IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Cécile HUBERT

Titre : Back-testing d'une norme tarifaire santé sur-mesure suite à la mise en place de l'OPTAM/OPTAM-CO et évaluation de sa fiabilité au moyen de techniques actuarielles modernes.

Confidentialité : ☐ NON ☑ OUI (Durée : ☐ 1an ☑ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : Swiss Life Prévoyance et Santé

Nom : _____

Signature et Cachet :

Directeur de mémoire en entreprise :

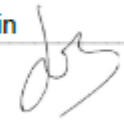
Nom : Jeannie Doukhan

Signature :

Invité :

Nom : Bastien Seguin

Signature :



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

Signature du responsable entreprise

Signature(s) du candidat(s)

SAS au capital de 96 550 euros - Déclaration d'activité enregistrée sous le n° 11 75 44451 75 auprès du Préfet de Région IDF
4, rue Chauveau-Lagarde - 75008 Paris - Tél : 01 44 51 72 79 - Fax : 01 44 51 72 73
Email : formation@institutdesactuaires.fr
Siret : 512 264 441 000 19 - Code NAF : 8559 A - TVA intracommunautaire : FR 25 512264441 - RCS Paris : 512 264 441

Résumé

En assurance, la tarification au plus juste est une condition essentielle pour la maîtrise du risque. Les outils de tarification, aujourd'hui très informatisés, avec une structure bien définie et un coût de développement suivi, occupent donc une place centrale dans le processus de commercialisation des contrats.

Chez Swiss Life, l'outil de tarification des contrats santé sur-mesure est un logiciel qui permet le paramétrage d'une prime pure pour un « assuré type » d'une part, et des coefficients multiplicatifs généraux ou par grands postes permettant un tarif spécifique à la population et les garanties étudiées d'autre part.

Lorsqu'on met à jour le tarif d'une garantie, il n'est pas rare de limiter son étude à la valeur de la prime pure, les coefficients étant contraints d'être communs à un groupe plus large de garanties, mais aussi pour maîtriser les coûts de développement.

Ce fut le cas lors de notre back-testing des garanties santé concernées par l'Option Pratique Tarifaire Maîtrisée (OPTAM/OPTAM-CO).

Lors de cette étude, nous nous sommes interrogés : notre modèle de tarification est-il toujours fiable, depuis sa création en 2014, et ses mises à jour, parfois partielles, progressives ?

Pour y répondre, nous avons décidé de tester la tarification de la garantie Spécialiste Non OPTAM, avec deux techniques modernes couramment citées et utilisées en actuariat : un Modèle Linéaire Généralisé et un modèle Machine Learning.

Les principes de ces deux techniques sont expliqués, puis nous analysons leurs résultats, que nous comparons à notre modèle interne. Nous concluons sur la fiabilité de notre modèle, ses points forts, ses points faibles, et les axes d'études à envisager dans le futur.

Mots clés : assurance santé, tarification, prime pure, GLM, apprentissage supervisé, machine learning, frequence x cout moyen

Abstract

In insurance, exact pricing is an essential condition to get risk under control. Nowadays, pricing tools are mainly computerized, with a precise structure and a followed-up development cost, so they are one of the main stage in the contracts mechandizing workflow.

In Swiss Life, the pricing tool for taylor made health contracts is a software which enable the set-up of a pure premium cost for a standard policyholder for one part, and for the other part of general or health major category multiplying coefficients, which enable the calculation of a specific price for the population and cover in question.

When a cover price is updated in the company, it happens frequently that the calculation is limited to the pure premium price, because these coefficients have to be set up for a major health category, and also to reduce development costs.

It happened during the update of health guarantees related to new controlled pricing practice option rules.

During this study, we wondered if our internal calculation model was still reliable, since its creation in 2014, and through its successive and sometimes partial updates ?

To answer the question, we have tested a specialist doctor cover pricing with two modernal methods frequently used and quoted in actuarial science : Generalized Linear Model and Machine Learning.

We explain these methods aims, then we analyse their results, that are compared to our internal model. At last we conclude on the liability of our model, its strenghs and weaknesses, and the line of thought to consider for our future analysis.

Key words: health insurance, pure premium cost, generalized linear model, machine learning, frequency x cost

Remerciements

Je tiens à adresser mes remerciements à Mme Jeannie Doukhan, qui m'a accordée sa confiance et permis d'accéder à la formation du Centre d'Etudes Actuarielles, et qui m'a ensuite toujours soutenu avec beaucoup de bienveillance jusqu'à la rédaction de ce mémoire.

Je remercie également mon manager et ami Bastien Seguin, pour son accompagnement et ses conseils toujours pertinents. Collaborer avec une telle personne est une source d'enrichissement permanente.

De nombreux collègues, de travail ou non m'ont également aidé par leur aide, leurs conseils, et leur écoute: je pense à Chahir, David, Aurélien, Soulaiman, Josh, Béatrice, Daria et Thibault. Je sais que j'en oublie. Merci à vous tous, je vous dois beaucoup.

Ce mémoire est l'aboutissement de nombreux mois de travail, et je suis fière de le présenter car j'ai essayé d'y inclure les différentes facettes opérationnelles, scolaires, mais aussi des recherches personnelles sur le sujet étudié.

Enfin, un remerciement tout particulier revient ma chère famille et mes précieux amis, qui m'ont soutenu sans relâche durant cette formation passionnante mais très exigeante. C'est grâce à eux que j'ai pu aller au bout de cette formation et de ce mémoire.

Table des matières

Introduction Générale

Dans tous les secteurs d'assurance, la tarification au plus juste des garanties est devenue essentielle afin de maîtriser les risques et les coûts des différents processus qui en découlent (régulation, résiliation, chiffre d'affaires généré, ratios…). L'assurance santé ne fait pas exception à ce constat, avec des évolutions réglementaires qui rendent le marché de plus en plus compétitif et limitent de plus en plus les possibilités de se différencier des autres assureurs.

C'est dans ce cadre que chez Swiss Life, au sein de la Direction des Assurances Collectives (DAC), les contrats sur-mesure collectifs santé sont tarifés à l'aide d'une norme tarifaire interne, créée en 2014.
En s'appuyant sur une liste de variables tarifaires précise, le modèle tarifaire mis en place consiste à calculer une prime pure définie pour un assuré qui aura des caractéristiques « types » précisément définies, puis de la personnaliser à tout assuré en la multipliant par des coefficients représentant les écarts de caractéristiques par rapport à l'assuré type. Les valeurs des primes pures et des coefficients sont ensuite stockées sous forme de grilles de données dans un outil de tarification créé exclusivement pour Swiss Life par un prestataire externe, selon un format standardisé pour toutes les garanties.
Ce modèle de tarification est surveillé annuellement au moyen d'un back-testing sur le portefeuille global, en regardant les résultats par contrat. Mais cette surveillance ne donne pas de détail sur la fiabilité et la robustesse de nos normes au sein d'une garantie précise.

C'est dans le cadre de la mise à jour des normes santé relatives aux soins OPTAM (Option Pratique Tarifaire Maîtrisée) / OPTAM-CO (Option Pratique Tarifaire Maîtrisée applicable aux spécialistes en Chirurgie ou en gynécologie Obstétrique), que nous nous sommes posés la question suivante : notre modèle tarifaire interne est-il fiable au sein d'une garantie santé isolée ?

Pour répondre à cette question, nous avons décidé de comparer notre nouveau tarif obtenu avec notre modèle interne avec un tarif obtenu via deux techniques de tarification actuarielles considérées comme des références en 2020 : un Modèle Linéaire Généralisé (GLM) et un modèle Machine Learning (ML). Le GLM est une méthode de tarification couramment utilisée en assurance santé notamment car elle permet de tenir compte des effets de corrélations entre les différentes variables explicatives.
Mais depuis quelques années les méthodes de tarification en Machine Learning se généralisent de plus en plus, car elles permettent de traiter un plus grand nombre de données, tant sur la profondeur que sur le nombre de variables explicatives, et de rendre accessible de nouveaux modèles mathématiques pour la prédiction de sinistres. Nous avons donc décidé d'utiliser également cette nouvelle méthode alternative, pour la modernité et le côté pratique qu'il semble représenter.

Nous réaliserons donc ces différents types de modélisation pour l'étude d'une garantie santé (Médecin Spécialiste non OPTAM), dont nous expliquerons les principes, et nous analyserons les écarts obtenus, qui nous permettront de conclure sur la fiabilité de nos normes, et de définir les adaptations à envisager en conséquence concernant notre modèle interne et nos grilles, dont la forme reste contrainte à l'outil de tarification dans lequel elles sont stockées.

1. Contexte et cadre de l'étude

1.1. Organisme d'accueil

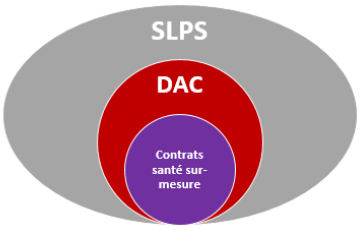
Le groupe Swiss Life a été fondé en 1857 à Zurich, est aujourd'hui l'un des leaders sur le marché de l'assurance, et un des principaux fournisseurs européens de produits d'assurance-vie, de prévoyance et retraite.

La société Swiss Life France créée en 1898 est un acteur dont le principal objectif est d'honorer sa réputation de sécurité en accompagnant ses clients durant chaque étape de leur vie dans la réalisation de leurs projets, tels que l'épargne, la retraite, la gestion de patrimoine, la banque privée, la santé, la prévoyance et le dommage.

Swiss Life Prévoyance et Santé (SLPS) est la filiale de Swiss Life France distribuant les produits assurances santé et prévoyance.

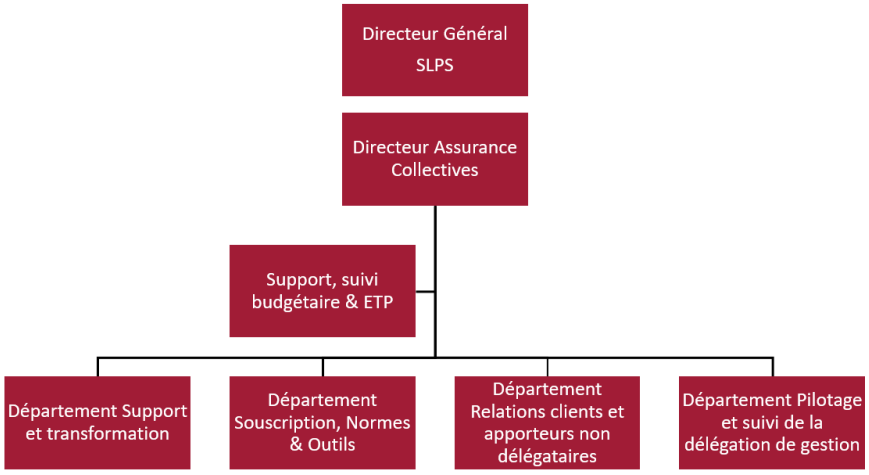
Il s'agit d'une société d'assurance telle que définie par l'article L.322-1-2 du Code des assurances.

La Direction des Assurances Collectives (DAC) distribue et gère les contrats collectifs sur-mesure proposés en santé, prévoyance et retraite.



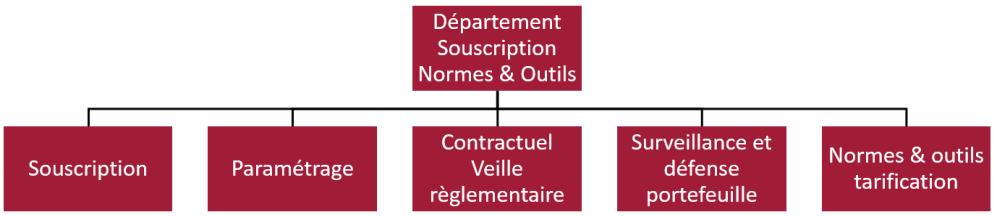
1-Organigramme macro SLPS

Depuis novembre 2019, l'organisation de la DAC est la suivante :



2-Organigramme de la DAC dans SLPS

C’est au sein du département Souscription, Normes et Outils que sont créées les normes de tarifications en santé et prévoyance, dans l’unité Normes et Outils. Le département est organisé comme suit :



3-Organigramme du département Souscription Normes & Outils

Les organismes assureurs comme Swiss Life Prévoyance et Santé peuvent gérer leur contrat d’assurance en interne, ou peuvent transférer une partie ou la totalité de la gestion d’un contrat à un tiers externe.

Il s’agit de confier tout ou partie des activités menant à l’exécution d’un contrat d’assurance à un autre organisme, qui effectuera ces tâches de façon autonome en utilisant ses propres ressources humaines, matérielles et financières.[1]

Le marché de l’assurance santé/prévoyance est marqué par une forte externalisation des actes de gestion. C’est notamment le cas pour la DAC qui externalise la gestion d’une part importante de la gestion de ses contrats.

1.2. L’assurance santé en France

En France, la sécurité sociale désigne un ensemble de dispositifs et d’institutions permettant de protéger les individus des conséquences d’évènements ou de situations diverses, généralement qualifiés de « risques sociaux » (maladie, accident de travail et maladie professionnelle, maternité et vieillesse)[2].
La Sécurité sociale est un système de solidarité collective et obligatoire entre citoyens.

Il existe 4 régimes principaux de Sécurité Sociale :

- Le Régime Général / Régime Local Alsace-Moselle : couvre les salariés du secteur privé de l’industrie, du commerce et des services.
- Le Régime Social des Indépendants.
- Le Régime Agricole.
- Autres Régimes Spéciaux.

Depuis 1967, le régime général est composé de quatre branches :

- La branche de l’assurance maladie, maternité, paternité, invalidité et décès gérée par la Caisse Nationale de l’Assurance Maladie des Travailleurs Salariés (CNAMTS) : permet essentiellement de couvrir les dépenses d’hospitalisation, de médicaments et de consultations de professionnels de santé.

- **La branche accidents du travail - maladies professionnelles** gérée également par la CNAMTS : prend en charge les frais liés aux maladies professionnelles et aux accidents de travail.
- **La branche famille** gérée par la Caisse Nationale des Allocations Familiales (CNAF) : verse des prestations liées à la naissance, à la garde d'enfants, aux aides à l'éducation ou au logement.
- **La branche retraite** gérée par la Caisse Nationale de l'Assurance Vieillesse des Travailleurs Salariés (CNAVTS) : prend en charge l'inscription des revenus sur le compte vieillesse de chacun durant sa vie active pour calculer les montants de retraites versés ultérieurement.

Principe de remboursement de la Sécurité Sociale : en assurance santé, le remboursement de l'Assurance Maladie dépend de plusieurs déterminants :

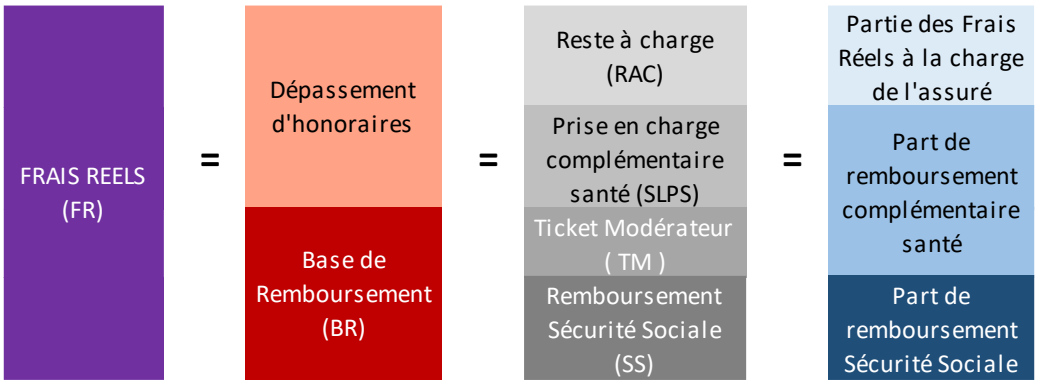
- Le type d'acte médical (consultation généraliste, prothèse dentaire…)
- Le conventionnement du médecin (dépassement d'honoraires ou non…)

Le tarif conventionné est utilisé comme base de calcul du montant des remboursements des frais de santé d'un acte médical par l'assurance maladie (appelé base de remboursement (BR) de la sécurité sociale). Exemple : médecin généraliste : BR=25€

Pour chaque régime, un taux est défini pour calculer la part de remboursement de la sécurité sociale (exemple : 70%BR soit 17,5€)

Le reste du montant de la BR est appelé ticket modérateur (exemple : médecin généraliste : 30%BR).

Les complémentaires santé remboursent le ticket modérateur agrégé à un éventuel dépassement d'honoraire, en fonction du niveau de garantie de l'assuré, qui peut protéger l'assuré d'un éventuel reste à charge sur le dépassement d'honoraires.



4- décomposition d'un remboursement santé

En 2018, la consommation de soins et de biens médicaux (CSBM) est évaluée à 203,5 milliards d'euros[3]. Celle-ci rassemble les soins hospitaliers, les soins courants (médecins, dentistes), les auxiliaires médicaux, les médicaments, autres biens médicaux (optique, prothèses, ... etc) et le transport des malades (comme le SAMU).

Ces dépenses importantes reflètent le haut niveau de protection offert par l'Assurance Maladie, qui prend en charge une part élevée des dépenses de santé (78,1%), mais également une progression non maîtrisée de ces dépenses.

[3] Source : DREES

En effet, dans la période des trente glorieuses, la montée en puissance du système de santé français a lieu dans un climat social, politique, économique et sanitaire prospère.
Cependant, à partir des années 1990, les dépenses de l'Assurance Maladie progressent à un rythme supérieur à celui du Produit Intérieur Brut (PIB), mais également à celui de ses recettes (cotisations sociales et impôts), ce qui provoque des déficits récurrents.

Presque la moitié des dépenses est faite pour les soins hospitaliers, et le deuxième poste le plus consommateur concerne les soins de ville (médecins généralistes, spécialistes…)

Pour faire face à l'augmentation de la part des dépenses de santé dans le PIB (11% depuis le milieu des années 2010[4], la France tente a créé des mesures par les prix, la négociation, l'étatisation et la privatisation.

L'assurance santé complémentaire est donc devenue au cours du temps un élément crucial du système de santé, pour l'accès aux soins.

1.3. L'évolution des niveaux de remboursement et le contexte de l'OPTAM/OPTAM-CO

Mise en place des contrats d'accès aux soins

Le Contrat d'Accès aux Soins (CAS) est entré en vigueur en 2013.

Il s'agit d'un contrat signé par les médecins de secteur 2 (pratiquant des dépassements d'honoraires) avec l'Assurance maladie. Son principe était d'engager les signataires à :

- Limiter leur dépassement d'honoraires à 100% du tarif de base de remboursement fixé par l'assurance maladie,
- Ne pas réduire leur nombre d'actes pratiqués sans dépassement d'honoraires en 2012.

Les garanties suivantes étaient concernées :

- Consultation et visite Généraliste
- Consultation et visite Spécialiste
- Actes Techniques Médicaux
- Imagerie (radiologie, échographie…)
- Actes de chirurgie hospitalière
- Actes de chirurgie maternité

Couplé avec un contrat santé responsable, mis en place en 2016, qui garantit un niveau de remboursement minimal et maximal par les complémentaires santé, les adhérents des complémentaires santé sont devenus mieux remboursés par celles-ci en allant consulter un médecin ayant signé le CAS plutôt qu'un médecin non signataire du CAS (remboursement minimum 20%BR supérieur), et les remboursements des consultations de médecins non signataires sont aujourd'hui plafonnés à un dépassement de 100% de la base de remboursement.

Le CAS était une première solution pour inciter les médecins à limiter leur dépassement d'honoraires. Cependant cette solution est limitée en terme de résultats car le système repose sur

[4] Source : Le Monde

la base du volontariat de ces derniers, et que ceux qui n'adhèrent pas peuvent continuer à pratiquer des dépassements d'honoraires parfois très couteux.

Mise en place de l'OPTAM en 2017

Début 2017, le contrat d'accès aux soins évolue et est remplacé par deux nouvelles options proposées à la signature des praticiens [3] :

- L'option pratique tarifaire maîtrisée (OPTAM) destinée aux médecins généralistes et spécialistes de secteur 2,
- L'option pratique tarifaire maîtrisée applicable aux spécialistes en chirurgie ou en gynécologie-obstétrique (OPTAM-CO) de secteur 2.

Ces nouvelles dispositions, plus souples et plus incitatives, ont contribué à augmenter le nombre de médecins signataires et à faire reculer les dépassements d'honoraires.

Aujourd'hui, les règles de remboursement des consultations médicales des soins du périmètre concerné par l'OPTAM sont les suivantes [4] :

ACTES	PRESTATIONS
MEDECINE COURANTE, HOSPITALISATION et MATERNITE	
Consultations, visites, actes	
Médecins ayant souscrit à l'OPTAM/OPTAM-CO	Min. : TM
Médecins n'ayant pas souscrit à l'OPTAM/OPTAM-CO	Min. : TM Max. : le plus petit des 2 montants ▪ OPTAM/OPTAM-CO - 20 % BR ▪ 200 % BR – SS (TM + 100 % BR)

Exemple de l'effet de l'OPTAM sur une consultation Généraliste à 90€ avec une garantie de 360%BR en DPTAM et 200%BR sinon :

	OPTAM	NON OPTAM
Frais réels (FR)	90 €	90 €
Base Remboursement (BR)	25 €	25 €
Rbt Sécurité Sociale (SS)	17,50 €	17,50 €
Rbt régime complémentaire*	**72,5€ (=360%BR)**	**32,50 €**
Reste à Charge (RAC)	**0 €**	**40 €**

5-*Exemple de décomposition de remboursement d'un médecin spécialiste*

2. La norme santé sur-mesure existante chez Swiss Life

2.1. Portefeuille étudié

La norme tarifaire des garanties santé des contrats collectifs sur-mesure de Swiss Life a été créée en 2013 et intégrée dans un nouvel outil de tarification en 2014.

Pour la construire, Swiss Life a utilisé son portefeuille de contrats en santé locale gérés en interne (16% des primes du portefeuille global), ainsi que celui de son principal délégataire de gestion (10% des primes du portefeuille global).

L'étude portait sur l'exercice technique 2013. Le portefeuille étudié contenait après retraitement 42 982 assurés principaux, 18 060 conjoints et 31 179 enfants soit un total de 92 221 assurés, dont 42% des assurés principaux et 67% des conjoints étaient des femmes.

L'âge moyen des assurés principaux était de 39,1 ans et celui des conjoints de 42,5 ans.

Le nombre de lignes de prestations étudiées était de 3,1 millions, tous postes confondus.

En 2013, le portefeuille étudié représentait 41,4 M€ de primes pour un S/P de 94,6% (et un R/C de 108,3%).

Précisions que ces données concernent l'ensemble des garanties santé proposées par Swiss Life à l'époque. Le périmètre était donc plus étendu que pour notre étude.

A cette date, les contrats d'accès aux soins n'étaient pas encore en place en France. Le modèle de construction des normes de l'époque n'a donc pas scindé les garanties selon les types de médecins. Il a été adapté plus tard au fil des réformes.

2.2. La contrainte de l'outil de tarification

La Direction des Assurances Collectives de Swiss Life (DAC) propose des contrats santé avec des garanties sur-mesure en terme de catalogue et de niveaux de garanties pour les entreprises. Pour réaliser ces tarifications à une granularité si fine, un outil d'avant-vente (OAV) nommé Antenia a été conçu, lui aussi sur-mesure, par un prestataire externe, en étroite collaboration avec les équipes de Swiss Life. Cet outil est en permanente évolution compte-tenu de son importance croissante au sein des métiers de Swiss Life.

Il permet à l'équipe Normes et Outils de Swiss Life de créer/modifier/supprimer n'importe quel type de garantie santé, et d'en stocker les grilles tarifaires manuellement pour être directement utilisables par les équipes de souscription. Cela permet une grande flexibilité en terme de catalogue de garanties, mais aussi une grande réactivité face aux besoins règlementaires et commerciaux.

L'outil de tarification est structuré de manière à calculer une « prime pure » qui dépend de la garantie, de son niveau de couverture, de la zone géographique à laquelle appartiennent les assurés de l'entreprise, et de l'assiette de remboursement (BR, BRSS, TM etc.). Cette prime pure

correspond au tarif de cette garantie pour un assuré type de sexe masculin, âgé de 40 ans, de type Adhérent principal et sur un contrat obligatoire de base.

La prime pure est stockée dans l'outil de tarification sous forme de grille de tarif propres à chaque garantie sous cette forme :

		Régime général			Alsace-Moselle	Monaco
		Zone unique	Zone 1	Zone 2	Zone unique	Zone unique
ASSIETTE	NIVEAU					
%BR	30					
	50					
	100					
	150					
	200					
	300					
	500					
% BR -SS	100					
	200					
	300					
	400					
	500					

6- Format de la grille de tarification des primes pures dans l'outil d'avant-vente

Puis on calcule une série de grilles de coefficients pour les variables explicatives (âge, sexe, option, rang…) qui seront multipliés à cette prime pure afin de la rendre personnalisable à chaque type d'assuré, quelles que soient ses caractéristiques.

Dans le prochain paragraphe nous allons décrire cette structure car elle correspond au modèle de tarification utilisé historiquement par Swiss Life pour calculer ses normes tarifaires santé sur-mesure.

2.3. Modèle de construction de la norme Swiss Life existante

2.3.1. Variables explicatives

L'ensemble des variables explicatives retenues à l'époque de la construction de la norme tenaient à la fois compte de l'influence objective qu'elles pouvaient avoir sur le tarif, mais aussi des contraintes de récupération des informations dans le cadre des appels d'offres qui sont traitées au quotidien dans l'entreprise.

On liste les variables suivantes :

- **Age** de l'assuré
- **Sexe** de l'assuré
- **Catégorie Socio-Professionnelle (CSP)**
- **Rang** du bénéficiaire (adhérent, conjoint, enfant)
- **Code NAF** de l'entreprise : Pour prendre en compte le secteur d'activité, les codes NAF ont été regroupés au sein de plusieurs classes de risque. Il y avait en effet trop de codes NAF différents pour espérer avoir suffisamment de données pour étudier chaque code séparément.
- **Département** du siège de l'entreprise

- **Régime de base** de la sécurité sociale (Régime général, Alsace-Moselle, Monaco) : Ces garanties sont, par défaut, souscrites à titre obligatoire pour tous les salariés. Elles peuvent être également optionnelles pour tous les salariés. Le régime demandé peut contenir une base obligatoire pour tous et des options facultatives.
- **Type de contrat** : obligatoire/facultatif/ à option
- **Mode d'expression tarifaire (assiette)** : sous quelle forme on va rembourser les assurés : %BR, %BRSS, %PMSS…
- **Taille de l'entreprise** (nombre de salariés présents) : cette donnée avait été ajoutée pour tenir compte d'un éventuel effet lié au nombre de salariés que compte l'entreprise souscriptrice.
- **Garantie souscrite** (médecin généraliste, médecin spécialiste…): cette donnée fera partie d'un groupe (qu'on appellera poste – exemple : Soins Courants)
- **Niveau de garantie** (autrement dit le niveau de remboursement de l'assuré par Swiss Life - exemple : 100%BR)

2.3.2. Calcul des primes pures

Parmi les paramètres qui interviennent dans la tarification, ceux dont l'effet est impossible à réduire à un coefficient multiplicatif ont été isolés dans la prime pure. :

- Niveau de garantie
- Assiette
- Zone de résidence des assurés
- Régime de Sécurité Sociale

En effet, le coût d'une garantie est directement lié au niveau de garantie proposé, mais cette dépendance est très variable selon les garanties.

De même, la zone de résidence a un impact direct sur les coûts des dispositifs de santé et la pratique de dépassements d'honoraires.

Les prestations versées par l'assureur sont complémentaires au régime de la Sécurité Sociale. Par construction, pour un même niveau de garantie, plus la Sécurité Sociale rembourse l'assuré, plus le coût pour l'assureur réduit. C'est notamment le cas pour les assurés qui ne sont couverts que pour le ticket modérateur et sont affiliés au régime local Alsace-Moselle pour lequel ce ticket modérateur est fortement réduit voir nul sur certaines garanties.

Les assiettes des garanties font également varier le tarif indépendamment des trois précédents paramètres.

Le principe de construction de la norme est de calculer une prime pure pour un assuré « standard », qu'on appellera par la suite assuré type: Homme adhérent principal âgé de 40 ans, assuré au moins 90 jours chez Swiss Life sur l'année étudiée sur un contrat de base (pas d'option).

Cette prime pure sera une fonction du niveau de garantie, de l'assiette de garantie, et du régime de base.

Pour la suite, on note la prime pure donnée par cette table PP_i pour une garantie i. On a donc $PP_i = PP_i(expression_i, niveau_i, régime, zone)$. Régime et zone dépendent de l'assuré pour lequel la tarification est faite.

Méthode théorique du modèle interne :

Nous incluons les assurés de sexe féminin pour s'assurer de pouvoir observer tous les effets des caractéristiques discriminantes (par exemple, on observera mieux les effets sur la garantie maternité).

Le tarif obtenu avec cet échantillon sera ensuite redressé via des coefficients (cf. 3.3.3.) afin d'obtenir un tarif pour homme de 40 ans assuré principal.

Pour chaque zone, nous sommons pour chaque niveau de garantie :

- Le total des jours risque couverts de notre échantillon d'assurés
- Le montant total des prestations versées sur la garantie pour ces mêmes assurés

Nous calculons à partir de ces deux valeurs, toujours sur chaque niveau de garantie :

- **Le poids :** nombre de jours risque par rapport au nombre total de jours risque observés.
- **Le coût moyen :** rapport entre les prestations versées et le nombre de jours risque multiplié ensuite par 365 pour être exprimé en année.

Notre portefeuille contient beaucoup de niveaux de garanties distincts, et nous obtenons des tables avec des poids très inégaux en fonction des niveaux. Un travail de retraitement est donc nécessaire :

- on sélectionne les points remarquables en fonction du critère de poids (poids > 10%), pour améliorer la fiabilité du résultat on considère les valeurs qui encadrent ces points en pondérant les coûts moyens par les poids respectifs.

- si cela aboutit à créer de trop grands intervalles sans point remarquable et en particulier pour les niveaux faibles et forts, nous ajoutons des points en regroupant plusieurs niveaux de poids moindre.

On calcule les points des niveaux des courbes à partir des regroupements de poids, par moyenne pondérées des niveaux correspondants au poids.

Exemple de grille :

Zone	Niveau de garantie	Jours risque	Prestations	POIDS	Coût moyen	Somme de poids	Niveau de garantie moyen	TARIF Coût moyen
1	0,3	40348	837 €	1%	7,57 €	1%	30%	7,57 €
1	0,35	13870	139 €	0%	3,67 €	1%	56%	13,55 €
1	0,35	3429	15 €	0%	1,57 €			
1	0,5	4315	502 €	0%	42,46 €			
1	0,6	3283	100 €	0%	11,07 €			
1	0,8	15089	729 €	0%	17,63 €			
1	1	69088	8 612 €	2%	45,50 €	57%	136%	31,15 €
1	1,1	16756	508 €	1%	11,06 €			
1	1,2	8022	1 113 €	0%	50,63 €			
1	1,3	1404272	81 533 €	42%	21,19 €			
1	1,5	219177	39 497 €	7%	65,78 €			
1	1,55	4379	484 €	0%	40,34 €			
1	1,7	365	110 €	0%	110,00 €			
1	1,8	137319	24 602 €	4%	65,39 €			
1	2	27896	4 703 €	1%	61,53 €			
1	2,1	970	76 €	0%	28,63 €	32%	370%	60,92 €
1	2,3	84089	14 406 €	3%	62,53 €			
1	2,55	10676	709 €	0%	24,25 €			
1	2,7	14928	2 953 €	0%	72,21 €			
1	3	133469	25 910 €	4%	70,86 €			
1	3,5	14199	4 264 €	0%	109,61 €			
1	4	808891	129 809 €	24%	58,57 €			
1	4,1	8760	1 423 €	0%	59,28 €			
1	4,3	13249	1 795 €	0%	49,44 €			
1	5	154260	21 316 €	5%	50,44 €	8%	538%	61,05 €
1	5,4	365		0%	- €			
1	5,6	3650	791 €	0%	79,12 €			
1	5,9	2190	18 €	0%	2,96 €			
1	6	93484	20 004 €	3%	78,10 €			

7-Exemple de construction d'une grille de primes pures

Nous verrons dans la suite qu'un lissage peut s'avérer nécessaire afin d'obtenir des courbes cohérentes avec notre modèle de tarification, qui induit les contraintes tarifaires suivantes :

- Les tarifs doivent respecter les ordres suivants en fonction des zones (contrainte commerciale pour cohérence tarifaire de la grille finale)

$$PP_{zone1} \geq PP_{zone\ unique} \geq PP_{zone2}$$

- Les tarifs de toutes les zones doivent être égaux pour le niveau du ticket modérateur (ou 100%TM égal à 30%BR en soins courants, 20%BR en hospitalisation et maternité). En réalité il existe des écarts, mais ils sont négligeables. De plus les dépassements d'honoraires sont justifiés seulement après le ticket modérateur

$$PP_{zone1}(100\%\text{TM}) = PP_{zone\ unique}(100\%\text{TM}) = PP_{zone2}(100\%\text{TM})$$

- Les courbes des tarifs doivent être croissantes avec le niveau de garantie. C'est une contrainte commerciale logique : on ne peut pas vendre moins de garanties plus cher.

2.3.3. Calcul des coefficients

Après calcul de la prime pure, des coefficients ont été calculés pour obtenir un tarif selon les différentes variables explicatives du tarif : âge, sexe, type de bénéficiaire, zone géographique, code NAF, classe CSP, taille de l'entreprise, et le caractère obligatoire ou facultatif du contrat.

Presque tous les coefficients ont été calculés à la maille du poste qui contient les garanties (on a donc le même coefficient pour toutes les garanties appartenant au Poste étudié) :

- **Age** de l'assuré
- **Sexe** de l'assuré
- **Rang** du bénéficiaire (adhérent, conjoint, enfant)
- **Catégorie Socio-Professionnelle (CSP)** : à l'époque, il a été estimé qu'aucun effet particulier de la catégorie de personnel n'a été identifié sur la consommation santé. Les différences de consommation sont en grande partie dues aux différences de garanties qu'il y a en général entre les cadres et les non-cadres. Ce coefficient est donc égal à 1 dans le modèle actuel.
- **Code NAF** de l'entreprise : Pour prendre en compte le secteur d'activité, on peut regrouper les codes NAF au sein de plusieurs classes de risque. Il y a en effet trop de codes NAF différents pour espérer avoir suffisamment de données pour étudier chaque code séparément. En santé, les différents essais réalisés n'ont pas permis d'identifier un effet particulier en fonction du secteur d'activité.
- **Département** du siège de l'entreprise: il ne s'agit pas d'un coefficient mais d'une grille attribuant à chaque département une zone de tarification, qui scinde les assurés selon 4 zones :
    - Zone 1 qui correspond aux département dont les niveaux d'honoraires des médecins sont les plus élevés,
    - Zone 2 qui correspond aux département dont les niveaux d'honoraires des médecins sont les plus élevés
    - Zone Alsace-Moselle qui correspond aux département situés en Alsace-Moselle et qui bénéficient donc de taux de remboursements différents des autres zones
    - Zone Monaco, pour la principauté
    - Zone unique : qui regroupe les départements de zone 1 et de zone 2 sans distinction
- **Type de contrat** : obligatoire/facultatif/ à option
- **Taille de l'entreprise** (nombre de salariés présents) : cette donnée avait été ajoutée pour tenir compte d'un éventuel effet lié au nombre de salariés que compte l'entreprise souscriptrice. Lors de la construction de la norme, l'impact de ce paramètre n'a pas été mis en évidence. Ce coefficient est depuis fixé à 1 pour tout poste.

Notre étude étant limitée à une partie seulement des garanties faisant partie des postes Hospitalisation et Soins courants, il ne nous est pas possible de recalculer ces coefficients de poste, car cela modifierait également la valeur des autres garanties qui font partie de ces mêmes postes mais qui sont hors du périmètre de notre étude.

Nous calculons cependant les coefficients suivants :

Coefficient d'assiette %BRSS :

Nous avons exprimé toutes les garanties en %BR (pourcentage de la base de remboursement) pour construire les grilles mais il est nécessaire de disposer également des tarifs l'assiette %BRSS (pourcentage de la base de remboursement moins la participation de la Sécurité Sociale). Ces deux expressions sont des assiettes couramment utilisées dans les contrats santé, et le fait d'être en sur-mesure nous empêche de pouvoir imposer l'un ou l'autre des modèles. Nous sommes donc obligés de proposer les deux assiettes aux assurés, selon leur choix.

Tout d'abord, on définit les pas qui vont être utilisés sur la grille %BR-SS à partir de ceux retenus pour la grille exprimée en %BR. A ce stade, on utilise une conversion en ajoutant le taux de couverture de la Sécurité sociale (estimé à 70%BR pour les soins courant et 80%BR pour l'hospitalisation et la médecine courante).

Concrètement, on ajoute ces taux constants à tous nos niveaux de garantie pour obtenir le taux en %BRSS

Pour chaque garantie de chaque type, on récupère la liste des prestations détaillées (Frais réels, base de remboursement, montant du remboursement de la sécurité sociale) pour les zones 1 et zones 2 (on exclut donc la zone Alsace-Moselle, dont les tarifs seront eux-aussi calculés par coefficient).

