GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES

INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

ADDING Groupe SIACI SAINT HONORE Actuariat et Conseil RH

Mémoire d'Actuariat

Réalisation d'un outil de tarification a priori en Santé grâce à des modèles de *machine learning*

Auteur :
Constance VILOTITCH

Tuteur professionnel :
Amara NDAO

Tuteur Pédagogique :
Pierrick PIETTE

Année 2022 - 2023

REMERCIEMENTS

Je tiens tout d'abord à remercier sincèrement Alexandre PLANQUART pour son accompagnement, sa disponibilité et sa relecture. Ses critiques et ses remarques m'ont permis d'avancer et m'ont menée vers de meilleures réflexions.

Je souhaite aussi adresser mes remerciements à Amara NDAO, pour les échanges enrichissants que j'ai pu avoir avec lui ainsi que pour son aide précieuse et pour le temps qu'il m'a accordé.

Je remercie également mes collègues d'Arkol et d'Adding, qui ont su partager leurs connaissances avec moi et m'ont conseillée durant mon parcours professionnel ; et je remercie plus particulièrement Papa SAR, qui m'a beaucoup guidée et encouragée pendant ma formation, ainsi que Berthille PIERRON pour sa relecture soignée.

Mes remerciements vont également à mon tuteur pédagogique, Pierrick PIETTE, pour m'avoir fait bénéficier de son expérience et de ses conseils tout au long de l'encadrement de ce mémoire, ainsi que pour sa relecture.

Enfin, je voudrais exprimer ma gratitude envers mon entourage personnel pour le soutien qu'ils m'ont apporté durant ma formation à l'ISFA.

RESUME

Mots clés : Tarification Santé, *Machine learning*, CART, Forêt aléatoire, GBM, Réseau de neurones, H2O, « Auto-ML », Interprétabilité des modèles, Graphique de dépendance partielle, Valeurs de Shapley, SHAP

Le rôle d'un conseiller en actuariat est d'accompagner au mieux ses clients dans le pilotage et la mise en place de leurs dispositifs de protection sociale, et notamment en Santé. Les équipes techniques peuvent alors être amenées à réaliser des tarifications. L'objectif est, à partir d'une population donnée et de garanties cibles, de déterminer un tarif pour les salariés de l'entreprise cliente.

Adding souhaite proposer des tarifications les plus pertinentes et adaptées possibles aux besoins de ses clients, et pour cela nécessite un outil de tarification performant et adaptable. Puissant et efficace, le *machine learning* s'est révélé être une approche particulièrement intéressante pour la construction d'un outil de tarification. Par ailleurs, la méthode « fréquence / coût moyen », sous poste par sous poste (un sous poste correspondant à une ligne d'une grille de garantie), a été retenue, puisqu'elle permet une grande adaptabilité.

Dans un premier temps, une base de données diversifiée, contenant les différentes dépenses Santé des bénéficiaires de plusieurs entreprises clientes, a été créée. Plusieurs traitements ont permis d'aboutir à la construction d'une base homogène et exploitable. L'analyse de ces données ainsi que de la corrélation entre les différentes variables a conduit à la détermination des différentes variables explicatives.

Par la suite, sur R, deux méthodes ont été utilisées afin de créer les différents modèles de fréquence et de coût moyen de chaque sous poste. Des indicateurs de performance ont aidé à comparer ces deux méthodes.

D'une part, la création des modèles a été testée en essayant et en paramétrant manuellement quatre algorithmes de *machine learning* sur deux sous postes choisis. Cette méthode a permis de proposer des modèles optimaux, mais est très chronophage, d'autant plus si elle doit être appliquée à l'ensemble des sous postes.

D'autre part, un algorithme de détermination « automatique » des meilleurs modèles, proposé par la plateforme H2O, a été utilisé sur l'ensemble des sous postes. L'utilisation de cet algorithme « auto-ML » a été couplée avec des méthodes d'interprétabilité, permettant de mieux comprendre les modèles construits. Cette méthode s'est révélée être plus efficace que la précédente, car elle est plus rapide et propose des modèles performants pour une partie des sous postes. Toutefois, certains modèles proposés ne sont pas appropriés et nécessitent d'être de nouveau modélisés, en modifiant notamment certains paramètres de l'algorithme « auto-ML » ou bien en améliorant la base de données.

Ainsi, la puissance et l'efficacité de l'algorithme proposé par H2O a permis d'établir rapidement une série de modèle de *machine learning* performants. La prime commerciale peut ensuite être calculées, en fonction des données d'entrée (population et garanties cibles) et de la structure de cotisation souhaitée.

ABSTRACT

Keywords : Healthcare insurance pricing, Machine learning, CART, Random forest, GBM, Neural networks, H2O, « Auto-ML », Models explainability, Partial dependence plot, Shapley values, SHAP

The actuarial consulting company's role is to best support its client in managing and implementing their social welfare, especially in health-insurance. The technical teams may then be required to carry out pricing. The goal is to determine a cost for the employees of the client company, from a given population and guarantees.

Adding wishes to offer pricing that is as relevant and adapted as possible to the needs of its customers, to this end, it requires a powerful and adaptable pricing tool. Machine learning seems to be a particularly interesting approach to build this pricing tool as it is powerful and efficient. Moreover, the "frequency / average cost" method, item by item (an item corresponding to a line of a guarantee grid), was chosen since it allows great adaptability.

First of all, a diversified database, containing the various health expenses of several client companies' beneficiaries, was created. Several treatments have led to the construction of a homogeneous and exploitable database. The analysis of these data and the correlation between the different variables led to the determination of the different predictor variables.

Then, on R, two methods were used to create the different frequency and average cost models for each item. Performance indicators helped to compare these two methods.

On the one hand, the creation of the models was tested by testing different parameters manually on four machine learning algorithms and two chosen items. This method has given high-performance models but is very time-consuming if it must be applied to all the items.

On the other hand, an algorithm for "automatic" determination of the best models, proposed by the H2O platform, was used on all the items. The use of this "auto-ML" algorithm has been coupled with explainability methods, allowing a better understanding of the models built. This method has proven to be more efficient than the previous one, as it is faster and offers good performing models for some item. However, some proposed models are not suited and need to be rebuilt, by modifying certain parameters of the "auto-ML" algorithm or by improving the database.

Thus, the power and efficiency of the algorithm proposed by H2O made it possible to quickly establish a series of high-performance machine learning models. The commercial premium can then be calculated, depending on the input data (given population and guarantees) and the desired premium structure.

SOMMAIRE

INTRODUCTION

Dans un contexte économique où l'inflation n'épargne aucun poste de dépense, piloter au plus juste les dispositifs d'avantages sociaux peut se révéler être un enjeu majeur pour les entreprises. En tant que société de conseil, le rôle d'Adding est d'aider les entreprises à limiter leurs coûts sur les différents dispositifs mis en place pour les salariés, et notamment sur leur régime de Frais de Santé.

Pour mener à bien cette mission, Adding utilise actuellement un outil de tarification interne, utilisant une méthode avec des barèmes et qui a tendance à surestimer le montant de la prime pure. Or, dans l'objectif de répondre au mieux aux besoins de ses client, Adding nécessite un outil de tarification plus adapté, et permettant de mieux coller avec la réalité.

D'autre part, avec l'essor du *Big Data,* le secteur de l'assurance voit ses techniques et méthodes évoluer. En effet, le *Big Data* et ses applications, telles que l'intelligence artificielle ou le *machine learning*, permettent de mieux piloter les risques grâce à leur précision, à leur vitesse et à leur adaptabilité. Grâce à cela, les produits disponibles sur le marché évoluent et permettent de prendre en compte les analyses comportementales.

C'est dans ce contexte que s'inscrit l'objectif de ce mémoire : la conception d'un outil de tarification a priori en Santé grâce à des modèles de *machine learning*.

Le *machine learning* va permettre de modéliser au plus juste le comportement des assurés et de leur famille, grâce à la prise en compte de plusieurs critères. Il va ainsi proposer une tarification plus adaptée et plus personnalisée, qui tentera de se rapprocher au maximum de la réalité.

Pour répondre à la problématique, nous allons tout d'abord donner des éléments de contexte fondamentaux à propos de l'assurance santé et du *machine learning*. Puis, nous parlerons de la constitution de la base de données ayant servi à l'apprentissage, ainsi que de ses caractéristiques intéressantes pour la suite de notre étude. Par la suite, nous étudierons la modélisation des différents postes de santé grâce au *machine learning*. Enfin, nous verrons comment tout ce travail peut être intégré dans un outil capable de déterminer des montants de cotisations pour une population et une grille de garanties données.

1 CONTEXTE GENERAL

L'objectif de ce chapitre est de présenter le contexte général du projet, ainsi que le cadre dans lequel il a été mené et les différents outils et modèles mathématiques qui ont permis de le réaliser.

1.1 PRESENTATION DE L'ENTREPRISE

1.1.1 Le Groupe Siaci Saint Honoré

Le Groupe Siaci Saint Honoré, ou S2H, est un groupe français né en 2007 de la fusion entre ACSH (Actuariat et Conseil Saint Honoré) et SIACI, et dirigé par Pierre Donnersberg. S2H une entreprise de conseil et courtage en assurance de biens et de personnes. Actuellement leader européen dans son domaine, le groupe a un champ d'action mondial puisqu'il n'emploie pas moins de 3 000 collaborateurs répartis dans quatre régions du globe : l'Europe, l'Asie, le Moyen-Orient et l'Amérique du Nord. En tant que courtier en assurance, le Groupe Siaci Saint Honoré se place comme un intermédiaire de confiance entre l'assuré et l'assureur, son but étant de conseiller ses clients et de les orienter vers les produits les mieux adaptés à leurs besoins. Il conçoit notamment des solutions pour les risques suivants :

- IARD,
- Transport,
- Protection sociale,
- Epargne salariale,
- Retraite,
- Rémunération et mobilité internationale.

S2H travaille pour des grandes entreprises, des ETI et des PME, et compte actuellement plus de 5 000 entreprises clientes à travers le monde. En 2019, S2H a réalisé un chiffre d'affaires d'environ 500 millions d'euros.

En 2021, le Groupe Siaci Saint Honoré et le Groupe Burrus Courtage, expert en assurance, courtage, technologie, gestion financière et conseil, signent un accord en vue de devenir le leader européen du courtage en assurance, d'envergure internationale. Avec cette fusion, le nouveau Groupe se positionnera parmi les dix plus gros acteurs mondiaux du secteur et renforcera sa position en France et en Europe, avec un chiffre d'affaires d'environ 700 millions d'euros. Il interviendra désormais auprès de ses clients grâce à plus de 5 000 collaborateurs répartis dans une quarantaine de pays. Il sera présidé par Pierre Donnersberg et le directeur général sera Christian Burrus. [1]

1.1.2 Adding

Adding a été créé en 1993 et appartenait avant 2019 au Groupe Addactis, expert proposant son savoir-faire actuariel et sa connaissance des risques et des marchés pour accompagner les organismes assureurs, les directions des ressources humaines et plus largement les entreprises sur leurs principaux enjeux actuels.

Dorénavant, Adding fait partie du Groupe Siaci Saint Honoré. Mais depuis toujours, Adding accompagne les entreprises et les branches professionnelles au travers de solutions Ressources Humaines personnalisées. L'intervention d'Adding nécessite un savoir-faire économique, financier, juridique et social. Elle met en place des dispositifs d'aide à la gestion optimisée dans ses domaines de prédilection : conseil en

rémunération, capital humain et avantages sociaux. Les actuaires répondent à tous types de besoins : audit, évaluations actuarielles, conseil…

Adding dispose de deux bureaux basés dans deux villes différentes : Paris et Lyon, et ayant des fonctions distinctes. Cependant, les bureaux communiquent régulièrement entre eux et travaillent ensemble.

Le bureau parisien s'occupe majoritairement de la partie communication mais peut répondre à des questions techniques. Il regroupe des chefs de projet et des consultants assurant le suivi avec les clients et l'interface avec les assureurs. Les métiers en contact direct avec le client ont été positionnés à Paris pour des contraintes pratiques car Adding traite majoritairement avec les Directeurs des Ressources Humaines ou Responsables des Ressources Humaines de grands groupes dont les locaux sont souvent situés à Paris. Bien que des responsables de chaque pôle supervisent les projets dans les bureaux lyonnais, les directeurs des différents pôles lyonnais se trouvent également à Paris.

Lyon correspond à la base technique de la société. On y retrouve tous les métiers nécessaires à ce travail. Le pôle innovation et projet assure le suivi et la création des projets. Il est souvent en lien avec le service informatique car de nombreuses nouveautés sont lancées sur le web (on peut notamment citer les sites *e-benefits,* qui sont paramétrables pour chaque entreprise et permettent aux salariés de connaitre l'ensemble de leur droits). Le pôle actuariat, composé d'actuaires et de chargés d'études répond à toutes les demandes liées à l'actuariat. Grâce à sa structure à taille humaine (environ 25 employés basés à Lyon), les services communiquent facilement entre eux.

Adding propose trois types de solution [2] :

Benefits

Le service *Benefits* répond aux questions sur l'ensemble des avantages que l'entreprise peut proposer à ses salariés. Ce service regroupe la santé, la prévoyance, la retraite, les engagements sociaux ainsi que l'épargne salariale. Adding aide les entreprises à mettre en place, évaluer et optimiser l'ensemble des avantages pour les salariés. Le pôle actuariat s'occupe majoritairement de la partie *Benefits*.

Conseil en rémunération

Adding offre aussi ses conseils en rémunération qui comprend l'intéressement, la rémunération variable de la performance, la modélisation et l'externalisation de la revue annuelle des salaires. Adding se trouve alors au cœur du pilotage et de la stratégie de rémunération de l'entreprise.

Capital humain

Adding est également présente pour accompagner les entreprises dans leur communication avec les salariés. Pour cela, elle fournit des Bilans Sociaux Individualisés (BSI) ajustés selon les envies de l'entreprise. Adding offre des solutions digitales à travers des applications et des sites internet dédiés, comme les sites *e-benefits*.

1.1.3 Les missions de l'équipe actuariat

L'équipe actuariat traite donc majoritairement de la partie *Benefits*. Ses différentes missions portent sur [2] :

La santé et la prévoyance

Adding aide les entreprises à mettre en place, évaluer et optimiser leurs régimes de protection sociale. Les entreprises sont accompagnées aussi bien sur le plan technique que juridique, grâce à des modélisations, des audits et des analyses, ce qui leur permet d'établir les stratégies les plus adaptées vis-à-vis de l'assureur. Cela permet aux entreprises d'avoir un régime qui convient à leurs salariés, en santé comme en prévoyance (arrêts de travail et décès). On parle ici de contrats de santé et prévoyance collectifs. C'est sur cette partie que portera le mémoire.

La retraite

Adding conseille les entreprises sur la mise en place ou la transformation de leur dispositif de régime de retraite. Grâce à des conseils simples et pédagogiques (notamment la mise en place d'outils web adaptés à chaque entreprise), Adding permet aux salariés de mieux comprendre la retraite, de mieux préparer leur départ, d'améliorer leur rente…

Les engagements sociaux

Adding conseille les entreprises sur tous les engagements que l'entreprise a envers ses salariés du fait de prestations différées qu'elle s'est engagée à leur fournir (indemnités de départ à la retraite, médaille du travail, compte épargne temps…). Adding leur permet de valoriser, modifier ou d'optimiser leur dispositif.

L'épargne salariale

Adding accompagne les entreprises en testant la compétitivité de leur dispositif d'épargne salariale. Grâce à sa communication auprès des salariés, Adding leur propose une approche pédagogique. Adding conseille également les entreprises sur les choix de fonds de placement, leur fonctionnement, leur performance, leur gestion… et communique également avec les partenaires sociaux.

1.2 LA PREVOYANCE COLLECTIVE

Etant donné que l'objet de ce mémoire porte sur les contrats de santé et prévoyance collectifs, il est important de bien définir les termes et le contexte. Dorénavant, lorsque l'on parlera de prévoyance cela impliquera la santé, l'arrêt de travail et le décès.

1.2.1 Principe de fonctionnement de la prévoyance collective

La « Prévoyance Collective » [3] permet de couvrir un groupe de personnes qui ont un lien objectif entre elles, ce groupe étant représenté par une personne morale qui va signer le contrat. Les différents acteurs d'un contrat de prévoyance collective sont :

- Le **souscripteur** : personne morale qui signe le contrat et paie les cotisations (par exemple l'entreprise).
- Les **affiliés** : ensemble des personnes appartenant au groupe assurable (par exemple les salariés).
- Les **assurés** : personnes soumises au risque. Il s'agit souvent des affiliés, mais attention dans le cadre d'un contrat de santé il peut s'agir de l'affilié et sa famille.
- Les **bénéficiaires** : personnes susceptibles de recevoir des prestations de la part de l'assureur.

Le contrat collectif peut être mis en place suivant quatre modes : par Convention Collective Nationale, par accord collectif d'entreprise, par référendum ou par décision unilatérale de l'employeur.

Les contrats collectifs présentent plusieurs avantages, que ce soit pour l'entreprise ou les salariés. Tout d'abord, il permet une meilleure couverture ainsi que la réduction des coût grâce à la mutualisation des risques. Il permet également de s'affranchir de certaines formalités médicales. Grâce aux contrats collectifs, l'entreprise peut fidéliser ses salariés. Enfin, il propose des règles sociales et fiscales avantageuses.

Les entreprises qui souscrivent ce type de contrat ont la possibilité de choisir parmi les modes d'adhésion suivants :

- Adhésion **facultative** : le salarié a le choix d'adhérer ou non au régime. Il y a alors un risque d'antisélection.
- Adhésion **obligatoire** : tous les salariés de la CSP concernée par la mise en place du contrat doivent y souscrire. L'assureur doit accepter tous les adhérents et les cotisations et prestations fournies seront les mêmes pour tous.
- Adhésion **facultative et obligatoire** : il s'agit des contrats dont l'adhésion est obligatoire mais dont certaines extensions peuvent être facultatives.

Plusieurs structures de cotisations sont possibles pour les entreprises ayants souscrit un contrat collectif. Elles seront détaillées par la suite.

1.2.2 Le rôle de la complémentaire santé

Dans les entreprises, la complémentaire santé peut être souscrite au moyen d'un contrat collectif. Son rôle est de venir compléter la Sécurité Sociale sur certains frais médicaux engagés par les patients. En France, depuis le 1er janvier 2016 [4], tous les salariés d'une entreprise doivent impérativement être couverts par une complémentaire santé. Trois types d'organismes peuvent proposer des complémentaires santé : les mutuelles, les compagnies d'assurance et les institutions de prévoyance.

Les remboursements des actes de soins médicaux fonctionnent de la manière suivante :

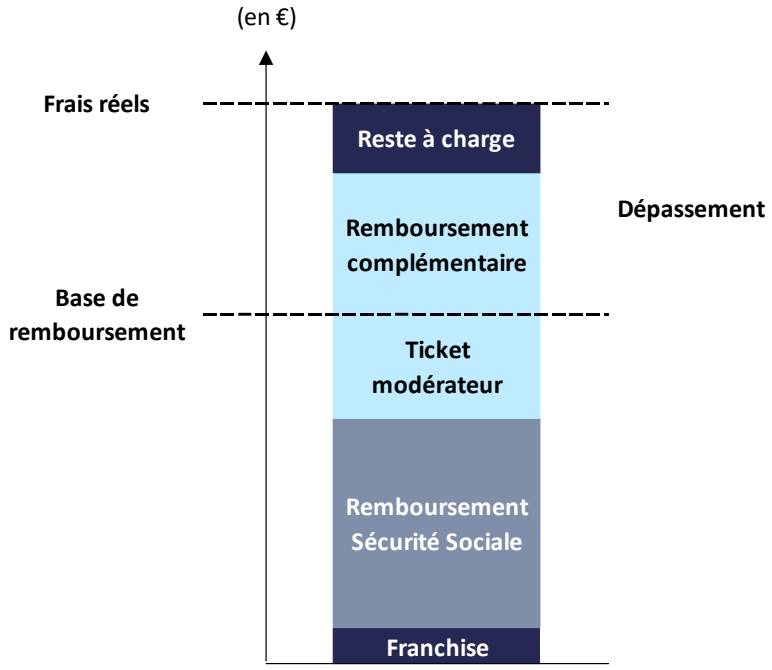


Figure 1.1 – Fonctionnement des remboursements en Santé

La Sécurité Sociale fixe pour chaque acte de santé :

- Une **base de remboursement**, qui est un tarif de référence pour les remboursements.
- Un **taux de remboursement**, compris entre 0% et 100%.

Le remboursement de la Sécurité Sociale s'exprime donc en multipliant la base de remboursement par le taux de remboursement.

Certains actes sont soumis à une franchise ou participation forfaitaire. Il s'agit d'une participation symbolique de la part des bénéficiaires sur des actes tels que des consultations médicales (1 €), des médicaments (0,50 €) …

La différence entre la base de remboursement et le remboursement de la Sécurité Sociale (+ l'éventuelle franchise) est appelé « ticket modérateur ». Cette partie est prise en charge par les régimes complémentaires responsables. Les régimes complémentaires responsables sont des régimes complémentaires qui favorisent le respect du parcours de soin par les bénéficiaires, qui respectent un cahier des charge (garanties minimales, encadrement de certains remboursements…) et qui sont solidaires (c'est-à-dire que les montants des cotisations ne sont pas fixés en fonction de l'état de santé des bénéficiaires). Les régimes complémentaires responsables bénéficient d'un cadre social et fiscal avantageux. Aujourd'hui, dans les entreprises, la plupart des régimes complémentaires collectifs mis en place sont des régimes responsables.

Parfois, les frais d'un acte médical peuvent aller au-delà de la base de remboursement. La partie supérieure à la base de remboursement s'appelle le « dépassement ». Il peut être pris en charge, intégralement ou non, par la complémentaire santé. La partie qui n'est pas prise en charge s'appelle le « reste à charge », et il est aux frais du bénéficiaire. Pour limiter son reste à charge, le bénéficiaire peut par exemple :

- Préférer les praticiens adhérents à l'**OPTAM** (Option Pratique Tarifaire Maîtrisée) : l'OPTAM est un dispositif mis en place ayant pour but de limiter les dépassements d'honoraire des praticiens qui y ont souscrit.
- Utiliser des **réseaux de soins** proposés par l'organisme assureur : un réseau de praticiens (opticiens, ostéopathes, dentistes…) est un ensemble de praticiens ayant conclu un partenariat avec l'organisme assureur et proposant des tarifs moins élevés aux bénéficiaires qui vont chez eux.

Le remboursement du dépassement dépend du niveau des garanties proposées par la complémentaire santé. Ainsi, plus les garanties sont élevées, mieux le bénéficiaire sera remboursé.

Les garanties proposées par la complémentaire santé peuvent être exprimées de manières différentes :

- En montant forfaitaire en ***€***.
- En ***%PMSS***: le PMSS est le Plafond Mensuel de la Sécurité Sociale. Il évolue légèrement chaque année avec le coût de la vie. En 2022, il était de 3 428 €.
- En ***%BRSS***: la BRSS est la Base de Remboursement de la Sécurité Sociale. Comme évoqué ci-dessus, elle est différente pour chaque acte. A priori, elle n'évolue pas chaque année contrairement au PMSS, mais elle peut être amenée à changer de temps en temps.
- En ***%FR***: les FR désignent les Frais Réels, il s'agit du montant facturé par le praticien.

Ces garanties peuvent être exprimées hors remboursement de la Sécurité Social ou y compris remboursement de la Sécurité Social. Excepté pour les garanties en %FR, puisque le total des remboursements ne peut jamais dépasser les frais réels.

1.3 PRESENTATION DU PROJET DE CREATION D'UN OUTIL DE TARIFICATION SANTE A PRIORI GRACE A DES MODELES DE MACHINE LEARNING

Afin de subvenir au mieux aux besoins de ses clients en Santé, Adding réalise diverses missions tout au long de l'année. Certaines missions sont régulières : les comptes de résultats, les projections de comptes de résultats et les analyse de consommation des frais de santé. D'autres missions sont plutôt ponctuelles et peuvent être réalisées en fonction des attentes et besoins des clients. Il s'agit par exemple d'analyses spécifiques sur un poste de soin, de tarifications ou bien d'études d'impact de réformes (par exemple, la réforme 100% Santé). L'étude qui nous intéresse plus particulièrement dans le cadre de ce mémoire est la tarification.

1.3.1 Définition de la tarification

La tarification se définit comme le processus de détermination des primes à payer par l'assuré, de sorte que l'assureur ait suffisamment de fonds pour régler les sinistres couverts par le contrat d'assurance. En d'autres termes, il s'agit de l'élaboration d'un tarif approprié à chaque risque, tenant compte de la solidarité entres assurés et de la solvabilité de l'assureur. Il existe deux types de tarification :

Tarification *a posteriori*

La tarification *a posteriori* se base sur des données de portefeuille. Elle est donc utilisée pour les assurés dont l'historique des sinistres est connu. Elle tient compte de variables explicatives exogènes relatives à la réalisation du risque. Cela permet à la fois d'adapter le tarif initial de l'assuré à sa sinistralité individuelle durant toute la durée de vie de son contrat (on parle de système *Bonus – Malus*), mais aussi de tenir compte de l'hétérogénéité du portefeuille. Dans le cadre de tarification *a posteriori*, les techniques actuarielles suivantes peuvent par exemple être utilisées :

- Les systèmes *Bonus – Malus* markovien
- La théorie de la crédibilité

Tarification *a priori*

La tarification *a priori*, quant à elle, est utilisée pour des assurés dont l'historique des sinistres est inconnu. En d'autres termes, cela impose à l'assureur d'essayer de prévoir, en amont du contrat, la sinistralité future de l'assuré, grâce à l'utilisation de variables explicatives endogènes propres au risque et à la segmentation en groupe de risques les plus homogènes possibles. Dans le cadre de tarification *a priori*, les techniques actuarielles suivantes peuvent par exemple être utilisées :

- Les modèles linéaires généralisés
- Les modèles de *machine learning* (arbres de régression, réseaux de neurones…)

Le sujet de ce mémoire porte plus particulièrement sur la tarification *a priori*, et notamment grâce à des modèles de *machine learning*.

1.3.2 Introduction au machine learning

Apparu dans les années 1950, le *machine learning* peut être défini comme une branche de l'intelligence artificielle. L'intelligence artificielle a pour but de d'appliquer des techniques permettant aux ordinateurs de reproduire des comportements humains. Le *machine learning* permet, plus spécifiquement, d'effectuer des prédictions, grâce à des modèles, en se basant sur des données et des statistiques. Il se

révèle être très efficace, par sa précision et sa vitesse, dans les ensembles de données volumineux et complexes.

Les données en entrées sont appelées « variables explicatives », notées X, et les éventuelles données en sortie sont appelées « variables à expliquer », notées Y.

Généralement les données d'entrée sont divisées en trois échantillons :

- Les *données d'entraînement* ou *d'apprentissage* : elles permettent d'entraîner le modèle, afin qu'il puisse par la suite réaliser des prédictions sur de nouvelles données.
- Les *données de validation* : elles sont composées de données qui n'étaient pas présentes dans les données d'entraînement. Elles permettent d'ajuster les paramètres du modèle et d'aboutir à la validation du modèle qui a été entraîné.
- Les *données test* : il s'agit des données n'ayant servi ni à l'apprentissage ni à la validation du modèle. Elles permettent d'évaluer les performances du modèle établi.

Les échantillons sont obtenus aléatoirement et doivent être représentatifs des données. La taille des échantillons dépend de la taille des données en entrée. Cependant, il est d'usage de garder au moins 70% / 80% des données pour l'échantillon d'apprentissage.

Afin d'éliminer en grande partie le biais créé par l'échantillonnage (le partitionnement des données en échantillons d'apprentissage et de validation), biais d'autant plus marqué lorsque la taille des données d'entrée n'est pas assez importante, une autre technique peut être utilisée. Cette technique permet de ne pas diviser les données en deux échantillons (apprentissage et validation). Il s'agit de la validation croisée, définie en annexe 1, qui permet de valider le modèle entraîné. L'évaluation du modèle se fera tout de même sur des données test, différentes des données d'apprentissage.

Le *machine learning* englobe plusieurs méthodes capables de créer automatiquement des modèles à partir de données. On distingue différents types d'algorithmes de *machine learning* :

Les algorithmes d'apprentissage supervisé

Dans le cadre de l'apprentissage supervisé, les données d'entrainement sont déjà labellisées. Ces données déjà labelisées avec les bonnes réponses permettent ensuite de prédire les réponses sur des nouveaux jeux de données non labelisés.

On peut écrire cela de la manière suivante :

Soit Y la variable à expliquer décrite par n objets pour lesquels nous connaissons p variables explicatives X. On définit les données d'entraînement par $D_{entraînement} = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$. L'objectif est de trouver une fonction φ des p prédicteurs telle que $Y = \varphi(X) + \epsilon$, où ϵ est l'erreur.

Suivant la nature des données à prédire Y, il existe deux sortes d'algorithmes d'apprentissage supervisé :

- Les algorithmes de classification pour des variables à expliquer qualitative : $Y \in \{y_1, \dots, y_k\}$.
- Les algorithmes de régression pour des variables à expliquer quantitatives : $Y \in \mathbb{R}$.

Pami les algorithmes supervisés, on peut citer par exemple des arbres de régression et de classification, les forêts aléatoires, les GBM (*Gradient Boosting Machine),* les réseaux de neurones… Ce sont ces types d'algorithmes qui seront utilisés dans le cadre de ce mémoire. Par ailleurs, les variables à modéliser étant la fréquence et le coût moyen, ce sont des algorithmes de régression qui seront utilisés.

Les algorithmes d'apprentissage non supervisé

Dans le cadre de l'apprentissage non supervisé, les données d'entrainement ne sont pas labellisées. Il n'existe donc pas de notion de variables à expliquer. L'algorithme doit lui-même déterminer la structure sous-jacente des données.

On définit cette fois la variable explicative X, décrite par n objets pour lesquels nous connaissons p variables explicatives. On définit également les données d'entraînement par $D_{entraînement} = \{X_1, X_2, \dots, X_n\}$, appartenant à l'ensemble χ. L'objectif est de trouver le partitionnement de χ en fonction de la ressemblance des caractéristiques entre les observations.

Pami les algorithmes non supervisés, on peut citer par exemple le *clustering* hiérarchique, les K-*means*…

1.3.3 Présentation du projet

Dans le cadre de certaines missions, il peut arriver que des demandes de tarification en santé soient faites sans que les données de consommation des frais de santé soient à la disposition d'Adding. C'est le cas par exemple lors de l'intégration d'un nouveau poste dans les garanties du régime en place, ou bien lorsqu'il s'agit d'un nouveau client pour lequel il faut réaliser une tarification *ex nihilo*. Dans ce cas, il est nécessaire d'utiliser une tarification *a priori*.

Les tarifications *a priori* en santé consistent donc en la détermination du coût de chaque bénéficiaire du régime, en fonction des garanties du régime et de données sur les bénéficiaires, dans le but de déterminer un montant de cotisation. Des logiciels existent déjà et permettent de réaliser ce genre de tarification. Ils s'appuient sur des approches dites « classiques » de tarification en utilisant des méthodes comme celles des barèmes par exemple. Ces logiciels sont en général coûteux, ne sont pas toujours faciles à prendre en main et ne sont pas toujours adaptables selon les besoins du client.

De plus, par la diversité de ses clients, Adding a besoin d'un outil de tarification adapté à ses données et capable de résoudre des problèmes complexes. L'utilisation d'algorithmes de *machine learning*, grâce à la diversité des algorithmes existants, peut alors apparaître comme une bonne solution, proposant une méthode plus moderne que celles déjà existantes dans les outils de tarification actuels.

Enfin, par expérience, on sait que la manière de consommer des bénéficiaires varie selon le poste de soin (Optique, Dentaire, Hospitalisation…). Par la suite, ces postes seront eux-mêmes segmentés en sous postes qui seront détaillés plus tard. Il existe de nombreux sous postes. La manière de consommer varie également à l'intérieur de ces sous postes. Il convient donc de réaliser une tarification qui prend en compte cet aspect.

Ainsi, pour répondre à toutes ces contraintes, l'objectif est de construire un outil de tarification *a priori* en santé grâce à des modèles de *machine learning*, tout en prenant

en compte l'hétérogénéité de la consommation au sein des postes. Ici, l'approche *machine learning* présente plusieurs avantages :

- L'entraînement des modèles est facile à mettre en œuvre et efficace.
- Il est possible d'automatiser la création des modèles, ce qui est très utile étant donné le grand nombre de postes de soin en santé existant.
- Cela permet de créer des modèles qui ne sont pas forcément linéaires, ce qui est le cas en tarification.
- Des variables explicatives à la fois qualitatives et quantitatives peuvent être prises en compte.

Pour terminer, l'outil devra être utilisable par tous les membres de l'équipe actuariat. Il sera donc réalisé sur le logiciel R.

1.3.4 Les étapes de réalisation du projet

Afin de répondre au mieux à la problématique, il convient de distinguer plusieurs étapes :

- **Etape 1** : construction d'une base de données de taille significative. Les données doivent être traitées et homogénéisées, afin d'optimiser la construction des modèles.

- **Etape 2** : sélection des variables explicatives tarifaires. Il s'agit des caractéristiques de l'assuré et du contrat souscrit, exerçant une influence sur la fréquence et le coût des sinistres. Elles seront déterminées grâce à une analyse détaillée des données.

- **Etape 3** : construction des modèles de *machine learning* permettant de prédire la fréquence et le coût (frais réels) des sinistres. Afin de prendre en compte au mieux l'hétérogénéité de la consommation entre les sous postes, un modèle devra être créé pour chaque sous poste. Comme il existe de nombreux sous postes, l'enjeu est d'automatiser au maximum cette étape.

- **Etape 4** : analyse des performances des modèles et interprétabilité.

- **Etape 5** : calcul de la prime pure à partir des modèles établis.

- **Etape 6** : détermination de la prime commerciale. Elle se fera grâce à la prime pure calculée précédemment, et tiendra compte des différents chargements et taxes appliquées, ainsi que de la structure de cotisation du régime.

Par la suite, nous verrons donc plus en détail les hypothèses et modèles utilisés pour la construction de l'outil.

1.4 METHODE FREQUENCE / COUT MOYEN INDIVIDUALISEE

Dans cette partie, l'objectif est d'expliquée la méthode de tarification utilisée dans la construction de l'outil. En assurance, la méthode la plus utilisée est celle par une approche « Fréquence / Coût moyen ». C'est cette méthode qui a été retenue car elle est facile à mettre en œuvre et elle permet de distinguer l'influence des variables sur ces deux grandeurs. Elle s'appuie sur l'hypothèse que la fréquence et le coût moyen sont indépendants.

Soit une variable aléatoire S exprimant le montant total des sinistres d'un portefeuille. En actuariat, on définit la prime pure π par : $\pi = \mathbb{E}(S)$.

On introduit maintenant les variables suivantes. Soient :

- n le nombre de bénéficiaires du portefeuille.
- p le nombre de sous postes distincts, couvert par le contrat d'assurance.
- $S_{i,j}$ une variable aléatoire désignant le montant total des sinistres pour le bénéficiaire j, sur le sous poste i.

Le montant total des sinistres du portefeuille s'écrit alors :

$$S = \sum_{j=1}^{n} \sum_{i=1}^{p} S_{i,j}$$

On peut écrire $S_{i,j}$ sous la forme suivante :

$$S_{i,j} = \sum_{k=1}^{N_{i,j}} X_{i,j,k}$$

Où $N_{i,j}$ et $X_{i,j,k}$ sont des variables aléatoires exprimant respectivement le nombre de sinistres pour le bénéficiaire j, sur le sous poste i et le montant du sinistre numéro k, pour le bénéficiaire j, sur le sous poste i. On a alors :

$$S = \sum_{j=1}^{n} \sum_{i=1}^{p} \sum_{k=1}^{N_{i,j}} X_{i,j,k}$$

Quelques soient i et j, les $X_{i,j,k}$ sont des variables aléatoires supposées indépendantes et identiquement distribuées et supposées indépendantes de la variable aléatoire discrète $N_{i,j}$. La prime pure du portefeuille s'exprime donc de la manière suivante :

$$\pi = \mathbb{E}(S) = \sum_{j=1}^{n} \sum_{i=1}^{p} \mathbb{E}\left(\sum_{k=1}^{N_{i,j}} X_{i,j,k} \right)$$

Soit :

$$\pi = \sum_{j=1}^{n} \sum_{i=1}^{p} \mathbb{E}(N_{i,j}) \times \mathbb{E}(X_{i,j})$$

La prime pure par bénéficiaire est obtenue en divisant la prime pure du portefeuille par le nombre de bénéficiaires, c'est-à-dire : $\pi_{bénéf} = {}^{\pi}/_{n}$.

1.5 PRESENTATION DE H2O

Afin de répondre le mieux à nos besoins, nous avons travaillé avec le package H2O qui va être présenté ci-dessous.

1.5.1 Introduction à H2O

H2O.ai est une start-up de *data scientists* fondée en 2012 aux Etats-Unis. Ils ont créé le logiciel *open source* H2O, qui est une plateforme gratuite de *data science* et *machine learning*. Leur objectif est de démocratiser l'intelligence artificielle. H2O permet d'effectuer des analyses statistiques sur des grands jeux de données. Il propose également des algorithmes de *machine learning*, des modèles linéaires… L'objectif d'H2O est d'être une plateforme utilisée pour les métiers. Parmi ses clients utilisant déjà l'intelligence artificielle en production, on compte déjà PayPal ou encore EDF. [5]

La plateforme est open source. C'est-à-dire qu'elle autorise la création de travaux dérivés, la libre redistribution et qu'elle donne l'accès au code source.

Le logiciel peut être utilisé dans divers langages : R, Python, Scala… et sur les systèmes d'exploitation Windwos, Mac et Linux. Il nécessite également l'installation de Java. En effet, on peut écrire un programme sur R par exemple, qui va faire appel aux modèles H2O dont le cœur du code est écrit en Java. Une connexion entre R et Java va donc se faire afin de faire tourner les modèles.

H2O est bien documenté sur internet, ce qui permet un apprentissage rapide et facile.

Un point d'attention particulier est accordé à la version de H2O. En effet, il faut faire très attention à bien avoir la même version pour entraîner des modèles et pour les utiliser.

1.5.2 Présentation de l'algorithme « auto-ML »

Le principal élément qui nous a intéressé pour ce mémoire est l'algorithme « auto-ML » [6]. Il s'agit d'un algorithme capable de calibrer des modèles, qui seront définis ci-après, et d'identifier automatiquement le plus performant. Créé en 2017, cet outil permet d'améliorer les performances d'un modèle. Il peut être très utile à la fois pour les personnes ayant peu d'expérience dans la mise en place de modèles de *machine learning*, et à la fois pour des personnes plus expérimentées dans ce domaine et voulant mettre en place un grand nombre de modèles.

Dans la suite de ce mémoire, nous utiliseront le langage R lorsqu'il s'agira de langage de programmation. Mais les algorithmes fonctionnent de la même manière sur Python ou dans d'autres langages. Avant toute chose, il sera nécessaire d'installer et de charger le package H2O.

La fonction *h2o.automl* propose divers paramètres à modifier. Voici la liste non exhaustive des paramètres pouvant être pris en arguments par la fonction *h2o.automl* :

- **x** : un vecteur contenant les noms de colonne des variables explicatives.
- ***y*** : le nom de colonne de la variable à expliquer.
- ***traning_frame*** : les données permettant au modèle de s'entrainer, qui doivent être sous forme d'un objet H2O.
- ***nfolds*** : nombre de séparations pour la validation croisée.
- ***seed*** : un entier permettant de « fixer l'aléa ».
- ***max_runtime_secs*** : le temps maximal pour tester les modèles.

- ***max_models*** : le nombre maximal de modèles à tester. Si *max_runtime_secs* et *max_models* sont renseignés, l'algorithme s'arrête dès qu'un seul des deux paramètres est atteint. Ici, seul le paramètre *max_runtime_secs* a été utilisé, pour des questions pratiques, afin de limiter l'entrainement des modèles dans le temps.
- ***stopping_metric*** : critère / mesure permettant de stopper l'algorithme plus tôt. Par exemple, MSE, RMSE, variance…
- ***exclude_algos*** : permet d'exclure certaines catégories de modèles à tester.
- ***include_algos*** : permet de déterminer explicitement les catégories de modèles à tester.

Dans le cadre de ce mémoire, seuls les paramètres *x*, *y*, *training_frame*, *nfolds*, *seed* et *max_runtime_secs* ont été utilisés. D'autres paramètres plus spécifiques et plus complexes sont également disponibles, mais ils ne seront pas présentés.

En sortie, on obtient le meilleur modèle parmi ceux testés. Grâce à la fonction *h2o.get_leaderboard*, on peut également voir tous les modèles qui ont été testés ainsi que tous les critères de test. Et avec la fonction *h2o.getModel*, on peut récupérer tous les hyper paramètres de chaque modèle que l'on souhaite.

Afin de déterminer le meilleur modèle, la fonction *h2o.automl* teste une liste de modèles et pratique à chaque fois une validation croisée. La fonction teste toujours la même liste de modèles, dans le même ordre. Elle teste soit des modèles isolés, soit des grilles de recherches. Une grille de recherche permet de tester plusieurs ensembles de paramètres sur un même modèle. Dans H2O, les grilles de recherches peuvent être soit prédéfinies à l'avance, soit totalement aléatoires. La liste ordonnée des modèles testés par la fonction *h2o.automl* est la suivante :

- 3 modèles de XGBoost pré-spécifiés
- 1 grille de recherche de GLM fixée
- 1 forêt aléatoire par défaut
- 5 modèles de GBM pré-spécifiés
- 1 réseau de neurones par défaut
- 1 forêt extrêmement aléatoire
- 1 grille de recherche aléatoire de XGBoost
- 1 grille de recherche aléatoire de GBM
- 1 grille de recherche aléatoire de réseau de neurones

S'il n'y a pas assez de temps ou un nombre de modèles à tester trop petit, certains algorithmes peuvent être absents du classement. De même si certaines catégories d'algorithmes sont exclues. En revanche, s'il reste encore du temps après tous les tests, les deux grilles aléatoires les plus performantes seront relancées afin de trouver encore d'autres modèles.

En parallèle, la fonction *h2o.automl* teste deux combinaisons par agrégation de modèles prédictifs différentes. Les catégories de modèles qui ont été exclues précédemment ne seront pas inclues dans les agrégations. La première combinaison de modèles testée est celle incluant tous les modèles testés cités ci-dessus. La seconde combinaison de modèles testée est celle avec les meilleurs modèles de chaque catégorie.

Les différents modèles testés sont ensuite classés selon un critère prédéfini. Il s'agit de la perte de log pour les algorithmes de classification et du RMSE (qui est détaillé au

paragraphe 1.6.) pour les algorithmes de régression. Le meilleur modèle sera retenu pour réaliser les prédictions.

1.5.3 Description des différents modèles utilisés par H2O

Les modèles [7] évoqués ci-dessus vont maintenant être détaillés dans cette partie, excepté les GLM et les XGBoost (variante améliorée du GBM), car ils ne seront pas utilisés dans ce mémoire.

Forêt aléatoire

La forêt aléatoire est un modèle de *machine learning* basé sur plusieurs CART ou arbres de décision. Les CART sont définis en annexe 2.

La construction d'une forêt aléatoire ressemble à la méthode du *bagging*, également définie en annexe 3. De la même manière que pour le *bagging*, on crée par *bootstrap* de nouveaux échantillons de taille n, à partir d'un échantillon d'apprentissage de la même taille. Avant l'agrégation des prédictions, une nouvelle étape vient s'ajouter. L'apprentissage ne se base plus sur l'ensemble des variables explicatives, mais plutôt sur une partie de ses variables explicatives tirées aléatoirement à chaque nœud de l'arbre. Cette étape permet de décorréler les arbres entre eux et de réduire significativement le temps de calcul.

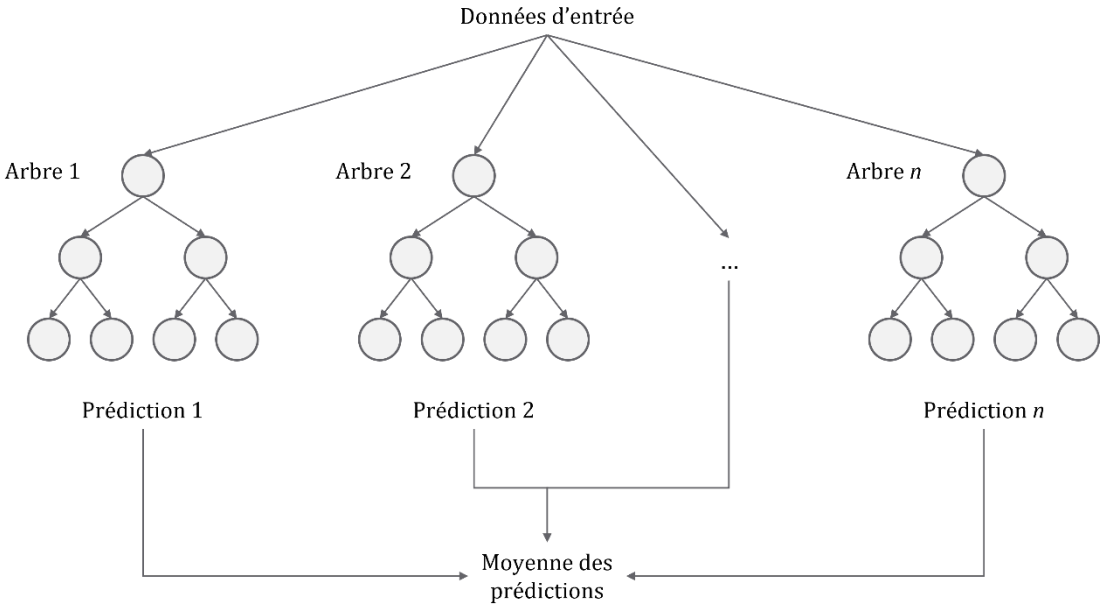


Figure 1.2 – Fonctionnement d'une forêt aléatoire

Soit m le nombre total de variables explicatives et m' le nombre de variables tirées aléatoirement sur un nœud. En général, on préférera choisir :

- $m' = \sqrt{m}$ pour un arbre de classification.
- $m' = {}^{m}/_{3}$ pour un arbre de régression.

Cette méthode d'agrégation permet de réduire la variance du modèle.

H2O utilise un modèle de forêt aléatoire déjà construit par défaut ; il n'y a donc qu'un seul jeu de paramètres qui est testé, présenté ci-dessous [6] :

Paramètres	Valeurs prises
ntrees	50
max_depth	20
min_rows	1
sample_rate	0,632
mtries	-1
stopping_rounds	3
stopping_metric	AUTO

Tableau 1.1 – Paramètres de la forêt aléatoire testée sur H2O

H2O propose également un autre type d'algorithme semblable à celui de la forêt aléatoire : l'algorithme de la forêt extrêmement aléatoire. La méthode d'agrégation est la même que pour la forêt aléatoire. Seule la construction de l'arbre diffère légèrement. A chaque nœud, les seuils de segmentation ne sont plus choisis de telle sorte qu'ils soient les plus discriminants, mais ils sont choisis aléatoirement. Cela permet de réduire encore plus la variance, mais augmente cependant le biais du modèle. De même que pour les forêt aléatoires, H20 ne teste qu'un seul jeu de paramètres par défaut, présenté ci-dessous :

Paramètres	Valeurs prises
ntrees	50
max_depth	20
min_rows	1
sample_rate	0,632
mtries	-1
col_sample_rate_per_tree	0,8
col_sample_rate_change_per_level	1
min_split_improvement	0,00001
stopping_rounds	3
stopping_metric	AUTO

Tableau 1.2 – Paramètres de la forêt extrêmement aléatoire testée sur H2O

GBM (*Gradient Boosting Machine*)

Le *gradient boosting machine* est un modèle de *machine learning* se définissant comme l'agrégation entre deux méthodes : d'un côté celle du *boosting*, et de l'autre celle de la descente de gradient. Il peut être utilisé aussi bien pour des problèmes de régression que de classification.

La technique du *boosting* consiste à utiliser plusieurs modèles, qui seront ensuite agrégés afin d'obtenir un résultat final :

- L'algorithme va d'abord déterminer un premier modèle M_0 en formant généralement un arbre de décision (ou plus rarement, un réseau de neurones ou un *Support Vector Machine*), puis il va ensuite l'évaluer. À la suite de cette évaluation, un poids est accordé à chaque individu en fonction de la performance de la prédiction. Les meilleures prédictions auront les poids les plus faibles et *vice versa*, permettant ainsi à l'algorithme de mieux se focaliser sur les individus dont les prédictions auront été les moins bonnes.
- L'algorithme déterminera ensuite un nouveau modèle M_1, dont l'objectif cette fois sera, non pas de prédire la variable à expliquer, mais plutôt l'erreur entre la prédiction du modèle précédent et la valeur réelle. La nouvelle prédiction sera la somme des prédictions des modèles M_0 et M_1. Le modèle $M_0 + M_1$ sera ensuite évalué et les poids seront alors corrigés.
- Puis l'algorithme va déterminer une succession de modèles M_i modélisant à chaque fois l'erreur entre la prédiction du modèle $M_0 + M_1 + \cdots + M_{i-1}$ et la valeur réelle, qui seront évalués et à partir desquels les poids seront corrigés.
- L'algorithme s'arrête lorsque i a atteint une valeur limite fixée au préalable.

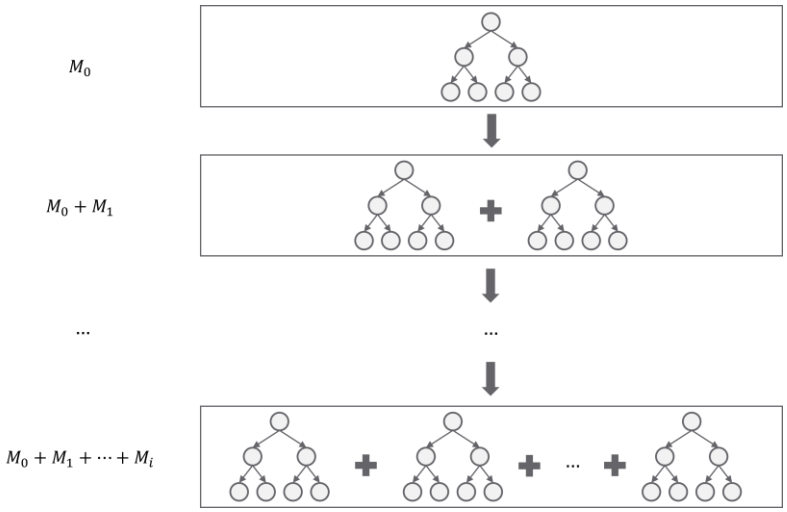


Figure 1.3 – Fonctionnement d'un GBM

Pour rappel, le gradient d'une fonction $f: (x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n)$, noté $\overrightarrow{grad}\, f$, est définit de la manière suivante :

$$\overrightarrow{grad}\, f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

La descente de gradient est un algorithme itératif d'optimisation différentiable servant à minimiser une fonction réelle différentiable convexe J. Il procède par amélioration successive en se déplaçant à l'opposé du gradient (c'est-à-dire le long d'une direction de descente), afin de faire décroitre la fonction. La descente de gradient intègre un coefficient de rétrécissement α, aussi appelé vitesse d'apprentissage. Plus α est élevé, plus l'algorithme sera rapide. Cependant, si α est trop grand, alors le pas de l'algorithme sera trop grand et le minimum risquerait de ne jamais être atteint. Un α trop petit rendrait, quant à lui, le temps de convergence de l'algorithme trop long. Il faut donc trouver un compromis entre les deux.

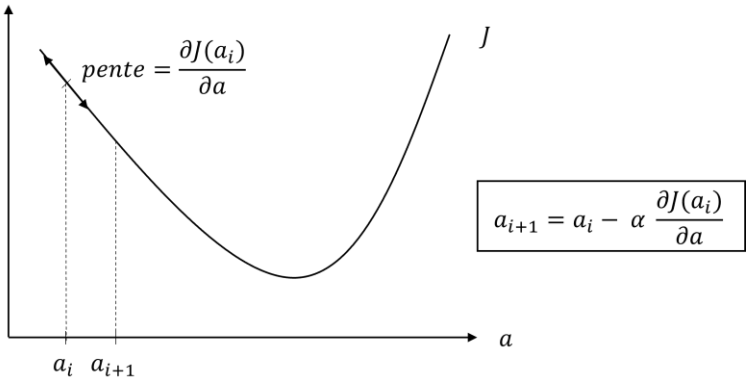


Figure 1.4 – Illustration de l'algorithme de descente de gradient

Dans le cas des GBM, l'algorithme de descente de gradient va être utilisé comme critère de sélection des modèles décrits ci-dessus (voir figure 1.4). A chaque étape de la procédure, l'algorithme va être utilisé pour minimiser la fonction de perte, qui permet le calcul des poids des individus. Minimiser la fonction de perte revient donc minimiser l'erreur d'estimation du modèle.

La fonction de perte dépend du type de variable à modéliser :

- Pour une variable binaire, la fonction de perte logistique est utilisée.
- Pour une variable qualitative, il s'agit de la fonction de perte multinomiale.
- Pour une variable quantitative discrète, on préférera prendre la log-vraisemblance de Poisson.
- Pour une variable quantitative continue, le MSE (il s'agit du carré du RMSE, qui est défini au paragraphe 1.6), est le plus souvent utilisé. Le MAE (également défini au paragraphe 1.6) est peut aussi être utilisé.

