta_1^{(1)}\right), ..., \left(Y_{n_1}^{(1)}, \delta_{n_1}^{(1)}\right)\right\} \\ \left(\mathbf{Y}^{(2)}, \mathbf{\Delta}^{(2)}\right) := \left\{\left(Y_1^{(2)}, \delta_1^{(2)}\right), ..., \left(Y_{n_2}^{(2)}, \delta_{n_2}^{(2)}\right)\right\} \end{cases}$$

Ils sont supposés indépendants et respectivement de même loi que les couples $\left(Y^{(1)}, \delta^{(1)}\right)$ et $\left(Y^{(2)}, \delta^{(2)}\right)$, avec $Y^{(j)} = \inf\left\{T^{(j)}, C^{(j)}\right\}$ et $\delta^{(j)} = \mathbb{1}_{\left\{T^{(j)} \leq C^{(j)}\right\}}$ pour $j \in \{1, 2\}$ où $T^{(j)}$ et $C^{(j)}$ représentent respectivement la durée de vie et la variable de censure $\big($cf. définition de la censure à droite (2.1.2)$\big)$.

De la même façon, les observations sont notées

$$\begin{cases} \left(\mathbf{y}^{(1)}, \mathbf{d}^{(1)}\right) = \left\{\left(y_1^{(1)}, d_1^{(1)}\right), ..., \left(y_{n_1}^{(1)}, d_{n_1}^{(1)}\right)\right\} \\ \left(\mathbf{y}^{(2)}, \mathbf{d}^{(2)}\right) = \left\{\left(y_1^{(2)}, d_1^{(2)}\right), ..., \left(y_{n_2}^{(2)}, d_{n_2}^{(2)}\right)\right\} \end{cases}$$

L'objectif du test est de tester l'égalité en loi des deux groupes. Pour cela, les fonctions de survie de ces deux groupes $S^{(1)}(\cdot) = \mathbb{P}(Y^{(1)} > \cdot)$ et $S^{(2)}(\cdot) = \mathbb{P}(Y^{(2)} > \cdot)$ sont comparées :

$$\mathcal{H}_0 : S^{(1)}(\cdot) = S^{(2)}(\cdot) \qquad \text{contre} \qquad \mathcal{H}_1 : S^{(1)}(\cdot) \neq S^{(2)}(\cdot)$$

Sous l'hypothèse nulle $\mathcal{H}_0$, le n-échantillon $(\mathbf{Y}, \mathbf{\Delta}) = \{(Y_1, \delta_1), ..., (Y_n, \delta_n)\}$ est i.i.d. de même loi que le couple (Y, δ) et les observations sont notées $(\mathbf{y}, \mathbf{d}) = \{(y_1, d_1), ..., (y_n, d_n)\}$.

Pour construire le test, les notations suivantes sont adoptées :

- ▷ $\{y_{(1)} < ... < y_{(n)}\}$ la statistique d'ordre de $\{y_1, ..., y_n\}$
- ▷ pour $j \in \{1, 2\}$, le vecteur $\mathbf{H}^{(j)} := \left(h_1^{(j)}, ..., h_n^{(j)}\right)$, avec pour $i \in [\![1, n]\!]$ $h_i^{(j)} = h^{(j)}(y_i)$ où $h^{(j)}(\cdot)$ est la fonction de hasard $\big($cf. (2.1.1)$\big)$ associée à la variable aléatoire $T^{(j)}$
- ▷ pour $i \in [\![1, n]\!]$, $O_i := O_i^{(1)} + O_i^{(2)}$ et $N_i := N_i^{(1)} + N_i^{(2)}$ respectivement le nombre d'évènements observés et le nombre d'individus à risque dans la réunion des deux groupes à la date $y_{(i)}$

De plus, pour $j \in \{1,2\}$, l'estimateur du maximum de vraisemblance de $\mathbf{H}^{(j)}$ est défini par

$$\widehat{\mathbf{H}}_m^{(j)} := \left(\widehat{h}_{m,1}^{(j)}, ..., \widehat{h}_{m,n}^{(j)}\right) = \left(\frac{O_1^{(j)}}{N_1^{(j)}}, ..., \frac{O_n^{(j)}}{N_n^{(j)}}\right) \tag{13}$$

Ainsi, sous $\mathcal{H}_0$, $\mathbf{H}^{(1)} = \mathbf{H}^{(2)} =: \mathbf{H}$, et en fusionnant les deux groupes, l'estimateur du maximum de vraisemblance de $\mathbf{H}$ s'écrit :

$$\widehat{\mathbf{H}}_m^{\cdot} = \left(\widehat{h}_{m,1}, ..., \widehat{h}_{m,n}\right) = \left(\frac{O_1}{N_1}, ..., \frac{O_n}{N_n}\right) \tag{14}$$

Donc sous $\mathcal{H}_0$, le résultat attendu est

$$\sum_{i=1}^{n} N_i^{(1)} \left(\widehat{h}_{m,i}^{(1)} - \widehat{h}_{m,i}\right) \approx 0 \approx \sum_{i=1}^{n} N_i^{(2)} \left(\widehat{h}_{m,i}^{(2)} - \widehat{h}_{m,i}\right)$$

Le test du *log-rank* est donc basé sur la variable aléatoire $K(\mathbf{Y}, \mathbf{\Delta})$ dont la valeur observée par

$$K := \sum_{i=1}^{n} \left(O_i^{(1)} - N_i^{(1)} \frac{O_i}{N_i}\right) \tag{15}$$

Si $\mathcal{H}_0$ est rejetée, K peut avoir deux types de comportements :

- ▷ Si les courbes de survie ne se croisent pas, *i.e.* si $\widehat{h}_{m,i}^{(1)} < \widehat{h}_{m,i}^{(2)}$ ou $\widehat{h}_{m,i}^{(1)} > \widehat{h}_{m,i}^{(2)}$ pour tout $i \in [\![1, n]\!]$, alors, sous l'hypothèse alternative $\mathcal{H}_1$, une valeur de K très différente de 0 dit être observée. K permet donc bien de conclure sur la validité de $\mathcal{H}_0$.
- ▷ Si les courbes de survie se croisent, une valeur de K proche de 0 peut être observée sous $\mathcal{H}_1$. Dans ce cas, le risque de seconde espèce* est plus important ce qui rend le test moins puissant.

Ainsi, pour décider si la valeur observée de la variable aléatoire $K(\mathbf{Y}, \mathbf{\Delta})$ est suffisamment éloignée de 0 pour rejeter $\mathcal{H}_0$, il faut connaître le comportement asymptotique de $K(\mathbf{Y}, \mathbf{\Delta})$ sous $\mathcal{H}_0$.
La statistique de test $Z(\mathbf{Y}, \mathbf{\Delta})$ est donc introduite et sa valeur observée est donnée par

$$Z(\mathbf{y}, \mathbf{d}) := \frac{\left(\sum_{i=1}^{n} O_i^{(1)} - N_i^{(1)} \frac{O_i}{N_i}\right)^2}{\mathcal{V}_m} \text{ où } \mathcal{V}_m := \sum_{i=1}^{n} \frac{N_i^{(1)} N_i^{(2)}}{N_i} \frac{O_i}{N_i} \left(1 - \frac{O_i}{N_i}\right) \tag{16}$$

En supposant que les lois des couples $\left(T^{(1)}, C^{(1)}\right)$ et $\left(T^{(2)}, C^{(2)}\right)$ sont indépendantes, et s'il existe $c \in]0,1[$ tel que $\frac{n_1}{n} \xrightarrow[n \to +\infty]{} c$, alors sous $\mathcal{H}_0$, $Z(\mathbf{Y}, \mathbf{\Delta})$ suit asymptotiquement une loi du Khi-Deux à 1 degré de liberté.

Par conséquent, l'hypothèse nulle est rejetée avec un risque d'erreur $\alpha \in]0,1[$ si $Z(\mathbf{Y}, \mathbf{\Delta}) > q_{1-\alpha}^{\chi_1^2}$, avec $q_{1-\alpha}^{\chi_1^2}$ le quantile d'ordre $1-\alpha$ de la loi du Khi-Deux à 1 degré de liberté.

La p-value est donc donnée par

$$p\text{-value} = \mathbb{P}\left(Z(\mathbf{y}, \mathbf{d}) > \chi_1^2\right) = 1 - F_{\chi_1^2}(Z(\mathbf{y}, \mathbf{d})) \tag{17}$$

*Risque de ne pas rejeter $\mathcal{H}_0$ à tort

où $F_{\chi^2_1}$ est la fonction de répartition d'une loi du Khi-Deux à 1 degré de liberté.