Puis à partir de cette grille on répète l'opération suivante : pour chaque niveau de garantie %BR, on calcule le niveau équivalent de garantie en %BRSS (soit niveau %BR +70% BR pour la garantie généraliste par exemple), ainsi que les prestations qui correspondraient au montant des remboursements de Swiss Life pour ces prestations.

Le coefficient de passage de l'assiette %BR à l'assiette %BRSS pour une garantie de type et de niveau donné est le rapport entre le coût total du taux converti et le coût total du taux exprimé en %BR.

$$\textbf{coeff assiette}\ (\textbf{niveau de garantie X}) = \frac{\sum \textbf{prestations}\ (\textbf{niveau de garantie X, assiette BR})}{\sum \textbf{prestations}\ (\textbf{niveau de garantie X, assiette BRSS})}$$

Exemple de grille de coefficient :

BR	BRSS	Coefficient assiette
0%	0%	0,000000
30%	100%	0,905122
100%	170%	0,976303
150%	220%	0,993272
200%	270%	0,998830
300%	370%	1,000000
400%	470%	1,000000
530%	600%	1,000000

Le calcul des tarifs pour l'assiette %TM ne nécessite pas de calcul de coefficient : en effet nous n'avons besoin que des niveaux 0 et 100%TM, le niveau 100%TM correspondant au niveau 30%BR pour les soins courants et à 20%BR pour les garanties hospitalisation et maternité.

Coefficient de Régime Alsace-Moselle

Pour adapter les grilles des primes pures obtenues avec le régime général au régime Alsace-Moselle, on utilise les prestations réglées pour les assurés du régime général zone 2 (qui contient le régime Alsace-Moselle et la zone 2 du régime général) exprimée en % BR –SS. On scinde cet échantillon en régime général et en régime Alsace-Moselle. Pour chacune de ces zones, on applique ensuite la garantie de chacun des pas à toutes les prestations. On calcule alors un coût moyen de la prestation pour les assurés affiliés à chacun des régimes (général et et Alsace-Moselle). En faisant le rapport entre les deux, on obtient un coefficient permettant de passer d'un régime à l'autre.

$$\textbf{coeff regime}\ (\textbf{niveau de garantie X}) = \frac{\left(\frac{\sum \textbf{prestations}\ (\textbf{niveau de garantie X, Alsace} - \textbf{Moselle})}{\textbf{nombre actes Alsace} - \textbf{Moselle}}\right)}{\left(\frac{\sum \textbf{prestations}\ (\textbf{niveau de garantie X, Zone 2})}{\textbf{nombre actes zone 2}}\right)}$$

BR	BRSS	Coeff assiette
0%	0%	0
30%	100%	0,40725562
100%	170%	0,66946327
150%	220%	0,73724154
200%	270%	0,75329693
300%	370%	0,74983571
400%	470%	0,74881222
530%	600%	0,74868296

Coefficient Age/sexe :

Le coefficient Age/sexe existant dans notre outil de tarification ne sera pas retouché pour cette étude. En effet, pour rappel, ces coefficients sont liés au niveau des postes et pas à la maille des garanties que nous modélisons.

Nous devons cependant recentrer notre échantillon pour qu'il corresponde à celui d'un homme de 40 ans.

On cherche à exprimer :

$$\text{coût H 40 ans}$$

Alors que l'on a observé :

$$\text{coût HF 35} - \text{45 ans}$$

On fait l'hypothèse suivante : $\text{Coût HF 35} - \text{45 ans} = \text{coût HF 40 ans}$

Cela revient à dire que le coût de la garantie sur la tranche d'âge correspond au coût sur l'âge moyen de la tranche d'âge.

Comme par ailleurs,

$$\text{coût HF 40 ans} = \%\,\text{H} * \text{coût H 40 ans} + \%\text{F} * \text{coût F 40 ans}$$

Notre hypothèse de tarification permet d'exprimer le coût de la garantie pour une femme à partir du coût pour un homme :

$$\text{coût F 40 ans} = \text{coût H 40 ans} * \text{coefficient F 40 ans}$$

On obtient :

$$\textbf{coût H 40 ans} = \frac{\textbf{coût HF 35} - \textbf{45 ans}}{\%\,\textbf{H} + (\textbf{1} - \%\textbf{H}) * \textbf{coefficient F 40 ans}}$$

- Bénéficiaire :

De la même manière, nous recentrons notre échantillon pour qu'il corresponde en plus de ces critères à un assuré principal :

$$\textbf{coût H 40 ans assuré principal} = \frac{\textbf{coût HF 35 - 45 ans}}{[\%\mathbf{H} + (\mathbf{1} - \%\mathbf{H}) * \textbf{coeff F 40 ans}] * [\%\mathbf{A} + (\mathbf{1} - \%\mathbf{A}) * \textbf{coeff C}]}$$

Où %A est le pourcentage d'adhérents principaux dans l'échantillon et coeff C le coefficient correspondant aux conjoints.

Coefficient option :

Nous avons fait notre étude sur un portefeuille contenant des contrats à niveau base ou options, et notre outil de tarification implante déjà un coefficient de majoration en fonction du niveau des options pour compenser l'anti sélection que cela implique. En effet, les contrats à option ont par définition des niveaux de garanties plus grands que les contrats de base auxquels ils sont rattachés. Les assurés qui choisissent une option sont par définition ceux qui « prévoient » de consommer plus que la garantie de base, ce qui créé une anti sélection, que nous tentons de contrôler via ce coefficient d'option.

Les prestations étudiées ont donc été classées par niveau d'option, puis à chaque niveau, nous avons retiré la part d'anti sélection estimée.

$$\textbf{Prestation finale (Option X)} = \frac{\textbf{Prestation option X}}{[(\mathbf{1} + \textbf{coeff\_optionX})]}$$

Ce montant correspond au montant de prestations estimé si tous les contrats étaient d'un niveau de base.

On obtient un coefficient « option » en divisant la prestation finale par la presta initiale.

2.3.4. Résultat technique

Le tarif technique pour un assuré i est obtenu en multipliant la prime pure par l'ensemble des coefficients

$$\boldsymbol{PR_i = PP_i(expression_i, niveau_i, régime, zone) * \prod_{i=1}^{Nombre\ de\ coeff} Coefficientk(i)}$$

Majoration des grilles obtenues

Le calcul de ces grilles a été fait à partir des prestations de l'année 2018. Comme les tarifs seront exprimés en % PASS, nous avons converti les tarifs obtenus à partir de la valeur du PASS 2018. Les grilles obtenues sont ensuite majorées pour tenir compte de différents facteurs :

- Majoration due à la Provision pour Sinistres à Payer (PSAP) pour tenir compte du fait que l'ensemble des sinistres de l'année étudiée ne sont pas connus au moment de l'étude : +0,6%

- Majoration liée à notre proportion d'assurés ANI : +5%

Au total, cela représente donc une majoration de 5,63% que l'on applique aux tarifs obtenus.

2.4. Tarifs et hypothèses actuellement en place sur les garanties OPTAM/OPTAM-CO

Les garanties concernées par le CAS puis par l'OPTAM sont les suivantes :

POSTE	GARANTIE
HOSPITALISATION	Honoraires et Actes
MATERNITE	Honoraires et Actes
SOINS COURANTS	Généralistes - consultations et visites
SOINS COURANTS	Spécialistes - consultations et visites
SOINS COURANTS	Actes techniques médicaux
SOINS COURANTS	Actes d'imagerie et actes d'échographie

8- Liste des garanties concernées par l'OPTAM

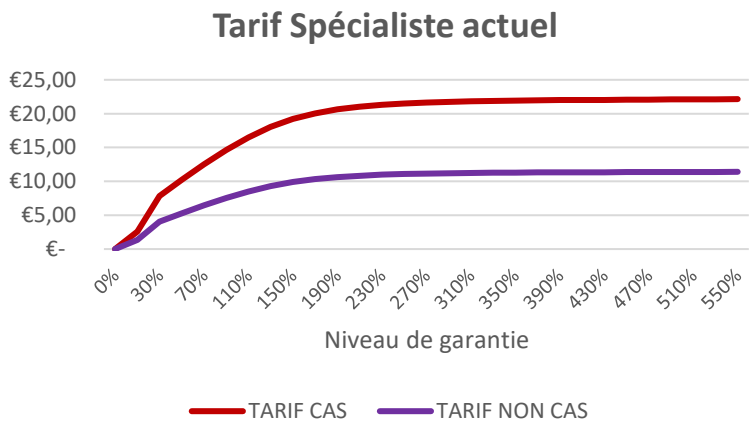
Les normes de Swiss Life ont été créées avant la mise en place du CAS et étaient donc sous forme de tarif unique pour chaque garantie.

A l'époque de la mise en place du CAS, Swiss Life ne possédait pas d'information sur la répartition des médecins signataires du CAS. Il a donc été décidé de répartir la norme globale de chaque garantie en affectant un pourcentage CAS/NON CAS (puis DPTAM/NON DPTAM) pour chaque garantie.

La répartition de l'époque était la suivante :

GARANTIE	%OPTAM
ATM	85%
CHIRURGIE	40%
GENERALISTE	94%
IMAGERIE	93%
MATERNITE	50%
SPECIALISTE	66%

9- Répartition en % des prestations liées à des médecins adhérents à l'OPTAM



10- Tarif spécialiste de la norme actuelle

Cette décision a plusieurs impacts dont le fait que les courbes de chaque classe de garantie gardent la même forme que la courbe de tarification, ce qui est à vérifier.

Dans le cadre de cette étude, les principales questions que nous nous posons sont :

- **La consommation a-t-elle évoluée depuis 2014 pour chaque garantie?**

- **La courbe garde-t-elle vraiment la même allure si on distingue le tarif OPTAM/OPTAM-CO des autres consultations?**

3. Back testing des norme OPTAM/OPTAM-CO

3.1. Récupération et nettoyage des données

3.1.1. Description des données

La récupération et le nettoyage des données est une étape essentielle du processus de création d'un tarif.

Nous avons récupéré les données 2018 des contrats gérés en interne, ainsi que celles de notre principal délégataire de gestion.

On distinguera 3 types de fichiers contenant chacun des données spécifiques :

- **Assurés :** données démographiques des assurés du portefeuille (identifiant, âge, sexe, nombre de jours risques, département, type de bénéficiaire, niveau d'option souscrit. Ces données doivent être anonymes
- **Garanties:** type de garanties et leurs niveaux par contrat
- **Prestations:** ligne à ligne des consommations des assurés et des remboursements Swiss Life associés : frais réels, base de remboursement, montant remboursé par la sécurité sociale, montant remboursé par une éventuelle autre couverture santé, montant remboursé par Swiss Life, nombre d'actes, type d'actes, assuré ayant consommé l'acte.

Ces données nous sont fournis sous formes différentes, et il faut les nettoyer et parfois les calculer à partir d'autres données pour les adapter à notre étude et nous permettre de les fusionner.

Après nettoyage de ces fichiers, on pourra procéder à une jointure des données via leurs points communs, afin d'obtenir un fichier unique contenant toutes les informations nécessaires à la construction du tarif.

Notre back-testing ne concernera que les garanties OPTAM citées dans la partie 2.3.

Nous allons réévaluer les normes de ces garanties en comparant 3 types de modélisations : le modèle interne (la norme Swiss Life dite « norme Antenia »), une modélisation linéaire généralisée (GLM), et un modèle Machine Learning qui nous semblera le mieux adapté à notre échantillon.

En ce qui concerne la modélisation de la norme Antenia, les garanties étudiées font partie des postes Hospitalisation, Maternité et Soins Courants dans notre outil de tarification. Les coefficients décrits dans la partie 2.3.3. sont construits dans l'outil pour être paramétrés au sein des postes des garanties, ce qui implique qu'ils soient valables pour toutes les garanties de chacun de ces postes. Notre étude n'est donc pas assez exhaustive pour permettre de réévaluer ces coefficients. On se contentera de réévaluer la prime pure $PP_i = PP_i(expression_i, niveau_i, régime, zone)$

On se limitera donc aux coefficients suivants :

- Coefficient d'assiette (passage de l'assiette %BR à l'assiette %BRSS)
- Coefficient de régime (passage du régime général au régime Alsace-Moselle)

Et on considèrera comme corrects les coefficients suivants :

- Coefficient âge/sexe

- Coefficient de type bénéficiaire
- Coefficient de poste
- Coefficient de taille d'entreprise
- Coefficient de CSP
- Coefficient de code NAF
- Coefficient d'option

Et on ne changera pas la répartition du zonier.

Cas du GLM et du Machine Learning :

La modélisation par GLM doit par définition calculer de nouveaux coefficients en fonction des variables tarifaires. Et les formules et paramètres obtenus par Machine Learning dépendent du modèle choisi. Nous calculerons donc tous les paramètres nécessaires pour ces modélisations, et nous créerons une grille de comparaison adaptée qui nous permettra de comparer les résultats des 3 modélisations au plus juste.

Les données finales que nous récupérons pour notre étude sont :

- Age
- Sexe
- Nombre de jours risque de l'assuré en 2018
- Type bénéficiaire (ou rang)
- Type de contrat (base/option 1/ option 2/ Option 3)
- Niveau de garantie pour chaque garantie (Médecin généraliste, médecin spécialiste etc) et chaque type (OPTAM / NON OPTAM)
- Zone (1, 2 ou Alsace-Moselle)

Ainsi que le ligne à ligne des prestations associées à chacun de ces assurés :

- Nombre d'actes
- Frais réels (FR)
- Base de remboursement (BR)
- Montant du remboursement de la Sécurité Sociale (SS)
- Montant du remboursement d'une éventuelle autre couverture souscrite par l'assuré
- Montant du remboursement Swiss Life (SLPS)

A partir de cette même base de données, nous calculerons la prime pure « Antenia » en calculant un coût annuel. Les GLM et Machine Learning, qui n'ont pas la même contrainte de format, seront eux faits sur un modèle fréquence x Coût moyen.

3.1.2. Données gestion interne Vs gestion déléguée

La récupération des données liées aux contrats gérés directement par Swiss Life et celle des contrats gérés par des délégataires est faite séparément, du fait de leur disponibilité et de leur format parfois bien différents, qui nécessitent des traitements distincts pour pouvoir être agrégés et utilisés en base commune de travail.

Pour chaque type de données (interne ou déléguée) nous devons faire 3 types d'extractions (prestations, assurés, niveaux de garantie) et trouver les éléments de jointures entre ces types de données pour créer une base de données unique. Ces jointures imposent un nettoyage des

données minutieux, long et fastidieux, car fait en majorité manuellement, compte-tenu de la forme de notre base de données.

Gestion interne : les données de notre portefeuille en gestion interne sont stockées dans un système d'information qui permettent aujourd'hui de les récupérer via des extractions SAS.

Gestion déléguée : Pour ce back-testing, et afin d'éviter une sous-représentativité de l'échantillon, nous avons décidé de récupérer également des données via notre principal délégataire externe.

Les données seront reçues sous formes de fichiers qui auront une format et donc un traitement différent de la gestion interne. Par exemple, les niveaux de garanties de ce fichier sont écrits manuellement (assiette + niveau), il n'est donc pas possible d'automatiser la récupération de cette information. Un autre exemple de retraitement : les assurés des contrats du délégataire qui ont souscrit une option sont référencés à la fois sur le contrat de base et sur le contrat d'option. Il faut donc nettoyer le fichier des assurés (principalement manuellement) pour ne pas avoir de doublon d'assurés (et faire l'erreur de diluer les coûts moyens).

3.1.3. Homogénéisation et retraitement des données

Pour homogénéiser les données nous avons dû adapter certaines variables pour qu'elles puissent être agrégées dans une même base de données commune:

- **Les types de bénéficiaires (ou rang)** doivent avoir le même label entre gestion interne et gestion déléguée, que nous noterons A (Assuré principal), C (Conjoint) E (Enfant). Nous ne ferons pas de distinction du énième enfant, cela n'ayant pas d'influence sur le coefficient bénéficiaire dans notre outil de tarification.
- **Les niveaux de garantie** ont tous été uniformisés dans l'assiette %BR (pourcentage de la base de remboursement de la sécurité Sociale).
  Certaines garanties non responsables étaient sous la forme 100%FR (frais réels). Nous avons défini ce niveau égal au maximal des niveaux de remboursements constatés dans les autres assiettes, qui était égal à 600%BR.
  Dans le cas où la garantie était sous la forme « X%BR – dans la limite de Y%FR », nous avons abaissé le niveau équivalent en %BR de la manière suivante :
  $$Niveau\ final\ \%BR = X\%*Y\%BR$$
  Cela nous permet de tenir compte de l'effet du plafond de garantie.
- **Nous avons utilisé le zonier de notre outil de tarification** pour définir dans quelle zone appartenaient les départements de nos assurés : Zone 1, Zone 2 ou Zone Alsace-Moselle (cf Annexe 1.). il s'agit de classer les taux moyens de dépassement des départements, la zone 1 ayant les taux les plus élevés que la zone 2 pour les assurés dans le régime général.
- **L'âge de nos assurés** a été définit avec la formule suivante : *âge=2018 – année_naissance*.
- **Les niveaux d'option** ont été ajoutés manuellement, en faisant la correspondance entre les contrats liés à un même assuré ou groupe de contrat ou via les champs donnant les informations quand cela était le cas (collège par exemple).
- Pour chaque assuré, on calcule le nombre de jour de présence en 2018, compte tenu de la date d'effet de leur contrat, et de leur éventuelle date de résiliation : il s'agit de la quantité de **jour risques (JR).**
- **Puis on en déduit un** poids qui représente la proportion de leur présence en 2018 : *poids = JR/365*

A l'issue du nettoyage et de l'homogénéisation des données, notre base de données est de la structure suivante :

AGE	SEXE	COD_RANG_	B/O	zone_AMO	SPECIALISTE	OPTAM_NO	CATEGORIE	SommeDeSommeDenb_acte	SommeDeSommeDeTOTAL_RBT	JR
55	F	A	B	1	1,3	NON OPTAM	SPECIALISTE	2	59,8	364
56	M	A	B	2	1,1					364
38	M	A	B	2	1,3	NON OPTAM	SPECIALISTE	1	30	365
9	M	E	B	2	1,3	NON OPTAM	SPECIALISTE	1	25	365
6	M	E	B	2	1,3	NON OPTAM	SPECIALISTE	1	1,5	365
33	M	A	B	2	1,3	NON OPTAM	SPECIALISTE	10	77,85	365
27	F	C	B	2	1,3	NON OPTAM	SPECIALISTE	2	2,1	365
1	M	E	B	2	1,3	NON OPTAM	SPECIALISTE	18	57,6	365
42	F	A	B	2	1,3	NON OPTAM	SPECIALISTE	4	51,5	364

11- Extrait de la base de données nettoyée pour modélisation interne

Cette base de données sera commune à tous nos types de modélisation. Elle sera transformée selon les besoins de chaque modélisation.

La base de données a été transformée pour convenir à tous les types de modèles, chacun ayant ses contraintes de format, ce qui a donc induit des biais différents, que nous avons accepté.

Pour le modèle Antenia, le back-testing du tarif de la garantie s'est limité à la prime pure

$$PP_i = PP_i(expression_i, niveau_i, régime, zone)$$

Les coefficients ont été conservés, car ils étaient communs au poste de la garantie, ce qui était trop contraignant à étudier dans le cadre de notre étude. Le biais qui en découle est que la nouvelle norme sera plus juste au niveau des caractéristiques de l'assuré type, et nous pouvons juste espérer que les coefficients d'âge, de sexe, d'option et de rang seront toujours justes.

3.1.4. Nettoyage des données pour Machine Learning et GLM

Remarque importante : dans la suite de ce document, nous présenterons les résultats pour la garantie Consultation Spécialiste Non Optam. Le niveau de garantie sera alors nommé SPECIALISTE_NON_OPTAM.

Pour les études utilisant le GLM et le Machine Learning, nous avons pu récupérer les mêmes caractéristiques que pour Antenia, sauf le rang Enfant et leur âge. En effet, nous ne disposions de l'âge des enfants dans les données que nous avons récupérées chez notre délégataire et dans notre gestion interne (ils sont faux ou absents). Cela n'a pas posé problème pour la norme Antenia, car la prime pure ne contient pas de données enfant. Mais pour les autres techniques, nous avons décidé de supprimer cette variable explicative, plutôt que de la tarifer fausse.

Ce filtre exclura donc les enfants de notre étude, et fera commencer l'âge des assurés à 18 ans.

Pour ces deux techniques, nous avons également décidé de ne garder que les assurés présents au moins 360 jours risques dans l'année 2018. Prendre en compte le poids des assurés était plus complexe en terme de programmation Python, langage que nous découvrons pour faire cette étude.

On aura donc moins d'assurés dans l'échantillon réalisé avec GLM et Machine Learning qu'avec le **modèle Antenia.**

3.1.4.a. Dummification

La dummification est un processus utilisé pour les modélisations GLM et Machine Learning.

Le GLM et la plupart des algorithmes de Machine Learning prennent des variables numériques en entrée. Les variables catégorielles comme le sexe ou la zone géographie sont pour le moment renseignées sous forme de modalité, et non sous forme numérique. Le processus de dummification des données permet de transformer les différentes modalités des variables catégorielles sous forme de variable numérique (valant 1 en présence de la modalité, 0 sinon).

Nous effectuons ce processus sur toutes nos variables catégorielles : le sexe, le rang, la zone, le niveau d'option.

Notre base de données est ainsi utilisable pour une modélisation en Machine Learning.

	RANG	sex	BASE_OPTION	zone_AMO	SPECIALISTE_NON_OPTAM	year	RANG_A	RANG_C	sex_F	sex_M	BASE_OPTION_B	BASE_OPTION_O1	I
0	A	F	B	1	0.3	65	1	0	1	0	1	0	
1	A	F	B	1	1.3	24	1	0	1	0	1	0	
2	A	F	B	2	1.3	37	1	0	1	0	1	0	
3	A	F	O1	1	4.0	34	1	0	1	0	0	1	
4	A	M	O1	2	1.5	36	1	0	0	1	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
4569	C	F	O1	2	2.1	46	0	1	1	0	0	1	
4570	A	M	O1	1	1.8	29	1	0	0	1	0	1	
4571	A	M	B	1	1.3	31	1	0	0	1	1	0	
4572	C	F	O2	2	2.0	47	0	1	1	0	0	0	
4573	A	M	O2	AMO	1.1	34	1	0	0	1	0	0	

4574 rows × 19 columns

12- Base de données nettoyée pour GLM et Apprentissage Supervisé

3.1.4.b. Echantillonnage

Pour la modélisation Antenia

Les primes pures stockées dans l'outil de tarification sont calculées pour un assuré dont les caractéristiques sont les suivantes : homme de 40 ans, assuré principal sur un contrat obligatoire de base.

Une fois notre base de données nettoyée, nous devons donc filtrer les assurés qui ont ces caractéristiques précises. Nous devons cependant inclure les assurés de sexe féminin pour s'assurer de pouvoir observer tous les effets des caractéristiques discriminantes (par exemple, on observera mieux les effets sur la garantie maternité). Nous incluons également les conjoints dans notre échantillon d'étude. La norme sera centralisée à l'assuré type en dernière étape de calcul de la prime pure.

La base de données globale contenant tous les types d'assurés servira ensuite à calculer les différents coefficients.

Pour le GLM et le Machine Learning

Lorsqu'on génère un modèle linéaire généralisé ou Machine Learning, on divise l'échantillon en deux :

- L'échantillon d'entraînement est la partie des données que le modèle va utiliser pour apprendre et trouver un lien entre les données. Il servira à calibrer et valider le meilleur modèle.
- L'échantillon de test contient l'autre partie des données, et sert quant à lui à vérifier que le modèle est en mesure de prédire des réponses sur lesquelles il n'a pas été entraîné.

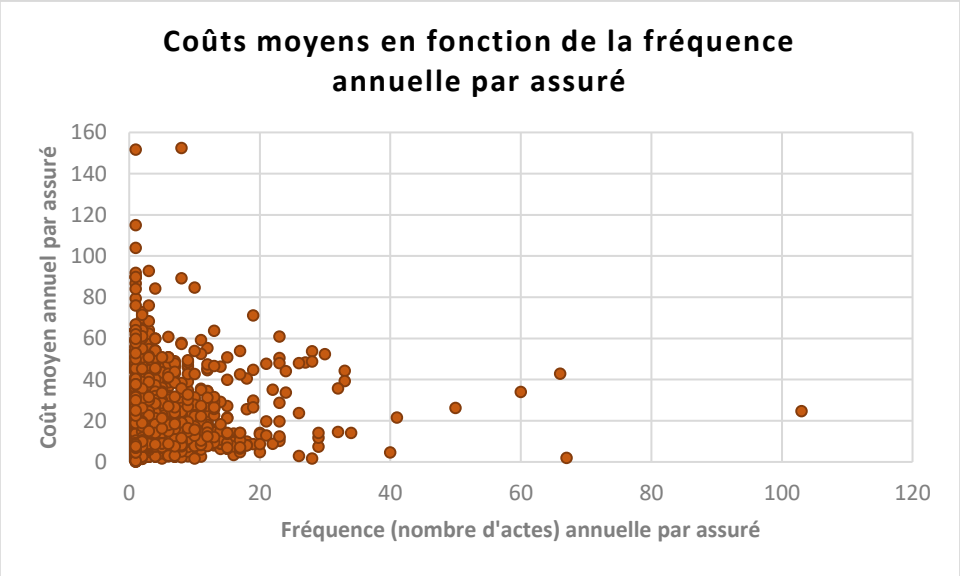
Notre base de données globale est donc divisée en deux échantillons (80% de données d'entraînement et 20% de données de test), pour effectuer ce type de modélisation.

3.1.4.c. Indépendance Fréquence Vs Coûts moyens

La tarification fréquence x Coût moyen présente l'avantage d'être simple à effectuer et permet une bonne estimation du risque, notamment pour les modèles GLM. En revanche, elle a pour inconvénients de ne pas permettre l'évaluation simple d'indicateurs de risques tels que la Value-at-Risk ou l'écart-type. De plus elle se base sur l'hypothèse forte d'indépendance entre la fréquence et le coût moyen, même si dans la réalité cette hypothèse n'est pas toujours vérifiée, notamment pour la variable liée aux régimes sur-complémentaires (variable BASE_OPTION dans notre modèle). Ces régimes sont anti sélectifs, et génèrent un effet fréquence sur le régime de base dont ils augmentent la consommation.

Autrement dit, plus une personne est couverte, plus elle consomme, et plus elle consomme souvent. Mais ce phénomène est surtout remarquable pour les postes Optiques, Dentaire, Aides auditives et Médecine Douce par exemple.

Nous testons donc cette hypothèse en représentant le coût moyen annuel des soins consommés chez le spécialiste Non OPTAM en fonction du nombre d'actes annuel (la fréquence) :



13- Analyse des coûts moyens du soin Spécialiste Non OPTAM en fonction de la fréquence pour chaque assuré

Les données sont globalement assez réparties uniformément. Une légère tendance se remarque: les données sont un peu réparties le long des axes du graphique et principalement le long de l'axe des ordonnées : plus le coût est élevé plus la fréquence est faible.

Au vu des résultats sur notre base de données, nous pressentons deux explications à cette répartition :

- Lorsque le coût est élevé, le comportement de l'assuré tend à contrôler la fréquence, notamment pour des raisons de niveaux de garanties inférieurs à ces coûts de consultation ou pour le problème d'avance des frais des assurés.

- Un assuré peu également voir avec une fréquence exceptionnelle, non régulière un ou plusieurs médecins spécialistes qui ont une expertise particulière comme les cardiologues, endocrinologue, chirurgien… dont les consultations sont plus élevées que les autres.

Pour confirmer notre intuition de corrélation entre les deux variables, nous calculons le coefficient de corrélation empirique entre la fréquence et le coût moyen, noté Cor(f,C):

$$Cor(F;C) = \frac{Cov(F;C)}{\sigma_F \sigma_C}$$

Où :

• F représente la variable aléatoire fréquence,

• C, la variable aléatoire coût moyen par acte,

• σ_F et σ_C représentent les écarts types des variables F et C

• $Cov(F;C)$ représente leur covariance

$$Cov(F;C) = \sum_{i=1}^{Ni} \sum_{k=1}^{Nk} f_i c_k P(F = f_i \; et \; C = c_k) - E(F)E(C)$$

Où :

• f_i , la valeur de la fréquence pour l'acte i,

• c_k, la valeur du coût moyen par acte pour l'acte k,

Pour nos données, on trouve les valeurs empiriques suivantes :

$$Cov\,(F;\,C) = \,-18{,}24$$

$$\sigma_F = 12{,}75$$

σ_C=15,61

Soit $Cor(F;C) = -0{,}092$

Ce résultat implique que F et C ne sont pas tout à fait indépendantes. De plus, le résultat confirme notre observation visuelle qui semblaient montrer que la fréquence décroissait avec le coût moyen croissant.

La faible valeur de la corrélation nous incite à la négliger, nous effectuons pour cela un test d'hypothèse de corrélation nulle pour confirmer l'indépendance voulue entre la fréquence et le coût moyen.

<u>Test d'indépendance du Khi2:</u>

Il s'agit d'un test qui détermine si la valeur observée d'une variable dépend de la valeur observée d'une autre variable.

L'hypothèse nulle (H_0) de ce test est la suivante : les deux variables X et Y sont indépendantes.

En termes de valeur p, l'hypothèse nulle est généralement rejetée lorsque $p \leq 0{,}05$.

L'analyse des corrélations par le critère du Khi2 entre les variables numériques (fréquence et coût moyen) confirme l'indépendance de ce type de variables entre elles (<5%).

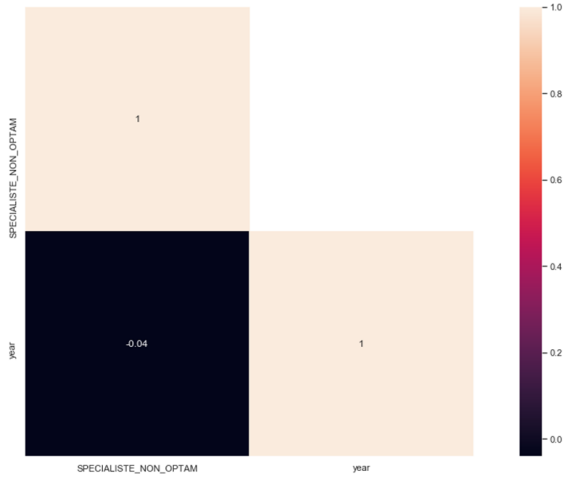
On trouve p= 0,0134, ce qui tend à accepter l'hypothèse d'indépendance.

3.1.4.d. Vérification de l'indépendance des variables explicatives

Les variables explicatives ayant été fixées par contrainte de l'outil de tarification, nous n'avons pas remis en question leur liste et leur exhaustivité. Nous vérifions cependant qu'elles ne sont pas corrélées entre elles en calculant un diagramme de corrélation sur les variables numériques représentant les niveaux de garantie et âge des assurés.

Test d'indépendance du Khi2:

L'analyse des corrélations par le critère du Khi2 entre les variables numériques (âge et niveau de garantie) confirme l'indépendance de ce type de variables entre elles (<5%).



14- Corrélation obtenues avec le Khi2 entre les variable numériques

Pour les variables qualitatives, nous analysons l'intensité des liaisons par le critère V Cramer, le critère du Khi2 ne permettant pas à lui seul de quantifier ce lien.

Le test de Cramer compare les variables 2 à 2 et donne une valeur comprise entre 0 et 1 qui montre l'intensité du lien qui existe entre les variables.

La formule mathématique du V de Cramer est la suivante:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{X^2/n}{\min(k-1, r-1)}}$$

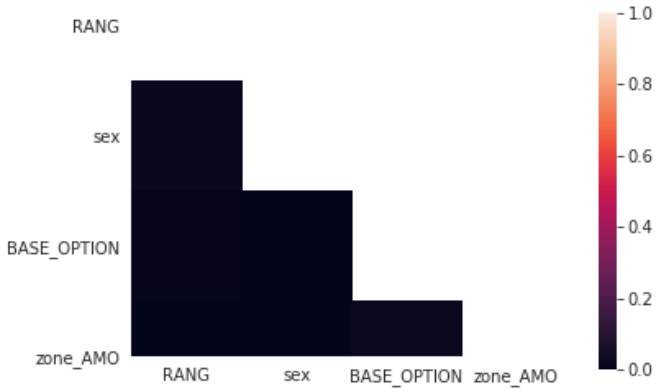
avec n_{ij} nombre de fois qu'apparaissent (A_i, B_j)) K le nombre de colonnes, r le nombre de lignes, et X (Khi2) tel que :

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$$

Résultats :

	RANG	sex	BASE_OPTION	zone_AMO
RANG	1.00	0.02	0.01	0.00
sex	0.02	1.00	0.00	0.00
BASE_OPTION	0.01	0.00	1.00	0.03
zone_AMO	0.00	0.00	0.03	1.00

15-Corrélations obtenues avec le test de Cramer pour les variables qualitatives



16-Corrélations entre les variables explicatives catégorielles

Ici aussi les variables explicatives sont bien indépendantes entre elles.

Nos variables explicatives sont donc bien considérées comme indépendantes les unes des autres, et nous ferons les différents types de modélisation avec cette même base de données, ce qui facilitera les comparaisons en les limitant à des mesures et critères statistiques objectifs.

3.1.5. Bilan données avant utilisation pour modélisation

- Notre étude part de la base de donnée globale (que nous appellerons base de données Antenia Globale).
- Pour le calcul des primes pures via la modélisation par la méthode Antenia, nous utilisons uniquement l'échantillon composé des assurés ayant les caractéristiques suivantes, qu'on appellera par la suite **assurés types**:

  - Assurés principaux ou conjoints
  - de sexe masculin et féminin
  - Ayant entre 35 et 45 ans
  - étant assurés au moins 90 jours risques consécutifs
  - habitant en zone 1 ou zone 2
  - assurés sur un contrat de base uniquement (pas d'option)

Rappel : Pour les études faites avec un GLM et un Machine Learning, nous avons supprimé les données liées aux enfants car nous ne disposons pas d'informations sur leur âge, et fixer un âge

moyen estimé fausserait l'analyse de la variable âge globale de l'échantillon. Cela change donc l'âge moyen calculé dans les différents échantillons.