Le *gradient boosting machine* est une technique rapide et précise, en particulier pour des données complexes et volumineuses. Mais il est sujet au surapprentissage, étant donné que son but est de réduire la fonction de perte. Afin d'éviter cela, il est important de jouer sur différents paramètres tels que la taille des arbres, le nombre d'itération, le coefficient de rétrécissement…

Pour les modèles de GBM pré-spécifiés, H2O fixe le nombre d'arbres à 10 000 ainsi que les paramètres d'échantillonnage des données à 80 %. Il va ensuite jouer sur le couple (profondeur de l'arbre maximale, nombre d'observations minimal dans une feuille) en proposant cinq couples de valeurs, puis repérer le meilleur couple. Voici la grille prédéfinie par H2O [6] :

Paramètres	Valeurs recherchées
max_depth	{6 ; 7 ; 8 ; 10 ; 15}
min_rows	{1 ; 10 ; 10 ; 10 ; 100}

Tableau 1.3 – Paramètres des GBM testés sur H2O

Réseau de neurones

Un réseau de neurones est un modèle de *machine learning*, et plus particulièrement de *deep learning*, qui s'inspire du système de fonctionnement du cerveau humain pour apprendre grâce à une approche connexionniste.

Pour expliquer brièvement son fonctionnement, le réseau de neurones reçoit des signaux en entrée qui envoient des informations qui vont se propager et permettre d'activer une fonction, appelée fonction d'activation. Cette fonction a pour argument les signaux d'entrée auxquels des poids ont été accordés, pour donner un signal en sortie. Il existe différents types de réseaux de neurones avec des caractéristiques différentes (positionnement des neurones, nombres de couches de neurones, sens de propagation de l'information…).

Avant de comprendre le fonctionnement d'un réseau, il est nécessaire de comprendre le fonctionnement d'un neurone au sein de ce réseau. On va dont voir comment fonctionne un neurone formel.

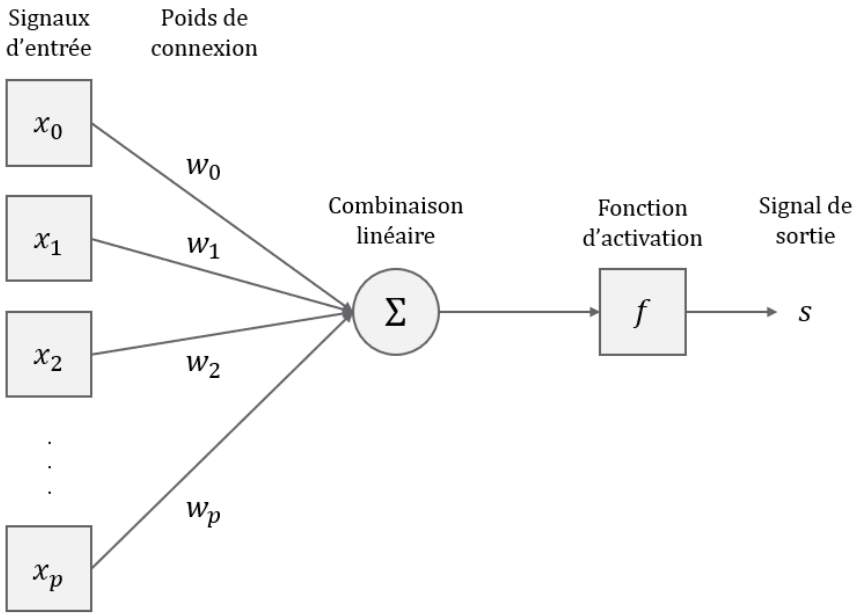


Figure 1.5 – Fonctionnement d'un neurone

Un neurone formel se définit par :

- Des **signaux d'entrée** $(x_0, x_1, \dots, x_p)$. Il s'agit de valeurs numériques qui doivent être normalisées. x_0 est appelé « biais du modèle » et a pour valeur 1.
- Des **poids de connexion** $(w_0, w_1, \dots, w_p)$. Ils sont estimés durant l'apprentissage du modèle, le but étant de trouver les poids optimaux.

- Un **signal de sortie** s, exprimé à l'aide d'une fonction d'activation.
- Une **fonction d'activation** f, telle que $s = f\left(w_0 + \sum_{i=1}^{p} w_i x_i\right)$. C'est une combinaison affine des signaux d'entrée, pondérée par les poids de connexion. On distingue différents types de neurones en fonction de f. Voici les différentes fonctions d'activation pouvant être utilisées :

Type de neurone	Fonction d'activation
Linéaire	$f(x) = x$
Sigmoïde	$f(x) = \dfrac{1}{1+e^{-x}}$
Seuil	$f(x) = \mathbb{I}_{[0,+\infty[}(x)$
Radial	$f(x) = \sqrt{\dfrac{1}{2\pi}} e^{-\frac{1}{2}x^2}$
Stochastique	$f(x) = \begin{cases} 1, & \text{avec une probabilité de } \dfrac{1}{1+e^{-\frac{x}{H}}} \\ 0, & \text{sinon} \end{cases}$

Tableau 1.4 – Fonctions d'activation des neurones

On s'intéresse maintenant au fonctionnement d'un perceptron multicouche (PMC). Même s'ils diffèrent par leur structure, le principe de fonctionnement des autres types de réseaux de neurones reste sensiblement le même. Il s'agit d'un réseau statique de neurones qui fonction par apprentissage supervisé.

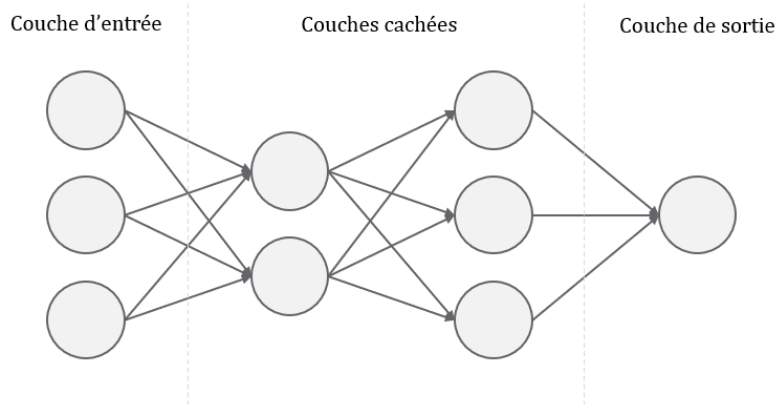


Figure 1.6 – Fonctionnement d'un réseau de neurone

Le PMC est composé de :

- Une couche en entrée qui va recevoir le signal d'entrée.
- Une ou plusieurs couche(s) cachée(s) qui vont à chaque fois appliquer la fonction d'activation. Plus il y a de couches cachées, plus le modèle est puissant mais plus le temps de calcul est long.
- Une couche en sortie qui va renvoyer la réponse en sortie.
- Des poids de connexion qui relient une couche à la suivante.

L'objectif de l'algorithme est de trouver les poids de connexion optimaux. Avant toute chose, il faut d'abord paramétrer la complexité de l'algorithme (à travers le nombre de

couches, le nombre de neurones par couche et le taux d'erreur maximal). Puis l'algorithme effectue les étapes suivantes :

1. Normalisation des données d'entrée
2. Tirage aléatoire des poids pour le premier essai.
3. Calcul de la valeur de sortie grâce à la fonction d'activation.
4. Calcul de l'erreur ou fonction de perte pour chacune des entrées.
5. Evaluation du gradient de la fonction de perte par rétropropagation jusqu'à l'entrée pour voir d'où vient l'erreur.
6. Mise à jour des poids.
7. Les étapes 3 à 6 sont répétées jusqu'à ce que le nombre d'itération dépasse la limite fixée, ou bien jusqu'à ce que la fonction de perte dépasse un certain seuil également fixé.

Dans la théorie des réseaux de neurones, le **théorème d'approximation universelle** [8] dit que toute fonction continue, sur des sous-ensembles compacts de $\mathbb{R}^n$, peut être approximée par un PMC à une seule couche cachée et un neurone de sortie de type linéaire.

Il existe deux types de réseaux de neurones :

- Les réseaux *feed-forward* : le réseau est non bouclé ; les informations ne circulent que dans un seul sans.

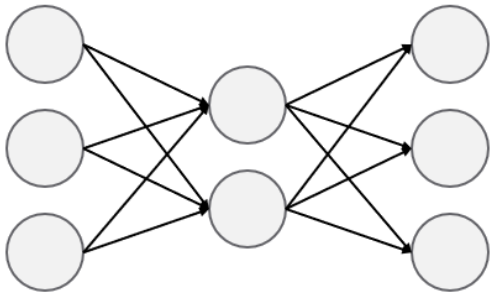


Figure 1.7 – Réseau de neurones feed-forward

- Les réseaux *feed-back* : le réseau est bouclé ; les informations circulent de manière cyclique.

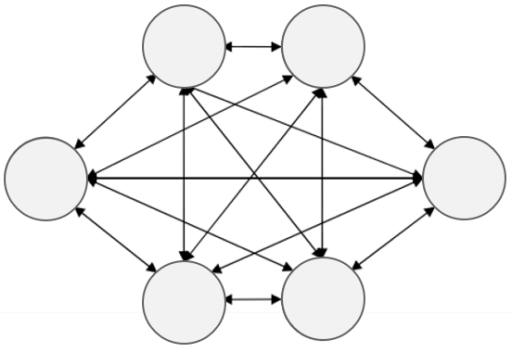


Figure 1.8 – Réseau de neurones feed-back

H2O utilise un PMC à descente de gradient par rétropropagation déjà construit par défaut ; il n'y a donc qu'un seul jeu de paramètres qui est testé [6].

Stacking ou agrégation de modèles prédictifs

Le *stacking* ou agrégation de modèles consiste à faire coopérer plusieurs modèles, dans le but de les rendre plus performants que chacun d'eux pris individuellement (attention : ce n'est pas toujours le cas). Plus les modèles seront performants individuellement, plus l'agrégation de ces modèles aura de chances d'être performante.

Le *stacking* fonctionne sur le même échantillon d'apprentissage, quelques soient les modèles. Il se différencie du *bagging*, par le fait d'utiliser des modèles issus de familles différentes (par exemple, un CART, un réseau de neurones, un GBM…), tandis que le bagging utilise toujours des modèles d'une même famille (par exemple, que des CART).

La combinaison crée s'appelle le métamodèle. La performance du métamodèle s'accroît lorsque les modèles sont diversifiés et complémentaires. En effet, on mise sur le fait que les modèles ne soient pas unanimes dans leur prédiction afin les faire coopérer et de permettre de corriger ceux qui donnent de mauvaises prédictions. La diversité des modèles s'obtient soit en prenant des modèles de différentes familles, soit en prenant deux modèles similaires mais avec des paramétrages très différents.

Les différents modèles sont construits indépendamment les uns des autres. Puis l'agrégation se fait selon une des deux manières suivantes :

- La **combinaison par vote simple** : on regarde le résultat de chaque modèle, le résultat final sera celui qui aura la majorité pour des variables à prédire qualitatives, ou bien la moyenne pour des variables à prédire qualitatives.
- La **combinaison par vote pondéré** : à chaque modèle des poids sont accordés. Le résultat final sera alors celui qui aura la majorité pondérée ou bien la moyenne pondérée des résultats. Ce type de combinaison a des limites car l'intérêt du *stacking* est d'utiliser des modèles diversifiés.

H2O propose deux modèles de stacking différents, tous deux utilisant la combinaison par vote simple :

- *All models* : cette agrégation de modèles inclue tous les algorithmes qui ont été testé au cours de l'algorithme *auto.ml*.
- *Best of family* : cette agrégation de modèle utilise les meilleurs de modèles de chaque famille présentée ci-dessus.

Recherche par grille aléatoire

H2O recherche également les meilleurs algorithmes avec des recherches par grille aléatoire. C'est le cas pour les GBM, les XGBoost et les réseaux de neurones. La recherche par grille aléatoire consiste à faire tourner un modèle avec des hyperparamètres choisis aléatoirement parmi une liste de potentielles valeurs. Ces listes de valeurs se trouvent en annexe 4 [6].

1.6 INDICATEURS DE PERFORMANCES DES MODELES

Afin de mesurer les performances et de valider les modèles établis par les algorithmes d'apprentissage supervisés, plusieurs indicateurs peuvent être utilisés. Ils permettent de savoir si le modèle a prédit des valeurs pertinentes, qui ne sont pas trop éloignées de la réalité. Les indicateurs varient en fonction du type d'algorithme :

- Algorithmes de classification : taux d'erreur, spécificité, sensibilité… obtenus à partir de la matrice de confusion. Ces indicateurs ne seront pas détaillés dans ce mémoire, car ce type d'algorithme n'est pas utilisé.

- Algorithmes de régression : il en existe plusieurs (RMSE, MAE, SCR…). Le RMSE et le MAE, utilisés dans ce mémoire, sont détaillés ci-dessous.

RMSE (Root Mean Square Error)

Le RMSE est la racine de l'erreur quadratique moyenne (MSE). Sur un échantillon de taille n, il se calcule de la manière suivante :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \text{avec :}$$

- y_i est la valeur de la $i^{ème}$ observation de l'échantillon.
- $\hat{y}_i$ est la valeur prédite pour la $i^{ème}$ observation.

Le RMSE est toujours positif. Plus il est faible, meilleure est la prédiction. Cependant, le RMSE peut être difficile à interpréter en tant que tel. Seul, on ne peut pas dire si un RMSE est bon ou mauvais. Afin de faciliter son interprétation, il peut être intéressant de le normaliser, par exemple grâce à l'écart-type ou à la moyenne. Ainsi, le RMSE normalisé [9] (par l'écart-type), noté $RMSE_{norm}$, vaut :

$$RMSE_{norm} = \frac{RMSE}{\sigma_y}, \text{avec } \sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} \text{ où } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

MAE (Mean Absolute Error)

Le MAE est très similaire au RMSE, cependant, on ne regarde pas l'erreur quadratique mais la valeur absolue de l'erreur. Sur un échantillon de taille n, il se calcule de la manière suivante :

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

Le MAE est également toujours positif et plus il est faible, meilleure est la prédiction. De même que pour le RMSE, le MAE est difficilement interprétable seul ; il faudra donc le normaliser. C'est la moyenne qui sera utilisée pour cela. Ainsi, le MAE normalisé (par la moyenne), noté MAE_{norm}, vaut :

$$MAE_{norm} = \frac{MAE}{\bar{y}}$$

1.7 INTERPRETABILITE DES MODELES DE MACHINE LEARNING

1.7.1 Nécessité d'interprétabilité des modèles

Les algorithmes de *machine learning* ne sont pas tous faciles à comprendre et à expliquer. Sur certains algorithmes, les résultats sont facilement explicables, car ils se rapprochent des algorithmes de régression linéaire. C'est le cas par exemple des GLM, ou bien des CART. En revanche, sur d'autres, malgré leur précision, il est plus difficile d'expliquer les prédictions, du fait de leur grande complexité, voire d'un effet boite noire pour certains. C'est le cas réseaux de neurones ou des agrégations de modèles par exemples. Ainsi, les algorithmes de *machine learning* gagnent en précision, mais perdent en explicabilité.

Toutefois, un actuaire doit pouvoir être capable d'expliquer simplement les résultats de l'algorithme, même si celui-ci est complexe. Il y a plusieurs raisons à cela. Tout d'abord, pour l'aspect légal : la décision ne doit pas être prise que par une machine, mais par un humain. Ensuite, cela permet de valider le modèle et de vérifier qu'il est bien cohérent avec la réalité. Et enfin, cela permet de pouvoir expliquer aux clients le résultat, et de pouvoir éventuellement leur apporter des conseils et des précisions.

Pour répondre à cette problématique, l'interprétabilité des modèles de *machine learning* [10] se développe depuis quelques années. Son but est de mieux comprendre les modèles et les rendre plus interprétables.

L'interprétabilité permet d'expliquer les prédictions de l'algorithme. Elle se caractérise par trois éléments :

- Son type :
  - ***A priori*** : réalisée avant la création du modèle, au moment du choix de la famille de modèle et sans hypothèses spécifiques sur les données.
  - ***A posteriori*** : réalisée après l'élaboration du modèle, permettant donc d'augmenter la précision descriptive du modèle.
- Son approche :
  - **Spécifique** au modèle : propre au modèle étudié.
  - **Agnostique** au modèle : qui ne dépend pas du modèle étudié, apportant ainsi une certaine flexibilité.
- Son critère d'explicabilité :
  - **Global** : qui permet d'expliquer le modèle dans sa globalité et de donner l'importance des variables, c'est-à-dire leur contribution dans la prédiction du résultat.
  - **Local** : qui permet d'expliquer la prévision pour un individu donné et comment les différentes variables permettent d'obtenir le résultat pour cet individu.

Il existe de nombreuses méthodes d'interprétabilité, mais toutes ne sont pas forcément démontrées. Dans le cadre de ce mémoire, nous nous intéresserons aux graphiques de dépendance partielle, ainsi qu'aux valeurs de Shapley, qui sont tous les deux des méthodes *a posteriori*.

1.7.2 Graphiques de dépendance partielle

Les graphiques de dépendance partielle ou *Partial Dependance Plot* (PDP) [11] sont une méthode d'interprétabilité globale agnostique au modèle. Leur analyse permet de montrer l'effet marginal d'une variable explicative sur la prédiction réalisée par un modèle.

On considère une base de données de taille n, composée des variables explicatives $(X_1 = (x_1^1, \dots, x_1^n), \dots, X_p = (x_p^1, \dots, x_p^n))$ indépendantes entre elles, et d'une variable à expliquer $Y = (y^1, \dots, y^n)$. La base de données peut être l'échantillon d'apprentissage ou bien l'échantillon test par exemple. On dispose également d'un modèle prédictif f ainsi que des prédictions $Y_{pred} = (y_{pred}^1, \dots, y_{pred}^n)$.

On veut connaître l'effet de certaines variables X_S sur la prédiction. Les variables explicatives restantes sont notées X_C. X_S peut par exemple être égal à X_1, et dans ce cas, $X_C = (X_2, \dots, X_p)$.

Pour cela, on définit la fonction de dépendance partielle :

$$f_{x_S}(x_S) = \mathbb{E}_{X_C}\big(f(x_S, X_C)\big) = \int f(x_S, x_C) d\mathbb{P}_{X_C}(x_C)$$

Celle-ci peut être approximée par l'expression suivante :

$$f_{x_S}(x_S) \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^i)$$

L'algorithme de construction de la courbe du graphique de dépendance partielle est le suivant :

> On détermine la ou les variables dont on veut connaître l'effet sur la prédiction : par simplification, on prendra ici $X_S = X_1$ et $X_C = (X_2, \dots, X_p)$.
>
> Pour i allant de 1 à n :
>
> - Pour chaque individu k de la base de données, x_1^k est remplacé par x_1^i. X_1 prendra donc toutes les valeurs possibles de la base de données.
> - La variable à expliquer est ensuite prédite grâce au modèle : calcul de $f\big(x_1^i, x_2^k, \dots, x_p^k\big) = f(x_1^i, x_C^k)$. On a donc pour chaque individu k, une courbe de prédiction de coordonnées $(x_1^i, f(x_1^i, x_C^k))$, appelée *Individual Conditionnal Expectations* (ICE).
> - La dépendance partielle est calculée grâce à son estimation : $f_{x_1}(x_1^i) \simeq \frac{1}{n}\sum_{k=1}^{n} f(x_1^i, x_C^k)$.
>
> En sortie, on obtient donc une courbe de cordonnées $(x_1^i, f(x_1^i))$. Il s'agit en fait de la moyenne des courbes de chaque individu.

Le PDP permet donc d'interpréter simplement et globalement un modèle prédictif. Il est également facile à implémenter. Toutefois, il prend en compte l'effet d'une variable sur la prédiction, indépendamment des autres variables, et pouvant ainsi masquer dans certains cas des effets hétérogènes (le PDP étant une moyenne de courbes). De plus, il repose sur une hypothèse forte d'indépendance des variables, qui n'est pas toujours évidente. Il peut donc être intéressant de compléter cette méthode avec d'autres méthodes d'interprétabilité.

1.7.3 Les valeurs de Shapley

Les valeurs de Shapley sont une méthode d'interprétabilité locale, basée sur la théorie des jeux.

La théorie des jeux

Les valeurs de Shapley, introduites en 1953 par Shapley, reposent sur la théorie des jeux coopératifs. N joueurs collaborent ensemble pour obtenir un gain total G. Comment répartir de manière équitable le gain G entre tous les joueurs, en tenant compte de la contribution de chacun (à la fois seul et en groupe) ?

Pour répondre à cela, on note $S \subseteq \{1, \dots, N\}$, une coalition (ou sous-ensemble) de joueurs, et on introduit une fonction caractéristique v, telle que $v(S)$ définie le gain maximal de la coalition S.

La contribution marginale du joueur i dans la coalition S est notée $\Delta_v(i, S)$ et est telle que :

$$\Delta_v(i, S) = v(S \cup \{i\}) - v(S)$$

Elle permet de mesurer le gain obtenu par la coalition avec et sans le joueur i.

Soit la fonction $\phi_v : \begin{cases} \{1, \dots, N\} \to \mathbb{R} \\ i \to \phi_v(i) \end{cases}$.

Le théorème de Shapley assure l'existence et l'unicité d'une fonction ϕ_v, qui garantit une répartition équitable du gain total entre tous les joueurs, et qui vérifie les quatre propriétés suivantes :

1. **Efficacité** : la somme des parts attribuées à chaque joueur est égale au gain total. Autrement dit :

$$\sum_{i \in \{1,\dots,N\}} \phi_v(i) = v(\{1, \dots, N\}) = G$$

2. **Symétrie** : si deux joueurs contribuent de la même manière à toutes les coalitions, alors leurs parts seront égales. C'est-à-dire :

$$\forall\, S \,/\, v(S \cup \{i\}) = v(S \cup \{j\}) \Rightarrow \phi_v(i) = \phi_v(j)$$

3. **Joueur nul** : si toutes les coalitions dans lesquelles un joueur est présent ont les mêmes gains, avec ou sans lui, alors le gain de ce joueur est nul. Autrement dit :

$$\forall\, S \,/\, v(S \cup \{i\}) = v(S) \Rightarrow \phi_v(i) = 0$$

4. **Additivité** : si un joueur participe à deux jeux, avec les mêmes joueurs, et avec des fonctions caractéristiques v et w, alors la somme de ses parts pour chaque jeu est égale à la part du gain total des deux jeux. Cela signifie que :

$$\phi_v(i) + \phi_w(i) = \phi_{v+w}(i)$$

Cette fonction ϕ_v est définie par l'expression suivante pour le joueur i :

$$\phi_v(i) = \sum_{S \subseteq \{1,\dots,N\} \setminus \{i\}} \frac{|S|! \times (N - |S| - 1)!}{N!} \times \Delta_v(i, S)$$

Où $|S|$ représente le nombre de joueurs de la coalition S.

$\phi_v(i)$ est appelée « valeur de Shapley du joueur i ». C'est sa contribution marginale pour toutes les coalitions de joueurs possible. [12]

Lien avec le *machine learning*

En *machine learning*, l'objectif est d'expliquer la valeur prédite $f(x)$, associée à une observation x. En 2017, Lundberg et Lee proposent une méthode permettant d'interpréter les prédictions d'un modèle de *machine learning* : SHapley Additive exPlanations (SHAP) [13], grâce à un parallèle entre la théorie des jeux et le *machine learning*, en considérant que :

- Les joueurs sont les modalités des variables explicatives prises par x, notées x_i. N est donc le nombre de variables explicatives.
- Le gain à partager est la différence entre la valeur prédite $f(x)$ et l'espérance de la prédiction $\mathbb{E}(f(X))$.

La fonction caractéristique de la coalition $s \subseteq \{1, \dots, N\}$ devient alors $v(s) = \mathbb{E}(f(X) \mid X_s = x_s)$. La valeur de Shapley associé à la modalité x_i est appelée « valeur de SHAP », et s'écrit donc de la manière suivante :

$$\phi_v(x_i) = \sum_{s \subseteq \{1,\dots,N\} \setminus \{i\}} \frac{|s|! \times (N - |s| - 1)!}{N!} \times \left(\mathbb{E}\left(f(X) \middle| X_{s \cup \{i\}} = x_{s \cup \{i\}}\right) - \mathbb{E}(f(X) | X_s = x_s) \right)$$

Tout comme en théorie des jeux, ϕ_v vérifie bien les propriétés d'efficacité, de symétrie, du joueur nul et d'additivité.

Bien qu'en théorie, les valeurs de SHAP s'expriment facilement, la mise en pratique est plus compliquée. En effet, le nombre de coalitions à parcourir étant en 2^N, le calcul peut rapidement devenir très lourd et complexe. De plus, il peut être parfois difficile de donner une expression des espérances conditionnelles.

Dans la pratique, ce sont donc des algorithmes d'estimation des valeurs de SHAP qui sont utilisés, comme le Kernel SHAP (agnostique), le *Tree* SHAP (spécifique), le *Deep* SHAP (spécifique)…

La méthode SHAP proposée par Lundberg et Lee utilise les valeurs de SHAP (obtenues à partir d'algorithmes d'estimation) et donne une expression simplifiée et interprétable de la prédiction. L'expression de la valeur prédite par l'approche SHAP est la suivante :

$$f(x) \simeq g(z) = \phi_0 + \sum_{i=1}^{N} z_i \phi_i$$

Avec ϕ_0 la valeur de base (c'est-à-dire la moyenne de toutes les prédictions du jeu de données), ϕ_i les valeurs de SHAP définie ci-dessus, z le vecteur de coalition tel que les $z_i \in \{0,1\}$ (0 pour une variable absente et 1 pour une variable observée) et N le nombre de variables.

Cette méthode permet de déterminer les effets des différentes variables sur la prédiction pour expliquer l'écart constaté par rapport à la valeur de base.

Pour passer d'une interprétabilité locale à une interprétabilité globale, il suffit de prendre la moyenne des valeurs absolues des valeurs de SHAP estimées pour chaque variable. Cela permet d'obtenir l'importance des variables sur le modèle.

Algorithme Tree SHAP

L'algorithme Tree SHAP permet de calculer les valeurs de SHAP exacte, pour des modèles basés sur des arbres de décision : arbres de régression, arbres de classification, forêts aléatoires et GBM.

Sans optimisation, le calcul des valeurs de SHAP pour un tel modèle aurait une complexité en $O(LT2^N)$, avec L le nombre maximal de feuilles sur tous les arbres, T le nombre d'arbres et N le nombre de variables explicatives.

L'algorithme Tree SHAP permet une optimisation du calcul des valeurs de SHAP grâce à la récursivité. Un algorithme récursif est un algorithme qui résout des problèmes, en s'appelant lui-même mais sur une instance plus petite, jusqu'à atteindre une condition d'arrêt. La récursivité de l'algorithme Tree SHAP lui permet donc de garder une trace de tous les sous-ensembles qui s'écoulent dans chaque nœud de l'arbre. Ainsi, au lieu de repartir du nœud racine à chaque fois, le calcul est effectué en descendant ou en remontant d'un cran dans l'arbre.

La complexité de l'algorithme Tree SHAP est donc bien meilleure que celle de l'algorithme de calcul des valeurs de SHAP sur un ensemble d'arbres non optimisé. Sa complexité est en $O(LTD^2)$, avec D la profondeur maximale des arbres. [14]

C'est cet algorithme qui sera utilisé dans ce mémoire pour calculer les valeurs de SHAP.

2 CREATION DE LA BASE DE DONNEES ET ANALYSES

L'objectif de ce chapitre est de présenter et analyser les données sur lesquelles se sont basés les algorithmes d'apprentissages pour la modélisation des postes.

2.1 PRESENTATION DU PORTEFEUILLE DE DONNEES

Dans le cadre de ce mémoire, nous disposons des lignes à lignes de prestations, des effectifs et des garanties de dix des clients d'Adding. Pour des raisons de confidentialités, ils ne seront pas nommés. Il s'agit de données de l'année de survenance 2018. Ces données proviennent pour neuf clients sur dix d'un seul gestionnaire. Ce sont des clients qui possèdent des garanties bonnes ou moyennes. Afin d'élargir le panel des niveaux de garanties, nous avons choisi de rajouter un client ayant des garanties moins bonnes. Les données ont déjà subi des premiers traitements, dans le cadre des analyses de consommation : vérification et retraitement des données aberrantes, correspondances entre les libellés gestionnaire et les libellés Adding pour certaines variables, calculs des variables numériques manquantes… Les données sont donc déjà homogènes.

A l'intérieur du portefeuille, on compte 154 996 bénéficiaires, pour un montant total de consommation de 36 822 322 €. Les paragraphes suivants vont permettre de détailler chaque type de données.

2.1.1 Effectifs

Il s'agit de toutes les données qui concernent les bénéficiaires du régime. Parmi les informations observées, nous avons choisi de garder les informations suivantes pour la suite de notre étude :

- Identifiant du bénéficiaire (unique pour chaque personne)
- Identifiant de l'assuré auquel le bénéficiaire est rattaché
- Société
- Catégorie (actifs, portables, retraités…)
- CSP (cadres, non cadres ou ensemble du personnel)
- Type de bénéficiaire (assurés, conjoints, enfants ou ascendants)
- Affiliation (base ou base + option)
- Âge
- Temps de présence (ou temps d'exposition)
- Structure familiale (assurés seuls, assurés avec conjoint, assurés avec conjoint et enfant(s), assurés avec 1 enfant, assurés avec 2 enfants, assurés avec 3 enfants et +)

D'autres informations auraient pu être intéressantes dans le cadre d'une tarification comme le genre, le régime ou le département des bénéficiaires par exemple, mais nous ne disposons pas de ces informations. Nous verrons par la suite les informations qui seront utilisées pour la tarification.

2.1.2 Prestations

Il s'agit de tous les actes de frais santé qui ont eu lieu au cours de l'année 2018 et qui ont été remboursés par la complémentaire Santé. Les bénéficiaires ont deux ans pour se faire rembourser à compter de la date de soin. Or les dates d'arrêté sont antérieures au 31 décembre 2020. Elles ne sont par ailleurs pas toutes les mêmes pour chaque client. Il faudra donc prendre cela en compte pour la suite en intégrant des Provisions

pour Sinistres À payer (PSAP). Parmi les informations à notre disposition, nous avons choisi de garder les informations suivantes :

- Identifiant du bénéficiaire (unique pour chaque personne)
- Identifiant de l'assuré auquel le bénéficiaire est rattaché
- Société
- Code acte
- Libellé acte
- Famille acte
- Grand poste
- Sous poste
- Spécialité Médecine alternative
- Date de soin
- Date de remboursement
- Adhésion à l'OPTAM (Option Pratique Tarifaire Maîtrisée)
- Réseau optique
- Secteur de conventionnement
- Quantité d'acte
- Frais réels
- Base de remboursement Sécurité Sociale
- Taux de remboursement Sécurité Sociale
- Remboursement Sécurité Sociale
- Remboursement autre mutuelle
- Remboursement complémentaire base
- Remboursement complémentaire option

Les codes acte, libellés acte et familles d'acte permettent de décrire en détail l'acte de soin qui a été effectué. Dans un souci d'harmonisation des études, Adding a classifié ces actes en grands postes et sous postes grâce à des tables de correspondances. Voici la classification des actes utilisée :

Optique
Montures Adulte
Montures Enfant
Verres simples Adulte
Verres simples Enfant
Verres complexes Adulte
Verres complexes Enfant
Verres hyper complexes Adulte
Verres hyper complexes Enfant
Chirurgie optique
Lentilles jetables
Lentilles prises en charge
Lentilles non prises en charge
Autres sous postes Optique

Dentaire
Soins dentaires
Orthodontie TO90
Orthodontie Autre
Implants dentaires
Inlays et Onlays
SPR 50 (couronnes)
SPR 57 (inlays core)
SPR autres (autres prothèses dentaires)
Parodontologie non prise en charge
Autres sous postes Dentaire

Consultations et Visites
Consultations et Visites Généralistes
Consultations et Visites Spécialistes
Consultations et Visites Majorations
Autres sous postes Consultations et Visites

Pharmacie
Pharmacie à 15%
Pharmacie à 30%
Pharmacie à 65%
Vaccins non remboursés
Autres sous postes Pharmacie

Soins courants
Actes techniques médicaux
Auxiliaires médicaux
Analyses médicales
Radiologie
Médecine alternative
Majorations Soins courants
Autres sous postes Soins courants

Hospitalisation
Honoraires
Forfait journalier
Frais de séjour
Transport
Chambre particulière
Lit d'accompagnant
Autres sous postes Hospitalisation

Autres postes	
Cures thermales	Prothèses auditives
Maternité Forfait	Prothèses orthopédiques
Maternité Chambre particulière	Prothèses autres
Maternité Autre	Autres sous postes Autres postes
Frais d'obsèques	

Tableau 2.1 – Liste des sous-postes classés par grand poste

2.1.3 Garanties

Les garanties permettent de décrire tous les remboursements de frais de santé auxquels ont droit les bénéficiaires grâce à leur complémentaire santé. Elles sont différentes dans chaque entreprise, et au sein d'une même entreprise, tous les bénéficiaires qui ont la même affiliation ont les mêmes garanties. Comme énoncé précédemment, les garanties peuvent être exprimer de différentes manières. Voici les informations retenues :

- Société
- Grand poste
- Sous poste
- Adhésion à l'OPTAM
- Réseau optique
- Affiliation (base ou base + option)
- Type de garantie (en complément du remboursement Sécurité Sociale ou y compris remboursement Sécurité Sociale)
- Assiette de garantie 1 (%PMSS, €, %FR, %BRSS)
- Garantie 1 (valeur de la garantie)
- Assiette de garantie 2 (%PMSS, €, %FR, %BRSS)
- Garantie 2 (valeur de la garantie)
- Assiette limite (%PMSS, €, %FR, %BRSS, « par an », pour la garantie limite ou plafond)
- Garantie limite (valeur du plafond)

Les informations contenues dans « Assiette garantie 2 » et « Garantie 2 » sont utilisées uniquement si la garantie s'exprime de deux manières différentes. Par exemple, une garantie pour des prothèses auditives à hauteur de 200 %BR + 500 €.

Pour la garantie limite, une nouvelle assiette de garantie est possible. Il s'agit de l'assiette « par an ». Pour cette assiette, le montant du plafond fait référence à la fréquence et non aux frais réels, contrairement aux autres assiettes de garanties. Il représente le nombre de fois maximal où le bénéficiaire peut être remboursé dans une année. Par exemple, une garantie pour des consultations ostéopathiques à hauteur de 50 € par séance, dans la limite de 3 séances par an.

2.2 TRAITEMENT ET NETTOYAGE DES DONNEES

L'étape de traitement et nettoyage des données est une étape très importante car elle permet l'obtention d'une base de données solide et exploitable pour la modélisation de la fréquence et du coût moyen. Comme expliqué précédemment, les données ont déjà subi des premiers traitements dans le cadre des analyses de consommation. Il n'y a donc pas de données aberrantes car elles ont déjà été repérées puis traitées. La majorité des traitements a ainsi déjà été effectuée. Toutefois, quelques traitements supplémentaires ont dû être réalisés.

Avant la réalisation de ces traitements, les données ont d'abord été agrégées. Etant donné qu'elles avaient toutes le même format (noms de colonnes et type de variables identiques), il a suffi d'utiliser la fonction *rbind* pour concaténer les effectifs entre eux, les prestations entre elles et les garanties entre elles, après importation sur R. Puis les traitements suivants ont été effectués.

2.2.1 Traitements sur les effectifs

Deux traitements ont été effectués sur les effectifs.

Regroupement des catégories

Après agrégation des données, on distingue 16 catégories différentes. Certaines contiennent très peu de bénéficiaires. De plus, certaines catégories sont plus détaillées que d'autres car cela dépend du niveau de détail des études qui avaient été menées pour chaque client. Afin de diminuer le nombre de catégories et de les homogénéiser, quatre nouvelles catégories sont créées :

- Actifs
- Portables
- Retraités
- Inactifs (autres que retraités)

Regroupement des âges par tranche

Dans les effectifs, tous les bénéficiaires du régime sont présents. Il y a des assurés, des conjoints, leurs enfants et parfois leurs ascendants. Il y a des personnes avec des âges très différents. Les âges vont de 0 à 106 ans. Comme peu de personnes ont plus de 70 ans, nous avons décidé de regrouper les âges en classe de 5 ans, de 0 à 70 ans, puis de faire une classe « 70 ans et + ». Nous avons estimé que des classes de 5 ans permettaient de conserver une cohérence de consommation au sein de chaque groupe.

2.2.2 Traitements sur les prestations

Quatre traitements ont été effectués sur les prestations. Ils concernent tous la modification des libellés des sous postes.

Intégration de l'information « OPTAM »

Les garanties sur les sous postes pour lesquels on peut distinguer OPTAM / hors OPTAM ne sont pas forcément les mêmes en OPTAM et en hors OPTAM. En effet, dans le cadre du contrat responsable, la complémentaire santé peut prévoir le remboursement de certains dépassements d'honoraires, à condition qu'ils soient inférieurs d'au moins 20 %BRSS aux remboursements des actes sans dépassements d'honoraires. La distinction OPTAM / hors OPTAM se fait sur les postes suivants :

- Honoraires
- Consultations et Visites Généralistes
- Consultations et Visites Spécialistes
- Actes techniques médicaux
- Radiologie

Nous avons donc rajouté dans le libellé du sous postes l'information OPTAM / hors OPTAM afin de pouvoir leur associer les bonnes garanties. Ces sous postes seront donc de type « Actes techniques médicaux Hors OPTAM », ou bien « Radiologie OPTAM ».

Intégration de l'information « Réseau optique »

Dans un régime de frais de santé, les garanties entre les équipements optiques dans le réseau et en dehors du réseau ne sont pas toujours les mêmes. Cette différence peut se faire sur les verres et les montures. Donc pour tous les verres et toutes les montures, l'information sur l'utilisation du réseau optique a été rajoutée dans le libellé du sous poste. Ces sous postes seront donc de type « Montures Adulte Hors réseau », ou bien « Verres simples Enfant Réseau ».

Intégration de la base de remboursement pour les verres

Les garanties ne sont pas toutes les mêmes suivant le type précis de verres. Cette information n'est pas forcément disponible dans le libellé d'acte mais on peut la retrouver grâce aux bases de remboursement. Donc la base de remboursement a été incluse dans le libellé du sous poste. Ces sous postes seront ainsi de type « Verres simples Enfant Réseau 12.04 ».

Détails sur la médecine alternative

Certains régimes proposent des garanties sur de la médecine alternative. Ces garanties ne sont pas forcément les mêmes en fonctions de la spécialité, et toutes les spécialités ne sont pas toujours prises en charge. Il est donc important de distinguer les spécialités. Pour cela, lors que le sous poste s'intitule « Médecine alternative », on tiendra compte de la colonne *Spécialité Médecine alternative*. Voici la liste des spécialités :

- Acupuncture
- Chiropractie
- Diététique
- Etiopathie
- Microkinésie
- Naturopathie
- Ostéopathie
- Pédicurie
- Podologie
- Psychologie
- Psychomotricité

Finalement, après toutes les modifications, on compte maintenant 98 sous postes distincts.

2.2.3 Traitements annexes sur les prestations

D'autres traitements sont réalisés, mais ils ne concernent pas la base de données directement. Ils serviront par la suite pour les calculs.

Création d'une table de bases de remboursement, de taux de Sécurité Sociale et frais réels

A la suite des modifications effectuées sur les sous postes, une table est créée à partir des prestations. Cette table indique pour chaque sous poste la base de remboursement moyenne, le taux de remboursement de Sécurité Sociale moyen et les frais réels maximaux. Parmi les sous postes, certains actes ont la même base de remboursement et le même taux de Sécurité Sociale au sein d'un même sous poste, comme les SPR 50 par exemple. En revanche, pour des sous postes qui contiennent plusieurs actes différents à, les bases de remboursements et les taux ne sont pas toujours les mêmes. C'est le cas par exemple des Auxiliaires médicaux, parmi lesquels on peut trouver les infirmiers, les orthophonistes, les kinésithérapeutes…

Calcul du coefficient global de PSAP

Comme évoqué dans le paragraphe 2.1.2, les données utilisées pour constituer une base de données de prestations ne sont pas complètes, puisque leurs dates d'arrêté sont antérieures au 31/12/2020. Par ailleurs, les dates d'arrêtés ne sont pas identiques pour tous les clients et leurs taux de PSAP ne sont pas les mêmes. Il faut donc calculer un coefficient de PSAP global qui permettra d'estimer la partie manquante des prestations et qui servira par la suite. Voici les différents taux de PSAP, calculés à l'aide des cadences propres à chaque client, et les différents volumes de prestations pour chaque client anonymisé :

Numéro client	Date d'arrêté	Remboursements complémentaires	Taux de PSAP
1	31/01/2019	6 085 640 €	3,99 %
2	31/03/2019	18 361 373 €	2,90 %
3	28/02/2019	455 683 €	5,08 %
4	31/01/2019	815 199 €	4,31 %
5	31/01/2019	1 700 433 €	4,64 %
6	28/02/2019	780 985 €	2,37 %
7	28/02/2019	2 960 802 €	3,06 %
8	28/02/2019	1 478 113 €	2,63 %
9	30/04/2019	3 427 842 €	1,63 %
10	28/02/2019	756 254 €	3,93 %
Total		***36 822 322 €***	***3,11 %***

Tableau 2.2 – Taux de PSAP par client et global

Pour chaque client k, on note RC_k le montant total des remboursements complémentaires à la date d'arrêté, et $t_{k,PSAP}$ son taux de PSAP. Le taux de PSAP global se calcule de la manière suivante (moyenne pondérée) :

$$t_{PSAP} = \frac{\sum_{k=1}^{10} RC_k \times t_{k,PSAP}}{\sum_{k=1}^{10} RC_k} = 3{,}11\ \%$$

2.2.4 Traitement sur les garanties

Cinq traitements ont été effectués sur les garanties. Les quatre premiers concernent la modification des libellés des sous postes avec l'utilisation du réseau, l'adhésion à l'OPTAM, les bases de remboursement et les spécialités de médecine alternative. Ils sont identiques à ceux mentionnés ci-dessus, donc ils ne seront pas redétaillés, contrairement au cinquième.

Conversion de toutes les garanties en euros en complément de la Sécurité Sociale

Afin de pouvoir les comparer et les utiliser en s'affranchissant des remboursements de la Sécurité Sociale, toutes les garanties doivent être converties dans une même unité de mesure et vont être exprimées en complément de la Sécurité Sociale. L'unité choisie est l'euro. On va donc calculer le montant de remboursement maximal en euro pour un acte, sur chaque sous poste. Voici les différents calculs et traitements réalisés pour faire la conversion :

On introduit tout d'abord les notations suivantes : $g_{i,€}$, $g_{i,PMSS}$, $g_{i,BRSS}$ et $g_{i,FR}$, pour désigner des garanties exprimées respectivement en €, en %PMSS, en %BRSS et en %FR, pour le sous poste i.

- Garantie en ***€***: la garantie est laissée telle quelle.
- Garantie en ***%PMSS***: $g_{i,€} = g_{i,PMSS} \times PMSS_{2018}$, avec $PMSS_{2018} = 3\,311$ €.
- Garantie en ***%BRSS***: $g_{i,€} = g_{i,BRSS} \times \overline{BRSS_i}$, avec $\overline{BRSS_i}$ la base de remboursement de la Sécurité Sociale moyenne pour le sous poste i, issue de la table construite ci-dessus (voir paragraphe 2.2.3).
- Garantie en ***%FR***: $g_{i,€} = g_{i,FR} \times FR_{max,i}$, avec $FR_{max,i}$ les frais réels maximums pour le sous poste i, issus de la table construite ci-dessus (voir paragraphe 2.2.3). Les garanties qui remboursent un pourcentage de frais réels, sans aucune autre condition, n'ont en réalité pas de limite de remboursement. Toutefois, afin de pouvoir les utiliser plus tard au même titre que les garanties qui sont exprimées de d'autres façon, nous sommes contraints de déterminer un montant. Ce montant a été déterminé en considérant le maximum dépensé par sous poste sur la base de données des prestations.

Pour les garanties exprimées de deux manières différentes, il suffira de les convertir en euros séparément, puis de sommer les montants convertis.

On note $g'_{i,€}$ la garantie exprimée en euros en complément de la Sécurité Sociale. Pour les garanties déjà exprimées en complément de la Sécurité Sociale, il n'y a pas besoin de faire de traitement supplémentaire : $g'_{i,€} = g_{i,€}$.

Pour les garanties exprimées y compris remboursement de la Sécurité Sociale, il faut soustraire la part de remboursement moyen de la Sécurité Sociale. Pour cela, le montant de remboursement moyen $\bar{R}_{SS,i}$ est calculé avec le taux moyen de remboursement de la Sécurité Sociale $\bar{t}_{SS,i}$, de la manière suivante : $\bar{R}_{SS,i} = \bar{t}_{SS,i} \times \overline{BRSS_i}$. A partir de là, la garantie en complément de la Sécurité s'exprime comme telle : $g'_{i,€} = g_{i,€} - \bar{R}_{SS,i}$.

2.3 JOINTURE DES DONNEES

Une fois que les données sont traitées, l'idée est de les joindre entre elles afin d'avoir une seule base de données contenant toutes les informations sur les effectifs, les garanties et les prestations.

Avant de réaliser la jointure et afin de simplifier la suite de cette étude, les remboursements complémentaires au titre des options ne seront plus considérés. En effet, il est difficile de modéliser un choix d'option, et donc une antisélection, dans un modèle collectif, car le choix de chaque bénéficiaire ne peut pas être connu à l'avance. Cela est d'autant plus difficile qu'il y a plusieurs choix d'options. On considérera désormais le montant de la garantie ainsi que les remboursements au titre de la base uniquement. La colonne « Affiliation » ne sera donc plus utilisée. Cette simplification peut entraîner un léger biais dans l'estimation des fréquences et des coûts moyens car les bénéficiaires qui choisissent une option ont tendance à plus consommer que ceux qui n'en prennent pas. Ainsi, pour les bénéficiaires ayant choisi l'option, la garantie de la base leur sera attribuée. Les remboursements de l'option ne seront, certes, pas pris en compte, mais leurs remboursements au titre de la base peuvent être probablement dans certains cas surestimés. Cette approximation nous paraît être toutefois correcte car seulement 11% des bénéficiaires ont adhéré à une option et que les options ne concernent pas tous les sous postes. Par ailleurs, dans le cas où ces bénéficiaires n'auraient pas pu choisir une option, ils auraient quand même consommé plus que les bénéficiaires n'ayant pas choisi l'option.

Pour réaliser la jointure, le but est de partir des effectifs et de dupliquer les effectifs autant de fois qu'il y a de sous postes différents. Les effectifs vont donc être dupliqués en 98 fois, et à chaque fois on va inscrire le nom d'un sous poste différent dans la colonne *Sous poste*, ainsi que le grand poste associé dans la colonne *Grand poste*. Afin de connaître la consommation de chaque bénéficiaire sur chaque sous poste, on va relier la table que l'on vient de créer aux prestations grâce à une clé unique pour chaque ligne. Cela va permettre de récupérer le nombre d'actes consommés et le total des frais réels et des remboursements par bénéficiaire et par sous poste. De même, afin de connaître la garantie associée à chaque sous poste pour chaque bénéficiaire, on va également créer une clé pour relier les garanties aux prestations. Voici les éléments utilisés pour les clés :

- Liaison **Effectifs – Prestations** : la clé qui va permettre de lier les deux jeux de données utilise l'identifiant bénéficiaire et la société. La société va permettre de résoudre le problème dans le cas où deux bénéficiaires peuvent avoir un même identifiant en étant dans deux sociétés différentes.
- Liaison avec les **Garanties** : une fois que les effectifs et les prestations sont liés, la clé contenant la société, le grand poste et le sous poste va permettre de les lier aux garanties.

Notre nouvelle base de données contient alors les informations suivantes :

- Identifiant du bénéficiaire
- Identifiant de l'assuré auquel le bénéficiaire est rattaché
- Société
- Catégorie
- CSP
- Type de bénéficiaire
- Tranche d'âge

- Temps de présence (ou temps d'exposition)
- Structure familiale
- Grand poste
- Sous poste (contenant également les informations sur l'adhésion à l'OPTAM, l'utilisation du réseau optique, la spécialité de médecine alternative, les bases de remboursement des verres)
- Nombre d'acte total par bénéficiaire par sous poste
- Montant total des frais réels par bénéficiaire et par sous poste
- Montant total des remboursements complémentaires par bénéficiaire et par sous poste
- Garantie maximale proposée

À la suite de cette jointure, d'autres traitements ont été réalisés sur la nouvelle base de données. Tout d'abord, certains bénéficiaires n'ont pas pu trouver de correspondance avec les prestations. Cela est tout à fait normal car ce sont en réalité les bénéficiaires qui n'ont pas consommé sur un ou plusieurs sous postes, ou qui n'ont pas consommé du tout. On leur affecte donc la valeur de 0 pour le nombre d'actes, les frais réels et les remboursements complémentaires.

A contrario, certains bénéficiaires présents dans les prestations n'ont pas trouvé de correspondances dans les effectifs. Il s'agit ici d'une anomalie des données. Cette anomalie concerne 298 bénéficiaires, soit 0,19% de l'effectif total. On choisit de les supprimer pour la suite.

Ensuite, certains bénéficiaires ont des montant totaux de frais réels et/ou de remboursements négatifs. Cela représente 134 bénéficiaires, soit 0,09% de l'effectif total. On choisit également de supprimer ces lignes pour la suite.

Les lignes qui ont des garanties nulles sont aussi supprimées. C'est le cas de certains sous postes chez certains clients.

Par ailleurs, deux colonnes supplémentaires ont été ajoutées. Les calculs sont à chaque fois réalisés par bénéficiaire et par sous postes :

- Frais réels moyens : il s'agit des frais réels divisés par le nombre d'actes. Si le nombre d'actes est nul, alors on considérera que les frais réels moyens le seront aussi.
- Remboursements moyens : il s'agit de la somme des remboursements complémentaires divisée par le nombre d'actes. De même que pour les frais réels moyens, si le nombre d'actes est nul, alors on considérera que le remboursement moyen sera nul.

Les colonnes des montants totaux de frais réels et de remboursements sont gardées pour l'élaboration des statistiques descriptives sur les données mais elles ne seront pas utilisées pour la modélisation.

2.4 STATISTIQUES DESCRIPTIVES DES DONNEES

En actuariat, l'analyse descriptive des données est très importante avant une tarification. Elle permet de mieux comprendre leur constitution ainsi que le comportement des bénéficiaires. Cette étape essentielle aide également au choix des variables explicatives.

2.4.1 Statistiques sur les effectifs

Dans un premier temps, des statistiques décrivant les effectifs uniquement seront présentées. Plusieurs variables descriptives seront détaillées : le type de bénéficiaire, la CSP, la catégorie, la tranche d'âge et la structure familiale.

Statistiques par Âge

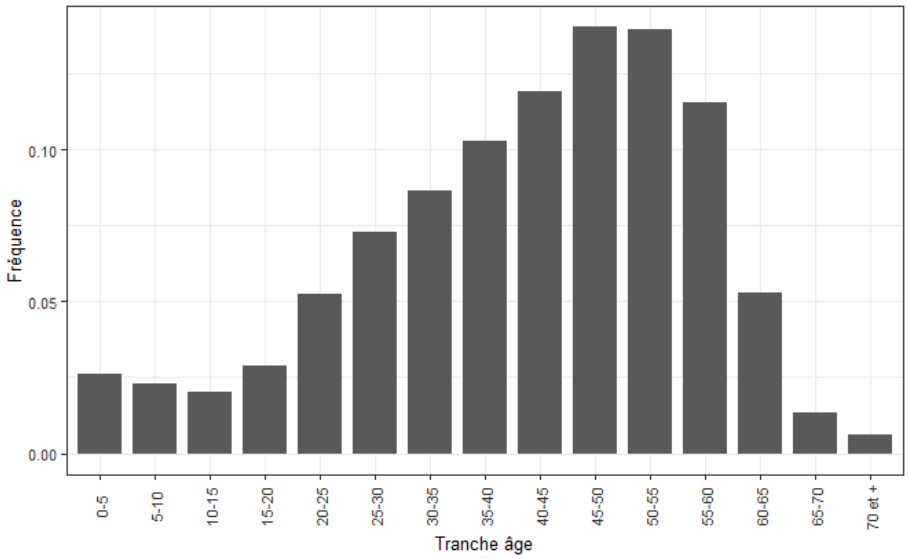


Figure 2.1 – Répartition des bénéficiaires par tranche d'âge

Ce graphique montre la répartition (au prorata du temps d'exposition) de tous les bénéficiaires en fonction de leur tranche d'âge. L'âge moyen de la population est de 42 ans. On peut voir que les deux tranches les plus représentées sont les « 45-50 ans » et les « 50-55 ans ». Elles représentent à elles seules un peu moins de 30% de la population. La population est donc plutôt âgée. Toutefois, moins de 3% des bénéficiaires ont plus de 65% ans.