COMPARAISON DE PLUSIEURS GROUPES :

Généralisons le test du *log-rank* pour comparer $g \geq 2$ groupes. Les hypothèses sont :

$$\mathcal{H}_0 : S^{(1)}(\cdot) = ... = S^{(g)}(\cdot) \qquad \text{contre} \qquad \mathcal{H}_1 : \exists (\ell, j) \in [\![1, g]\!], \ell \neq j \text{ tel que } S^{(\ell)}(\cdot) \neq S^{(j)}(\cdot)$$

Les individus des groupes sont supposés différents (*i.e.* groupes indépendants).

Sous les mêmes notations que dans le cas à 2 groupes, et en posant $\mathbf{K} = \left(K^{(1)}, ..., K^{(g-1)}\right)^\top$, avec pour $j \in [\![1, g-1]\!]$, un élément de $\mathbf{K}$ s'écrit

$$K^{(j)} := K^{(j)}(\mathbf{y}, \mathbf{d}) = \sum_{i=1}^{n} \left(O_i^{(j)} - N_i^{(j)} \frac{O_i}{N_i} \right) \text{ où } \left\{ \begin{array}{l} O_i = \sum_{j=1}^{g} O_i^{(j)} \\ N_i = \sum_{j=1}^{g} N_i^{(j)} \end{array} \right. \tag{18}$$

Intuition :

Pour toute matrice V symétrique définie positive, donc inversible, de dimension $(g-1 \times g-1)$, $\|\cdot\|_V : \mathbb{R}^{g-1} \to \mathbb{R}_+$ est la norme définie pour tout $x \in \mathbb{R}^{g-1}$ par $\|x\|_V = x^\top V x$. Or sous $\mathcal{H}_0$, le résultat attendu est $K^{(1)}(\mathbf{y}, \mathbf{d}) \approx ... \approx K^{(g-1)}(\mathbf{y}, \mathbf{d}) \approx 0$, donc $\|\mathbf{K}\|_V \approx 0$. Il faut donc trouver la bonne matrice V qui permettra de reconnaître la loi asymptotique de $\|\mathbf{K}\|_V$ sous $\mathcal{H}_0$.

En reprenant les notations précédentes, pour $j \in [\![1, g-1]\!]$, la valeur de l'estimateur de la variance asymptotique de $K^{(j)}(\mathbf{Y}, \mathbf{\Delta})$ est donnée par

$$\mathcal{V}_n^{(j)} := \sum_{i=1}^{n} \frac{N_i^{(j)} N_i^{(-j)}}{N_i} \frac{O_i}{N_i} \left(1 - \frac{O_i}{N_i} \right) \text{ où } N_i^{(-j)} = N_i - N_i^{(j)}$$

La valeur observée de l'estimateur de la covariance asymptotique entre les variables aléatoires $K^{(\ell)}(\mathbf{Y}, \mathbf{\Delta})$ et $K^{(j)}(\mathbf{Y}, \mathbf{\Delta})$, pour $\ell \neq j$, est donnée par

$$\mathcal{C}_n^{(\ell,j)} := -\sum_{i=1}^{n} \frac{N_i^{(\ell)} N_i^{(j)}}{N_i} \frac{O_i}{N_i} \left(1 - \frac{O_i}{N_i} \right)$$

La matrice de variance-covariance asymptotique de $\mathbf{K} = \left(K^{(1)}, ..., K^{(g-1)}\right)^\top$ est donnée par

$$\mathbf{\Sigma_K} = \begin{pmatrix} \mathcal{V}_n^{(1)} & \mathcal{C}_n^{(1,2)} & \cdots & \mathcal{C}_n^{(1,g-2)} & \mathcal{C}_n^{(1,g-1)} \\ \mathcal{C}_n^{(2,1)} & \mathcal{V}_n^{(2)} & \cdots & \mathcal{C}_n^{(2,g-2)} & \mathcal{C}_n^{(2,g-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{C}_n^{(g-2,1)} & \mathcal{C}_n^{(g-2,2)} & \cdots & \mathcal{V}_n^{(g-2)} & \mathcal{C}_n^{(g-2,g-1)} \\ \mathcal{C}_n^{(g-1,1)} & \mathcal{C}_n^{(g-1,2)} & \cdots & \mathcal{C}_n^{(g-1,g-2)} & \mathcal{V}_n^{(g-1)} \end{pmatrix}$$

La statistique de test $\mathbf{Z}(\mathbf{Y}, \mathbf{\Delta})$ est donc introduite et sa valeur observée est donnée par

$$\mathbf{Z}(\mathbf{y}, \mathbf{d}) := \mathbf{K}^\top \mathbf{\Sigma_K} \mathbf{K} = \begin{pmatrix} K^{(1)} & \cdots & K^{(g-1)} \end{pmatrix} \begin{pmatrix} \mathcal{V}_n^{(1)} & \cdots & \mathcal{C}_n^{(1,g-1)} \\ \vdots & \ddots & \vdots \\ \mathcal{C}_n^{(g-1,1)} & \cdots & \mathcal{V}_n^{(g-1)} \end{pmatrix} \begin{pmatrix} K^{(1)} \\ \vdots \\ K^{(g-1)} \end{pmatrix} \tag{19}$$

En supposant que les lois des couples $\left(T^{(1)}, C^{(1)}\right), ..., \left(T^{(g)}, C^{(g)}\right)$ sont indépendantes, et s'il existe $(c_1, ..., c_g) \in]0,1[^g$ tels que $c_1 + ... + c_g = 1$ et pour tout $j \in [\![1, g]\!]$, $\frac{n_j}{n} \xrightarrow[n\to+\infty]{} c_j$, alors sous $\mathcal{H}_0$, $\mathbf{Z}(\mathbf{Y}, \mathbf{\Delta})$ suit asymptotiquement une loi du Khi-Deux à $g-1$ degrés de liberté.

Par conséquent, l'hypothèse nulle est rejetée avec un risque d'erreur $\alpha \in]0,1[$ si $\mathbf{Z}(\mathbf{Y}, \mathbf{\Delta}) > q_{1-\alpha}^{\chi^2_{g-1}}$, avec $q_{1-\alpha}^{\chi^2_{g-1}}$ le quantile d'ordre $1-\alpha$ de la loi du Khi-Deux à $g-1$ degrés de liberté.
La p-value est donc donnée par

$$p\text{-value} = \mathbb{P}\left(\mathbf{Z}(\mathbf{y}, \mathbf{d}) > \chi^2_{g-1}\right) = 1 - F_{\chi^2_{g-1}}(\mathbf{Z}(\mathbf{y}, \mathbf{d})) \tag{20}$$

où $F_{\chi^2_{g-1}}$ est la fonction de répartition d'une loi du Khi-Deux à $g-1$ degrés de liberté.

Test du rapport de vraisemblances

Le test du rapport de vraisemblance (*likelihood ratio test*) est une méthode statistique utilisée pour comparer la qualité d'ajustement de deux modèles imbriqués, c'est-à-dire un modèle restreint (ou réduit) et un modèle plus général (ou complet). Voici comment se formalise ce test.

Notons $\beta = (\beta_1, ..., \beta_p)$ le vecteur des paramètres à estimer. Il faut tester

$$\mathcal{H}_0 : \beta \in \mathcal{B}_0 \qquad \text{contre} \qquad \mathcal{H}_1 : \beta \notin \mathcal{B}_0$$

Pour cela, la statistique de test se calcule avec la formule suivante

$$R = -2 \log \left(\frac{\mathcal{L}(\widehat{\beta}_0)}{\mathcal{L}(\widehat{\beta})} \right) \tag{21}$$

où $\widehat{\beta} \in \mathbb{R}^p$ est l'estimateur du maximum de vraisemblance, $\widehat{\beta}_0 \in \mathcal{B}_0 \subset \mathbb{R}^p$ (où $r \leq p$ est le nombre de paramètres contraints dans $\mathcal{B}_0$) est l'estimateur du maximum de vraisemblance sous $\mathcal{H}_0$ et $\mathcal{L}$ la fonction de vraisemblance*.

Celle-ci suit asymptotiquement une loi du Khi-Deux à r degrés de liberté. Par conséquent, l'hypothèse nulle est rejetée avec un risque d'erreur $\alpha \in]0,1[$ si $R > q_{1-\alpha}^{\chi^2_r}$, avec $q_{1-\alpha}^{\chi^2_r}$ le quantile d'ordre $1-\alpha$ de la loi du Khi-Deux à r degrés de liberté.
La p-value est donc donnée par

$$p\text{-value} = \mathbb{P}\left(R > \chi^2_r\right) = 1 - F_{\chi^2_r}(R) \tag{22}$$

où $F_{\chi^2_r}$ est la fonction de répartition d'une loi du Khi-Deux à r degrés de liberté.