Bilan des données utilisées :

		Antenia - Global	Antenia – Assuré Type	GLM & ML
Age moyen	Global	42	40	41,44
	Hommes	44	40	44
	Femmes	40	40	40

17- Age moyen entre les différentes bases de données

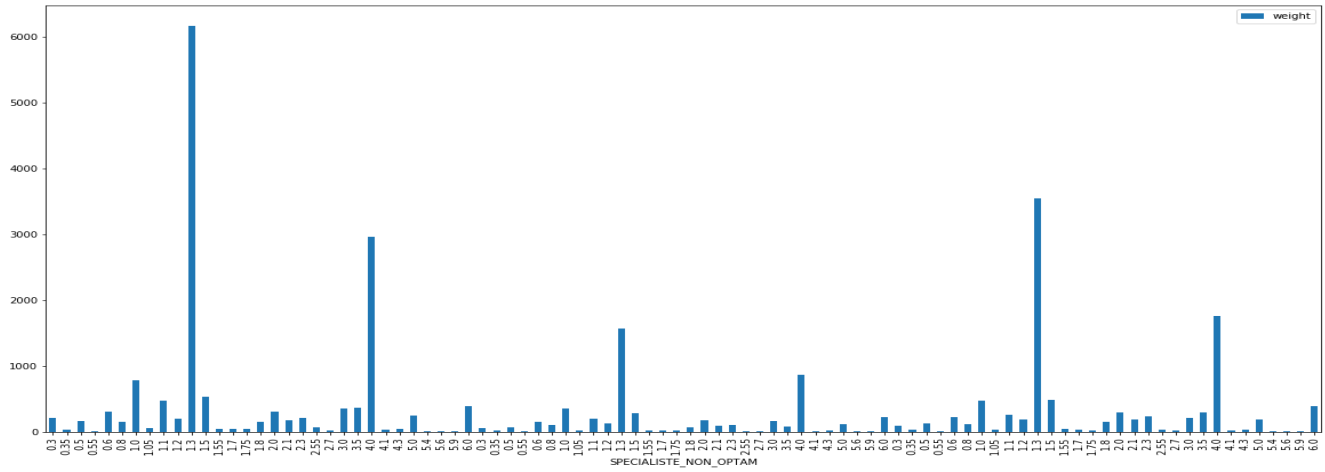
Les assurés sont représentés selon la répartition suivante :

Variable	Modalité	% BDD GLOBALE	% BDD Assuré Type	% BDD GLM - ML
SEXE	Masculin (M)	59,80%	53,49%	50,10%
	Féminin (F)	40,20%	46,51%	49,90%
TYPE BENEFICIAIRE	Adhérent (A)	50,30%	66,4%	72,10%
	Conjoint (C)	16,90%	33,6%	27,90%
	Enfant (E)	32,70%		
ZONE	Zone 1 (Z1)	35,30%	41,04%	33,80%
	Zone 2 (Z2)	54,70%	58,96%	55,90%
	Zone Alsace-Moselle (AMO)	10,10%		10,40%
BASE_OPTION	Base (B)	44,70%	49,70%	49,20%
	Option 1 (O1)	43,10%	42 ,5%	39,50%
	Option 2 (O2)	11,90%	7,70%	11,20%
	Option 3(O3)	0,20%	0,10%	0,10%

18-Répartition des assurés par variable explicative catégorielle

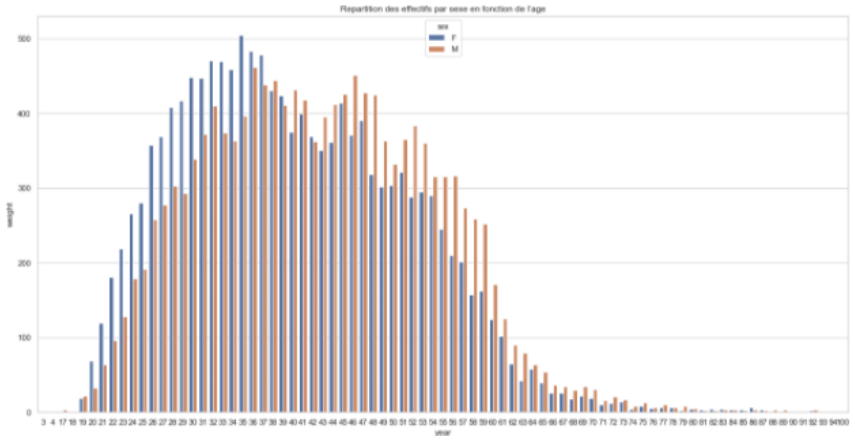
Description des données globales :

Répartition par niveau de garanties



19- répartition des assurés par niveau de garantie (en BR)

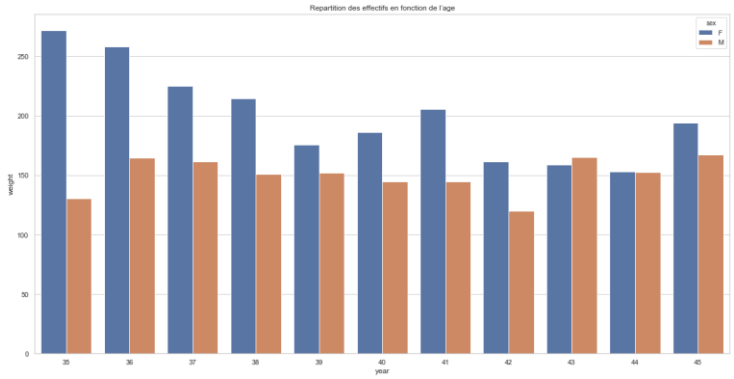
Répartition Age/Sexe :



20- Répartition âge/sexe des assurés de la BDD globale

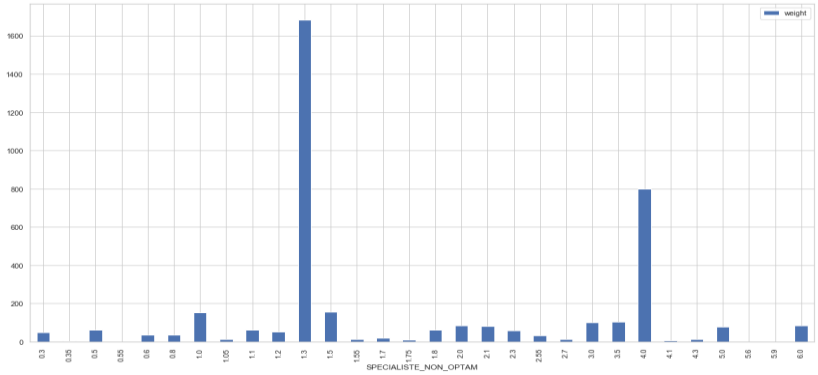
Etude de l'échantillon central (assuré type):

Répartition des effectifs en fonction de l'âge et du sexe :



21- Répartition âge/sexe de l'échantillon d'assurés types

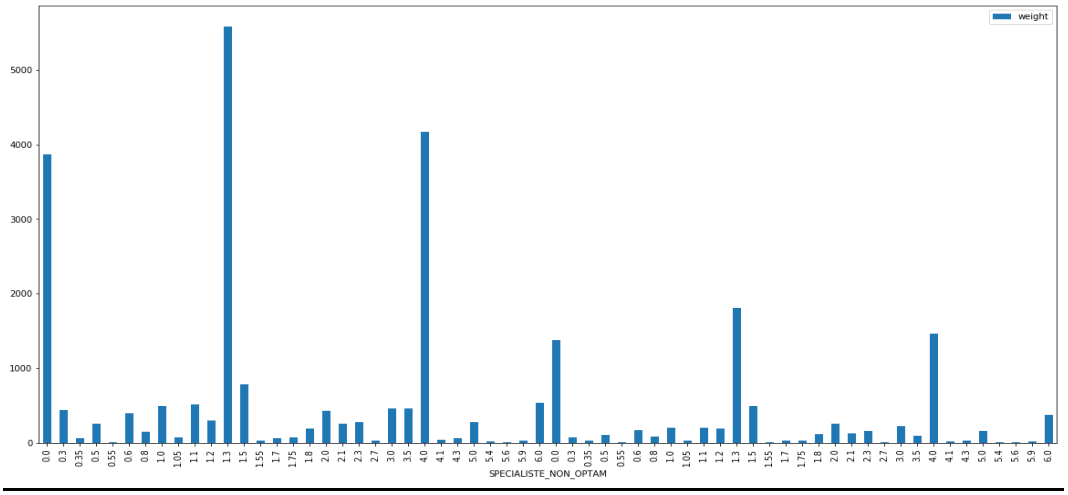
Répartition des assurés par niveau de garantie :



22- Répartition des assurés types par niveau de garantie

Echantillon hors enfant (GLM et Apprentissage supervisé) :

L'âge moyen global est de : 41,4 années



23- Répartition des assurés de la BDD du GLM et de l'apprentissage supervisé par niveau de garantie

Les échantillons utilisés lors des différentes tarifications ne sont pas les mêmes, mais ils sont répartis de manières homologues. Il n'y a pas de différences notables qui pourrait créer un biais évident. Nous pouvons commencer les modélisations.

3.2. Modélisation par via la méthode Antenia (méthode en place chez Swiss Life)

3.2.1. Résultats bruts du Spécialiste Non OPTAM

Construction de la prime pure :

Durant cette étape de construction, nous nous sommes confrontés à plusieurs problèmes :

- D'une part les garanties de type OPTAM représentant une faible part de l'échantillon (5% des prestations environ), leur grille de tarif sont parfois incohérentes : courbe non monotones, et/ou ne respectant pas l'ordre du zonage.
- Mais ces effets sont parfois également observés sur les garanties NON OPTAM.

Les règles de remboursements imposées peuvent expliquer ces résultats : alors que les garanties OPTAM sont les plus élevées, les prestations associées à ce type de soins sont à des coûts « faibles » puisqu'à dépassement limité par définition. Au contraire, les soins non OPTAM sont limités en niveau de garantie (pour les contrats de base responsables principalement) mais les prestations associées ne sont pas limitées en dépassement d'honoraires.

On obtient parfois des grilles avec des niveaux à poids très forts (c'est-à-dire avec beaucoup d'assurés) qui déséquilibrent les autres niveaux et prestations. C'est notamment le cas du point 200%BRSS (ou 130%BR), qui correspond à la limite d'encadrement du contrat responsable : les contrats collectifs sur mesure qui sont plutôt haut de gamme, ont alors tendance à tous avoir ce niveau de couverture.

Quant à la faible représentation des soins à tarif maîtrisés, elle peut être expliquée par deux principaux facteurs : notre portefeuille pourrait ne pas être représentatif des consommations moyennes en France (peu probable compte-tenu de la taille de l'échantillon et de notre cible de clients) ou la proportion de médecins adhérents à l'OPTAM/OPTAM-CO pouvait être faible en 2018 ou mal renseignée.

La garantie maternité, qui avait un faible nombre de points dans notre échantillon, avait une courbe très difficilement exploitable, avec points de niveaux très élevés (>400%BR par exemple) à des coûts moyens très faibles par rapport aux points de niveau très bas.

Les tarifs segmentés par zone n'ont parfois pas donné un résultat satisfaisant. En effet, d'après notre modèle de zonage, les tarifs de la zone 1 (Z1) doivent être supérieurs ou égaux aux tarifs de la zone 2 (Z2), eux-mêmes supérieurs ou égaux à ceux de la zone unique (ZU).

Ces mauvais résultats peuvent être expliqués également par un faible échantillonnage de certaines garanties, notamment les garanties de type OPTAM, mais il est aussi possible que notre modèle de zonage soit à rafraichir également.

Nous décidons tout de même de conserver ce zonage.

3.2.2. Lissage/Retraitement des résultats

Lissage :

Un travail manuel de lissage a été réalisé en plusieurs temps pour corriger ces problèmes, sur les courbes OPTAM comme sur les non OPTAM. Tout en essayant de conserver le maximum d'information des grilles des primes pures obtenues, nous avons ajouté les contraintes suivantes pour arriver à la courbe finale

- Les tarifs doivent respecter les ordres suivants en fonction des zones

$$PP_{zone1} \geq PP_{zone\ unique} \geq PP_{zone2}$$

- Les tarifs de toutes les zones doivent être égaux pour le niveau du ticket modérateur (ou 100%TM égal à 30%BR en soins courants, 20%BR en hospitalisation et maternité).

$$PP_{zone1}(100\%\text{TM}) = PP_{zone\ unique}(100\%\text{TM}) = PP_{zone2}(100\%\text{TM})$$

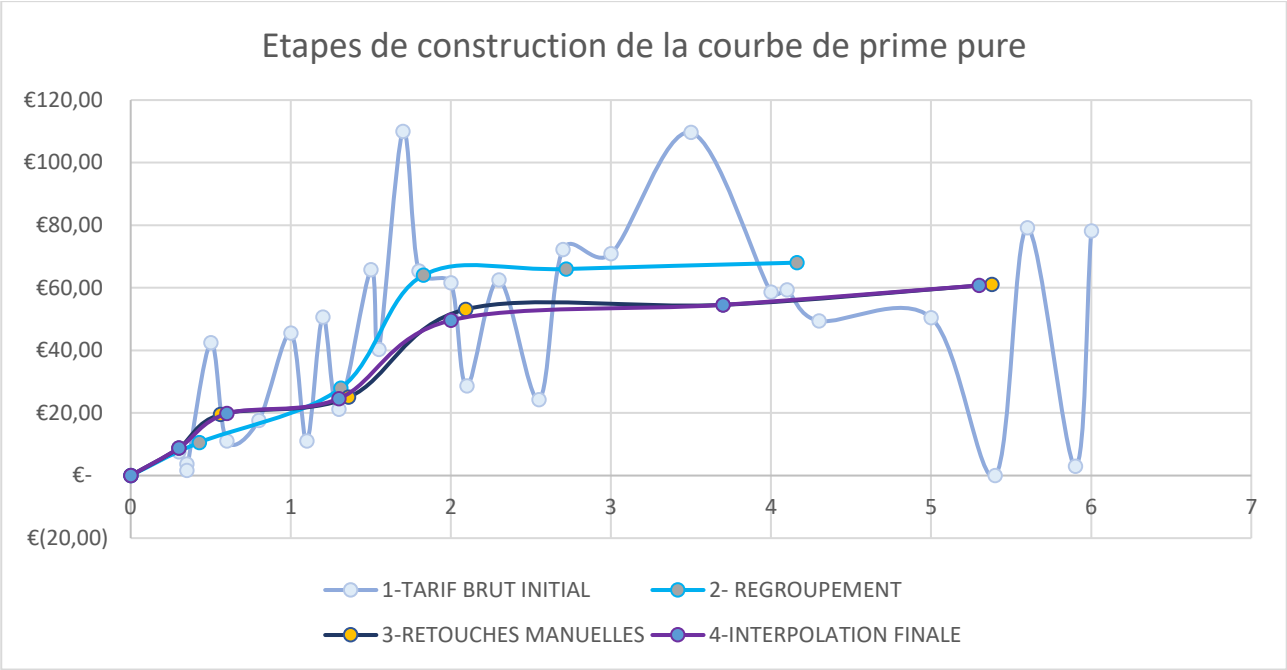
- Les courbes des tarifs doivent être croissantes avec le niveau de garantie

A partir des points sélectionnés dans la grille décrite au paragraphe 3.2.1., nous avons ajusté les courbes en fonction de ces contraintes.

Puis nous avons défini une liste constante de niveaux %BR qui constituera les points de notre grille finale de tarif : 0%, 30%, 100%, 150%, 200%, 300%, 400%, 530%. Les tarifs associés à ces points ont été calculés par interpolations linéaires à partir de la courbe déjà lissée.

Cette étape peut paraître approximative car elle biaise parfois beaucoup les informations récupérées dans notre étude, pourtant il est essentiel que nos courbes respectent ces contraintes, car elles représentent un tarif. Il serait par exemple incohérent de vendre un niveau A de garantie plus cher qu'un niveau B de niveau inférieur à A.

Nous obtenons ainsi des courbes en zone 1, zone 2 et zone unique pour chaque garantie et chaque type de soins OPTAM ou non OPTAM avec une forme cohérente.

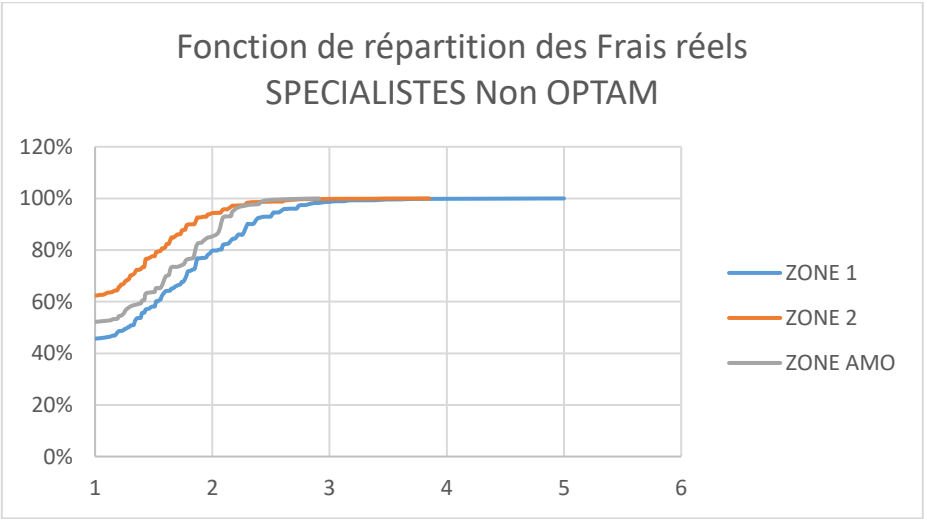


24- les différents retraitements de la prime pure

Coefficients de régime :

Par soucis de cohérence, on choisit de ne pas utiliser de coefficients supérieurs à 1 car la garantie en régime général est nécessairement supérieure à la garantie identique en régime Alsace – Moselle, puisque par définition ce dernier régime rembourse à un taux plus élevé. L'observation de coefficients supérieurs à 1 vient du fait que les tarifs moyens des prestations en Alsace-Moselle sont probablement légèrement supérieurs à ceux des autres départements de la zone 2.

Ce fut le cas notamment pour la garantie spécialiste non OPTAM.



25- Fonction de répartition des frais réels des soins médecins spécialistes non OPTAM

4.2.3. Résultats et analyse de l'évolution du tarif

Nous avons calculé les primes pures associées aux différentes garanties du back-testing. D'une manière générale, on constate que les estimations faites lors de la mise en place du C.A.S n'étaient pas forcément toutes correctes. En effet, les proportions de consommation de soins OPTAM / non OPTAM en 2018 sont très différentes, tout comme les allures des courbes de tarif de primes pures.

Les coefficients d'assiette (BRSS) et de régime sont eux restés relativement homogènes.

1. Répartition des consommations OPTAM / NON OPTAM

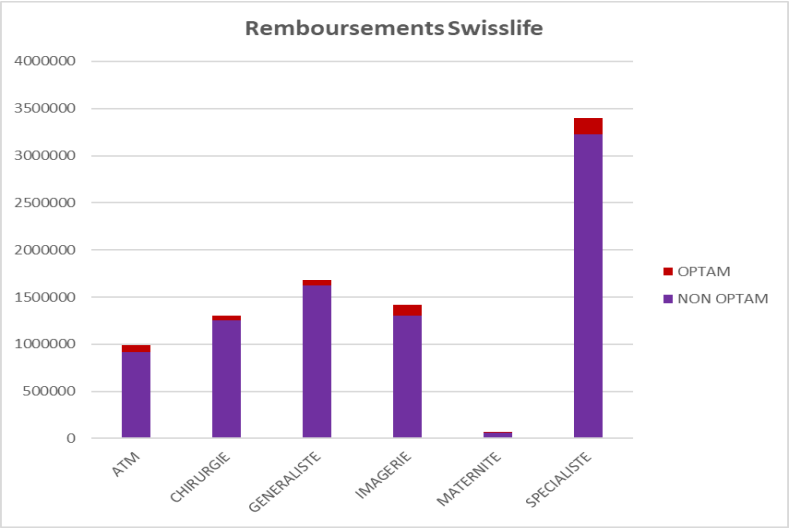
Lors de la mise en place du C.A.S, nous avions estimé que 65% des soins des médecins seraient consommés chez des praticiens ayant adhéré à l'OPTAM/OPTAM-CO.

Selon des données communiquées par l'assurance maladie et la DREES (Direction de la recherche, des études, de l'évaluation et des statistiques), la part des médecins adhérents à l'OPTAM 18% des prestations globales des soins Spécialistes sont des dépassement d'honoraires [5] [6] .

Pourtant, les résultats que nous récupérons de notre portefeuille en gestion interne et du portefeuille en gestion déléguée montrent des résultats bien différents :

	% Prestions OPTAM
ATM	7%
CHIRURGIE	4%
GENERALISTE	3%
IMAGERIE	8%
MATERNITE	7%
SPECIALISTE	5%

26- Taux de prestations de la BDD globale consommées en soin OPTAM



27- Répartition des montants des remboursements SLPS en soins OPTAM/ Non OPTAM

On constate que les soins OPTAM sont minoritaires dans la proportion de soins consommés.

La répartition est relativement homogène entre les types de garanties (entre 3% et 8% des prestations remboursées par Swiss Life).

Les consultations et soins généralistes sont étonnamment ceux qui sont le moins représentés par des praticiens adhérant à l'OPTAM/OPTAM-CO. Ceci peut paraitre étonnant car c'est aussi la garantie qui contient le moins de dépassement d'honoraires.

Cela est notamment dû au fait que dans les systèmes d'information, seuls les médecins spécialistes de secteur 2 ayant adhéré à l'OPTAM sont catégorisés en tant que tels, mais ceux qui n'ont pas besoin d'y adhérer (par exemple les médecins des secteurs 1) sont catégorisés Non OPTAM.

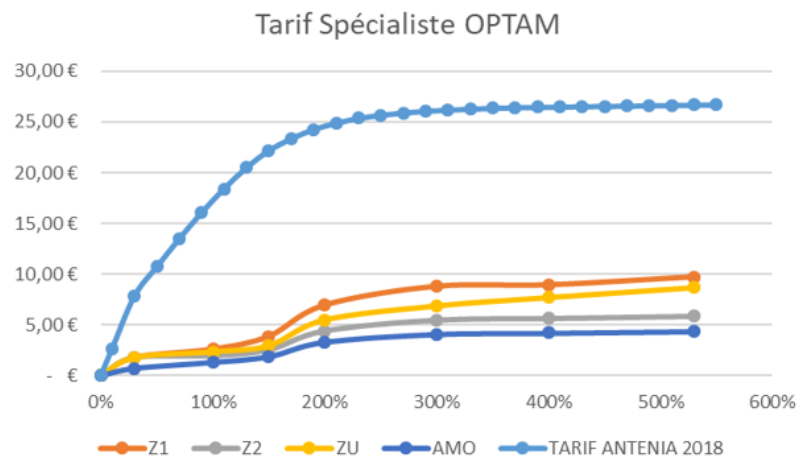
Cela explique la différence de proportions entre ce qui était attendu et ce qui fut observé en réalité, mais faute de meilleures informations nous conserverons ces classifications pour effectuer notre back-testing.

Il reste alors à analyser le montant des tarifs et l'allure globale de ces courbes.

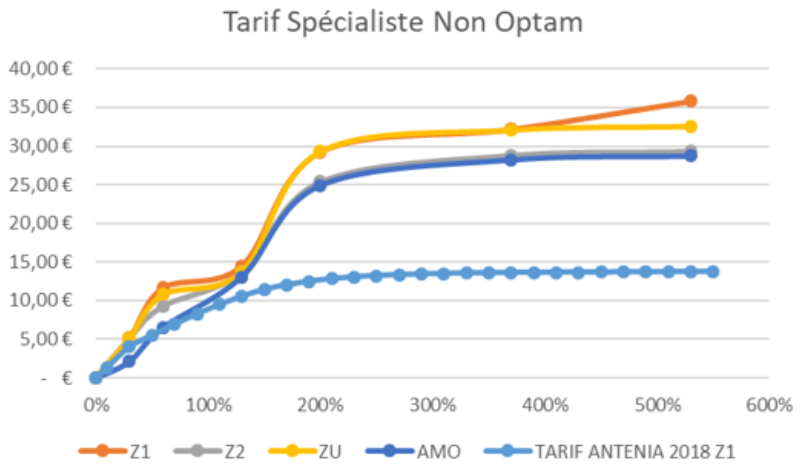
2. Courbes des tarifs

Les tarifs finaux après application des coefficients sont différents des courbes actuellement en production chez Swiss Life :

Les tarifs non OPTAM modélisés sont à l'inverse très supérieurs aux tarifs actuels.



28-Tarif de la garantie Médecin Spécialiste OPTAM avant et après back-testing sur plusieurs zones distinctes



29-Tarif de la garantie Médecin Spécialiste non OPTAM avant et après back-testing sur plusieurs zones distinctes

Cette différence est principalement expliquée par la très forte proportion de soins NON OPTAM par rapport à celle qui était attendue.

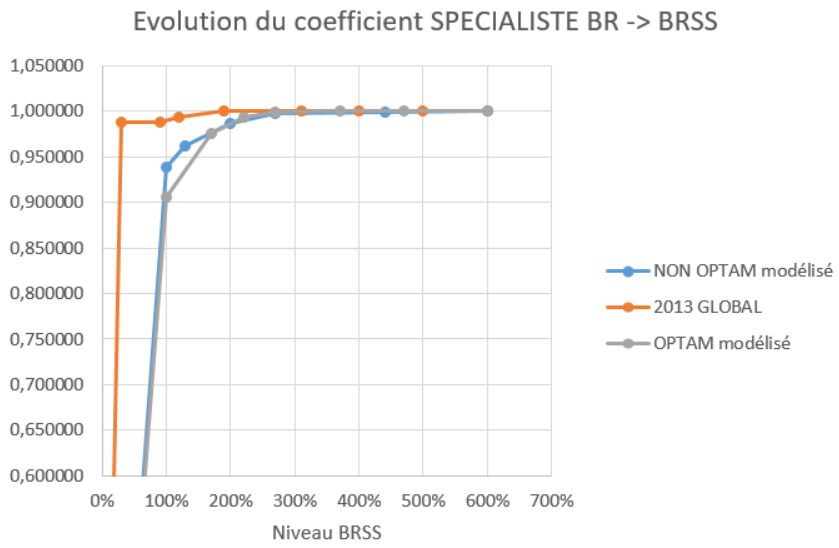
Les différences se compensent au global quand on additionne les tarifs OPTAM et Non OPTAM pour obtenir le tarif global. On observe cependant que le tarif modélisé est inférieur à la courbe en place pour les faibles niveaux de garantie (<200%BR) tandis qu'elle devient supérieure pour les hauts niveaux de garantie.

On remarque également que la forme de la courbe change par rapport à la courbe actuellement en place : en effet, contrairement à l'ancienne courbe dont l'allure était concave, on observe dans le tarif modélisé un point d'inflexion pour les niveaux de garantie autour de 150%BR.

3. Coefficient d'assiette

Le tarif en assiette BRSS est calculé à partir du tarif en BR via un coefficient de passage. Décrit dans le paragraphie 2.2.2.

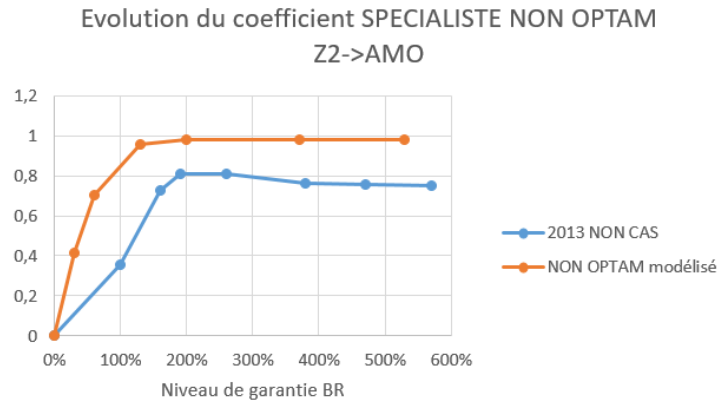
Comme attendu, les coefficients obtenus sont du même ordre de grandeur que ceux qui avaient été calculés en 2013, tous très proche de 1. Ceci montre que la proportion des taux réels de remboursements de la sécurité sociale est restée globalement la même.



30-Evolution du coefficient BR-> BRSS pour la garantie Spécialiste

4. Coefficient de régime

Les coefficients de passage de la zone 2 à la zone Alsace-Moselle sont changeants selon les garanties. Les allures des courbes restent globalement stables entre 2013 et aujourd'hui.



31- Evolution du coefficient de zone pour la garantie spécialiste non OPTAM

5. Options

On constate que dans notre échantillon d'assurés, les prestations de base sont consommées uniformément (environ 80%) pour toutes les garanties et de tous types, sauf pour la maternité donc les prestations de base constituent presque la totalité des prestations globales.

Les prestations liées à des contrats de niveau 1 d'option représentent environ 18% des prestations globales, et celles de niveau 2 d'option environ 4%.

Les contrats à 3 options sont très minoritaires et donc très dilués dans l'échantillon des assurés principaux et conjoints de 35-45 ans, leur représentation a un poids quasi nul dans notre modélisation.

Les coefficients d'option finaux calculés ont tous la même tendance : les garanties OPTAM ont un coefficient proche de 99% tandis que les non OPTAM ont un coefficient proche de 95%.

Cela diminue encore le niveau global des tarifs OPTAM et accentue leur écart par rapport aux tarifs non OPTAM.

4.2.4. Conclusion

La mise à jour des tarifs a mis en lumière la proportion très élevée de soins NON OPTAM par rapport à ce qui était espéré. La conséquence de cette proportion est l'augmentation significative du tarif associé à la garantie.

Mais lorsqu'on ajoute le tarif lié aux consultations OPTAM, Le tarif Spécialiste global est globalement stable.

Ces résultats sont cependant à remettre dans le contexte particulier de notre étude, qui ne prend pas toutes les variables en considération, puisque nous avons figé les coefficients âge/sexe, bénéficiaire, d'option et le zonage déjà paramétré dans notre outil de tarification.

L'ordre de grandeur du tarif obtenu est satisfaisante, nous cherchons maintenant à savoir s'il est vraiment fiable. Et plus largement, si notre technique de mise à jour de tarif est une méthode fiable, et éventuellement d'en évaluer les limites.

Pour répondre à cette question, nous allons tarifer cette même garantie avec deux types de modélisations techniques régulièrement utilisées dans le domaine de la tarification santé : un Modèle Linéaire Généralisé (GLM) et un modèle Machine Learning (ML).

3.3. Modélisation via une méthode GLM

3.3.1. Principe théorique de la tarification [7]

La construction d'un tarif en assurance santé est classiquement faite via un calcul des primes pures par une approche fréquence x coût moyen.

Les modèles de régression de type GLM (Generalized Linear Models) sont souvent sélectionnés pour la modélisation des tarifs santé car ils sont adaptés à la tarification d'une variable dépendant de plusieurs facteurs, ou variables explicatives.

Leur estimation n'est fiable que si certaines hypothèses fortes sont préalablement reconnues. Si nous considérons un échantillon de deux séquences de variables aléatoires : $(X_i)_{1\leq i\leq n}$ et $(N_i)_{1\leq i\leq A}$, alors ces hypothèses sont :

- L'indépendance et l'identique distribution des variables aléatoires Xi représentatives des « Coûts moyens de consommation » d'un assuré i.

- L'indépendance entre les variables aléatoires Xi représentatives des « Coûts moyens de consommation » et les variables aléatoires Ni représentatives des « Fréquences de consommation annuelles » relativement à un assuré i.

Soit A Le nombre total d'assurés exposé au risque santé. Alors, le nombre de consommations total en biens et services de santé est :

$$N = \sum_{i=1}^{A} \mathrm{N_i}$$

Et le coût total de toutes les consommations est :

$$S = \sum_{i=1}^{N} \mathrm{X_i}$$

L'espérance du coût est alors définie comme suit :

$$E(S) = E\left(\sum_{i=1}^{N} \mathrm{X_i}\right)$$

$$E(S) = \sum_{k=1}^{\infty} \Pr(\mathrm{N} = \mathrm{k}) * E\left(\sum_{i=1}^{k} \mathrm{X_i}\right)$$

$$E(S) = \sum_{k=1}^{\infty} \Pr(\mathrm{N} = \mathrm{k}) * k * E(\mathrm{X_1})$$

$$E(S) = E(N) * E(\mathrm{X_1})$$

Notre approche consiste donc percevoir la prime pure π qui reflète le coût annuel moyen de toutes les consommations en santé, égale au produit du nombre moyen de consommations et de leur coût moyen.

$$\pi = E(S) = E(N) * E(\mathrm{X_1})$$

On en déduit la variance de la prime pure, égale à la variance de la variable aléatoire S définie par :

$$\mathrm{Var}(S) = \mathrm{Var}\big(E(S|N)\big) + \mathrm{E}\big(Var(S|N)\big)$$

Sachant que :

$$E(S|N) = \mathrm{E}\left(\sum_{i=1}^{N} \mathrm{X_i}\,|N\right)$$
$$E(S|N) = N * \mathrm{E}(\mathrm{X_1})$$

Et que :

$$Var(S|N) = \mathrm{Var}\left(\sum_{i=1}^{N} \mathrm{X_i}\,|N\right)$$
$$= \mathrm{Var}\left(\sum_{i=1}^{N} \mathrm{X_i}\right) \text{ Par indépendance de N et } X_i$$
$$= \sum_{i=1}^{N} \mathrm{Var}(\mathrm{X_i}) \quad \text{Par indépendance des } X_i$$
$$= N * \mathrm{Var}(\mathrm{X_1})$$

Donc :

$$Var(E\big((S|N)\big) = \mathrm{Var}\big(N * \mathrm{E}(\mathrm{X_1})\big) = E^2(X) * Var(N)$$

En conclusion :

$$Var(S) = E^2(X) * Var(N) + E(N) * var(X)$$

La variance du coût total des consommations dépend donc de deux composantes : la fréquence annuelle de consommation et le coût de consommation.

3.3.2. Modèles linéaires généralisés

3.3.2.a. Modèle linéaire gaussien

Soit un n-échantillon d'une variable aléatoire cible $(Y_i)_{1\leq i\leq n}$. Le modèle linéaire gaussien est un modèle statistique qui sert à expliquer une variable aléatoire Y par k variable(s) exogène(s) Xi supposées déterministes (k ϵ [1 : n]) :

$$Y = \alpha_0 + \sum_{i=1}^{N} \alpha_i \mathrm{X_i} + \varepsilon \text{ avec } \alpha_i \text{ les paramètres du modèle à estimer}$$

ε est la variable aléatoire non observée qui correspond au terme d'erreur du modèle (écart entre valeur observée et valeur estimée) et pour laquelle on pose les hypothèses suivantes :
ε suit une loi normale telle que :

- $E(\varepsilon) = 0\ (centralité\ des\ erreurs\ distinctes)$
- $Var(\varepsilon) = \sigma^2 (homoscédasticité\ des\ erreurs\ distinctes)$ où σ^2 une constante inconnue à estimer
- $Cov\big(\varepsilon_i, \varepsilon_j\big) = 0\ \forall i \neq j$ *(Absence de corrélation entre les erreurs distinctes)*

Il en découle l'ensemble des caractéristiques suivantes sur Y :

$$E(Y) = \alpha_0 + \sum_{i=1}^{N} \alpha_i X_i$$

$$Var(Y) = \sigma^2$$

$$Cov(Y_i, Y_j) = Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall i \neq j$$

Une des méthodes d'estimation classique du vecteur des paramètres α_i est celle du maximum de vraisemblance, que nous utiliserons pour notre modélisation. Son principe est de maximiser la fonction de vraisemblance par rapport aux paramètres α_i ce qui revient à résoudre une système d'équations. Les solutions trouvées sont les estimateurs de vraisemblance des paramètres.

3.3.2.b. Modèle linéaire généralisé [8]

Le modèle « fréquence-coût moyen » que nous allons utiliser nous amène à modéliser deux variables aléatoires : le nombre d'occurrence d'un sinistre et le coût moyen des sinistres.

Pour cette modélisation nous proposons d'utiliser les modèles linéaires généralisés. Ils sont fréquemment choisis grâce à leur flexibilité vis-à-vis des hypothèses sur la loi de la variable aléatoire à expliquer. Ils permettent en effet de généraliser les modèles linéaires gaussiens lorsque la variable aléatoire suit une autre loi que la loi gaussienne.

L'autre qualité de ces modèles est leur adaptabilité aux erreurs corrélées entre elles (hétéroscédasticité des erreurs). Ces modèles sont conciliables avec la réalité des distributions des fréquences et des coûts annuels.

Le principe est le même que le modèle linéaire gaussien, mais il s'agit de modéliser une fonction de son espérance appelée fonction de lien notée *g*, au lieu de modéliser la variable aléatoire à expliquer.

Les modèles linéaires généralisés établissent la relation suivante entre la transformation de l'espérance de la variable expliquée et la combinaison linéaire des variables prédictives :

$$g(E(Y_i)) = \alpha_0 + \sum_{i=1}^{k} \alpha_i X_i$$

Soit

$$E(Y_i) = g^{-1}(\alpha_0 + \sum_{i=1}^{k} \alpha_i X_i)$$

Les fonctions de lien les plus fréquemment utilisées en modélisation linéaire généralisée sont :

Fonction de lien	Expression
Identité	g(x)=x
Log	g(x)=log(x)
Logit	g(x)=log(x/(1-x))
Inverse	g(x)=1/x
Probit	g(x)=ϕ(x)

Famille exponentielle

Soit X une variable aléatoire réelle, dont la loi de probabilité dépend d'un paramètre $\theta \epsilon R^d$

On dit que la loi de X appartient à la famille exponentielle si et seulement si $P(X = x, \theta)$ (cas discret) ou $f_x(x, \theta)$ (cas continu) est de la forme :

$$f(\mathrm{y}, \Theta, \varphi) = c_\varphi(y) exp(\frac{\mathrm{y}.\Theta - \mathrm{b}.\Theta}{\varphi})$$

Où

θ : paramètre naturel,

a(Θ)convexe

φ : paramètre de dispersion

$c_\varphi(y)$ ne dépend pas de Θ

L'espérance et la variance d'une telle famille de lois peuvent être facilement calculées grâce à des dérivés au lieu d'intégrales :

$$E(\mathrm{Y}) = \mathrm{b}'(\Theta)$$

$$Var(\mathrm{Y}) = \mathrm{a}(\varphi).\mathrm{b}''(\Theta)$$

La famille exponentielle est nécessaire pour la mise en place d'un modèle linéaire généralisé et comprends plusieurs lois paramétriques usuelles comme la loi normale, la loi Exponentielle, la loi Gamma, la loi Poisson, la loi Binomiale, la loi binomiale négative etc…

Estimation par maximum de vraisemblance

Développée par le statisticien Ronald Aylmer Fisher en 1922, la méthode du maximum de vraisemblance est une méthode d'estimation paramétrique connue pour sa simplicité de mise en œuvre et sa faculté d'adaptation à une modélisation complexe, c'est-à-dire une loi ayant pour paramètre un réel Θ ou un vecteur de plusieurs paramètres inconnus $\Theta = (\Theta_1, \Theta_2, \Theta_3 \dots)$

Concrètement, considérons un n-échantillon aléatoire de variables $(\mathrm{Y_i})_{1 \le i \le n}$ indépendantes et identiquement distribuées selon une loi paramétrique associée à une loi de distribution connue.

Le maximum de vraisemblance repose sur l'idée de considérer Θ comme une variable réelle (ou un vecteur de variables réelles) et chercher les valeurs de Θ qui rendent l'observation des valeurs $\mathrm{y}_1, \dots, \mathrm{y}_n$ « la plus vraisemblable possible » ou maximiser les chances de réalisation de l'évènement $\{(\mathrm{y}_1, \dots, \mathrm{y}_n)\}$ qui est traduit par la probabilité $\mathrm{P}_\Theta((Y_1, \mathrm{Y}_2, \dots \mathrm{Y}_n) = (y_1, y_2, \dots, \mathrm{y_n})$.