Statistiques par Type de bénéficiaire

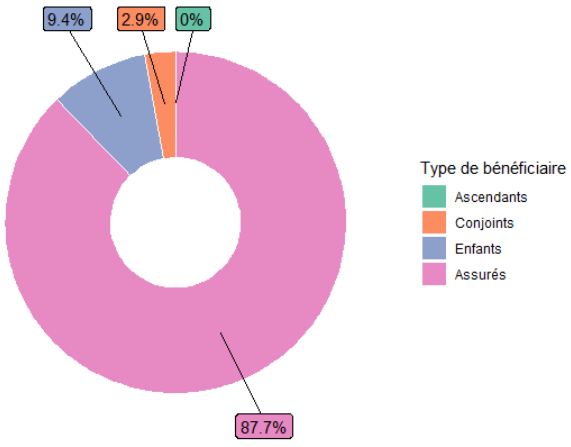


Figure 2.2 – Répartition des bénéficiaires par type de bénéficiaire

Ce diagramme représente la répartition des bénéficiaires selon le type de bénéficiaire. Il révèle que 88% des bénéficiaires sont des assurés, soit une grande majorité d'entre eux. A contrario, les ascendants sont très peu nombreux.

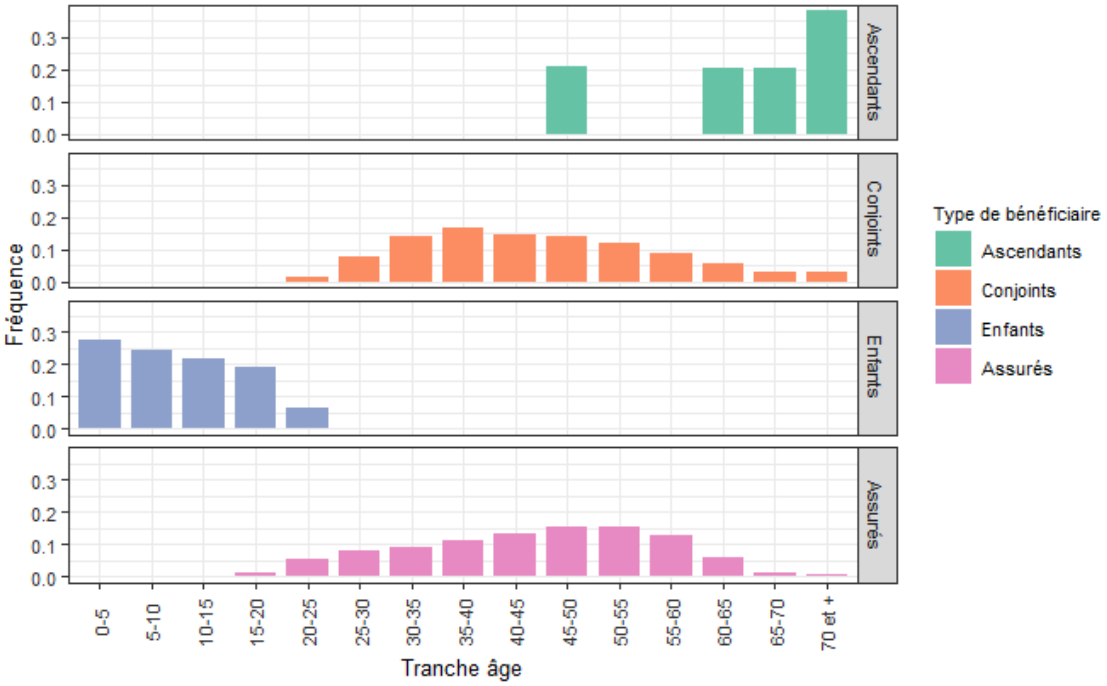


Figure 2.3 – Répartition des bénéficiaires par tranche d'âge en fonction du type de bénéficiaire

Ces graphiques montrent la répartition des bénéficiaires par tranche d'âge en fonction de leur type. De même que sur l'ensemble de la population, chez les assurés, les tranches d'âge les plus représentées sont les « 45-50 ans » et les « 50-55 ans ». Cela vient du fait que la grande majorité des bénéficiaires sont des assurés, donc ce sont eux qui donnent la tendance. On peut également voir que globalement les conjoints sont plus jeunes que les assurés. Par ailleurs, on remarque de plus de la moitié des enfants ont moins de 10 ans ; ils sont donc assez jeunes.

Statistiques par CSP

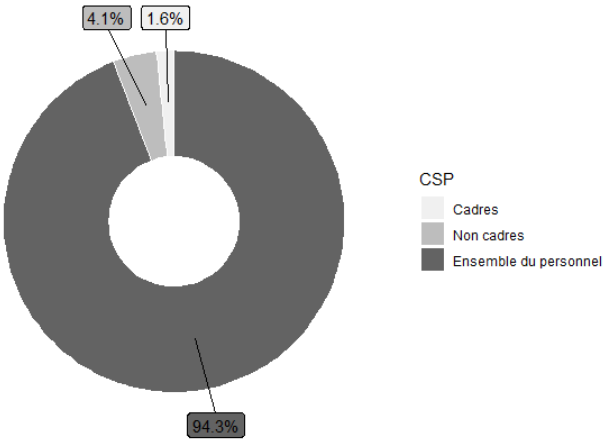


Figure 2.4 – Répartition des bénéficiaires par CSP

Cette répartition des bénéficiaires par CSP nous indique que la quasi-totalité de la population (94%) est ensemble du personnel. Cette CSP regroupe aussi bien des cadres que des non cadres. Elle ne nous apporte donc pas d'information. Cette variable ne sera donc pas utilisée dans la suite de l'étude.

Statistiques par Catégorie

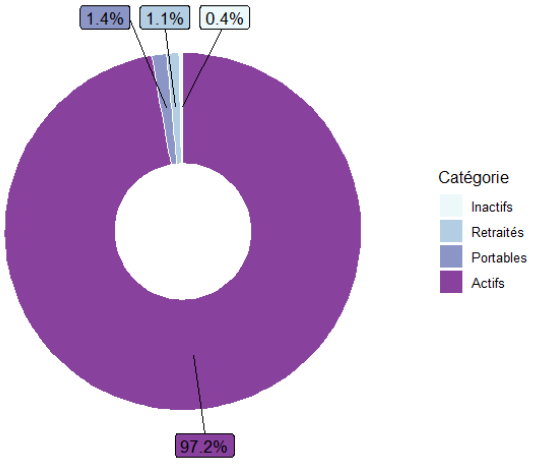


Figure 2.5 – Répartition des bénéficiaires par catégorie

Dans ce diagramme circulaire sur la répartition des bénéficiaires par catégorie, on peut voir qu'environ 97% de la population est active. Les 3% restants se répartissent dans les catégories suivantes : inactifs, portables et retraités.

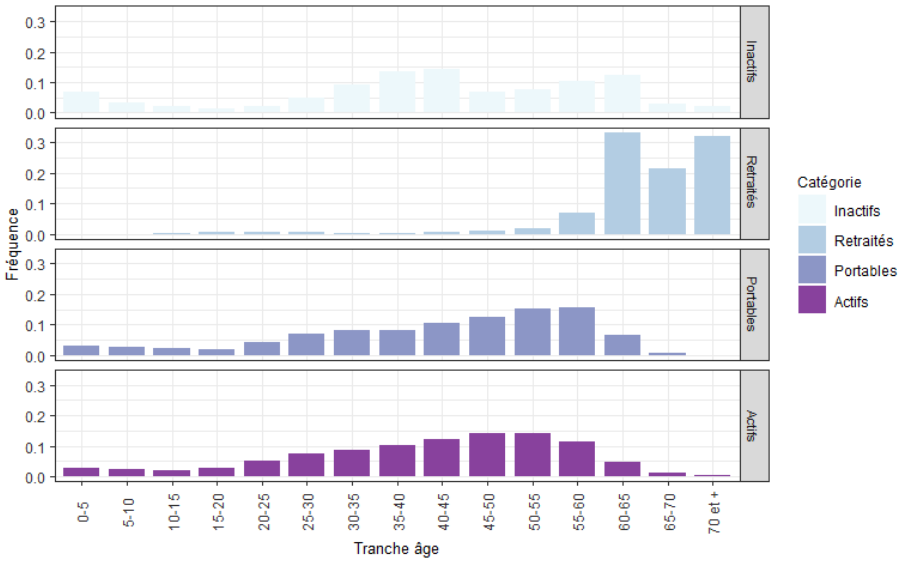


Figure 2.6 – Répartition des bénéficiaires par tranche d'âge en fonction de la catégorie

Ces graphiques montrent la répartition des bénéficiaires par tranche d'âge en fonction de leur catégorie. De même que sur l'ensemble de la population, chez les actifs, les tranches d'âge les plus représentées sont les « 45-50 ans » et les « 50-55 ans ». Cela vient du fait que la grande majorité des bénéficiaires sont des actifs, donc ce sont eux qui donnent la tendance. La répartition des portables s'apparente également à celle des actifs. En revanche, les retraités se concentrent sur les tranches d'âge de plus de 60 ans, ce qui semble tout à fait logique (les retraités dans les tranches plus jeunes correspondent aux enfants d'assurés retraités). On peut également remarquer que la répartition par tranche d'âge des inactifs semble assez hétérogène.

Statistiques par Structure familiale

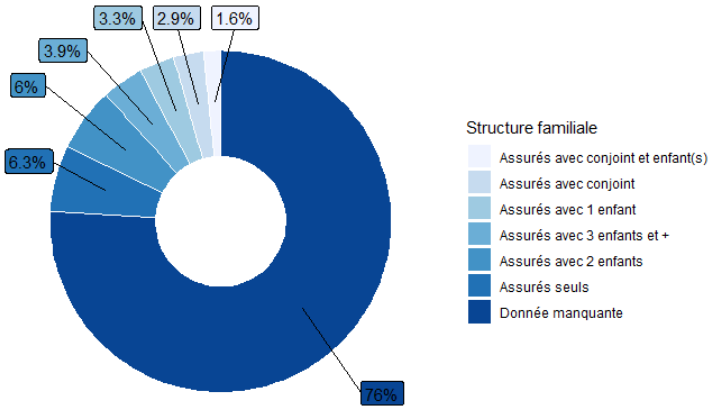


Figure 2.7 – Répartition des bénéficiaires par structure familiale

Ce diagramme montre la répartition des assurés en fonction de leur structure familiale. Il révèle que pour les trois quarts des assurés l'information est manquante. La variable « Structure familiale » ne peut donc pas être utilisée par la suite. Cependant, ce n'est pas un problème car intuitivement, on peut penser cette variable n'a pas d'influence sur la fréquence de consommation de l'assuré, ou sur les frais réels.

2.4.2 Statistiques sur les prestations

Dans un second temps, les statistiques décrivant les prestations, et plus exactement les quantités d'actes et les frais réels, vont être présentées. Elles vont faire appel à plusieurs variables descriptives, issues des effectifs, des prestations et des garanties. Il s'agit des variables suivantes : le type de bénéficiaire, la catégorie, la tranche d'âge, le grand poste, le sous poste et le montant maximal de la garantie.

Statistique par Âge

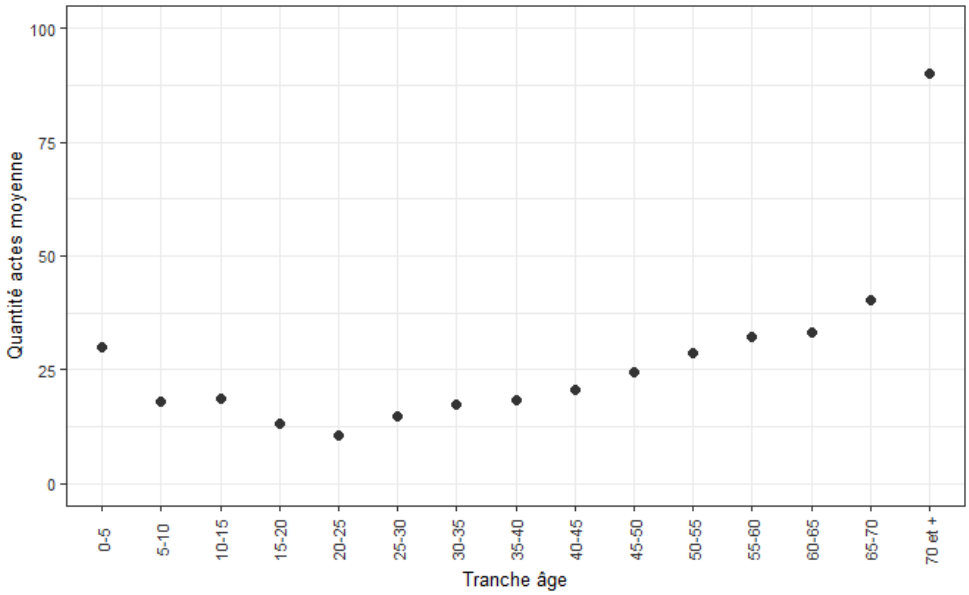


Figure 2.8 – Quantité d'actes moyenne en fonction de la tranche d'âge

Le graphique ci-dessus représente la quantité d'acte moyenne (tous sous postes confondus) par bénéficiaire en fonction de la tranche d'âge du bénéficiaire. A partir de 20-25 ans, le nombre d'actes moyen croît avec l'âge. Cela vient du fait que plus un

individu est âgé, plus il aura besoin de soins et donc plus il va consommer des prestations de santé. On peut également remarquer un nombre d'actes moyens un peu plus élevé chez les très jeunes. Cela est dû à la fois à la fragilité des bébés qui ont eux aussi besoins de plus de soins, mais également aux soins dont les enfants et adolescents en pleine croissance ont besoin : orthodontie, vaccin…

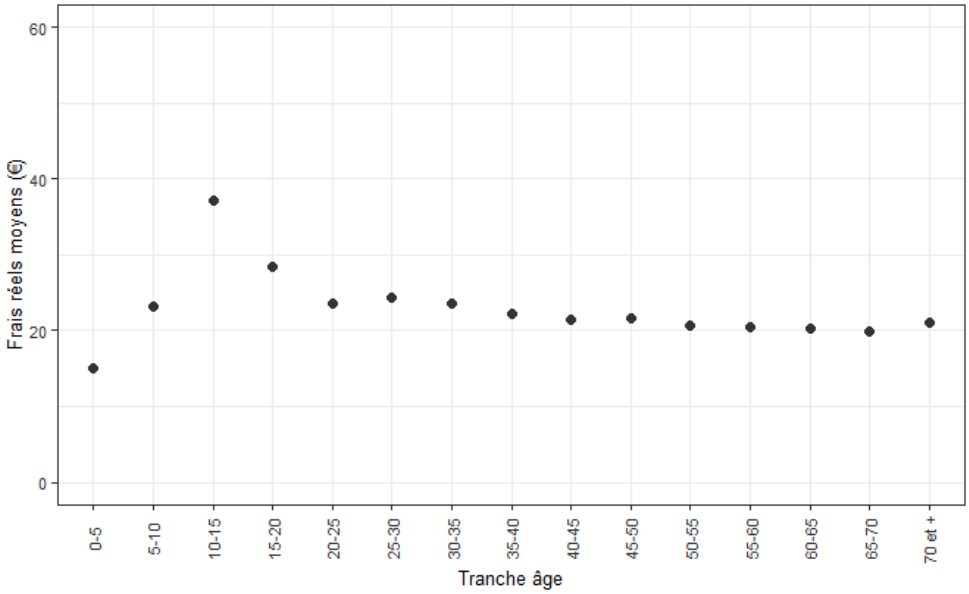


Figure 2.9 – Frais réels moyens en fonction de la tranche d'âge

Sur ce graphique représentant les frais réels moyens par acte (tous sous postes confondus) en fonction de la tranche d'âge des bénéficiaires, on peut voir un comportement totalement différent des quantités d'actes moyennes. En effet, on constate une stabilité sur les coûts moyens, exceptés chez les enfants et adolescents. Cela est en grande partie dû à l'orthodontie qui touchent un grand nombre d'enfants et adolescents et dont les coûts sont particulièrement élevés. Il serait donc intéressant de voir plus en détail ces graphiques, notamment en distinguant les grands postes, voire les sous postes.

Statistiques par Type de bénéficiaire

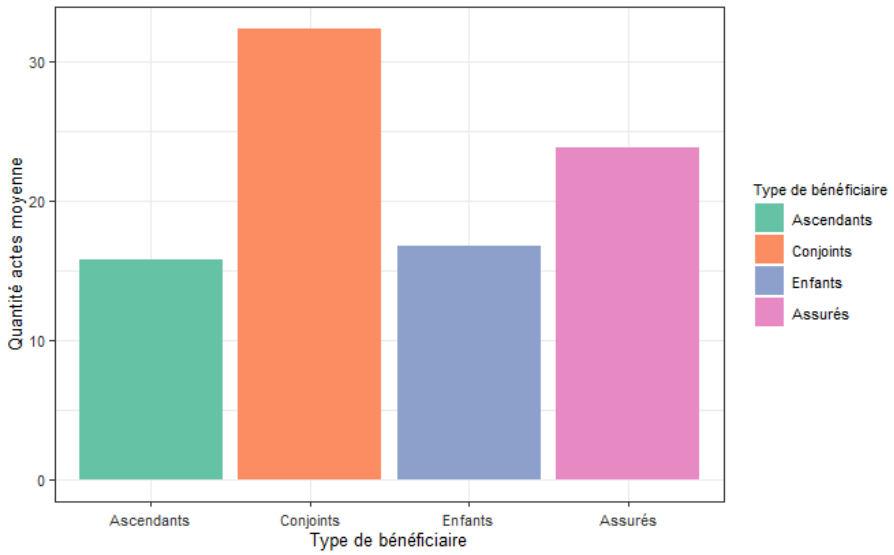


Figure 2.10 – Quantités d'acte moyenne en fonction du type de bénéficiaire

L'histogramme ci-dessus représente la quantité d'acte moyenne par bénéficiaire en fonction du type de bénéficiaire. Avant toute chose, on peut remarquer que les ascendants ont l'air d'avoir une consommation similaire à celle des enfants, et donc assez faible. Cette observation semble peu intuitive au vu des âges élevés des ascendants. Or, comme on a pu le remarquer précédemment, il y a très peu d'ascendants et donc leur consommation est plus volatile. Enfin, ce graphique montre sans surprise que les adultes (assurés et conjoints) consomment plus d'actes que les enfants. En effet, ils consomment en moyenne autour de 25-30 actes, tandis que les enfants consomment en moyenne une quinzaine d'actes. Par ailleurs, chez les adultes, les conjoints consomment plus d'actes que les assurés.

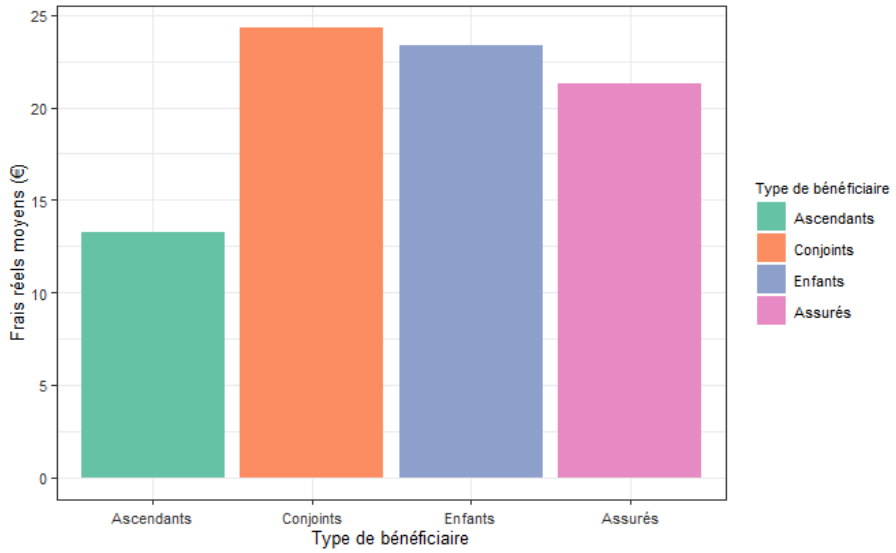


Figure 2.11 – Frais réels moyens en fonction du type de bénéficiaire

Ce graphique illustre les frais réels moyens par acte (tous sous postes confondus) chez les différents types de bénéficiaires. Une fois de plus, on constate que les ascendants se distinguent du reste des bénéficiaires par de faibles coûts moyens, ce qui est dû à la volatilité de leur consommation. Par ailleurs, il y a très peu d'écarts sur les frais réels des assurés, conjoints et enfants. En effet, ils sont tous compris entre 21 et 24 €. Cette variable semble donc avoir moins d'influence sur les frais réels moyens.

Statistiques par Catégorie

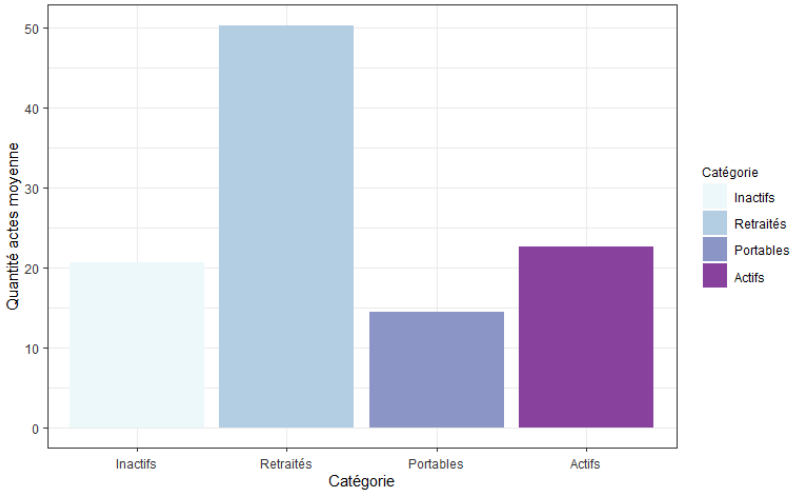


Figure 2.12 – Quantité d'actes moyenne en fonction de la catégorie

Sur ce graphique montrant les quantités d'actes moyennes par bénéficiaire pour chaque catégorie, on peut voir que les retraités se détachent largement des autres catégories. Leur quantité d'acte moyenne est environ deux fois plus élevée : 50 actes environ en moyenne contre 20/25 actes en moyenne pour les autres catégories. Cela peut s'expliquer par l'âge assez élevé des retraités.

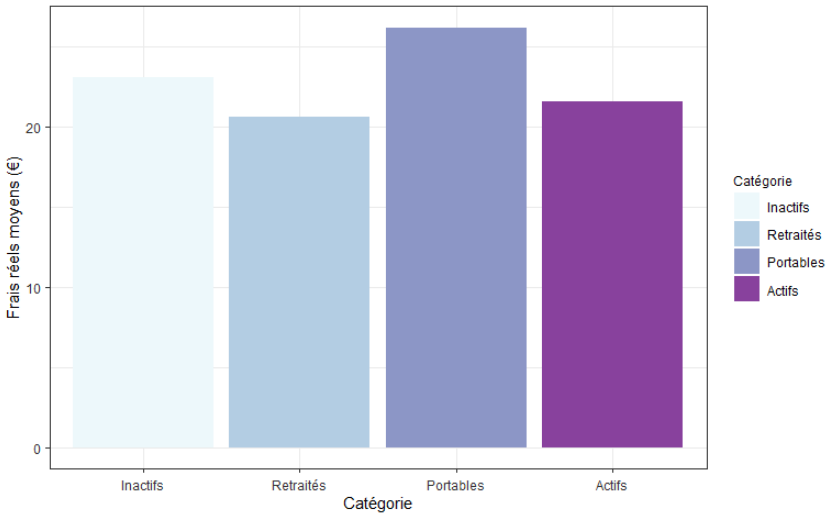


Figure 2.13 – Frais réels moyens en fonction de la catégorie

Ce graphique illustre les frais réels moyens par acte chez les différentes catégories. On constate qu'il y a très peu d'écarts sur les frais réels moyens, contrairement aux quantités d'actes moyennes. De même que pour le type de bénéficiaire, cette variable semble donc avoir moins d'influence sur les frais réels moyens.

Statistiques par Grand poste

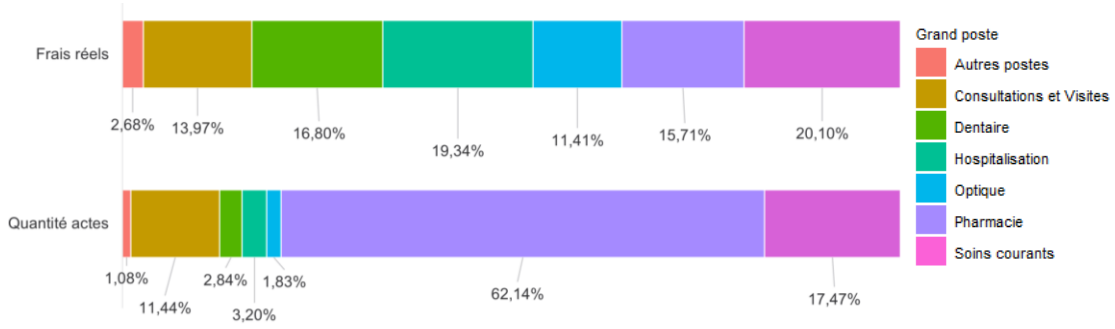


Figure 2.14 – Répartition des frais réels et des quantités d'actes par grand poste

Ces graphiques représentent les répartitions des frais réels et des quantités d'actes par grand poste. Les grands postes ont des comportements très différents. Les frais réels semblent répartis de manière à peu près égale (sauf pour les Autres postes qui sont moins représentés). En revanche, 62% des actes sont des actes de Pharmacie. Cette sur-représentation des actes de Pharmacie peut s'expliquer par le fait que ce sont des actes très courants qui concernent de nombreux bénéficiaires, et qu'il n'est pas rare d'acheter plusieurs médicaments à la fois lors d'un passage à la pharmacie.

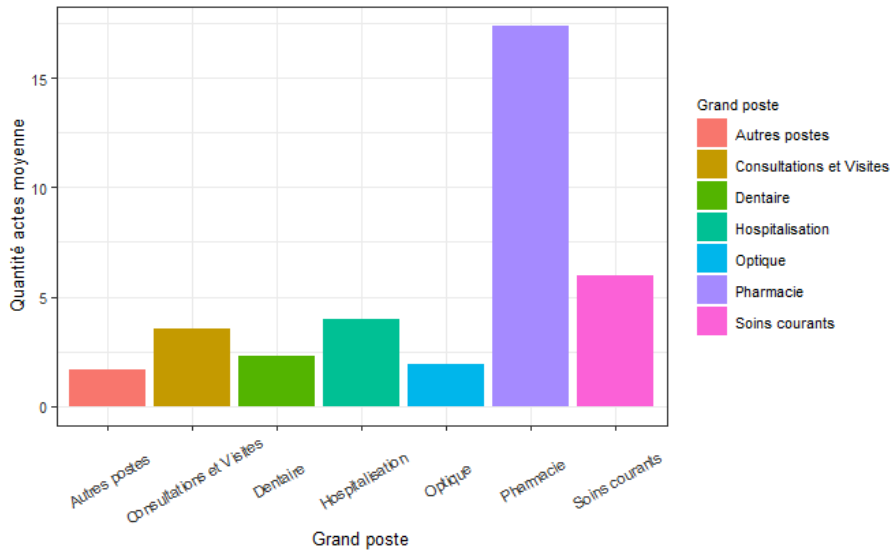


Figure 2.15 – Quantité d'actes moyenne en fonction du grand poste

Sur ce graphique qui représente les quantités d'actes moyennes par grand poste par bénéficiaire, on retrouve bien la fréquence élevée en Pharmacie (environ 20 actes par bénéficiaires) qui se démarque des autres postes (moins de 7 actes par bénéficiaire). Les autres postes ont des quantités d'actes moyennes bien plus faibles, mais qui sont toutefois très différentes les unes des autres. Cela illustre bien le fait que le comportement des bénéficiaires est très différent d'un grand poste à l'autre.

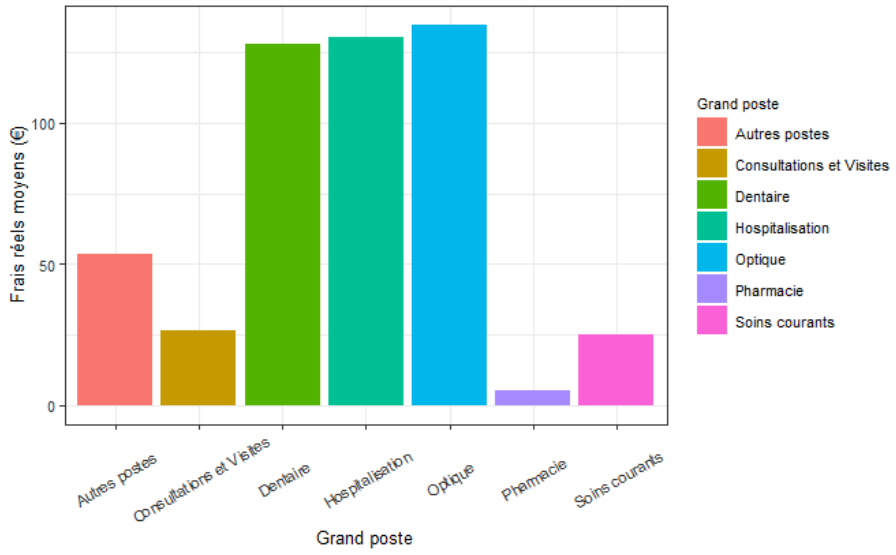


Figure 2.16 – Frais réels moyens en fonction du grand poste

Ce graphique illustre les frais réels moyens par grand poste par bénéficiaire. Trois grands postes se distingues des autres. Il s'agit du Dentaire, de l'Hospitalisation et de l'Optique. Leurs frais réels moyens sont très élevés : environ 150 € par acte. En effet, dans ces grands postes, on sait que certains sous postes ont des coûts particulièrement élevés. C'est le cas notamment des prothèses dentaires, des lunettes (verres et montures), de certains honoraires d'hospitalisation… On distingue ensuite la Pharmacie, dont les frais réels moyens sont très faibles (de l'ordre de 5 € par acte). Les trois autres grands postes se situent plutôt en milieu de tableau avec des frais réels allant de 25 à 50 € par acte.

Statistiques par Âge et par Grand poste

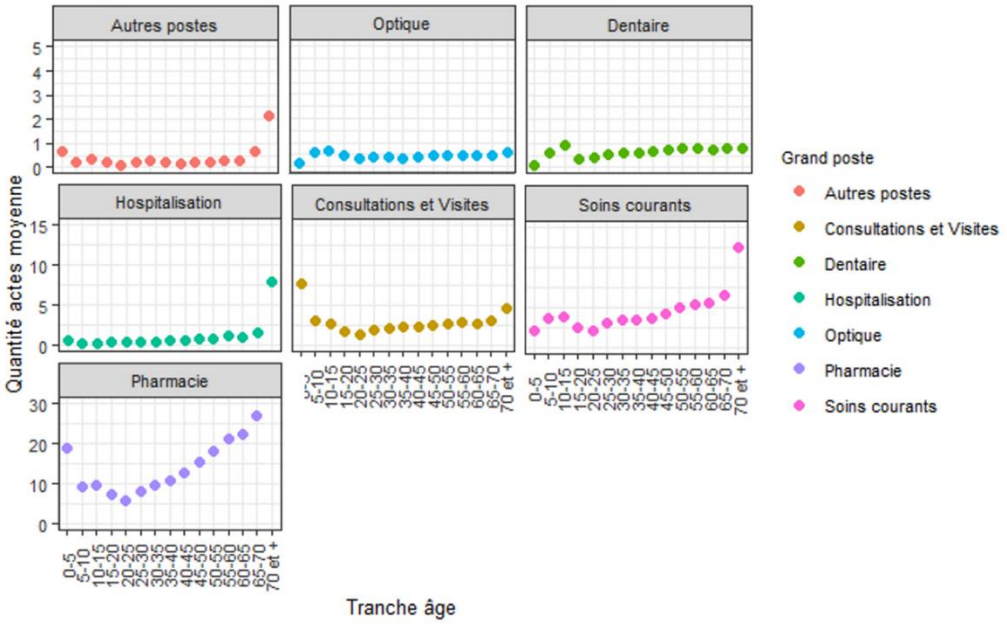


Figure 2.17 – Quantité d'actes moyenne par grand poste en fonction de la tranche d'âge

Ces graphiques montrent les différentes quantités d'actes moyennes par bénéficiaire, par tranche d'âge et par grand poste. Les échelles de quantités sont différentes sur chaque ligne. On peut voir des comportements des bénéficiaires très différents en fonction des grands postes. En effet, les quantités d'actes moyennes semblent augmenter avec l'âge en Autres postes, en Hospitalisation, en Soins courants et en Pharmacie. On retrouve ici le fait que plus les individus sont âgés, plus ils nécessitent des soins. En Optique, la consommation semble rester assez stable avec l'âge. En Dentaire, on constate un pic à l'adolescence (lié à l'orthodontie) ainsi qu'une très légère augmentation des quantités d'actes moyennes en fonction de l'âge. Et en Consultations et Visites, la fréquence semble être plus élevée lorsque les bénéficiaires sont jeunes que lorsqu'ils sont âgés.

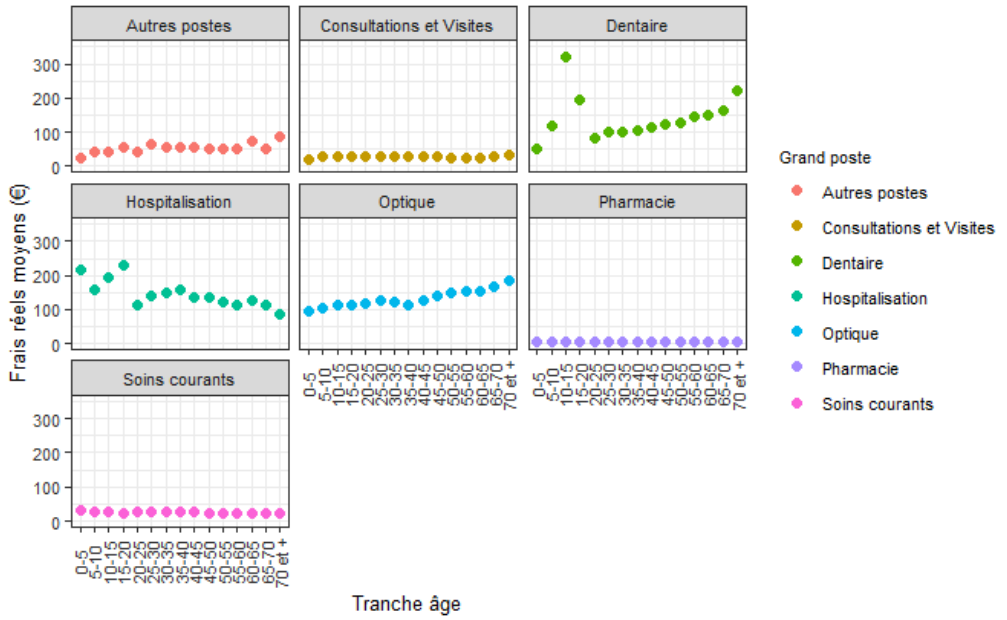


Figure 2.18 – Frais réels moyens par grand poste en fonction de la tranche d'âge

De même que pour les quantités d'actes moyennes, sur ces graphiques représentant les frais réels moyens par tranche d'âge et par grand poste, on peut voir que les comportements des bénéficiaires sont très variés en fonction des grands postes. Les coûts moyens des Consultations et Visites, de Pharmacie et des Soins courants semblent assez stables quel que soit l'âge. Ils sont plus volatiles en Hospitalisation. En Optique et en Dentaire, on note une augmentation des frais réels moyens avec l'âge. Cela signifie que plus les bénéficiaires sont âgés, plus ils ont besoin de soins conséquents. On remarque également un pic à l'adolescence en Dentaire. Enfin, on peut voir une légère augmentation des coûts moyens des Autres postes avec l'âge.

Statistiques par Âge et Sous poste

Les mêmes graphiques que ceux présentés ci-dessus dans le paragraphe « Statistiques par Âge et par Grand poste » sont présentés en annexe 5, non plus par grand poste mais par sous poste. Ils offrent encore plus de détails. Etant donné qu'il y a de nombreux sous postes, on se focalisera sur deux sous postes, qui seront analysés plus particulièrement tout au long de ce mémoire. Il s'agit des SPR 50 (couronnes dentaires) et des Consultations et Visites Spécialistes OPTAM.

- *SPR 50*

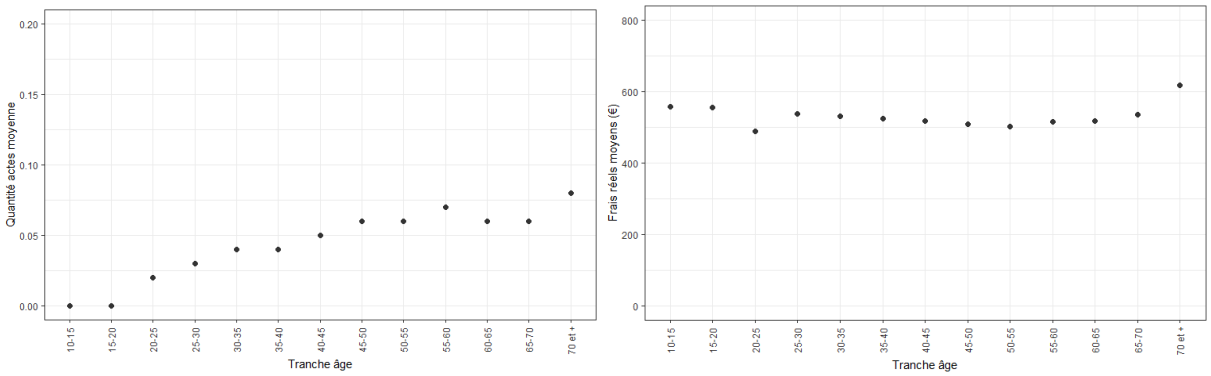


Figure 2.19 – Quantité d'actes moyenne (à gauche) et frais réels moyens (à droite) en fonction de la tranche d'âge pour le sous poste SPR 50

Le graphique de gauche illustre le nombre moyen de couronnes par bénéficiaire, en fonction de la tanche d'âge. Il montre que les bénéficiaires plus âgés consomment plus de couronnes que les plus jeunes, ce qui semble assez logique. En revanche, le graphique de droite, montrant les frais réels moyens par acte selon les tranches d'âge, nous montre que l'âge semble avoir peu d'importance sur le coût de l'acte (excepté pour les plus âgés).

- *Consultations et Visites Spécialistes OPTAM*

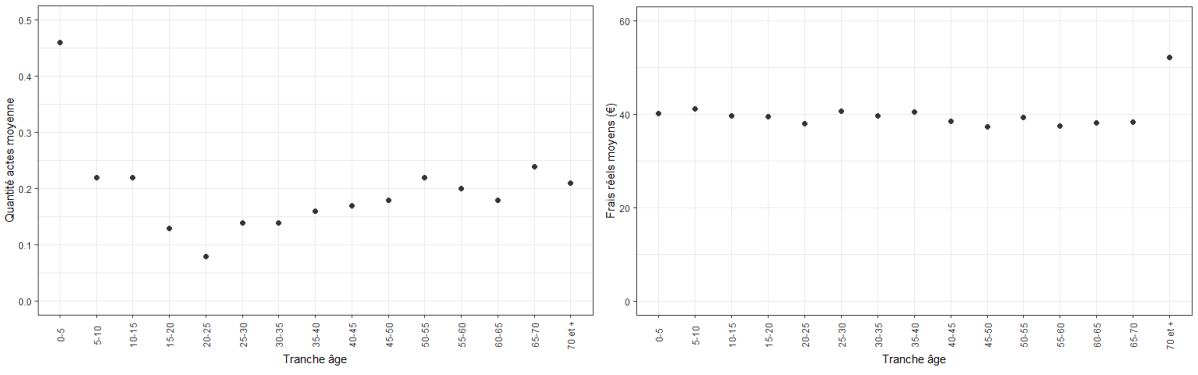


Figure 2.20 – Quantité d'actes moyenne (à gauche) et frais réels moyens (à droite) en fonction de la tranche d'âge pour le sous poste Consultations et Visites Spécialistes OPTAM

D'après le graphique de gauche, on peut voir que les bénéficiaires ayant réalisé le plus de consultations chez les spécialistes sont les plus jeunes et les plus âgés : ce sont ceux qui sont les plus fragiles. De même que pour les couronnes dentaires, le graphique de droite illustre que le fait que les frais réels dépendent peu de l'âge du bénéficiaire. Ils sont légèrement plus élevés chez les bénéficiaires de 70 ans et plus.

Statistiques par Niveau de garantie et par Sous poste

Tout comme les statistiques par âge et par sous poste, le nombre de graphique avec les niveaux de garantie est trop élevé, à cause du grand nombre de sous postes. Nous ne présenterons ici que ceux sur les SPR 50 et les Consultations et Visites Spécialistes OPTAM.

- *SPR 50*

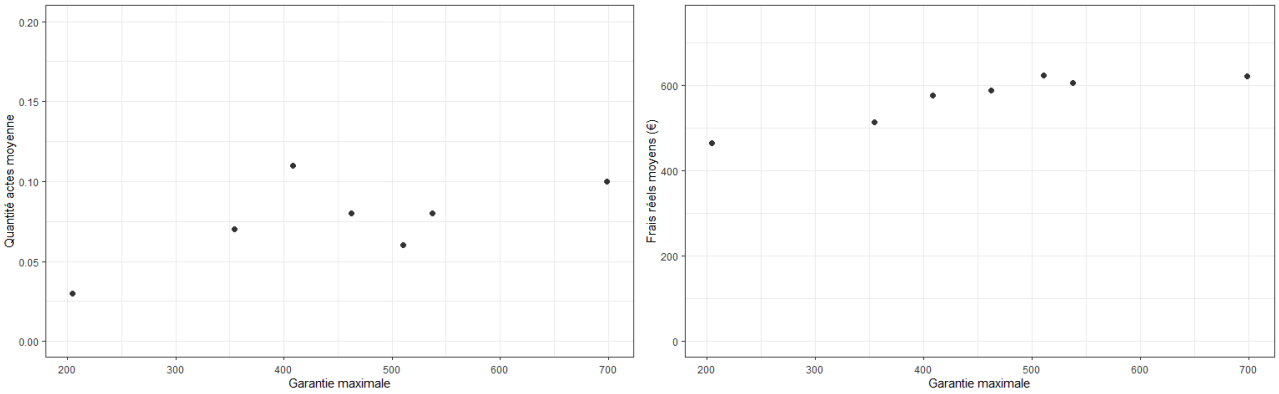


Figure 2.21 – Quantité d'actes moyenne (à gauche) et frais réels moyens (à droite) en fonction du niveau de garantie pour le sous poste SPR 50

Le graphique de gauche montre le nombre moyen de couronnes par bénéficiaire en fonction du niveau de garantie. Celui-ci semble augmenter quand le montant de la garantie augmente, mais avec une certaine volatilité. Le graphique de droite représente quant à lui les frais réels moyens des couronnes en fonction du niveau de garantie. Il illustre clairement que le montant de la garantie a un impact sur le coût d'un acte : le coût d'un acte croit avec le niveau de garantie.

- *Consultations et Visites Spécialistes OPTAM*

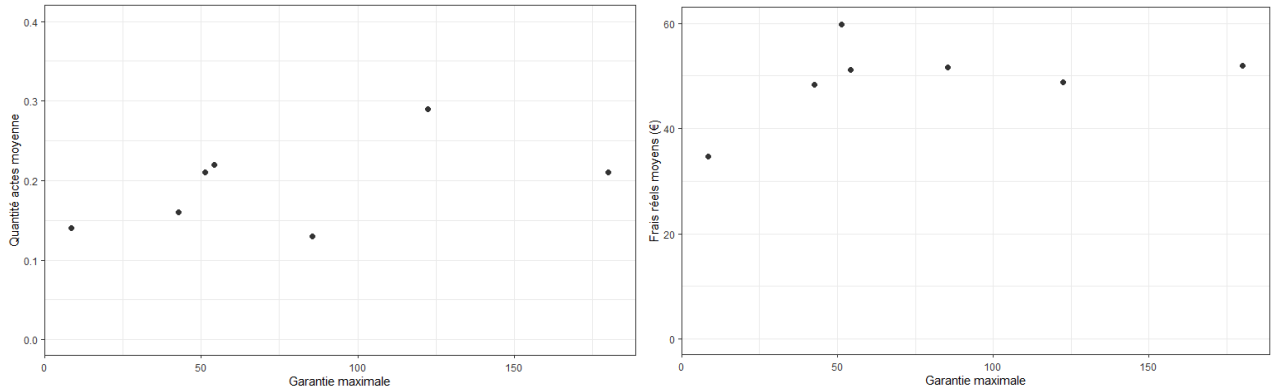


Figure 2.22 – Quantité d'actes moyenne (à gauche) et frais réels moyens (à droite) en fonction du niveau de garantie pour le sous poste Consultations et Visites Spécialistes OPTAM

Sur le graphique de gauche, de la même manière que pour les couronnes, on peut voir que le nombre de consultations chez les spécialistes a tendance à augmenter quand la garantie augmente, mais avec une certaine volatilité. Le graphique de droite nous

montre que les frais réels augmentent également avec le niveau de garantie, mais qu'à partir d'un certain palier (environ 50€), ils restent assez stables. Ainsi, d'un sous poste à l'autre, le comportement des bénéficiaires en fonction du niveau de garantie est totalement différent.

2.4.3 Corrélation entre les variables

Dans cette partie, nous allons étudier la corrélation entre les variables. Nous séparerons les variables qualitatives des variables quantitatives.

Corrélation entre les variables qualitatives

La corrélation entre les variables qualitatives est étudiée grâce au test du χ^2, dont l'aspect théorique est rappelé en annexe 6. Ce test permet de vérifier l'absence de dépendance entre les variables. Il est effectué sur les effectifs pour chaque couple de variables quantitatives (X, Y). Les variables qualitatives sont les suivantes :

- Catégorie
- Tranche d'âge
- Type de bénéficiaire

Avant d'effectuer le test, il faut vérifier que chaque couple de modalités de variables (x_i, y_j) contient au moins 5 valeurs, ce qui est bien le cas ici.

L'hypothèse nulle de ce test est la suivante : X et Y sont indépendantes. On définit également le seuil de significativité $\alpha = 0{,}05$.

Intuitivement, on pourrait penser que ces variables sont dépendantes. Par exemple, entre la tranche d'âge et le type de bénéficiaire : les enfants ont en général moins de 20 ans et les adultes (assurés et conjoints) ont pour la plupart plus de 20 ans. Ou bien entre la catégorie et la tranche d'âge : les retraités ont souvent plus de 60 ans, les actifs eux ont en général moins de 60 ans.

Les *p-values* renvoyées lors du test du χ^2 sont les suivantes :

Couple de variables	p-value
Catégorie - Tranche d'âge	< 2,2 e-16
Catégorie - Type de bénéficiaire	1,3 e-15
Tranche d'âge - Type de bénéficiaire	< 2,2 e-16

Tableau 2.3 – p-values renvoyées pour chaque couple de variables

Les *p-values* sont toutes bien inférieures à 0,05. On peut donc rejeter l'hypothèse d'indépendance entre les variables avec un risque d'erreur de moins de 5%. Cela confirme bien notre intuition. Il ne faudra donc par la suite garder qu'une seule de ces trois variables ; il s'agira de la tranche d'âge qui est plus précise et plus pertinente.

Corrélation entre les variables quantitatives

La corrélation entre les variables quantitatives est quant à elle étudiée grâce à des matrices de corrélation. Elles permettent de quantifier les dépendances entre les variables grâce aux coefficients de corrélation de Pearson, définis en annexe 7. Les variables quantitatives sont les suivantes :

- Âge
- Temps de présence
- Montant maximal remboursé par la garantie

- Quantité d'actes
- Frais réels moyens

L'âge ne sera pas utilisé par la suite (puisqu'il est évidement corrélé avec la tranche d'âge), mais il permet d'étudier les corrélations des autres variables quantitatives avec la tranche d'âges, qui est une variable qualitative.

L'objectif est de vérifier que les quantités d'actes sont bien indépendantes des frais réels (comme évoqué dans la partie 1.4.), et que les potentielles variables explicatives sont également indépendantes entre elles.

Etant donné qu'un même montant de garantie ne représente pas le même niveau de couverture pour des sous postes différents, les matrices de corrélations sont réalisées par sous postes. On va analyser plus précisément les matrices de corrélation pour les SPR 50 et pour les Consultations et Visites Spécialistes OPTAM.

Les coefficients des matrices sont calculés à partir des lignes à lignes de prestations sur chaque sous poste, pour chaque couple de variables. Ils sont compris entre -1 et 1. Un coefficient proche de 1 en valeur absolue indiquera une forte corrélation entre les deux variables, tandis qu'un coefficient proche de 0 révèlera plutôt l'indépendance des variables. En théorie, il n'existe pas de valeurs spécifiques limite des coefficients de corrélation, permettant de déterminer si la corrélation entre les variables est faible ou forte. Cela dépend plutôt du contexte et des objectifs. [15] On considérera qu'en dessous de 0,5 (en valeur absolue), la corrélation entre les variables est suffisamment faible pour supposer qu'elles sont indépendantes.

Les représentations graphiques des matrices sont tracées grâce à la fonction *cor_plot*. La taille des disques et l'intensité de leur couleur donne une indication sur la force de la corrélation : plus le disque est large et plus la couleur est intense, plus la corrélation est forte. La couleur du disque (bleue ou rouge) donne le sens de la corrélation (positive ou négative).

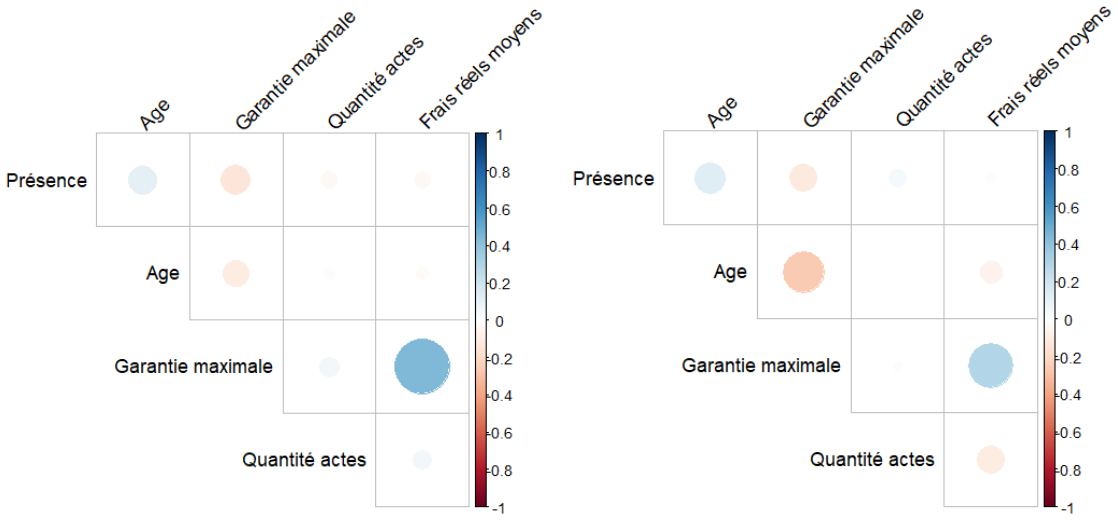


Figure 2.23 – Matrices de corrélation des SPR 50 (à gauche) et des Consultations et Visites Spécialistes OPTAM (à droite)

La matrice de gauche est celle des SPR 50 et celle de droite, la matrice des Consultations et Visites Spécialistes OPTAM. On peut observer certaines similitudes entre les deux matrices. Tous les coefficients de corrélation semblent compris entre -0,5 et 0,5, ce qui nous permet de conclure que les quantités d'actes sont bien indépendantes des frais réels. Cela est d'autant plus vrai que les coefficients de

Pearson sont particulièrement faibles entre les quantités d'actes et les frais réels moyens : 0,05 pour les couronnes dentaires et -0,1 pour les consultations spécialistes. On en conclue également que l'âge, le temps de présence et le montant de la garantie maximal sont également indépendants.

On peut toutefois remarquer que le coefficient de corrélation entre la garantie et les frais réels moyens des couronnes dentaires est proche de 0,5, ce qui pourrait laisser penser qu'il existe tout de même une certaine corrélation positive entre ces deux variables (plus la garantie est haute, plus le prix des couronnes est élevé).

Pour chaque sous poste, on vérifie que les coefficients de corrélation entre l'âge et le temps de présence, l'âge et la garantie maximale, le temps de présence et la garantie maximale, les frais réels et les quantités d'actes sont tous bien compris entre -0,5 et 0,5. C'est bien le cas en pratique, donc on considère que pour chaque sous poste la fréquence et les frais réels sont indépendants, ce qui nous permet de conclure que le modèle « fréquence / coût moyen » peut être utilisé ici. Et on considère également que le temps de présence, le montant de la garantie maximale et l'âge sont indépendants. Ainsi, ces trois variables explicatives peuvent être utilisées. Toutefois, ce n'est pas l'âge qui sera retenu pour ce mémoire, mais bien la tranche d'âge.

3 MODELISATION DES POSTES

L'objectif de ce chapitre est d'expliquer en détail les étapes de la modélisation de la fréquence et du coût moyen de chaque sous poste.

3.1 « TARIFABILITE » DES POSTES

3.1.1 Détermination des postes tarifables

Avant de commencer la tarification, il est important de vérifier si tous les sous postes peuvent être tarifés, et si ce n'est pas le cas, de voir ce qui les empêche d'être tarifés.