*De façon équivalente, la statistique peut s'écrire par différence des log-vraisemblances : $-2\left(\ell(\widehat{\beta}_0) - \ell(\widehat{\beta})\right)$

A.3 Résultat du test de Wald du modèle à risques proportionnels de Cox

Les résultats pour chaque covariable du modèle à risques proportionnels de Cox sont disponibles dans le tableau 7 ci-dessous. Pour le résumer, presque toutes les variables sont très significatives puisqu'elles ont une p-value inférieures à 0,05. Seules la catégorie socio-professionnelle "Profession libérale" et le niveau de franchise bris de glace "Sans objet" ne respectent pas seuil de 5% mais ces covariables restent quand même assez significative avec des p-values respectivement égales à 0,138 et 0,062. On peut donc en conclure que toutes ces variables semblent être importantes pour le modèle.

Covariables	$\beta_{\mathbf{k}}$	$\mathbb{V}\mathrm{ar}(\beta_{\mathbf{k}})$	$\mathbf{W_k}$	p-**value**	**Significativité**
COEF_BONUS_MALUS	0,2	$3{,}32\times 10^{-4}$	118,67	$< 0,05$	⋆⋆⋆
AGE_COND	-0,0	$4{,}87\times 10^{-8}$	32,44	$< 0,05$	⋆⋆⋆
ANCIENNETE_VEHICULE	0,02	$1{,}37\times 10^{-7}$	2014,45	$< 0,05$	⋆⋆⋆
PRIME_COMM	0,0	$9{,}51\times 10^{-11}$	4263,09	$< 0,05$	⋆⋆⋆
NBR_CTR_DIFF	0,07	$1{,}79\times 10^{-5}$	280,14	$< 0,05$	⋆⋆⋆
NBR_GARANTIES	-0,14	$3{,}52\times 10^{-5}$	526,69	$< 0,05$	⋆⋆⋆
FRAC_PAIEMENT_Annuel	-0,14	$2{,}16\times 10^{-5}$	944,58	$< 0,05$	⋆⋆⋆
RESIL_PRECEDENT_ASS	0,36	$1{,}47\times 10^{-4}$	861,91	$< 0,05$	⋆⋆⋆
CSP_Agriculteur	-0,09	$1{,}14\times 10^{-4}$	68,93	$< 0,05$	⋆⋆⋆
CSP_Artisan - Commerçant	0,12	$6{,}82\times 10^{-5}$	204,02	$< 0,05$	⋆⋆⋆
CSP_Etudiant	-0,06	$1{,}02\times 10^{-4}$	31,09	$< 0,05$	⋆⋆⋆
CSP_Fonctionnaire ou assimilé	-0,09	$5{,}62\times 10^{-5}$	144,31	$< 0,05$	⋆⋆⋆
CSP_Profession libérale	-0,02	$2{,}20\times 10^{-4}$	2,2	0,138	⋆
CSP_Retraité	-0,1	$5{,}07\times 10^{-5}$	189,63	$< 0,05$	⋆⋆⋆
CSP_Sans profession	0,09	$1{,}63\times 10^{-4}$	49,44	$< 0,05$	⋆⋆⋆
RESEAU_DIST_Autres	0,3	$2{,}17\times 10^{-4}$	425,22	$< 0,05$	⋆⋆⋆
RESEAU_DIST_Courtiers	0,16	$2{,}07\times 10^{-5}$	1293,86	$< 0,05$	⋆⋆⋆
OPT_PK_MOB	-0,04	$8{,}48\times 10^{-5}$	15,51	$< 0,05$	⋆⋆⋆
OPT_PK_INDEM	0,02	$5{,}99\times 10^{-5}$	6,65	$< 0,05$	⋆⋆⋆
OPT_PK_CONT	-0,16	$2{,}93\times 10^{-5}$	925,15	$< 0,05$	⋆⋆⋆
OPT_SCD_CTR_AUTO	-0,45	$4{,}89\times 10^{-5}$	4359,3	$< 0,05$	⋆⋆⋆
OPT_CTR_HAB	-0,65	$6{,}64\times 10^{-5}$	6421,2	$< 0,05$	⋆⋆⋆
TYPE_HAB_App-Loc	0,73	$6{,}25\times 10^{-5}$	8628,93	$< 0,05$	⋆⋆⋆
TYPE_HAB_App-NR	0,73	$1{,}26\times 10^{-3}$	426,49	$< 0,05$	⋆⋆⋆
TYPE_HAB_App-Prop	0,66	$1{,}95\times 10^{-4}$	2221,96	$< 0,05$	⋆⋆⋆
TYPE_HAB_Autre-Loc	0,77	$2{,}04\times 10^{-2}$	29,14	$< 0,05$	⋆⋆⋆
TYPE_HAB_Autre-Prop	0,84	$5{,}41\times 10^{-3}$	129,3	$< 0,05$	⋆⋆⋆
TYPE_HAB_Mais-Loc	0,75	$7{,}46\times 10^{-5}$	7512,99	$< 0,05$	⋆⋆⋆
TYPE_HAB_Mais-NR	0,69	$2{,}01\times 10^{-4}$	2396,82	$< 0,05$	⋆⋆⋆
TYPE_HAB_Mais-Prop	0,7	$4{,}02\times 10^{-5}$	12209,97	$< 0,05$	⋆⋆⋆
TYPE_COND_SCD_Enfant	-0,12	$5{,}85\times 10^{-5}$	259,3	$< 0,05$	⋆⋆⋆
TYPE_COND_SCD_Conjoint	0,06	$4{,}66\times 10^{-5}$	68,41	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_Sans Franchise	-0,07	$6{,}17\times 10^{-5}$	74,95	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_Faible	-0,04	$7{,}21\times 10^{-5}$	26,22	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_Elevé	0,03	$2{,}81\times 10^{-5}$	40,2	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_Doublé	0,17	$1{,}14\times 10^{-4}$	253,93	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_NR	-0,1	$2{,}09\times 10^{-3}$	4,96	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_DMG_Sans objet	-0,32	$5{,}18\times 10^{-4}$	195,52	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_BDG_Standard	0,08	$2{,}01\times 10^{-5}$	300,64	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_BDG_Elevé	0,09	$2{,}19\times 10^{-4}$	39,69	$< 0,05$	⋆⋆⋆
NIVEAU_FRANCHISE_BDG_Sans objet	-0,02	$1{,}19\times 10^{-4}$	3,48	0,062	⋆⋆

TABLE 7 : Résultats du test de Wald pour chaque variable du modèle à risques proportionnels de Cox

A.4 Indices d'agrégation classiques pour la C.A.H.

Voici quelques exemples d'indices d'agrégation classiques utilisés dans la classification ascendante hiérarchique :

- ▷ Indice du lien simple (saut minimal) : $\delta_{\min}(A,B) = \min_{x\in A, y\in B} d(x,y)$
- ▷ Indice du lien complet (saut maximal) : $\delta_{\max}(A,B) = \min_{x\in A, y\in B} d(x,y)$
- ▷ Indice du lien moyen : $\delta_{\text{moyen}}(A,B) = \dfrac{1}{n_A n_B} \displaystyle\sum_{x\in A, y\in B} d(x,y)$ où $n_A = \text{Card}(A)$ et $n_B = \text{Card}(B)$

où $d(x,y)$ est la dissimilarité entre $x \in A$ et $y \in B$, correspondant à une distance entre x et y. Voici des exemples de distances usuelles :

- ▷ Distance euclidienne : $d(x,y) = \sqrt{\displaystyle\sum_{i=1}^{d} (x_i - y_i)^2}$
- ▷ Distance de Manhattan : $d(x,y) = \displaystyle\sum_{i=1}^{d} |x_i - y_i|$
- ▷ Distance de Sebestyen : $d(x,y) = \sqrt{\displaystyle\sum_{i=1}^{d} w_i (x_i - y_i)^2}$ où les w_i sont des poids.
- ▷ Distance de Tchebychev : $d(x,y) = \max_{i\in[\![1,d]\!]} |x_i - y_i|$

A.5 Exemple de C.A.H. avec l'indice d'aggrégation de Ward

Rappelons la formule de l'indice d'aggrégation de Ward

$$\delta_{\text{Ward}}(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B) \tag{23}$$

où n_A et n_B sont les effectifs des classes A et B, g_A et g_B sont les centres de gravité de A et B, et d est la distance euclidienne.