Nous définissons la vraisemblance du modèle notée $\mathrm{L}_n(\Theta)$ par :

$$\mathrm{L}_n(y_1, \dots, y_\mathrm{n}, \Theta) = \mathrm{P}_\Theta(\bigcap_{i=1}^{n} \{Y_\mathrm{i} = y_\mathrm{i}\}) = \prod_{i=1}^{n} \mathrm{P}_\Theta(Y_\mathrm{i} = y_\mathrm{i}))$$

Si de plus la distribution en question est continue et appartient à la famille exponentielle, la vraisemblance s'écrit sous la forme :

$$\mathrm{L}_n(y_1, \dots, y_\mathrm{n}, \Theta) = \exp(\sum_{i=1}^{n} \frac{y_\mathrm{i}\Theta_{\mathrm{i}-b(\Theta_\mathrm{i})}}{a_\mathrm{i}(\varphi)} + c(y_\mathrm{i}, \varphi))$$

$$\text{Et } L = \log(\mathrm{L}_n(y_1, \dots, y_\mathrm{n}, \Theta)) = \sum_{i=1}^{n} \frac{y_\mathrm{i}\Theta_{\mathrm{i}-b(\Theta_\mathrm{i})}}{a_\mathrm{i}(\varphi)} + c(y_\mathrm{i}, \varphi)$$

La grandeur L est généralement dérivable (mais ce n'est pas toujours le cas) et facile à maximiser du fait de la simplicité du calcul de ses dérivées première et seconde par rapport aux α_i (somme de dérivés de fonctions usuelles) :

$$\frac{\partial L}{\partial \alpha_\mathrm{i}} = \log(\mathrm{L}_n(y_1, \dots, y_\mathrm{n}, \Theta)) = \sum_{i=1}^{n} \frac{y_\mathrm{i}\Theta_{\mathrm{i}-b(\Theta_\mathrm{i})}}{a_\mathrm{i}(\varphi)} + c(y_\mathrm{i}, \varphi)$$

$$\text{Sachant que } E(Y_\mathrm{i}) = g^{-1}(\alpha_0 + \textstyle\sum_{i=1}^{k} \alpha_\mathrm{i}\, \mathrm{Y_i}) = b'(\Theta)$$

En notant $\mu_\mathrm{i} = E(Y_\mathrm{i})$ et $\nu_\mathrm{i} = \alpha_0 + \sum_{i=1}^{n} \alpha_\mathrm{i}\, Y_\mathrm{i}$ et étant donné une fonction de lien *g* nous aurons

$$\mu_\mathrm{i} = g^{-1}(\nu_\mathrm{i})$$

Ce qui veut dire que :

$$\frac{\partial L}{\partial \alpha_\mathrm{i}} = \frac{\partial L}{\partial \mu_\mathrm{i}} * \frac{\partial \mu_{\mathrm{i}_\mathrm{i}}}{\partial \nu_\mathrm{i}} * \frac{\partial \nu_\mathrm{i}}{\partial \alpha_\mathrm{i}}$$

L'équation ci-dessus donne lieu à un système d'équations appelées équations de vraisemblance. Pour les résoudre, il faut faire appel à des méthodes itératives, à savoir celle de Newton-Raphson.

L'estimateur obtenu du vecteur des α_i par la méthode du maximum de vraisemblance est caractérisé par des propriétés fortes comme sa convergence et sa variance minimale. Il est asymptotiquement efficace et distribué selon une loi normale. En revanche, il peut être biaisé en échantillon fini.

Conclusion :

Afin de réaliser la modélisation linéaire généralisée il faut auparavant choisir les éléments constitutifs de cette modélisation :

- Le choix de la loi de la variable à expliquer parmi les lois de la famille exponentielle.
- Les variables explicatives les plus significatives à la variable réponse
- La fonction de lien

On appelle fonction de lien canonique, la fonction qui lie le paramètre μ au paramètre Θ.

Ci-dessus les fonctions de lien associées à quelques lois usuelles de la famille exponentielle :

Loi	Fonction de lien	Expression
Binomiale	Logit	g(x)=log(x/(1-x))
Poisson	Log	g(x)=log(x)
Binomiale négative	Logit	g(x)=log(x/(1-x))
Gamma	Inverse	g(x)=1/x
Normale	Identité	g(x)=x

32- Fonctions de lien associées aux lois usuelles de la famille exponentielle

3.3.3. Méthode de sélection du modèle

Dans cette partie nous expliquerons les étapes et les critères de décision qui nous permettront de choisir les éléments constitutifs de notre modélisation linéaire généralisé.

Ce choix est en effet une étape délicate et essentielle dans la modélisation.

Nous élaborerons notre modèle sous la forme du produit de la fréquence des consommations (nombre d'actes annuels sur une garantie) et du coût moyen des prestations sur cette garantie.

3.3.3.a Etude graphiques et tests non paramétriques

La première étape des modèles linéaires généralisés est usuellement de visualiser graphiquement l'influence des variables explicatives potentielles étudiées sur la variable à expliquer (dans notre cas, la fréquence des consommations, et le coût moyen d'une consommation d'une garantie).

Notre périmètre d'étude et notre outil de tarification nous imposent une liste de variables explicatives définie. Sans les remettre en question, nous pouvons cependant confirmer ou infirmer leur impact sur nos variables à expliquer à travers cette première étape.

Nous représentons les variables à expliquer en fonction des variables explicatives. Nous calculons également les corrélations entre toutes ces variables.

Il est également pertinent d'utiliser des graphes d'ajustement des variables cibles. Cette étape nous permettra de présélectionner la loi qui s'ajusterait potentiellement le mieux aux distributions de ces variables, tout en donnant une estimation de la pertinence de correspondance de notre distribution empirique quant à sa modélisation par un modèle théorique. Pour cela on utilisera notamment la fonction *getattr* de python.

3.3.3.b Modèles classiques de comptage

Les études de tarification en assurance santé utilisent très fréquemment une approche fréquence x coût moyen. La fréquence de consommation est habituellement modélisée en utilisant une loi de Poisson ou une loi Binomiale négative.

Cependant ces modèles de comptages classiques sont souvent remis en question lorsqu'on regarde leur application sur un échantillon réel de données, notamment à cause de leur hypothèse d'équidispersion (Poisson : égalité entre l'espérance et la variance), mais aussi par l'abondance des valeurs nulles (absence de consommation d'une grande partie des assurés) ainsi que l'éventuelle présence de valeurs extrêmes.

Nous verrons en effet dans la suite de l'étude que nous retrouvons ces problèmes d'adéquation dans notre échantillon. Les consommations en assurance santé sont en effet marquées par une sur-dispersion, notamment expliquée par l'absence de consommation de certains assurés.

Nous envisageons donc d'utiliser des modèles alternatifs de comptage : les modèles modifiés en zéro (Zero-Inflated Models) qui permettent de gérer ces caractéristiques particulières.

3.3.3.c Modèles alternatifs de comptage

Il existe deux types de modèles modifiés en zéro : le modèle Zero-Inflated Poisson (ZIP) et le modèle Zéro-Inflated Negative Binomial (ZINB). Ces deux modèles modélisent séparément :

- La présence/absence de consommation (processus binaire) : modélisation de la probabilité de consommation $\pi_{i,}$ fournie par une loi de Bernoulli
- La quantité de consommations, conditionnellement à la présence d'une consommation, par une loi de Poisson ou binomiale négative :

$$P(X_{i=}x_i) = \begin{cases} \pi_i + (1-\pi).f(0) \;\; si\; y_i = 0 \\ (1-\pi_i).f(y_i) \;\; sinon \end{cases}$$

Puis combinent ces deux modélisations.

Par la suite, la densité de distribution d'une loi ZIP s'écrit :

$$P(X_{i=}x_i) = \begin{cases} \pi_i + (1-\pi).exp(-\mu_i) \;\; si\; y_i = 0 \\ (1-\pi_i).exp(-\mu_i).\dfrac{{\mu_i}^{y_i}}{y_i!} \;\; sinon \end{cases}$$

L'espérance et la variance du modèle sont données par :

$$E(X_i) = (1-\pi).\mu_i \; et \; Var(X_i) = \; (1-\pi).(\mu_i + \pi_i\mu_i^2) = E(X_i).(1-\pi_i\mu_i)$$

Et pour la loi ZINB :

Densité de la distribution :

$$P(X_{i=}x_i) = \begin{cases} \pi_i + (1-\pi).(1+\nu\mu_i)^{-\frac{1}{\nu}} \;\; si\; y_i = 0 \\ (1-\pi_i).\dfrac{\Gamma\left(y_i + \frac{1}{\nu}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\nu}\right)}\left(\dfrac{(\frac{1}{\nu})}{\frac{1}{\nu}+\mu_i}\right)^{\frac{1}{\nu}}\left(\dfrac{\mu_i}{\frac{1}{\nu}+\mu_i}\right)^{y_i} \;\; sinon \end{cases}$$

L'espérance et la variance du modèle sont données par :

$$E(X_i) = (1-\pi_i).\mu_i \; et \; Var(X_i) = \; (1-\pi).(\mu_i + \pi_i\mu_i^2) = E(X_i).(1-\pi_i\mu_i)$$

3.3.3.d. Modèles classiques des coûts moyens

La modélisation du coût moyen doit se faire par une loi de probabilité de la famille exponentielle qui soit continue à valeur positive.

Les lois généralement utilisées pour ce genre de modèle sont les lois Log-normale (qui ne fait pas partie de la famille exponentielle mais peut être modélisée par une transformation logarithmique du coût, qui sera modélisée par une loi normale avec une fonction de lien identité), Gamma et Exponentielle.

3.3.4.e Qualité d'ajustement du modèle

Déviance résiduelle

La déviance résiduelle compare la vraisemblance maximisée du modèle à celle d'un modèle saturé.

Un modèle saturé est un modèle de mêmes composantes que notre modèle estimé en terme de distribution et fonction de lien, mais sa vraisemblance est fonction des observations y_i au lieu du paramètre $\mu = E(y_i)$

Via l'estimation d'un paramètre α_i pour chaque observation de sorte que les données soient correctement ajustées (ou « saturées »)

Elle est définie par l'écart entre les valeurs observées et les valeurs prédites. La qualité d'ajustement sera d'autant meilleure que cet écart est faible :

$$D = 2 * \phi.\{logL(y, y, \phi) - logL(y, \mu, \phi)\} = \phi * D^*$$

avec ϕ le paramètre de dispersion et D^* la déviance standardisée

Notons que D^* est une variable aléatoire qui suit asymptotiquement une loi X^2_{n-p} où :

- n est le nombre d'observations de la variable y_i

- p est le nombre de variables explicatives

Cette propriété nous permet de construire un test paramétrique pour tester la qualité d'ajustement du modèle.

Critères AIC et BIC

Le critère d'information d'Akaike (« Akaike Information Criterion » noté « AIC ») est une mesure de la qualité d'ajustement d'un modèle statistique qui pénalise le modèle en fonction du nombre de paramètres.

On note :

$$AIC = -2 * log(L) + 2 * k$$

Avec $log(L)$ la log-vraisemblance maximisée

Et k le nombre de paramètres

Plus les variables sont nombreuses et la vraisemblance minimale, plus la valeur de l'AIC est grande.

C'est un critère qui permet d'arbitrer le nombre de variables explicatives qu'on souhaite conserver et le biais du modèle qu'on chercher à minimiser notamment via l'augmentation du nombre de variables explicatives. On choisit alors le modèle dont l'AIC est le plus faible.

Néanmoins, l'AIC doit être calculé avec un échantillon dont la taille est ni trop grande, ni trop faible, au risque de valider les modèles avec de nombreuses variables explicatives. Dans un tel cas on utilise alors plutôt le critère d'information Bayésien (« Bayesian Information Criterion » ou « BIC »), qui est un dérivé de l'AIC, qui en diffère par une pénalisation plus importante, en tenant compte de la taille de l'échantillon en plus du nombre de paramètres explicatifs. Ici aussi le meilleur modèle sera celui qui minimise le critère BIC.

On note :

$$BIC = -2 * log(L) + 2 * k * \log(n)$$

Avec $log(L)$ la log-vraisemblance

k le nombre de paramètres

et n le nombre d'observations dans l'échantillon

3.3.4.f Analyse des résidus

Les résidus sont les écarts entre les valeurs $\widehat{y_i}$ prédites par le modèle et les valeurs observées y_i.

La qualité d'ajustement d'un modèle linéaire généralisé nécessite une analyse détaillée de ces résidus. Il en existe de divers types, construits par diverses transformations :

Les résidus de déviance sont définis à partir de la contribution de la i-ème observation y_i à la déviance D, notée d_i :

On note

$$r_{Di} = signe(y_i - \widehat{y_i})\sqrt{2(log(L(y_i, y_i, \phi)) - \log(L(y_i, \widehat{y_i}, \phi)))} = signe(y_i - \widehat{y_i})\sqrt{d_i}$$

Les résidus de Pearson sont définis à l'aide de la variance estimée du modèle :

$$r_{Pi} = \frac{(y_i - \widehat{y_i})}{\widehat{\sigma_{\widehat{y_i}}}}$$

3.3.4. Modélisation de la fréquence

Rappel: dans la suite de ce rapport, la fréquence est ici représentée par le nombre d'actes annuels constaté pour les adhérents et leur conjoint présents au moins 360 jours en 2018. Ils ont tous le même poids dans notre modélisation.

Les statistiques descriptives de la fréquence de notre échantillon sont les suivantes :

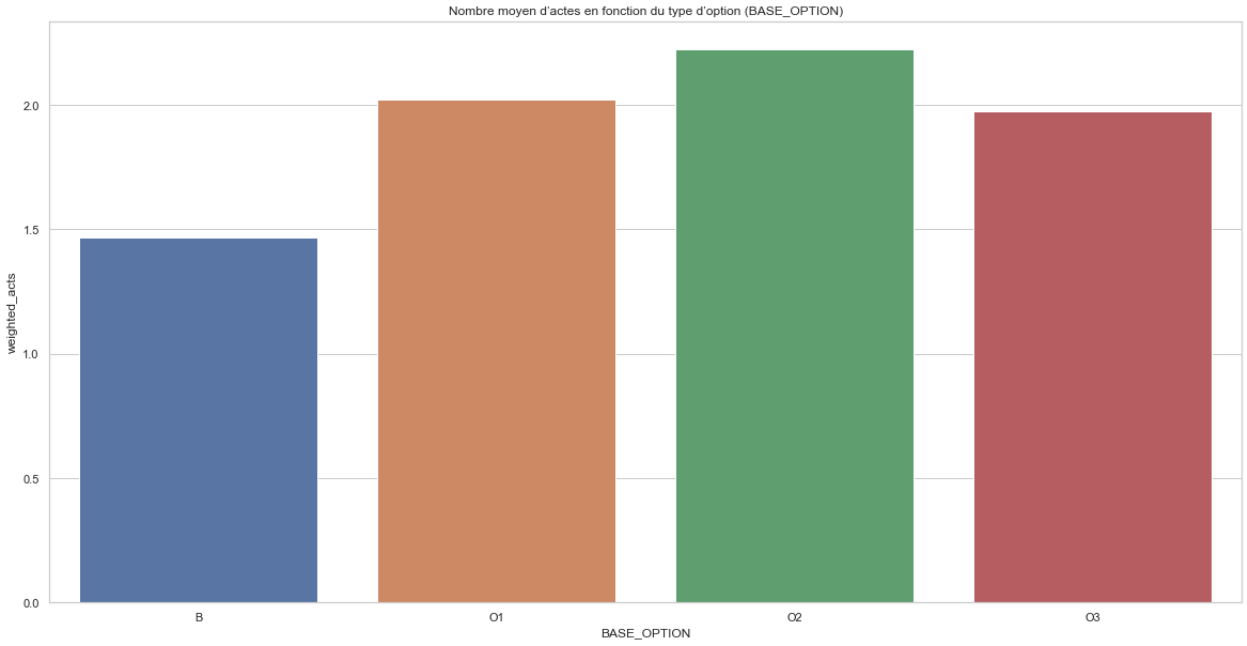
Répartition du nombre d'acte par type de niveau de garantie (base, option) :

Les assurés ayant des options consomment plus que les assurés qui sont sur une base seule, ce qui est le comportement attendu en théorie (anti sélection).

En revanche, la logique voudrait que la fréquence de consommation augmente avec le niveau de l'option, mais on ne l'observe pas dans notre échantillon.

Cela peut être dû à la faible représentation des niveaux supérieurs à 1, et à la faible opportunité de consommer « vraiment plus » sur le type de poste étudié (on ne va à priori chez le médecin spécialiste que si on a un vrai problème de santé, il ne s'agit pas ici de soins esthétiques ou de confort à priori). Une autre explication peut être la faible différence entre les niveaux de garantie proposés sur les différentes options (l'option 1 est surement suffisante pour ne pas consommer davantage).

Les consommations croissantes avec les niveaux d'options sont plus représentées sur des postes tels que l'Optique ou le Dentaire.

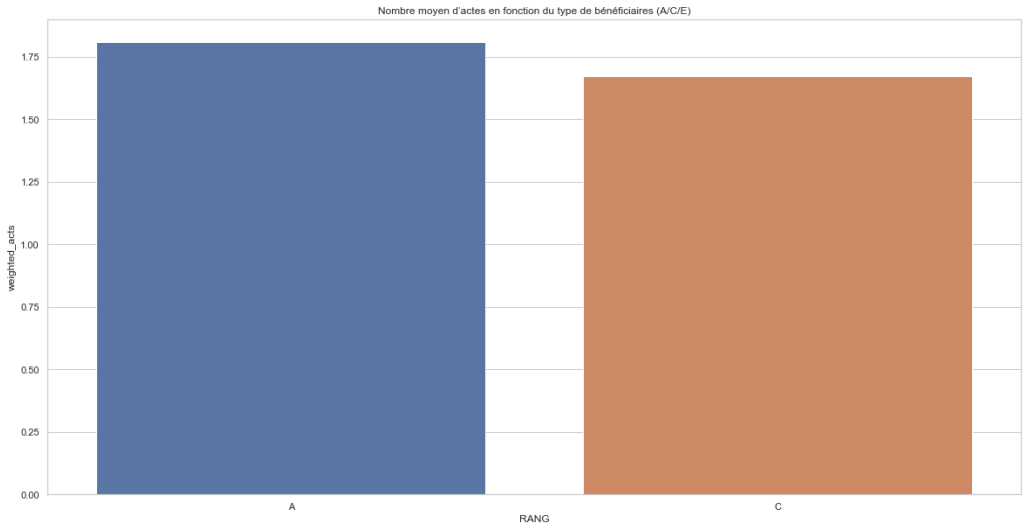


33-Nombre moyen d'actes en fonction du niveau d'option

Répartition du nombre d'acte par type de bénéficiaire (assuré/conjoint/enfant) :

RANG	weight	weighted_acts
A	20266	1.809731
C	7845	1.670746

34-Répartition des actes et du nombre d'assurés par Rang (assuré/Conjoint)



35- Histogramme de répartition du nombre d'actes par type de rang

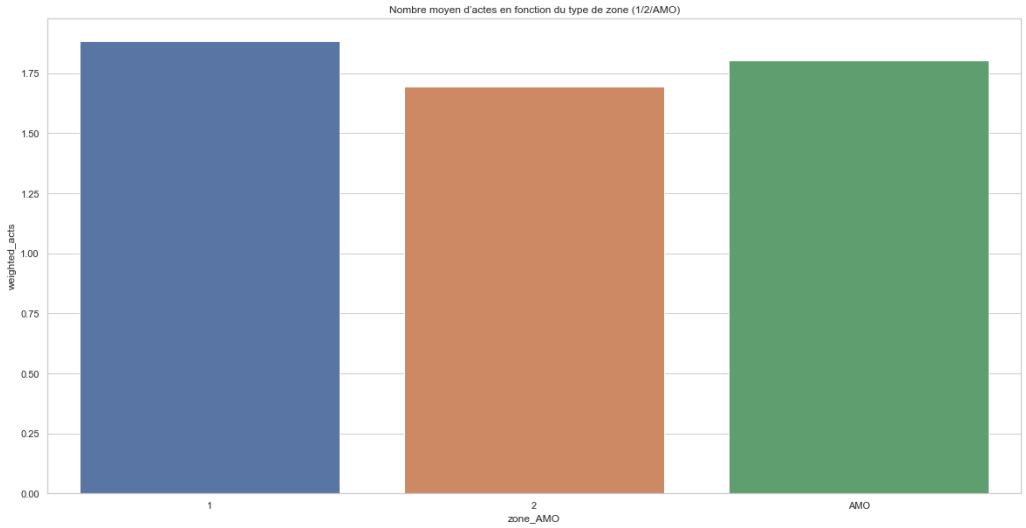
Les conjoints consomment un peu moins d'actes que les adhérents principaux, ce qui est une constatation normale du fait que certains conjoints ont déjà une complémentaire santé, et aussi car

nous ne voyons donc pas toutes leurs consultations (ils n'envoient toujours leur demande de remboursement que s'ils ont du reste à charge après intervention de leur complémentaire santé).

Répartition du nombre d'actes par zone géographique (zone 1, zone 2, zone Alsace-Moselle):

Les assurés habitant en zone 2 consomment plus souvent que les assurés habitant en en zone 1, ce qui est à priori surprenant. Mais les coûts moyens de consommation des actes, à priori plus élevés en zone 1, vont sûrement compenser ces chiffres pour un coût annuel qui aura l'effet inverse.

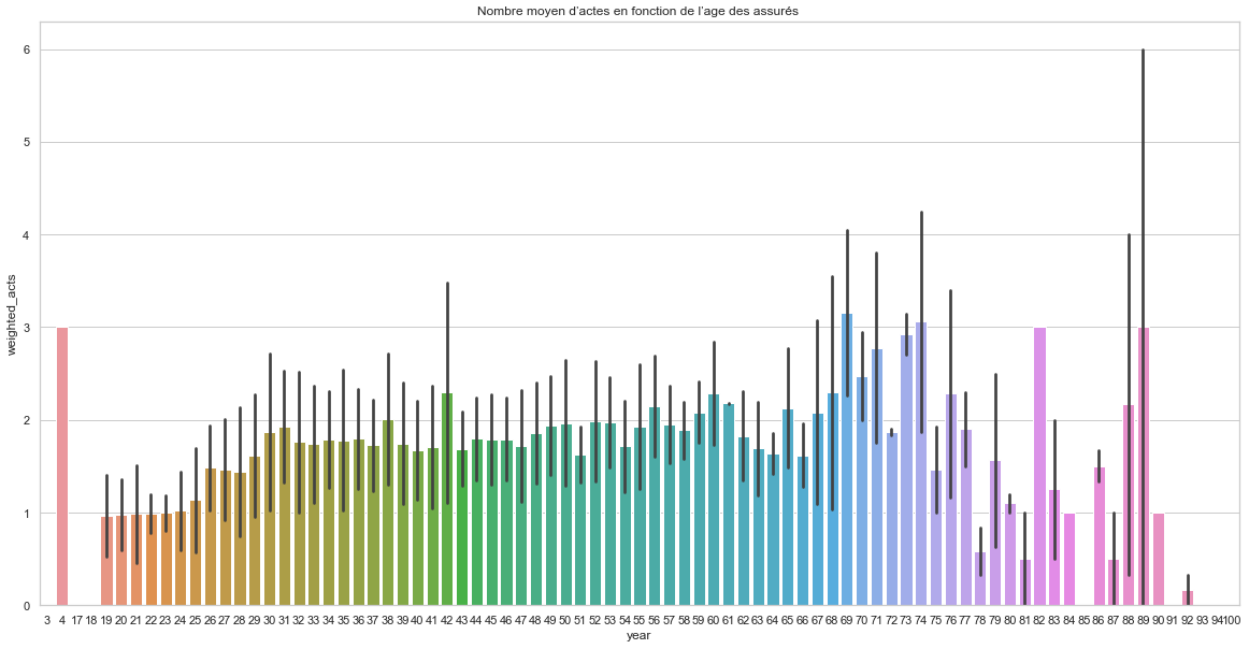
La zone Alsace-Moselle contient également une fréquence de consommation élevée, cela peut être dû au meilleur remboursement des soins dans ces départements par le régime obligatoire incitant à se soigner sans délai, ou à une tarification plus courante d'actes codifiés par la Sécurité Sociale majorants la consultation de base.



36-Répartition du nombre d'actes par zone géographique

Répartition du nombre d'acte par âge :

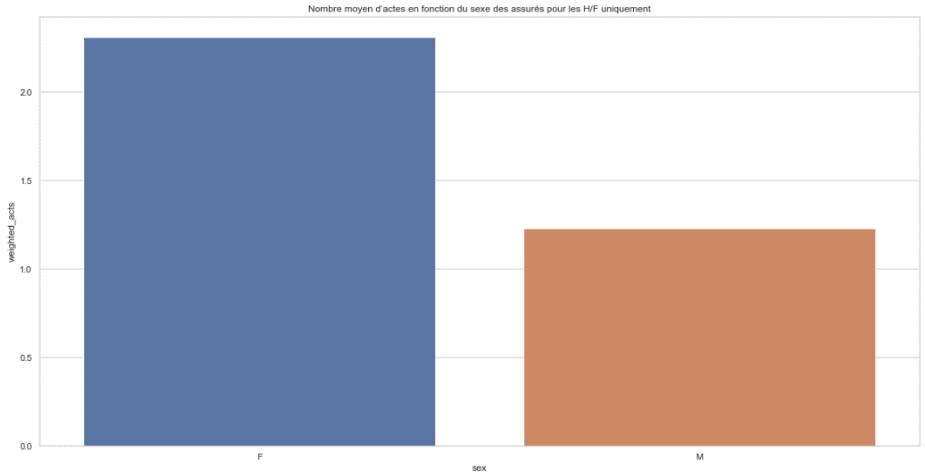
L'âge moyen global est de : 41.45 années.
L'âge moyen des hommes est de : 42.74 années.
L'âge moyen des femmes est de : 40.17 années.



37- Répartition du nombre moyen d'actes par âge

La courbe des fréquences a une tendance haussière avec l'âge, jusqu''à environ 75 ans. Les fortes variations observées pour les âges plus élevés ne sont pas significatives, car nous avons trop peu de données avec ces valeurs dans l'échantillon.

Répartition du nombre d'acte par sexe :



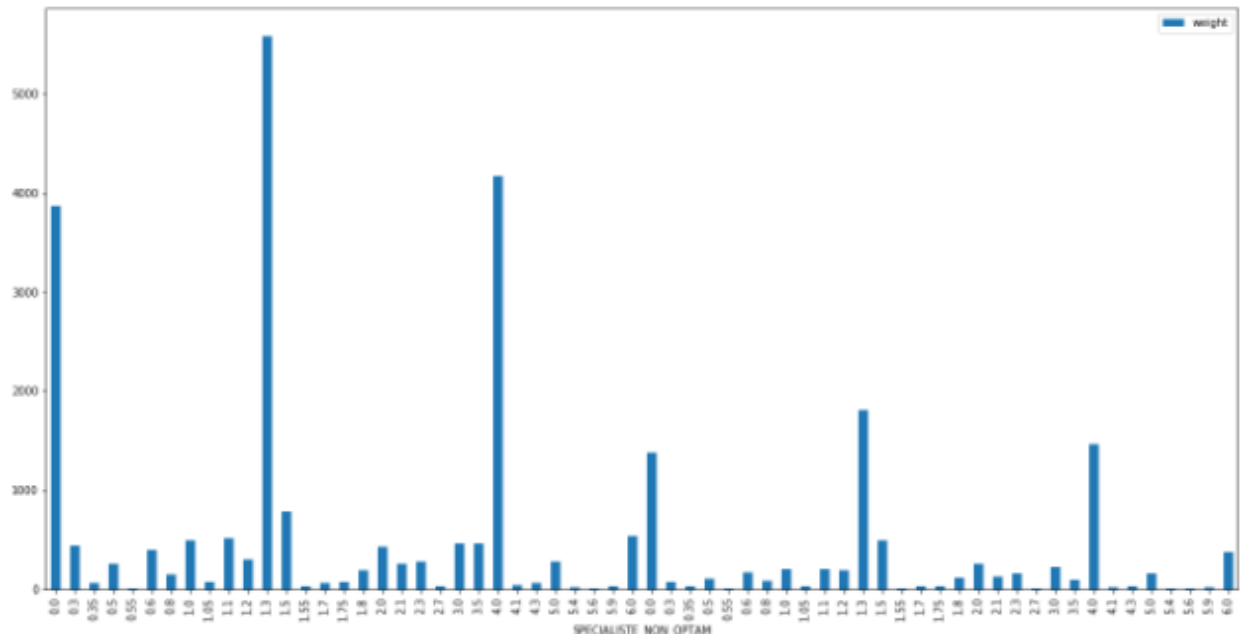
38- Répartition du nombre moyen d'actes par sexe

sex	weight	weighted_acts
F	14086	2.312083
M	14025	1.227451

39- Répartition du nombre moyen d'actes et du nombre d'assurés par sexe

Les femmes ont une fréquence de consommation presque deux fois plus élevée que les hommes, ce qui est cohérent avec notre modèle actuel et les tendances observées sur le marché.

Répartition des assurés par niveau de garantie :



40- Répartition des assurés de la BDD GLM et apprentissage supervisé par niveau de garantie

Les niveaux de garantie étudiés auraient pu être regroupés par tranche pour éviter la dispersion qu'on observe. Toutes autres variables explicatives confondues, on n'observe pas une fréquence croissante avec le niveau de garantie. Cela risque de poser problème lors de la modélisation du GLM ou du Machine Learning, qui pourrait définir une fréquence décroissante avec le niveau de garantie. Mais ces données sont explicables d'une part par la forte granularité que nous avons laissé dans l'échantillon, et ici aussi par la faible opportunité de consommer « vraiment plus » comme expliqué pour les niveaux de garantie.

Conclusion sur les fréquences observées : l'échantillon montre les tendances classiques observées en tarification santé, excepté sur la variable niveau de garantie, et niveau d'option.

Ces deux exceptions peuvent nuire à la qualité des modélisations qui seront faites, mais le coût moyen sera de toute façon multiplié à cette fréquence, et une compensation se fera peut-être sur ces modalités.

3.3.4.a Analyse graphique

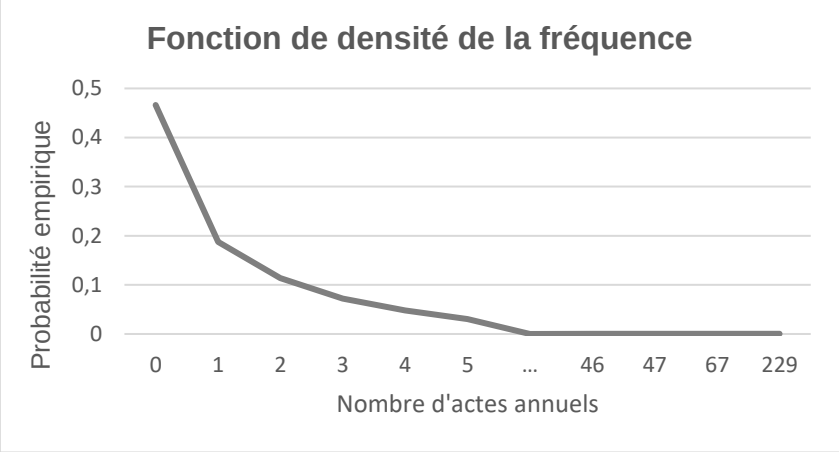
La répartition du nombre d'actes est la suivante :

Nombre d'actes annuels	Poids	Probabilité
0	13107	0,466259
1	5276	0,187685
2	3196	0,113692
3	2023	0,071965
4	1349	0,047988
5	862	0,030664
…	…	…
46	2	0,000071
47	2	0,000071
67	3	0,000107
229	1	0,000036

41- répartition des fréquences de consommation de l'échantillon

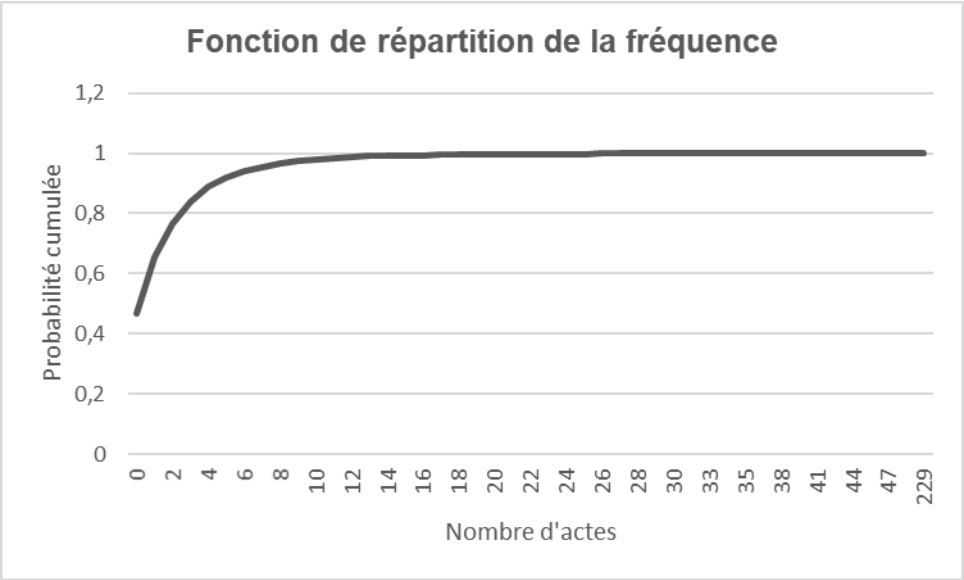
Fonction de densité :

La fonction de densité est en forme de coude, caractéristique classique des variables de comptage.



42-Fonction de densité de la fréquence

Fonction de répartition



43- Fonction de répartition de la fréquence

Moyenne	1,77
Ecart-type	3,42

44- Moyenne et écart-type de la fréquence

La garantie Spécialiste non OPTAM est consommée en moyenne 1,77 fois. Plus de 90% des assurés ont consommé moins de 6 soins dans l'année.

Comme c'est souvent le cas dans la représentation des fréquences en assurance santé, on trouve un grand nombre d'assurés n'ayant pas consommé la garantie spécialiste non OPTAM, ce qui provoque une surreprésentation des zéros (47% de l'échantillon ici).

Avec la simple étude de la moyenne et de l'écart-type du nombre d'actes, on constate qu'il y a un écart entre la moyenne et l'écart-type. La modélisation de cette variable via une loi de Poisson ne paraît donc d'avance pas tout à fait adaptée à notre courbe. Nous la testerons tout de même.

La forme coudée de la fonction de densité de la fréquence empirique nous incite à comparer la courbe de densité avec celles des lois de Poisson et Binomiale Négative. Nous testons le GLM sur ces lois, ainsi que leur version Zero-Inflated (ZIP et ZINB).

3.3.4.b Résultats du GLM

Nous construisons les modèles linéaires généralisés du nombre d'actes annuels pour les lois de Poisson et Binomiale Négative, en la faisant dépendre des variables de régression « RANG » (type de bénéficiaire), « sex » (sexe), « BASE_OPTION » (niveau d'option), « zone_AMO » (zone géographique), « SPECIALISTE_NON_OPTAM » (niveau de garantie) et « year » (âge).

Résultats pour la fréquence :

	Log vraissemblance	Déviance résiduelle	Critère AIC	Critère BIC
Poisson	- 45 697	62 903	91 413	- 116 531
Negative Binomial	- 45 697	62 903	65 816	- 163 290
ZIP	- 38 846		77 711	77 790
ZINB	- 32 732		65 483	65 561

45- Résultats des critères de performance du GLM sur la fréquence

A première vue le modèle ZINB est le meilleur modèle, suivi du modèle ZIP.

Mais lorsque nous regardons les fréquences modélisées par ces modèles, elles nous semblent inexploitables car toutes arrondies à l'unité. La fréquence n'augmente donc pas avec le niveau de garantie (ou très peu) et paraît donc constante. Pour nous ces modèles ne sont donc pas à prioriser.

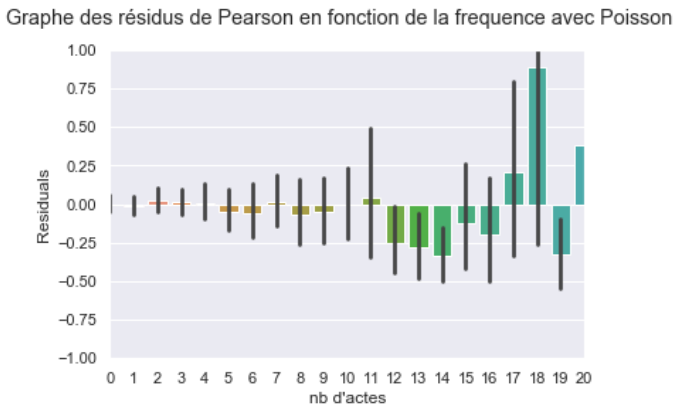
Le tableau ci-dessous présente les fréquences modélisées avec les lois ZIP et ZINB pour des assurés de différentes caractéristiques. On voit que les fréquences sont quasi constantes :

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	**Niveau Garantie (BR)**												
ZERO INFLATED POISSON	0	1	2	**1**	1	1	3	2	1	1	2	1	1
	0,3	1	2	**1**	1	1	3	2	1	1	2	1	2
	0,6	1	2	**1**	1	1	3	2	1	1	2	1	2
	1,3	1	2	**1**	1	1	3	2	1	1	2	1	2
	2	1	2	**1**	1	1	3	2	1	1	2	1	2
	3,7	1	2	**1**	1	1	3	2	1	1	3	1	2
	5,3	1	2	**1**	1	2	3	2	1	1	3	2	2
ZERO INFLATED NEGATIVE BINOMIALE	0	1	2	**1**	0	1	3	2	1	1	2	1	1
	0,3	1	2	**1**	1	1	3	2	1	1	2	1	1
	0,6	1	2	**1**	1	1	3	2	1	1	2	1	1
	1,3	1	2	**1**	1	1	3	2	1	1	2	1	1
	2	1	2	**1**	1	1	3	2	1	1	2	1	1
	3,7	1	2	**1**	1	2	3	2	1	1	2	2	2
	5,3	1	2	**1**	1	2	3	2	1	1	2	2	2

***46**-Comparaison des fréquences ZIP et ZINB modélisées selon plusieurs profils d'assurés*

Nous décidons alors de modéliser la fréquence des consommations avec une loi de Poisson, meilleur choix parmi les modèles « simples » selon le critère BIC. Les résultats seront présentés dans la suite.