Tout d'abord, on peut voir que dans chaque grand poste, il y a un libellé de sous poste intitulé « Autre sous postes Grand poste ». Il correspond à tous les actes qui n'ont pas pu être classés parmi les autres sous postes. Par exemple, le sous poste « Autres sous postes Soins courants » contient à la fois les indemnités kilométriques des auxiliaires médicaux, des soins à l'étranger ou bien des actes divers non identifiés. Il y a ainsi plusieurs actes très différents pour un même libellé, qui peuvent, par conséquent, avoir des garanties distinctes. Ce n'est donc pas possible de déterminer une seule et même garantie pour ce genre de sous postes. On va donc supprimer ces postes dans la suite de notre étude, mais on devra les prendre en compte plus tard dans la tarification.

Par ailleurs, nous remarquons que certains sous postes sont très peu représentés dans la consommation. Cela peut provenir de deux raisons. La première est que certains de ces sous postes ne sont tout simplement pas remboursés par la complémentaire santé chez la plupart des clients qui constituent la base de données. En conséquence, ils n'apparaissent pas dans les lignes à lignes de prestations de ces clients. La seconde raison peut être liée au fait que ces sous postes sont très peu consommés. La question à se poser est alors : quelle est le nombre d'observations minimal à avoir par sous poste pour pouvoir espérer construire un modèle prédictif ? Après plusieurs recherches documentaires, la « règle des 1 sur 10 » [16] est souvent apparue comme une solution simple au problème. Cette règle indique que pour des problèmes de régression, la taille minimale pour la base de données qui permette de construire un modèle performant, est de 10 fois le nombre de paramètres. Dans notre cas, on compte 3 paramètres : la tranche d'âge, le temps d'exposition et le montant maximal remboursé par la garantie. Cela veut donc dire une taille minimale de 30 données. Par mesure de précaution, nous avons donc décidé d'éliminer tous les sous postes qui comptaient moins de 100 observations. Comme précédemment, il faudra en tenir compte plus tard dans la tarification. Cela concerne les sous postes suivants :

- Frais d'obsèques
- Lit d'accompagnant
- Maternité Autre
- Maternité Chambre particulière
- Certains verres complexes et hyper complexes, pour adultes et enfants, avec des bases de remboursement spécifiques

Une table de sous postes non tarifables est créée. Elle sera utilisée par la suite pour la détection de ces cas particuliers.

Finalement, le nombre de sous postes à tarifer s'élève à 67.

3.1.2 Solution envisagée pour les sous postes non tarifables

Afin de tenir compte des sous postes non tarifables dans la tarification, nous avons calculé la part que représentaient ces sous postes, en termes de remboursement complémentaires, sur les remboursements totaux des sous postes tarifables. Nous nous sommes appuyés sur l'hypothèse que la consommation des sous postes non tarifables serait la même que celle qui va être calculée ci-dessous pour chaque bénéficiaire.

On introduit les notations suivantes :

- q_{nt} la part de remboursement des sous postes non tarifables
- R_{comp} les remboursements complémentaires totaux
- $R_{comp,nt}$ les remboursements complémentaires des sous postes non tarifables

On a alors :

$$q_{nt} = \frac{R_{comp,nt}}{R_{comp} - R_{comp,nt}}$$

Le coefficient q_{nt} trouvé est de 1,1 %. Il faudra donc multiplier le résultat final par $1+1{,}1\% = 101{,}1$ %, pour inclure les sous postes non tarifables.

Cette méthode présente tout de même des limites puisqu'on considère que le poids des sous postes non tarifables sera le même pour tous les clients pour lesquels une tarification sera réalisée, quel que soit le nombre de sous postes non tarifables couvert par le contrat d'assurance et quel que soit le montant de leurs garanties.

Cependant, en calculant le coefficient q_{nt} pour chaque client de la base de données, on trouve des valeurs comprises entre 0,8% et 1,3%. Ces valeurs restent très proches du coefficient trouvé sur l'ensemble du portefeuille, malgré le fait que les clients n'ont pas tous le même nombre de sous postes non tarifables couverts par leurs garanties, ni les mêmes montants de garantie. Utiliser le coefficient calculé sur l'ensemble du portefeuille semble donc être une bonne approximation.

3.2 PREMIERS ESSAIS DE MODELISATION : PARAMETRAGE MANUEL DES MODELES

Une première version de la modélisation a été réalisée sur les sous postes « SPR 50 » et « Consultations et Visites Spécialistes OPTAM ». Pour chacun d'eux, plusieurs modèles de *machine learning* ont été testés, pour modéliser la fréquence et le coût moyen. Les modèles testés sont les suivants :

- CART
- Forêts aléatoires
- GBM
- Réseau de neurones

Le paramétrage des modèles, ainsi que l'analyse de leur performance sont énoncés si dessous.

3.2.1 Echantillonnage des données

Avant de passer à l'étape de modélisation, les données doivent être divisées en plusieurs échantillons. Le découpage se fera de la manière suivante : un échantillon d'apprentissage et un échantillon test. Il n'y aura pas d'échantillon de validation car la validation des modèles se fera avec la validation croisée. En effet, cette méthode permet d'éliminer en partie le biais lié à l'échantillonnage.

L'échantillon test devra être représentatif de la base de données, pour éviter que les prédictions ne soient biaisées. L'échantillonnage se fera sous poste par sous poste, pour la fréquence et le cout moyen.

Avant toute chose, on affecte à *seed* la valeur 1 (choisie arbitrairement), pour figer l'aléa lié à l'échantillonnage. Les échantillons tests seront ainsi les mêmes, quelle que soit la méthode de construction des modèles.

3.2.2 Modélisation de la fréquence

La base de données est filtrée sur le sous poste étudié (Consultations et Visites Spécialistes OPTAM ou SPR 50). La fréquence sera la variable à expliquer, et la tranche d'âge, le temps d'exposition et la garantie seront les variables explicatives. Les autres variables ne sont pas prises en compte. La base de données est ensuite divisée aléatoirement en deux parties :

- Echantillon d'apprentissage : 90% des données
- Echantillon test : 10% des données

Pour les deux sous postes, on vérifie que les deux échantillons aient des caractéristiques similaires : fréquence moyenne, minimale, et maximale, répartition parmi les tranches d'âge…

On pourrait intuitivement penser que les trois variables explicatives vont avoir de l'influence sur la fréquence de consommation sur ces deux sous postes, mais que celles du temps d'exposition et de la tranche d'âge seront plus importantes. L'analyse des modèles qui seront établis ci-dessous permettra de valider, ou non, ces hypothèses.

Arbres de régression

Un modèle d'arbre de régression est d'abord testé pour la fréquence, grâce à la fonction *rpart*, du package du même nom. Tout d'abord, l'arbre maximal est construit, c'est-à-dire avec un coefficient de complexité *cp* égal à 0. Il est également possible de

modifier d'autres paramètres pour le contrôle de la construction des arbres. Nous décidons cependant de les laisser par défaut :

- Nombre minimal d'observations dans chaque nœud : 20
- Nombre minimal d'observations dans chaque feuille : 7
- Profondeur maximale de l'arbre : 30
- …

Les arbres obtenus pour les deux sous-postes comportent un grand nombre de ramifications, ce qui rend leur interprétation complexe. Ils se trouvent en annexe 8. De plus, cela peut engendrer un problème de surapprentissage. Afin d'éviter cela, on va diminuer la complexité de l'arbre, en jouant sur le *cp*. Le *cp* le plus pertinent est celui qui minimise l'erreur de validation croisée.

Les graphiques représentent les erreurs de validations croisées des Consultations et Visites Spécialistes OPTAM et des SPR 50 en fonction du coefficient de complexité (*cp-plot*) sont très difficiles à lire à cause du grand nombre de données ; ils sont disponibles en annexe 8. En zoomant et en utilisant la fonction de R *locator*, qui permet de récupérer les coordonnées d'un point sur un graphique, on trouve pour les consultations spécialistes : $cp = 2{,}63 \times 10^{-4}$, et pour les couronnes dentaires : $cp = 9{,}99 \times 10^{-4}$.

On procède ensuite à l'élagage des arbres maximaux, grâce aux coefficients de complexité trouvés. Les arbres obtenus sont tracés ci-dessous :

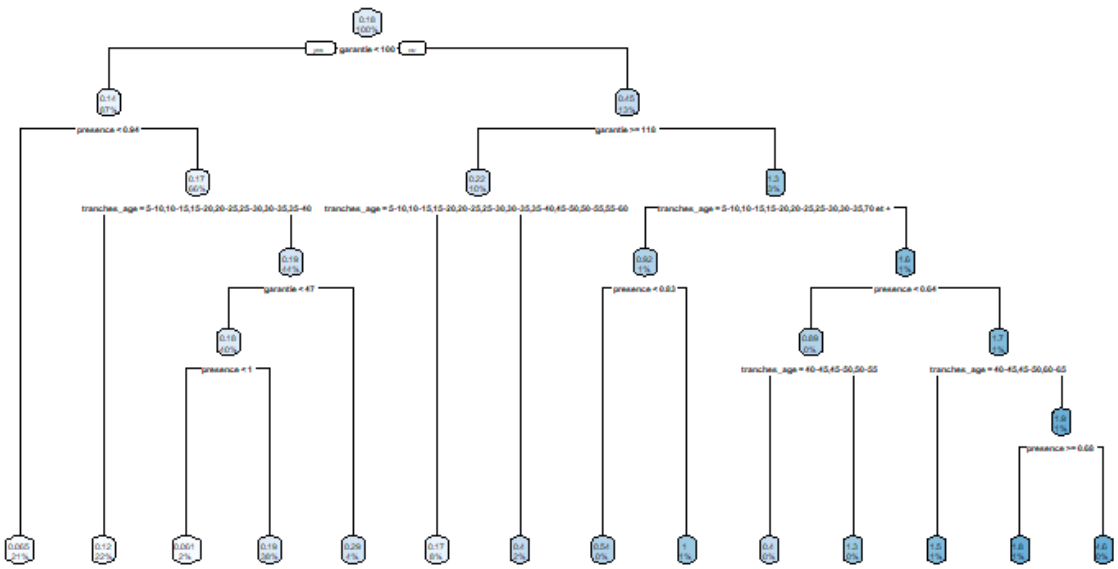


Figure 3.1 – Arbre optimal modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

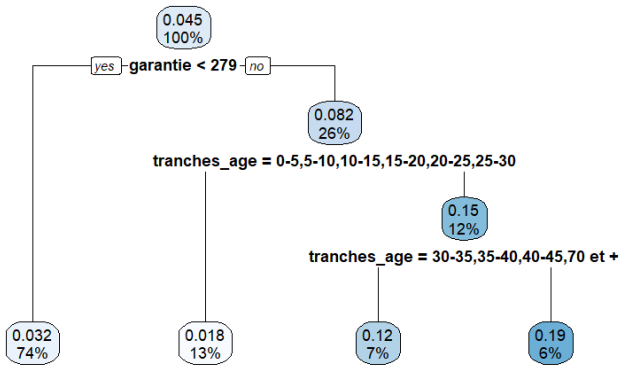


Figure 3.2 – Arbre optimal modélisant la fréquence des SPR 50

On remarque que pour le sous poste Consultations et Visites Spécialistes OPTAM, toutes les variables sont utilisées. En revanche, pour le sous poste SPR 50 toutes ne sont pas forcément utilisées. L'élagage de l'arbre a permis de sélectionner les variables qui avaient le plus d'influence. Par ailleurs, l'arbre des couronnes dentaires (4 feuilles) est plus petit, et donc moins complexe, que l'arbre des consultations spécialistes (14 feuilles).

Sous poste	Garantie	Temps de présence	Tranche d'âge
Consultations et Visites Spécialistes OPTAM	24%	38%	38%
SPR 50	33%	0%	67%

Tableau 3.1 – Importance des variables pour les arbres modélisant la fréquence

En effet, on peut voir que l'importance des variables est très différente entre les deux sous postes. Pour les consultations spécialistes, toutes les variables semblent avoir une importance du même ordre de grandeur, avec une importance tout de même un peu plus faible pour la garantie, ce qui peut sembler assez cohérent puisqu'une garantie plus élevée ne va a priori pas pousser les assurés à consommer sur ce type de poste. Une garantie faible pourra cependant générer un peu de renoncement aux soins.

En revanche, pour les couronnes dentaires, c'est clairement la tranche d'âge qui a le plus d'importance. Sur les couronnes dentaires, un point a particulièrement attiré notre attention : le temps de présence ne semble pas avoir d'importance sur cet arbre ; en effet, il n'est pas utilisé comme critère de séparation au niveau des nœuds. Cela est surprenant, car on peut penser qu'une personne présente toute l'année aura plus de chance de consommer qu'une personne présente seulement une partie de l'année. Cependant, l'arbre étant de petite taille (seulement 4 feuilles), il y a peu de critères de séparation. En augmentant le paramètre *cp*, le temps de présence apparaît comme critère de sélection sur plusieurs nœuds. Toutefois, l'erreur de validation croisée étant plus élevée, on préfèrera conserver le premier modèle, tout en gardant en tête cet aspect.

Forêts aléatoires

Nous allons maintenant modéliser la fréquence grâce aux forêts aléatoires du package *randomForest*, et plus particulièrement à la fonction du même nom. Afin d'optimiser les modèles, il y a deux paramètres à régler : le nombre d'arbres dans la forêt et le nombre de variables explicatives tirées aléatoirement sur chaque nœud.

Comme évoqué dans le paragraphe sur les forêts aléatoires du chapitre 1.3.5., le nombre de variables explicatives tirées aléatoirement sur chaque nœud m' dépend du nombre de variables explicatives m. Pour un cas de régression, il est recommandé de prendre $m' = m/3$. Etant donné qu'il y a 3 variables explicatives, il faudrait prendre $m' = 1$. Il a toutefois été décidé de tester également $m' = 2$, mais pas $m' = 3$ car cela enlèverait l'aléatoire (et augmenterait la corrélation entre les arbres). Après avoir testé les deux valeurs de m' pour les deux sous postes, avec le nombre d'arbres laissé à sa valeur par défaut ($ntree = 500$), il apparait que $m' = 2$ minimise l'erreur quadratique moyenne ($MSE = RMSE^2$), obtenue par validation croisée. C'est donc cette valeur qui sera retenue.

Afin d'optimiser le nombre d'arbres, le MSE, obtenu par validation croisée, a été tracé en fonction du nombre d'arbres pour les deux sous postes :

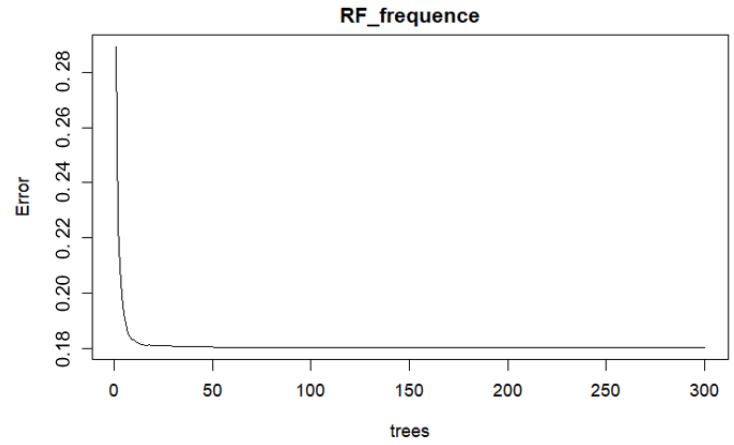


Figure 3.3 – MSE en fonction du nombre d'arbres dans la forêt aléatoire modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

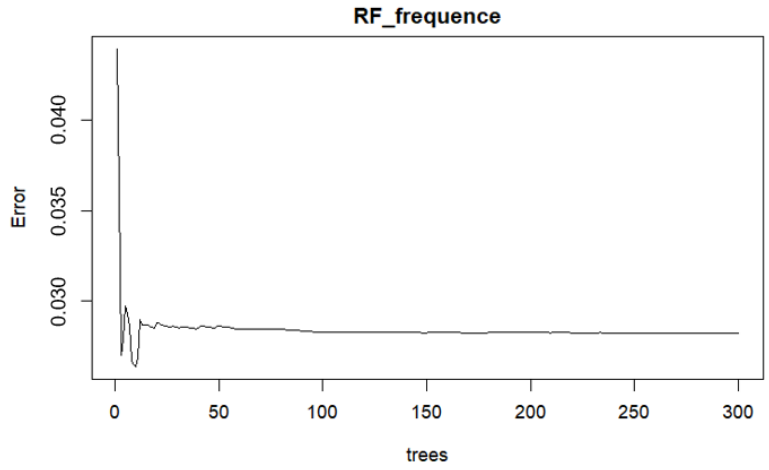


Figure 3.4 – MSE en fonction du nombre d'arbres dans la forêt aléatoire modélisant la fréquence des SPR 50

Le nombre d'arbres retenu est le plus petit nombre à partir duquel le MSE se stabilise au minimum. Pour les Consultations et Visites Spécialistes OPTAM, le MSE semble se stabiliser au minimum autour de 50 arbres. Pour les SPR 50 en revanche, cela est plus compliqué. L'endroit où le MSE se stabilise, à 100 arbres, n'est pas la valeur minimum. On retiendra toute de même cette valeur, mais on gardera bien en tête cette spécificité.

La forêt aléatoire ne sera sans doute pas le meilleur modèle pour modéliser la fréquence des SPR 50.

Un tunage automatique est également possible sur R, mais le temps d'optimisation est très long et cela est déconseillé pour les grosses bases de données, comme celle que nous utilisons.

Les graphiques d'importance des variables sont tracés :

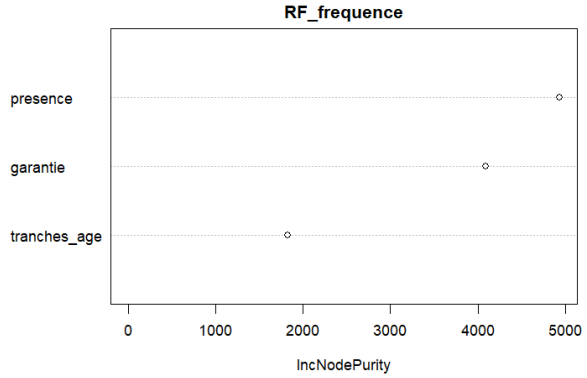


Figure 3.5 – Importance des variables de la forêt aléatoire modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

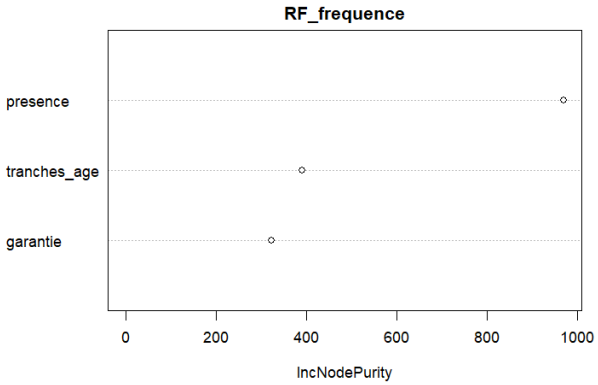


Figure 3.6 – Importance des variables de la forêt aléatoire modélisant la fréquence des SPR 50

Pour les deux sous postes, le temps de présence semble avoir une grande importance : c'est en effet la variable qui en a le plus. Pour les couronnes dentaires, cette conclusion est contraire à celle tirée sur l'arbre de décision, mais elle semble plus cohérente. Le fait de tirer aléatoirement deux variables sur trois à chaque nœud a donc dû forcer l'utilisation de la variable temps de présence comme critère de sélection ; il s'est révélé très discriminant.

Puis le modèle des consultations spécialistes accorde également une grande importance à la garantie, mais une faible importance à la tranche d'âge. Tandis que le modèle des SPR 50 accorde une importance moyenne et similaire à la garantie et à la tranche d'âge.

GBM

Un GBM va maintenant être testé pour modéliser la fréquence des deux sous postes. Plusieurs paramètres vont permettre de pouvoir optimiser le modèle : la fonction de perte, le nombre d'arbres, le coefficient de rétrécissement et la profondeur des arbres. La fonction *gbm*, du package du même nom, sera utilisée dans cette partie.

La variable à modéliser étant la fréquence, il s'agit d'une variable quantitative discrète. La fonction de perte choisie est donc celle associée à la loi de Poisson. Il s'agit de la log-vraisemblance de Poisson.

Le coefficient de rétrécissement, noté *shrinkage*, permet de jouer sur la vitesse de convergence de l'algorithme en donnant une influence plus forte aux corrections apportées à chaque arbre. Il est d'usage de forcer cette valeur à 0,1, car cela permet un bon compromis entre un coefficient trop faible qui créerait du sous-apprentissage et un coefficient trop élevé qui, au contraire, engendrerait du surapprentissage.

Pour déterminer le nombre d'arbres *ntree* et leurs profondeurs *interaction.depth*, plusieurs couples de valeurs vont être testés. Celui qui sera retenu sera celui qui donnera le meilleur MSE en validation croisée. Pour le nombre d'arbres, les valeurs 50, 200 et 500 vont être testées. Pour la profondeur des arbres, on va tester les valeurs des profondeurs des arbres élagués obtenus précédemment (3 et 6), ainsi que la valeur 9. Les résultats obtenus sont répertoriés dans les tableaux suivants :

interaction.depth \ ntree	50	200	500
3	0.18438	0.18444	0.18417
6	0.18416	0.18386	0.18438
9	0.18428	0.18474	0.18569

Tableau 3.2 – MSE en fonction du nombre d'arbres et de leur profondeur pour le GBM modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

Pour le sous poste Consultations et Visites Spécialistes OPTAM, le meilleur couple testé est le couple avec 200 arbres et une profondeur de 6.

interaction.depth \ ntree	50	200	500
3	0.02902	0.02904	0.02923
6	0.02904	0.02913	0.02945
9	0.02907	0.02919	0.02957

Tableau 3.3 – MSE en fonction du nombre d'arbres et de leur profondeur pour le GBM modélisant la fréquence des SPR 50

Pour le sous poste SPR 50, le meilleur couple testé est le couple avec 50 arbres et une profondeur de 3.

Les graphiques ci-dessous montrent l'importance relative des variables pour les deux modèles obtenus :

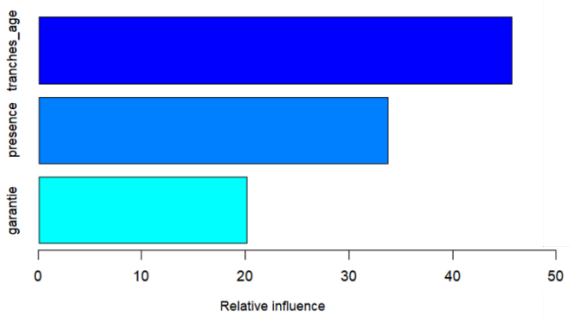


Figure 3.7 – Importance des variables du GBM modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

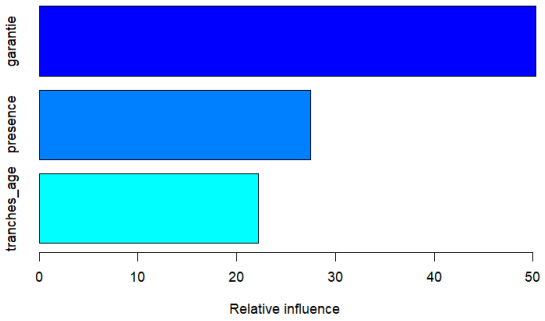


Figure 3.8 – Importance des variables du GBM modélisant la fréquence des SPR 50

L'importance accordée aux variables est très différente pour les deux modèles. Pour les consultations spécialistes, c'est la tranche d'âge qui a le plus d'importance, tandis que pour les couronnes dentaires, c'est la garantie. Mais dans chaque cas, une importance non négligeable (de plus de 20%) est accordée aux deux autres variables. Ainsi, quel que soit le sous poste, toutes les variables explicatives ont été prises en compte dans le modèle. Etant donné qu'il y a plusieurs arbres dans un GBM, cela augmente les chances de toutes les retrouver.

Réseau de neurones

Pour finir, un réseau de neurones est testé pour la modélisation de la fréquence. Pour cela, la fonction *nnet*, du package du même nom, est utilisée. Il permet de modéliser un perceptron multicouche à une seule couche cachée. Ce package a été choisi car il est performant, robuste et facile à utiliser. Pour optimiser le modèle, plusieurs paramètres doivent être définis : le nombre de d'itérations maximal, le nombre de neurones sur la couche cachée et le paramètre de régularisation.

Pour déterminer le nombre de neurones sur la couche cachée et le paramètres de régularisation, on fait appel à la fonction *tune.nnt*, qui permet d'optimiser ces paramètres grâce à la validation croisée. Pour le nombre de neurones sur la couche cachée, noté *size*, les valeurs 1, 2 et 3 vont être testées. On ne peut pas tester plus de valeurs car l'algorithme est très coûteux en termes de temps de calcul et les bases de données pour la modélisation de la fréquence sont très volumineuses. Le paramètre de régularisation *decay* prend en général des petites valeurs, bien inférieures à 1. On va donc tester les valeurs suivantes : 0,1, 0,01, 0,001 et 0,0001. Les graphiques des performances des algorithmes (MSE) en fonction de *size* et *decay* sont représentés ci-dessous :

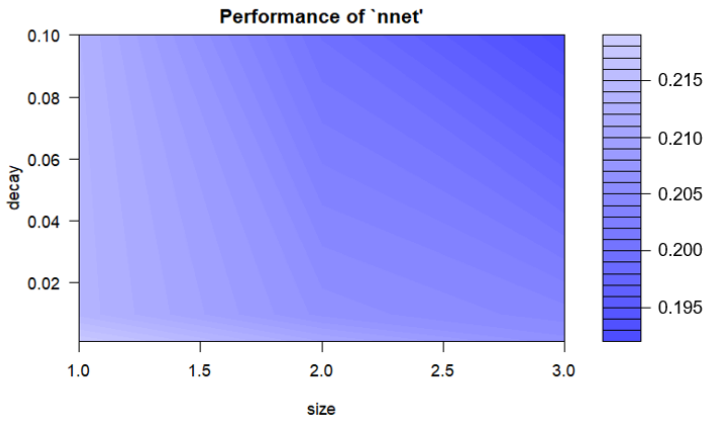


Figure 3.9 – MSE en fonction du nombre de neurones et du paramètre de régularisation pour le réseau de neurones modélisant la fréquence des Consultations et Visites Spécialistes OPTAM

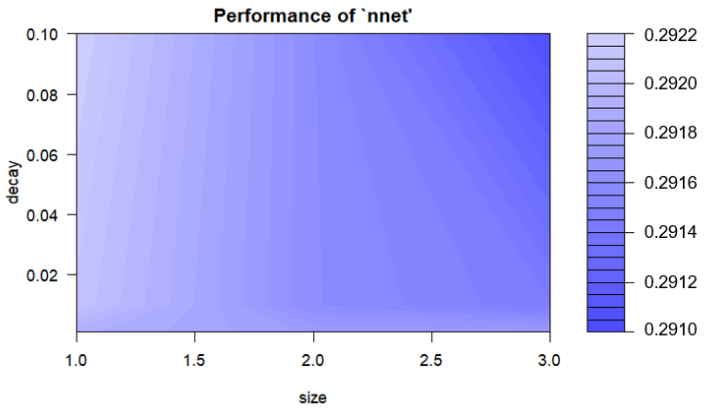


Figure 3.10 – MSE en fonction du nombre de neurones et du paramètre de régularisation pour le réseau de neurones modélisant la fréquence des SPR 50

Pour le sous poste Consultations et Visites Spécialistes OTPAM, on trouve $size = 3$ et $decay = 0{,}1$. Et pour le sous poste SPR 50, on trouve aussi $size = 3$ et $decay = 0{,}1$. Les valeurs de $size$ et $decay$ permettant d'obtenir le meilleur modèle se trouvent dans les deux cas à l'extrémité en haut à droite du graphique. On pourrait donc penser qu'en augmentant ces valeurs, on obtiendrait un modèle encore meilleur. Etant donné que $decay$ prend en général des valeurs bien inférieures à 1, on ne va pas tester de valeurs plus grandes. On gardera $decay = 0{,}1$. En revanche, pour le paramètre $size$, la valeur 4 va être testée. Les MSE obtenus pour les deux sous postes sont plus élevés que ceux pour les paramètres $size = 3$ et $decay = 0{,}1$, donc ce sont bien ces deux paramètres qui seront gardés.

Une fois que ces deux paramètres ont été déterminés, on cherche à trouver le meilleur nombre d'itération maximal *maxit*. Pour cela, on teste plusieurs valeurs de *maxit*, et on prend la plus petite valeur pour laquelle l'algorithme converge. Le nombre d'itération maximal optimal trouvé pour les consultations spécialistes est de 100, et il est de 150 pour les couronnes dentaires.

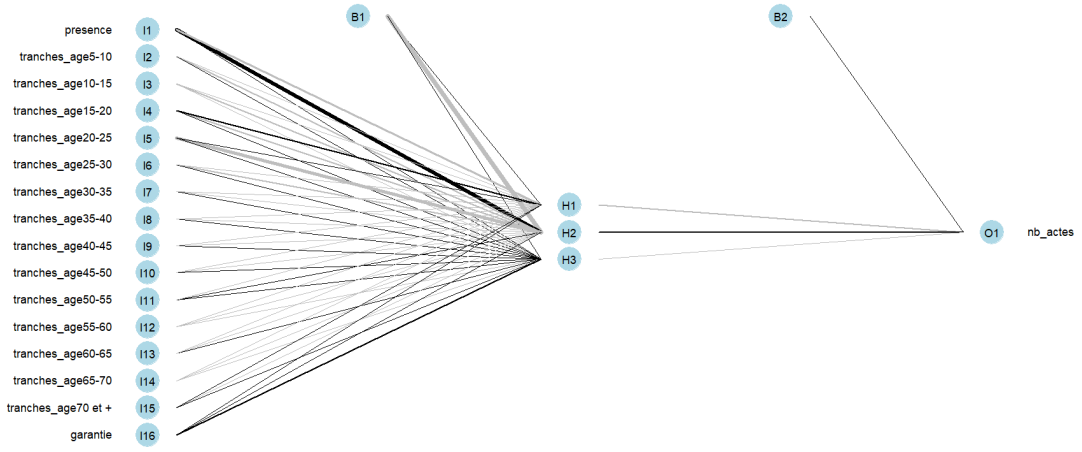


Figure 3.11 – Représentation graphique du réseau de neurones modélisant la fréquence des Consultations et Visites Spécialistes OPTAM et des SPR 50

Le graphique ci-dessus représente le réseau de neurones tracé pour chaque sous poste. Ils sont identiques, puisqu'ils ont les mêmes paramètres.

L'importance des variables est donnée dans le tableau suivant :

Sous poste	Garantie	Temps de présence	Tranche d'âge
Consultations et Visites Spécialistes OPTAM	30%	27%	43%
SPR 50	69%	1%	30%

Tableau 3.4 – Importance des variables pour les réseaux de neurones modélisant la fréquence

Le modèle des consultations spécialistes semble accorder une importance plutôt similaire aux trois variables, avec une légère préférence pour la tranche d'âge. A contrario, le modèle des couronnes dentaires ne semble accorder presque aucune importance au temps de présence, une importance très forte à la garantie et une importance moyenne à la tranche d'âge. On se retrouve dans un cas similaire à celui de l'arbre de décision.

Comparaison des modèles

Le tableau ci-dessous récapitule les RMSE et les MAE calculés pour les différents modèles de la fréquence sur l'échantillon test :

Modèle	Consultations et Visites Spécialistes OPTAM		SPR 50	
	RMSE	MAE	RMSE	MAE
Arbre de régression	0.4498	0.1087	0.1708	0.0314
Forêt aléatoire	0.4324	0.1083	0.1725	0.0313
GBM	0.4288	0.1067	0.1703	0.0306
Réseau de neurones	0.4356	0.1090	0.1704	0.0325

Tableau 3.5 – RMSE et MAE des modèles de fréquence paramétrés manuellement

Pour modéliser la fréquence des consultations spécialistes et les couronnes dentaires, le meilleur modèle semble être le GBM, car c'est celui qui a le RMSE et le MAE les plus petits dans les deux cas.

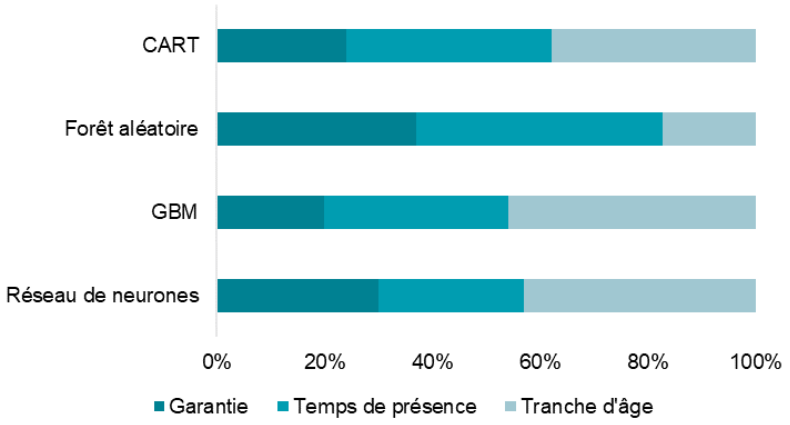


Figure 3.12 – Importance des variables pour les modèles de fréquence des Consultations et Visites Spécialistes OPTAM

Pour les Consultations et Visites Spécialistes OPTAM, l'importance des variables semble être assez similaire d'un modèle à l'autre. La garantie, le temps de présence et la tranche d'âge semblent avoir toutes les trois presque autant d'importance dans chaque modèle, bien que la tranche d'âge paraisse en avoir légèrement plus (entre

35% et 40%). Mais il y a une exception avec la forêt aléatoire qui accorde moins d'importance à la tranche d'âge (moins de 20%). Le GBM est le modèle que nous considérons comme étant le plus performant ; il accorde un peu moins d'importance à la garantie, qui n'est toutefois pas négligeable (environ 20%). Cela est donc bien en accord avec notre première intuition.

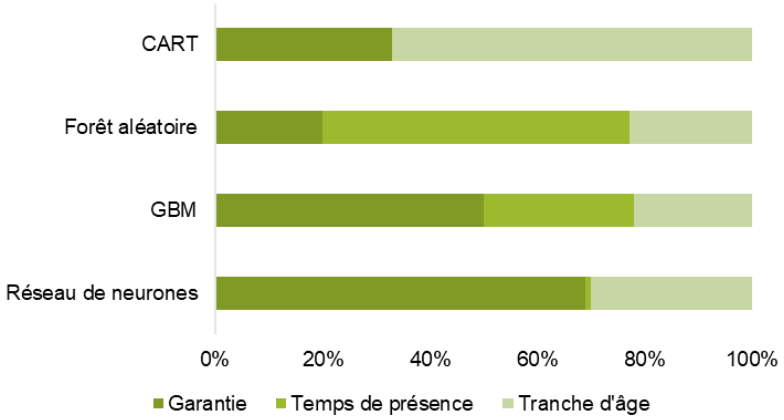


Figure 3.13 – Importance des variables pour les modèles de fréquence des SPR 50

Pour les SPR 50, l'importance des variables diffère d'un modèle à l'autre. Certains modèles (CART et réseaux neurones) n'accordent aucune importance au temps de présence, tandis que d'autres en accordent une majeure (forêt aléatoire). L'arbre de régression et la forêt aléatoire en voient leurs performances pénalisées avec des $RMSE$ et MAE élevés, tandis que le réseau de neurones se classe second au niveau de la performance, juste derrière le GBM. Ces deux modèles les plus performants accordent une grande importance à la garantie, ce qui n'est pas totalement en accord avec notre première intuition. En revanche, les résultats du GBM confirment ce que nous pensions à propos de l'importance de toutes les variables : aucune n'est négligeable.

3.2.3 Modélisation du coût moyen

De même que pour la modélisation de la fréquence, la base de données est filtrée sur le sous poste étudié. Elle est également filtrée sur les fréquences non nulles. Les frais réels moyens (ou coût moyen) seront la variable à expliquer, et les variables explicatives seront les mêmes que précédemment. La base de données est ensuite divisée en deux parties :

- Echantillon d'apprentissage : 90% des données
- Echantillon test : 10% des données

Enfin, la similarité entre les deux échantillons est vérifiée.

On pourrait intuitivement penser, pour les deux sous postes, que le temps d'exposition aura peu d'importance, que la tranche d'âge en aura moyennement, tandis que la garantie en aura une grande. L'analyse des modèles qui seront établis ci-dessous permettra de valider, ou non, ces suppositions.

Arbres de régression

Un modèle d'arbre de régression est testé en premier lieu pour le coût moyen. Tout d'abord, l'arbre maximal ($cp = 0$) est construit pour ces deux sous postes. Les autres paramètres de contrôle sont laissés par défaut. Mêmes s'ils sont plus petits que pour la

modélisation de la fréquence, les arbres obtenus pour le coût moyen comportent un grand nombre de ramifications. Leur représentation graphique se trouvent également en annexe 8. Les graphiques des erreurs de validation croisée en fonction du *cp* sont tracés :

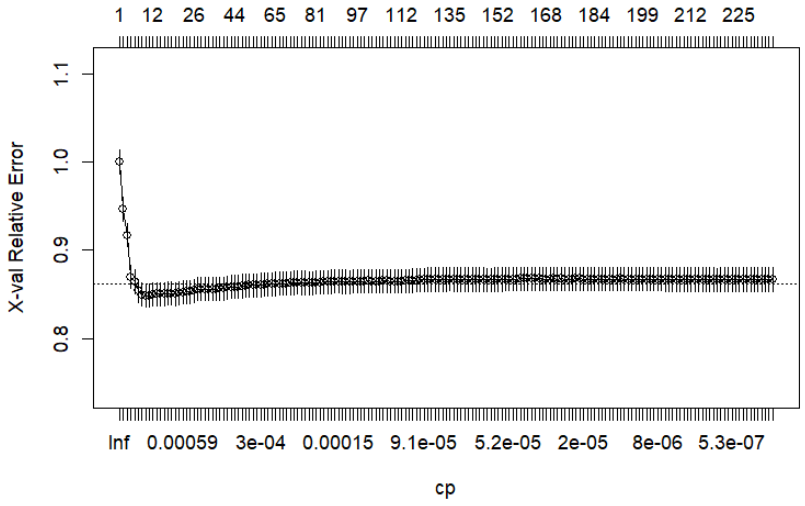


Figure 3.14 – Erreur de validation croisée en fonction du coefficient de complexité pour l'arbre maximal modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

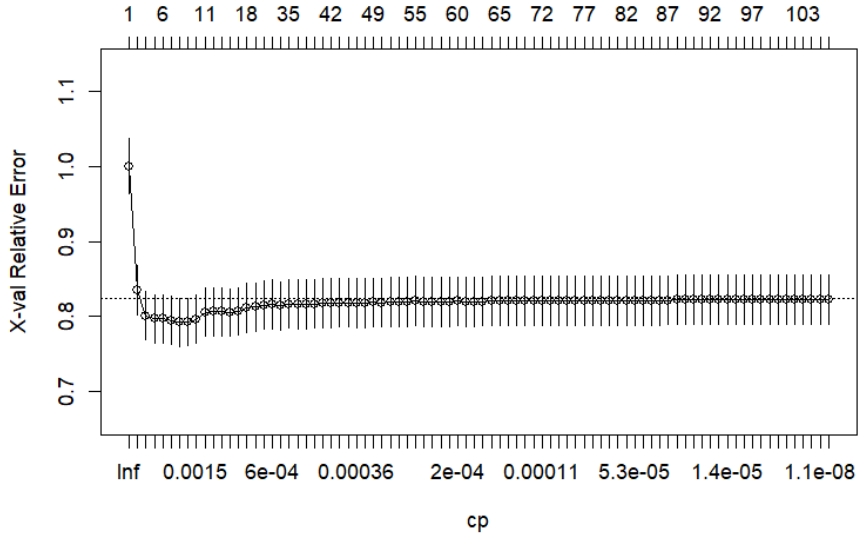


Figure 3.15 – Erreur de validation croisée en fonction du coefficient de complexité pour l'arbre maximal modélisant le coût moyen des SPR 50

On trouve les coefficients de complexité suivants : pour les consultations spécialistes $cp = 2{,}05 \times 10^{-3}$, et pour les couronnes dentaires $cp = 2{,}59 \times 10^{-3}$.

On procède ensuite à l'élagage des arbres maximaux, grâce aux coefficients de complexité trouvés. Les arbres obtenus sont tracés ci-dessous :

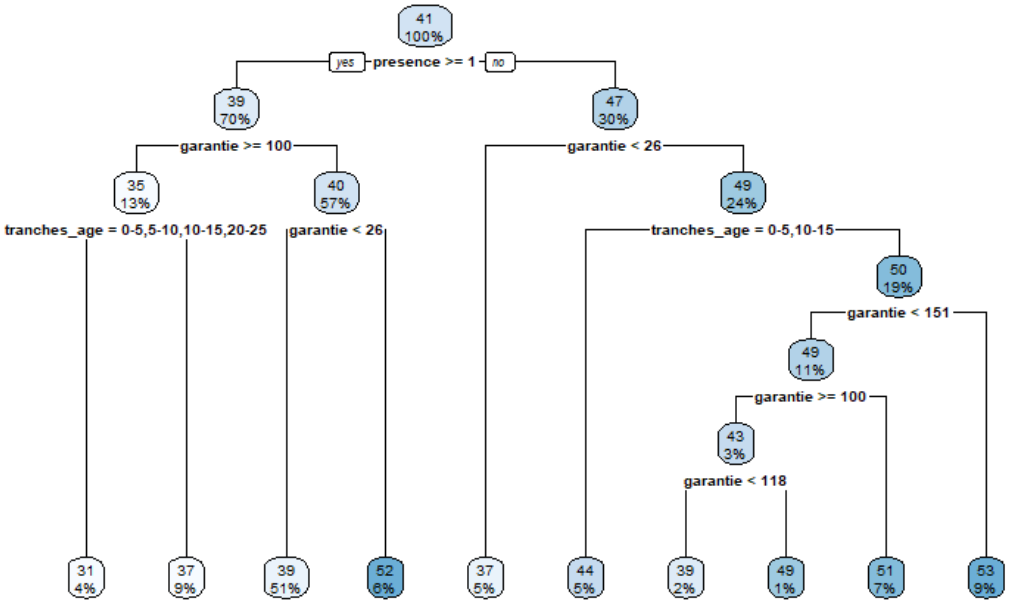


Figure 3.16 – Arbre optimal modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

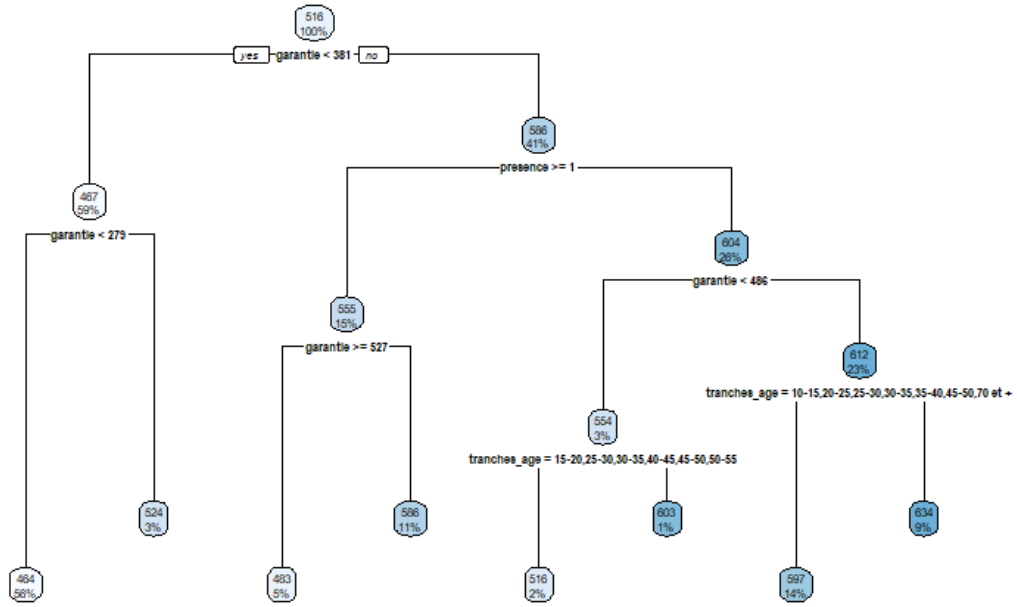


Figure 3.17 – Arbre optimal modélisant le coût moyen des SPR 50

Ils sont de taille similaire pour les deux sous postes : 10 feuilles pour l'arbre des consultations spécialistes et 8 feuilles pour l'arbre des couronnes dentaires. Par ailleurs, on peut remarquer que toutes les variables ont été utilisées comme critère de sélection sur les deux arbres, contrairement à l'arbre de modélisation de la fréquence des couronnes dentaires. En effet, les arbres modélisant le coût moyen sont plus grands que ce dernier, ce qui laisse plus de place aux variables moins importantes.

Sous poste	Garantie	Temps de présence	Tranche d'âge
Consultations et Visites Spécialistes OPTAM	67%	11%	22%
SPR 50	57%	14%	29%

Tableau 3.6 – Importance des variables pour les arbres modélisant le coût moyen

Contrairement à l'arbre de régression pour la modélisation de la fréquence, l'importance des variables semble être assez similaire entre les deux sous postes, bien que la garantie ait légèrement plus d'importance pour les consultations spécialistes que pour les couronnes dentaires.

Forêts aléatoires

Nous allons maintenant modéliser le coût moyen grâce aux forêts aléatoires. On règle également le nombre d'arbres dans la forêt et le nombre de variables explicatives tirées aléatoirement sur chaque nœud.

De même que pour la fréquence, les modèles sont testés avec le nombre de variables explicatives tirées aléatoirement sur chaque nœud m' égal à 1 et 2, avec le nombre d'arbres laissé à sa valeur par défaut. C'est également $m' = 2$ qui est retenu pour les deux sous postes.

Afin d'optimiser le nombre d'arbres, le MSE, obtenu par validation croisée, a été tracé en fonction du nombre d'arbres pour les deux sous postes :

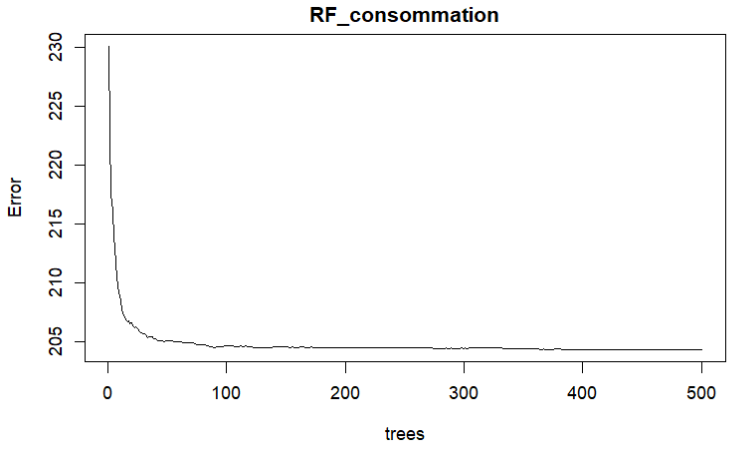


Figure 3.18 – MSE en fonction du nombre d'arbres dans la forêt aléatoire modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

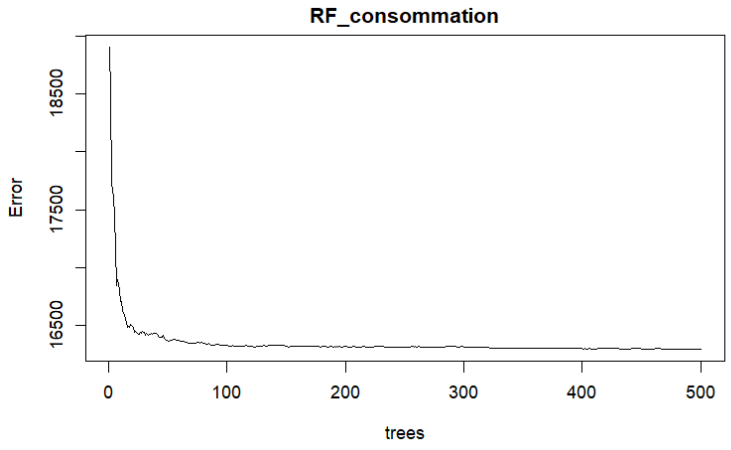


Figure 3.19 – MSE en fonction du nombre d'arbres dans la forêt aléatoire modélisant la fréquence des SPR 50

En procédant de la même manière que pour la fréquence, le nombre d'arbres retenu pour le sous poste Consultations et Visites Spécialistes OPTAM est de 120, et pour le sous poste SRP 50, il est de 100.

Les graphiques d'importance des variables sont tracés :

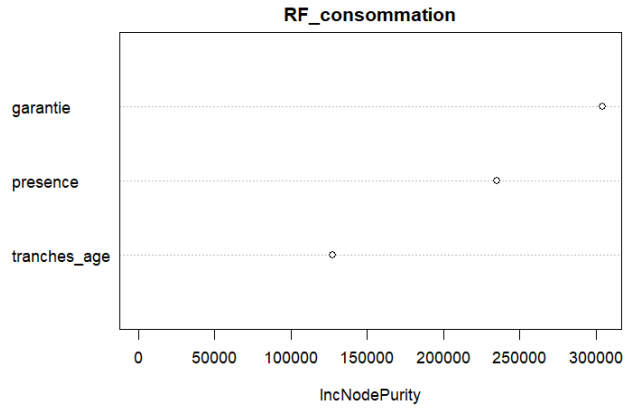


Figure 3.20 – Importance des variables de la forêt aléatoire modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

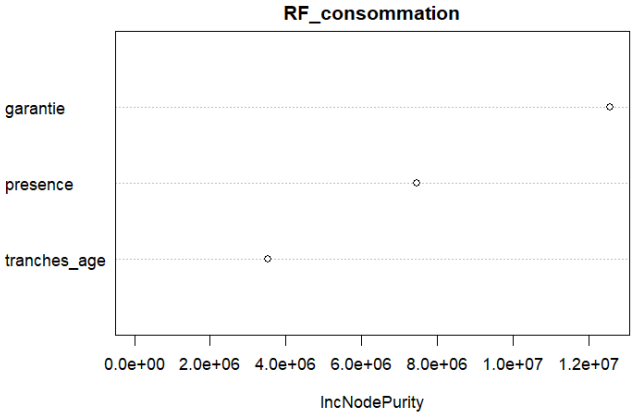


Figure 3.21 – Importance des variables de la forêt aléatoire modélisant le coût moyen des SPR 50

Les modèles des deux sous postes sont très similaires : la garantie est la variable qui a le plus d'importance, suivie par le temps de présence puis par la tranche d'âge. L'importance accordée au temps de présence est plutôt élevée, ce qui semble assez surprenant.

GBM

Un GBM va maintenant être testé pour modéliser le coût moyen des deux sous postes. Les mêmes paramètres que ceux utilisés pour la fréquence vont être réglés, afin d'optimiser le modèle.

La variable à modéliser étant le coût moyen, il s'agit d'une variable quantitative continue. La fonction de perte choisie est celle associée à la loi de Gauss. Il s'agit du MSE.

On affecte au coefficient de rétrécissement la valeur 0,1, comme pour la modélisation de la fréquence.

Plusieurs couples de valeur (nombre d'arbres, profondeur des arbres) sont testés. Les résultats obtenus sont répertoriés dans les tableaux suivants :

interaction.depth \ ntree	50	200	500
3	227.149	212.095	209.003
6	223.850	209.687	208.504
9	223.523	209.738	208.944

Tableau 3.7 – MSE en fonction du nombre d'arbres et de leur profondeur pour le GBM modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

Pour le sous poste Consultations et Visites Spécialistes OPTAM, le meilleur couple testé est le couple avec 500 arbres et une profondeur de 6.

interaction.depth \ ntree	50	200	500
3	14 554.806	13 333.367	13 293.629
6	14 437.897	13 228.358	13 265.119
9	14 417.069	13 225.874	13 281.179

Tableau 3.8 – MSE en fonction du nombre d'arbres et de leur profondeur pour le GBM modélisant le coût moyen des SPR 50

Pour le sous poste SPR 50, le meilleur couple testé est le couple avec 200 arbres et une profondeur de 9.