Nota Bene :

> *Afin de simplifier les notations, la notation condensée sans accolades pour écrire un ensemble est adopté. Par exemple, ABC est écrit au lieu de $\{A, B, C\}$ et A au lieu de $\{A\}$.*

	A	B	C	D	E	F
x	-2	-2	-2	2	2	1
y	3	1	-1	-1	1	0

TABLE 8 : Coordonnées des points de $\mathcal{P}$

Ces points peuvent être représentés dans le plan :

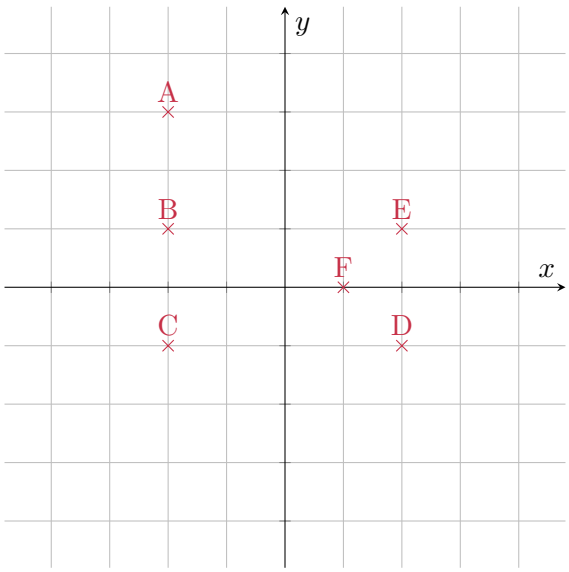


FIGURE 11 : Représentation graphique des points de $\mathcal{P}$

▷ **Initialisation :** On dispose de 6 classes formées par chaque point de $\mathcal{P}$.

▷ **Itération 1 :** On calcule la matrice de Ward $\mathbb{W}$ selon la formule définie à l'équation (23) :

	A	B	C	D	E	F
A	0	2	8	16	10	9
B		0	2	16	8	9
C			0	8	10	5
D				0	2	**1**
E					0	**1**
F						0

Table 9 : Matrice de Ward à l'étape 1

La plus petite valeur de l'indice est égale 1 pour la réunion de D et F, et aussi pour la réunion de E et F. Par choix arbitraire, on choisit de réunir E et F, ce qui nous donne la classe EF ayant pour centre de gravité $\left(\frac{3}{2}, \frac{1}{2}\right)^\top$.

▷ **Itération 2 :** On recalcule la matrice de Ward $\mathbb{W}$ selon la formule définie à l'équation (23) :

	A	B	C	D	EF
A	0	2	8	16	$\frac{37}{3}$
B		0	2	16	$\frac{25}{3}$
C			0	8	$\frac{29}{3}$
D				0	$\mathbf{\frac{5}{3}}$
EF					0

Table 10 : Matrice de Ward à l'étape 2

La plus petite valeur de l'indice est égale $\frac{5}{3}$ pour la réunion de EF et D, ce qui nous donne la classe DEF ayant pour centre de gravité $\left(\frac{5}{3}, 0\right)^\top$.

▷ **Itération 3 :** On recalcule la matrice de Ward $\mathbb{W}$ selon la formule définie à l'équation (23) :

	A	B	C	DEF
A	0	**2**	8	$\frac{101}{6}$
B		0	**2**	$\frac{65}{6}$
C			0	$\frac{65}{6}$
DEF				0

Table 11 : Matrice de Ward à l'étape 3

La plus petite valeur de l'indice est égale 2 pour la réunion de A et B, et aussi pour la réunion de B et C. Par choix arbitraire, on choisit de réunir A et B, ce qui nous donne la classe AB ayant pour centre de gravité $(-2, 2)^\top$.

▷ **Itération 4 :** On recalcule la matrice de Ward $\mathbb{W}$ selon la formule définie à l'équation (23) :

	AB	C	DEF
AB	0	**6**	$\frac{314}{15}$
C		0	$\frac{65}{6}$
DEF			0

TABLE 12 : Matrice de Ward à l'étape 4

La plus petite valeur de l'indice est égale 6 pour la réunion de AB et C, ce qui nous donne la classe ABC ayant pour centre de gravité $(-2, 1)^\top$.

▷ **Itération 5 :** On calcule une dernière fois la matrice de Ward $\mathbb{W}$ selon la formule définie à l'équation (23) (même s'il est clair qu'on va réunir les deux classes restantes) :

	ABC	DEF
ABC	0	$\frac{65}{3}$
DEF		0

TABLE 13 : Matrice de Ward à l'étape 5

On réunit les deux dernières classes ABC et DEF pour obtenir $\mathcal{P}$.

▷ **Conclusion :** On peut ensuite représenter la classification sous forme de dendrogramme :