3.3.4.c Analyse des résidus



47- Résidus de Pearson des fonctions modélisées avec loi de Poisson

Les résidus de Pearson obtenus avec la loi Poisson sont satisfaisant, mais on constate que les forts niveaux de consommation sont moins bien prédits.

3.3.4.d Estimation des paramètres

Pour les variables explicatives étudiées, nous représentons les paramètres estimés, obtenus avec la fonction *summary* de python :

Modalité	Paramètre estimé	Erreur standard	t-value Z	p-value
RANG_A	0.3021	0.011	26.509	0.000
RANG_C	0.0632	0.012	5.191	0.000
sex_F	0.5222	0.011	45.953	0.000
sex_M	-0.1570	0.012	-13.361	0.000
BASE_OPTION_B	-0.1613	0.028	-5.670	0.000
BASE_OPTION_O1	0.0587	0.028	2.121	0.034
BASE_OPTION_O2	0.2031	0.029	6.974	0.000
BASE_OPTION_O3	0.2647	0.100	2.645	0.000
zone_AMO_1	0.1388	0.010	13.363	0.000
zone_AMO_2	0.1006	0.010	10.480	0.000
zone_AMO_AMO	0.1258	0.014	9.214	0.000
SPECIALISTE_NON_OPTAM	0.0414	0.007	6.098	0.000
year	0.1043	0.005	20.296	0.000

Définition des grandeurs observées :

Std err: c'est l'écart type de l'estimation ponctuelle du coefficient dans le GLM. C'est une mesure de l'incertitude sur cette estimation - si elle est trop grande, alors on a une estimation ponctuelle de coefficient calculée avec beaucoup d'imprécision.

P (> | z |): ou "p-value" du test pour savoir si l'estimation ponctuelle du coefficient est significativement différente de 0. Intuitivement, elle nous indique si notre estimation ponctuelle a été calculée suffisamment précisément pour la distinguer de zéro. Nous définissons généralement « assez précisément » en utilisant la maxime p <0,05.

[0.025 0.975] : intervalle de confiance à 95%

Les scores Z sont des mesures de l'écart type. Par exemple, si un outil renvoie un score Z de +2,5, il est interprété comme « + 2,5 écarts types par rapport à la moyenne ».

Les valeurs P (ou p-value) sont des probabilités. Les deux statistiques sont associées à la distribution normale standard. Cette distribution relie les écarts-types aux probabilités et permet d'attacher une signification et une confiance aux scores Z et aux valeurs p.

Qualité d'ajustement du modèle de régression de Poisson

Une source courante d'échec du modèle de régression de Poisson est que les données ne satisfont pas au critère *moyenne* = *variance* imposé par la distribution de Poisson.

La méthode *summary ()* de la classe *statsmodels.GLMResults* montre quelques statistiques de qualité d'ajustement utiles pour nous aider à évaluer si notre modèle de régression de Poisson a réussi à ajuster les données d'entraînement. Regardons leurs valeurs:

Les valeurs trouvées pour la log vraisemblance (-4,57 e+05) et les critères BIC sont plus grands que pour les modèles ZIP et ZINB, ce qui induit que l'ajustement est moins bon pour le modèle Poisson en théorie.

L'erreur standard de chaque paramètre est faible, ce qui indique une bonne précision et un intervalle de confiance étroit.

La significativité du modèle est déterminée en lisant la p-value et en la comparant au niveau de significativité souhaité. Les étoiles accompagnées de la p-value indiquent le degré de significativité.
Dans notre modélisation, toutes les variables semblent être significatives à un niveau de confiance 95%.

Ces résultats confortent le choix des variables explicatives, qui ont été conservées identiques au modèle Antenia.

Le GLM modélise une fréquence croissante avec toutes les variables, excepté le sexe masculin, ce qui se traduit par le fait qu'un homme consommerait moins la garantie Spécialiste qu'une femme. De même es assurés sur les niveaux Base consommeraient moins que les autres. Ces conclusions sont tout à fait cohérentes avec notre modèle Antenia.

Tous les coefficients sont ordonnés de la même manière que nos coefficients Antenia, sauf les tarifs zone AMO et zone 2 : dans notre modèle interne, la zone Alsace-Moselle a un tarif inférieur à la zone 2.

Si nous simulons les coûts moyens pour des assurés ayant des caractéristiques différentes, nous obtenons les résultats suivants :

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	**Niveau Garantie (BR)**												
Fréquence POISSON	0	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	0,3	0,892	2,112	**1,071**	0,843	1,334	3,041	2,033	1,031	1,057	2,285	1,269	1,640
	0,6	0,899	2,128	**1,079**	0,850	1,345	3,064	2,048	1,039	1,065	2,302	1,279	1,652
	1,3	0,915	2,166	**1,098**	0,865	1,369	3,119	2,085	1,057	1,084	2,343	1,302	1,682
	2	0,931	2,205	**1,118**	0,880	1,393	3,175	2,122	1,076	1,104	2,385	1,325	1,712
	3,7	0,972	2,302	**1,167**	0,919	1,455	3,314	2,216	1,123	1,152	2,490	1,384	1,787
	5,3	1,012	2,398	**1,216**	0,957	1,515	3,452	2,308	1,170	1,200	2,594	1,441	1,861

48- Fréquence modélisée avec le GLM avec une loi de Poisson

Si les ordres de grandeurs sont cohérents avec la réalité observée en moyenne, les fréquences modélisées ne sont pas étendues entre les garanties de faible niveau et les garanties de niveau élevé. Il y a moins de 0,5 actes de différences entre les faibles et les forts niveaux de garantie.

On observe plutôt bien les effets d'âge, de sexe, des options et des zones, et ce dans les ordres de grandeurs attendus (fréquence croissante avec un niveau d'option croissant, un âge croissant, avec un sexe féminin…), sauf la zone AMO dont les tarifs apparaissent supérieurs à la zone 2.

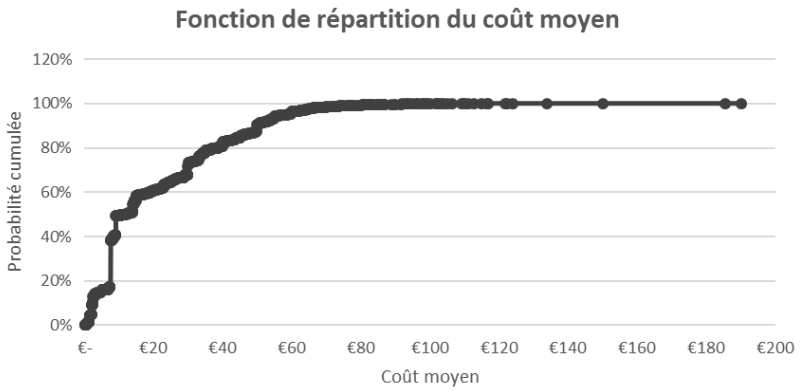
Les résultats de la modélisation du coût moyen joueront un grand rôle dans la confirmation de ces influences des variables.

3.3.5. Modélisation du cout moyen

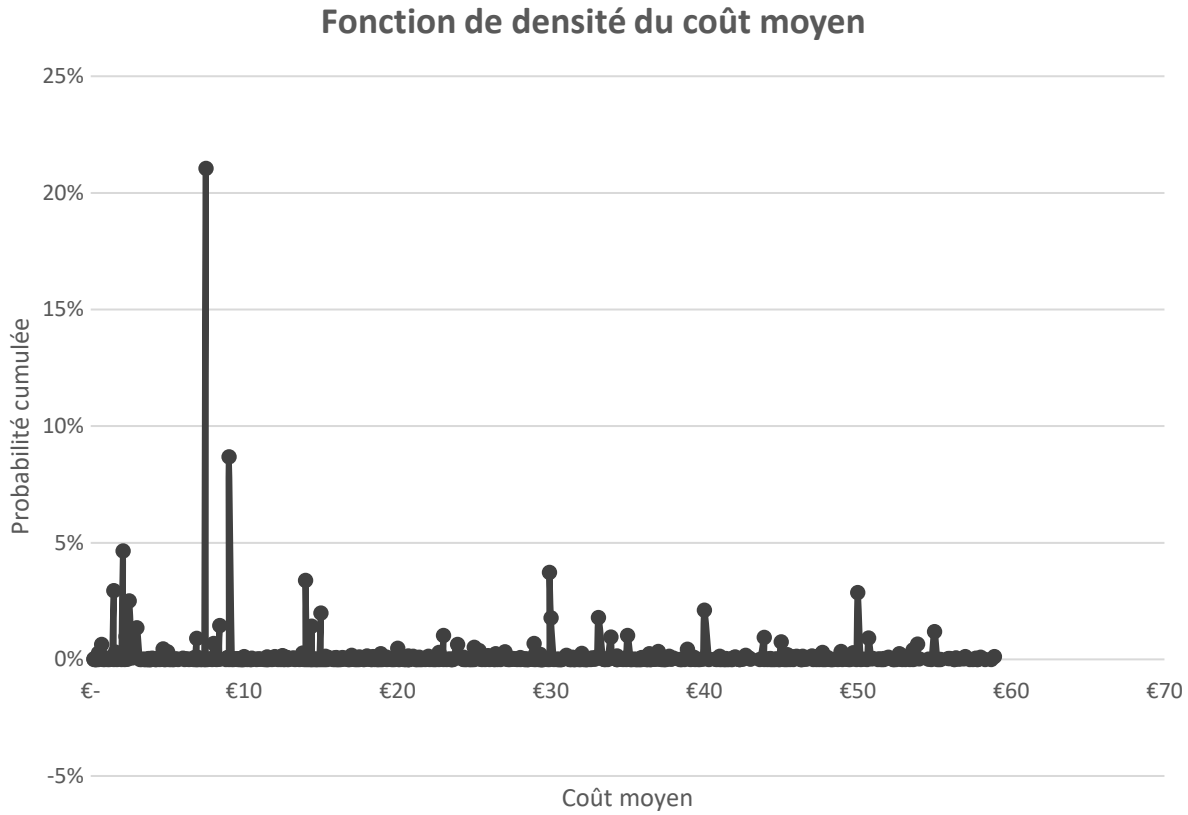
3.3.5.a Analyse graphique

Coût moyen	Poids	Probabilité	Probabilité cumulée
0,19 €	1	0,002%	0,002%
0,20 €	12	0,024%	0,026%
0,25 €	1	0,002%	0,028%
0,30 €	1	0,002%	0,030%
0,33 €	1	0,002%	0,032%
0,40 €	1	0,002%	0,034%
0,50 €	136	0,274%	0,308%
…	…	…	…
121,70 €	1	0,002%	99,984%
122,50 €	1	0,002%	99,986%
123,90 €	1	0,002%	99,988%
133,90 €	3	0,006%	99,994%
150,00 €	1	0,002%	99,996%
185,40 €	1	0,002%	99,998%
190,30 €	1	0,002%	100,000%

49- Répartition des assurés par montant du coût moyen



50- Fonction de répartition du coût moyen



51- Fonction de densité du coût moyen

Moyenne	16,85
Ecart-type	15,4
Maximum	200

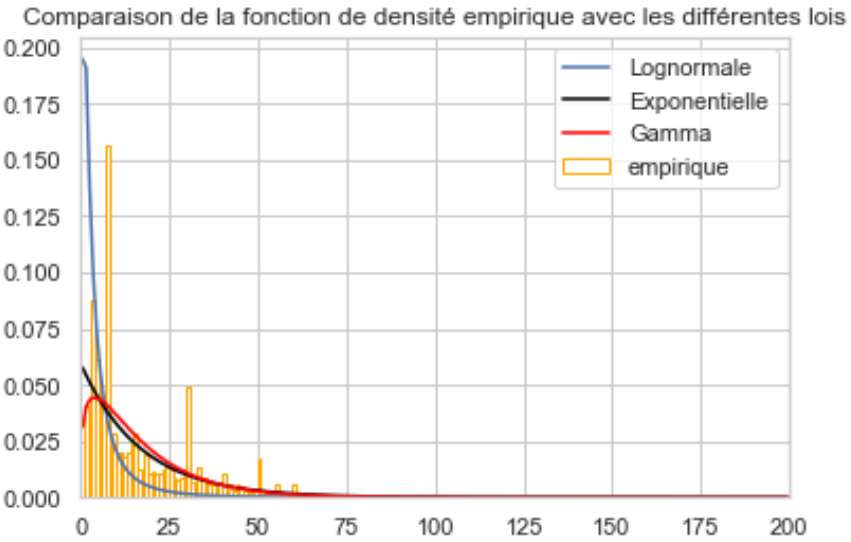
52- Caractéristiques du coût moyen

Le pic des coûts moyens du poste SPECIALISTE NON OPTAM sont généralement autour de 7,5€ (valeur du ticket modérateur). Un assuré moyen de l'échantillon consomme 16,85€ de ce poste par soin et l'écart-type de cette distribution est égale à 15,4€.

L'allure de la distribution empirique des coûts moyens diffère de celle de la fréquence, celle-ci est plutôt en forme de cloche, mais le fort échantillon de données ne permet pas une visualisation très nette.

Nous remarquons que les coûts moyens les plus probables sont inférieurs à 10€. A partir de cette valeur, nous observons une baisse du nombre d'observations des coûts moyens plus élevés.

Compte-tenu de ces arguments nous allons comparer la courbe de densité avec celles des lois log-normale, Gamma, et exponentielle avec la fonction *getattr* et son fit dans Python :



53- Comparaison de la fonction de densité empirique avec les différentes lois testées

Avec cette première visualisation, nous pouvons envisager d'exclure la loi Exponentielle : en effet, elle n'a pas l'allure d'une cloche comme celle de notre distribution, ce qui fausserait les résultats des petits coûts moyens, qui sont les plus probables.

Les lois Gamma et Log-normale fournissent des résultats satisfaisants, même si la première sous-estime les coûts moyens faibles, et la seconde les surestime.

Les résultats des tests non paramétriques de Kolmogorov-Smirnov semblent montrer que la loi Log-normale est la seule qui serait ajustable aux données. En effet c'est la seule des deux lois qui obtient une p-value assez haute, signifiant la validation de l'hypothèse d'ajustement.
Les lois Gamma et Exponentielle sont éliminées d'après ce test.

Loi	Statistique	p-value	Conclusion
Log-Normale	0.00079	0.960332039907191	Non Rejet
Gamma	0.96033	0.0	Rejet
Exponentielle	0.07053	7.391e-137	Rejet

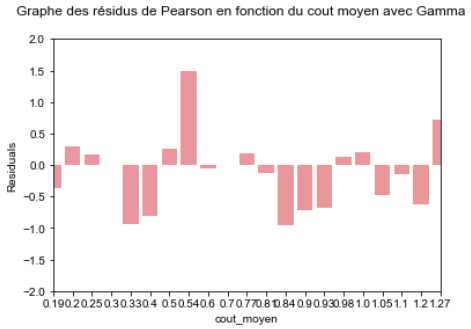
54-Résultats des tests non paramétriques sur le coût moyen

3.3.5.b Résultats du GLM :

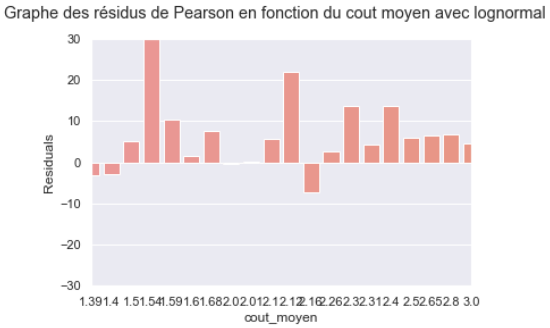
Analyse des résidus

Nous réalisons les deux modélisations linéaires généralisées, l'une avec la loi Gamma, l'autre avec la loi log-normale.

Nous commençons par une analyse de la distribution des résidus de chaque modèle pour vérifier leur symétrie par rapport au zéro et valider l'hypothèse d'espérance nulle et d'homoscédasticité.



55- Distribution des résidus des prédiction du coût moyen avec une loi Gamma



56- Distribution des résidus des prédiction du coût moyen avec une loi Lognormale

Selon les graphiques qui représentent les résidus de Pearson, il apparaît que ceux de Gamma sont dans un ordre de grandeur bien plus petit. Le modèle Gamma donne des résultats plus proches de la réalité que le modèle Log-normal., même s'il sous-estime en moyenne les coûts des soins spécialiste Non OPTAM.

Les paramètres du tableau suivant nous permettront de conclure sur le meilleur modèle.

	Déviance résiduelle	Critère AIC	Critère BIC
Gamma	97 547	423 249	-323 231
Log-normale	177 490 000	355 294	17 328 498

57- Critères de performance des modélisations GLM sur le coût moyen

Le modèle Gamma apparaît comme le plus adapté pour représenter les coûts moyens de notre échantillon.

Sa déviance résiduelle est plus faible que pour le modèle Log-normal, et sa Déviance nulle est également inférieure, ce qui prouve sa meilleure capacité à expliquer le coût moyen par les variables tarifaires utilisées plutôt que par un modèle nul associé à une constante.

Le critère BIC pénalise également davantage le modèle Log-normal.

Seul le critère AIC pénalise plus le modèle Gamma, mais compte-tenu de la taille de notre échantillon le critère BIC reste à privilégier pour le choix du modèle.

Les ajustements graphiques ne permettaient pas de conclure avec certitude quel modèle serait le plus adapté, mais ces critères de validations permettent d'affirmer que le modèle Gamma est à choisir pour représenter les coûts moyens.

Compte-tenu de toutes ces informations, nous devrions privilégier le modèle Gamma pour la prédiction des coûts moyens. Mais l'analyse des paramètres décrite dans le paragraphe suivant nous poussera à choisir finalement le modèle Log-normale.

Estimation des paramètres

Pour les variables explicatives étudiées, nous représentons les paramètres estimés, obtenus avec la fonction *summary* de python :

Modalité	Paramètre estimé	Erreur standard	t-value	p-value
RANG_A	1.475211	0.031	47.118	0.000***
sex_M	0.503498	0.036	13.969	0.000***
BASE_OPTION_O1	2.249496	0.039	57.239	0.000***
BASE_OPTION_O2	2.649034	0.057	46.668	0.000***
BASE_OPTION_O3	3.131969	0.430	7.283	0.000***
zone_AMO_2	0.870251	0.032	26.871	0.000***
zone_AMO_AMO	0.360510	0.059	6.157	0.000***
SPECIALISTE_NON_OPTAM	-0.547922	0.022	-25.241	0.000***
year	-0.002221	0.017	-0.127	0.899

58- Résultats de la modélisation GLM du coût moyen avec loi Gamma

L'erreur standard des paramètres est faible, ce qui implique une bonne précision et un intervalle de confiance large.

La significativité du modèle est déterminée en lisant la p-value et en la comparant au niveau de significativité souhaité. Les étoiles accompagnées de la p-value indiquent le degré de significativité.
Dans notre modélisation, toutes les variables semblent être significatives à un niveau de confiance 95%, sauf l'âge, ce qui pose évidemment un problème pour une tarification santé.

Le GLM modélise un coût moyen décroissant avec le niveau de garantie, ce qui est totalement contraire à notre principe de tarification.

Ces deux remarques nous font penser que le modèle Gamma est inexploitable pour modéliser le coût moyen d'un soin Médecin Spécialiste. Nous choisissons donc de faire notre GLM avec une loi log-normale.

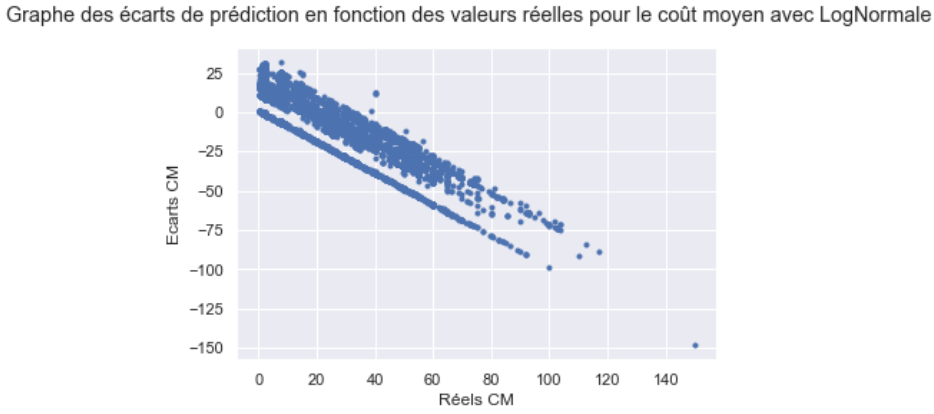
Avec cette loi, les résultats des estimations des paramètres sont les suivants :

Modalité	Paramètre estimé	Erreur standard	t-value	p-value
RANG_A	2.982062	0.015	204.602	0.000***
sex_M	0.025363	0.012	2.083	0.037***
BASE_OPTION_O1	0.360136	0.017	20.874	0.000***
BASE_OPTION_O2	0.475541	0.021	22.493	0.000***
BASE_OPTION_O3	0.947216	0.074	12.761	0.000***
zone_AMO_2	-0.279907	0.012	-23.356	0.000***
zone_AMO_AMO	-0.581438	0.028	-21.107	0.000***
SPECIALISTE_NON_OPTAM_norm	0.000313	0.008	0.042	0.967
year_norm	0.002733	0.006	0.458	0.647

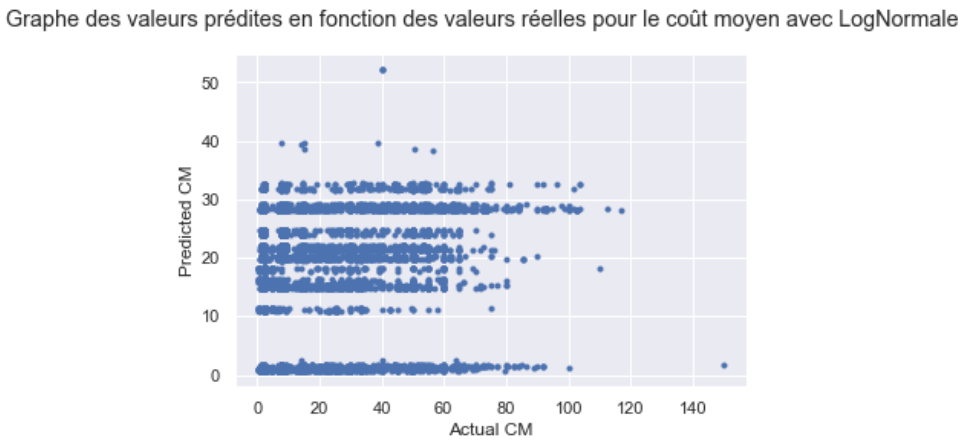
59- Résultats de la modélisation GLM du coût moyen avec loi Log-normale

Les coefficients obtenus avec la loi log-normale sont plus représentatifs d'une tarification santé, du moins en terme d'influence. On retrouve bien ici un tarif plus élevé si l'assuré est adhérent principal, ou de sexe masculin, et d'autant plus croissant que le niveau d'option augmente. Le tarif est également moins cher dans les zones2 et Alsace-Moselle, comme défini avec Antenia.

Les variables Age et Niveau de Garantie (SPECIALISTE_NON_OPTAM) ne sont pas significatives avec ce modèle. Nous le conservons tout de même, car il s'agit ici du calcul du coût moyen, qui sera multiplié par la fréquence modélisée. Nous verrons par la suite que le coût annuel obtenu est satisfaisant, la fréquence permettant de maîtriser l'influence des variables niveau de garantie et zone.



60- Graphique des écarts de prédiction pour le coût moyen avec la loi Log-normale



61- graphique des valeurs prédites avec un GLM via loi Log-normale en fonction des valeurs réelles

Les graphiques présentant les résidus de Pearson en fonction des valeurs estimées. Nous pouvons voir que les résidus du modèle Log-normal ne sont globalement pas bien répartis autour de zéro. Le modèle présente une tendance : les écarts diminuent avec l'augmentation du coût moyen, et ne sont pas répartis autour de zéro, ce qui contredit la théorie de l'homogénéité de la variance.
Ces résultats nous laissent penser que notre modélisation des coûts moyens n'est pas fiable, ce qui va poser un problème en la confiance du modèle GLM global.

Rappelons que nous cherchons à savoir si notre norme Antenia est fiable en la comparant à des techniques modernes telles qu'un GLM. On doit pour cela la comparer à un GLM jugé fiable également.

Prédiction du coût moyen sur l'échantillon d'assurés types

Si nous simulons les coûts moyens pour des assurés ayant des caractéristiques différentes, nous obtenons les résultats suivants :

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	Niveau Garantie (BR)												
COUT MOYEN GLM LOI LOGNORMALE	0	20,12 €	19,71 €	**20,22 €**	1,02 €	28,98 €	31,71 €	14,90 €	15,28 €	11,30 €	11,05 €	11,36 €	52,13 €
	0,3	20,12 €	19,71 €	**20,22 €**	1,02 €	28,98 €	31,71 €	14,90 €	15,28 €	11,30 €	11,05 €	11,36 €	52,13 €
	0,6	20,12 €	19,71 €	**20,22 €**	1,02 €	28,98 €	31,71 €	14,90 €	15,28 €	11,30 €	11,05 €	11,36 €	52,13 €
	1,3	20,12 €	19,71 €	**20,22 €**	1,02 €	28,99 €	31,72 €	14,90 €	15,28 €	11,31 €	11,05 €	11,36 €	52,14 €
	2	20,12 €	19,72 €	**20,22 €**	1,03 €	28,99 €	31,72 €	14,90 €	15,29 €	11,31 €	11,05 €	11,36 €	52,15 €
	3,7	20,13 €	19,72 €	**20,23 €**	1,03 €	29,00 €	31,73 €	14,91 €	15,29 €	11,31 €	11,05 €	11,37 €	52,16 €
	5,3	20,14 €	19,73 €	**20,24 €**	1,03 €	29,01 €	31,74 €	14,91 €	15,30 €	11,31 €	11,06 €	11,37 €	52,18 €

62-modélisation du coût moyen avec GLM sur plusieurs profils d'assurés

Comme le prévoyait les résultats de la modélisation, le coût moyen prédit n'est pas significativement dépendant du niveau de garantie, ni de l'âge. Cette modélisation n'aura donc pas la précision que nous attendions pour comparer notre tarif Antenia à un modèle fiable, cependant la part de la fréquence dans le coût annuel comparé compensera au moins partiellement ce manque.

Les coûts moyens modélisés ont presque la même valeur entre les faibles niveaux et les niveaux élevés. Cela aura surement un impact sur le coût annuel final. Ils sont croissants avec le niveau d'option, l'âge et le niveau de garantie (malgré une faible influence de ces dernières variables)

décroissants avec les zones 2 et Alsace-Moselle, ou pour les conjoints. Toutes ces tendances sont celles que nous attendions.

Les femmes ont un coût moyen moins élevé que les hommes, mais la différence n'est pas significative et sera surement compensée lors de la multiplication à la fréquence.

L'étude des coefficients (coût moyen d'un assuré avec des caractéristiques précises ci-dessous divisé par le coût moyen de l'assuré type d'Antenia) montre bien l'analyse faite des variables explicatives. Les ordres de grandeur sont clairement discutables : par exemple un conjoint ne consommerait que 5% du coût moyen d'un adhérent principal, et les femmes consomment 2,5% de moins qu'un homme.

En revanche, le fait que le niveau de garantie n'influence pas le tarif rend cette modélisation plus proche du modèle de tarification interne, qui multiplie la prime de l'assuré type par un même coefficient par variable explicative, quel que soit le niveau de garantie.

	RANG	A	A	A	C	A	A	A	A	A	A	A	A
	sex	M	F	M	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	B	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	1	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	40	40	40	40	40	40	40	50	60	40
	IMPACT	AGE 20 ANS	SEXE FEMME	ASSURE MOYEN	CONJOINT	OPTION 1	FEMME OPTION 2	FEMME ZONE 2	ZONE 2	ZONE AMO	FEMME ZONE AMO	AGE 60 ANS	OPTION 3
	Niveau de garantie (BR)												
Coefficient Cout moyen LOGNORMALE	0	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	0,3	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	0,6	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	1,3	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	2	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	3,7	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579
	5,3	0,995	0,975	1,000	0,051	1,434	1,569	0,737	0,756	0,559	0,546	0,562	2,579

63- Coefficients des coûts moyens modélisés par GLM sur plusieurs profils d'assurés par rapport à l'assuré type

Le résultat final (le coût annuel) déterminera réellement la qualité du modèle via GLM.

3.3.6. Tarification d'une grille avec le meilleur modèle

La méthode retenue pour la prédiction de notre tarif de consultation Spécialiste NON OPTAM est le Log-normal pour le coût moyen et le Poisson pour la fréquence.

Il nous faut maintenant obtenir le tarif modélisé avec le GLM en multipliant la prédiction de la fréquence par celle du coût moyen.

Pour analyser les résultats, nous prenons un échantillon type d'assurés avec des caractéristiques différentes en terme d'âge, sexe, zone, niveau d'option, type de bénéficiaires, et nous simulons un tarif sur les mêmes niveaux que ceux que nous avons dans notre outil de tarification. Nous pouvons ainsi faire nos premiers commentaires sur le tarif qui serait proposé via un GLM, en le comparant avec notre tarificateur, d'un point de vue prime pure, et d'un point de vue coefficient.

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	**Niveau Garantie (BR)**												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
POISSON x LOGNORMALE	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	17,94 €	41,63 €	21,65 €	0,86 €	38,67 €	96,42 €	30,29 €	15,75 €	11,95 €	25,24 €	14,42 €	85,47 €
	0,6	18,08 €	41,95 €	21,82 €	0,87 €	38,97 €	97,17 €	30,52 €	15,87 €	12,04 €	25,43 €	14,53 €	86,13 €
	1,3	18,41 €	42,71 €	22,21 €	0,89 €	39,67 €	98,92 €	31,07 €	16,16 €	12,26 €	25,89 €	14,79 €	87,69 €
	2	18,74 €	43,48 €	22,61 €	0,90 €	40,39 €	100,70 €	31,63 €	16,45 €	12,48 €	26,36 €	15,06 €	89,27 €
	3,7	19,57 €	45,41 €	23,61 €	0,94 €	42,18 €	105,17 €	33,04 €	17,18 €	13,03 €	27,53 €	15,73 €	93,23 €
	5,3	20,39 €	47,31 €	24,60 €	0,98 €	43,94 €	109,56 €	34,41 €	17,90 €	13,58 €	28,68 €	16,38 €	97,12 €

64- Comparaison des coûts annuels Antenia Vs GLM sur plusieurs profils d'assurés

Les différences entre les deux tarifs sont notables : d'une part, les tarifs obtenus via GLM ne dépendent presque pas du niveau de garantie, ce qui pose des problèmes, notamment aux bornes des niveaux : tarifs trop élevés sur les niveaux faibles, et tarifs surement trop faibles pour les niveaux élevés. Les coûts moyens n'ont pas été compensés par les fréquences pour ce problème.

Une différence notable se voit aussi dans les montants des niveaux hauts, qui sont soit plus élevés soit plus bas de manière significative.

Le tarif annuel est aussi extrêmement faible pour les assurés conjoints via le GLM.

En revanche toutes les variables explicatives font évoluer le tarif annuel dans le même sens.

Nous comparons les tarifs par rapport à celui de l'assuré type dont nous nous servons pour calculer la prime pure dans Antenia, en faisant la division du tarif par le tarif de la prime pure. Cela nous permet de comparer simplement et intuitivement les effets des variables explicatives (hors niveau de garantie).

A noter que si un coefficient dépasse 1, alors le profil de l'assuré entraîne un tarif plus élevé que celui de l'assuré des primes pures, et inversement.

	RANG	A	A	A	C	A	A	A	A	A	A	A	A
	sex	M	F	M	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	B	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	1	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	40	40	40	40	40	40	40	50	60	40
	IMPACT	AGE 20 ANS	SEXE FEMME	ASSURE MOYEN	CONJOINT	OPTION 1	FEMME OPTION 2	FEMME ZONE 2	ZONE 2	ZONE AMO	FEMME ZONE AMO	AGE 60 ANS	OPTION 3
	Niveau de garantie (BR)												
Cout annuel ANTENIA	0,3	0,621	2,011	1	0,93	1,05	2,11	2,01	1,00	0,42	0,86	0,57	1,10
	0,6	0,621	2,011	1	0,93	1,05	2,11	1,61	0,80	0,56	1,15	0,76	1,10
	1,3	0,621	2,011	1	0,93	1,05	2,11	1,90	0,94	0,90	1,85	1,23	1,10
	2	0,621	2,011	1	0,93	1,05	2,11	1,74	0,87	0,85	1,75	1,16	1,10
	3,7	0,621	2,011	1	0,93	1,05	2,11	1,80	0,90	0,88	1,80	1,19	1,10
	5,3	0,621	2,011	1	0,93	1,05	2,11	1,65	0,82	0,80	1,64	1,09	1,10
MEILLEUR MODELE GLM	0												
POISSON x LOGNORMAL	0,3	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9
	0,6	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9
	1,3	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9
	2	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9
	3,7	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9
	5,3	0,8	1,9	1,0	0,04	1,8	4,5	1,4	0,7	0,6	1,2	0,7	3,9

65- Comparaison des coefficients des profils entre modèle Antenia et modèle GLM

Les coefficients obtenus avec le GLM ne sont pas tous cohérents : les ordres de grandeurs des effets des variables explicatives vont dans le sens des coefficients Antenia, et certains ordres de

grandeurs sont très variés. Les variables sexe et zone sont plutôt homogènes entre la tarification Antenia et celle du GLM.

En revanche les variables âge et niveau d'option, s'ils évoluent parfois dans le même sens, prennent des valeurs beaucoup plus élevées pour le GLM, ce qui va faire augmenter les tarifs beaucoup plus rapidement. Par exemple les assurés qui ont un contrat avec une option de troisième niveau auraient un coût annuel 3,9 fois plus élevé que celui d'un assuré sur un contrat de base si on fait un tarif via GLM, alors qu'il n'augmenterait que de 10% avec le tarif Antenia.

Quant au rang, le coefficient du conjoint donné via GLM (0,04) est trop faible pour être considéré comme fiable.

Conclusion :

Les résultats de la tarification de la garantie médecin spécialiste mettent en lumière les avantages et inconvénients d'une tarification faite via un GLM.

La forme de la fonction qu'il génère est pratique car elle permet de voir clairement l'influence de chaque variable explicative dans les coefficients. En revanche les résultats obtenus ne sont pas forcément en accord avec les principes et les ordres de grandeur des tarification santé, et lorsqu'il y a des écarts avec la réalité observée, la correction ne pourra se faire que manuellement. Considérer le GLM comme modèle de fiabilité est donc partiellement remis en question suite à cette analyse de tarif, dans le sens où son modèle n'est lui-même pas parfait pour une tarification, sans un sérieux nettoyage et lissage des résultats, ce qui peut demander beaucoup de temps.

On peut tout de même conserver l'idée que les ordres de grandeur des coefficients (femme plus chère qu'un homme, conjoint moins cher qu'un adhérent principal etc.) sont cohérents.

Il faudra confirmer la qualité plus précisément avec l'analyse d'un échantillon réel, décrite en dernière partie.

3.4. Modélisation via une méthode Machine Learning

En tarification santé, la méthode de tarification très souvent utilisée par les actuaires se base sur les GLM (Modèles Linéaires Généralisés) [10] .

Mais depuis quelques années, les méthodes de tarification utilisées en Machine Learning se généralisent de plus en plus et sont en train de transformer le métier d'actuaire. Ces méthodes pourraient permettre de traiter un plus grand volume de données, de rendre accessibles de nouveaux modèles mathématiques pour la prédiction de sinistres, et d'optimiser l'utilisation des variables exogènes, permettant de segmenter les risques dans une maille toujours plus fine et d'apporter de la non linéarité aux modèles.

L'objectif de cette partie est de tester quelques modèles de tarifications qui serviront à analyser la fiabilité de notre modèle interne. Il doit donc être simple d'utilisation et d'interprétation (on doit pouvoir analyser l'influence des variables explicatives). Ce modèle devra également tenir compte des contraintes opérationnelles en place chez Swiss Life, en l'occurrence l'outil de tarification utilisé, et donc les variables qu'il est capable de stocker, ainsi que la forme générale de construction du tarif (primes pures et coefficients). Ces contraintes vont forcément limiter le modèle choisi et donc sa précision.

Néanmoins cet outil peut être amené à évoluer et on peut donc envisager une certaine déviance par rapport au modèle existant.

3.4.1. Définition du Machine Learning

Le Machine Learning, ou apprentissage automatique, consiste à utiliser des outils issus de l'intelligence artificielle (algorithmes, réseaux de neurones…) d'une base de données, afin d'en obtenir une analyse prédictive. Dans ce type de modélisation, on cherche les corrélations entre les variables explicative d'une base de données d'évènements, et non plus une causalité.

Il existe plusieurs modèles, nous en décrirons certains dans la suite de ce chapitre, qui diffèrent par leur méthode, mais la méthode de modélisation reste la même pour tous, et se résume en 3 grandes étapes:

- La construction de la base de données à analyser, qu'on fractionne en deux échantillons : un échantillon d'apprentissage (80% de l'échantillon global dans notre étude), et un échantillon test (20% de l'échantillon).