Les graphiques ci-dessous montrent l'importance relative des variables pour les deux modèles obtenus :

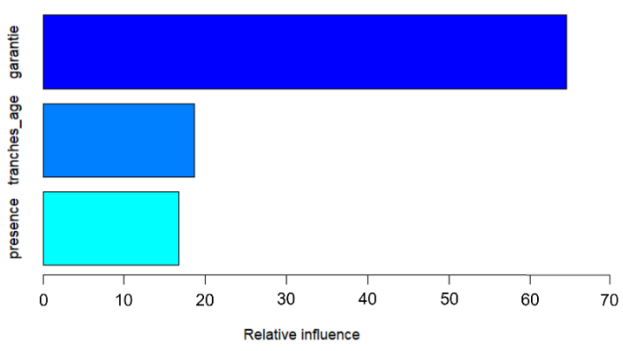


Figure 3.22 – Importance des variables du GBM modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

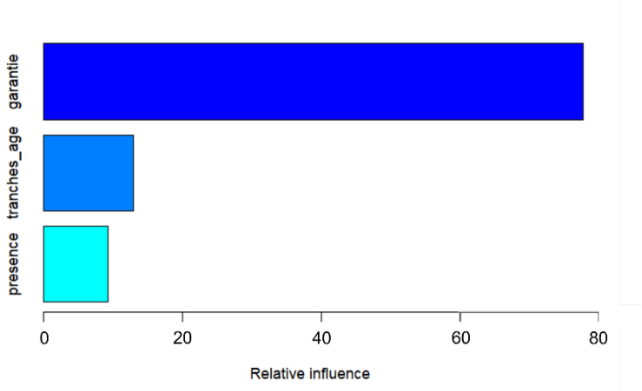


Figure 3.23 – Importance des variables du GBM modélisant le coût moyen des SPR 50

L'importance des variables est assez similaire entre les deux sous postes : la garantie compte le plus (importance relative de plus de 60%), puis la tranche d'âge et enfin le temps de présence, qui ont une importance relative très proche dans chaque cas. Toutefois, on peut soulever le fait que l'importance accordée à la variable garantie est plus élevée pour le sous poste SPR 50. Cette importance plus élevée pour les couronnes dentaires peut venir du fait que, sur ce poste, outre le résultat médical, il y a également un résultat esthétique suivant le matériau utilisé. Les coûts sont différents en fonction des matériaux utilisés donc cela paraît cohérent que l'impact de la garantie soit plus important pour les couronnes dentaires que pour les consultations spécialistes, qui n'ont pas cette dimension esthétique.

Réseau de neurones

Pour finir, un réseau de neurones est testé pour la modélisation du coût moyen. Tout comme pour la fréquence, on cherche à optimiser le modèle en définissant le nombre de d'itérations maximal, le nombre de neurones sur la couche cachée et le paramètre de régularisation.

Pour déterminer le nombre de neurones sur la couche cachée et le paramètres de régularisation, on fait également appel à la fonction *tune.nnt*. Pour *size*, étant donné que la taille de la base de données qui sert à la modélisation du coût moyen est plus petite, on va pouvoir tester plus de valeurs : les valeurs 1 à 10 sont donc testées. Et pour *decay,* on teste les valeurs suivantes : 0,1, 0,01, 0,001 et 0,0001. Les graphiques des performances (MSE) des algorithmes en fonction de *size* et *decay* sont représentés ci-dessous :

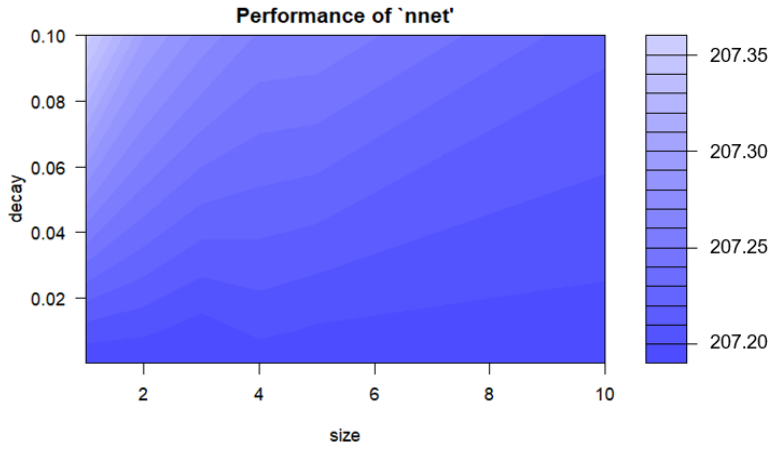


Figure 3.24 – MSE en fonction du nombre de neurones et du paramètre de régularisation pour le réseau de neurones modélisant le coût moyen des Consultations et Visites Spécialistes OPTAM

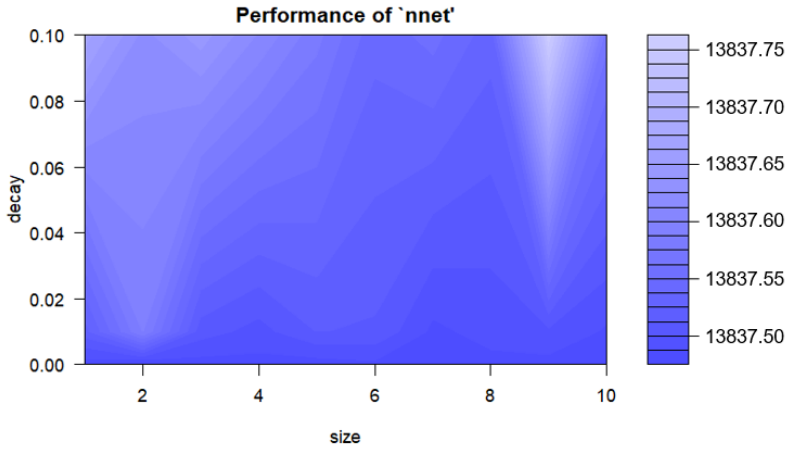


Figure 3.25 – MSE en fonction du nombre de neurones et du paramètre de régularisation pour le réseau de neurones modélisant le coût moyen des SPR 50

Pour le sous poste Consultations et Visites Spécialistes OTPAM, on trouve $size = 1$ et $decay = 0{,}0001$. Et pour le sous poste SPR 50, on trouve $size = 4$ et $decay = 0{,}0001$.

Une fois que ces deux paramètres ont été déterminés, on cherche à trouver le meilleur nombre d'itération maximal *maxit*, de la même manière que pour les modèles de la fréquence. Le nombre d'itération maximal optimal trouvé pour les consultations spécialistes est de 70, et il est de 50 pour les couronnes dentaires.

Les représentations des réseaux de neurones des deux sous postes se trouvent en annexe 8. L'importance des variables est donnée dans le tableau suivant :

Sous poste	Garantie	Temps de présence	Tranche d'âge
Consultations et Visites Spécialistes OPTAM	62%	10%	28%
SPR 50	28%	4%	68%

Tableau 3.9 – Importance des variables des réseaux de neurones modélisant le coût moyen

Pour les deux sous postes, le temps de présence ne semble pas avoir beaucoup d'importance, ce qui semble assez réaliste. Par ailleurs, pour les consultations spécialistes, une importance forte est accordée à la garantie, tandis que pour les couronnes dentaires, une importance forte est accordée à la tranche d'âge. La plus faible importance accordée par le réseau de neurones pour le sous poste SPR 50 est contraire à ce qui a pu être illustré sur les modèles précédents et paraît surprenant.

Comparaison des modèles

Le tableau ci-dessous récapitule les RMSE et les MAE calculés pour les différents modèles du coût moyen sur l'échantillon test :

Modèle	Consultations et Visites Spécialistes OPTAM		SPR 50	
	RMSE	MAE	RMSE	MAE
Arbre de régression	14.53	11.66	115.19	90.45
Forêt aléatoire	14.58	11.76	116.51	91.02
GBM	14.44	11.62	115.00	88.92
Réseau de neurones	14.39	11.63	117.63	91.83

Tableau 3.10 – RMSE et MAE des modèles de coût moyen paramétrés manuellement

Pour modéliser le coût moyen des consultations spécialistes, deux modèles semblent convenir : le GBM et le réseau de neurones. En effet, le GBM a le MAE le plus faible et le réseau de neurones a le RMSE le plus faible. On retiendra cependant plutôt le réseau de neurones car les valeurs des MAE sont très proches. Concernant les couronnes dentaires, le meilleur modèle semble être, une fois de plus, le GBM, car les deux critères (RMSE et MAE) sont plus faibles pour ce modèle.

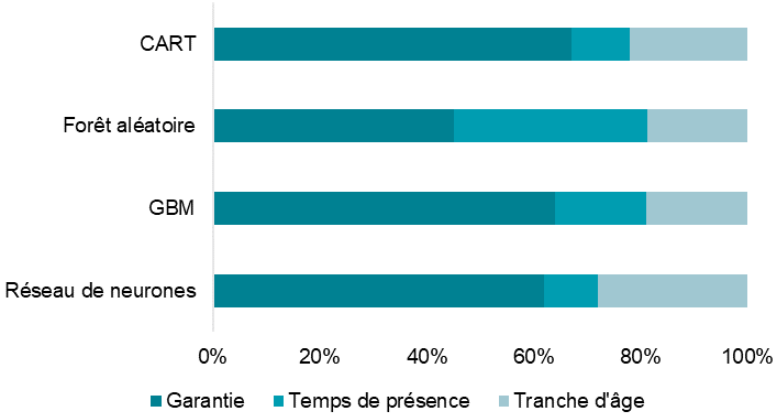


Figure 3.26 – Importance des variables pour les modèles de coût moyen des Consultations et Visites Spécialistes OPTAM

Pour les Consultations et Visites Spécialistes OPTAM, de même que pour la fréquence, les résultats sont similaires entre les modèles, excepté pour la forêt aléatoire qui est celui que nous considérons comme le moins performant. Pour l'arbre de régression, le GBM et le réseau de neurones, la garantie a une très grande importance, et le temps de présence et la tranche d'âge ont une importance faible. L'importance de la tranche d'âge est toutefois plus importante que celle du temps de présence.

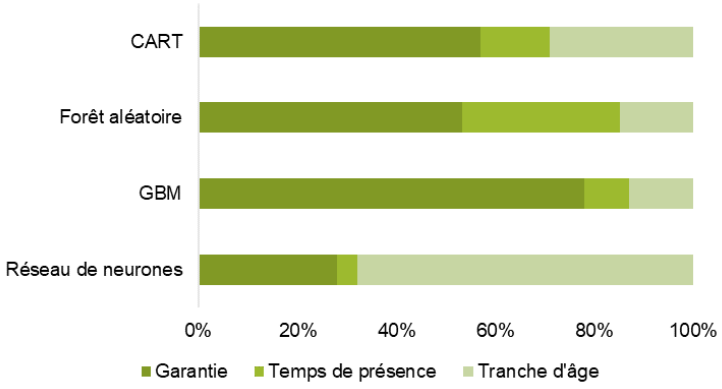


Figure 3.27 – Importance des variables pour les modèles de coût moyen des SPR 50

Pour les SPR 50, un modèle se distingue des autres : il s'agit du réseau de neurones qui accordent une importance plus faible à la garantie. Cela est contraire à notre intuition. Mais ce modèle est classé dernier au niveau de la performance, ce qui nous amène à penser qu'il n'est pas vraiment adapté. Les autres modèles accordent une importance forte à la garantie, et en particulier le GBM (environ 80%) qui est le modèle le plus performant.

Ainsi, pour les deux sous postes, les modèles sont unanimes et accordent le plus d'importance à la garantie. A contrario, le temps de présence semble être la variable la moins importante. Cela semble cohérent et confirme notre première intuition.

3.2.4 Importance de l'utilisation d'un algorithme « auto-ML »

Comme évoqué précédemment, le nombre de sous postes à tarifer est de 67. Pour chacun des sous postes, il faut modéliser la fréquence et le coût moyen. Il y a donc au total 134 modèles à construire. On ne peut donc pas se permettre de chercher pour chaque sous poste le modèle utilisé, puis d'optimiser chaque paramètre car cela est particulièrement chronophage.

L'objectif est donc de pouvoir automatiser le choix et la création des modèles pour la fréquence et le coût moyen. Pour cela, nous avons besoin d'une fonction permettant de modéliser la fréquence pour un sous poste donnée, et une similaire pour le coût moyen, que l'on fera tourner pour chaque sous poste. Pour la création des modèles à proprement parlé, la fonction *h2o.automl*, présentée dans le premier chapitre, sera utilisée. En effet, elle répond à nos besoins car elle va permettre de modéliser les sous postes de façon automatique, tout en donnant le meilleur modèle.

Les fonctions de modélisation de la fréquence et du coût moyen devront sélectionner :

- La base de données utilisée pour l'apprentissage
- Les variables explicatives
- La variable à expliquer
- Les paramètres de la fonction *h2o.automl*

Ensuite, il faudra récupérer les résultats, à savoir les modèles paramétrés, pour chaque sous poste, et les stocker.

Il faudra enfin exécuter les fonctions de fréquence et de coût à l'aide d'une boucle sur les sous postes. Afin de gérer au mieux l'automatisation, les sous postes ont été numérotés de 1 à 67. Leurs numéros peuvent être retrouvés grâce à une table de correspondance, présente en annexe 9.

3.3 MODELISATION A L'AIDE D'UN ALGORITHME « AUTO-ML »

La fonction *h2o.automl* va être utilisée dans cette partie, afin de trouver les meilleurs modèles pour la fréquence et le coût moyen de chaque sous poste.

De même que pour les premiers essais de modélisation, les données sont divisées en deux parties pour chaque sous poste, pour la fréquence et le coût moyen :

- Echantillon d'apprentissage : 90% des données
- Echantillon test : 10% des données

Pour séparer les données, on garde $seed = 1$, car cela va nous permettre d'obtenir les mêmes échantillons tests que dans le paragraphe précédent, ce qui sera utile pour comparer les performances des modèles.

3.3.1 Modélisation de la fréquence

Dans cette sous-partie, les étapes de modélisation de la fréquence vont être décrites. Dans un premier temps, nous allons détailler la fonction permettant de modéliser la fréquence, puis nous verrons ensuite son application.

Tout d'abord, la fonction qui modélise la fréquence prend en argument le nom du sous poste à modéliser. Puis, elle effectue dans l'ordre les étapes ci-dessous :

1. **Récupération du numéro du sous poste** : cette étape se fait grâce à la table de correspondance. Le numéro sera utile pour la sauvegarde du modèle.

2. **Sélection de l'échantillon d'apprentissage** : l'objectif de cette fonction est de modéliser la fréquence ou nombre d'actes, qui est supposée indépendante du coût moyen (voir paragraphe 2.4.3). Un filtre est effectué, pour ne garder que les lignes qui concerne le sous poste en argument. Le découpage des données en échantillon d'apprentissage et échantillon de validation n'est pas nécessaire pour la modélisation avec H2O, puisque que la fonction *h2o.automl* utilise la validation croisée. Les variables suivantes sont gardées :
   - Identifiant du bénéficiaire
   - Tranche d'âge
   - Temps d'exposition
   - Garantie maximale remboursée
   - Nombre d'actes

3. **Détermination des variables explicatives** : ce sont les variables définies dans le chapitre précédent, à savoir : la tranche d'âge, le temps d'exposition et la garantie maximale remboursée. Les noms de colonne des variables explicatives sont enregistrés dans un vecteur qui servira par la suite.

4. **Détermination de la variable à expliquer** : ici, on cherche à modéliser la fréquence donc la variable à modéliser est le nombre d'actes.

5. **Conversion de l'échantillon d'apprentissage au format H2O** : pour pouvoir utiliser la fonction *h2o.automl*, il faut convertir les données au format H2O grâce à la fonction *as.h2o*. La fonction renverra en sortie un objet H2O, qui ne fonctionnera à son tour qu'avec des objet H2O.

6. **Détermination des paramètres de la fonction *h2o.automl***: les paramètres utilisés par cette fonction sont les suivants :
    - *max_runtime_secs* : nous avons choisi de laisser 1 000 secondes à la fonction pour trouver le meilleur algorithme. Cela représente entre 16 et 17 minutes. Ce temps nous paraît être un bon compromis en laissant un temps raisonnable à la fonction pour tourner, sans pour autant prendre trop de temps, étant donné que la fonction va tourner 67 fois.
    - *nfolds* : en général, le nombre de sous échantillons pour la validation croisée varie entre 5 et 10. Comme la taille des données est très importante, on choisit de diviser l'échantillon en 5, afin que les calculs ne soient pas trop lourds.
    - *seed* : la même valeur que pour la modélisation manuelle est choisie, c'est-à-dire 1. Cela permettra de pouvoir « enlever » l'aléa lié au choix de l'échantillon et pouvoir comparer les modèles entre eux.

7. **Affichage du nom du sous poste** : cette étape sera utile lorsque la fonction tournera en boucle sur tous les sous postes. Si jamais un problème survient, cela va permettre de voir quel sous poste est responsable et de faire des modifications le cas échéant.

8. **Exécution de la fonction *h2o.automl***: elle est exécutée avec tous les paramètres prédéfinis aux étapes précédentes :
    - *x* : vecteur des noms de variables explicatives déterminé à l'étape 3
    - *y* : variable à expliquer déterminée en étape 4
    - *training_frame* : échantillon d'apprentissage déterminé en étape 2
    - *max_runtime_secs* : 1 000
    - *nfolds* : 5
    - *seed* : 1

9. **Sauvegarde du modèle** : le modèle défini grâce à la fonction *h2o.automl* est sauvegardé dans un dossier, sur l'ordinateur, avec tous ses paramètres, grâce à la fonction *h2o.saveModel*, sous un format H2O. Afin de mieux organiser les dossiers pour la suite et d'éviter les erreurs, un dossier est créé pour les modèles de la fréquence. Puis chaque modèle est rangé dans un dossier intitulé avec le numéro du sous poste. Les noms des modèles sont ensuite récupérés manuellement et inscrit dans une table de correspondance, avec le nom et le numéro du sous poste.

10. **Affichage du nom du sous poste** : pour compléter l'étape 7, cette étape permet d'afficher le fait que la modélisation s'est bien déroulée pour ce sous poste.

Ensuite, la fonction est exécutée de manière automatique pour chaque sous poste. Les temps de calcul étant longs et la fonction demandant beaucoup de travail à l'ordinateur, nous avons décidé de ne pas faire tourner tous les sous postes d'un seul coup, mais plutôt de les séparer par groupe, afin d'éviter d'éventuels bugs. Ils seront donc séparés par grand poste.

Avant de faire tourner une fonction provenant du package H2O, il faut ouvrir une session H2O, qu'il faudra refermer à la fin du travail effectué. Pour chaque grand poste, on ouvre une session H2O, grâce à la fonction *h2o.init*. Puis la fonction créée ci-dessus est appliquée à chaque sous poste qui composent le grand poste, avec la fonction *map*.

Cette fonction permet d'appliquer une fonction plusieurs fois, à une liste d'arguments renseignés. On remarque que plusieurs types de modèles sont utilisés pour la fréquence. Le tableau des fréquences, disponible en annexe 10, illustre leur répartition dans les différents sous postes.

Pour les deux sous postes étudiés plus particulièrement tout au long de ce mémoire, les modèles suivants ont été choisis :

- Consultations et Visites Spécialistes OPTAM : *Stacked Ensemble (All models)*,
- SPR 50 : *GBM*.

On peut résumer les fréquences d'utilisations des différents modèles dans le tableau suivant :

Modèle	Nombre d'utilisations
DRF	5
GBM	15 (dont 6 par grille)
GLM	1
Stacked Ensemble All models	32
Stacked Ensemble Best of family	13
XRT	1

Tableau 3.11 – Utilisation des différents types de modèles pour la fréquence

Plusieurs constats peuvent être tirés de ce tableau. Tout d'abord, il n'y a pas de réseaux de neurones, ni de XGBoost. Cela peut être soit dû au fait que la fonction *h2o.automl* n'ait pas eu le temps de tester tous les types de modèles avec différents paramètres, les grilles de recherche de réseaux de neurones et de XGBoost étant les derniers algorithmes testés. Ou bien cela peut tout simplement être dû au fait que ces types de modèles ne conviennent pas pour modéliser la fréquence. Par ailleurs, on peut voir que certains modèles de GBM ont été trouvés par recherche par grille aléatoire. Cela signifie que la fonction *h2o.automl* a au moins pu aller jusqu'à ce point-là dans ses recherches des modèles. Enfin, il est important de remarquer que la plupart des modèles (environ 2/3) sont en réalité des agrégations de modèles (voir chapitre 1.5.3.), que ce soit du stacking avec tous les modèles testés, ou bien seulement avec les meilleurs de chaque type d'algorithme. Cela montre ainsi la puissance des combinaisons de modèles.

3.3.2 Modélisation du coût moyen

Dans cette sous-partie, les étapes de modélisation du coût moyen vont être décrites. Dans un premier temps, nous allons détailler la fonction permettant de modéliser le coût moyen, puis nous verrons ensuite son application. Certaines étapes sont très similaires à celles effectuées pour la fréquence ; elles ne sont donc pas autant détaillées dans cette partie.

Tout d'abord, la fonction qui modélise le coût moyen prend en argument le nom du sous poste à modéliser. Puis, elle effectue dans l'ordre les étapes ci-dessous :

1. **Récupération du numéro du sous poste**.

2. **Sélection de l'échantillon d'apprentissage** : l'objectif de cette fonction est de modéliser le coût moyen, qui est supposée indépendant de la fréquence (voir paragraphe 2.4.3). Un filtre est effectué, pour ne garder que les lignes qui concerne le sous poste en argument. Un deuxième filtre est fait pour ne garder

que les bénéficiaires qui ont consommé. Le coût moyen est calculé à partir du nombre de sinistres, donc on ne tiendra pas compte de cette variable. On va donc garder les variables suivantes :

- Identifiant du bénéficiaire
- Tranche d'âge
- Temps d'exposition
- Garantie maximale remboursée
- Frais réels moyens

3. **Détermination des variables explicatives** : ce sont les mêmes variables que pour la modélisation de la fréquence : la tranche d'âge, le temps d'exposition et la garantie maximale remboursée.

4. **Détermination de la variable à expliquer** : ici, on cherche à modéliser le coût moyen ou les frais réels moyens. Attention toutefois à ne pas confondre avec les remboursements moyens. La modélisation des frais réels ne donne donc pas directement le résultat final. Pour l'obtenir, il faudra le calculer à l'aide des garanties. Cela implique l'utilisation de l'hypothèse suivante : « nous considérons qu'il n'y a pas de remboursements d'autres mutuelles ». En effet, il est trop compliqué d'estimer ces montants et de plus ils représentent souvent une faible part très faible de la consommation (0,52% des remboursements complémentaires sur notre base de données). La consommation totale sera donc légèrement surestimée.

5. **Conversion de l'échantillon d'apprentissage au format H2O**.

6. **Détermination des paramètres de la fonction *h2o.automl***: les paramètres utilisés par cette fonction sont les mêmes que ceux utilisés pour la modélisation de la fréquence, à savoir *max_runtime_secs*, *nfolds* et *seed*. Les mêmes valeurs que pour la fréquence seront utilisées.

7. **Affichage du nom du sous poste**.

8. **Exécution de la fonction *h2o.automl***: elle est exécutée avec tous les paramètres prédéfinis aux étapes précédentes :
   - *x* : vecteur des noms de variables explicatives déterminé à l'étape 3
   - *y* : variable à expliquer déterminée en étape 4
   - *training_frame* : échantillon d'apprentissage déterminé en étape 2
   - *max_runtime_secs* : 1 000
   - *nfolds* : 5
   - *seed* : 1

9. **Sauvegarde du modèle** : la sauvegarde fonctionne de la même manière que pour la fréquence. Cette fois-ci, un dossier pour le coût moyen est créé.

10. **Affichage du nom du sous poste**.

De la même manière que précédemment, les sous postes tournent par grand postes. La fonction créée ci-dessus est appliquée à chaque sous poste qui composent le grand poste. Le tableau des coûts moyens, disponible en annexe 10, illustre leur répartition dans les différents sous postes.

Pour les deux sous postes étudiés plus particulièrement tout au long de ce mémoire, les modèles suivants ont été choisis :

- Consultations et Visites Spécialistes OPTAM : *Stacked Ensemble (Best of family)*,
- SPR 50 : *GBM*.

Les fréquences d'utilisation des différents modèles sont répertoriées ci-dessous :

Modèle	Nombre d'utilisations
Deep learning	19 (dont 16 par grille)
DRF	1
GBM	16 (dont 7 par grille)
GLM	1
Stacked Ensemble All models	13
Stacked Ensemble Best of family	13
XRT	4

Tableau 3.12 – Utilisation des différents types de modèles pour le coût moyen

On peut voir que, comme pour la modélisation de la fréquence, les XGBoost n'ont pas été retenus. Cependant, les réseaux de neurones (*deep learning*) semblent avoir fait leurs preuves cette fois-ci. On remarque également que plus de la moitié des réseaux de neurones et des GBM ont été trouvés grâce à la recherche par grille aléatoire. Cela peut laisser penser que la fonction *h2o.automl* a eu le temps d'aller plus loin dans sa liste de recherche. Cela vient sans doute du fait que les base de données sont plus petites pour le coût moyen que pour la fréquence car seuls les lignes des bénéficiaires qui ont consommé ont été gardées. Enfin, on peut voir qu'une part plus faible des modèles sont des agrégations de modèles. Mais ils forment tout de même une bonne partie des algorithmes puisqu'ils représentent un peu moins de la moitié.

3.4 PERFORMANCE DES MODELES

Les performances des modèles obtenus grâce à la fonction *h2o.automl* sont ensuite mesurées avec le RMSE normalisé par l'écart-type. Pour cela, on récupère le RMSE calculé sur les échantillons test pour chaque modèle, ainsi que l'écart-type de chaque échantillon test. Le RMSE est ensuite divisé par l'écart-type pour obtenir le RMSE normalisé, noté $RMSE_{norm}$. Il représente le rapport entre le RMSE du modèle étudié et le RMSE du prédicteur simple (c'est-à-dire la moyenne). Si le RMSE normalisé est supérieur à 1, alors utiliser la moyenne comme estimation de la fréquence ou du coût moyen aurait donné de meilleurs résultats.

Un modèle performant aura un RMSE normalisé très proche de zéro. Cependant, il n'existe pas de valeur exacte fixant une limite entre un bon et un mauvais RMSE normalisé. Cette valeur dépend du problème. Les répartitions des RMSE normalisés des modèles de fréquence et de coût moyen sont représentées dans les graphiques ci-dessous.

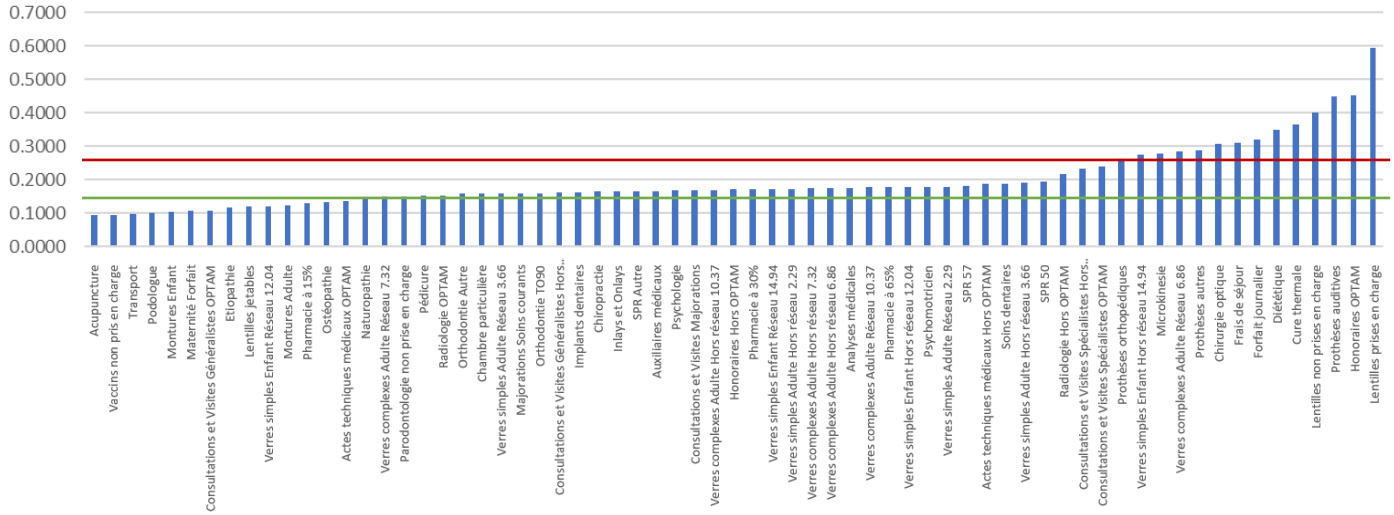


Figure 3.28 – Répartition des RMSE normalisés pour les modèles de fréquence

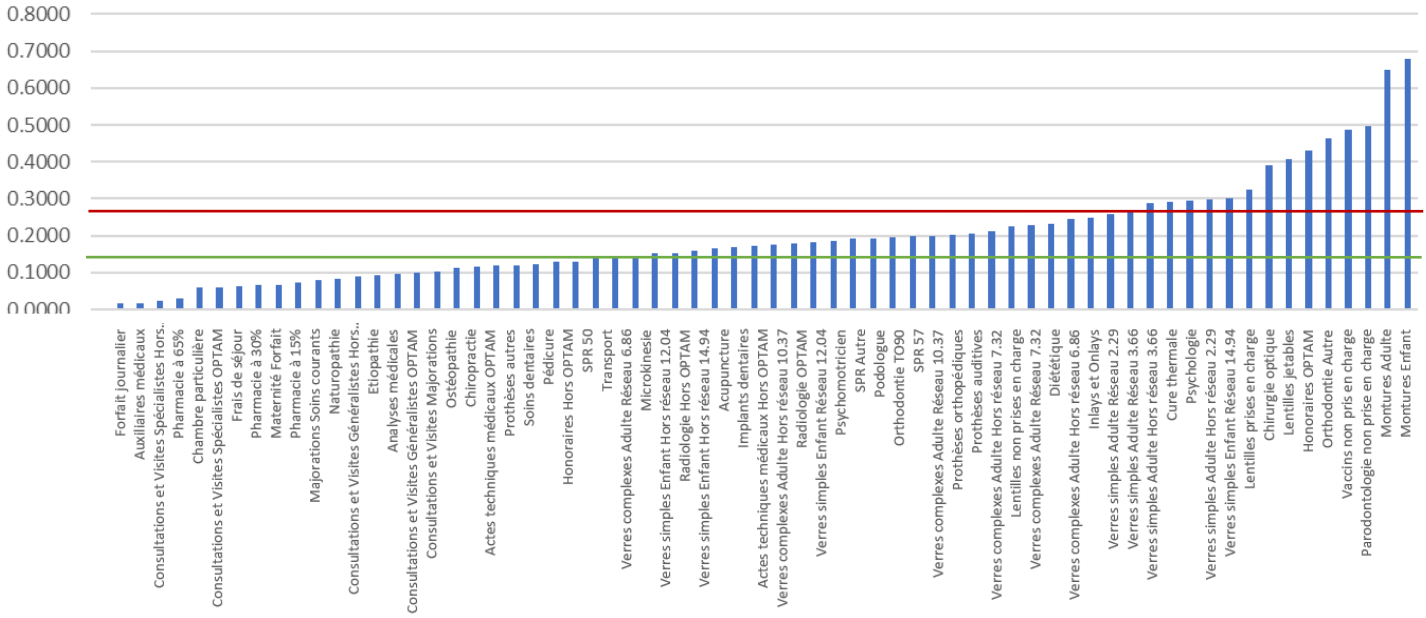


Figure 3.29 – Répartition des RMSE normalisés pour les modèles de coût moyen

Comme le montrent les graphiques ci-dessus, tous les RMSE normalisés sont inférieurs à 1. Les modèles établis sont donc plus performants que le prédicteur simple.

Les modèles de chaque sous postes sont classés dans l'ordre croissant des RMSE normalisés. Ils sont ensuite répartis en trois groupes :

- Groupe 1 : $RMSE_{norm} < 0{,}14$,
- Groupe 2 : $0{,}14 \leq RMSE_{norm} < 0{,}27$,
- Groupe 3 : $RMSE_{norm} \geq 0{,}27$.

Les valeurs limites d'appartenance à un groupe ont été déterminées à la lecture des graphiques. Au sein des groupes 1 et 2, les modèles ont des RMSE normalisés proches. Les modèles du groupe 1 sont a priori des modèles plutôt performants, tandis que ceux du groupe 2 le sont un peu moins. Le groupe 3 regroupe le reste des modèles, avec des valeurs de RMSE plus élevées et plus dispersées. Ce sont a priori des modèles peu performants.

Afin de réaliser une meilleure analyse des performances des modèles, on va également regarder les MAE des modèles normalisés par la moyenne. Pour cela, on récupère le MAE de chaque modèle, également calculé sur les échantillons test, ainsi que la moyenne de chaque échantillon test. Le MAE est ensuite divisé par la moyenne pour obtenir le MAE normalisé, noté MAE_{norm}. Il permet d'exprimer le pourcentage d'erreur absolue par rapport à la moyenne.

De même que pour le RMSE normalisé, plus le MAE_{norm} est proche de zéro, meilleur est le modèle, mais il n'existe pas de valeur exacte fixant une limite entre un bon et un mauvais MAE normalisé. Les répartitions des MAE normalisés des modèles de fréquence et de coût moyen sont représentées dans les graphiques ci-dessous.

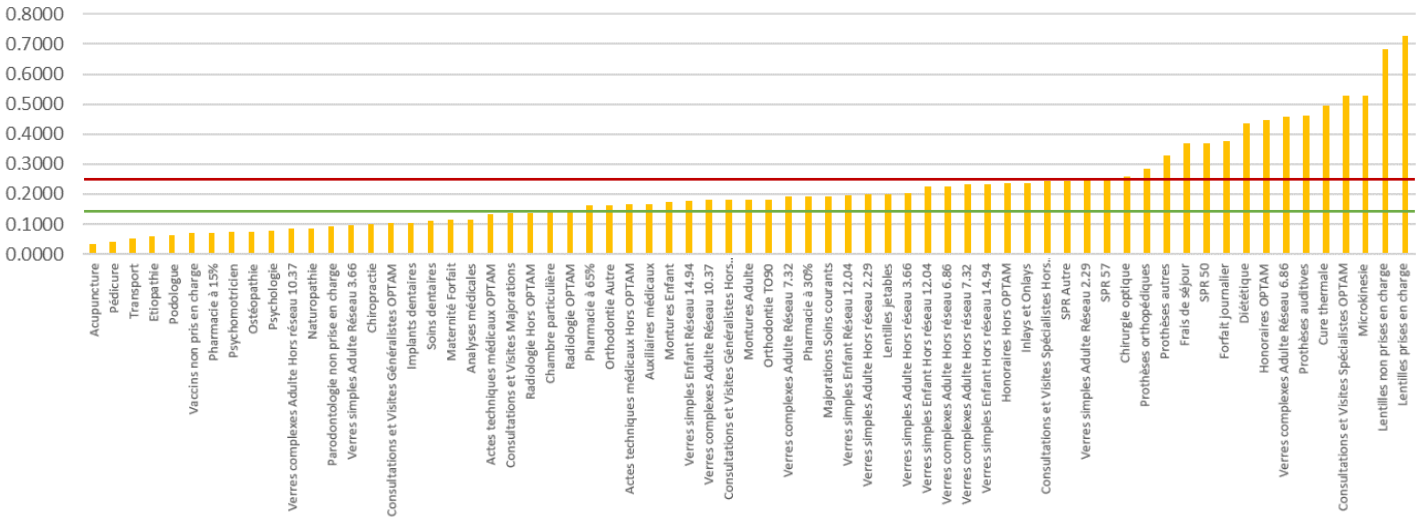


Figure 3.30 – Répartition des MAE normalisés pour les modèles de fréquence

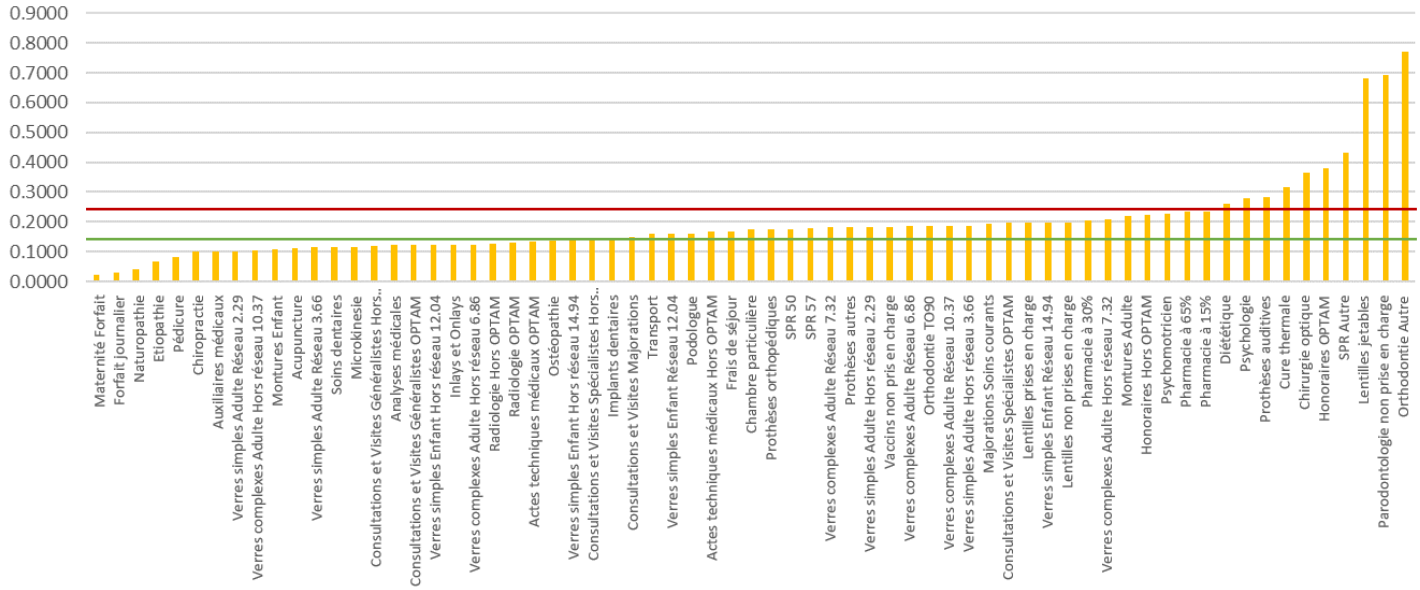


Figure 3.31 – Répartition des MAE normalisés pour les modèles de coût moyen

Les modèles de chaque sous postes sont classés dans l'ordre croissant des MAE normalisés. Ils sont également répartis en trois groupes :

- Groupe 1 : $MAE_{norm} < 0{,}13$,
- Groupe 2 : $0{,}13 \leq MAE_{norm} < 0{,}23$,
- Groupe 3 : $MAE_{norm} \geq 0{,}23$.

Les RMSE, $RMSE_{norm}$, MAE et MAE_{norm} de chaque sous poste sont disponibles en annexe 11, pour la fréquence et pour le coût moyen. Les tableaux suivants récapitulent la répartition des modèles selon leur groupe décrit par le $RMSE_{norm}$ et le MAE_{norm}.

Modélisation de la fréquence	Groupe 1	Groupe 2	Groupe 3
Performance déterminée par le $RMSE_{norm}$	21%	60%	19%
Performance déterminée par le MAE_{norm}	30%	36%	34%

Tableau 3.13 – Répartition des modèles de fréquence selon leur performance

Modélisation du coût moyen	Groupe 1	Groupe 2	Groupe 3
Performance déterminée par le $RMSE_{norm}$	37%	42%	21%
Performance déterminée par le MAE_{norm}	33%	49%	18%

Tableau 3.14 – Répartition des modèles de coût moyen selon leur performance

Selon nos critères, environ un tiers des modèles de fréquence et de coût moyen appartiennent au groupe 1 et sont probablement assez performants. De manière générale, un modèle appartenant au groupe 1 par le $RMSE_{norm}$ appartient aussi au groupe 1 avec le MAE_{norm}. Ce n'est toutefois pas le cas sur tous les modèles.

Globalement, il semblerait que les modèles de fréquence soient moins bons que les modèles de coût moyen. Cela vient probablement du fait que les bases de données des fréquences sont considérablement plus volumineuses que celles des coûts moyens et donc que l'apprentissage d'un modèle est plus long pour la fréquence. La fonction *h2o.automl* a pu tester plus de modèles pour le coût moyen et a donc eu plus de chance de tomber sur un meilleur modèle. Un autre élément qui pourrait expliquer

qu'il soit plus difficile de modéliser la fréquence que le coût moyen est l'hétérogénéité de la population qui n'a pas consommé. En effet, parmi ces personnes, certaines n'avaient réellement pas besoin de soins, en revanche, d'autres n'ont pas consommé à cause du renoncement aux soins, d'autres n'ont pas effectué de demande de remboursements… Cette population est plus hétérogène et il peut donc être plus difficile de constituer des modèles.

Dans l'annexe 11, on peut voir que le modèle de fréquence du sous poste Consultations et Visites Spécialistes OPTAM (*Stacked Ensemble All models*) appartient au groupe 2 avec le $RMSE_{norm}$ (0,2377), et au groupe 3 avec le MAE_{norm} (0,5279). On peut donc penser que ce n'est pas le modèle le plus adapté pour modéliser la fréquence. Le modèle de fréquence du sous poste SPR 50 (GBM) appartient aux mêmes groupes, mais avec toutefois des indicateurs plus faibles ($RMSE_{norm} = 0{,}1942$ et $MAE_{norm} = 0{,}3699$). Même s'il semble plus performant que le modèle des consultations spécialistes, celui des couronnes dentaires ne semble pas être non plus le modèle le plus optimal.

En revanche, les modèles de coût moyen de ces deux sous postes (*Stacked Ensemble Best of family* pour les consultations spécialistes et GBM pour les couronnes dentaires) appartiennent tous les deux au groupe 1 par le $RMSE_{norm}$ (consultations spécialistes : 0,0598 et couronnes dentaires : 0,1289) et au groupe 2 par le MAE_{norm} (consultations spécialistes : 0,1960 et couronnes dentaires : 0,1765). Ils semblent donc plus adaptés que les modèles de fréquence.

On peut donc en conclure que l'automatisation du paramétrage des modèles avec H2O s'est révélée plutôt efficace, puisque cela a permis de construire de nombreux modèles différents sans intervention humaine. Parmi les modèles établis, environ un tiers, soit une quarantaine de modèles, semblent performants au regard des $RMSE_{norm}$ et MAE_{norm}. Cependant, il reste une majorité de modèles qui ne paraissent pas assez satisfaisants. Ceux-ci doivent donc être améliorés.

Pour cela, la méthode suivante a été pensée :

1. La fréquence et le coût moyen sont modélisés une première fois pour chaque sous poste, grâce aux fonctions définies aux paragraphes 3.3.1 et 3.3.2. Pour rappel, le temps laissé à l'algorithme pour trouver un modèle est de 1 000 secondes.
2. Les performances des modèles obtenus sont mesurées à l'aide du $RMSE_{norm}$ et du MAE_{norm} sur les échantillons test. Les modèles ayant a minima un des deux indicateurs les classant dans le groupe 1 et un autre dans le groupe 2 sont conservés. Les autres modèles vont être améliorés.
3. Pour cela, les fonctions définies aux paragraphes 3.3.1 et 3.3.2 sont de nouveau utilisées, mais cette fois le temps pour trouver les modèles est quadruplé. Il est alors de 4 000 secondes, ce qui représente environ une heure.
4. Les performances sont de nouveau mesurées et classés de la même manière que précédemment. Les modèles ayant a minima un des deux indicateurs les classant dans le groupe 1 et un autre dans le groupe 2 sont conservés.
5. Pour les modèles qui n'ont pas été retenus, la méthode va dépendre du nombre de modèles restants. Si on compte un nombre faible de modèles, ces derniers peuvent être paramétrés manuellement, de la même manière que dans le paragraphe 3.2. En revanche, s'ils sont nombreux, il faudra refaire tourner les fonctions définies aux paragraphes 3.3.1 et 3.3.2, en augmentant encore le temps laissé à l'algorithme pour trouver un modèle. Cette étape peut donc

s'avérée être très longue, voire compliquée à mettre en œuvre sans ordinateur très puissant. C'est donc une limite à cette procédure.

Ayant quitté l'entreprise et n'ayant plus les données à disposition, cette procédure n'a pas pu être réalisée.

Nous allons maintenant regarder plus en détail le RMSE et le MAE des sous postes Consultations et Visites Spécialistes OPTAM et SPR 50, afin de les comparer avec les RMSE et MAE des modèles paramétrés manuellement.

Modèle	Consultations et Visites Spécialistes OPTAM		SPR 50	
	RMSE	MAE	RMSE	MAE
Arbre de régression	0.4498	0.1087	0.1708	0.0184
Forêt aléatoire	0.4324	0.1083	0.1725	0.0183
GBM	0.4288	0.1067	0.1704	0.0176
Réseau de neurones	0.4356	0.1090	0.1704	0.0195
Modèle H2O	**0.4496**	**0.1097**	**0.1707**	**0.0187**

Tableau 3.15 – RMSE et MAE des modèles de fréquence

Les modèles de fréquence mis en place par H2O ne sont pas les meilleurs modèles. Le modèle des consultations spécialistes (*Stacked Ensemble*) se classe en avant dernière position au regard du RMSE et en dernière position au regard du MAE. Le modèle des couronnes dentaires (GBM) est, quant à lui, classé avant dernier avec le RMSE et en milieu de tableau avec le MAE. Cela confirme l'étude des RMSE et MAE normalisés, qui ne qualifiaient pas ces modèles de performants (groupes 2 et 3).

Modèle	Consultations et Visites Spécialistes OPTAM		SPR 50	
	RMSE	MAE	RMSE	MAE
Arbre de régression	14.53	11.66	115.19	90.45
Forêt aléatoire	14.58	11.76	116.51	91.02
GBM	14.44	11.62	115.00	88.92
Réseau de neurones	14.39	11.63	117.63	91.83
Modèle H2O	**11.38**	**8.14**	**116.03**	**90.96**

Tableau 3.16 – RMSE et MAE des modèles de coût moyen

Les modèles de coût moyen semblent en revanche plus adaptés. En effet, le modèle des consultations spécialistes (*Stacked Ensemble*) semble être de loin le meilleur modèle, que ce soit avec le RMSE ou le MAE. Le modèle des couronnes dentaires (GBM) est également bien classé ; il est en deuxième position avec le RMSE et en troisième position avec le MAE. Cela confirme, une fois de plus, l'étude des RMSE et MAE normalisés qui qualifiaient plutôt ces modèles de performants (groupes 1 et 2).

Ainsi, même si H2O semble apporter des résultats plutôt bons pour certains modèles, ils ne sont pas toujours plus performants qu'un modèle paramétré manuellement. Une procédure sélectionnant les modèles les plus performants et trouvant de meilleurs modèles pour ceux qui le sont moins permettrait de remédier à ce problème et d'obtenir un ensemble de modèles plus performants.

3.5 INTERPRETABILITE DES MODELES

Afin d'interpréter les modèles générés par H2O, nous allons utiliser les graphiques de dépendance partielle, et plus spécifiquement sur les sous postes Consultations et Visites Spécialistes OPTAM et SPR 50.

SPR 50

- *Modélisation de la fréquence*

La fréquence des SPR 50 est modélisée grâce à un GBM.

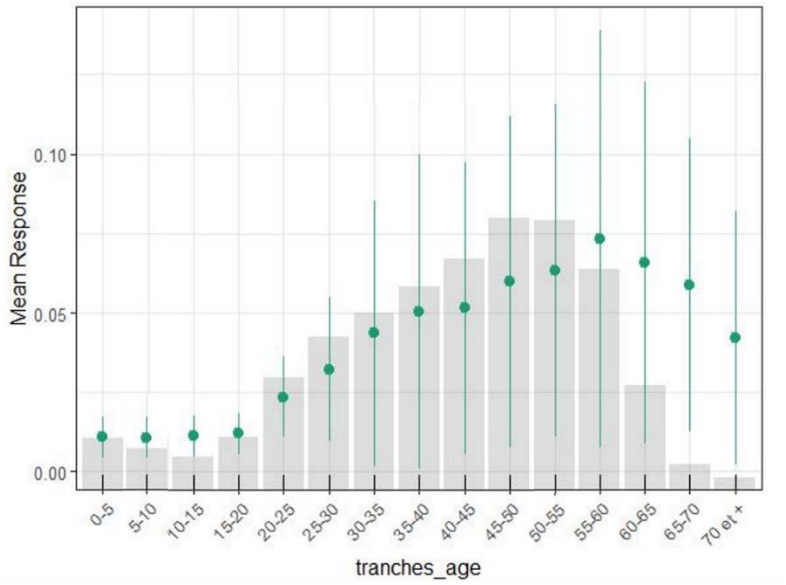


Figure 3.32 – Graphique de dépendance partielle de la variable tranche d'âge pour le modèle de fréquence H2O des SPR 50

Le graphique ci-dessus montre la courbe de dépendance partielle (points verts) en fonction de la tranche d'âge. La courbe de dépendance partielle est la moyenne des courbes ICE (*Individual Conditional Expectations*). L'intervalle [*moyenne ICE* $\pm$ *écart-type ICE*] est également représenté sur ce graphique (barres vertes). Enfin, on distingue la distribution des bénéficiaires par tranche d'âge (histogramme gris).

Ce graphique nous montre que la fréquence modélisée augmente avec l'âge à partir de 20 ans, puis elle diminue à partir de 60 ans. En effet, les bénéficiaires les plus jeunes ont moins besoin de couronnes dentaires. Les plus âgés ont, dans certains cas, besoin de d'autres types de prothèses dentaires, tels que des dentiers.

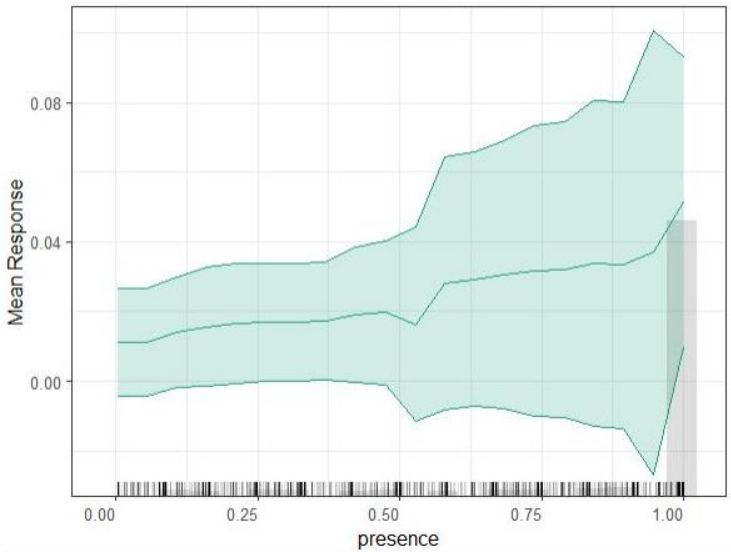


Figure 3.33 – Graphique de dépendance partielle de la variable temps de présence pour le modèle de fréquence H2O des SPR 50

Sur ce graphique, on peut voir pour le temps de présence les mêmes éléments que sur le graphique précédent, mais présentés d'une manière différente car la variable est quantitative continue cette fois et non discrète. La courbe verte centrale est la courbe de dépendance partielle en fonction du temps de présence, et les courbes vertes aux extrémités sont les bornes de l'intervalle [*moyenne ICE* $\pm$ *écart-type ICE*].

Ce graphique illustre une augmentation plus ou moins affine de la fréquence modélisée avec la présence. Cependant, on remarque un faible creux pour les bénéficiaires qui ont une exposition autour de 0,5 : cela peut être lié au faible nombre de données à ce niveau ou bien à un facteur externe (par exemple la région). C'est la limite des graphiques de dépendances partielles : on regarde l'impact marginal d'une variable.

On peut également observer une excroissance pour les bénéficiaires qui sont présents toute l'année (donc a priori ceux qui sont présents depuis plus d'un an), et qui connaissent probablement mieux leurs garanties et sont bien informés sur leur fonctionnement. Cela peut s'avérer d'autant plus vrai que les couronnes dentaires sont un acte pour lequel les bénéficiaires vérifient leur niveau de garantie avant d'aller chez le dentiste.

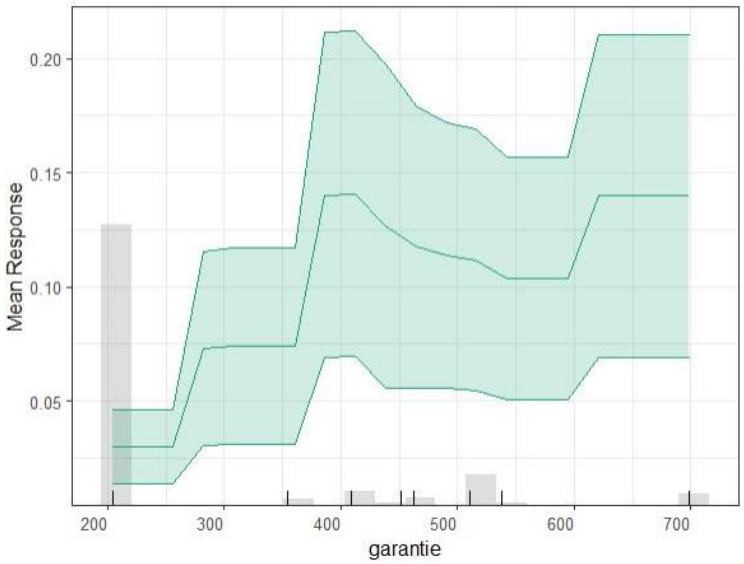


Figure 3.34 – Graphique de dépendance partielle de la variable niveau de garantie pour le modèle de fréquence H2O des SPR 50

Sur ce graphique de dépendance partielle liée à la garantie, on peut voir une augmentation de la fréquence modélisée par pallier jusqu'à 400 € environ, puis une redescente de la fréquence au niveau des garanties à hauteur de 500 €. Enfin, pour des garanties autour de 700 € la fréquence revient au niveau qu'elle avait pour des garanties à 400€. Il y a donc un creux entre 700 € et 400 €. On ne s'attendait pas forcément à ce que la fréquence augmente entre 400 € et 700 €, mais au moins à ce qu'elle se stabilise.