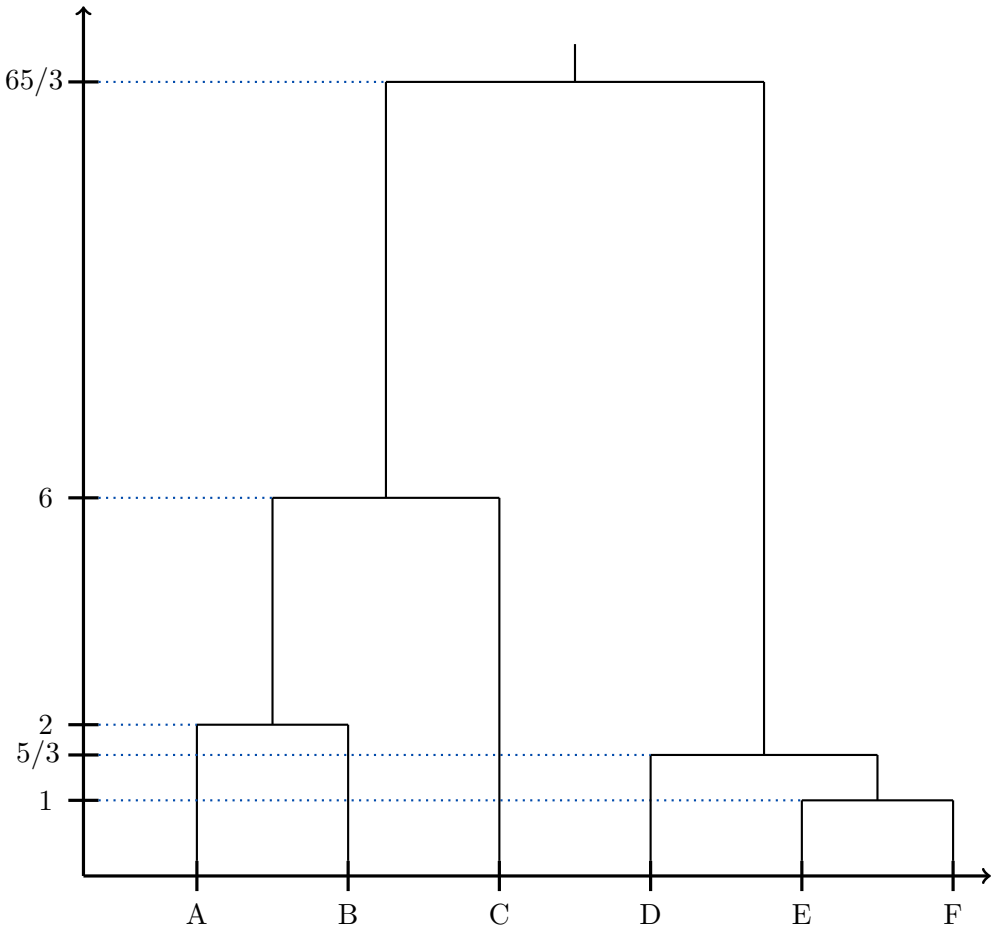


FIGURE 12 : Dendrogramme issu de la C.A.H. précédente

A.6 Classes de durée de vie des contrats selon la CSP

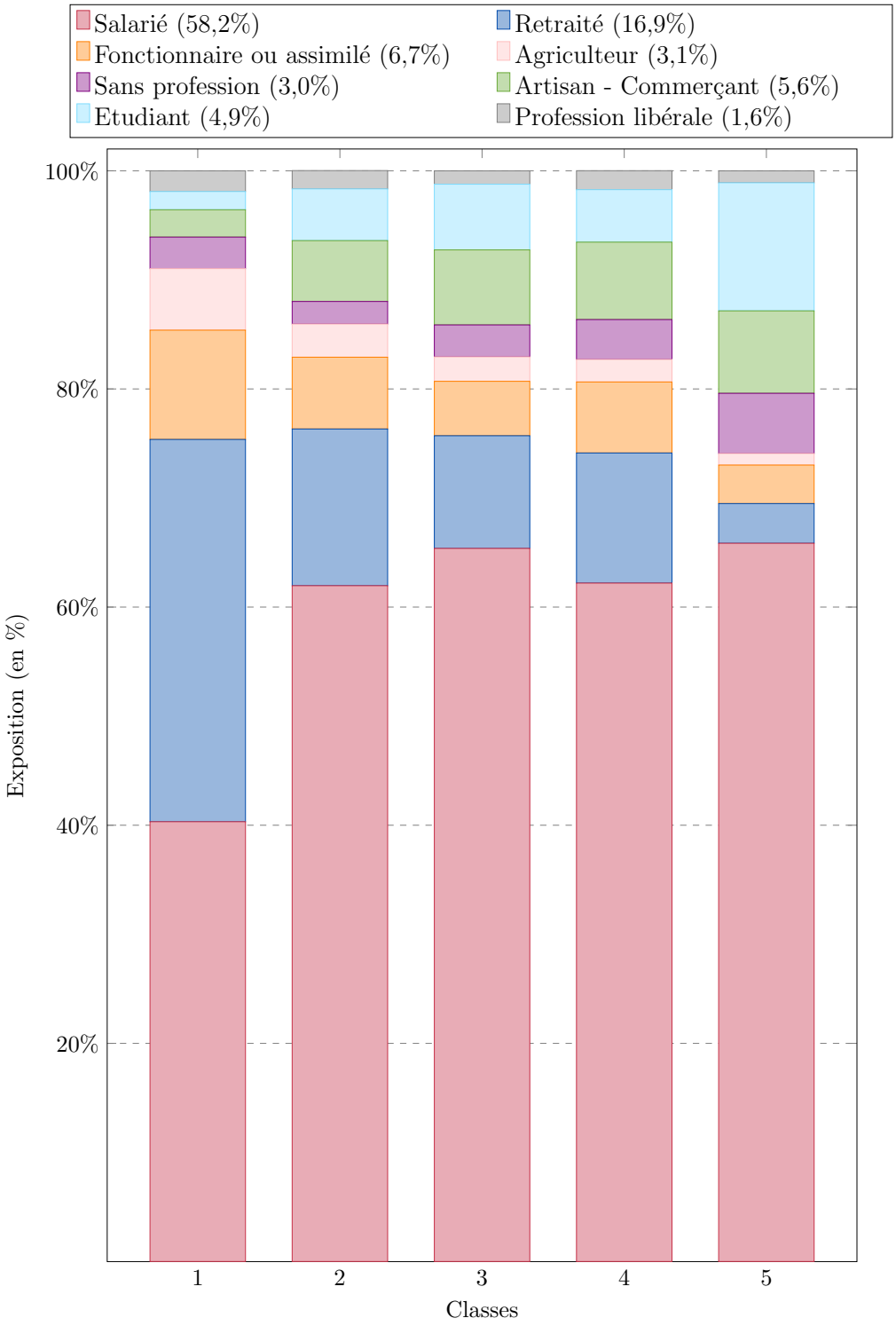
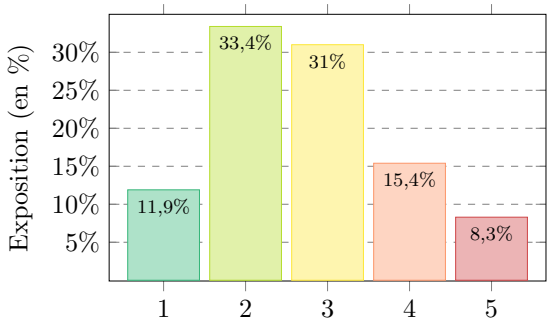
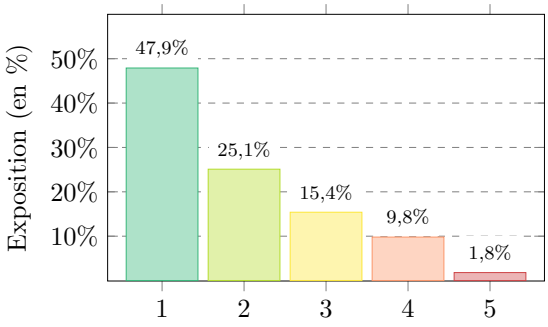


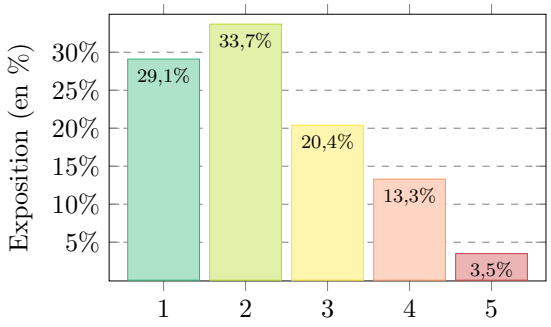
Figure 13 : Répartition des catégories socio-professionnelles selon les classes de durée de vie des contrats



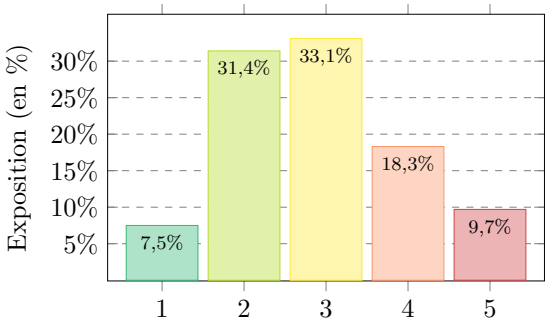
(a) Pour les salariés (58,2%)



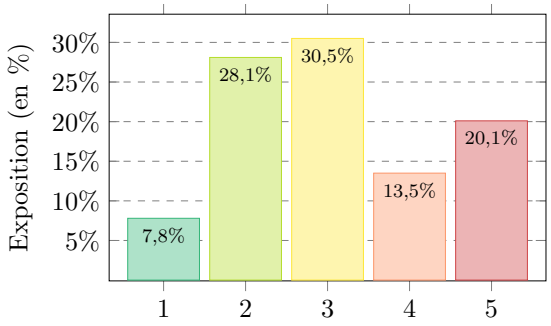
(b) Pour les retraités (16,9%)



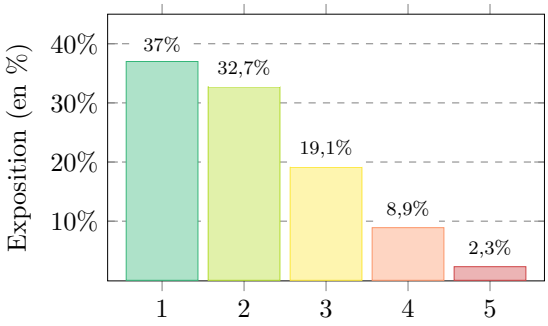
(c) Pour les fonctionnaires ou assimilés (6,7%)



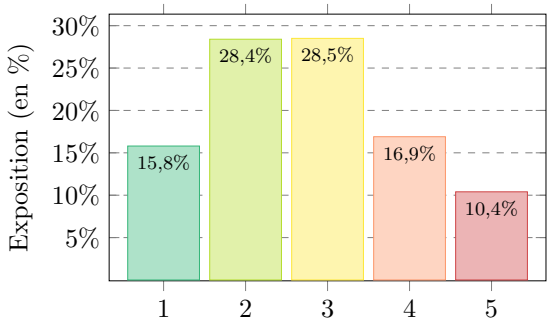
(d) Pour les artisans et commerçants (5,6%)



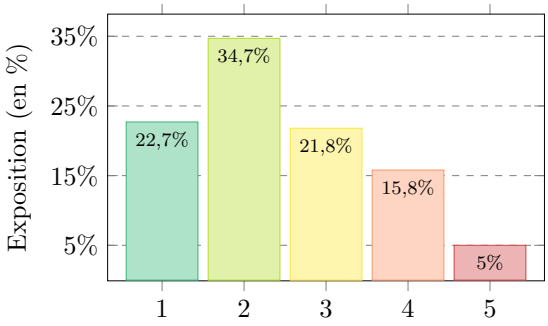
(e) Pour les étudiants (4,9%)



(f) Pour les agriculteurs (3,1%)



(g) Pour les personnes sans profession (3,0%)



(h) Pour les personnes ayant une profession libérale (1,6%)

FIGURE 14 : Répartition des catégories socio-professionnelles dans les classes de durée de vie