- L'apprentissage (ou la modélisation) des données, qui s'effectue sur l'échantillon d'apprentissage afin d'en comprendre la logique

- La réalisation de la prédiction, sur l'échantillon test, avec en général un score de confiance associé

Dans cette partie, la base de données est la même que pour la modélisation des Modèles Linéaires Généralisés.

Nous allons donc directement décrire les modèles utilisés et les étapes de réalisation d'une tarification en Machine Learning.

3.4.2. Définition des méthodes utilisées

3.4.2.a Apprentissage supervisé Vs apprentissage non supervisé :

Il existe deux types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé.

Dans l'apprentissage supervisé, on utilise le résultat de l'échantillon d'apprentissage, ou données de sortie (notre y, autrement dit ici la fréquence ou le coût moyen) afin de produire une fonction inférée, qui peut être utilisée pour mapper de nouveaux exemples (phase de prédiction).

L'apprentissage non supervisé au contraire, les données ne sont pas étiquetées, on ne se sert pas du résultat de l'échantillon d'apprentissage pour faire la modélisation.

3.4.2.b Modèles de prédiction

Notre base de données contient les valeurs constatées en 2018 des fréquences et des coûts moyens associés aux caractéristiques des assurés, nous pouvons donc envisager des modèles d'apprentissage supervisé. Notre objectif étant de modéliser des valeurs numériques continues, nous devons trouver des modèles de régression. Il en existe de nombreux, nous avons limité notre étude aux suivants :

- **Arbre de décision (Decision Tree ou DT)** [11]**:** il consiste à représenter une hiérarchie de la structure des données sous forme de séquence de décisions (tests) pour prédire un résultat.

Ce modèle a l'avantage d'être facile à comprendre, car plutôt intuitif. Il est également facile à interpréter. Les résultats peuvent être présentés et les règles de décision peuvent être ensuite reproduites pour de futurs travaux similaires.

Enfin, l'arbre de décision a l'avantage non négligeable de pouvoir être réalisé dans un temps d'exécution raisonnable. Dans notre quotidien, si nous envisageons de calculer ou comparer nos normes tarifaires via une méthode Machine Learning, il nous faut un algorithme facile d'utilisation.

Nous avons sélectionné l'arbre de décision pour toutes ces raisons.

Néanmoins les défauts de ses qualités sont à prendre en compte dans la fiabilité du modèle : l'arbre de décision a en général un risque de faible performance par rapport à d'autres modèles plus complexes. Il a également un risque de sur apprentissage. Mais ce dernier risque bien que toujours présent sera relativement atténué par le fait que nous avons sélectionné peu de variables explicatives.

Pour notre étude, l'interprétation des résultats est tout aussi importante que la performance du modèle. De plus notre équipe n'est pas formée en data science, il s'agit pour nous d'un concept nouveau sur lequel nous sommes débutant. Commencer par ce genre de modèle est conseillé pour toutes ces raisons.

- Foret aléatoire (ou Random Forest ou RF) [12] **:** l'algorithme des forêts aléatoires effectue un apprentissage sur de multiples arbres de décision entrainés sur des sous-ensembles de données légèrement différents. Les prédictions sont ensuite moyennées lorsque les données sont quantitatives (ou utilisés pour un vote pour des données qualitatives, dans le cas des arbres de classification). L'algorithme des forêts aléatoires est connu pour être un des classificateurs les plus efficaces « out-of-the-box » (c'est-à-dire nécessitant peu de prétraitement des données).

Les Random Forest donnent de bons résultats surtout en grande dimension, et sont simples à mettre en œuvre. Elles ont également peu de paramètres.

On s'attend à priori à de meilleurs résultats qu'avec l'arbre de décision, tout en ayant un modèle exploitable d'un point de vue opérationnel, et encore relativement simple d'utilisation.

Les avantages de ce type de données sont nombreux : facile d'utilisation comme l'arbre de décision, nécessite peu de préparation des données, temps d'exécution raisonnable, ils ont également un bon comportement pour gérer les valeurs extrêmes et les données manquantes.

En revanche ils nécessitent d'avoir une base de données équilibrée, sinon ils ne génèreront pas d'arbre équilibré. Les forêts aléatoires ont également un effet boîte noire », c'est-à-dire qu'elles sont difficilement interprétables et améliorables. Elles nécessitent enfin un entraînement plus long.

- Ridge [13] **:** il s'agit d'un type de régression qui ajoute une contrainte sur les coefficients lors de la modélisation pour maîtriser l'amplitude de leur valeur.

Le modèle Ridge nous semble bien adapté à notre recherche d'une modélisation au plus juste et qui saura gérer toutes les variables explicatives qu'on lui fournira. Il s'agit d'une régression donc elle se rapproche du concept des Modèles Linéaires Généralisés, mais est néanmoins paramétrable avec une méthode Machine Learning.

- Lasso [14] **:** il s'agit d'un modèle linéaire similaire à Ridge, mais avec une fonction de pénalité plus forte :

L'intérêt de Lasso par rapport à Ridge est que ce modèle permet de sélectionner uniquement les variables explicatives pertinentes pour réaliser le modèle, en annulant certains coefficients.

Mais pour notre étude, ce modèle pourrait s'avérer inexploitable. On ne peut en effet envisager de faire un tarif qui serait indépendant du sexe, de l'âge, ou des autres variables que nous étudions ici. Nous avons préalablement limité les variables explicatives de notre étude à celles qui se sont avérées être nécessaires et suffisantes à un tarif fiable jusqu'à aujourd'hui. Mais on peut envisager certaines adaptations, en l'occurrence ici certaines atténuations de nos coefficients si les modèles montrent qu'elles sont nécessaires.

En conséquence, si Lasso s'avère être le meilleur modèle et qu'il ne retient pas certaines variables ou modalités essentielles, nous ne pourrons pas retenir ce modèle car nous devrons tout de même intégrer toutes les variables explicatives dans nos tarifs. Néanmoins l'information pourrait être intéressante afin d'éventuellement remettre en question les valeurs de nos coefficients.

4.4.2.c Critères de performance des modèles

Pour déterminer quel modèle donne la meilleure prédiction, nous calculons les critères suivants pour un échantillon de taille N et de variable observée y_i et de variable estimée $\hat{y}_i$:

La moyenne des erreurs absolues (Mean Absolute Error ou MAE) : moyenne arithmétique des valeurs absolues des écarts entre la valeur réelle et la valeur estimée. Cet indicateur est à la fois le plus simple à calculer et donne un bon aperçu de la qualité du modèle. En effet pour une modélisation du tarif sur un échantillon prospectif connu, il reste le meilleur indicateur, compte-tenu du fait qu'il faudra également comparer les résultats obtenus avec ceux du GLM calculés dans la partie précédente.

Nous retiendrons donc cet indicateur pour définir quel sera le meilleur modèle entre tous.

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

L'erreur quadratique moyenne (en anglais Mean Squared Error ou MSE) : il s'agit d'une mesure caractérisant la précision d'un estimateur. Pour calculer l'erreur quadratique moyenne MSE, les erreurs individuelles sont tout d'abord élevées au carré, puis additionnées les unes aux autres. On divise ensuite le résultat obtenu par le nombre total d'erreurs individuelles, puis on en prend la racine carrée. Cette erreur nous donne une mesure synthétique de l'erreur globale dans une seule valeur.
Sous python le MSE est calculé de manière inversé de manière à obtenir un nombre négatif (NMSE ou Negative Mean Squared Error) : la convention dans les fonctions d'évaluation de modèles sous Python (sklearn) est en effet d'envoyer le modèle ayant la métrique la plus élevée, cela ne pourrait pas marcher avec un MSE normal, d'où le fait de le rendre négatif, où le maximum sera en effet de 0.

Selon ce critère de performance, le meilleur modèle aura un NMSE qui se rapproche le plus de zéro.

$$MSE = \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|^2}{N}$$

La moyenne des erreurs (Mean Error ou ME) : indicateur qui ne donnera pas vraiment de détail sur la qualité du modèle si on cherche un tarif individuel, mais qui peut donner une estimation de la précision dans un modèle collectif, qui lui cherche sa précision en moyenne.

$$ME = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)}{N}$$

Le meilleur modèle est celui qui minimise ces grandeurs.

3.4.3. Optimisation des modèles

Chaque modèle doit être optimisé, c'est-à-dire fournir la meilleure prédiction possible. Pour cela, il faut trouver les paramètres de chaque modèle qui permettent d'obtenir le meilleur résultat (la plus faible erreur absolue dans notre cas).

Pour trouver ces paramètres, nous utilisons la méthode de validation croisée *GridSearchCV* dans python. Il s'agit d'une technique souvent utilisée en Machine Learning pour évaluer la variabilité d'un jeu de données et la fiabilité de tout modèle entrainé avec des données. Cette méthode vise aussi à optimiser les hyper-paramètres du modèle en faisant un balayage exhaustif sur un ensemble de valeurs d'hyper-paramètres.

Pour chaque modèle à tester, les données de l'échantillon d'entrainement sont divisées n fois de manière aléatoire en 2 sous-échantillons : le premier servira à l'apprentissage (représentant un pourcentage des données de l'échantillon global). Le deuxième servira à valider la prédiction.

La sélection du meilleur modèle dépend de la performance qui sera calculée sur l'échantillon d'apprentissage via la méthode de la validation croisée.

Les avantages de cette méthode sont nombreux :

1. La validation croisée utilise la totalité du jeu de données d'entrainement
2. Elle évalue à la fois le jeu de données et le modèle
3. Elle ne mesure pas simplement la justesse d'un modèle. Elle donne également une idée du degré de représentativité du jeu de données et de la sensibilité du modèle aux variations des données.

En revanche, comme elle entraine et valide le modèle plusieurs fois sur un jeu de données plus grand, elle est plus gourmande en ressource de calcul et prends plus de temps que la validation sur un découpage aléatoire.

On obtient pour chaque modèle une version optimisée des hyper-paramètres que nous avons choisi de faire varier :

	Parametre 1	Valeur frequence	Valeur cout moyen	Parametre 2	Valeur frequence	Valeur cout moyen
Decision Tree	max_depth	5	15	min_samples_leaf	100	10
Random Forest	max_depth	5	15	min_samples_leaf	100	10
Lasso	n_alphas	200	5			
Ridge	alphas	0,001	0,01			

66- valeurs des paramètres optimisées des modèles d'apprentissage supervisés

Il nous reste maintenant à tester ces modèles avec l'échantillon test pour déterminer lequel sera le plus fiable pour estimer la fréquence et le coût moyen.

3.4.4. Test sur échantillon réel

Après avoir obtenu les modèles les plus optimisés possibles pour chacun de nos modèles, nous effectuons une prédiction de l'échantillon test qui sert à estimer la qualité réelle de la prédiction.

Nous comparons les résultats obtenus des modèles avec les fréquences et les coûts moyens réellement observés afin de calculer les ME, MSE et MAE et choisir le meilleur modèle de prédiction en Machine Learning.

Résultats pour la fréquence :

	Decision Tree	Random Forest	Ridge	Lasso
ME	0,0277	**0,0269**	0,0352	0,026
MSE	9,4876	9,558	**9,4577**	9,5568
MAE	1,7999	1,814	**1,7998**	1,812

67- résultats des critères de performance pour la fréquence avec les modèles machine learning

Selon les critères de performance que nous avons choisis (le MAE), le modèle de régression du Ridge semble le meilleur pour effectuer notre modélisation de la fréquence.

Néanmoins remarquons que les écarts sont très faibles, notamment avec le modèle Decision Tree. Nous pourrions donc envisager de choisir un autre modèle par soucis pratique ou opérationnel, sans trop affecter la performance de notre estimation.

Ces résultats semblent encourageant lorsqu'on regarde la moyenne des erreurs, qui est faible pour tous les modèles, et qui laisse penser qu'on aura une bonne prédiction de la fréquence en moyenne sur notre portefeuille collectif.

Résultats pour le cout moyen

	Decision Tree	Random Forest	Ridge	Lasso
ME	-0,0639	**-0,017**	-0,0619	-0,0173
MSE	311,067	344,1233	**303,6488**	344,1646
MAE	**13,7175**	15,2994	13,7637	15,2984

68- résultats des critères de performance pour le coût moyen avec les modèles machine learning

Regarder les moyennes des erreurs permet de voir comment le modèle tarifie un groupe, ce qui est plus intéressant pour une tarification collective. Les écarts de tarif négatifs et positifs peuvent se compenser au global, ce qui n'a pas d'incidence sur le tarif final collectif.

Ici le Decision Tree semble être le meilleur choix de prédiction et les écarts sont toujours homogènes entre les modèles, même si on voit qu'à priori Random Forest et Lasso donneront de moins bons résultats en terme de MAE, leurs erreurs moyennes se compensent mieux sur l'ensemble de l'échantillon.

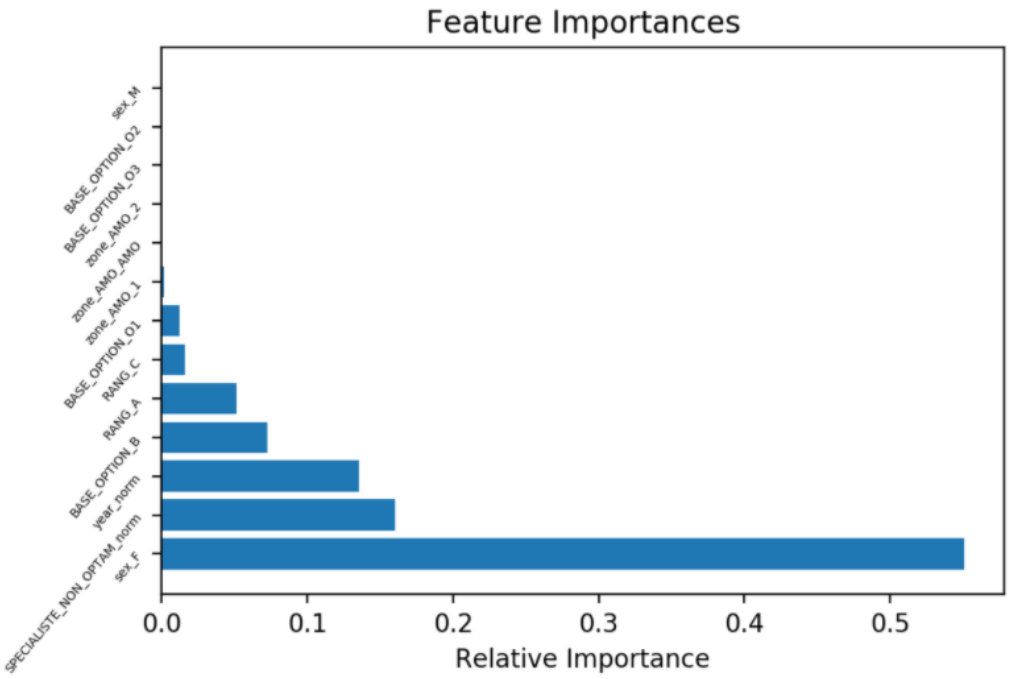
3.4.5. Analyse du meilleur modèle Machine Learning et conclusion sur cette méthode

Les meilleures méthodes pour la prédiction de notre tarif de consultation Spécialiste NON OPTAM sont à priori le Ridge pour la fréquence et le Decision Tree pour le coût moyen.

Notre approche cherche autant à estimer le coût annuel de cette garantie le plus justement possible, que d'avoir un modèle simple d'utilisation. Dans cette considération, et compte-tenu du très faible écart de performance observé pour l'estimation de la fréquence entre le Decision Tree et le Ridge, nous choisissons de faire notre modélisation globale de la fréquence et du coût moyen avec le modèle du Decision Tree.

Nous souhaitons analyser la qualité de cette prédiction d'un point de vue tarifaire, c'est-à-dire en voyant les nuances qu'il aura sur les différentes caractéristiques des assurés.

Impact des variables explicatives sur la fréquence :



69- feature importance obtenue via Decision Tree sur la fréquence

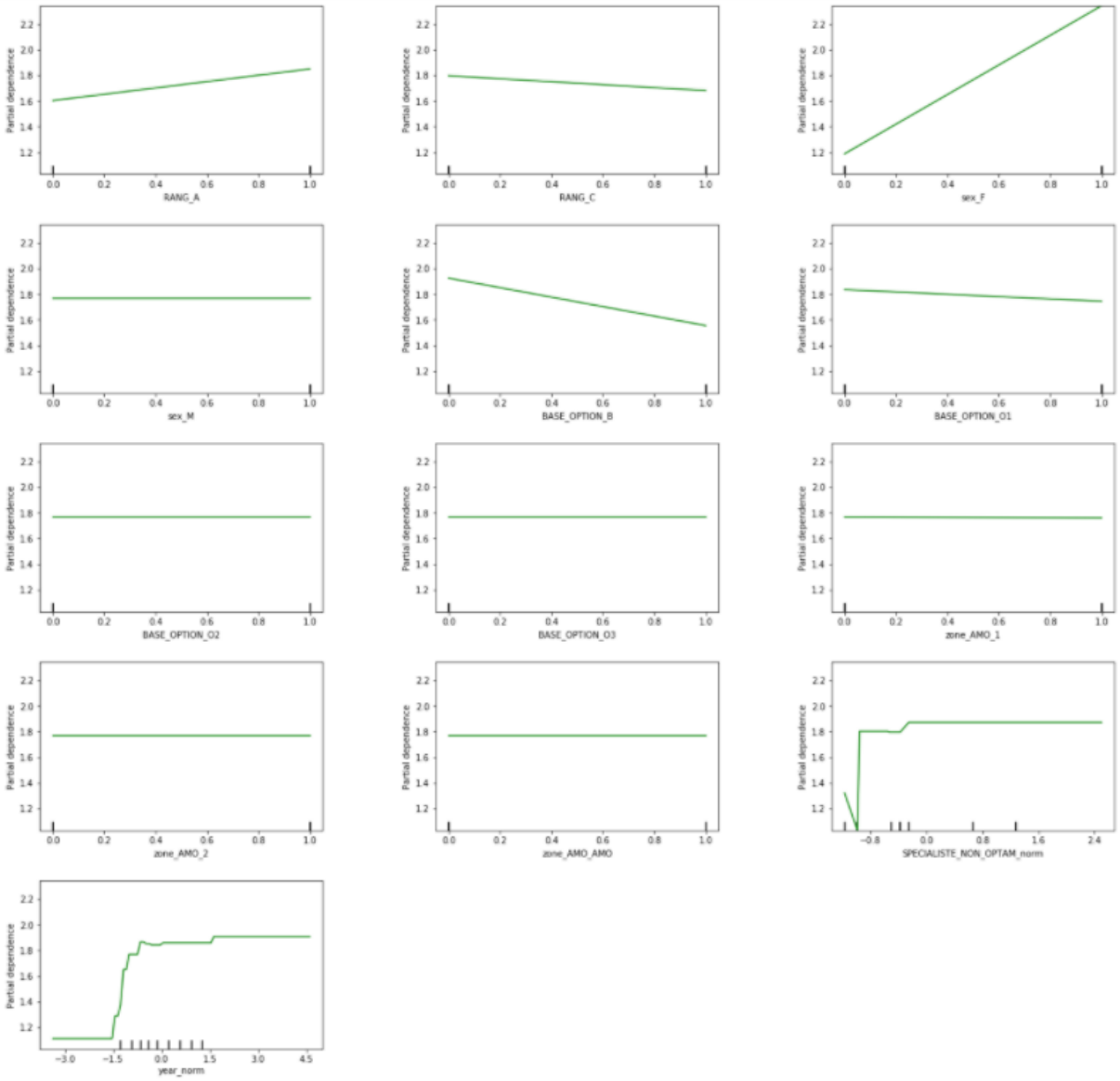
La fonction *feature_importances* de python permet de connaître les variables les plus importantes dans la modélisation d'une variable.

Pour la fréquence, le sexe (féminin), le niveau de garantie (SPECIALIST_NON_OPTAM) ainsi que l'âge sont les variables les plus significatives. En revanche on remarque que la zone de soin ne semble pas affecter beaucoup la fréquence. Ces caractéristiques sont cohérentes avec la modélisation d'une fréquence.

Les graphiques suivants montrent l'influence de chaque variable sur la fréquence (croissance ou décroissance). Ils sont complémentaires du graphique précédent. On peut observer que la fréquence est croissante avec le sexe féminin, l'âge et le niveau de garantie (en moyenne). Certaines variables ne font pas ou très peu varier la fréquence, comme le sexe masculin, l'option 2 et l'option 3.

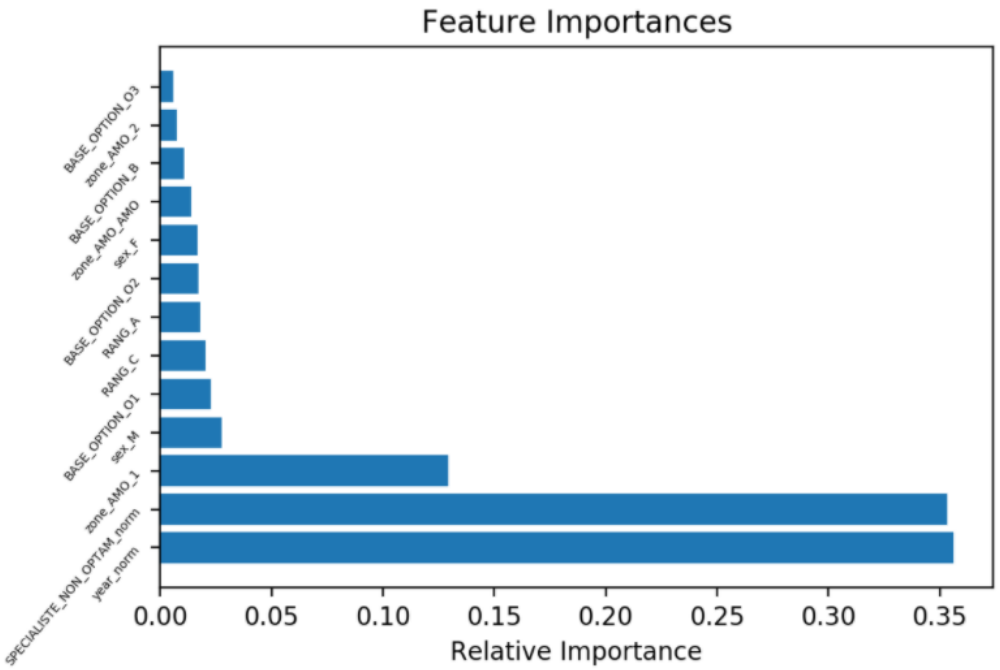
Cela pourra poser des problèmes de calculs pour envisager d'utiliser cette technique pour un tarif : en effet nous avons déjà filtré nos variables explicatives au strict minimum, et nous avons besoin qu'elles aient toutes un impact, même très faible, sur le tarif annuel.

Les mêmes résultats appliqués au coût moyen et le reste de l'analyse confirmeront ces limites.



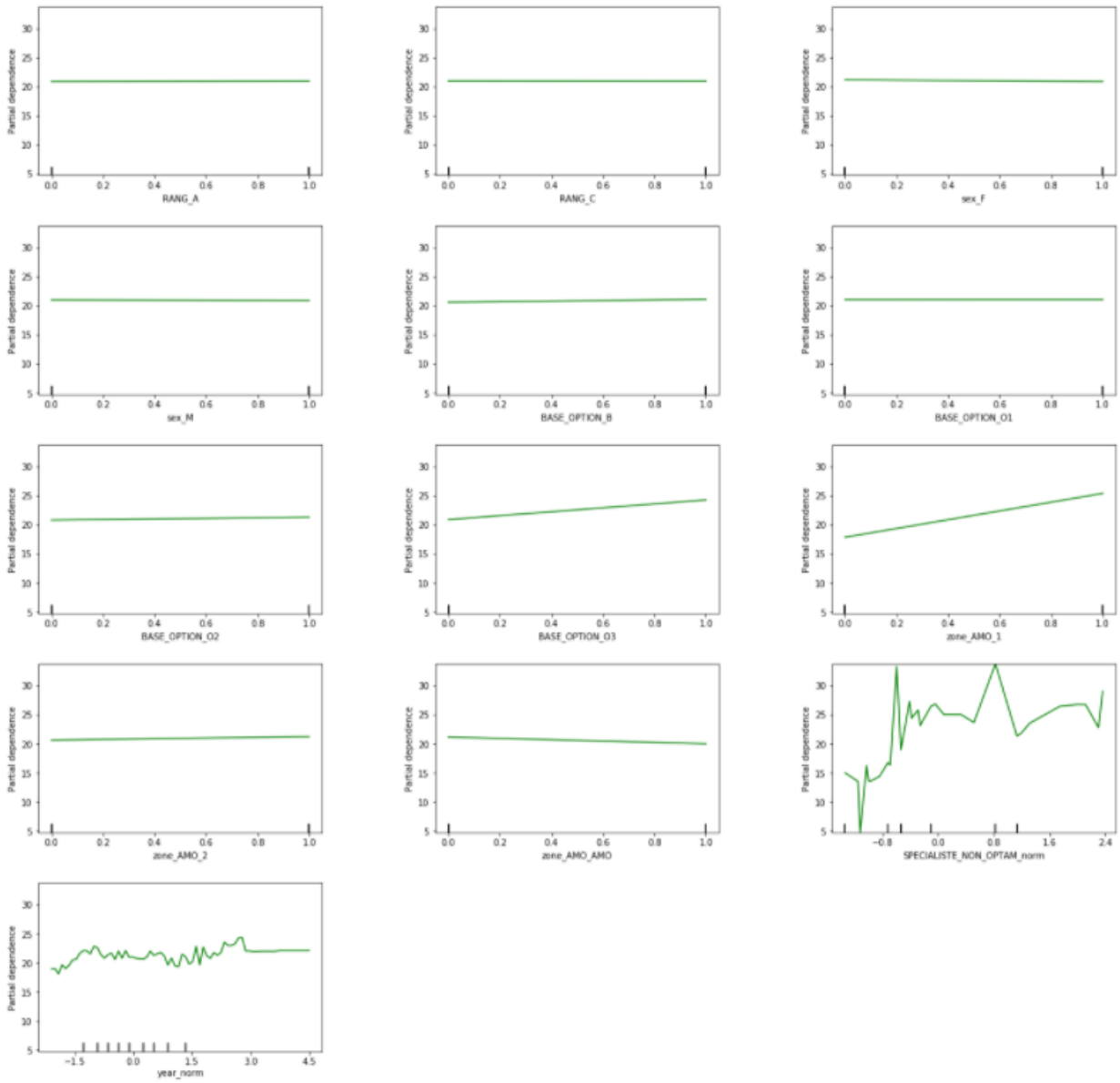
***70**-Influence des modalités des variables explicatives sur la fréquence*

Impact des variables explicatives sur le cout moyen :



71- feature importance obtenue via Decision Tree sur le coût moyen

L'âge et le niveau de garantie sont également les variables les plus significatives pour expliquer les coûts moyens d'un soin spécialiste non OPTAM, avec cette fois-ci la zone (zone 1 dans le graphique) qui arrivent avant le sexe. L'apparition de la zone géographique en tant de variable déterminante est quelque chose d'attendu dans la modélisation d'un coût moyen. On observe en effet des coûts bien différents selon les zones géographiques chez les médecins spécialistes.



72-Influence des modalités des variables explicatives sur le coût moyen

L'influence de l'âge est moins nette côté coût moyen, et la courbe de croissance du coût moyen selon le niveau de garantie est moins lisse que pour la fréquence. Ce genre de courbe peut rendre le tarif annuel final inexploitable car notre tarif doit impérativement être croissant avec le niveau de garantie. Un travail de lissage pourra s'avérer nécessaire. On peut cependant confirmer que le coût moyen sera croissant comme nous l'avons modélisé avec Antenia selon les zones, et selon les niveaux d'option (même si l'option 1 et l'option 2 ont une courbe de croissance très faible). L'influence du rang du bénéficiaire n'est pas significative.

Au final, malgré une bonne représentativité de certaines variables, le coût moyen n'apparait pas proportionnel à toutes nos variables tarifaires, comme pour la fréquence. Cela peut poser problème si les effets ne se compensent pas lorsqu'on les multiplie pour obtenir le tarif annuel final.

Analyse des coûts annuels modélisés:

Nous avons observé les caractéristiques de la fréquence et du coût moyen, mais pour bien analyser la pertinence de la modélisation effectuée via un Décision Tree, nous analysons le coût annuel final obtenu par leur produit. Pour cela, nous prenons un échantillon type d'assurés avec des caractéristiques précises sur leur âge, sexe, zone, niveau d'option, type de bénéficiaires, et nous simulons un tarif sur les mêmes niveaux que ceux que nous avons dans notre outil de tarification. Nous pouvons ainsi faire nos premiers commentaires sur le tarif qui serait proposé via un Decision Tree, en le comparant avec notre tarificateur, d'un point de vue prime pure, et d'un point de vue coefficient.

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
	Niveau Garantie (BR)												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
Machine Learning	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	9,34 €	9,60 €	**12,08 €**	12,30 €	15,18 €	9,60 €	14,49 €	25,29 €	12,08 €	10,38 €	12,08 €	15,18 €
	0,6	17,70 €	59,60 €	**30,64 €**	9,82 €	15,71 €	38,26 €	59,60 €	30,64 €	17,92 €	36,31 €	18,99 €	15,71 €
	1,3	19,76 €	53,32 €	**41,94 €**	8,73 €	40,23 €	64,04 €	51,18 €	21,80 €	16,07 €	35,22 €	14,18 €	40,23 €
	2	39,26 €	104,11 €	**52,52 €**	18,98 €	42,50 €	125,05 €	62,73 €	40,48 €	40,48 €	77,58 €	10,83 €	69,82 €
	3,7	23,64 €	171,78 €	**88,31 €**	28,32 €	117,41 €	206,34 €	100,63 €	20,65 €	20,65 €	94,53 €	37,22 €	117,41 €
	5,3	35,55 €	76,53 €	**39,34 €**	12,90 €	52,31 €	91,93 €	57,47 €	20,60 €	20,60 €	29,57 €	16,58 €	103,71 €

73- Comparaison du coût annuel obtenu entre Antenia et Décision Tree

Les primes annuelles calculées via fréquence x coût moyen du modèle Decision Tree optimisé sont intéressantes. Elles sont globalement dans le même ordre de grandeur que notre tarificateur en terme d'intervalle.

Quelques points d'attentions s'observent, notamment sur le point correspondant à notre prime pure (assuré principal de 40 ans de sexe masculin):

- Les tarifs de certains niveaux de garanties ne respectent pas la monotonie de croissance de la courbe selon les niveaux de garantie. Néanmoins rappelons que le tarif Antenia a été lissé (parfois plusieurs fois) pour que cette règle soit respectée. Mais cela reste un désavantage de cette méthode face au GLM que nous observerons dans la partie suivante. On peut expliquer ce résultat par la répartition de nos assurés sur les niveaux de garantie : nous manquons d'observations sur les garanties très élevées, et les personnes qui sont assurées sur ces hauts niveaux ne consomment pas forcément à la hauteur de leur garantie.
  La fiabilité de cette méthode de modélisation ne doit pas affecter le côté pratique de cette tarification, et doit respecter les règles implicites d'un tarif santé, notamment sa croissance avec le niveau de garantie.
- Les tarifs sont globalement plus élevés dans la modélisation via Machine Learning qu'avec notre tarificateur, sauf pour le point 5,3xBR.
- Les niveaux de garantie faibles ont parfois des tarifs très élevés, pas forcément cohérents avec la réalité des courbes de consommation et de tarifs pratiqués en assurance. La répartition des assurés sur les niveaux de garantie faibles est ici aussi une explication de ce résultat.

- Le tarif manque globalement « d'étalement », c'est-à-dire qu'il y a peu d'écart entre le tarif de la garantie minimale et le tarif d'une garantie élevée.

D'une manière générale, les résultats sont cohérents, mais demanderont un travail de lissage supplémentaire pour avoir une courbe fiable et exploitable en terme de tarif.

Pour évaluer l'effet des caractéristiques des assurés, il est intéressant d'observer les coefficients liés à chaque changement par rapport au tarif de la prime pure de l'assuré type :

	RANG	A	A	A	C	A	A	A	A	A	A	A	A
	sex	M	F	M	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	B	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	1	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	40	40	40	40	40	40	40	50	60	40
	IMPACT	AGE 20 ANS	SEXE FEMME	SSURE MOYE	CONJOINT	OPTION 1	OPTION 2	EMME ZONE	ZONE 2	ZONE AMO	/ME ZONE A	AGE 60 ANS	OPTION 3
ANTENIA	NIVEAU GARANTIE												
Cout annuel	0,3	0,621	2,011	1	0,93	1,05	2,11	2,01	1,00	0,42	0,86	0,57	1,10
	0,6	0,621	2,011	1	0,93	1,05	2,11	1,61	0,80	0,56	1,15	0,76	1,10
	1,3	0,621	2,011	1	0,93	1,05	2,11	1,90	0,94	0,90	1,85	1,23	1,10
	2	0,621	2,011	1	0,93	1,05	2,11	1,74	0,87	0,85	1,75	1,16	1,10
	3,7	0,621	2,011	1	0,93	1,05	2,11	1,80	0,90	0,88	1,80	1,19	1,10
	5,3	0,621	2,011	1	0,93	1,05	2,11	1,65	0,82	0,80	1,64	1,09	1,10
Machine Learning													
Cout annuel	0,3	0,8	0,8	1,0	1,0	1,3	0,8	1,2	2,1	1,0	0,9	1,0	1,3
	0,6	0,6	1,9	1,0	0,3	0,5	1,2	1,9	1,0	0,6	1,2	0,6	0,5
	1,3	0,5	1,3	1,0	0,2	1,0	1,5	1,2	0,5	0,4	0,8	0,3	1,0
	2	0,7	2,0	1,0	0,4	0,8	2,4	1,2	0,8	0,8	1,5	0,2	1,3
	3,7	0,3	1,9	1,0	0,3	1,3	2,3	1,1	0,2	0,2	1,1	0,4	1,3
	5,3	0,9	1,9	1,0	0,3	1,3	2,3	1,5	0,5	0,5	0,8	0,4	2,6

74- Comparaison des coefficients de variables explicatives obtenus entre Antenia et Decision Tree

Les tarifs du modèle Decision Tree ne sont pas linéaires avec le niveau de garantie, donc on ne peut pas retrouver de coefficient unique pour chaque caractéristique. C'est une autre limite de ce type de modélisation compte-tenu du format de notre outil de tarification.

Néanmoins les coefficients observés nous permettent de vérifier si les tarifs vont dans le même sens selon les caractéristiques des assurés. On peut ainsi confirmer que les femmes consomment presque deux fois plus de consultations spécialistes en cout annuel. Les jeunes (20 ans) consomment moins que les assurés types. Les effets des zones, des options et de l'âge sont ainsi confirmés par la modélisation. Ces résultats sont encourageants pour plusieurs raisons : la première, c'est que notre tarificateur paraît cohérent en terme de tarif, du moins autant qu'un tarificateur plus technique tel que le Decision Tree. La seconde, c'est qu'on peut commencer à envisager de d'utiliser ce modèle de Machine Learning pour tester la fiabilité de notre modélisation de primes pures, étant donné que les résultats vont dans le bon sens, et ce même sans lisser la courbe du Decision Tree.

Les écarts que nous observons entre les deux types de modèles, par exemple sur le tarif des conjoints qui sont sous-tarifés avec une modélisation Machine Learning, seront à confirmer avec le produit du coût moyen modélisé, et à analyser globalement avec un échantillon d'assurés qui servira à comparer les trois types de modélisation étudiées dans ce sujet.

Conclusion :

Le modèle Machine Learning est un modèle intéressant car il permet, une fois en place, de calculer des primes pures cohérentes et retransmet bien les impacts des variables explicatives.

Il demeure cependant plus complexe sur plusieurs points :

- en terme de code : il faut connaitre les modèles à étudier et leur spécificité, tenir compte du temps de calcul, et éviter le sur-apprentissage, et aussi maîtriser le langage de programmation dans lequel il est codé.

- Contrairement au GLM, le Decision Tree ne permet pas d'avoir des coefficients fixes pour expliquer les variables. Il restera donc parfois un gros travail de lissage pour la courbe des primes pures, qui paraît plus difficile que pour notre modélisation interne actuelle, et pour le modèle Machine Learning notamment.

- selon les modèles choisis, certaines variables explicatives peuvent être négligées par construction, ce qui ne serait pas cohérent avec la réalité de nos tarifs. L'ajustement des coefficients ou des primes pures serait alors à faire, ce qui constituerait un travail supplémentaire non négligeable.

- un autre sujet apparaît complexifier un petit peu notre travail : le modèle fréquence x cout moyen demande deux fois plus de ressources (code, temps de travail, résultats, calculs à mettre en place pour obtenir le coût annuel).

Toutes ces remarques ainsi que la fiabilité du modèle Machine Learning doivent être confirmées avec une comparaison objective, que nous ferons à la fin de cette étude, afin de savoir quel modèle apparaît de manière générale comme le plus juste et le plus exploitable pour une tarification d'une norme santé.

3.5. Comparaison des méthodes de modélisation

Nous avons estimé un tarif annuel pour les consultations spécialistes NON OPTAM via trois techniques très différentes, et toutes pertinentes.

Afin de comparer objectivement ces techniques, et donc vérifier que notre outil de tarification est fiable et exploitable, nous devons vérifier les écarts de chacun de ces modèles avec la réalité.