Cela vient des données : peut-être que les personnes qui ont cette garantie consomment de manière « atypique » (moins de besoin de couronnes dentaires ?), ou peut être qu'un autre facteur inconnu (région ?) intervient.

Hormis la baisse de la fréquence au niveau de la garantie à 500 €, ces résultats semblent assez cohérents puisque le coût moyen d'une couronne dentaire dans notre base de données est de 518 €. En dessous de 400 € de garantie, en comptant le remboursement de la Sécurité Sociale de 75 €, les bénéficiaires ont un reste à charge moyen au moins égal à 43 €, ce qui peut être conséquent et donc provoquer du renoncement aux soins qui va impacter la fréquence. Pour rappel, les données utilisées dans la conception des modèles sont des données de survenance 2018, c'est-à-dire avant la mise en place de la réforme « 100% Santé », qui a notamment impacté les prothèses dentaires. Cette réforme a pour but de diminuer le renoncement aux soins en diminuant les restes à charge des bénéficiaires qui utilisent les paniers « RAC 0 » et « RAC maîtrisé ». En revanche, pour des garanties au-delà de 400 € (soit un reste à charge moyen inférieur à 43 €), le phénomène de renoncement aux soins peut avoir tendance à disparaître et donc diminuer l'impact du montant de garantie sur la fréquence.

Etant donné que les graphiques de dépendance partielle ne montrent que l'influence marginale d'une variable sur le modèle, il peut être intéressant de compléter l'interprétation des modèles avec les valeurs de SHAP.

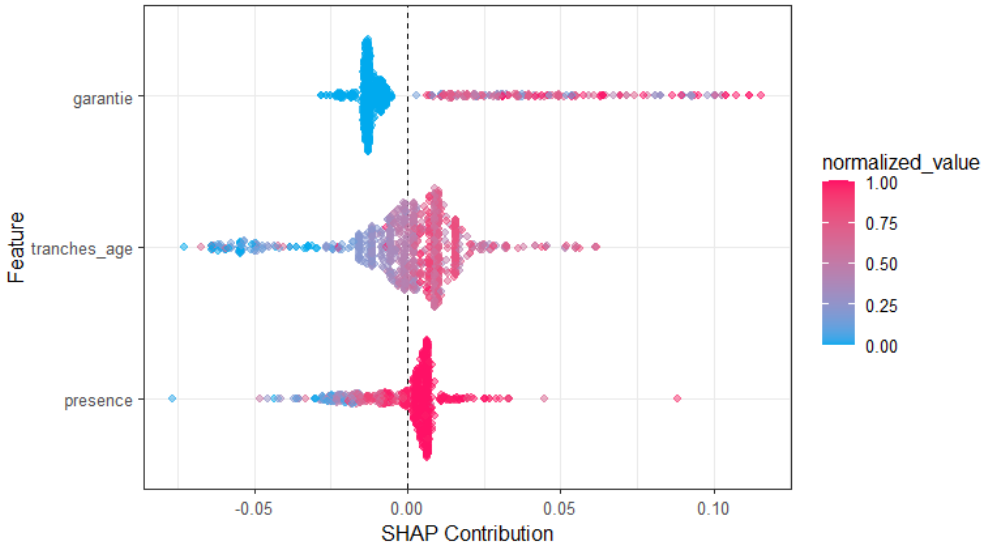


Figure 3.35 – SHAP summary plot du modèle de fréquence H2O des SPR 50

Ce graphique représente la contribution des variables sur la prédiction du modèle, au travers des valeurs de SHAP. Les variables sont classées par ordre d'importance : de la plus importante en haut, à la moins importante en bas. Pour chaque variable, un point du graphique correspond à une donnée, soit un bénéficiaire. Une épaisseur importante de point indique une forte densité de bénéficiaires. La couleur représente la valeur de la variable, qui a été normalisée : du bleu (valeur faible) au rouge (valeur élevée), en passant par le violet (valeur moyenne). Les différentes variables se lisent sur l'axe des ordonnées et les valeurs de SHAP sur l'axe des abscisses. Plus une valeur de SHAP est élevée en valeur absolue, plus la contribution à la prédiction par le modèle (ici la fréquence) est importante. Une valeur de SHAP négative aura un impact à la baisse sur la fréquence, tandis qu'une valeur de SHAP positive aura un impact à la hausse. [17]

Sur ce modèle, on peut donc voir, grâce au graphique, qu'une garantie faible diminue légèrement la fréquence modélisée (phénomène de renoncement aux soins) par rapport à la fréquence moyenne, et qu'une garantie moyenne ou élevée a tendance à augmenter plus ou moins la fréquence. Le niveau d'augmentation de fréquence est difficile à déterminer car les bénéficiaires avec des garanties moyennes et élevées sont dispersées, alors que les bénéficiaires avec des garanties faibles sont concentrées autour des valeurs de SHAP de l'ordre de -0,01.

Une tranche d'âge basse diminue la fréquence modélisée. En revanche, une tranche d'âge élevée a tendance à l'augmenter. Une tranche d'âge moyenne, quant à elle, a un faible impact (positif ou négatif), voire pas du tout d'impact sur la prédiction.

On peut également remarquer que plus le temps de présence est faible, plus son impact à la baisse sur la fréquence est fort élevé. Un temps d'exposition élevé, en revanche, augmente très légèrement la prédiction.

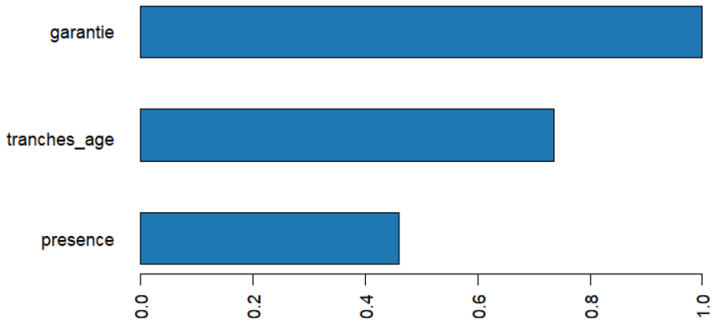


Figure 3.36 – Importance des variables pour le modèle de fréquence H2O des SPR 50

L'importance des variables a également été tracée. De même que pour le GBM paramétré manuellement, la garantie est la variable la plus importante. Les deux autres variables ont aussi une importance non négligeable. Cependant, le GBM paramétré automatiquement par H2O accorde plus d'importance à la tranche d'âge, tandis que le GBM paramétré manuellement accord une importance plus grande au temps de présence.

- *Modélisation du coût moyen*

Le coût moyen des SPR 50 est modélisée grâce à un GBM.

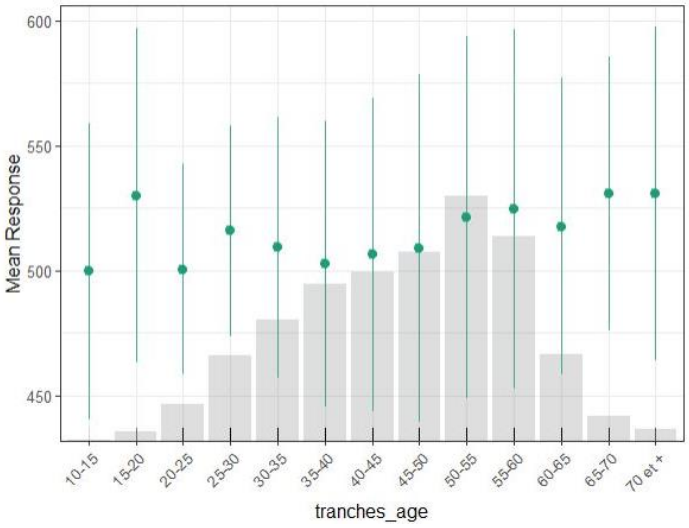


Figure 3.37 – Graphique de dépendance partielle de la variable tranche d'âge pour le modèle de coût moyen H2O des SPR 50

La courbe de dépendance partielle liée à la tranche d'âge montre que le coût moyen modélisé est assez stable en fonction de l'âge. Il varie entre 500 € et 540 €. Les coûts moyens semblent toutefois plus élevés pour les plus jeunes et les plus âgés. Les bénéficiaires jeunes vont probablement garder ces couronnes dentaires toute leur vie donc ils préfèrent peut-être investir dedans. Les bénéficiaires plus âgés ont probablement des salaires plus élevés et ont donc plus de moyens pour payer le reste à charge des couronnes dentaires.

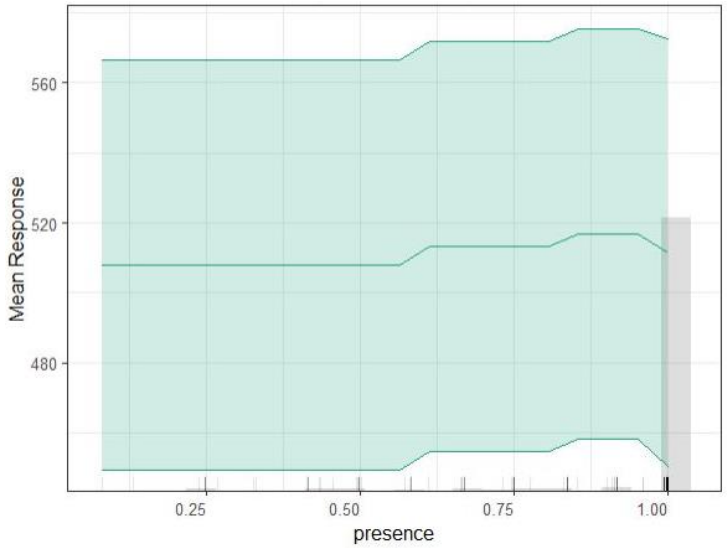


Figure 3.38 – Graphique de dépendance partielle de la variable temps de présence pour le modèle de coût moyen H2O des SPR 50

Sur ce graphique de dépendance partielle liée au temps de présence, on peut observer que celui-ci ne semble pas vraiment avoir d'influence sur la modélisation du coût moyen.

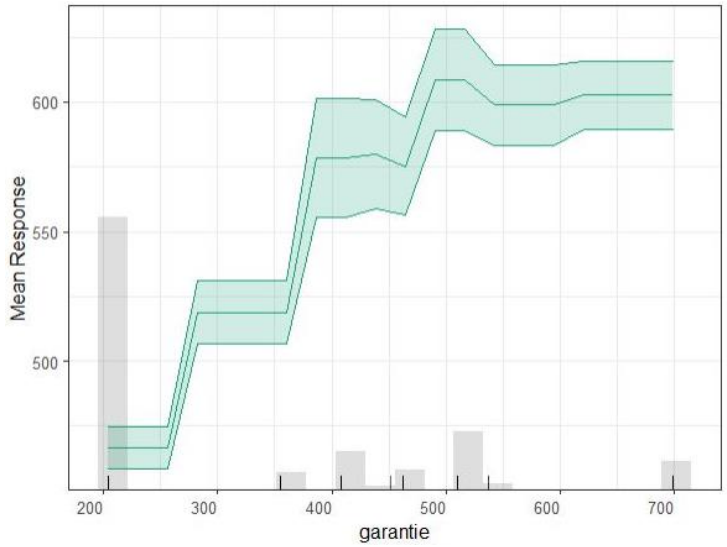


Figure 3.39 – Graphique de dépendance partielle de la variable niveau de garantie pour le modèle de coût moyen H2O des SPR 50

Cette courbe de dépendance partielle en fonction de la garantie montre clairement une augmentation du coût moyen modélisé par pallier. Un bénéficiaire avec des garanties plus élevées consomme donc des couronnes dentaires plus coûteuses, ce qui semble assez logique.

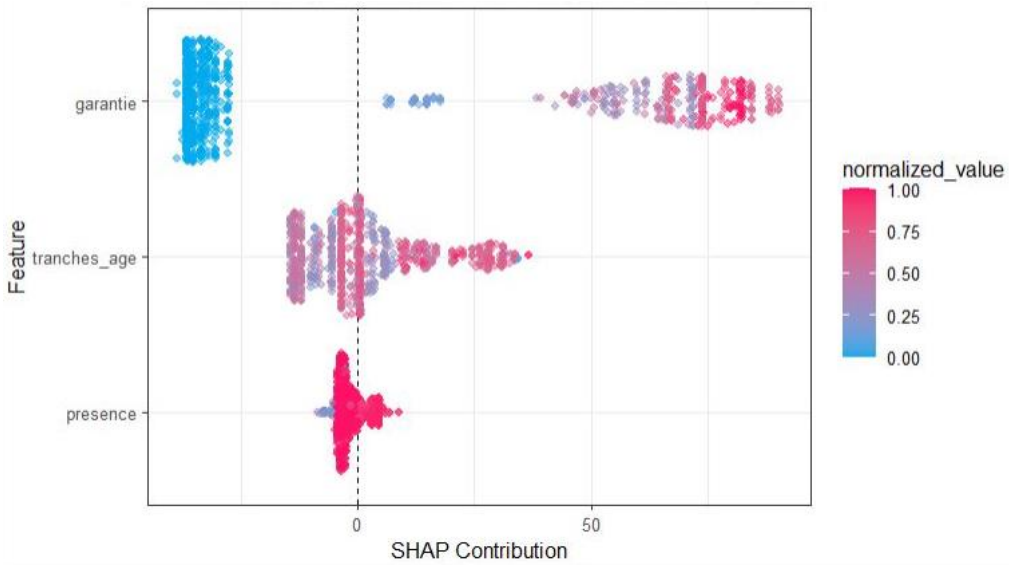


Figure 3.40 – SHAP summary plot du modèle de coût moyen H2O des SPR 50

Sur ce graphique, on observe très nettement qu'une garantie faible diminue fortement le coût moyen modélisé et qu'une garantie moyenne ou haute l'augmente fortement.

Il est plus compliqué de tirer des conclusions sur l'influence de la tranche d'âge sur la prédiction. En effet, les valeurs hautes, moyennes et basses semblent répartis aléatoirement sur l'axe des abscisses.

Concernant le temps de présence, un très faible impact négatif est constaté pour une valeur basse. Il est en revanche plus difficile de tirer des conclusions pour des valeurs hautes car leurs valeurs de SHAP sont très proches de zéro.

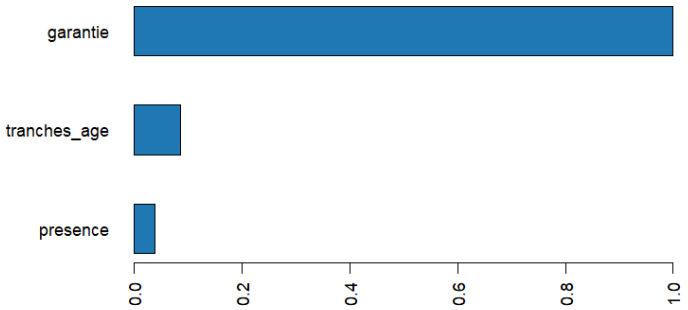


Figure 3.41 – Importance des variables pour le modèle de coût moyen H2O des SPR 50

Sur ce graphique montrant l'importance des variables du modèle, on peut voir que la garantie est la variable qui est de loin la plus importante. La tranche d'âge, puis le temps de présence viennent ensuite, avec une importance très faible. Les proportions de ce modèle semblent très similaires à celles du GBM paramétré manuellement.

Consultations et Visites Spécialistes OPTAM

Dans ce paragraphe, les valeurs de SHAP ne pourront pas être utilisées pour interpréter les modèles car ils ne sont pas basés sur des arbres. L'importance des variables ne sera pas donnée non plus car elle ne peut pas être utilisée pour des agrégations de modèles.

- *Modélisation de la fréquence*

La fréquence des Consultations et Visites Spécialistes OPTAM est modélisée grâce à une agrégation de modèles.

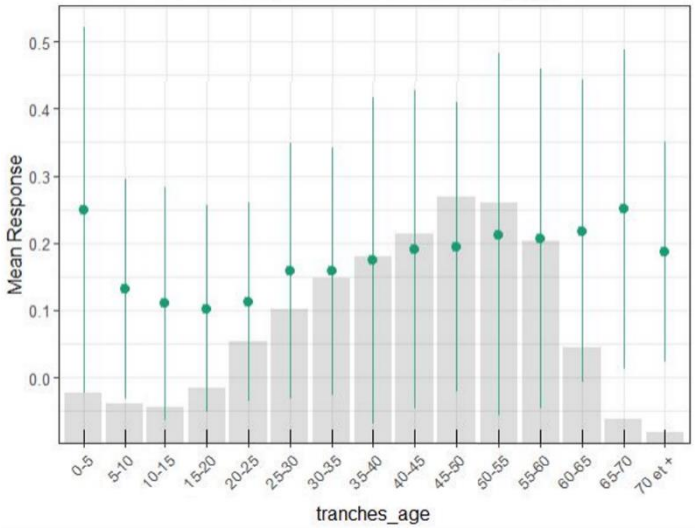


Figure 3.42 – Graphique de dépendance partielle de la variable tranche d'âge pour le modèle de fréquence H2O des Consultations et Visites Spécialistes OPTAM

Le graphique ci-dessus représente la courbe de dépendance partielle en fonction de la tranche d'âge. On peut remarquer que la fréquence modélisée est plus élevée chez les bénéficiaires jeunes et chez les plus âgés : ce sont en effet ceux qui consultent le plus les spécialistes (pédiatres, gériatres, ophtalmologistes…).

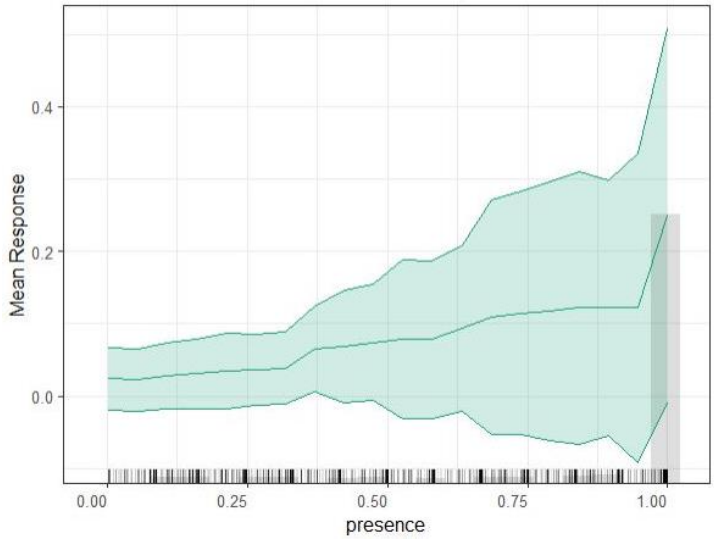


Figure 3.43 – Graphique de dépendance partielle de la variable temps de présence pour le modèle de fréquence H2O des Consultations et Visites Spécialistes OPTAM

Sur ce graphique de dépendance partielle liée au temps de présence, de même que pour les SPR 50, on constate une augmentation affine de la fréquence modélisée en fonction de la présence. On peut également voir un pic pour les bénéficiaires présents toute l'année. On peut donc aussi supposer une bonne connaissance des garanties pour ces bénéficiaires.

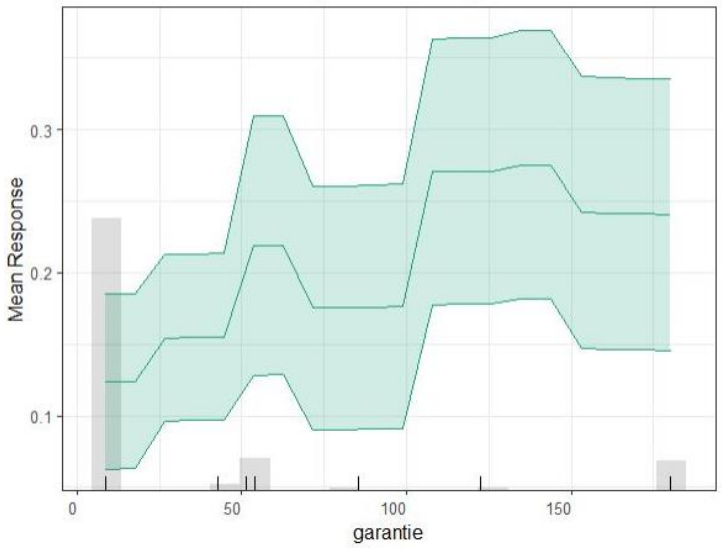


Figure 3.44 – Graphique de dépendance partielle de la variable niveau de garantie pour le modèle de fréquence H2O des Consultations et Visites Spécialistes OPTAM

Cette courbe de dépendance partielle montre globalement une augmentation de la fréquence modélisée avec la garantie. Cependant, de la même manière que pour les SPR 50, on constate des creux à certains niveaux de garanties. Cela vient encore une fois des données qui ont servi à l'apprentissage du modèle.

- *Modélisation du coût moyen*

Le coût moyen des Consultations et Visites Spécialistes OPTAM est modélisé grâce à une agrégation de modèles.

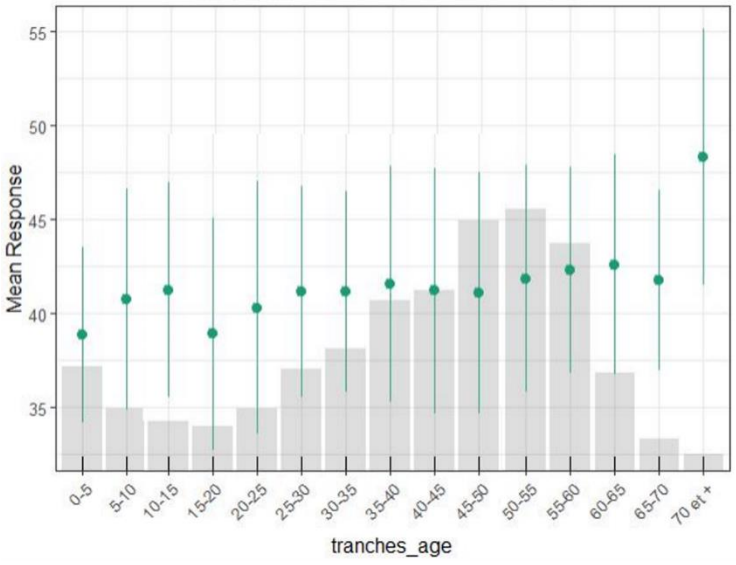


Figure 3.45 – Graphique de dépendance partielle de la variable tranche d'âge pour le modèle de coût moyen H2O des Consultations et Visites Spécialistes OPTAM

Le graphique ci-dessus représente la courbe de dépendance partielle en fonction de la tranche d'âge. Le coût moyen modélisé semble assez stable selon l'âge. Un pic est tout de même constaté pour les bénéficiaires les plus âgés. Il est probablement lié à la complexité de leurs consultations spécialistes.

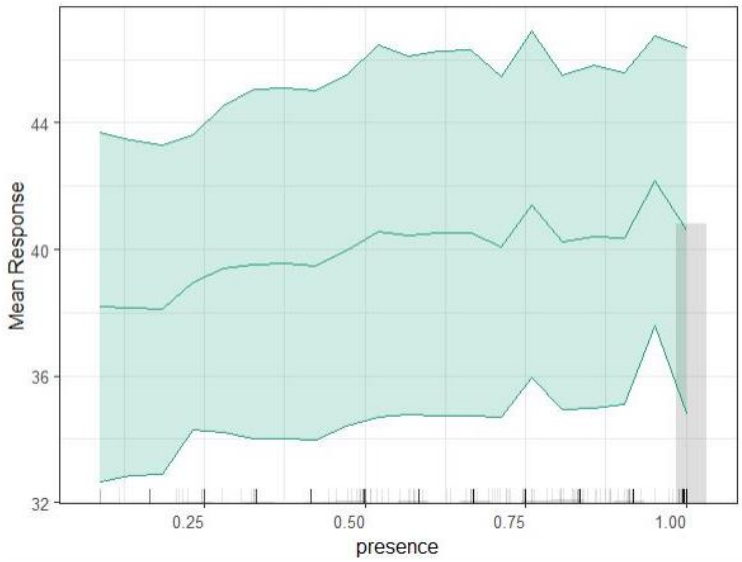


Figure 3.46 – Graphique de dépendance partielle de la variable temps de présence pour le modèle de coût moyen H2O des Consultations et Visites Spécialistes OPTAM

Cette courbe de dépendance partielle montre une légère augmentation du coût moyen modélisé en fonction du temps de présence, avec quelques pics (notamment vers 0,75 et 0,9). Cela vient des données car aucun élément ne peut réellement expliquer cette caractéristique du modèle.

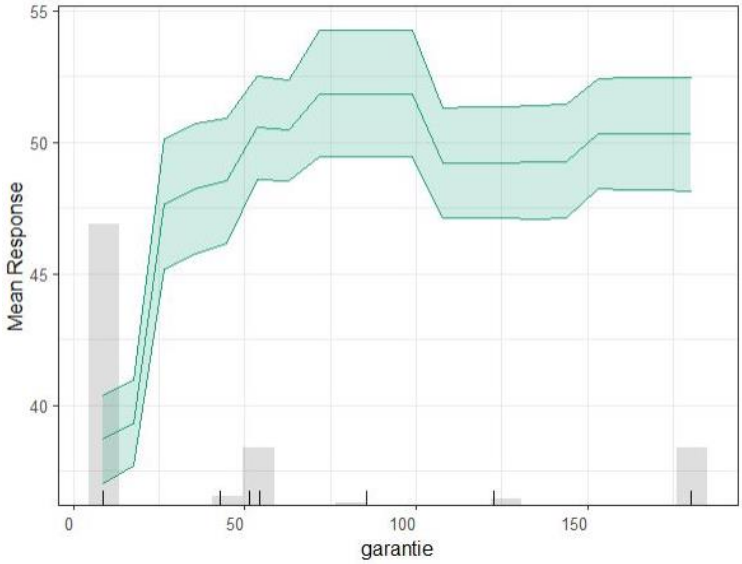


Figure 3.47 – Graphique de dépendance partielle de la variable niveau de garantie pour le modèle de coût moyen H2O des Consultations et Visites Spécialistes OPTAM

Sur ce graphique de dépendance partielle lié à la garantie, on peut observer un coût moyen très faible pour les garanties en-dessous de 50 €, et, au-delà de 50 €, on peut distinguer un palier. On remarque cependant un creux autour de 125 €, sans doute dû aux données également.

Les graphiques de dépendance partielle et les valeurs de SHAP ont permis de mieux interpréter les modèles, mais également de montrer leurs limites : les modèles sont très liés aux données d'entrées.

4 APPLICATION A L'OUTIL DE TARIFICATION

L'objectif de ce chapitre est d'expliquer comment est calculée la prime pure et comment on peut construire un outil de tarification, à partir des modèles établis ci-dessus.

4.1 CALCUL DE LA PRIME PURE

Grâce aux modèles de fréquence et de coût moyen établis pour chaque sous postes, on va pouvoir calculer la prime pure d'une population donnée. Ce calcul se fera au travers d'une fonction, qui prendra en argument les éléments suivants :

- **Une liste de bénéficiaires** : elle devra être au format Excel et contenir en colonne les informations suivantes : identifiant du bénéficiaire, âge, catégorie, type de bénéficiaire, temps de présence et structure familiale. L'âge et le temps de présence sont essentiels pour le calcul de la prime pure. Sans ces éléments, le calcul ne pourra être effectué. La catégorie, le type de bénéficiaire et la structure familiale sont facultatifs. Ces éléments permettront juste de donner quelques statistiques sur les données, mais ne seront pas utilisés dans le calcul de la prime pure. En cas d'absence d'un ou plusieurs de ces éléments, il suffit de remplir la colonne par le libellé « Donnée manquante ».
- **Des garanties à tarifer** : elles devront également être au format Excel. Les informations qu'elles devront contenir en colonne sont : le nom du grand poste, le nom du sous poste, le type de garantie, l'assiette de garantie, la valeur de la garantie, l'assiette de garantie limite et la valeur de la garantie limite. Deux autres colonnes permettront éventuellement de remplir une deuxième assiette de garantie et une deuxième valeur de garantie pour les garanties qui s'expriment de plusieurs manières. Tous les éléments sont essentiels au calcul de la prime pure. On peut remplir autant de lignes que de garanties nécessaires à tarifer. Ainsi, si la tarification ne concerne que l'ajout d'une garantie au régime, il n'y aura qu'une seule ligne de remplie pour le sous poste concerné. En revanche, si la tarification concerne un régime entier, plus de ligne seront remplies.
- **Le nom et prénom de l'utilisateur du logiciel** : il sera utilisé pour le chemin permettant d'accéder aux dossiers contenant les modèles stockés.

Un modèle de liste de bénéficiaires et de grille de garantie est proposé, afin de fournir les données sous le bon format. La fonction de calcul de la prime pure, effectuera les étapes énoncées ci-dessous :

1. **Importation des tables de correspondances** : quatre tables de correspondance sont importées. La première comporte les numéros de sous postes. Elle sert à retrouver le chemin d'un modèle pour un sous poste donné sur l'ordinateur. La seconde est celle déterminée dans le paragraphe 3.1.1., contenant la liste des sous postes tarifables. Les deux dernières sont celles sur les modèles de fréquence et de coût moyen. Elles permettent de récupérer le nom du modèle pour un numéro de sous poste donné.

2. **Initialisation du montant du PMSS et des bases de remboursement et taux de la Sécurité Sociale** : le PMSS est initialisé à 3 311 € (PMSS de l'année 2018). Les bases de remboursement et les taux sont initialisés en important la table créée

dans le paragraphe 2.2.2. Ces informations serviront plus tard pour le calcul des garanties.

3. **Test des assiettes de garanties** : un test est effectué afin de voir si les assiettes des garanties remplies en input respectent bien les formats demandés : « € », « %PMSS », « %BRSS » ou « %FR ». Si ce n'est pas le cas, un message d'erreur est envoyé à l'utilisateur, sinon on passe à l'étape d'après. Cela permet d'empêcher un bug de la fonction au moment du calcul des remboursements.

4. **Création de la colonne de tranches d'âge** : la population rentrée en input contient l'âge des bénéficiaires. Une colonne de tranches d'âge est rajoutée. A partir de l'âge, on calcule les mêmes tranches d'âge que celles utilisées pour l'apprentissage des modèles. A savoir, des tranches de 5 ans, de 0 à 70 ans, et une tranche pour ceux ayant 70 ans et plus. C'est la tranche d'âge, et non l'âge, qui sera utilisée au moment de la prédiction.

5. **Conversion des garanties en euros en complément de la Sécurité Sociale** : les garanties sont toutes converties en euros et en complément de la Sécurité Sociale, de la même manière que dans le paragraphe 2.2.3.

6. **Détermination des sous postes tarifables** : avant de se lancer dans les calculs, il faut déterminer quels sous postes peuvent être tarifer. Pour cela, une recherche est effectuée dans la table de correspondance contenant la liste des sous postes tarifables. Les sous postes qui ne sont pas tarifables sont stocké dans une liste, qui sera affichée plus tard.

7. **Création d'un tableau de données par sous poste et ajout du montant de la garantie** : cette étape est une étape d'initialisation. Elle s'apparente à l'étape précédant la jointure des données. La population est dupliquée autant de fois qu'il y a de sous postes à tarifer. Pour chaque duplication, le nom d'un sous poste différent est rajouté, ainsi que sa garantie associée calculée en étape 5, l'assiette limite de garantie et la valeur limite de garantie. Pour n sous postes à tarifer, on aura n duplications de la population initiale auxquelles ont été rajoutées des informations sur les sous postes et les garanties. Ces n duplication seront rangées dans une liste de taille n.

8. **Boucle permettant de calculer la prime pure** : cette étape est une boucle qui sera effectuée pour chaque sous poste i. Il s'agit d'une boucle « pour », elle est donc finie. Au sein de la boucle, les étapes suivantes ont lieu :

    8.1. **Sélection des données** : sur la liste de populations créée en étape 6, on sélection l'élément de la liste associé au sous poste i.

    8.2. **Conversion des données au format H2O** : de même que pour la création des fonctions permettant de modéliser la fréquence et le coût moyen, les données sont converties au format H2O, car les modèles sont des objets H2O qui ne fonctionnent qu'avec d'autres objets H2O.

8.3. **Récupération des noms des modèles** : pour la fréquence et pour le coût moyen, le nom du modèle du sous poste i est récupéré grâce au numéro de sous poste et à la table de correspondance.

8.4. **Chargement des modèles** : grâce au nom de l'utilisateur rentré en input, au nom du modèle et au numéro de sous poste, le modèle est récupéré dans le dossier où il a été stocké sur l'ordinateur, en utilisant la fonction *h2o.loadModel*.

8.5. **Prédiction de la fréquence et du coût moyen** : à partir des modèles récupérés ci-dessus, on peut prédire la fréquence grâce à la fonction *h2o.predict*. Cette fonction prend en argument un modèle et des variables explicatives, et elle renvoie la fréquence prédite à partir du modèle. Il s'agit de la fréquence pour le sous poste i, qui est calculée pour chaque bénéficiaire. L'élément de la liste qui contient les n duplications de population est modifié, par l'ajout d'une colonne « Fréquence ». Le même travail est effectué sur le coût moyen. Ainsi, on obtient pour le sous poste i, une table qui contient les colonnes suivantes :

- Identifiant du bénéficiaire
- Tranche d'âges
- Garantie maximale remboursée
- Assiette limite de garantie
- Valeur limite de garantie
- Fréquence
- Coût moyen

8.6. **Conversion des données au format classique** : la table de l'étape 8.5 est au format H2O. Elle est convertie en un format plus classique sur R, afin de faciliter sa manipulation.

9. **Agrégation des données** : tous les éléments qui composent la liste sont ensuite agrégés pour ne former plus qu'une seule table de données.

10. **Calcul des remboursements complémentaires modélisés** : il s'agit ici du calcul de la prime pure, mais qui ne tient pas compte des PSAP énoncées au paragraphe 2.1.2, ni du coefficient pour les sous postes non tarifables du paragraphe 3.1.2. Les notations suivantes sont introduites :

- $freq_{i,j}$ la fréquence prédite pour le sous poste i et pour le bénéficiaire j. Certaines fréquences prédites sont négatives, mais néanmoins très proches de zéro. C'est le cas pour les actes qui sont très peu consommé. Par la suite, on prendra plutôt $\max(freq_{i,j}, 0)$, car une fréquence ne peut pas être négative en pratique.
- $cout_{i,j}$ la fréquence prédite pour le sous poste i et pour le bénéficiaire j. De même que pour la fréquence, on prendra plutôt $\max(cout_{i,j}, 0)$, car un coût moyen ne peut pas être négatif en pratique.

- $g'_{i,€}$ la garantie exprimée en euros en complément de la Sécurité Sociale.
- $\pi^*_{j,i}$ la prime pure hors PSAP et sous postes non tarifables pour le sous poste i et pour le bénéficiaire j.
- $g_{i,lim}$ le nombre maximum de remboursements de la garantie pour le sous poste i, dans le cas où l'assiette de garantie limite est par an.
- $g'_{i,lim,€}$ la garantie limite exprimée en euros en complément de la Sécurité Sociale, dans le cas où sont assiette n'est pas par an.

Trois cas vont alors être distingués en fonction de l'assiette de la garantie limite :

→ S'il n'y a pas de garantie limite, alors

$$\pi^*_{i,j} = \max(freq_{i,j}, 0) \times \min(\max(cout_{i,j}, 0), g'_{i,€})$$

→ Si l'assiette de garantie limite est par an, alors

$$\pi^*_{i,j} = \min(\max(freq_{i,j}, 0), g_{i,lim}) \times \min(\max(cout_{i,j}, 0), g'_{i,€})$$

→ Sinon $\pi^*_{i,j} = \min(max(freq_{i,j}, 0) \times min(max(cout_{i,j}, 0), g'_{i,€}), g'_{i,lim,€})$

11. **Intégration des PSAP et du coefficient des sous postes non tarifables** : la prime pure calculée n'inclue pas les PSAP, ni la part des sous postes non tarifables. En les prenant en compte, la prime pure pour le sous poste i et pour le bénéficiaire j devient : $\pi_{i,j} = \pi^*_{i,j} \times (1 + q_{nt}) \times (1 + t_{PSAP})$, avec, pour rappel, $q_{nt} = 1{,}1\%$ et $t_{PSAP} = 3{,}11\%$. En calculant la prime pure de cette façon, on sous-entend que tous les sous postes ont les mêmes taux de PSAP.

À la suite de toutes ces étapes, la fonction renvoie en sortie une liste de bénéficiaires avec pour chaque sous poste leur prime pure. Avant d'utiliser cette fonction, il ne faut pas oublier d'ouvrir une session H2O, et de la refermer après utilisation.

4.2 CALCUL DES MONTANTS DE COTISATION EN FONCTION DES STRUCTURES DE COTISATION

Nous allons voir ici comment sont calculés les montants de cotisations (ou primes commerciales), à partir de la prime pure et en fonction de la structure de cotisation. Pour passer de la prime pure à la prime commerciale, il faut prendre en compte les taxes et les chargements futurs. Pour la suite, on considérera que les contrats de frais de santé sont responsables. Le taux de taxes sera donc de 13,27 %. On considérera également un taux de chargements unique pour tous les bénéficiaires de θ, ainsi qu'un taux de FMT de 0,80 %.

Il existe plusieurs types de structures de cotisation pour les entreprises qui ont souscrit à un contrat collectif à adhésion obligatoire. Quatre d'entre eux seront détaillés ci-dessous ; ce sont les plus courantes chez les clients d'Adding. Mais il en existe d'autres.

4.2.1 Structure unique « Famille »

Dans cette structure de cotisation, tous les salariés payent la même cotisation, quelle que soit la composition familiale. La cotisation couvre le salarié et tous les éventuels membres de la famille à la charge du salarié : le conjoint, les enfants et les autres ayants droit. Le conjoint peut éventuellement déjà être couvert par ailleurs. Dans ce cas, il bénéficiera d'une double couverture (le remboursement cumulé y compris Sécurité Sociale ne pourra pas excéder le montant des frais réels). La structure unique « Famille » instaure une solidarité entre les salariés de l'entreprise, mais peut également paraître désavantageux pour les salariés célibataires et sans enfants.

Pour calculer le montant de la cotisation, il faut d'abord calculer la prime pure globale de chaque assuré π_j, en sommant la prime pure des n sous postes à tarifer. La prime pure totale π, s'obtient en sommant les primes pures globales de chaque assuré pour les m assurés. On obtient donc :

$$\pi = \sum_{j=1}^{m} \pi_j \text{ , avec } \pi_j = \sum_{i=1}^{n} \pi_{i,j}$$

Le montant de la cotisation unique « Famille » mensuelle $c^{unique\ fam}$, pour un régime à l'équilibre, c'est-à-dire avec un S/P égal à 100 %, s'écrit de la manière suivante :

$$c^{unique\ fam} = \frac{\pi\ (1 + 13{,}27\ \%)}{12\ m_{cotisants}\ (1 - \theta - 0{,}80\ \%)},$$

avec $m_{cotisants}$ le nombre d'assurés qui cotisent.

4.2.2 Structure « Isolé / Famille »

Cette structure de cotisation permet de faire la distinction entre les salariés célibataires et sans enfants, avec les salariés avec conjoint et / ou enfant(s) à charge. Le salarié seul souscrit alors la cotisation « Isolé », et les autres salariés souscrivent la cotisation « Famille ». De même que pour un régime de structure unique « Famille », les conjoints qui sont déjà couverts bénéficieront d'une double couverture. La structure « Isolé/Famille » permet de trouver une forme d'équilibre entre cotisation individuelle et solidarité entre les salariés, sans pour autant désavantager les salariés seuls. Cette structure de cotisation est très utilisée chez les actifs mais rarement chez les inactifs.

Afin de calculer les deux montants de cotisations, il faut d'abord distinguer la prime pure des assurés isolés, notée π^{iso}, de la prime pure des assurés avec conjoint et / ou enfant(s), de leurs conjoints et de leurs enfants, notée π^{fam}, telles que $\pi = \pi^{iso} + \pi^{fam}$.

Dans les structures « Isolé / Famille », les assurés isolés sont la plupart du temps solidaires avec les familles, c'est-à-dire qu'ils ne payent pas uniquement ce qu'ils consomment, mais ils payent également une partie de la consommation des familles. Il est donc important d'introduire un coefficient de solidarité x^{fam}, compris entre 0 et 1 (1 étant la valeur maximale de solidarité, où les familles auraient une cotisation nulle, et 0 la valeur pour laquelle les familles payeraient la totalité de ce qu'elles consomment). La cotisation mensuelle « Famille » c_{fam} se calcule donc ainsi :

$$c^{fam} = \frac{\pi^{fam}\,(1 + 13{,}27\,\%)\,(1 - x^{fam})}{12\,m^{fam}_{cotisants}\,(1 - \theta - 0{,}80\,\%)},$$

avec $m^{fam}_{cotisants}$ le nombre d'assurés qui cotisent et qui ont une famille qui bénéficie du régime.

La cotisation mensuelle « Isolé » c^{iso} est alors égale à :

$$c^{iso} = \frac{(\pi^{iso} + \pi^{fam} \times x^{fam})\,(1 + 13{,}27\,\%)}{12\,m^{iso}_{cotisants}\,(1 - \theta - 0{,}80\,\%)},$$

avec $m^{iso}_{cotisants}$ le nombre d'assurés isolés qui cotisent, et tel que $m_{cotisants} = m^{iso}_{cotisants} + m^{fam}_{cotisants}$.

4.2.3 Structure « Assuré + Enfant(s) / Conjoint facultatif »

Dans cette structure de cotisation, tous les assurés souscrivent à la cotisation « Assuré + Enfant(s) », qu'ils aient ou non des enfants à charge. Pour ceux qui ont des conjoints, ils ont la possibilité de souscrire à la cotisation « Conjoint facultatif », qui est, comme son nom l'indique, facultative. Cette cotisation est entièrement à la charge de l'assuré.

Le calcul de ces cotisations ressemble à celui de la structure « Isolé / Famille ». Il faut tout d'abord distinguer la consommation des conjoints, notée π^{conj}, de celle des assurés et enfants $\pi^{ass+enf}$, telles que $\pi = \pi^{conj} + \pi^{ass+enf}$. Il y a également une solidarité des assurés avec ou sans enfant(s) auprès des conjoints. Le coefficient de solidarité est noté x^{conj}. Contrairement à la cotisation « Famille » qui est obligatoire, la cotisation des conjoints est facultative et cela entraîne de l'antisélection. Pour pallier cela, un coefficient d'antisélection de 15% est ajouté. Ce coefficient a été calculé grâce à des données de portefeuille, en comparant la consommation de conjoints qui ont adhéré de manière facultative à la consommation de notre base de données. Ce coefficient signifie qu'en moyenne, les conjoints qui adhèrent de manière facultative consomment 15% de plus par rapport à la prime pure calculée.

Le nombre de conjoints cotisants et le nombre d'assurés cotisants sont respectivement notés $m^{conj}_{cotisants}$ et $m^{ass+enf}_{cotisants}$, tel que $m^{ass+enf}_{cotisants} = m_{cotisants}$.

La cotisation mensuelle « Conjoint facultatif » s'écrit donc :

$$c^{conj} = \frac{\pi^{conj}\,p\,(1 + 15\%)\,(1 - x^{conj})\,(1 + 13{,}27\,\%)}{12\,m^{conj}_{cotisants}\,(1 - \theta - 0{,}80\,\%)}$$

Etant donné que l'adhésion des conjoints est facultative, tous ne vont pas adhérer au régime. Seule une proportion p va adhérer et donc consommer. La cotisation mensuelle « Assuré + Enfant(s) » devient donc :

$$c^{ass+enf} = \frac{(\pi^{ass+enf} + \pi^{conj} \times x^{conj}\ (1 + 15\%)\ p)\ (1 + 13{,}27\ \%)}{12\ m_{cotisants}^{ass+enf}\ (1 - \theta - 0{,}80\ \%)}$$

4.2.4 Structure « Adulte / Enfant »

La spécificité de cette structure de cotisations est qu'il existe un tarif pour les adultes et un tarif pour les enfants. Chaque membre de la famille couvert paye donc une cotisation. Ainsi, une famille composée d'un assuré, de son conjoint et de deux enfants, payera deux cotisations « Adulte » et deux cotisations « Enfant ». Cette structure de cotisation permet d'adapter le montant de la cotisation en fonction de la structure familiale. Dans le cadre de ce mémoire, elle ne sera utilisée que pour les inactifs (retraités et autres inactifs) car Adding réalise ce type tarification uniquement pour des inactifs.

Dans ce cadre, l'adhésion des assurés, des conjoints et des enfants est facultative. Etant donné que l'adhésion des inactifs est toujours facultative, il n'est pas nécessaire d'introduire une proportion d'adhérent. En effet, on considérera que les proportions d'assurés, de conjoints et d'enfants resteront identiques malgré le fait que les garanties et les tarifs changent, puisque nous avons déjà à faire à des adhésions facultatives.

De même que pour les deux structures de cotisations précédentes, il peut y avoir une solidarité des adultes envers les enfants, qui sera représentée par le coefficient de solidarité x^{enf}. On distingue également la consommation des adultes (assurés et conjoints), π^{ad}, de celle des enfants, π^{enf}, avec $\pi = \pi^{ad} + \pi^{enf}$. Le nombre d'adultes cotisants (assurés et conjoints) et le nombre d'enfants cotisants sont respectivement notés $m_{cotisants}^{ad}$ et $m_{cotisants}^{enf}$.

Par ailleurs, comme vu au chapitre 2.4.3., la tranche d'âge et la catégorie sont très corrélées. En effet, la majeure partie de la population inactive est composée de retraités et de préretraités, et donc de personnes plus âgées. On peut donc considérer que dans la base de données ayant permis la réalisation des modèles, les personnes les plus âgées sont des inactifs, qui sont donc des adhérents facultatifs. Il n'y a donc pas besoin de rajouter de coefficient d'antisélection lors d'une tarification d'inactifs, puisque celle-ci se base sur une population d'adhérents facultatifs qui consomme déjà plus en moyenne qu'une population identique mais à adhésion obligatoire.

La cotisation mensuelle « Enfant » se calcule de la manière suivante :

$$c^{enf} = \frac{\pi^{enf}\ (1 + 13{,}27\ \%)\ (1 - x^{enf})}{12\ m_{cotisants}^{enf}\ (1 - \theta - 0{,}80\ \%)}$$

Et la cotisation mensuelle « Adulte » s'écrit :

$$c^{ad} = \frac{(\pi^{ad} + \pi^{enf} \times x^{enf})\ (1 + 13{,}27\ \%)}{12\ m_{cotisants}^{ad}\ (1 - \theta - 0{,}80\ \%)}$$

4.2.5 Calcul du budget annuel employeur

Les cotisations des régimes de frais de Santé obligatoires sont financées en partie par l'employeur. Il prend en charge une part y des cotisations pour tous les assurés actifs. Les assurés financent l'autre partie. Pour déterminer le budget annuel employeur $b_{employeur}$, il suffit de multiplier le montant de la cotisation mensuelle par 12 pour obtenir une cotisation annuelle, puis de multiplier cela par le nombre de cotisants m et par la part du budget employeur y et additionner les différents types de cotisations i :

$$b_{employeur} = \sum_{i} 12 \times y_i \times c_i \times m_i$$

4.3 MISE EN FORME DE L'OUTIL

Pour l'étape de mise en forme de l'outil, le package *rshinny* a été utilisé. Il permet de réaliser des interfaces dans lesquelles on peut insérer des graphiques, du texte, des éléments permettant d'interagir avec l'utilisateur (boutons, espace d'écriture…) … Grâce à cela, l'outil de tarification pourra être utilisé par des personnes qui ne maîtrisent pas forcément le langage R.

L'outil créé contient une colonne de paramètres, ainsi que 6 onglets distincts, qui sont détaillés ci-dessous.

Colonne de paramètres

Cette colonne permet de sélectionner le fichier sur lequel on travaille. Il permet également de rentrer les paramètres suivants :

- Le taux de chargement du régime
- Le S/P visé, qui permet de piloter le régime
- Le taux de participation employeur, qui va permettre le calcul du budget employeur

Onglet « Initialisation »

Cet onglet donne les principales indications sur la manière d'utiliser l'outil : remplissage de l'input et paramètres à modifier. Il indique aussi les spécialités de médecine alternative qui peuvent être tarifées. Dans cet onglet, l'utilisateur doit également renseigner des indications pour obtenir le chemin du dossier où sont rangés les modèles.

Onglet « Statistiques population »

Cet onglet donne quelques statistiques sur la population pour laquelle la tarification est réalisée. Les statistiques portent principalement le nombre de bénéficiaires et sur l'âge moyen. Elles sont présentées par catégorie ou par type de bénéficiaires, sous forme de tableaux ou de graphiques.

Onglet « Coût d'un bénéficiaire »

Cet onglet renvoie la prime pure calculée par bénéficiaire, globale et détaillée par sous poste. Il renvoie également la prime pure chargée et taxée. Une liste des sous postes qui n'ont pas pu être tarifés est également donnée.

On peut aussi voir que l'onglet propose un graphique de prime pure par type de bénéficiaire, et un par catégorie.

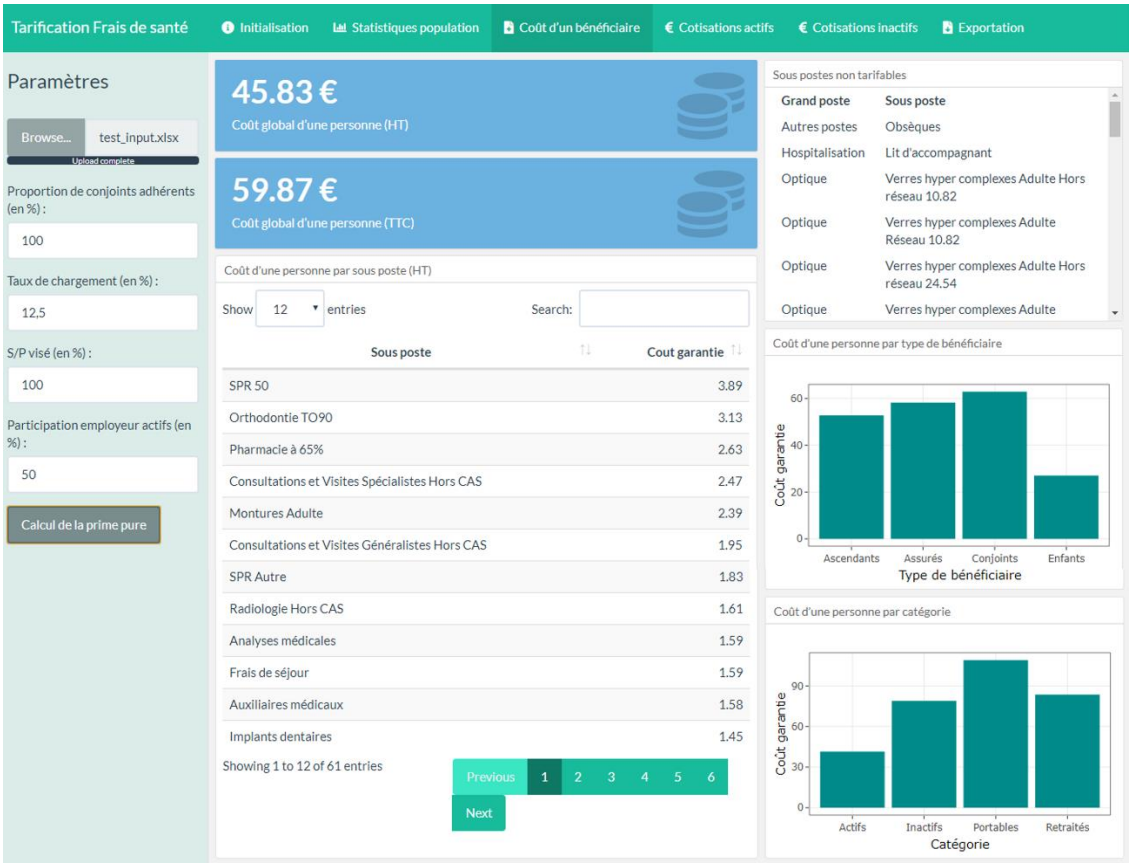


Figure 4.1 – Onglet « Coût d'un bénéficiaire » de l'outil de tarification réalisé

Onglet « Cotisations actifs »

Cet onglet permet le calcul des cotisations, pour le type de structure de cotisations souhaitée. Trois types de structures de cotisations sont proposés :

- Structure unique « Famille »
- Structure « Isolé / Famille »
- Structure « Adulte + Enfant(s) / Conjoint facultatif »

L'utilisateur a le choix de la structure. Il pourra ensuite, au sein de chaque structure, choisir des paramètres influents sur le calcul des cotisations. Il s'agit de la proportion de conjoints adhérents et de la minoration de la cotisation Famille ou Conjoint facultatif, dans le cas où il y aurait plusieurs types de cotisations. Dans cet onglet, l'utilisateur trouveras également le budget employeur pour chaque structure de cotisations.

Onglet « Cotisations inactifs »

Cet onglet est très similaire à celui des actifs. Toutefois, deux éléments diffèrent. Tout d'abord, il n'y a pas de budget employeur car ce dernier n'intervient que pour les actifs. Ensuite, les structures de cotisations ne sont pas toutes les mêmes que pour les actifs. On retrouvera la structure unique « Famille » et la structure « Adulte + Enfant(s) / Conjoint facultatif ». La structure « Isolé / Famille » n'étant pas très fréquente chez les inactifs, elle est remplacée par la structure « Adulte / Enfant ».