A.7 Classes de transformation selon la CSP

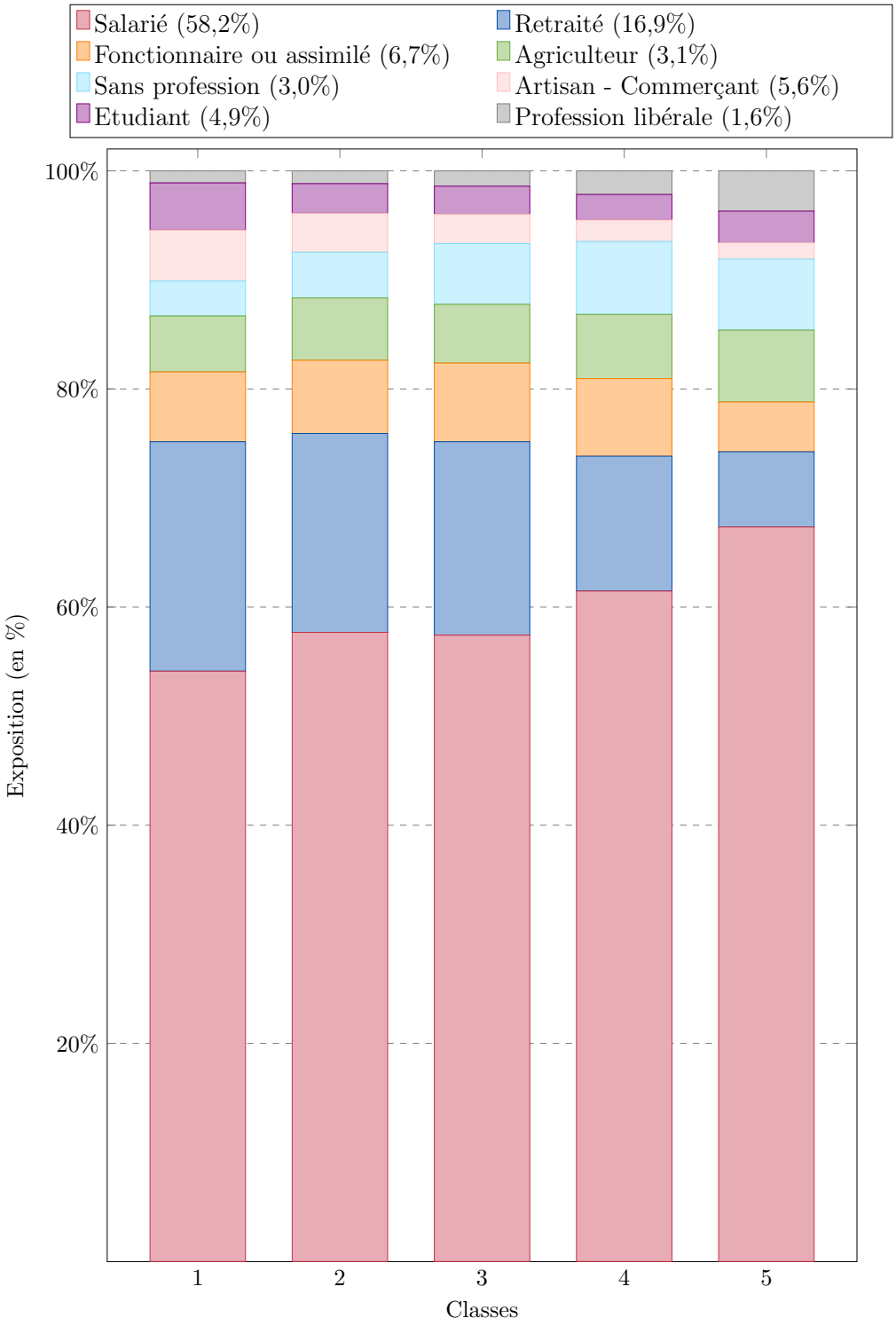
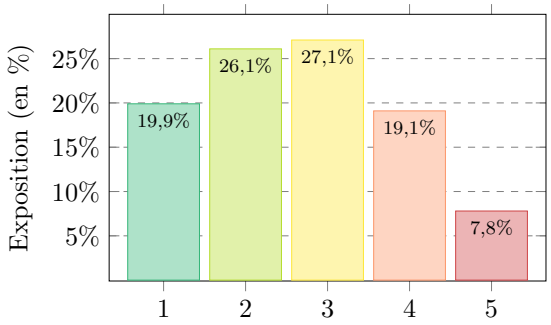
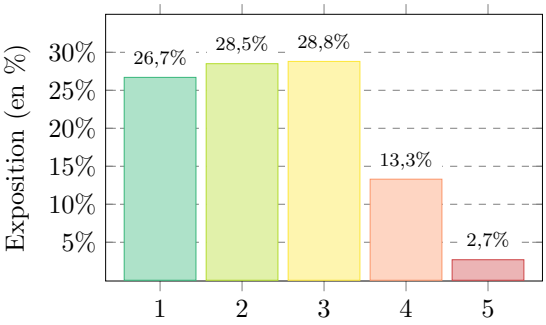


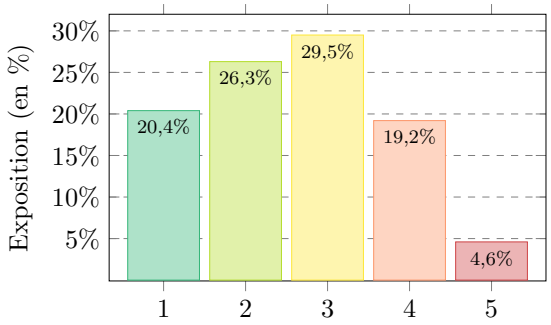
Figure 15 : Répartition des catégories socio-professionnelles selon les classes de transformation



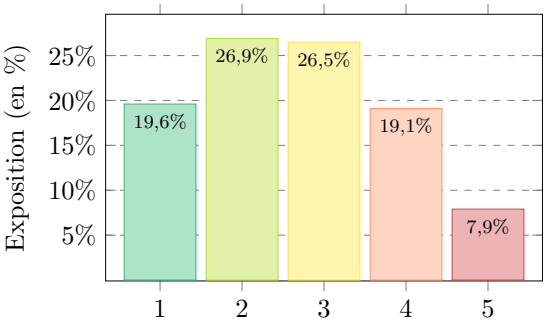
(a) Pour les salariés (58,2%)



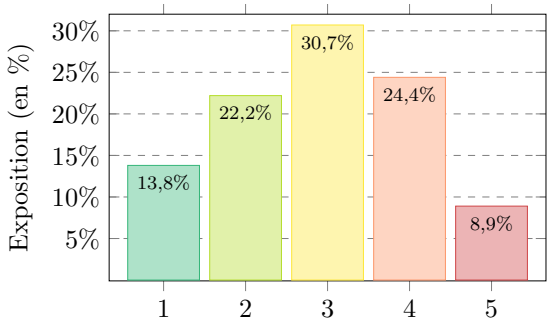
(b) Pour les retraités (16,9%)



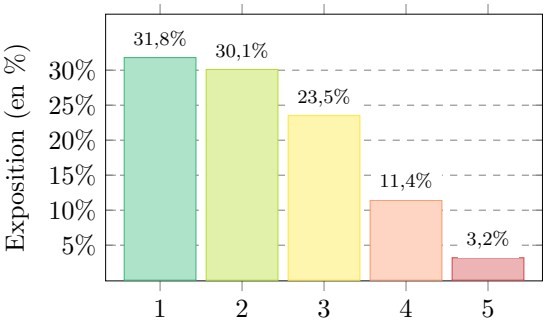
(c) Pour les fonctionnaires ou assimilés (6,7%)



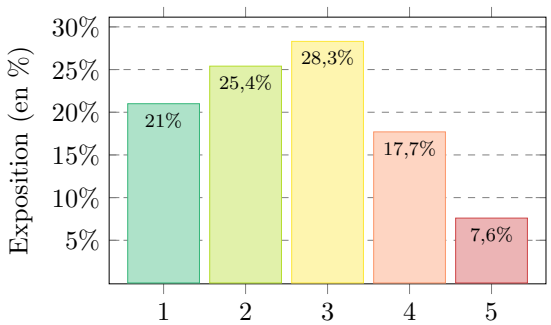
(d) Pour les artisans et commerçants (5,6%)



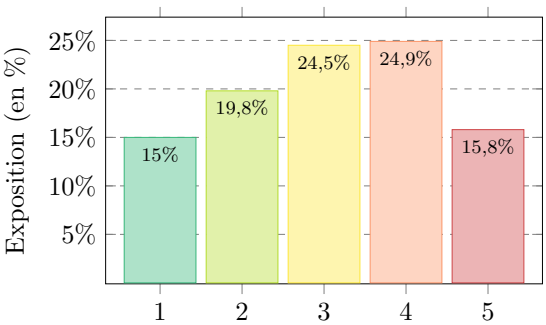
(e) Pour les étudiants (4,9%)