Pour cela, nous prenons notre échantillon 2018 de coûts annuels constatés sur notre portefeuille d'assurés (hors enfant) et nous simulons un tarif annuel avec chacun des modèles retenus :

- **Antenia** (outil de tarification interne) : coût annuel
- **GLM** : fréquence estimée avec Poisson, coût moyen estimé avec une loi Log-normale
- **Machine Learning** : fréquence et coût moyen estimés avec Decision Tree

3.5.1. Résultats globaux

Nous calculons ensuite la MAE, MSE, et ME des grilles de résultats obtenus :

	Réels	Antenia	Machine Learning	GLM
Moyenne coûts annuels	32,78	34,13	44,73	32,52
ME		1,36	11,96	- 0,26
MAE		25,64	29,93	29,15
MSE		1 223,10	1 496,46	1 587,52

75-Résultats des critères de performance pour chaque modélisation

Si on regarde la moyenne des erreurs, c'est finalement le GLM qui est le modèle le plus proche de la réalité avec une moyenne proche de zéro (-0,26€), suivi de près par Antenia (1,36€ d'écart de tarif en moyenne). Le modèle en Machine Learning est beaucoup plus éloigné du coût annuel réellement constaté. Il surtarifie les coûts comme nous l'avions analysé dans la partie 4.4.5. Mais son MAE est dans le même ordre de grandeur que les autres modèles. On peut donc estimer que la précision des modèles est bonne, et nous pouvons tous les qualifier de fiable pour analyser la qualité de notre modèle de tarification interne.

Si on analyse les résultats par rapport au MAE, c'est Antenia qui a les meilleurs résultats.

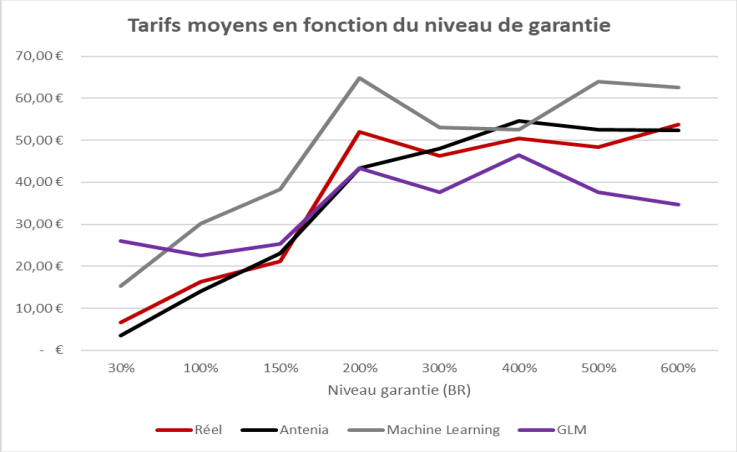
A partir de ces premiers résultats, nous pouvons estimer que au global sur l'ensemble du portefeuille étudié notre modélisation interne est fiable et de précision satisfaisante par rapport aux modèles obtenus avec des méthodes modernes. C'est déjà une bonne nouvelle, mais il faut la confirmer avec une analyse plus poussée des résultats.

3.5.2. Résultats par segment de variable explicative

On peut regarder les résultats des MAE par segment de modalité de variable explicative : ainsi on fige une modalité précise, par exemple le niveau de garantie, et on regarde quel modèle sera le plus juste sur cette tranche d'assurés.

NIVEAU GARANTIE (BR)		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
30%	1%	6,68 €	3,49 €	15,34 €	26,00 €	- 3,19 €	8,66 €	19,32 €	-48%	130%	289%	6,75 €	11,78 €	24,25 €
100%	7%	16,33 €	14,01 €	30,12 €	22,56 €	- 2,32 €	13,79 €	6,24 €	-14%	84%	38%	13,27 €	21,65 €	20,89 €
150%	50%	21,25 €	23,07 €	38,30 €	25,29 €	1,82 €	17,05 €	4,04 €	9%	80%	19%	19,67 €	26,90 €	22,55 €
200%	5%	52,09 €	43,34 €	64,92 €	43,41 €	- 8,74 €	12,83 €	- 8,68 €	-17%	25%	-17%	35,95 €	39,19 €	41,75 €
300%	6%	46,35 €	48,11 €	53,09 €	37,57 €	1,77 €	6,74 €	- 8,78 €	4%	15%	-19%	33,09 €	33,11 €	35,99 €
400%	24%	50,46 €	54,54 €	52,52 €	46,45 €	4,08 €	2,06 €	- 4,02 €	8%	4%	-8%	35,47 €	34,64 €	38,14 €
500%	3%	48,32 €	52,63 €	64,03 €	37,68 €	4,31 €	15,71 €	- 10,64 €	9%	33%	-22%	35,01 €	38,31 €	37,83 €
600%	4%	53,83 €	52,32 €	62,58 €	34,62 €	- 1,51 €	8,75 €	- 19,21 €	-3%	16%	-36%	39,36 €	37,90 €	42,31 €

76-Etude des écarts par tranche de niveaux de garanties



77- Tarif moyen en fonction du niveau de garantie pour les différentes modélisations

L'analyse des erreurs par tranche de niveau de garantie est très importante. Un bon modèle doit fournir un prix croissant avec les niveaux de garantie. Ici les tarifs sont parfois décroissants mais

cela est dû à l'hétérogénéité des modalités des autres variables explicatives qu'on a mélangée (hommes et femmes en ensemble, tout niveau d'options confondus, tout âge confondu etc.). Les courbes ont toutes une tendance croissante. On voit nettement qu'Antenia fournit la courbe la plus proche de la réalité, suivi du GLM.

Le modèle Machine Learning sur tarifie pour tous les niveaux de garantie. Antenia et GLM sont antagonistes sur leur modélisation : GLM sur tarifie les niveaux faibles, et sous-tarifie les niveaux élevés (à partir de 200%BR), tandis qu'Antenia a tendance à sous tarifer les niveaux faibles et sur tarifie les niveaux élevés. Cependant ces écarts ne sont pas significatifs.

La moyenne des erreurs absolue est meilleure pour Antenia quel que soit le niveau de garantie, et surtout ce modèle reste plus proche de la réalité quand on regarde les moyennes des erreurs.

Les constats sont globalement les mêmes pour l'analyse des variables type bénéficiaire, zone, niveau d'option et âge :

BENEFICIAIRE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
A	72%	33,67 €	33,95 €	46,88 €	44,53 €	0,28 €	13,21 €	10,87 €	1%	39%	32%	25,23 €	30,82 €	29,08 €
C	28%	30,52 €	34,60 €	39,28 €	1,96 €	4,09 €	8,76 €	- 28,56 €	13%	29%	-94%	26,69 €	27,68 €	29,33 €

78-Etude des écarts par rang de l'assuré

ZONE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Tarif réel	Tarif Antenia	Tarif Machine Learning	Tarif GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
1	39%	38,49 €	36,51 €	55,01 €	42,24 €	- 1,99 €	16,51 €	3,74 €	-5%	43%	10%	28,24 €	34,65 €	34,13 €
2	50%	29,93 €	33,64 €	39,62 €	27,82 €	3,71 €	9,69 €	- 2,11 €	12%	32%	-7%	23,76 €	27,26 €	26,14 €
AMO	11%	25,34 €	27,77 €	31,41 €	19,23 €	2,43 €	6,07 €	- 6,11 €	10%	24%	-24%	25,03 €	25,35 €	25,17 €

79-Etude des écarts par zone

<u>Remarque sur les options :</u> Les options en tant que variable explicatives sont relativement contestables. En effet, chaque niveau d'option est lié au niveaux de sa garantie de base. Un niveau d'option à 200%BR n'aura donc pas le même impact si la base est à 150%BR ou à 30%BR. Néanmoins nous avons choisi de les représenter sous cette forme afin de les représenter en tant que variable explicatives, car nous n'avions pas la qualité des données nécessaire par salarié pour analyser les options autrement, ni la quantité de données suffisante pour chiffrer le coût de l'option en fonction de chaque niveau de couverture possible croisé avec chaque niveau de base possible.

Les tarifs des options sont tous sous tarifés tandis que les tarifs des contrats de base sont sur-tarifés (+50% au moins). Pour cette variable explicative aussi, il faudrait pousser l'analyse, pour les mêmes raisons que pour le sexe. Nous n'avons probablement pas eu un échantillon assez grand pour représenter au mieux les tarifs des options, néanmoins comparé au GLM qui fournit le meilleur modèle statistique de référence, Antenia s'aligne plutôt bien avec la réalité, compte-tenu du fait que ses coefficients datent eux aussi de la création du modèle tarifaire interne.

BASE/OPTION		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
B	46%	14,43 €	21,49 €	34,63 €	22,92 €	7,06 €	20,20 €	8,49 €	49%	140%	59%	15,53 €	24,67 €	18,22 €
O1	42%	47,89 €	45,38 €	51,63 €	40,93 €	- 2,50 €	3,74 €	- 6,95 €	-5%	8%	-15%	34,03 €	33,80 €	37,46 €
O2	12%	49,73 €	43,18 €	58,82 €	39,49 €	- 6,54 €	9,09 €	- 10,24 €	-13%	18%	-21%	34,89 €	36,45 €	41,52 €
O3	0%	69,50 €	42,86 €	70,51 €	64,64 €	- 26,63 €	1,01 €	- 4,86 €	-38%	1%	-7%	41,49 €	35,60 €	53,55 €

80- Etude des écarts par niveau d'option

AGE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
0-24 ans	4%	21,01 €	20,28 €	21,00 €	34,56 €	- 0,72 €	- 0,01 €	13,56 €	-3%	0%	65%	17,99 €	19,42 €	24,84 €
25-34 ans	25%	33,22 €	35,32 €	44,53 €	33,70 €	2,10 €	11,32 €	0,49 €	6%	34%	1%	26,50 €	30,62 €	28,34 €
35-44 ans	29%	34,05 €	34,54 €	45,14 €	30,72 €	0,49 €	11,09 €	- 3,32 €	1%	33%	-10%	26,73 €	30,38 €	29,10 €
45-54 ans	26%	32,35 €	33,01 €	46,08 €	31,24 €	0,67 €	13,74 €	- 1,11 €	2%	42%	-3%	24,77 €	30,11 €	29,15 €
55-64 and	13%	34,11 €	36,09 €	48,18 €	36,47 €	1,98 €	14,07 €	2,36 €	6%	41%	7%	25,39 €	30,09 €	31,14 €
65 ans et +	2%	28,43 €	39,86 €	45,36 €	31,20 €	11,42 €	16,92 €	2,77 €	40%	60%	10%	26,35 €	30,91 €	34,11 €

81- Etude des écarts par tranche d'âge

Globalement Antenia reste le modèle le plus homogène et proche de la réalité selon le MAE.

Si on analyse les écarts selon les erreurs moyennes en revanche, le GLM fournit de meilleurs résultats pour le sexe. Antenia semble sur-tarifer les femmes et sous-tarifer les hommes en moyenne. C'est un résultat intéressant qu'il faudra analyser, notamment car les coefficients utilisés pour la modélisation Antenia ont été calculés lors de la création de la norme, et n'ont pas été remis à jour depuis.

Ce qui est intéressant est également de constater que ces erreurs de tarification se compensent en moyenne dans Antenia, permettant une erreur de tarif globale tout sexe confondu proche de zéro.

SEXE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
F	60%	35,93 €	43,01 €	55,03 €	39,11 €	7,08 €	19,11 €	3,19 €	20%	53%	9%	27,43 €	33,64 €	32,18 €
M	40%	28,01 €	20,68 €	29,12 €	22,52 €	- 7,33 €	1,11 €	- 5,50 €	-26%	4%	-20%	22,95 €	24,31 €	24,55 €

82- Etude des écarts par sexe

3.5.3. Analyse des grands écarts

Nous pouvons terminer l'analyse de la fiabilité du modèle par une étude des grands écarts entre valeurs prédites et coûts annuels réellement constatés. Il est normal que certains assurés aient parfois été modélisés avec des erreurs importantes, étant donné que tous les modèles fournissent un tarif plafonné, alors que dans la réalité les assurés peuvent être amenés à consommer de nombreux soins courants, selon leur état de santé.

Le nombre de cas pour lesquels les modèles ont fourni des erreurs supérieures à 2 fois le tarif réellement observé (en valeur absolue) représente soit l'hétérogénéité de la modalité observée dans cet échantillon, et/ou les « grandes erreurs » que les modèles ont pu faire. Ce n'est pas un indicateur de stabilité précis, car les grands écarts peuvent être dus à des consommations exceptionnellement élevées de la part de certains assurés, que nos modèles ne sauraient représenter que dans la moyenne de toutes les modalités. Mais l'analyse reste intéressante car elle permet de voir où il y a eu des pics d'erreurs.

L'analyse montre les erreurs supérieures à deux fois le tarif réel, par niveau de modalité étudié, en nombre et en pourcentage, pour un total de 41417 assurés étudiés.

SEXE		Nbre valeurs réelles dont ecart > 200%					
	Poids de la modalité	Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
F	60%	779	3%	1599	7%	1898	8%
M	40%	1563	10%	1341	8%	1599	10%

BENEFICIAIRE		Nbre valeurs réelles dont ecart > 200%					
		Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
A	72%	1589	6%	2195	8%	1807	6%
C	28%	446	4%	1095	10%	1473	13%

AGE		Nbre valeurs réelles dont ecart > 200%					
	Poids de la modalité	Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
24	3,75%	132	9%	138	9%	193	13%
34	25,17%	574	6%	713	7%	693	7%
44	29,34%	741	6%	829	7%	1020	9%
54	26,26%	609	6%	956	9%	1121	11%
64	13,08%	267	5%	402	8%	544	10%
120	2,40%	76	8%	131	14%	193	20%

NIVEAU GARANTIE		Nbre valeurs réelles dont ecart > 200%					
0	Poids de la modalité	Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
30%	1%	43	9%	172	36%	370	77%
100%	7%	261	9%	640	22%	593	20%
150%	50%	1434	7%	3913	19%	2493	12%
200%	5%	43	2%	89	4%	131	6%
300%	6%	49	2%	111	5%	121	5%
400%	24%	207	2%	221	2%	362	4%
500%	3%	23	2%	44	4%	53	5%
600%	4%	39	2%	46	3%	87	6%

BASE/OPTION		Nbre valeurs réelles dont ecart > 200%					
	Poids de la modalité	Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
B	46%	2013	11%	5907	32%	3853	21%
O1	42%	522	3%	481	3%	845	5%
O2	12%	123	2%	195	4%	325	7%
O3	0%	3	5%	0	0%	2	3%

ZONE		Nbre valeurs réelles dont ecart > 200%					
	Poids de la modalité	Nbre Antenia	% de la modalité	Nbre Machine Learning	% de la modalité	Nbre GLM	% de la modalité
1	39%	790	13%	997	38%	1299	25%
2	50%	1265	3%	1776	2%	1891	4%
AMO	11%	614	3%	632	5%	628	8%

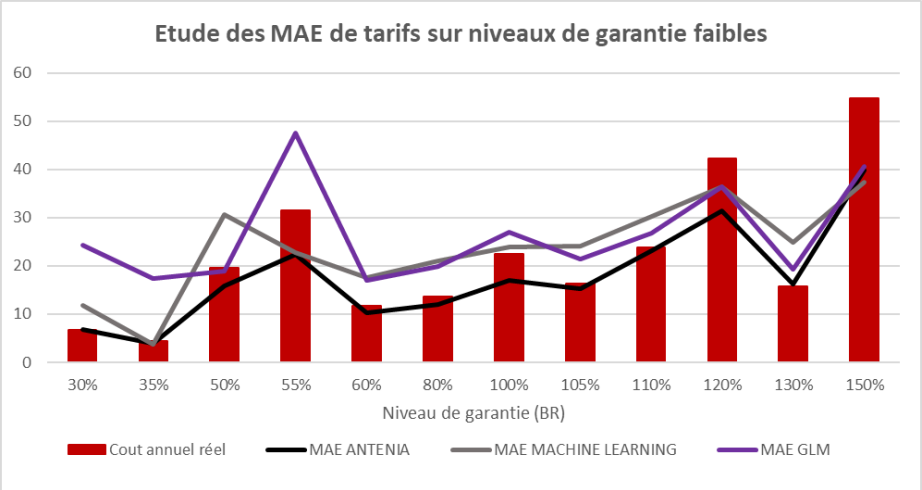
83- Etude des valeurs extrêmes sur les différentes variables explicatives

Antenia ne contient pas de grands écarts notables (maximum 13% de chaque modalité avec des écarts dépassant le seuil de l'analyse). Du côté des modèles modernes, les seules variables qui retiennent notre attention pour le GLM et le Machine Learning sont les faibles niveaux de garantie (jusqu'à 150%BR), ainsi que les modalités « Base » et « Zone 1 ».

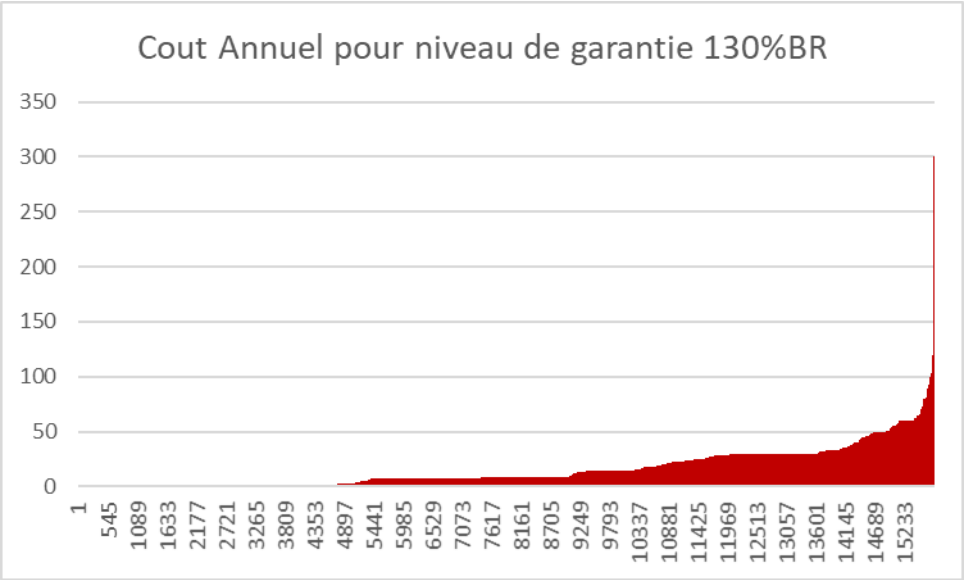
En zoomant sur les faibles niveaux de garantie, on voit qu'il y a certains niveaux remarquables avec de forts écarts. Ces données mettent en relief la partie délicate du lissage des points et du regroupement des niveaux de garantie. On voit que le lissage effectué coté Antenia permet d'atténuer les pics de prestations en les moyennant.

NIVEAU GARANTIE (BR)		Moyenne Tarif				Nbre valeurs réelles dont ecart < 200%					
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Nbre Antenia	% Antenia	Nbre Machine Learning	% Machine Learning	Nbre GLM	% GLM
30%	1%	6,68 €	3,49 €	15,34 €	26,00 €	43	9%	172	36%	370	77%
35%	0%	4,34 €	4,58 €	5,65 €	20,14 €	8	8%	12	12%	69	70%
50%	1%	19,63 €	8,46 €	43,73 €	19,53 €	82	19%	118	27%	38	9%
55%	0%	31,42 €	11,73 €	43,66 €	64,58 €	1	10%	0	0%	2	20%
60%	2%	11,71 €	10,93 €	22,03 €	17,75 €	93	10%	225	24%	246	27%
80%	1%	13,70 €	14,77 €	29,28 €	19,54 €	40	7%	168	30%	147	26%
100%	2%	22,50 €	20,78 €	35,02 €	31,05 €	31	4%	125	15%	161	19%
105%	0%	16,24 €	19,00 €	33,21 €	25,62 €	11	6%	52	28%	32	17%
110%	3%	23,88 €	19,80 €	34,43 €	22,63 €	160	12%	226	17%	214	16%
120%	2%	42,31 €	22,52 €	48,77 €	29,90 €	64	9%	54	7%	56	8%
130%	39%	15,80 €	22,56 €	35,27 €	24,41 €	919	6%	4479	28%	2869	18%
150%	5%	54,78 €	29,87 €	61,23 €	32,31 €	71	4%	26	1%	90	4%

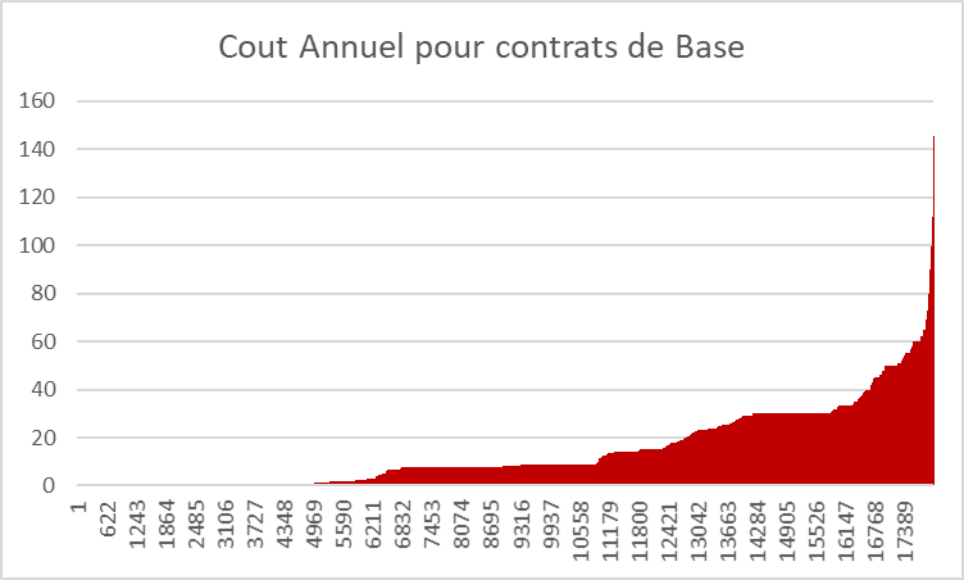
84- Détail des valeurs extrêmes sur les niveaux bas de garantie



85- Etude des MAE des tarifs sur les bas niveaux de garanties



86- Fonction de répartition des coûts annuels réels pour les assurés sur un niveau 130%BR



87- Fonction de répartition des coûts annuels réels des assurés sur un contrat de base

Les écarts extrêmes des zones 1 et de la modalité BASE_OPTION sont certainement explicables par la forte représentativité de ces données dans les échantillons, et du biais qui ont dû être faits pour avoir ces informations (zonier non remis en question et manque d'information sur les niveaux d'options).

4.5.4. Conclusion sur l'analyse des résultats

Les résultats montrent sans équivoques que notre modélisation interne est fiable.

Les techniques modernes de modélisation (Machine Learning et GLM) donnent des résultats satisfaisants pour les échantillons, mais sont finalement moins proches de la réalité que notre modélisation Antenia. Compte-tenu de leur complexité (hypothèses fortes, calculs, code), il aurait fallu de bien meilleurs résultats de leur part pour envisager une véritable remise en question de notre processus de création de primes pures santé.

Si le lissage effectué uniquement pour notre modélisation interne peut être la cause de ces meilleurs résultats, il n'empêche qu'avec tous les biais acceptés au départ (non remise en question des coefficients de sexe, âge, bénéficiaire, zonier, option), le résultat prouve la bonne fiabilité et la maîtrise de notre technique de tarification.

Les techniques modernes ont tout de même soulevé quelques retouches à envisager dans notre modélisation : les coefficients de sexe sont à revoir (sur-tarification des femmes, sous-tarification des hommes), et des niveaux de garantie (sous –tarification des niveaux faibles et sur-tarification des niveaux forts).

Ces corrections sont exactement ce que nous cherchions à analyser lorsque nous avons pensé notre étude : vérifier la robustesse de notre modèle de normes, et trouver les éventuelles points d'attention, pour l'amélioration continue de nos processus de tarification et de statistiques.

La comparaison a été limitée à la seule garantie des médecins spécialistes non OPTAM, par soucis évident de temps (nettoyage des données et analyse) et d'efficacité (création d'un processus de vérification de fiabilité opérationnel). Il restera à réutiliser ce type d'étude pour une gamme de garanties plus larges, afin d'officialiser son efficacité et envisager son automatisation dans nos processus de création et de suivi des normes tarifaires santé, toujours dans l'optique d'une amélioration continue, autant pour la norme que pour la vérification de la fiabilité.

Conclusion

La tarification d'un contrat sur-mesure est un processus en perpétuelle évolution, les assureurs cherchant à estimer le coût du risque au plus juste face à la grande compétitivité qui existe entre eux.

Cette concurrence pousse les assureurs à se renouveler sans cesse, cherchant des techniques de plus en plus sophistiquées tout en s'assurant que les caractéristiques fondamentales propres à ce secteur sont respectées. On ne remettra pas en cause les variables explicatives telles que le sexe ou l'âge. On ne contestera pas non plus que le tarif doit être croissant avec le niveau de garantie de l'assuré.

Lorsqu'une modélisation est faite, quelle que soit la méthode puisqu'il en existe aujourd'hui une grande variété, il faut sans cesse s'assurer de sa fiabilité. Cette partie du travail est tout aussi complexe que de réaliser une nouvelle tarification, car une méthode mathématique a été choisie et un processus opérationnel a été mis en place, qui implique souvent un outil de tarification contenant les calculs associés à la méthode. Il s'agit donc de ne pas tout remettre en question, afin que le coût et le temps passé sur l'étude tarifaire soient contrôlés et pérennes. C'est tout l'objet de l'étude réalisée et décrite dans ce rapport : comment s'assurer que le tarif prédit, avec une technique choisie et intégrée à l'organisation d'une entreprise, soit fiable et au plus juste ?

Nous avons tenté de répondre à cette question à l'aide des techniques de tarification telles que les Modèles Linéaires Généralisés et le Machine Learning, qui sont très en vogue ces dernières années.

Mais il a fallu cadrer cette étude, en limitant le choix des variables tarifaires notamment, contrainte imposée par la forme de la modélisation de notre outil de tarification interne, qui isole la prime pure d'un assuré principal de 40 ans de sexe masculin, sur un contrat de base dans une zone géographique définie, et qui calcule la prime pure d'un assuré de caractéristiques différentes en multipliant cette prime pure de base avec chaque coefficient correspondant aux caractéristiques qui diffèreraient de cet assuré type. Ces coefficients sont pour certains communs à un même poste de santé (Hospitalisation, Soins courants…) et nous avons donc dû faire le choix de les conserver lors du back-testing de notre tarificateur, afin d'éviter trop de dispersion dans notre étude. Cela a forcément constitué un biais dans le résultat de la prime, dont nous cherchons à étudier la fiabilité.

Le modèle Linéaire Généralisé et le machine Learning sont des techniques assez puissantes, qui permettent d'obtenir des résultats facilement explicables. Cependant comparer notre nouveau tarif avec les résultats de ces techniques s'est avéré très long, notamment étant donné qu'il a fallu nettoyer à nouveau les données pour les mettre au format fréquence x cout moyen. Ce choix est assumé mais il est vrai que nous aurions pu envisager de modéliser un coût annuel. Ce sera sûrement le cas lors de nos prochaines études de fiabilité.

La prise en main du langage python a également ajouté une couche de complexité à notre étude. Il s'agit d'un langage choisi pour ses performances coté Machine Learning et aussi réputé simple d'utilisation. Il a tout de même fallu se former à ce langage et comprendre les résultats qui ont été générés lors des modélisations. Néanmoins une fois le langage écrit, nous pouvons envisager de le réutiliser lors de nos études de fiabilité futures.

Enfin, il a fallu analyser les résultats obtenus via ces deux méthodes statistiques afin de s'assurer qu'elles étaient bien des modèles fiables pour la tarification et qu'elles pouvaient servir de référence à l'analyse de notre norme Antenia. Les résultats ont d'ailleurs montré que le Decision Tree, modèle du Machine Learning, bien qu'obtenu sans les difficultés observées pour le GLM, donne de moins bons résultats que le GLM. Ce dernier a donc été retenu en référence plutôt que le Machine Learning lors de l'analyse finale de la norme Antenia.

Les résultats de cette étude permettent de conclure sur la fiabilité de notre norme interne. Les modèles ont donné des résultats satisfaisants pour un tarif collectif, mais aussi pour l'analyse des influences de la plupart des variables explicatives étudiées. Cela encourage la confiance que nous avons pour notre technique de modélisation interne.

L'analyse de fiabilité a remonté quelques retouches à envisager, notamment celle des coefficients sexe et les tarifs de faibles niveaux de garantie. Pour confirmer ces retouches il conviendra de refaire la même analyse avec les autres garanties du poste Soins courant dont fait partie la Consultation du Médecin Spécialiste (Consultation Médecin Généraliste, Imagerie Médicale, Actes Techniques Médicaux…).

C'était la première fois que nous étudiions la fiabilité de nos normes et plus largement de notre modèle de tarification, et pour cela, l'étude était aussi intéressante pour les résultats observés que par l'instauration de la méthode qui nous a servi à vérifier cette fiabilité avec les techniques actuarielles modernes. Il faudra la développer au fur et à mesure des études qui seront faites dans le futur, notamment en y incluant les bénéficiaires de type Enfant, qui ont manqué à l'étude par manque de qualité de données.

Bibliographie

[1] « La délégation de gestion en assurances de personnes » : Institut Français de l'Audit et du Contrôle Internes, Paris mai 2012, p.92.

[2] Site web vie-publique.fr : www.vie-publique.fr

[3]La maîtrise des dépassements d'honoraires: https://www.ameli.fr/medecin/exercice-liberal/remuneration/maitrise-depassements/maitrise-depassements

[4] Les dépassements d'honoraires et l'OPTAM : https://www.ameli.fr/fileadmin/user_upload/documents/DP_OPTAM.pdf

[5] : Médecins libéraux: le taux de dépassement d'honoraires moyen a encore reculé en 2018 : https://www.medecin-occitanie.org/medecins-liberaux-le-taux-de-depassement-dhonoraires-moyen-a-encore-recule-en-2018/?cli_action=1585818092.342

[6] https://drees.solidarites-sante.gouv.fr/IMG/pdf/cns2019.pdf

[7] Introduction au modèle linéaire général : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf

[8] Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité : https://www.institutdesactuaires.com/global/gene/link.php?news_link=mem%2F46dd866818e6465b650f8cec38a75327.pdf&fg=1

[9] Tests non paramétriques: http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-inf-np.pdf

[10] LE MACHINE LEARNING ET SON APPLICATION DANS LE SECTEUR BANQUE ET ASSURANCE : https://www.insurancespeaker-wavestone.com/2016/07/machine-learning-application-domaine-banque-assurance/#:~:text=Sa%20d%C3%A9finition%20n%C3%A9cessite%20d'%C3%AAtre,pr%C3%A9dictive%20%C3%A0%20partir%20de%20donn%C3%A9es.

[11] Arbre de décision, comment ça marche? - Lovely Analytics : https://www.lovelyanalytics.com/2016/08/16/decision-tree-comment-ca-marche/#:~:text=Qu%27est%20ce%20qu%27un,ou%20une%20cat%C3%A9gorie%20(classement).

[12] Random Forest - Data Analytics Post : https://dataanalyticspost.com/Lexique/random-forest/

[13]Régression régularisée - Ridge, Lasso, Elasticnet (univ-lyon2.fr) : https://eric.univ-lyon2.fr/~ricco/cours/slides/regularized_regression.pdf

[14] Lasso (statistiques) — Wikipédia (wikipedia.org): https://fr.wikipedia.org/wiki/Lasso_(statistiques)

Liste des figures

Note de synthèse

Sujet du mémoire : Back-testing d'une norme tarifaire santé sur-mesure suite à la mise en place de l'OPTAM/OPTAM-CO et évaluation de sa fiabilité au moyen de techniques actuarielles modernes.

Par: Cécile HUBERT

Introduction et contexte

En assurance, la tarification au plus juste est une condition essentielle pour la maîtrise du risque. Chez Swiss Life, au sein de la Direction des Assurances collectives (DAC), l'outil de tarification des contrats santé collectifs sur-mesure permet un paramétrage manuel par les équipes chargées d'études actuarielles des primes pures $\boldsymbol{PP_i}$ de chaque garantie associée à un assuré type (adhérent principal âgé de 40 ans, de sexe masculin et sur un contrat de base) pour une zone, un niveau de garantie et une assiette donnée, ainsi que des coefficient multiplicatifs $\boldsymbol{Coefficientk(i)}$ communs aux grands postes santé qui personnaliseront le tarif de tous les profils d'assuré selon leur caractéristiques (âge, sexe, rang, niveau d'option, CSP, Code NAF, taille de l'entreprise). La multiplication des éléments donne la prime de risque annuelle (ou norme tarifaire) :

$$\boldsymbol{PR_i = PP_i(assiette_i, niveau_i, régime, zone) * \prod_{i=1}^{Nombre\ de\ coeff} Coefficientk(i)}$$

Après une construction globale des primes pures et des coefficients dans l'outil en 2014, le format imposé par l'OAV a souvent eu pour impact de limiter les études de tarification à la prime pure de l'assuré type.

Ce fut notamment le cas lors des évolutions réglementaires concernant l'Option de Pratique Tarifaire Maîtrisée (OPTAM), qui a changé les règles de remboursements des complémentaires santé sur quelques garanties des postes Hospitalisation et Soins Courants.

Nous nous sommes alors posés la question suivante : dans un contexte concurrentiel où la nécessité d'avoir un tarif toujours actualisé se confronte aux difficultés de la réalité opérationnelle (format de saisie, disponibilité des données, temps de travail), comment s'assurer que les tarifs sont toujours justes suite aux ajustements progressivement effectués ?

Nous avons alors décidé d'analyser la fiabilité de la norme tarifaire médecin spécialiste non OPTAM.

Pour cela, nous avons choisi de comparer notre tarification avec celle que nous aurions obtenu via deux techniques modernes couramment citées et utilisées en actuariat : un Modèle Linéaire Généralisé et un modèle d'apprentissage supervisé (Machine Learning).

Données utilisées :

Pour effectuer l'analyse, nous avons récupéré les données de l'exercice 2018 des contrats collectifs sur-mesure gérés en interne chez Swiss Life ainsi que ceux gérés par son plus grand délégataire (Génération).

Après homogénéisation, nettoyage et analyse des données, il a été décidé de retenir uniquement les variables tarifaires suivantes :

- Age
- Sexe
- Rang
- Niveau d'option
- Zone géographique de l'assuré
- Niveau de garantie

Les autres variables tarifaires dont l'OAV tient compte étant soit inaccessibles, soit sans impact observé pour le tarif d'une garantie santé.

Les modélisations faites via GLM et apprentissage supervisé seront faites selon un modèle fréquence x coût moyen. Il a de plus été remarqué que les enfants n'avaient pas leur âge et leur sexe bien paramétré dans les bases de données, nous les avons donc exclus de la modalité Rang pour ces deux modélisations.

Bilan des données utilisées :

		Antenia - Global	Antenia – Assuré Type	GLM & ML
Age moyen	Global	42	40	41,44
	Hommes	44	40	44
	Femmes	40	40	40

88- Age moyen entre les différentes bases de données

Les assurés sont représentés selon la répartition suivante :

Variable	Modalité	% BDD GLOBALE	% BDD Assuré Type	% BDD GLM - ML
SEXE	Masculin (M)	59,80%	53,49%	50,10%
	Féminin (F)	40,20%	46,51%	49,90%
TYPE BENEFICIAIRE	Adhérent (A)	50,30%	66,4%	72,10%
	Conjoint (C)	16,90%	33,6%	27,90%
	Enfant (E)	32,70%		
ZONE	Zone 1 (Z1)	35,30%	41,04%	33,80%
	Zone 2 (Z2)	54,70%	58,96%	55,90%
	Zone Alsace-Moselle (AMO)	10,10%		10,40%
BASE_OPTION	Base (B)	44,70%	49,70%	49,20%
	Option 1 (O1)	43,10%	42 ,5%	39,50%
	Option 2 (O2)	11,90%	7,70%	11,20%
	Option 3(O3)	0,20%	0,10%	0,10%

89-Répartition des assurés par variable explicative catégorielle

Critères de performance :

Si nous analysons la fiabilité de notre modélisation interne avec des modélisations statistiques plus classique, il faut cependant valider la propre validité de ces modèles. Pour cela nous réalisons les tests nécessaires pour définir quel type de modèle convient le mieux à nos échantillons de données.

Pour le GLM, la meilleure loi sera choisie en regardant la log vraisemblance, la déviance résiduelle, les critères AIC et BIC.

Pour le modèle machine learning, chaque type de modèle sera optimisé selon ses propres paramètres (exemple: maximum de profondeur pour le Decision Tree), puis le meilleur modèle sera choisi selon les critères du Mean Absolute Error (ou MAE – moyenne des erreurs absolues entre les valeurs estimées et les

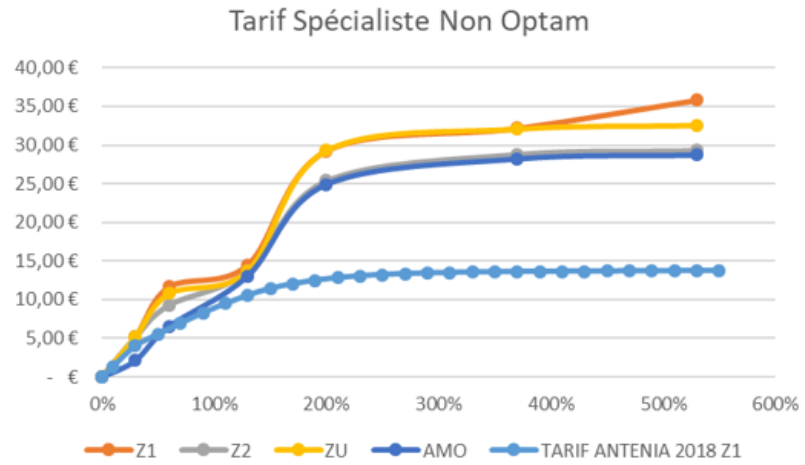
valeurs réelles), Mean Squarred Error (ou MSE – moyenne des carrés des erreurs) ainsi que du Mean Error (ou ME - moyenne des erreurs). Le meilleur modèle sera celui qui donnera les indicateurs les plus proches de zéro.