Onglet « Exportations »

Ce dernier onglet permet d'exporter les résultats au format Excel, afin qu'ils soient utilisés plus tard.

4.4 EXEMPLE D'APPLICATION ET COMPARAISON AVEC LA REALITE

L'outil est ensuite testé sur un client qui n'a pas servi à l'apprentissage des modèles, pour évaluer sa performance sur un cas concret. En entrée, on a une liste de bénéficiaires contenant leur identifiant, leur âge et leur temps de présence sur l'année. On a également une grille de garantie d'un régime complet.

Les résultats par grand poste et le résultat global sont retranscrits ci-dessous. Cette fois, la comparaison se fait sur la consommation dans son ensemble et non en séparant la fréquence et le coût moyen.

Grand poste	Consommation réelle totale	Consommation estimée totale	Consommation réelle par bénéficiaire	Consommation estimée par bénéficiaire	Ecart
Autres postes	120 779 €	184 736 €	19,88 €	30,41 €	53,0%
Consultations et Visites	357 077 €	404 150 €	58,77 €	66,52 €	13,2%
Dentaire	932 591 €	988 310 €	153,49 €	162,66 €	6,0%
Hospitalisation	529 719 €	374 678 €	87,19 €	61,67 €	-29,3%
Optique	604 736 €	538 621 €	99,53 €	88,65 €	-10,9%
Pharmacie	322 975 €	275 783 €	53,16 €	45,39 €	-14,6%
Soins courants	528 356 €	538 748 €	86,96 €	88,67 €	2,0%
Autres	34 832 €	36 355 €	5,73 €	5,98 €	4,4%
Total	**3 431 065 €**	**3 341 381 €**	**564,71 €**	**549,95 €**	**-2,6%**

Tableau 4.1 – Comparaison de la consommation réelle et estimée par grand poste

Globalement, les modèles prédisent un résultat qui n'est pas trop loin de la réalité, puisque l'écart constaté est de -2,6%. En revanche, lorsque l'on regarde de plus près, les résultats par grand postes semblent hétérogènes, et les écarts semblent finalement se compenser. Les soins courants et les actes dentaires sont ceux qui ont le moins d'écarts. Les actes d'hospitalisation et les autres postes sont ceux qui ont les écarts les plus marqués. Par ailleurs, le coefficient des sous postes non tarifables semble bien calibrés, puisqu'un écart de 4,4% est constaté.

Il faut également regarder les résultats sur une maille encore plus fine, afin de voir concrètement quels sont les modèles qui ont le mieux fonctionné. Les résultats par sous postes se trouvent en annexe 12.

Une fois de plus, on constate des écarts très différents en fonction des sous postes, qui se compensent pour donner un écart global faible. Plus spécifiquement, les résultats pour les deux sous postes étudiés en détail tout au long de ce mémoire sont les suivants :

Grand poste	Sous poste	Consommation réelle par bénéficiaire	Consommation estimée par bénéficiaire	Ecart
Consultations et Visites	Consultations et Visites Spécialistes OPTAM	7,59 €	6,47 €	-14,7%
Dentaire	SPR 50	44,33 €	46,67 €	5,3%

Tableau 4.2 – Comparaison de la consommation réelle et estimée par sous poste

On constate un écart modéré de -14,7% pour les Consultations et Visites Spécialistes OPTAM. Dans la partie 3.4, nous avions constaté que le modèle de fréquence de ce sous poste ne semblait pas être le plus adapté, et que le modèle du coût moyen était le

meilleur parmi ceux testés. En regardant en détail les modélisations de fréquence et de coût moyen, on s'aperçoit qu'effectivement le coût moyen a été mieux modélisé que la fréquence : l'écart sur la fréquence globale est de -19,2%, alors que celui sur le coût moyen n'est que de 5,5%.

On peut voir que l'écart entre la consommation réelle des SPR 50 et la consommation estimée sur ce sous poste est assez bon ; en effet, il est de 5,3%. Dans la partie 3.4, nous avions constaté que les modèles de fréquence et de coût moyen des SPR 50 étaient plutôt performants, même s'ils n'étaient pas les meilleurs. Cet exemple semble donc confirmer cette affirmation.

En conclusion, nous pouvons voir que même si le résultat global est bon, les résultats individuels ne sont pas tous satisfaisants, car des écarts importants sont constatés sur certains sous postes. Certains points doivent donc être améliorés avant que l'outil puisse être utilisé.

4.5 PISTES D'AMELIORATION

Afin d'améliorer l'outil, plusieurs pistes peuvent être explorées. On distingue principalement trois types d'amélioration :

- **Elargissement de la base de données** : les modèles étant très dépendants de la base de données, il peut être intéressant de la modifier pour impacter les modèles. Agrandir la base de données avec d'autres clients pourrait permettre de la diversifier. Une autre solution intéressante pourrait être de trouver des clients pour lesquels nous aurions plus de variables à notre disposition. Des variables telles que le département, le genre, le salaire… peuvent exercer une influence importante sur la manière de consommer des bénéficiaires.

- **Mise à jour des données** : les données utilisées pour créer les modèles sont celles de l'année de survenance 2018. Depuis, les tarifs ont, pour certains postes, augmenté et les bases de remboursements ont également évolué. De plus, la réforme 100% Santé est également venue bouleverser les bases de remboursement et les plafonds limites de vente sur certains postes dentaires, d'optique et d'audiologie. Utiliser des données d'une survenance plus récente permettrait d'intégrer ces évolutions.

- **Modification des paramètres d'apprentissage** : outre la modification de la base de données, l'apprentissage des modèles peut aussi être amélioré. Pour cela, on peut modifier des paramètres de la fonction *h2o.automl*. On peut jouer, par exemple, sur la durée maximale de l'apprentissage (comme cela a été évoqué au paragraphe 3.4), sur le nombre d'algorithmes à tester, restreindre l'apprentissage à certaines familles d'algorithmes… Cela pourrait permettre de trouver de meilleurs modèles.

CONCLUSION

L'objectif de ce mémoire est de réaliser un outil de tarification Santé, grâce à des modèles de *machine learning*. La tarification réalisée se base sur une approche « Fréquence / Coût moyen » individualisée et mise en œuvre poste par poste.

Une analyse détaillée des données à notre disposition nous a conduit à prendre en compte, pour l'apprentissage des modèles, les variables explicatives suivantes : tranche d'âge des bénéficiaires, temps d'exposition et niveau de garantie. En effet, ce sont les variables qui impactent le plus la consommation en Santé.

L'objectif de ce mémoire est de produire une tarification aussi pertinente que possible, en tenant compte des particularités de chacun des sous postes. Une modélisation différente a ainsi été appliquée pour modéliser les fréquences et cout moyens de chaque sous postes. Dans un premier temps, un paramétrage manuel des modèles a été effectué sur deux exemples de postes (consultations spécialistes et couronnes dentaires) ; il s'est révélé être très chronophage s'il doit être appliqué à l'ensemble des sous postes. La fonction *h2o.automl*, proposée par H2O, semblait alors être une bonne solution. En effet, cette fonction est capable de tester plusieurs familles de modèles (forêts aléatoires, GBM, réseaux de neurones, agrégations de modèles…), avec plusieurs jeux de paramètres, et de déterminer le meilleur modèle parmi ceux testés.

Les modèles des fréquences et des coûts moyens de chaque sous poste ont donc été déterminés grâce à la fonction *h2o.automl*. Les performances des modèles ainsi obtenus ont ensuite été analysées, et comparées avec celles des modèles paramétrés manuellement pour les deux sous postes concernés, grâce aux RMSE et aux MAE.

Que ce soit pour la modélisation de la fréquence ou du coût moyen, il s'est avéré qu'environ un tiers des modèles produits pouvaient être a priori considérés comme performants. Cependant, une partie non négligeable des modèles ne semblent pas assez satisfaisants et nécessitent d'être améliorés. Un test effectué sur les données d'un client a permis de confirmer cela, en comparant les prédictions de nos modèles avec la consommation réelle des bénéficiaires : bien que le résultat global fût bon, des écarts étaient constatés sur certains des sous postes pris individuellement.

Afin de compléter notre analyse, nous avons interprété les modèles produits par H2O grâce aux graphiques de dépendance partielle et aux valeurs de SHAP. L'interprétabilité nous a permis de mieux comprendre l'influence des variables sur les modèles, et de constater que ces derniers étaient étroitement liés aux données.

Ainsi, H2O permet de produire de nombreux modèles de manière automatisée mais la procédure n'est pas parfaite puisque tous les modèles ne sont pas performants. Une amélioration de la procédure visant à faire tourner la fonction *h2o.automl* de plus en plus longtemps sur un nombre de sous postes de plus en plus restreint permettrait de rendre la tarification plus performante globalement, sans pour autant être aussi chronophage qu'un paramétrage manuel sous poste par sous poste. La puissance et la performance d'H2O s'avère donc être utile pour réaliser ce genre de tarificateur mais il ne faut pas se contenter des premiers résultats.

Par ailleurs, afin d'améliorer les modèles, l'enrichissement de la base de données peut se révéler être une option intéressante.

Suite à l'intégration d'Adding dans le Groupe Sciaci Saint Honoré, une base de données, comprenant environ un million de bénéficiaires, a été mise à notre disposition. Cette

base semble être une bonne solution pour améliorer les modèles, puisqu'elle contient une quantité de données bien plus grande que celle que nous avons utilisée dans ce mémoire. De plus, elle permettra d'intégrer le département et le genre dans les variables explicatives, ce qui va peut-être pouvoir aider à affiner les modèles.

BIBLIOGRAPHIE

[1] «Site officiel de Siaci Saint Honoré,» [En ligne]. Available: https://www.s2hgroup.com/fr/accueil.html.

[2] «Site officiel d'Adding,» [En ligne]. Available: https://www.adding.fr/.

[3] V. Benoit, *Cours Prévoyance collective,* ISFA, 2021.

[4] *Loi Accord National Interprofessionnel (ANI),* 2013.

[5] «Site officiel d'h2o,» 2021. [En ligne]. Available: https://www.h2o.ai/fr/.

[6] «AutoML: Automatic Machine Learning,» [En ligne]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

[7] X. Milhaud, *Cours Data Science,* ISFA, 2021.

[8] S. Mallat, «Sciences des données / L'apprentissage par réseaux de neurones profonds,» Collège de France, 2019.

[9] S. Otto, «How to normalize the RMSE,» 2019. [En ligne]. Available: https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/.

[10] P. Piette, *Cours Introduction à l'Apprentissage Statistique - Interprétabilité,* ISFA, 2022.

[11] D. Delcaillau, A. Ly, F. Vermet et A. Papp, «Interprétabilité des modèles : Etat des lieux des méthodes et applications à l'assurance,» 2020.

[12] T. S. Nicolas Peltre, «Les valeurs de Shapley en intelligibilité locale des modèles,» 2021. [En ligne]. Available: https://www.quantmetry.com/blog/valeurs-de-shapley/.

[13] S. Lundberg et S.-I. Lee, «A unified appoach to interpreting model predictions,» 2017.

[14] S. Lundberg et S. Lee, «Consistent Individualized Feature Attribution for Tree Ensemble,» 2019.

[15] J. Cohen, Statistical power analysis for the behavioral sciences (2nd ed.), 1988.

[16] T. Mitsa, «How Do You Know You Have Enough Training Data?,» 2019. [En ligne]. Available: https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee.

[17] A. Léon, «INTERPRÉTABILITÉ DES MODÈLES DE MACHINE LEARNING,» 2020. [En ligne]. Available: https://www.aquiladata.fr/insights/interpretabilite-des-modeles-de-machine-learning/.

[18] S. Sandoval, A. Momplot et L. Bacot, «Critères & indicateurs d'auto-évaluation des modèles dans le cadre de l'autosurveillance réglementaire des systèmes

d'assainissement pour permettre au maître d'ouvrage de juger son modèle,» Lyon, 2018.

[19] T. Lagos, Mémoire : *Etude des méthodes innovantes de machine learning permettant la tarification de produits Santé en mobilité internationale*, 2018.

[20] J.-M. Aouizerate, Mémoire : *Alternative neuronale en tarification Santé*, 2010.

[21] D. Henoc Akaffou, Mémoire : *Méthode alternative de tarification Santé : GLM/XGBoost*, 2020.

[22] C. H. Karamoko Fofana, Mémoire : *Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage*, 2017.

Annexe 1 : Validation croisée

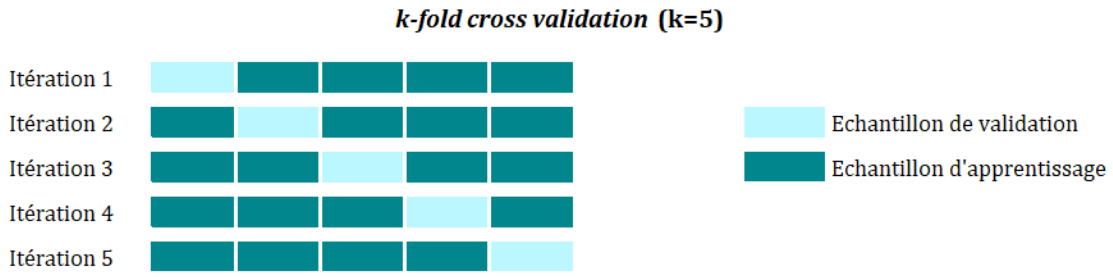
La validation croisée (ou *cross validation*) est une méthode qui permet de tester la performance d'un modèle, fondée sur des techniques d'échantillonnage. Elle permet de s'affranchir du découpage de l'échantillon initial, en échantillon d'apprentissage et échantillon de validation. Cela est très utile lorsque l'échantillon est de petite taille. La validation croisée permet également d'obtenir une estimation plus robuste de la performance du modèle. Il existe plusieurs types de méthodes de validation croisée :

- La *k-fold cross validation*
- La *leave-one-out cross validation*
- La *leave-v-out cross validation*

La méthode qui sera détaillée ci-dessous est la *k-fold cross validation* ; c'est celle qui est utilisée par la fonction *auto_ml* de H2O.

On considère un modèle entraîné sur un échantillon donné. Cet échantillon est ensuite divisé en k sous-échantillons de taille égale. Un des k sous-échantillons est sélectionné comme ensemble de validation. Les $k-1$ autres constitueront l'échantillon d'apprentissage. Le modèle est ensuite entraîné sur ce nouvel échantillon. Après apprentissage, la performance du modèle est calculée grâce à l'échantillon de validation.

Pour calculer la performance d'un modèle, plusieurs indicateurs peuvent être utilisés : le RMSE, le MAE, le MSE… C'est le RMSE qui est utilisé par la fonction *auto_ml* de H2O.



La performance du modèle est gardée en mémoire. Puis, l'étape est répétée $k-1$ fois, où tour à tour, les k sous-échantillons deviennent les échantillons de validation. La performance globale du modèle est la moyenne des RMSE de chaque modèle.

Annexe 2 : Arbres de classification et de régression (CART)

L'objectif de cet algorithme est de regrouper des individus hétérogènes en classes les plus homogènes possibles. On distingue les arbres de classification (variable à expliquer qualitative) et de régression (variable à expliquer quantitative). Le principe de fonctionnement est le même. Un arbre est composé de trois éléments :

- Un **nœud racine** : c'est le point de départ de l'arbre. Il contient l'ensemble de la population que l'on veut segmenter.
- Des **nœuds internes** : ils contiennent les règles de division qui segmentent la population. Il y a deux règles importantes à respecter pour la segmentation : chaque question de segmentation doit avoir une réponse binaire, et elle ne doit être basée que sur une seule variable explicative.
- Des **feuilles** (ou nœuds terminaux) : il s'agit de la segmentation finale de l'arbre qui contient des sous populations homogènes. Chaque individu de départ appartient à une seule feuille.

Pour construire l'arbre, on part de la racine. On cherche la meilleure segmentation, en permettant d'obtenir deux ensembles les plus homogènes possibles. On segmente puis on fait de même sur chaque segment. Puis on réitère jusqu'à ce qu'on ne puisse plus segmenter. On obtient alors l'arbre maximal.

De manière plus théorique, on introduit les notations suivantes :

- Y_i la réponse observée du $i^{ème}$ individu.
- $X_i = (X_{i1}, \dots, X_{ik})$ le vecteur des facteurs de risque de l'individu i.
- χ l'espace des covariables (c'est-à-dire les facteurs de risque).

Dans le cas d'un arbre de régression, on cherche à exprimer $\pi_0(x) = \mathbb{E}_0(Y|X = x)$. Pour maximiser l'homogénéité, on va maximiser le critère de division. La solution $\pi_0(x)$ est donnée grâce à la méthode des moindres carrés ordinaires comme étant égale à : $\pi_0(x) = arg \min_{\pi(x)} \mathbb{E}\left(\left(Y - \pi(x)\right)^2 \middle| X = x\right)$.

Il s'agit ici de minimiser l'erreur quadratique moyenne. L'hétérogénéité va diminuer à chaque segmentation. Elle sera minimale quand l'arbre sera maximal. Toutefois, construire l'arbre maximal conduit très souvent à du surapprentissage. Pour éviter cela, il faudra ensuite l'élaguer.

L'élagage permet d'éviter d'avoir un estimateur trop complexe. Pour élaguer l'arbre, il faut donc jouer sur un paramètre de complexité α, et essayer de trouver la meilleure adéquation complexité / prédiction. La complexité α peut prendre des valeurs entre 0 et $+\infty$, 0 correspondant à l'arbre maximal et l'infini à l'arbre composé uniquement de la racine. L'élagage consiste donc à créer une suite emboitée de sous-arbres de l'arbre maximal grâce à un critère pénalisé, et de chercher l'arbre optimal de cette suite, à l'aide la méthode de validation croisée.

Afin de mesurer la performance du modèle, on peut utiliser les mesures MAE (*Mean Absolute Error*) et MSE (*Mean Square Error*).

Annexe 3 : Bagging

Le mot *bagging* est la concaténation des mots *bootstrap* et *aggregating*. Il s'agit d'une technique d'agrégation permettant d'améliorer la stabilité et la précision des algorithmes d'apprentissage. Le *bagging* est très souvent utilisé avec les arbres de décision.

Le principe du *bagging* est le suivant : on part d'un échantillon d'apprentissage de taille n. On va ensuite construire x nouveaux échantillons de taille n, par tirage au sort avec remise dans l'échantillon de départ. Certaines observations peuvent donc être répétées. Ce procédé est appelé *bootstrap*.

L'algorithme d'apprentissage va ensuite être entraîné sur les x échantillons.

La prédiction de ce nouveau modèle est obtenue en agrégeant les résultats de chaque algorithme :

- Par vote à la majorité, pour des algorithmes de classification.
- Par moyenne, pour des algorithmes de régression.

Le bagging permet de réduire la variance du modèle grâce à la combinaison de nombreux estimateurs indépendants. Plus le nombre d'estimateurs x est élevé, plus la variance sera maîtrisée.

Annexe 4 : Grilles de valeurs H2O

Grille de valeurs GBM

Paramètres	Valeurs recherchées
col_sample_rate	{0,4 ; 0,7 ; 1}
col_sample_rate_per_tree	{0,4 ; 0,7 ; 1}
learn_rate	0,1
max_depth	{3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12 ; 13 ; 14 ; 15 ; 16 ; 17}
min_rows	{1 ; 5 ; 10 ; 15 ; 30 ; 100}
min_split_improvment	$\{1e^{-4} ; 1e^{-5}\}$
ntrees	10 000
sample_rate	{0,5 ; 0,6 ; 0,7 ; 0,8 ; 0,9 ; 1}

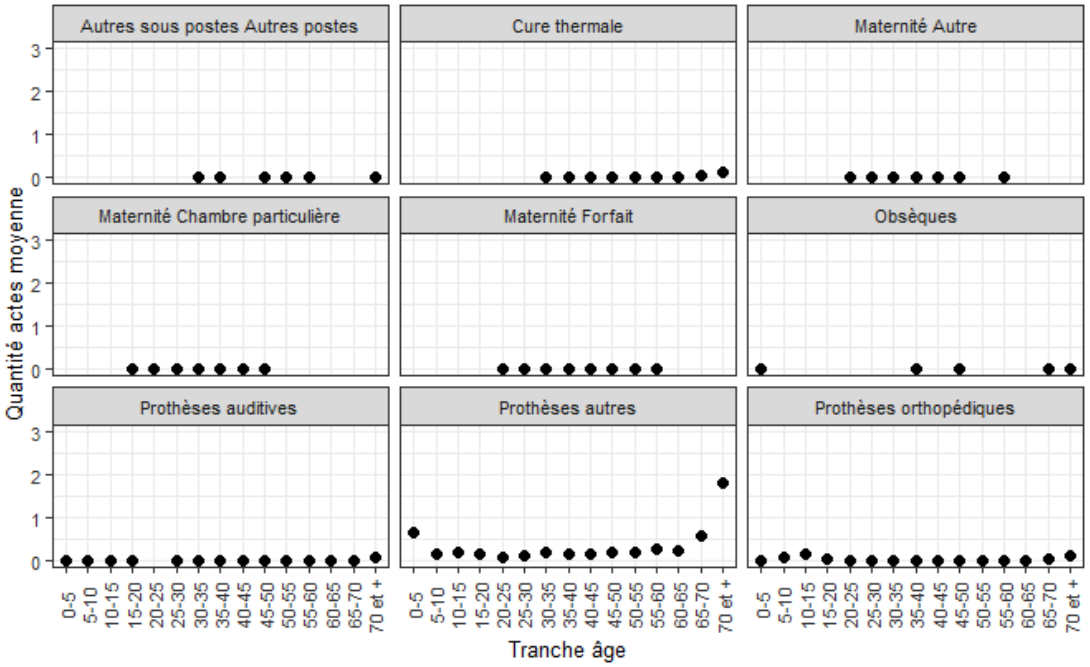
Grille de valeurs Réseau de neurones

Paramètres	Valeurs recherchées
epochs	10 000
epsilon	$\{1e^{-6} ; 1e^{-7} ; 1e^{-8} ; 1e^{-9}\}$
hidden	Grille de recherche 1 : {20} ; {50} ; {100} Grille de recherche 2 : {20 ;20} ; {50 ; 50} ; {100 ; 100} Grille de recherche 3 : {20 ;20 ; 20} ; {50 ; 50 ; 50} ; {100 ; 100 ; 100}
hidden_dropout_ratios	Grille de recherche 1 : {0,1} ; {0,2} ; {0,3} ; {0,4} ; {0,5} Grille de recherche 2 : {0,1 ; 0,1} ; {0,2 ; 0,2} ; {0,3 ; 0,3} ; {0,4 ; 0,4} ; {0,5 ; 0,5} Grille de recherche 3 : {0,1 ; 0,1 ; 0,1} ; {0,2 ; 0,2 ; 0,2} ; {0,3 ; 0,3 ; 0,3} ; {0,4 ; 0,4 ; 0,4} ; {0,5 ; 0,5 ; 0,5}
input_dropout_ratios	{0 ; 0,05 ; 0,1 ; 0,15 ; 0,2}
rho	{0,9 ; 0,95 ; 0,99}

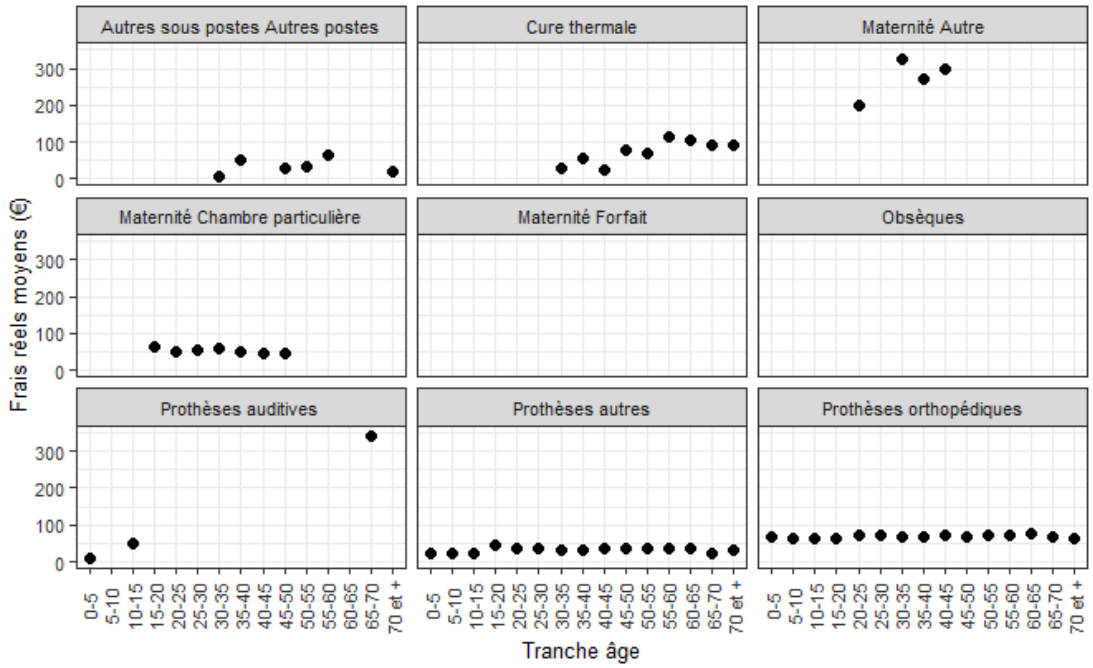
Annexe 5 : Statistiques par Sous poste sur les prestations

Autres postes

- *Quantité d'actes moyenne*

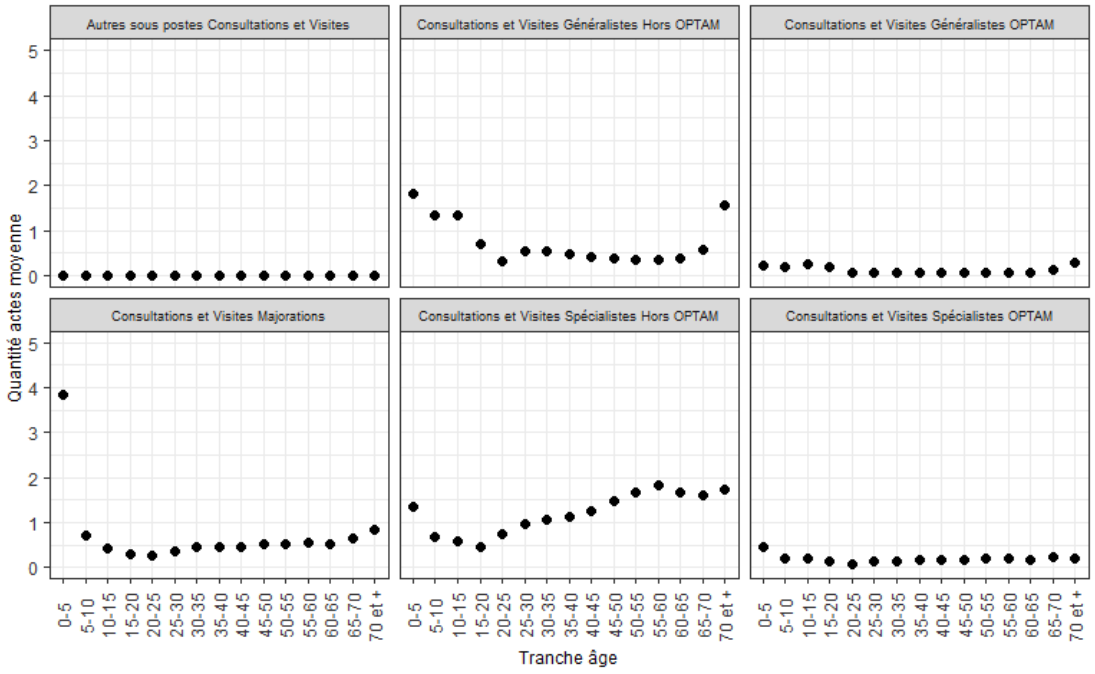


- *Frais réels moyens*

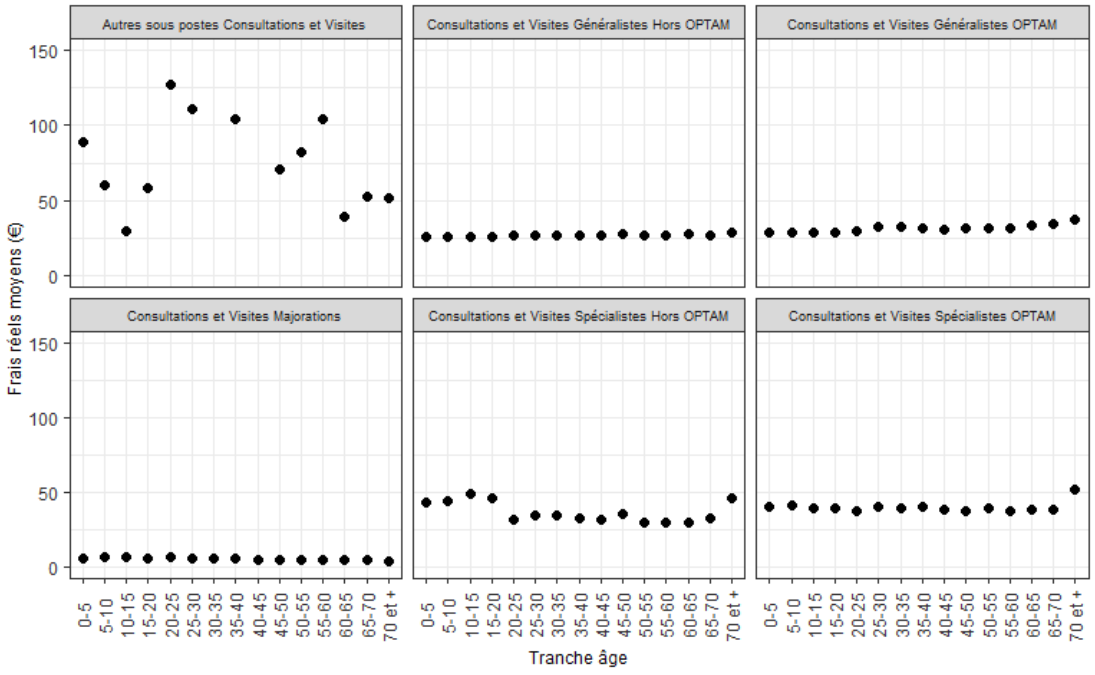


Consultations et Visites

- *Quantité d'actes moyenne*

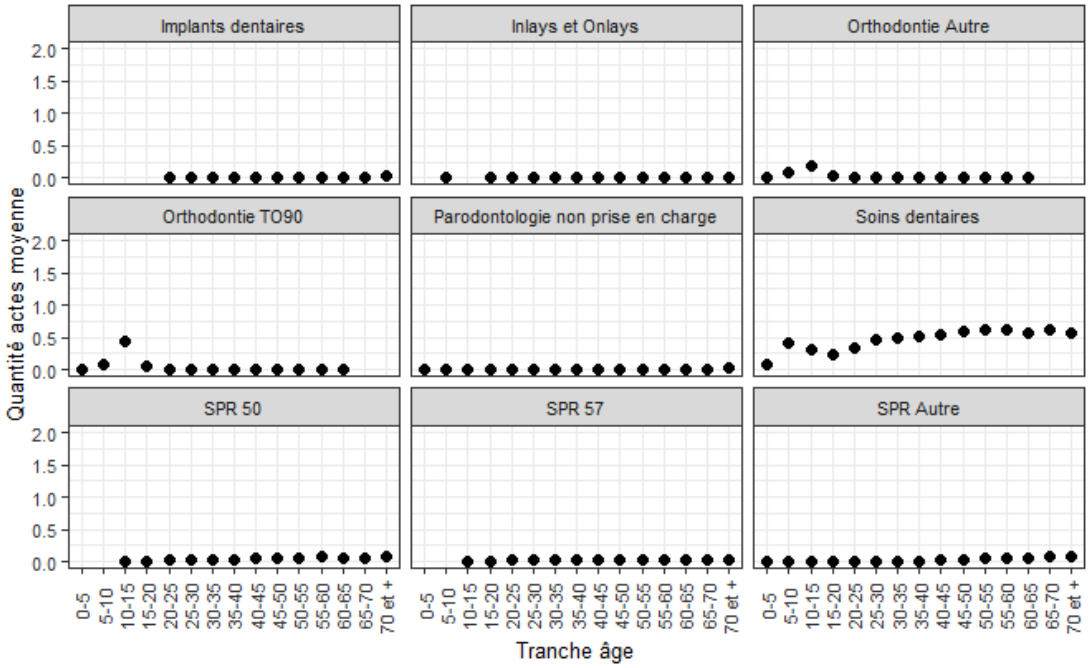


- *Frais réels moyens*

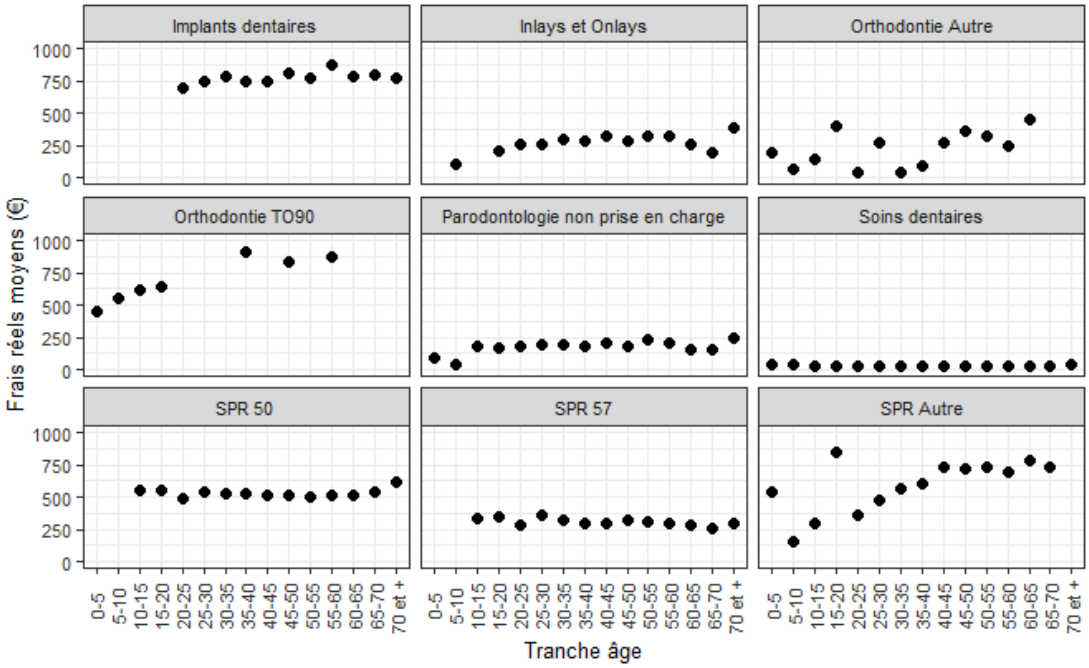


Dentaire

- *Quantité d'actes moyenne*

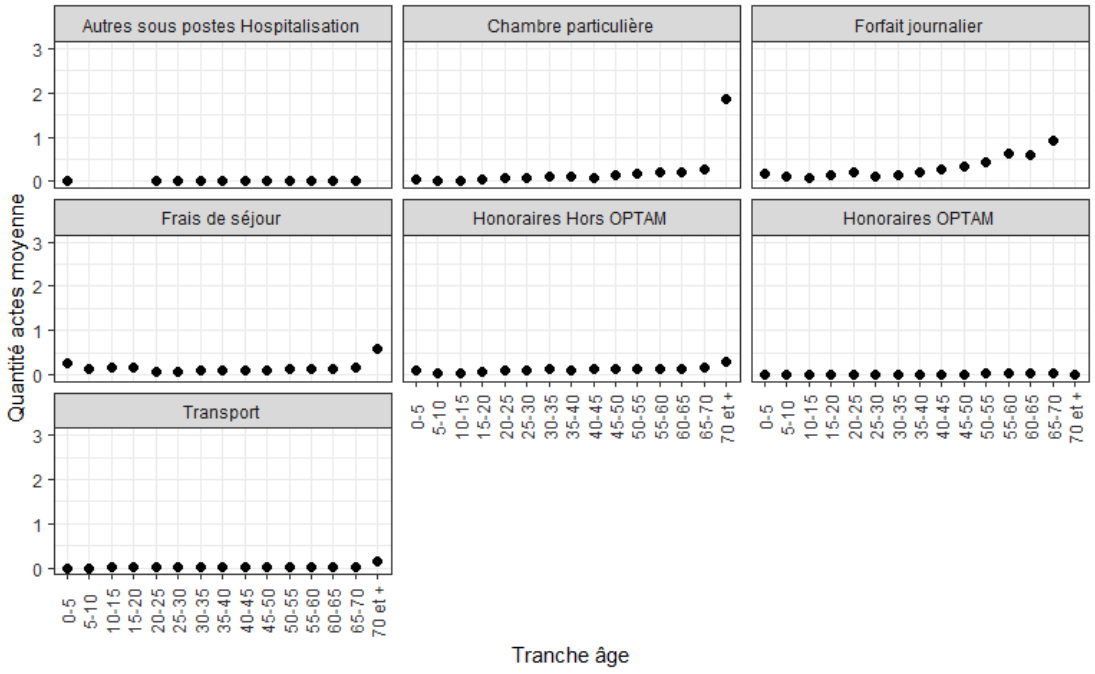


- *Frais réels moyens*

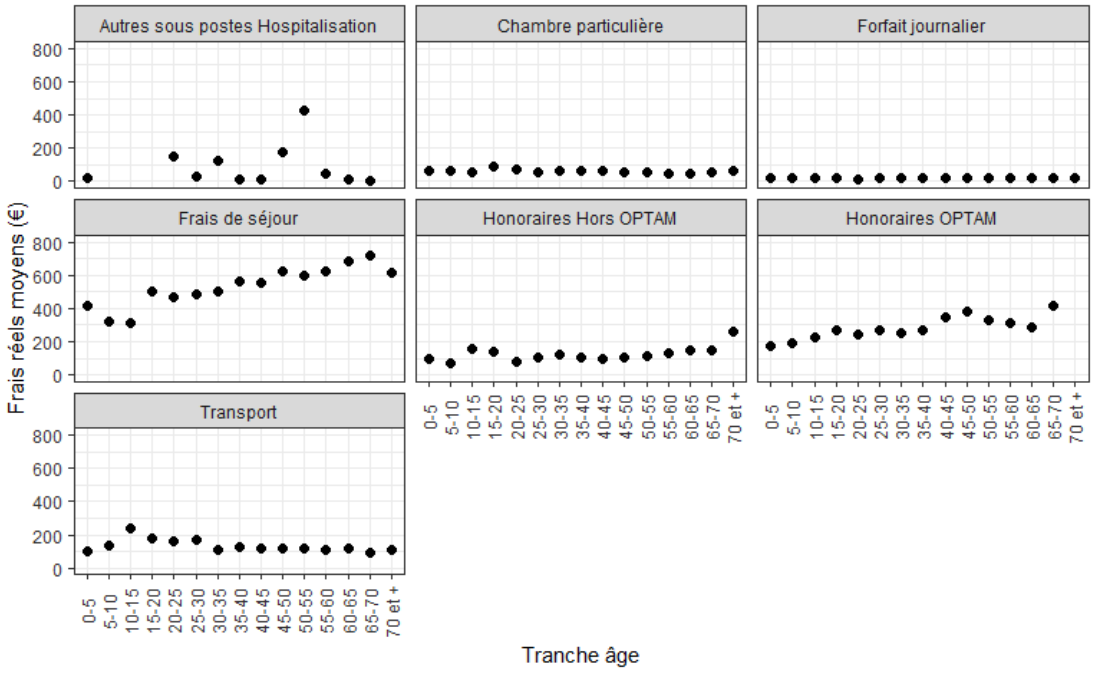


Hospitalisation

- *Quantité d'actes moyenne*

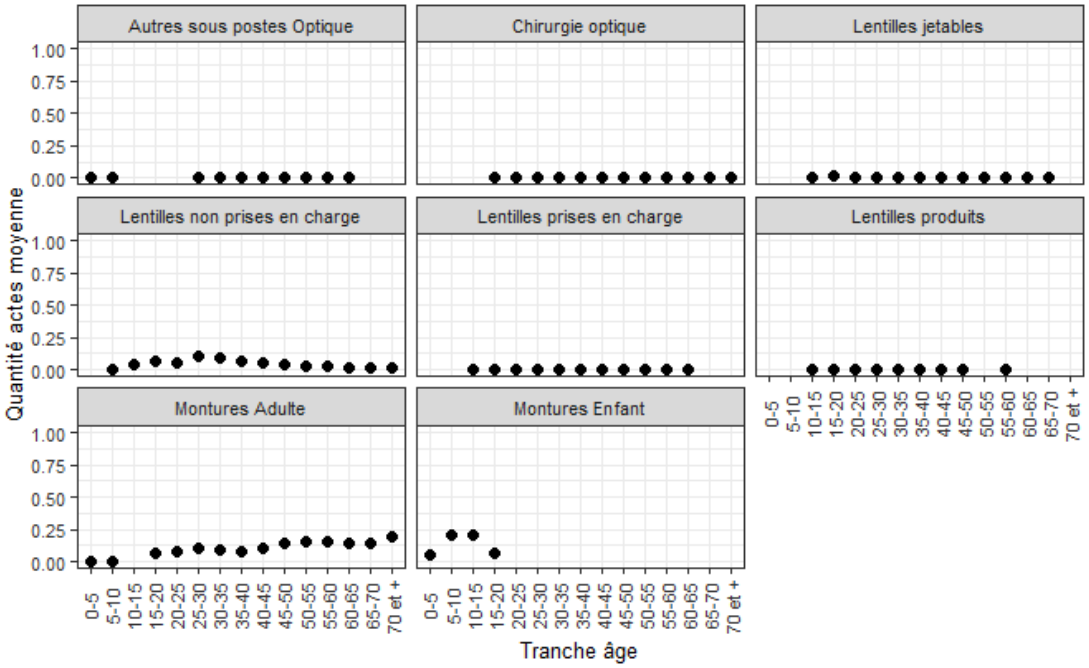


- *Frais réels moyens*

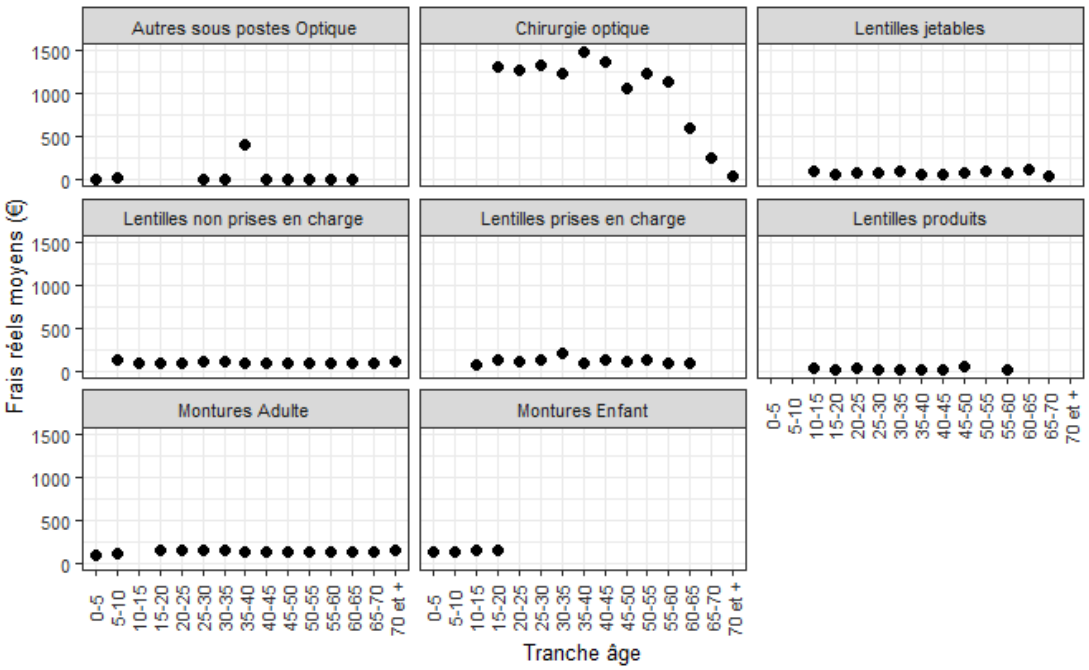


Optique (hors verres)

- *Quantité d'actes moyenne*

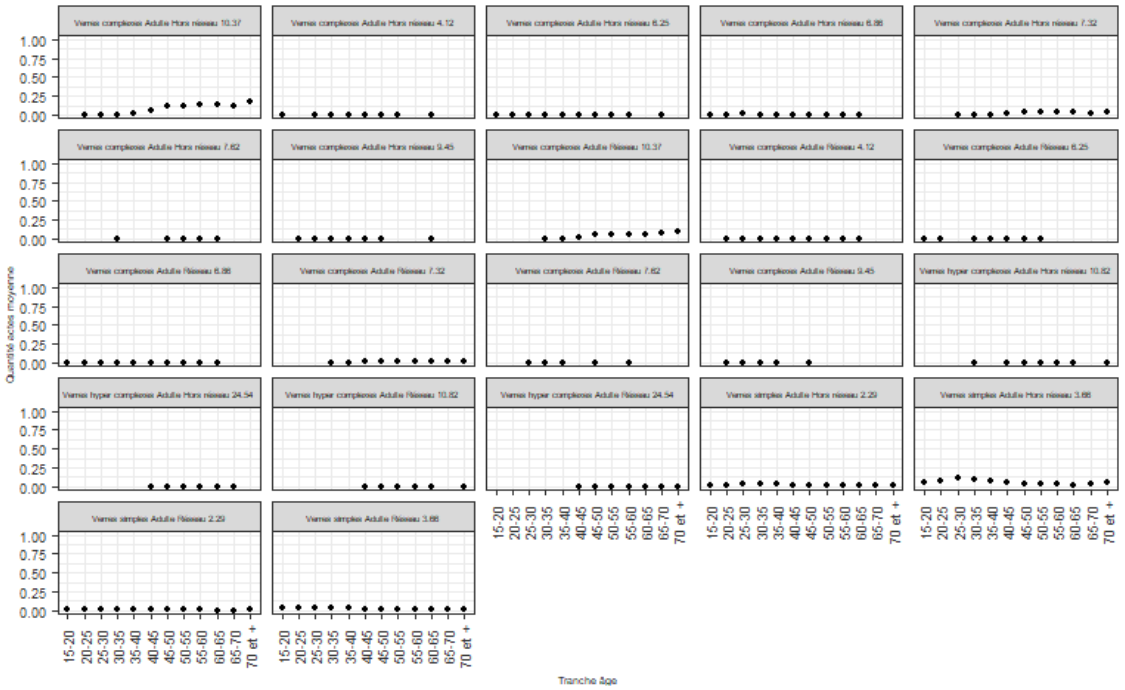


- *Frais réels moyens*

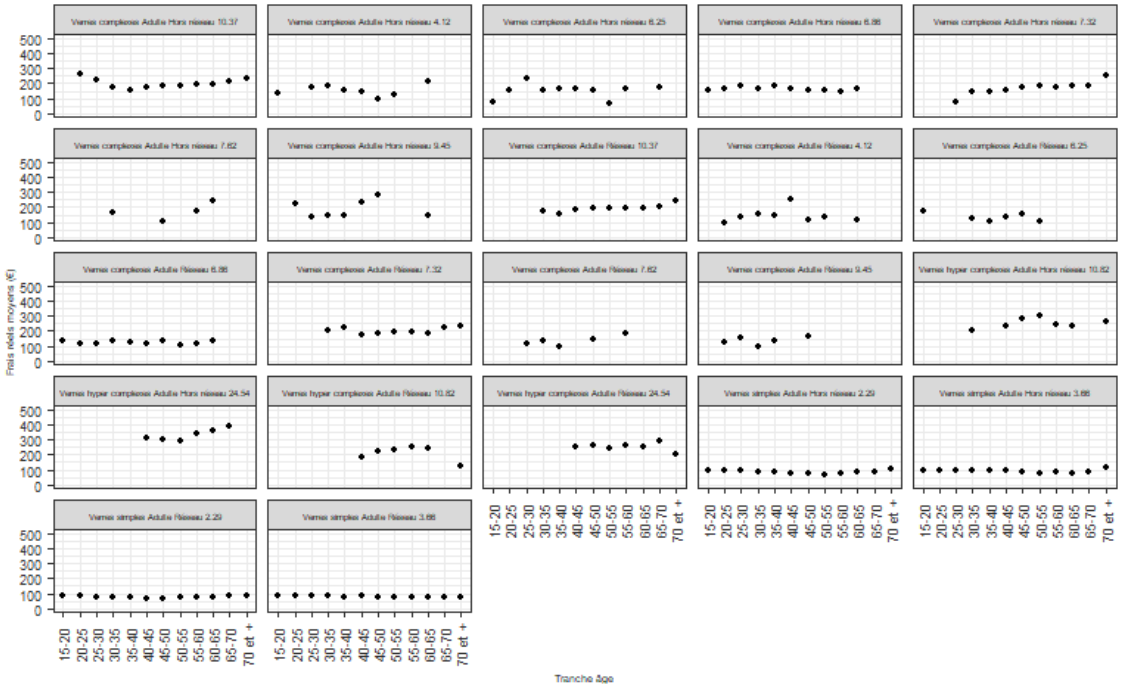


Optique (verres Adulte)

- *Quantité d'actes moyenne*

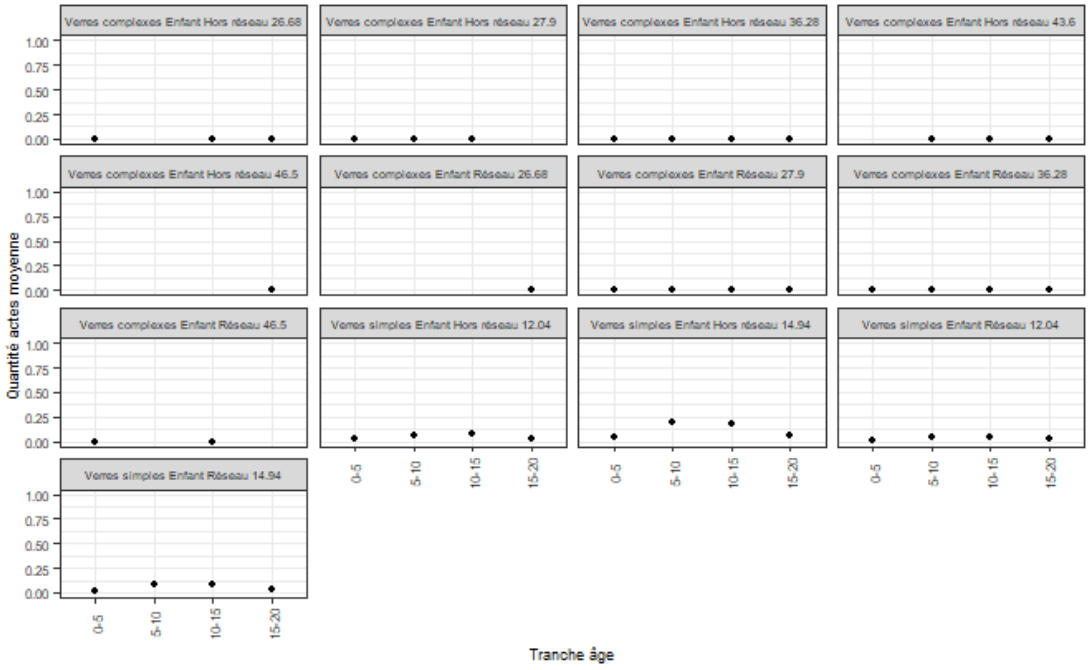


- *Frais réels moyens*

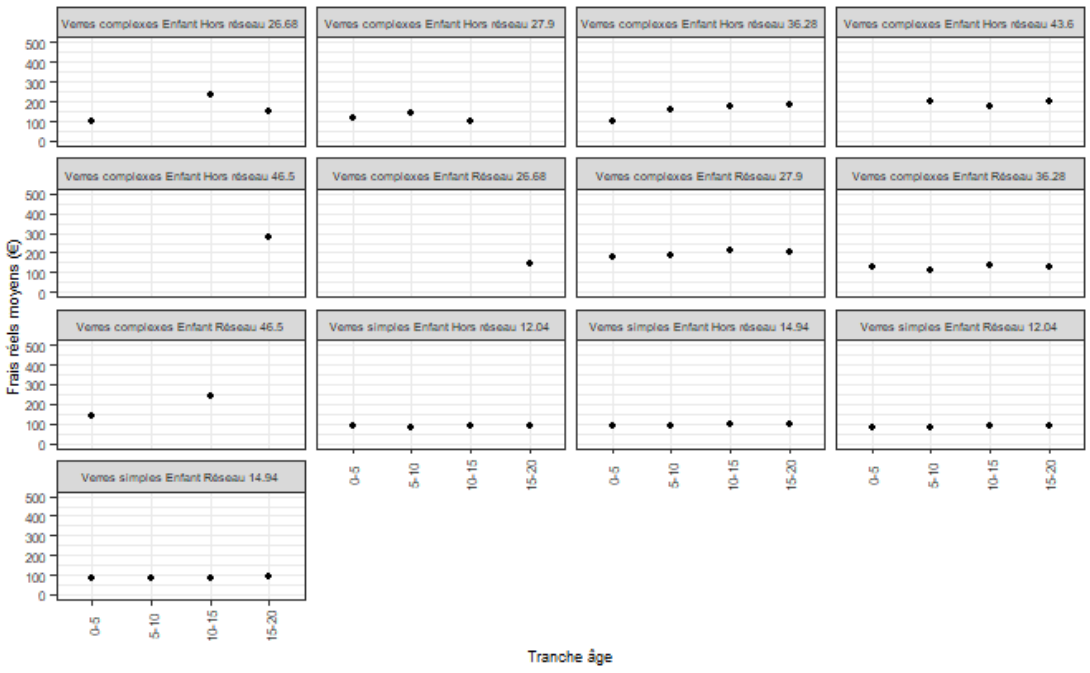


Optique (verres Enfant)