(f) Pour les agriculteurs (3,1%)



(g) Pour les personnes sans profession (3,0%)



(h) Pour les personnes ayant une profession libérale (1,6%)

FIGURE 16 : Répartition des catégories socio-professionnelles dans les classes de transformation

A.8 Clusters selon la catégorie socio-professionnelle

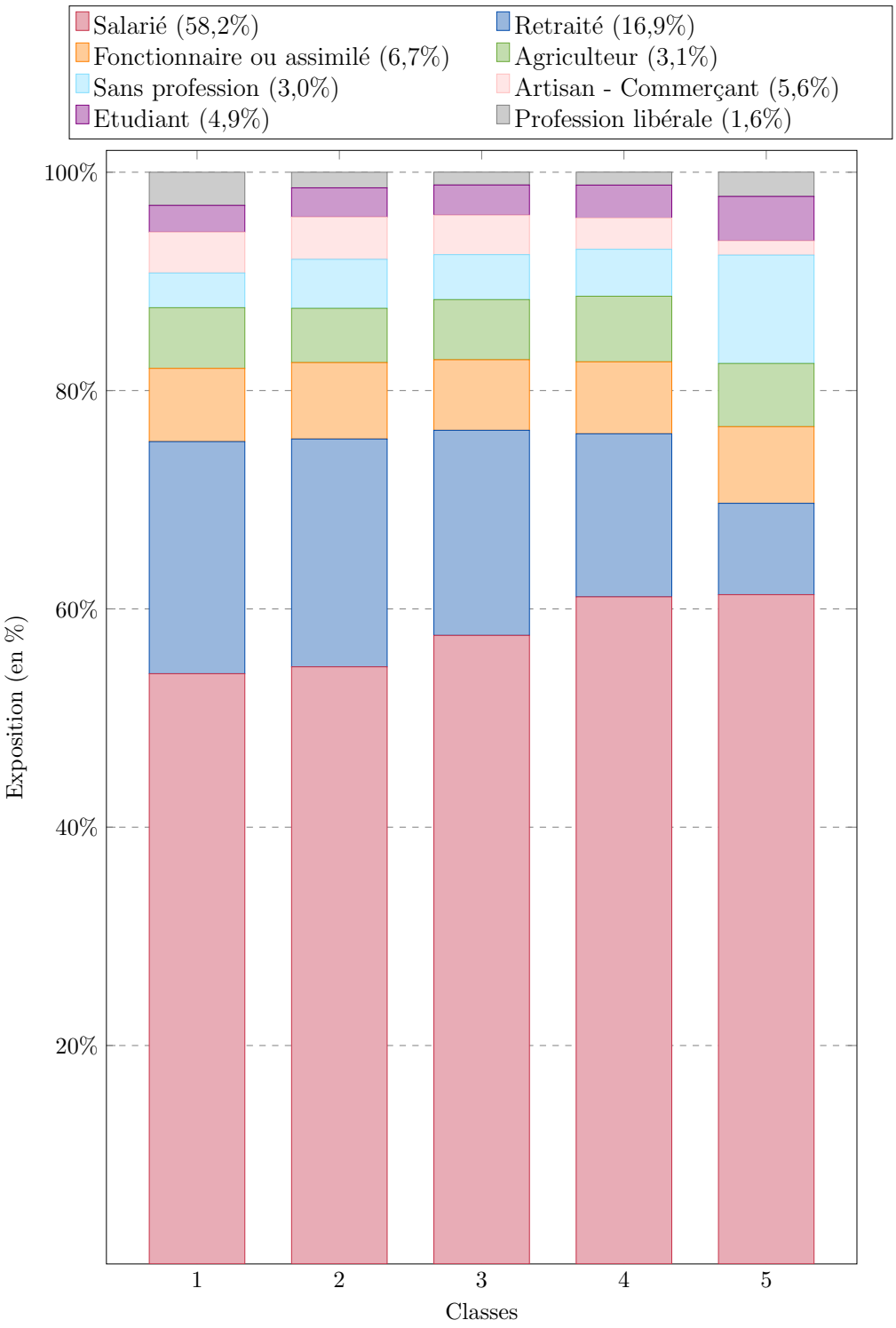
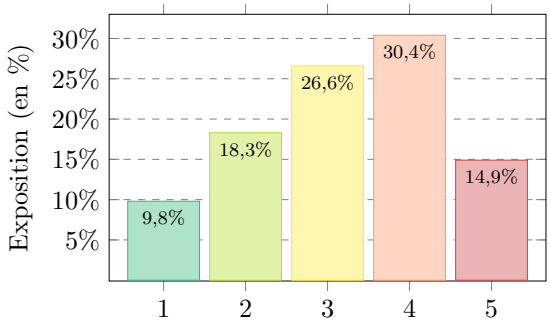
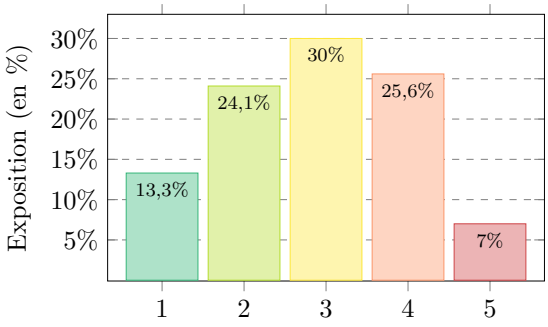


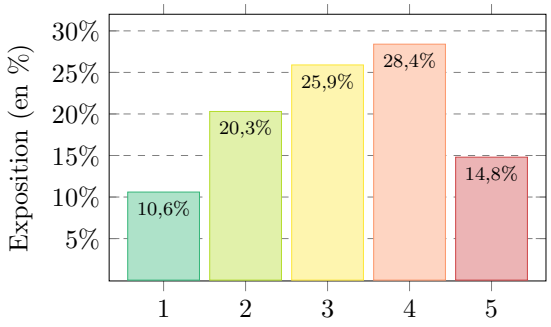
Figure 17 : Répartition des catégories socio-professionnelles selon les clusters



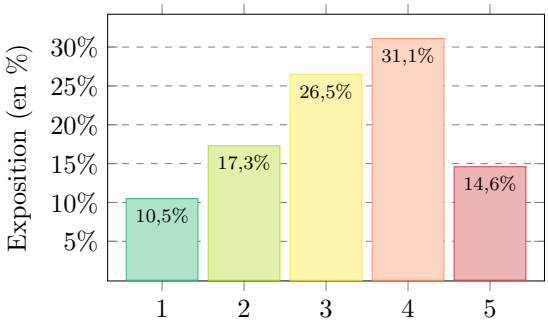
(a) Pour les salariés (58,2%)



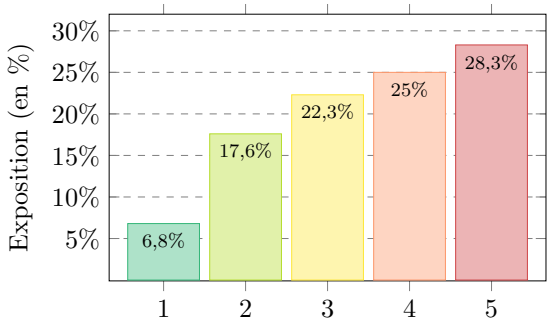
(b) Pour les retraités (16,9%)



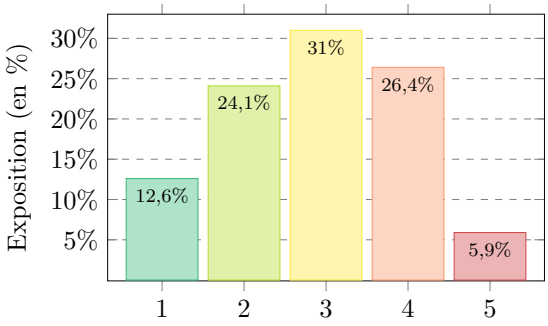
(c) Pour les fonctionnaires ou assimilés (6,7%)



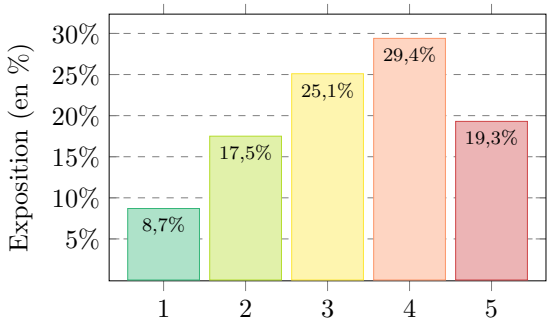
(d) Pour les artisans et commerçants (5,6%)