Le MAE et MSE sont des indicateurs simples pour estimer la qualité de prédiction du modèle. Le ME permet d'avoir la précision sur un modèle collectif, qui cherche une précision moyenne et pas individuelle.

Enfin, **l'analyse de fiabilité** sera faite à l'aide de ces 3 derniers indicateurs.

Résultats obtenus avec le modèle interne :

Une fois lissée, on projette le tarif pour l'assuré type qui définit notre prime pure et sur un éventail d'assurés avec des caractéristiques différentes pour tenir compte des coefficients de variables explicatives:



90-Tarif de la garantie Médecin Spécialiste non OPTAM avant et après back-testing sur plusieurs zones distinctes

	RANG	A	A	A	C	A	A	A	A	A	A	A	A	
	sex	M	F	M	M	M	F	F	M	M	F	M	M	
	BASE_OPTION	B	B	B	B	O1	O2	B	B	B	B	B	O3	
	zone_AMO	1	1	1	1	1	1	2	2	AMO	AMO	AMO	1	
	year	20	40	40	40	40	40	40	40	40	50	60	40	
MODELE	Niveau Garantie (BR)													
ANTENIA	0	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €	
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €	
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €	
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €	
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €	
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €	
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €	

Nouveaux coûts Antenia calculés sur plusieurs profils d'assurés

Par rapport au tarif en place, le tarif est en forte croissance, et la courbe montre une inflexion au niveau du point 130%BR. C'est précisément la limite des remboursements autorisés dans le cadre d'un contrat responsable pour les complémentaires santé lorsque le spécialiste n'est pas adhérent à l'OPTAM. On observe donc bien un changement de comportement.

Résultats obtenus avec le GLM :

Malgré les critères de performances choisis, la réalité opérationnelle d'un tarif nous a forcé à choisir de modéliser la fréquence avec une loi de Poisson et les coûts moyens avec une loi log-normale.

Les résultats obtenus sur un éventail d'assurés de caractéristiques différentes sont les suivants:

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	**Niveau Garantie (BR)**												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
POISSON x LOGNORMALE	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	17,94 €	41,63 €	21,65 €	0,86 €	38,67 €	96,42 €	30,29 €	15,75 €	11,95 €	25,24 €	14,42 €	85,47 €
	0,6	18,08 €	41,95 €	21,82 €	0,87 €	38,97 €	97,17 €	30,52 €	15,87 €	12,04 €	25,43 €	14,53 €	86,13 €
	1,3	18,41 €	42,71 €	22,21 €	0,89 €	39,67 €	98,92 €	31,07 €	16,16 €	12,26 €	25,89 €	14,79 €	87,69 €
	2	18,74 €	43,48 €	22,61 €	0,90 €	40,39 €	100,70 €	31,63 €	16,45 €	12,48 €	26,36 €	15,06 €	89,27 €
	3,7	19,57 €	45,41 €	23,61 €	0,94 €	42,18 €	105,17 €	33,04 €	17,18 €	13,03 €	27,53 €	15,73 €	93,23 €
	5,3	20,39 €	47,31 €	24,60 €	0,98 €	43,94 €	109,56 €	34,41 €	17,90 €	13,58 €	28,68 €	16,38 €	97,12 €

91- Comparaison des coûts annuels Antenia Vs GLM sur plusieurs profils d'assurés

Les différences entre les deux tarifs sont notables : d'une part, les tarifs obtenus via GLM ne dépendent presque pas du niveau de garantie, ce qui pose des problèmes, notamment aux bornes des niveaux : tarifs trop élevés sur les niveaux faibles, et tarifs surement trop faibles pour les niveaux élevés. Les couts moyens n'ont pas été compensés par les fréquences pour ce problème.

Une différence notable se voit aussi dans les montants des niveaux hauts, qui sont soit plus élevés soit plus bas de manière significative.

Le tarif annuel est aussi extrêmement faible pour les assurés conjoints via le GLM.

En revanche toutes les variables explicatives font évoluer le tarif annuel dans le même sens.

Résultats obtenus avec l'apprentissage supervisé :

Tout comme pour le GLM, la recherche du meilleur modèle d'apprentissage supervisé a dans la pratique reconduit à choisir le Décision Tree pour la fréquence et pour le coût moyen, pour sa simplicité et pour le faible écart de performance qu'il présentait face au meilleur modèle observé en fréquence.

L'analyse du coût annuel obtenu sur l'éventail d'assurés est la suivante :

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
	Niveau Garantie (BR)												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
Machine Learning	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	9,34 €	9,60 €	**12,08 €**	12,30 €	15,18 €	9,60 €	14,49 €	25,29 €	12,08 €	10,38 €	12,08 €	15,18 €
	0,6	17,70 €	59,60 €	**30,64 €**	9,82 €	15,71 €	38,26 €	59,60 €	30,64 €	17,92 €	36,31 €	18,99 €	15,71 €
	1,3	19,76 €	53,32 €	**41,94 €**	8,73 €	40,23 €	64,04 €	51,18 €	21,80 €	16,07 €	35,22 €	14,18 €	40,23 €
	2	39,26 €	104,11 €	**52,52 €**	18,98 €	42,50 €	125,05 €	62,73 €	40,48 €	40,48 €	77,58 €	10,83 €	69,82 €
	3,7	23,64 €	171,78 €	**88,31 €**	28,32 €	117,41 €	206,34 €	100,63 €	20,65 €	20,65 €	94,53 €	37,22 €	117,41 €
	5,3	35,55 €	76,53 €	**39,34 €**	12,90 €	52,31 €	91,93 €	57,47 €	20,60 €	20,60 €	29,57 €	16,58 €	103,71 €

92- Comparaison du coût annuel obtenu entre Antenia et Décision Tree

Les primes annuelles calculées via fréquence x cout moyen du modèle Decision Tree optimisé sont intéressantes. Elles sont globalement dans le même ordre de grandeur que notre tarificateur en terme d'intervalle.

Quelques points d'attentions s'observent, notamment sur le point correspondant à notre prime pure (assuré principal de 40 ans de sexe masculin): les tarifs ne sont pas croissants avec les niveaux de garantie pour certains points, ils sont également plus élevés (sauf pour le point 5,3xBR), les faibles et hauts niveaux de garanties. Enfin, le tarif manque globalement « d'étalement », c'est-à-dire qu'il y a peu d'écart entre le tarif de la garantie minimale et le tarif d'une garantie élevée, et les faibles niveaux sont très élevés

D'une manière générale, les résultats sont cohérents, mais demanderaient un travail de lissage supplémentaire pour avoir une courbe fiable et exploitable en terme de tarif.

Etude de la fiabilité de la norme interne Swiss Life

Nous avons obtenu les meilleures modélisations possibles via le GLM et l'apprentissage supervisé, et nous en connaissons les avantages et les limites. Nous allons maintenant analyser objectivement la qualité de tarification de notre norme interne en la comparant avec ces deux autres méthodes.

Pour cela, nous tarifons un même échantillon d'assurés (hors enfant, et nous analysons les écarts entre les différentes prédictions.

	Réels	Antenia	Machine Learning	GLM
Moyenne coûts annuels	32,78	34,13	44,73	32,52
ME		1,36	11,96	- 0,26
MAE		25,64	29,93	29,15
MSE		1 223,10	1 496,46	1 587,52

93-Résultats des critères de performance pour chaque modélisation

Si on regarde la moyenne des erreurs, c'est finalement le GLM qui est le modèle le plus proche de la réalité avec une erreur moyenne proche de zéro (-0,26€), suivi de près par Antenia (1,36€ d'écart de tarif en moyenne). Le modèle en Machine Learning est beaucoup plus éloigné du coût annuel réellement constaté. Il sur tarifie les coûts comme nous l'avions analysé. Mais son MAE est dans le même ordre de grandeur que les autres modèles. On peut donc estimer que la précision des modèles est bonne, et nous pouvons tous les qualifier de fiable pour analyser la qualité de notre modèle de tarification interne.

Si on analyse les résultats par rapport au MAE, c'est Antenia qui a les meilleurs résultats.

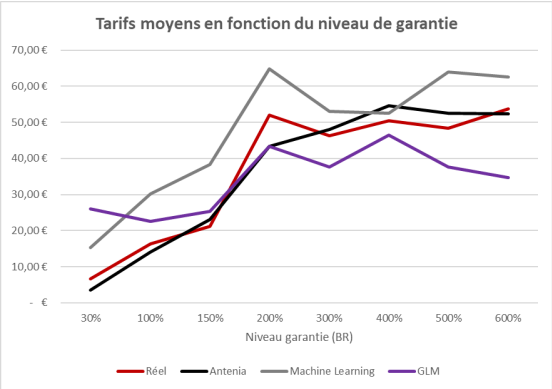
A partir de ces premiers résultats, nous pouvons estimer que au global sur l'ensemble du portefeuille étudié notre modélisation interne est fiable et de précision satisfaisante par rapport aux modèles obtenus avec des méthodes modernes. C'est déjà une bonne nouvelle, mais il faut la confirmer avec une analyse plus poussée des résultats.

Résultats par segment de variable explicative

On peut regarder les résultats des MAE par segment de modalité de variable explicative : ainsi on fige une modalité précise, par exemple le niveau de garantie, et on regarde quel modèle sera le plus juste sur ces tranches d'assurés.

NIVEAU GARANTIE (BR)		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
30%	1%	6,68 €	3,49 €	15,34 €	26,00 €	- 3,19 €	8,66 €	19,32 €	-48%	130%	289%	6,75 €	11,78 €	24,25 €
100%	7%	16,33 €	14,01 €	30,12 €	22,56 €	- 2,32 €	13,79 €	6,24 €	-14%	84%	38%	13,27 €	21,65 €	20,89 €
150%	50%	21,25 €	23,07 €	38,30 €	25,29 €	1,82 €	17,05 €	4,04 €	9%	80%	19%	19,67 €	26,90 €	22,55 €
200%	5%	52,09 €	43,34 €	64,92 €	43,41 €	- 8,74 €	12,83 €	- 8,68 €	-17%	25%	-17%	35,95 €	39,19 €	41,75 €
300%	6%	46,35 €	48,11 €	53,09 €	37,57 €	1,77 €	6,74 €	- 8,78 €	4%	15%	-19%	33,09 €	33,11 €	35,99 €
400%	24%	50,46 €	54,54 €	52,52 €	46,45 €	4,08 €	2,06 €	- 4,02 €	8%	4%	-8%	35,47 €	34,64 €	38,14 €
500%	3%	48,32 €	52,63 €	64,03 €	37,68 €	4,31 €	15,71 €	- 10,64 €	9%	33%	-22%	35,01 €	38,31 €	37,83 €
600%	4%	53,83 €	52,32 €	62,58 €	34,62 €	- 1,51 €	8,75 €	- 19,21 €	-3%	16%	-36%	39,36 €	37,90 €	42,31 €

94-Etude des écarts par tranche de niveaux de garanties



95- Tarif moyen en fonction du niveau de garantie pour les différentes modélisations

Pour la plupart des analyses par variable explicative, on voit nettement qu'Antenia fourni la courbe la plus proche de la réalité, suivi du GLM. Le modèle Machine Learning sur tarifie pour tous les niveaux de garantie. Antenia et GLM sont antagonistes sur leur modélisation : GLM sur tarifie les niveaux faibles, et sous-tarifie les niveaux élevés (à partir de 200%BR), tandis qu'Antenia a tendance à sous tarifer les niveaux faibles et sur tarifie les niveaux élevés. Cependant ces écarts ne sont pas significatifs.

La moyenne des erreurs absolue est meilleure pour Antenia quel que soit le niveau de garantie, et surtout ce modèle reste plus proche de la réalité quand on regarde les moyennes des erreurs.

Les constats sont globalement les mêmes pour l'analyse des variables type bénéficiaire, zone, niveau d'option et âge. En revanche Antenia semble sur tarifer les femmes et sous tarifer les hommes, mais avec des niveaux qui se compensent globalement.

SEXE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
F	60%	35,93 €	43,01 €	55,03 €	39,11 €	7,08 €	19,11 €	3,19 €	20%	53%	9%	27,43 €	33,64 €	32,18 €
M	40%	28,01 €	20,68 €	29,12 €	22,52 €	- 7,33 €	1,11 €	- 5,50 €	-26%	4%	-20%	22,95 €	24,31 €	24,55 €

96- Etude des écarts par sexe

Etude des grands écarts

Nous pouvons terminer l'analyse de la fiabilité du modèle par une étude des grands écarts. Il est normal que certains assurés aient parfois été modélisés avec des erreurs importantes, étant donné que tous les modèles fournissent un tarif plafonné, alors que dans la réalité les assurés peuvent être amenés à consommer de nombreux soins courants, selon leur état de santé.

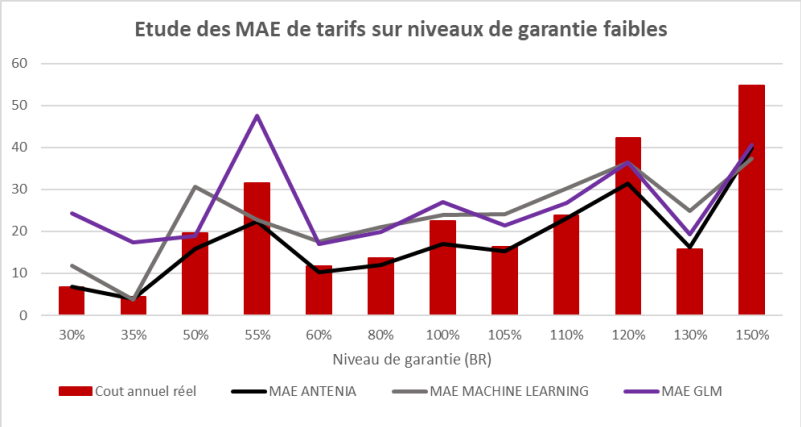
Le nombre de cas pour lesquels les modèles ont fourni des erreurs supérieures à 2 fois le tarif réellement observé (en valeur absolue) peut représenter l'hétérogénéité de la modalité observée dans cet échantillon, et/ou les « grandes erreurs » que les modèles ont pu faire. Ce n'est pas un indicateur de stabilité précis, car les grands écarts peuvent être dus à des consommations exceptionnellement élevées de la part de certains assurés, que nos modèles ne sauraient représenter que dans la moyenne de toutes les modalités. Mais l'analyse reste intéressante car elle permet de voir où il y a eu des pics d'erreurs.

L'analyse montre les erreurs supérieures à deux fois le tarif réel, par niveau de modalité étudié, en nombre et en pourcentage, pour un total de 41417 assurés étudiés

La conclusion de cette analyse est que Antenia ne contient pas de grands écarts notables (maximum 13% de chaque modalité avec des valeurs extrêmes). Du côté des modèles modernes, les seules variables qui retiennent notre attention pour le GLM et le Machine Learning sont les faibles niveaux de garantie (jusqu'à

150%BR), ainsi que les modalités « Base » et « Zone 1 ». Le zoom sur une partie des modalités de chaque variable explicative précise permet d'analyser plus précisément les éventuels écarts remarqués.

On voit notamment que le lissage effectué coté Antenia sur les niveaux de garantie permet d'atténuer les pics de prestations en les moyennant.



97- Etude des MAE des tarifs sur les bas niveaux de garanties

Les autres écarts remarqués sur la zones 1 et de la modalité BASE_OPTION sont certainement explicables par la forte représentativité de ces données dans les échantillons, et du biais qui ont dû être faits pour avoir ces informations (zonier non remis en question et manque d'information sur les niveaux d'options).

Conclusion

Après avoir réalisé le back-testing d'une garantie, qui a dû être limité à un choix réduit de variables explicatives pour des raisons opérationnelles (format de l'outil de paramétrage final) et techniques (disponibilité et qualité des données, études passées montrant l'indépendance du tarif par rapport à certaines variables explicatives dans notre portefeuille), nous avons étudié sa fiabilité en comparant le nouveau tarif avec celui que nous aurions obtenu avec deux techniques de tarification : le GLM et l'apprentissage supervisé.

L'analyse de fiabilité demande tout d'abord de s'assurer que les techniques dites « de référence » sont bien fiables elles-mêmes. Nous avons vu dans notre analyse qu'un GLM et un apprentissage supervisé avaient à la fois leurs avantages et leurs inconvénients en terme d'interprétation des résultats.

L'étude montre que notre norme interne est plutôt fiable, et met en relief quelques mises à jour potentielles de nos coefficients (sexe et niveau d'option notamment).

Mots clés : *Mots clés : assurance santé, tarification, prime pure, GLM, apprentissage supervisé, machine learning, fréquence x cout moyen*

Synthesis note

Subject: Health pure premium Back-testing following the OPTAM/OPTAM-CO set-up and study of its reliability with modern actuarial techniques

Author: Cécile HUBERT

Introduction and context

In insurance, correct pricing is an essential condition to manage risk. At Swiss Life, within the Directorate of Group Insurance (DAC), the bespoke group health contract pricing tool allows manual setting by the actuarial research teams of pure premiums $\boldsymbol{PP_i}$ associated with a typical insured (main member aged 40, male and on a basic contract) for a given area, guarantee level and base, as well as multi-factors $\boldsymbol{Coefficientk(i)}$ common to large health posts that will customize the rate of all insured profiles according to their characteristics (age, gender, rank, option level, CSP, NAF code, company size) for each guarantee, The multiplication of the elements gives the annual risk premium (or tariff standard):

$$\boldsymbol{PR_i = PP_i(assiette_i, niveau_i, régime, zone) * \prod_{i=1}^{Nombre\ de\ coeff} Coefficientk(i)}$$

After a global construction of pure premiums and coefficients in the tool in 2014, the format imposed by the OAV has often had the impact of limiting pricing studies to the pure premium of the typical insured.

This was particularly the case during regulatory developments concerning the Controlled Tariff Practice Option (OPTAM), which changed the rules for reimbursement of health supplements on some guarantees of hospitalization and current care positions.

It leaded us to ask ourselves the following question: in a competitive environment where the need for an always-updated tariff deals with the difficulties of operational reality (entry format, availability of data, working time), how can we ensure that rates are always fair considering the adjustments gradually made?

We then decided to analyze the reliability of the no-OPTAM specialist doctor price.

To do this, we chose to compare our pricing with the one we would have obtained using two modern techniques commonly cited and used in actuarial science: a Generalized Linear Model (GLM) and a Supervised Learning Model (Machine Learning).

Data used:

To carry out the analysis, we collected data from the 2018 financial year of the in-house custom collective contracts at Swiss Life as well as those managed by its largest delegate.

After data homogenization, cleaning and analysis, it was decided to retain only the following pricing variables:

- age
- sex
- rank

- Option level
- The insured's geographical area
- Guarantee level

The other pricing variables that the tool takes into account are either inaccessible or have no observed impact on a health guarantee price.

Modelling done via GLM and supervised learning will be done using a frequency x average cost model. It was also noted that children were not their age and gender well set up in the databases, so we excluded them from the Rank modality for these two models.

Review of the data used:

		Antenia - Global	Antenia - Type Insured	GLM and ML
Average age	Global	42	40	41,44
	Men (M)	44	40	44
	Women (F)	40	40	40

98- Average age between the different data bases used

Policyholders are split according to the following distribution:

Variable	Modalité	% BDD GLOBALE	% BDD Assuré Type	% BDD GLM - ML
SEXE	Masculin (M)	59,80%	53,49%	50,10%
	Féminin (F)	40,20%	46,51%	49,90%
TYPE BENEFICIAIRE	Adhérent (A)	50,30%	66,4%	72,10%
	Conjoint (C)	16,90%	33,6%	27,90%
	Enfant (E)	32,70%		
ZONE	Zone 1 (Z1)	35,30%	41,04%	33,80%
	Zone 2 (Z2)	54,70%	58,96%	55,90%
	Zone Alsace-Moselle (AMO)	10,10%		10,40%
BASE_OPTION	Base (B)	44,70%	49,70%	49,20%
	Option 1 (O1)	43,10%	42 ,5%	39,50%
	Option 2 (O2)	11,90%	7,70%	11,20%
	Option 3(O3)	0,20%	0,10%	0,10%

99-policy holders slit by categorical explanatory variable

Performance criteria:

If we analyze the reliability of our internal modeling with more conventional statistical modeling, but we must validate the own validity of these models first. To do this, we carry out the necessary tests to define which type of model is best suited to our data samples.

For the GLM, the best law will be chosen by looking at the likelihood log, residual deviance, AIC and BIC criteria.

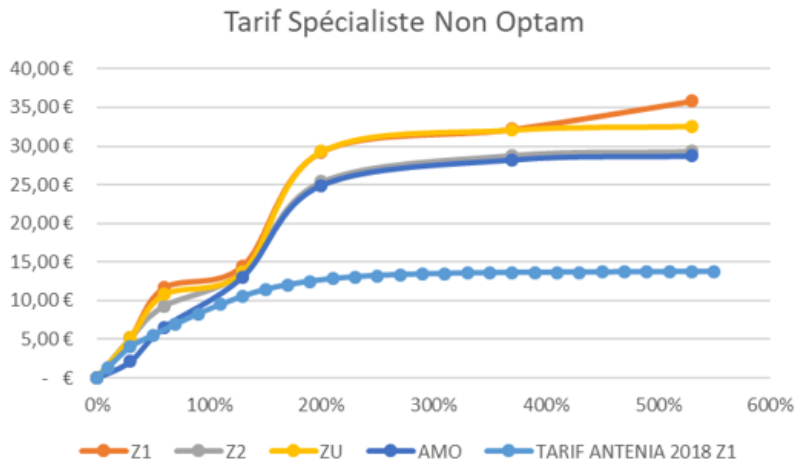
For the machine learning model, each type of model will be optimized according to its own parameters (example: maximum depth for the Decision Tree), then the best model will be chosen according to the criteria of the Mean Absolute Error (or MAE - average of the absolute errors between the estimated values and the actual values), Mean Squared Error (or MSE - average of the error squares) as well as the Mean Error (or ME- average errors). The best model will be the one that will give the indicators closest to zero.

The MAE and MSE are simple indicators for estimating the predictive quality of the model. The ME allows having precision on a collective model, which seeks an average precision and not an individual one.

Finally, **the reliability analysis** will be done using these last 3 indicators.

Results obtained with the internal model:

Once smoothed, we project the rate for the typical insured who defines our pure premium and on a range of insureds with different characteristics to take into account the coefficients of explanatory variables:



100-Pricing obtained for Non-OPTAM specialist doctor before and after the back-testing on the distinct geographical areas

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	Niveau Garantie (BR)												
ANTENIA	0	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €

New Antenia prices calculated on several policy holders profiles

Compared to the current price, the premium is growing rapidly, and the curve shows an inflection at the 130% reimbursement rate point. This is precisely the limit of reimbursements authorized under a responsible contract for health supplements when the specialist is not a member of OPTAM. We can then confirm there is a change in behaviour.

Results obtained with the GLM:

Despite the performance criteria chosen, the pricing operational reality forced us to choose to model the frequency with a Poisson Law and the average costs with a lognormal law.

The results obtained on a range of policyholders of different characteristics are:

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
MODELE	**Niveau Garantie (BR)**												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
POISSON x LOGNORMALE	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	17,94 €	41,63 €	21,65 €	0,86 €	38,67 €	96,42 €	30,29 €	15,75 €	11,95 €	25,24 €	14,42 €	85,47 €
	0,6	18,08 €	41,95 €	21,82 €	0,87 €	38,97 €	97,17 €	30,52 €	15,87 €	12,04 €	25,43 €	14,53 €	86,13 €
	1,3	18,41 €	42,71 €	22,21 €	0,89 €	39,67 €	98,92 €	31,07 €	16,16 €	12,26 €	25,89 €	14,79 €	87,69 €
	2	18,74 €	43,48 €	22,61 €	0,90 €	40,39 €	100,70 €	31,63 €	16,45 €	12,48 €	26,36 €	15,06 €	89,27 €
	3,7	19,57 €	45,41 €	23,61 €	0,94 €	42,18 €	105,17 €	33,04 €	17,18 €	13,03 €	27,53 €	15,73 €	93,23 €
	5,3	20,39 €	47,31 €	24,60 €	0,98 €	43,94 €	109,56 €	34,41 €	17,90 €	13,58 €	28,68 €	16,38 €	97,12 €

101- Antenia and GLM annuel cost comparison on several policy holders profiles

The differences between the two pricings are notable: on the one hand, the premium obtained by GLM depend hardly on the level of guarantee. That is a problem, especially at the levels borders: pricing too high for low levels, and probably too low for high levels. Average costs were not offset by frequencies for this problem.

A noticeable difference can also be seen in the high levels, which are either significantly higher or lower than internal pricing.

The annual rate is also extremely low for partner policyholders' rank by the GLM.

On the other hand, all the explanatory variables change the annual rate in the same direction.

Results from supervised learning:

As with the GLM, the search for the best supervised machine learning model has led us to choose the Tree Decision for frequency and average cost, for its simplicity and for the small performance gap it presented against the real best theoretical model.

The analysis of the annual cost obtained on the range of policyholders is as follows:

	RANG	A	A	**A**	C	A	A	A	A	A	A	A	A
	sex	M	F	**M**	M	M	F	F	M	M	F	M	M
	BASE_OPTION	B	B	**B**	B	O1	O2	B	B	B	B	B	O3
	zone_AMO	1	1	**1**	1	1	1	2	2	AMO	AMO	AMO	1
	year	20	40	**40**	40	40	40	40	40	40	50	60	40
	Niveau Garantie (BR)												
ANTENIA	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	3,50 €	11,34 €	**5,64 €**	5,24 €	5,92 €	11,91 €	11,34 €	5,64 €	2,35 €	4,82 €	3,20 €	6,20 €
	0,6	7,93 €	25,68 €	**12,77 €**	11,88 €	13,41 €	26,96 €	20,50 €	10,20 €	7,16 €	14,70 €	9,76 €	14,05 €
	1,3	9,85 €	31,90 €	**15,86 €**	14,75 €	16,66 €	33,50 €	30,11 €	14,97 €	14,30 €	29,35 €	19,49 €	17,45 €
	2	19,87 €	64,36 €	**32,00 €**	29,76 €	33,60 €	67,58 €	55,83 €	27,76 €	27,24 €	55,91 €	37,12 €	35,20 €
	3,7	21,86 €	70,78 €	**35,20 €**	32,73 €	36,96 €	74,32 €	63,39 €	31,52 €	30,85 €	63,34 €	42,05 €	38,72 €
	5,3	24,34 €	78,81 €	**39,19 €**	36,44 €	41,15 €	82,75 €	64,48 €	32,06 €	31,38 €	64,42 €	42,77 €	43,11 €
Machine Learning	0	- €	- €	**- €**	- €	- €	- €	- €	- €	- €	- €	- €	- €
Cout annuel	0,3	9,34 €	9,60 €	**12,08 €**	12,30 €	15,18 €	9,60 €	14,49 €	25,29 €	12,08 €	10,38 €	12,08 €	15,18 €
	0,6	17,70 €	59,60 €	**30,64 €**	9,82 €	15,71 €	38,26 €	59,60 €	30,64 €	17,92 €	36,31 €	18,99 €	15,71 €
	1,3	19,76 €	53,32 €	**41,94 €**	8,73 €	40,23 €	64,04 €	51,18 €	21,80 €	16,07 €	35,22 €	14,18 €	40,23 €
	2	39,26 €	104,11 €	**52,52 €**	18,98 €	42,50 €	125,05 €	62,73 €	40,48 €	40,48 €	77,58 €	10,83 €	69,82 €
	3,7	23,64 €	171,78 €	**88,31 €**	28,32 €	117,41 €	206,34 €	100,63 €	20,65 €	20,65 €	94,53 €	37,22 €	117,41 €
	5,3	35,55 €	76,53 €	**39,34 €**	12,90 €	52,31 €	91,93 €	57,47 €	20,60 €	20,60 €	29,57 €	16,58 €	103,71 €

102- Annual cost comparison between internal and Decision Tree modelling

The annual premiums calculated via the average cost of the optimized Decision Tree model are interesting. They are generally in the same rough size as our pricing tool.

A few points of attention are observed, especially on the point corresponding to our pure premium (main insured of 40 years male): the prices are not increasing with the guarantee level for some points, they are

also higher (except for the 5.3xRB rate point), the low and high levels of guarantees. Finally, the price is generally lacking in "sprawl", i.e. there is little difference between the minimum guarantee price and the high warranty rate, and the low levels are very high.

In general, the results are consistent, but would require additional smoothing to have a reliable and usable curve in terms of price.

Study of the reliability of the internal Swiss Life standard

We have achieved the best possible modelling through GLM and supervised learning, and we know the benefits and limitations. We will now objectively analyze the pricing quality of our internal method by comparing it with these other two methods.

To do this, we rate the same sample of policyholders (excluding children), and analyze the differences between the different predictions.

	Réels	Antenia	Machine Learning	GLM
Moyenne coûts annuels	32,78	34,13	44,73	32,52
ME		1,36	11,96	- 0,26
MAE		25,64	29,93	29,15
MSE		1 223,10	1 496,46	1 587,52

103-Performance indicators results for each model

If we look at the average of errors, it is finally the GLM that is the closest model to reality with an average error close to zero (-0.26€), followed closely by Antenia (1.36€). The Machine Learning model is much further from the actual annual cost. It overprices the premium as we had analyzed it. But its MAE is in the same rough size as other models. We can therefore consider the accuracy of the models to be good, and we can all describe them as reliable in analyzing the quality of our internal pricing model.

If we analyze the results in relation to the MAE, Antenia has the best results.

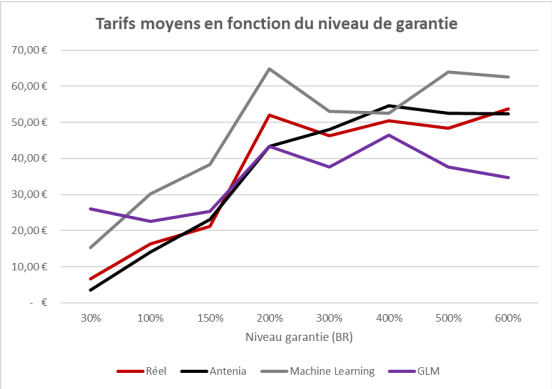
From these initial results, we can estimate that over the entire portfolio studied, our internal modeling is reliable and its accuracy is satisfactory compared to models obtained with modern methods. This is already good news, but it needs to be confirmed with further analysis of the results.

Results by explanatory variable segment

We can look at the results of the MAE by explanatory variable segment: so we freeze a specific modality, for example the guarantee level, and we look at which model will be the fairest on these groups of insureds.

NIVEAU GARANTIE (BR)		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
30%	1%	6,68 €	3,49 €	15,34 €	26,00 €	- 3,19 €	8,66 €	19,32 €	-48%	130%	289%	6,75 €	11,78 €	24,25 €
100%	7%	16,33 €	14,01 €	30,12 €	22,56 €	- 2,32 €	13,79 €	6,24 €	-14%	84%	38%	13,27 €	21,65 €	20,89 €
150%	50%	21,25 €	23,07 €	38,30 €	25,29 €	1,82 €	17,05 €	4,04 €	9%	80%	19%	19,67 €	26,90 €	22,55 €
200%	5%	52,09 €	43,34 €	64,92 €	43,41 €	- 8,74 €	12,83 €	- 8,68 €	-17%	25%	-17%	35,95 €	39,19 €	41,75 €
300%	6%	46,35 €	48,11 €	53,09 €	37,57 €	1,77 €	6,74 €	- 8,78 €	4%	15%	-19%	33,09 €	33,11 €	35,99 €
400%	24%	50,46 €	54,54 €	52,52 €	46,45 €	4,08 €	2,06 €	- 4,02 €	8%	4%	-8%	35,47 €	34,64 €	38,14 €
500%	3%	48,32 €	52,63 €	64,03 €	37,68 €	4,31 €	15,71 €	- 10,64 €	9%	33%	-22%	35,01 €	38,31 €	37,83 €
600%	4%	53,83 €	52,32 €	62,58 €	34,62 €	- 1,51 €	8,75 €	- 19,21 €	-3%	16%	-36%	39,36 €	37,90 €	42,31 €

104-Errors study by guarantees levels groups



105- Average price in function of the guarantee level obtained with each model

For most explanatory variable analyses, it is clear that Antenia provides the closest curve to reality, followed by the GLM. The Machine Learning model overprices all warranty levels. Antenia and GLM are antagonistic on their modeling: GLM on low levels prices, and sub-price high levels (from 200%BR), while Antenia tends to underprice low levels and to overprice high levels. However, these discrepancies are not significant.

The average absolute errors are better for Antenia regardless of the guarantee level, and above all this model stays closer to reality when looking at average errors.

The findings are broadly the same for the analysis of beneficiary, area, option level and age variables. On the other hand, Antenia seems to overprice the women category and underprice the men, but these errors compensate each other globally.

SEXE		Moyenne Tarif				Moyenne Ecarts			Moyenne %ecart			Moyenne erreurs absolues		
	Poids de la modalité	Réel	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM	Antenia	Machine Learning	GLM
F	60%	35,93 €	43,01 €	55,03 €	39,11 €	7,08 €	19,11 €	3,19 €	20%	53%	9%	27,43 €	33,64 €	32,18 €
M	40%	28,01 €	20,68 €	29,12 €	22,52 €	- 7,33 €	1,11 €	- 5,50 €	-26%	4%	-20%	22,95 €	24,31 €	24,55 €

106- Errors study by sex category

Extreme values

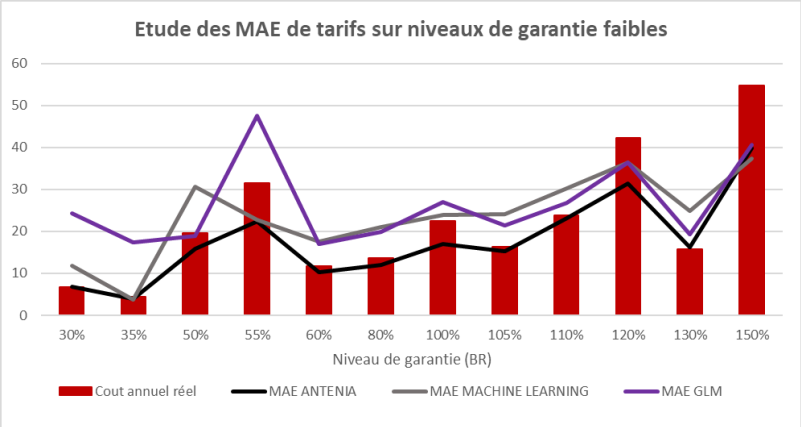
We can complete the analysis of the reliability of the model by studying the biggest errors. It is normal that some policyholders have sometimes been modelled with significant errors, since all models provide a capped rate, whereas in reality policyholders may have to consume many routine cares, depending on their health status.

The number of lines for which the models reported errors greater than 2 times the actual observed rate (in absolute value) may represent the heterogeneity of the modality observed in this sample, and/or the "big errors" that the models were able to make. This is not a precise indicator of stability, as large deviations may be due to exceptionally high consumption by some policyholders, which our models can only represent in the average of all modalities. However, the analysis remains interesting because it allows seeing where there have been peaks of errors.

The analysis shows errors greater than twice the actual rate, by level of modality studied, in number and percentage, for the 41417 insured studied

The conclusion of this analysis is that Antenia does not contain any significant extreme values (maximum 13% of each modality with extreme values). On the modern model side, the only variables that catch our attention for GLM and Machine Learning are the low warranty levels (up to 150%RB rate), as well as the "Base" and "Zone 1" (area 1) modalities. Zooming in on part of the modalities of each specific explanatory variable allows you to analyze more precisely any discrepancies noticed.

In particular, the Antenia-rated smoothing on the guarantee levels helps to mitigate the peaks in benefits by averaging them.



107- MAE prices study zoomed on low guarantees levels

The other extremes noted on Area 1 and the modality BASE_OPTION are certainly explicable by the strong representation of these data in the samples, and the bias that had to be made to have this information (not questioned area classification and lack of information on the levels of options).

Conclusion

After completing the back-testing of a health warranty, which had to be limited to a reduced choice of explanatory variables for operational (format of the final setting tool) and techniques (availability and quality of data, past studies showing the independence of the pricing compared to certain explanatory variables in our portfolio) reasons, we studied its reliability by comparing the new premium with that we would have obtained with two modern pricing methods : GLM and supervised machine learning.

The reliability analysis first requires ensuring that the so-called "reference" methodes are reliable themselves. We saw in our analysis that a GLM and supervised learning had both their advantages and disadvantages in interpreting the results.

The study shows that our internal standard is rather reliable, and highlights some potential updates of our coefficients (sex and option level in particular).

Key words: health insurance, pure premium cost, generalized linear model, machine learning, frequency x cost