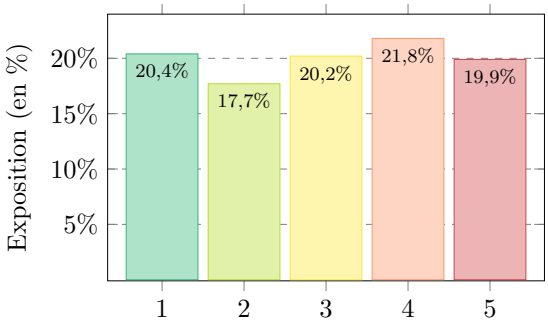
(e) Pour les étudiants (4,9%)



(f) Pour les agriculteurs (3,1%)



(g) Pour les personnes sans profession (3,0%)



(h) Pour les personnes ayant une profession libérale (1,6%)

FIGURE 18 : Répartition des catégories socio-professionnelles dans les clusters

A.9 Résultats de l'optimisation sans contrainte pour chaque cluster

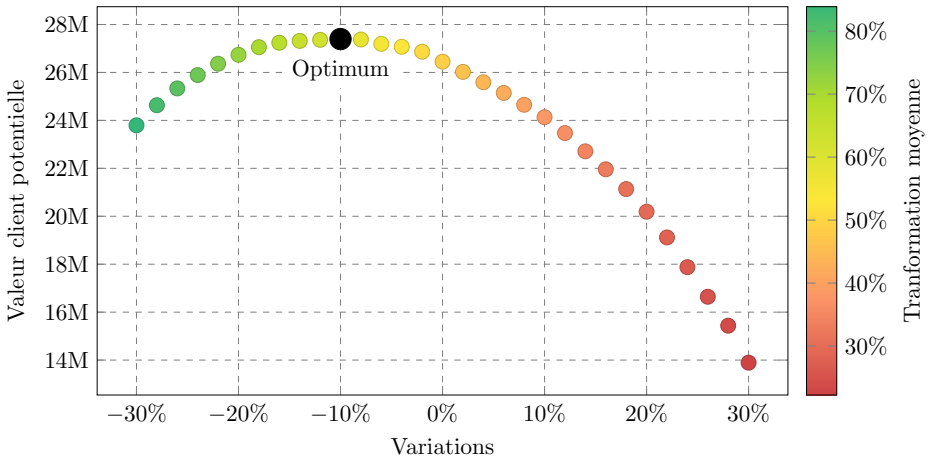


FIGURE 19 : Résultats de l'optimisation sans contrainte du cluster 1

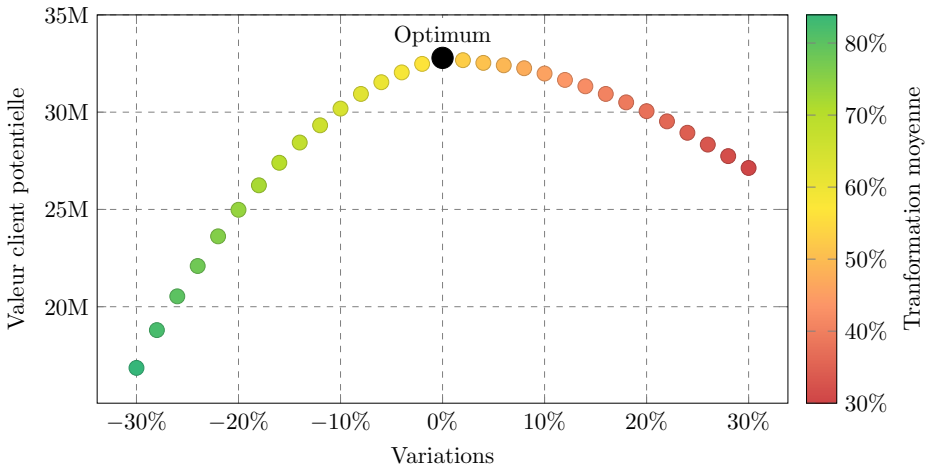


FIGURE 20 : Résultats de l'optimisation sans contrainte du cluster 2

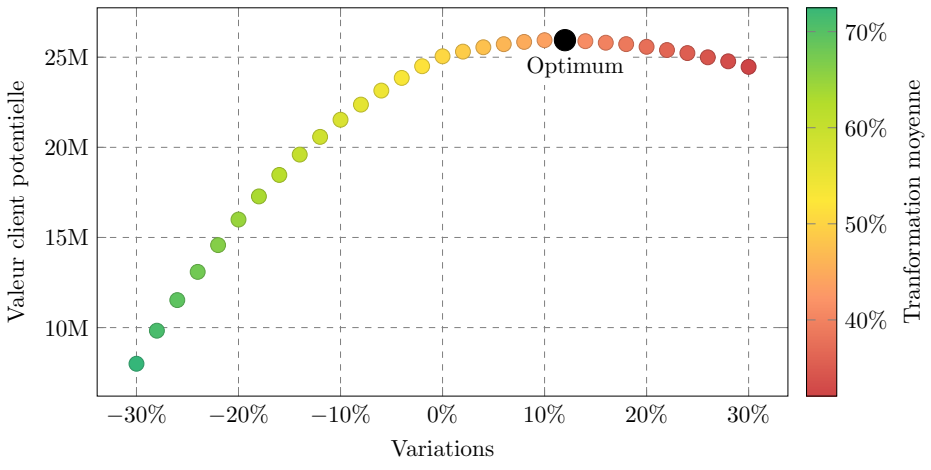


FIGURE 21 : Résultats de l'optimisation sans contrainte du cluster 3

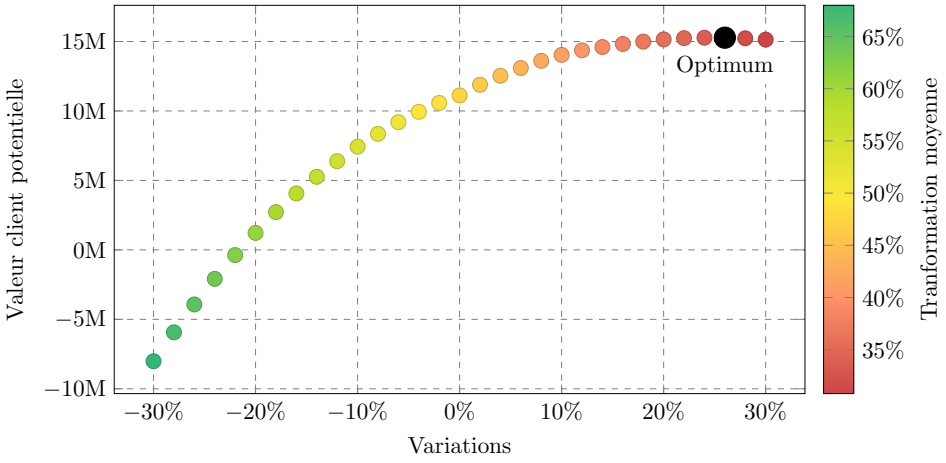


FIGURE 22 : Résultats de l'optimisation sans contrainte du cluster 4

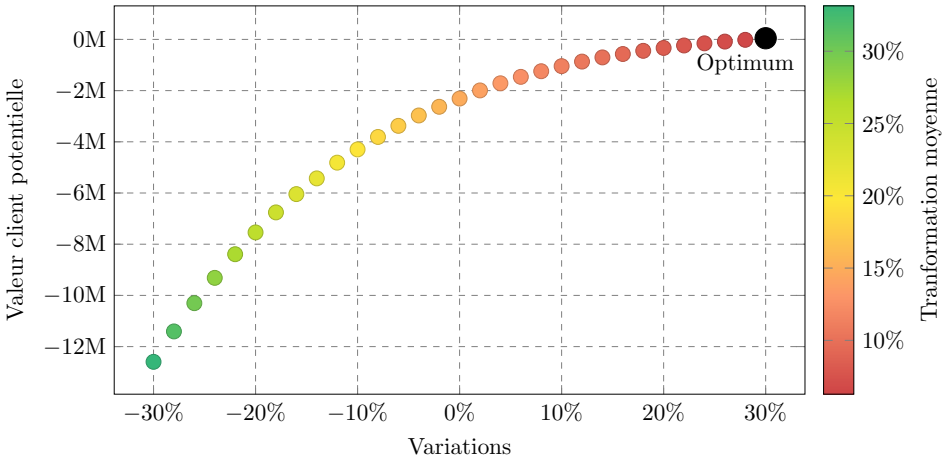


FIGURE 23 : Résultats de l'optimisation sans contrainte du cluster 5