- *Quantité d'actes moyenne*

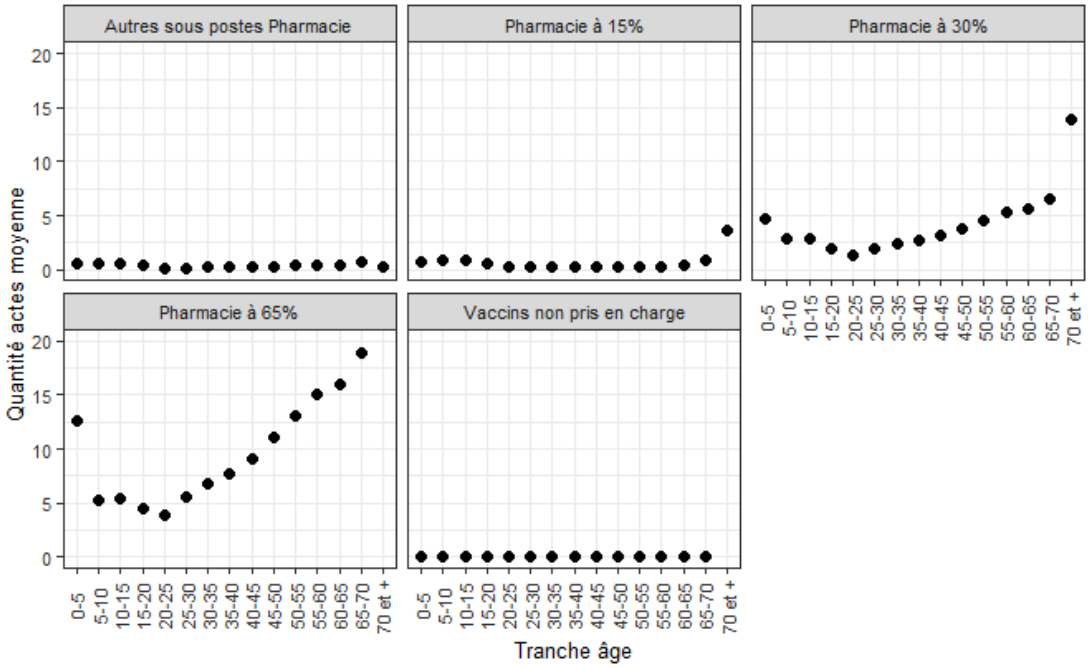


- *Frais réels moyens*

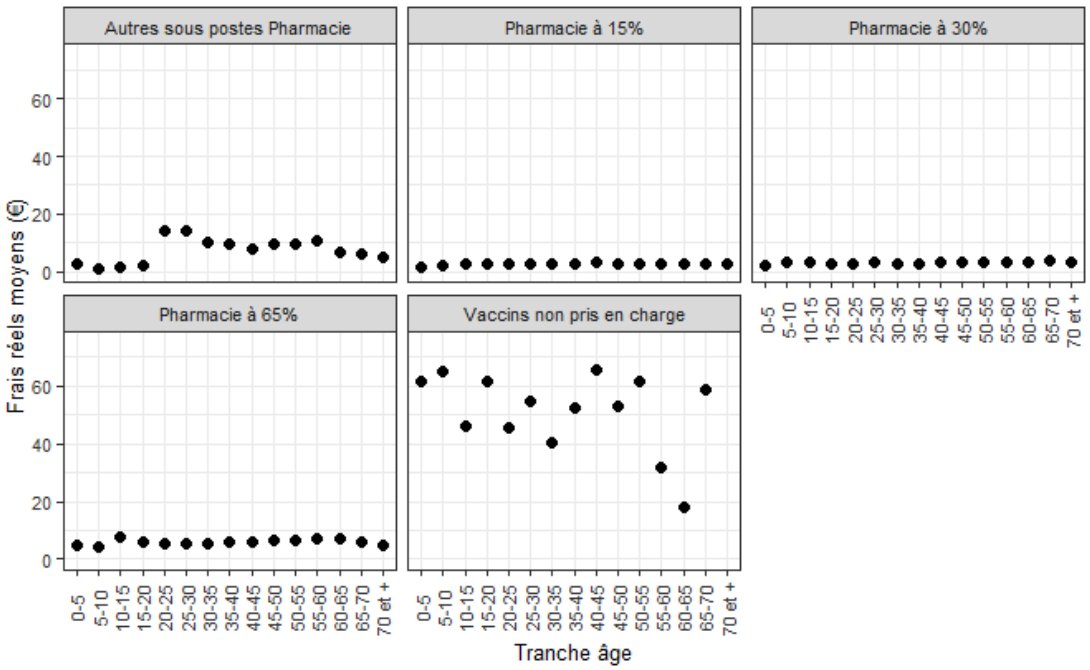


Pharmacie

- *Quantité d'actes moyenne*

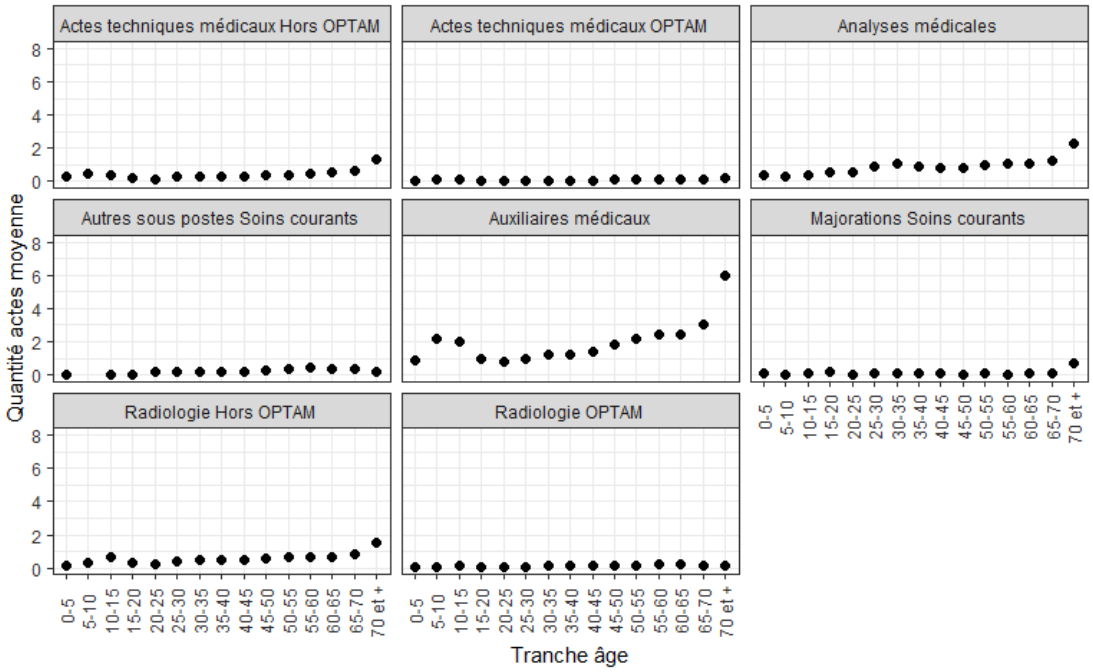


- *Frais réels moyens*

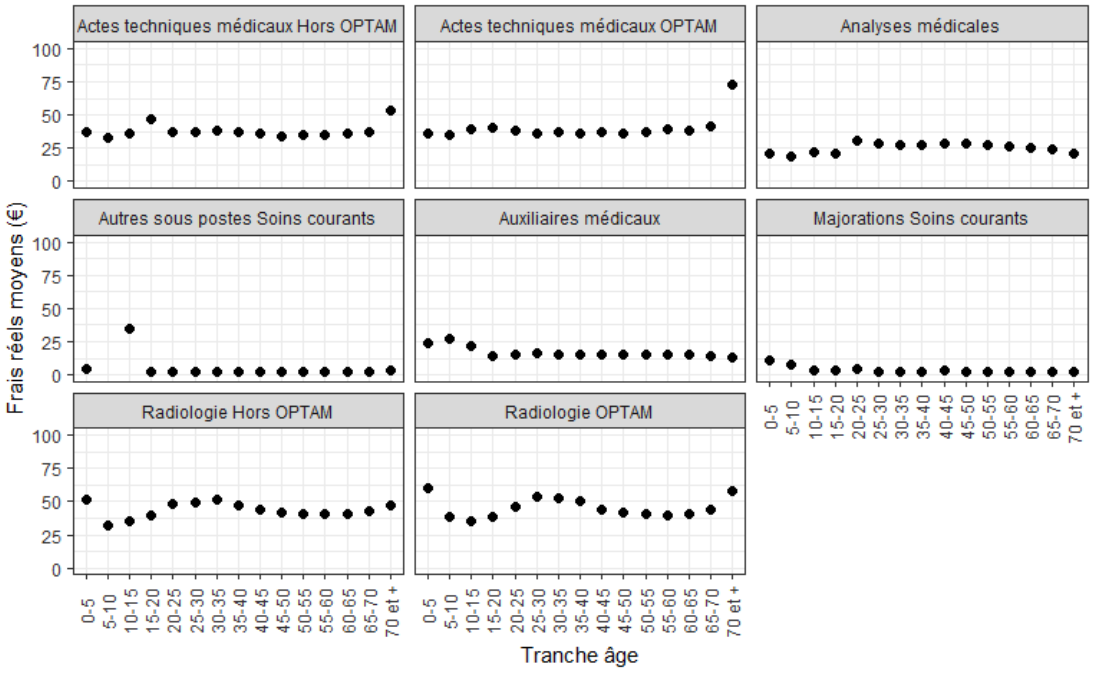


Soins courants (hors Médecine alternative)

- *Quantité d'actes moyenne*

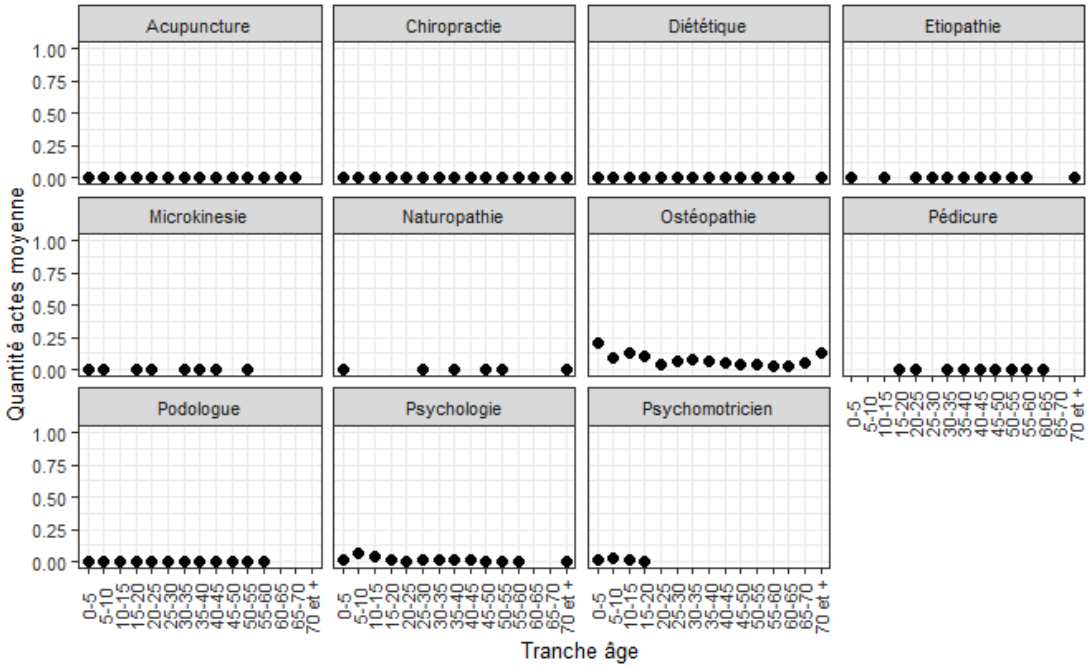


- *Frais réels moyens*

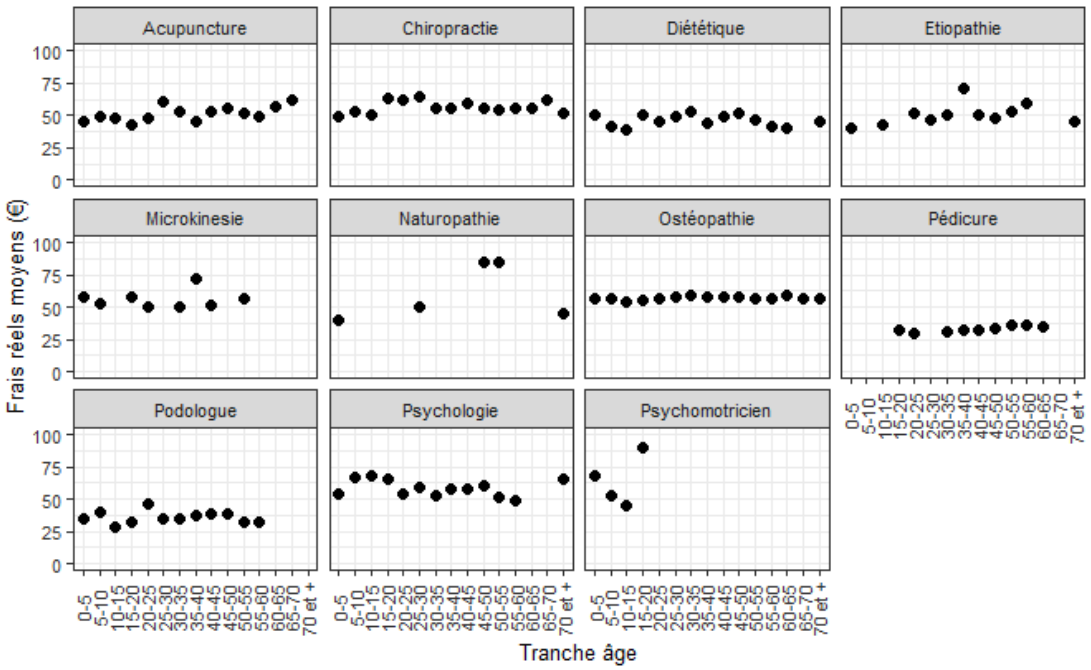


Soins courants (Médecine alternative)

- *Quantité d'actes moyenne*



- *Frais réels moyens*



Annexe 6 : Test d'indépendance du χ^2

Le test d'indépendance du χ^2 permet de déterminer si des variables qualitatives sont indépendantes ou non. L'hypothèse nulle de ce test est l'indépendance des deux variables.

Soient deux variables aléatoires qualitatives X et Y, prenant chacune un nombre de valeurs finies : I pour X et J pour Y. Soit N la taille de l'échantillon.

On réalise le tableau de contingence des données observées de notre échantillon. On note $O_{i,j}$ le nombre observé de données pour lesquelles X prend la valeur x_i, et Y prend la valeur y_j. Ce test ne peut être réalisé que si tous les $O_{i,j} > 5$.

Pour réaliser le tableau de contingence espéré sous l'hypothèse d'indépendance, on fait en sorte que les totaux des lignes et des colonnes restent identiques à ceux observés. Le nombre espéré de données pour lesquelles X prend la valeur x_i, et Y prend la valeur y_j, est noté $E_{i,j}$ et est tel que :

$$E_{i,j} = \frac{O_{i,+} \times O_{+,j}}{N}, \qquad \text{où } O_{i,+} = \sum_{j=1}^{J} O_{i,j} \text{ et } O_{+,j} = \sum_{i=1}^{I} O_{i,j}$$

La distance entre les valeurs théoriques et empiriques est calculée par la formule suivante :

$$T = \sum_{i,j} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

Sous l'hypothèses d'indépendance, T tend en loi vers une loi du χ^2, dont de degré de liberté df vaut $(I-1) \times (J-1)$.

Pour vérifier si l'hypothèse est vraie, T est comparé à la valeur tabulée du χ^2 au seuil α (seuil de significativité, généralement égal à 5%). Si T est supérieur à cette valeur, l'hypothèse d'indépendance est rejetée, sinon elle est validée.

Annexe 7 : Coefficient de corrélation de Pearson

Soient deux variables aléatoires réelles, X et Y, de variance finie. On note σ_X et σ_Y les écarts types respectifs de X et Y, et $\mathrm{Cov}(X, Y)$, la covariance entre les deux variables. Le coefficient de corrélation de Pearson, noté $r_{X,Y}$, est défini par :

$$r_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

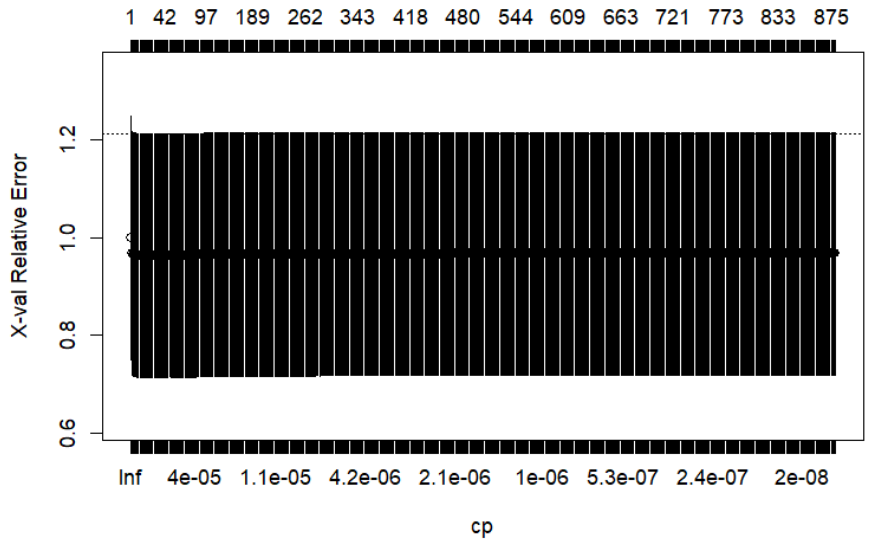
La matrice de corrélation du vecteur de N variables aléatoire $\begin{pmatrix} X_1 \\ \dots \\ X_N \end{pmatrix}$, dont chacune possède une variance finie, est la matrice dont les coefficients sont les r_{X_i,X_j}. Il s'agit d'une matrice carrée, symétrique, dont les coefficients diagonaux sont tous égaux à 1.

Le coefficient de corrélation de Pearson permet de quantifier la dépendance linéaire entre deux variables quantitatives.

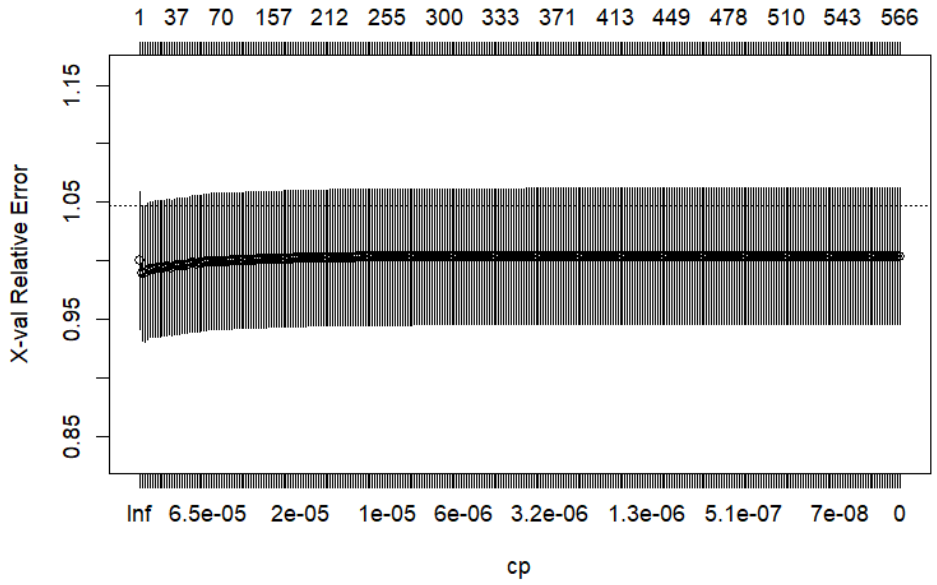
Annexe 8 : Graphiques annexes des premiers essais de modélisation

Arbres de régression : Fréquence

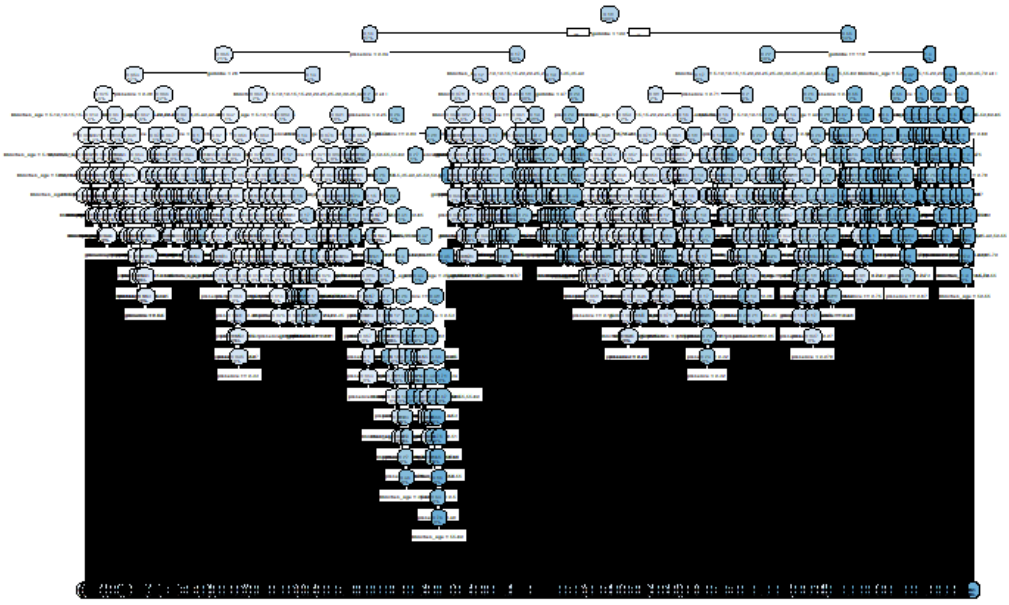
- *cp-plot : Consultations et Visites Spécialistes OPTAM*



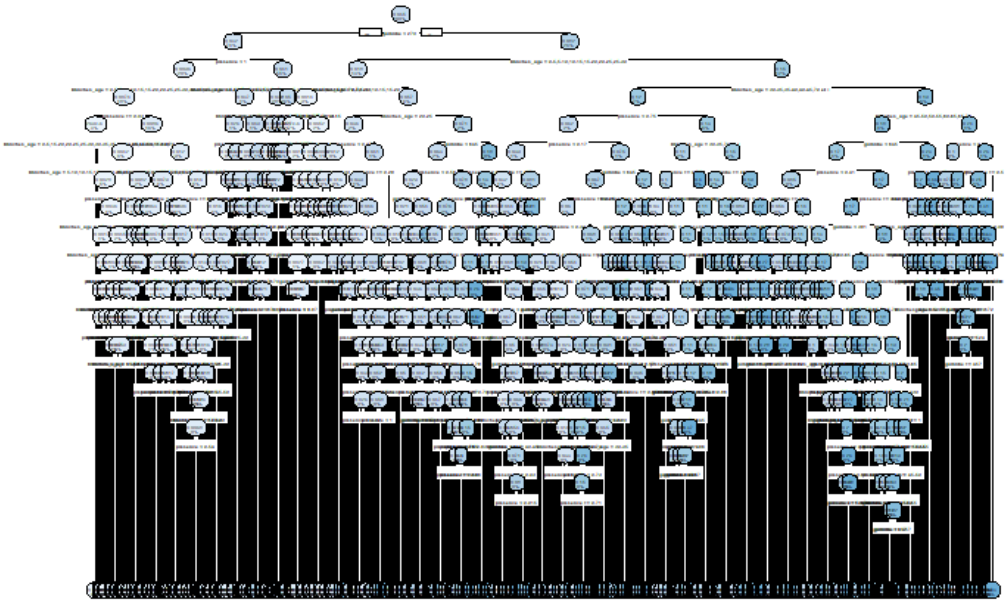
- *cp-plot : SPR 50*



- *Arbre maximal : Consultations et Visites Spécialistes OPTAM*

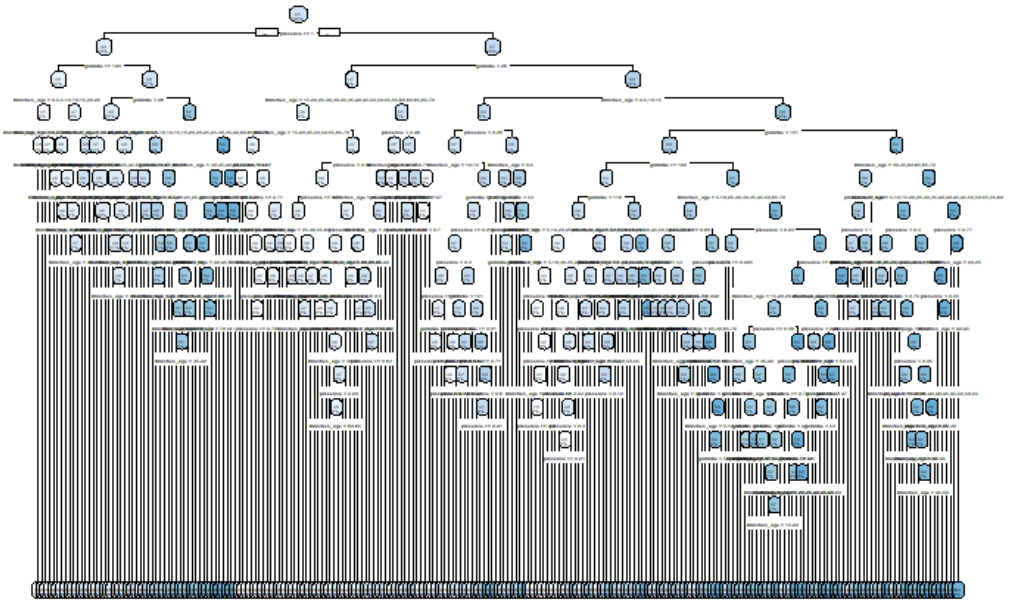


- *Arbre maximal : SPR 50*

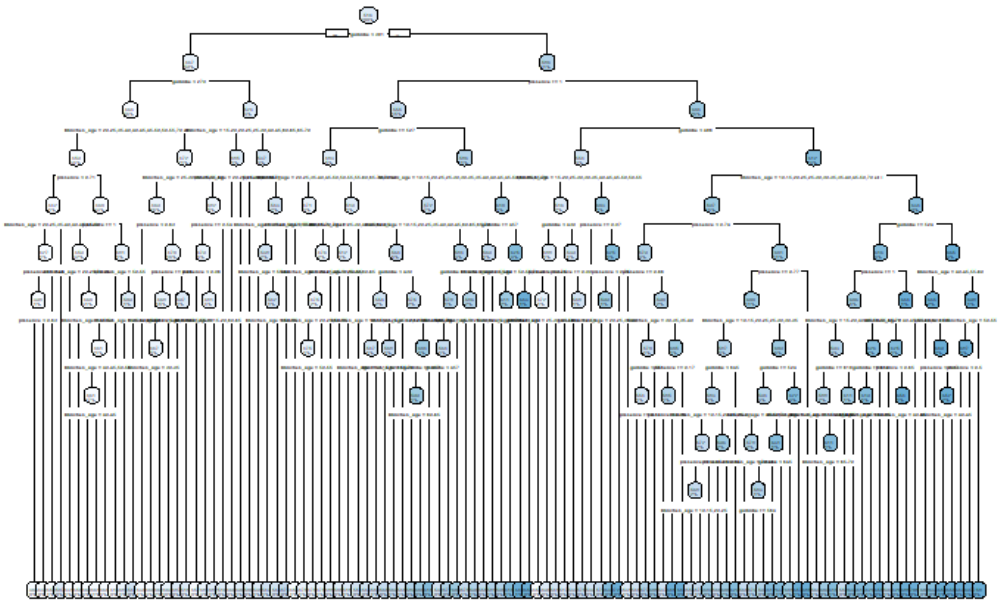


Arbres de régressions : Coût moyen

- *Arbre maximal : Consultations et Visites Spécialistes OPTAM*

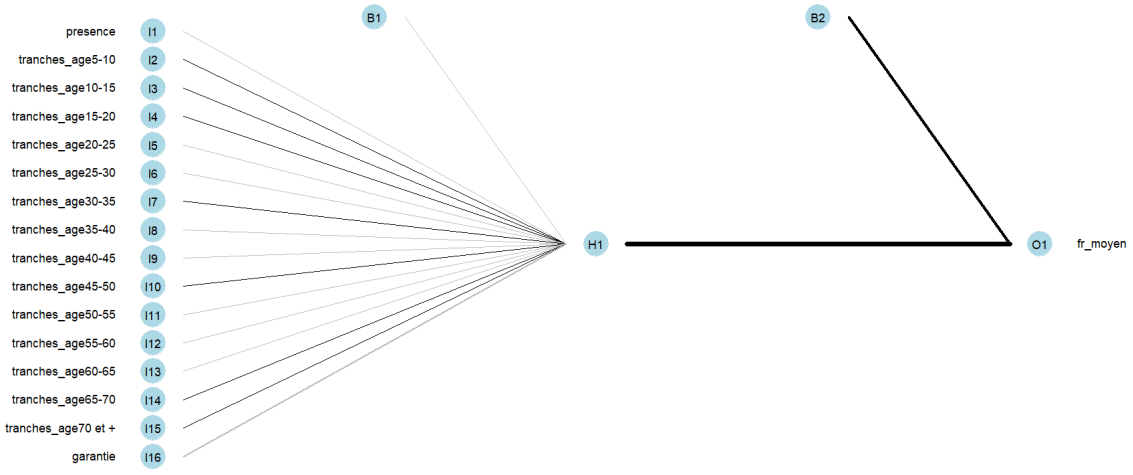


- *Arbre maximal : SPR 50*

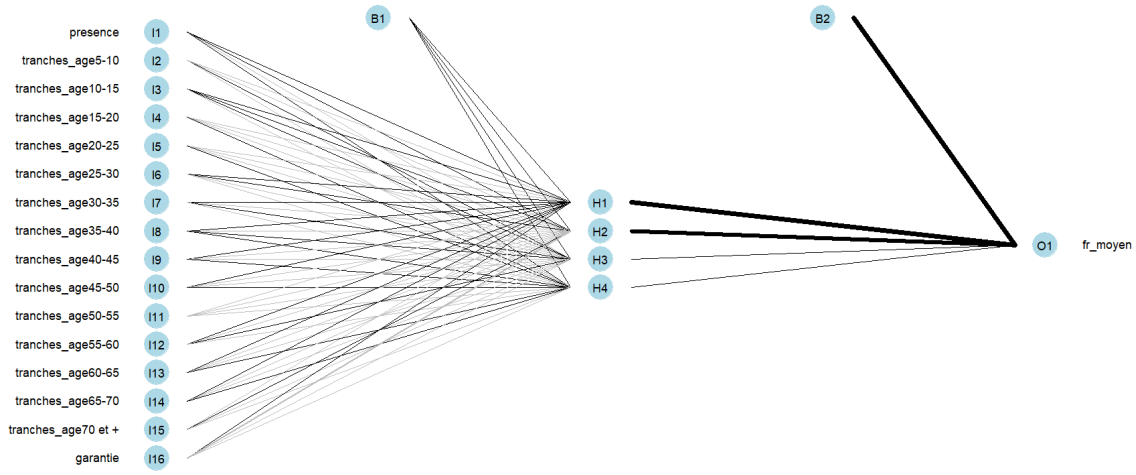


Réseaux de neurones : Coût moyen

- *Représentation du réseau de neurones : Consultations et Visites Spécialistes OPTAM*



- *Représentation du réseau de neurones : SPR 50*



Annexe 9 : Table des sous postes tarifables

Numéro Sous poste	Sous poste	Grand poste
1	Cure thermale	Autres postes
2	Maternité Forfait	Autres postes
3	Prothèses auditives	Autres postes
4	Prothèses autres	Autres postes
5	Prothèses orthopédiques	Autres postes
6	Consultations et Visites Généralistes OPTAM	Consultations et Visites
7	Consultations et Visites Généralistes Hors OPTAM	Consultations et Visites
8	Consultations et Visites Majorations	Consultations et Visites
9	Consultations et Visites Spécialistes OPTAM	Consultations et Visites
10	Consultations et Visites Spécialistes Hors OPTAM	Consultations et Visites
11	Implants dentaires	Dentaire
12	Inlays et Onlays	Dentaire
13	Orthodontie Autre	Dentaire
14	Orthodontie TO90	Dentaire
15	Parodontologie non prise en charge	Dentaire
16	Soins dentaires	Dentaire
17	SPR 50	Dentaire
18	SPR 57	Dentaire
19	SPR Autre	Dentaire
20	Chambre particulière	Hospitalisation
21	Forfait journalier	Hospitalisation
22	Frais de séjour	Hospitalisation
23	Honoraires OPTAM	Hospitalisation
24	Honoraires Hors OPTAM	Hospitalisation
25	Transport	Hospitalisation
26	Chirurgie optique	Optique
27	Lentilles jetables	Optique
28	Lentilles non prises en charge	Optique
29	Lentilles prises en charge	Optique
30	Montures Adulte	Optique
31	Montures Enfant	Optique
32	Verres complexes Adulte Hors réseau 10.37	Optique
33	Verres complexes Adulte Hors réseau 6.86	Optique
34	Verres complexes Adulte Hors réseau 7.32	Optique
35	Verres complexes Adulte Réseau 10.37	Optique
36	Verres complexes Adulte Réseau 6.86	Optique
37	Verres complexes Adulte Réseau 7.32	Optique
38	Verres simples Adulte Hors réseau 2.29	Optique
39	Verres simples Adulte Hors réseau 3.66	Optique
40	Verres simples Adulte Réseau 2.29	Optique
41	Verres simples Adulte Réseau 3.66	Optique
42	Verres simples Enfant Hors réseau 12.04	Optique
43	Verres simples Enfant Hors réseau 14.94	Optique
44	Verres simples Enfant Réseau 12.04	Optique
45	Verres simples Enfant Réseau 14.94	Optique
46	Pharmacie à 15%	Pharmacie
47	Pharmacie à 30%	Pharmacie
48	Pharmacie à 65%	Pharmacie
49	Vaccins non pris en charge	Pharmacie
50	Actes techniques médicaux OPTAM	Soins courants
51	Actes techniques médicaux Hors OPTAM	Soins courants
52	Acupuncture	Soins courants
53	Analyses médicales	Soins courants
54	Auxiliaires médicaux	Soins courants
55	Majorations Soins courants	Soins courants
56	Radiologie OPTAM	Soins courants
57	Radiologie Hors OPTAM	Soins courants
58	Chiropractie	Soins courants
59	Diététique	Soins courants
60	Etiopathie	Soins courants
61	Microkinésie	Soins courants
62	Naturopathie	Soins courants
63	Ostéopathie	Soins courants
64	Pédicure	Soins courants
65	Podologue	Soins courants
66	Psychologie	Soins courants
67	Psychomotricien	Soins courants

Annexe 10 : Classement des sous postes par type de modèle

Fréquence

GLM	DRF (Forêts aléatoires)	XRT (Forêts extrêmement aléatoires)
Soins courants - Podologie - Verres simples Adulte Réseau 2.29	*Soins courants* - Chiropractie - Psychologie - Psychomotricité *Dentaire* - SPR 57 *Optique* - Verres simples Adulte Réseau 2.29	*Soins courants* - Microkinésie

GBM	Stacked Ensemble All models	Stacked Ensemble Best of family
Autres postes - Prothèses auditives *Dentaire* - Implants dentaires - SPR 50 *Optique* - Chirurgie optique - Lentilles jetables - Lentilles prises en charge - Tous les Verres simples Enfant - Verres complexes Adulte Réseau 6.86 *Pharmacie* - Vaccins non pris en charge *Soins courants* - Diététique - Etiopathie - Majorations Soins courants	*Autres postes* - Cure thermale - Maternité Forfait - Prothèses autres - Prothèses orthopédiques *Consultations et Visites* - Consultations et Visites Généralistes Hors OPTAM Consultations et Visites Spécialistes OPTAM - Consultations et Visites Majorations *Dentaire* - Orthodontie TO90 et Autre - Parodontologie non prise en charge - Soins dentaires - SPR autres *Hospitalisation* - Chambre particulière - Honoraires OPTAM - Transport *Optique* - Montures Adulte - Montures Enfant - Verres complexes Adulte Hors réseau 10.37 - Verres complexes Adulte Hors réseau 6.86 - Verres complexes Adulte Hors réseau 7.32 - Verres complexes Adulte Réseau 10.37 - Verres simples Adulte Hors réseau 3.66 *Pharmacie* - Pharmacie à 15 % - Pharmacie à 30 % - Pharmacie à 65 % *Soins courants* - Actes techniques médicaux Hors OPTAM - Analyses médicales - Auxiliaires médicaux - Naturopathie - Ostéopathie - Radiologie OPTAM - Radiologie Hors OPTAM	*Consultations et Visites* - Consultation et Visites Généralistes OPTAM - Consultation et Visites Spécialistes Hors OPTAM *Dentaire* - Inlays et Onlays *Hospitalisation* - Forfait journalier - Frais de séjour - Honoraires Hors OPTAM *Optique* - Lentilles non prises en charge - Verres complexes Adulte Réseau 7.32 - Verres simples Adulte Hors réseau 2.29 - Verres simples Adulte Réseau 3.66 *Soins courants* - Actes techniques médicaux OPTAM - Acupuncture - Pédicure

Coût moyen

GBM	XRT
Autres postes - Maternité Forfait ***Dentaire*** - Orthodontie TO90 - SPR 50 - SPR 57 - SPR autres ***Hospitalisation*** - Chambre particulière - Frais de séjour ***Optique*** - Verres simples Adulte Réseau 3.66 - Verres simples Enfant Hors réseau 12.04 - Verres simples Enfant Hors réseau 14.94 - Verres simples Enfant Réseau 14.94 ***Soins courants*** - Actes techniques médicaux Hors OPTAM - Chiropractie - Majorations Soins courants - Ostéopathie - Psychologie	***Consultations et Visites*** - Consultations et Visites Généralistes OPTAM ***Hospitalisation*** - Honoraires OPTAM ***Optique*** - Lentilles jetables - Verres complexes Adulte Réseau 6.86

DRF	GLM
Consultations et Visites - Consultations et Visites Généralistes Hors OPTAM	***Optique*** - Verres simples Adulte Réseau 3.66

Deep learning (Réseaux de neurones)	Stacked Ensemble All models	Stacked Ensemble Best of family
Autres postes - Cure thermale - Prothèses auditives ***Consultations et Visites*** - Consultation et Visites Spécialistes Hors OPTAM ***Dentaire*** - Inlays et Onlays ***Hospitalisation*** - Forfait journalier - Transport ***Optique*** - Chirurgie optique - Lentilles prises en charge - Verres complexes Adulte Hors réseau 6.86 - Verres simples Enfant Réseau 12.04 ***Pharmacie*** - Vaccins non pris en charge ***Soins courants*** - Acupuncture - Diététique - Etiopathie - Microkinésie - Naturopathie - Pédicurie - Podologie - Psychomotricité	***Consultations et Visites*** - Consultation et Visites Majorations **Dentaire** - Parodontologie non prise en charge - Soins dentaires ***Optique*** - Montures Adulte - Verres complexes Adulte Hors réseau 10.37 - Verres complexes Adulte Réseau 10.37 - Verres complexes Adulte Réseau 7.32 - Verres simples Adulte Réseau 2.29 ***Pharmacie*** - Pharmacie à 30 % - Pharmacie à 65 % ***Soins courants*** - Analyses médicales - Auxiliaires médicaux - Radiologie Hors OPTAM	***Autres postes*** - Prothèses autres - Prothèses orthopédiques ***Consultations et Visites*** - Consultation et Visites Spécialistes OPTAM ***Dentaire*** - Implants dentaires - Orthodontie Autre ***Hospitalisation*** - Honoraires Hors OPTAM ***Optique*** - Lentilles non prises en charge - Montures Enfant - Verres complexes Adulte Hors réseau 7.32 - Verres simples Adulte Hors réseau 2.29 ***Pharmacie*** - Pharmacie à 15 % ***Soins courants*** - Actes techniques médicaux OPTAM - Radiologie OPTAM

Annexe 11 : RMSE et MAE des modèles

Fréquence

Grand poste	Sous poste	RMSE	RMSE normalisé	MAE	MAE normalisé
Autres postes	Cure thermale	0.1748	0.3641	0.0037	0.4960
	Maternité Forfait	0.0211	0.1065	0.0017	0.1134
	Prothèses auditives	0.0605	0.4504	0.0041	0.4603
	Prothèses autres	1.1288	0.2892	0.1325	0.3303
	Prothèses orthopédiques	0.1016	0.2628	0.0157	0.2856
Consultations et Visites	Consultations et Visites Généralistes OPTAM	0.2292	0.1083	0.0287	0.1036
	Consultations et Visites Généralistes Hors OPTAM	0.7086	0.1611	0.2445	0.1798
	Consultations et Visites Majorations	0.9764	0.1678	0.3359	0.1370
	Consultations et Visites Spécialistes OPTAM	0.4496	0.2377	0.1097	0.5279
	Consultations et Visites Spécialistes Hors OPTAM	1.3751	0.2334	0.6598	0.2435
Dentaire	Implants dentaires	0.0566	0.1627	0.0042	0.1044
	Inlays et Onlays	0.0439	0.1648	0.0028	0.2380
	Orthodontie Autre	0.1685	0.1576	0.0061	0.1616
	Orthodontie TO90	0.0760	0.1600	0.0091	0.1811
	Parodontologie non prise en charge	0.0612	0.1490	0.0053	0.0941
	Soins dentaires	1.1288	0.1867	0.4960	0.1091
	SPR 50	0.1707	0.1942	0.0187	0.3699
	SPR 57	0.1519	0.1823	0.0226	0.2488
	SPR Autre	0.1209	0.1655	0.0210	0.2436
Hospitalisation	Chambre particulière	1.1010	0.1582	0.0790	0.1422
	Forfait journalier	2.3088	0.3209	0.2507	0.3761
	Frais de séjour	0.5630	0.3096	0.0804	0.3679
	Honoraires OPTAM	0.0662	0.4515	0.0056	0.4482
	Honoraires Hors OPTAM	0.4244	0.1711	0.0869	0.2372
	Transport	0.2467	0.0971	0.0178	0.0508
Optique	Chirurgie optique	0.0201	0.3065	0.0007	0.2572
	Lentilles jetables	0.0408	0.1197	0.0016	0.1988
	Lentilles non prises en charge	0.2644	0.4005	0.0590	0.6844
	Lentilles prises en charge	0.0428	0.5945	0.0022	0.7270
	Montures Adulte	0.1489	0.1223	0.0711	0.1799
	Montures Enfant	0.0466	0.1035	0.0074	0.1733
	Verres complexes Adulte Hors réseau 10.37	0.2266	0.1687	0.0681	0.0844
	Verres complexes Adulte Hors réseau 6.86	0.0310	0.1750	0.0018	0.2262
	Verres complexes Adulte Hors réseau 7.32	0.0853	0.1740	0.0138	0.2322
	Verres complexes Adulte Réseau 10.37	0.1020	0.1762	0.0178	0.1794
	Verres complexes Adulte Réseau 6.86	0.0236	0.2858	0.0010	0.4598
	Verres complexes Adulte Réseau 7.32	0.0515	0.1473	0.0055	0.1918
	Verres simples Adulte Hors réseau 2.29	0.0902	0.1724	0.0144	0.1985
	Verres simples Adulte Hors réseau 3.66	0.1446	0.1896	0.0350	0.2026
	Verres simples Adulte Réseau 2.29	0.0589	0.1792	0.0065	0.2462
	Verres simples Adulte Réseau 3.66	0.1325	0.1589	0.0228	0.0955
	Verres simples Enfant Hors réseau 12.04	0.0395	0.1780	0.0028	0.2259
	Verres simples Enfant Hors réseau 14.94	0.0633	0.2732	0.0071	0.2323
	Verres simples Enfant Réseau 12.04	0.0273	0.1204	0.0014	0.1961
	Verres simples Enfant Réseau 14.94	0.0438	0.1723	0.0034	0.1787
Pharmacie	Pharmacie à 15%	1.0950	0.1302	0.2074	0.0713
	Pharmacie à 30%	4.4660	0.1720	1.8158	0.1931
	Pharmacie à 65%	9.1497	0.1770	4.5728	0.1611
	Vaccins non pris en charge	0.0138	0.0947	0.0005	0.0688
Soins courants	Actes techniques médicaux OPTAM	0.2067	0.1365	0.0447	0.1316
	Actes techniques médicaux Hors OPTAM	0.5604	0.1863	0.2311	0.1648
	Acupuncture	0.0224	0.0927	0.0007	0.0340
	Analyses médicales	1.2914	0.1751	0.4809	0.1146
	Auxiliaires médicaux	4.0183	0.1657	1.0650	0.1665
	Majorations Soins courants	0.4794	0.1591	0.0429	0.1932
	Radiologie OPTAM	0.2901	0.1529	0.0824	0.1457
	Radiologie Hors OPTAM	0.6792	0.2170	0.3062	0.1384
	Chiropractie	0.0388	0.1641	0.0019	0.1005
	Diététique	0.0365	0.3503	0.0017	0.4345
	Etiopathie	0.0142	0.1177	0.0003	0.0599
	Microkinesie	0.0070	0.2796	0.0001	0.5298
	Naturopathie	0.0066	0.1432	0.0002	0.0848
	Ostéopathie	0.1515	0.1336	0.0346	0.0758
	Pédicure	0.0172	0.1515	0.0006	0.0426
	Podologue	0.0161	0.0997	0.0006	0.0628
	Psychologie	0.0732	0.1675	0.0051	0.0778
	Psychomotricien	0.0321	0.1784	0.0008	0.0753

Coût moyen

Grand poste	Sous poste	RMSE	RMSE normalisé	MAE	MAE normalisé
Autres postes	Cure thermale	150.35	0.2924	98.38	0.3151
	Maternité Forfait	24.32	0.0649	10.38	0.0210
	Prothèses auditives	571.42	0.2041	487.07	0.2820
	Prothèses autres	24.94	0.1190	13.95	0.1814
	Prothèses orthopédiques	36.36	0.2028	27.90	0.1741
Consultations et Visites	Consultations et Visites Généralistes OPTAM	9.28	0.0950	6.95	0.1235
	Consultations et Visites Généralistes Hors OPTAM	6.85	0.0821	3.20	0.1194
	Consultations et Visites Majorations	2.54	0.1009	1.43	0.1510
	Consultations et Visites Spécialistes OPTAM	11.38	0.0598	8.14	0.1960
	Consultations et Visites Spécialistes Hors OPTAM	13.29	0.0160	8.15	0.1458
Dentaire	Implants dentaires	273.61	0.1691	196.46	0.1471
	Inlays et Onlays	128.97	0.2491	98.55	0.1240
	Orthodontie Autre	177.90	0.4627	117.55	0.7712
	Orthodontie TO90	247.25	0.1945	189.23	0.1870
	Parodontologie non prise en charge	226.66	0.4959	153.56	0.6920
	Soins dentaires	13.11	0.1190	3.69	0.1166
	SPR 50	116.03	0.1289	91.41	0.1765
	SPR 57	150.85	0.1975	85.78	0.1791
	SPR Autre	472.17	0.1920	375.47	0.4317
Hospitalisation	Chambre particulière	29.61	0.0315	15.63	0.1736
	Forfait journalier	1.85	0.0042	0.56	0.0295
	Frais de séjour	231.02	0.0602	93.47	0.1672
	Honoraires OPTAM	596.39	0.4292	150.49	0.3788
	Honoraires Hors OPTAM	127.43	0.1281	54.87	0.2221
	Transport	82.94	0.1417	31.15	0.1587
Optique	Chirurgie optique	267.94	0.3917	204.16	0.3641
	Lentilles jetables	85.54	0.4077	65.44	0.6834
	Lentilles non prises en charge	64.64	0.2257	47.93	0.1980
	Lentilles prises en charge	92.48	0.3249	58.28	0.1965
	Montures Adulte	58.91	0.6483	43.71	0.2207
	Montures Enfant	39.56	0.6778	29.68	0.1079
	Verres complexes Adulte Hors réseau 10.37	72.82	0.1765	57.75	0.1057
	Verres complexes Adulte Hors réseau 6.86	49.34	0.2465	37.88	0.1246
	Verres complexes Adulte Hors réseau 7.32	67.91	0.2128	53.91	0.2079
	Verres complexes Adulte Réseau 10.37	45.20	0.1982	36.16	0.1875
	Verres complexes Adulte Réseau 6.86	27.73	0.1430	22.89	0.1849
	Verres complexes Adulte Réseau 7.32	43.40	0.2292	34.34	0.1810
	Verres simples Adulte Hors réseau 2.29	29.99	0.2967	23.58	0.1817
	Verres simples Adulte Hors réseau 3.66	34.76	0.2890	26.61	0.1881
	Verres simples Adulte Réseau 2.29	18.49	0.2578	14.92	0.1023
	Verres simples Adulte Réseau 3.66	21.26	0.2692	17.33	0.1146
	Verres simples Enfant Hors réseau 12.04	26.30	0.1511	19.66	0.1238
	Verres simples Enfant Hors réseau 14.94	30.04	0.1575	22.57	0.1399
	Verres simples Enfant Réseau 12.04	18.15	0.1815	13.55	0.1588
	Verres simples Enfant Réseau 14.94	21.38	0.2998	17.15	0.1972
Pharmacie	Pharmacie à 15%	2.63	0.0666	0.89	0.2355
	Pharmacie à 30%	4.03	0.0632	0.96	0.2051
	Pharmacie à 65%	6.16	0.0239	2.12	0.2355
	Vaccins non pris en charge	25.39	0.4856	14.43	0.1838
Soins courants	Actes techniques médicaux OPTAM	19.65	0.1154	8.95	0.1353
	Actes techniques médicaux Hors OPTAM	23.67	0.1709	9.71	0.1670
	Acupuncture	14.70	0.1655	10.93	0.1105
	Analyses médicales	8.91	0.0916	6.10	0.1233
	Auxiliaires médicaux	4.52	0.0159	3.17	0.1013
	Majorations Soins courants	1.57	0.0738	0.88	0.1921
	Radiologie OPTAM	13.70	0.1783	10.01	0.1295
	Radiologie Hors OPTAM	30.94	0.1512	9.80	0.1255
	Chiropractie	7.40	0.1140	5.72	0.1008
	Diététique	18.59	0.2319	12.36	0.2616
	Etiopathie	5.12	0.0884	3.45	0.0656
	Microkinesie	10.36	0.1441	6.84	0.1166
	Naturopathie	3.53	0.0785	2.78	0.0393
	Ostéopathie	11.29	0.1027	7.86	0.1369
	Pédicure	3.77	0.1236	2.75	0.0817
	Podologue	8.67	0.1931	5.94	0.1615
	Psychologie	52.67	0.2962	22.84	0.2809
	Psychomotricien	38.22	0.1861	18.18	0.2284

Annexe 12 : Comparaison par sous poste sur un client

Grand poste	Sous poste	Consommation réelle totale	Consommation estimée totale	Consommation réelle par bénéficiaire	Consommation estimée par bénéficiaire	Ecart
Autres postes	Cure thermale	4 724.27 €	11 001.52 €	0.78 €	1.81 €	132.9%
	Maternité Forfait	13 720.50 €	16 599.34 €	2.26 €	2.73 €	21.0%
	Prothèses auditives	15 510.15 €	34 596.63 €	2.55 €	5.69 €	123.1%
	Prothèses autres	65 390.40 €	92 668.60 €	10.76 €	15.25 €	41.7%
	Prothèses orthopédiques	21 433.52 €	29 869.98 €	3.53 €	4.92 €	39.4%
Consultations et Visites	Consultations et Visites Généralistes CAS	23 301.88 €	26 267.29 €	3.84 €	4.32 €	12.7%
	Consultations et Visites Généralistes Hors CAS	116 584.52 €	141 930.11 €	19.19 €	23.36 €	21.7%
	Consultations et Visites Spécialistes CAS	46 087.71 €	39 290.34 €	7.59 €	6.47 €	-14.7%
	Consultations et Visites Spécialistes Hors CAS	158 259.48 €	179 985.56 €	26.05 €	29.62 €	13.7%
	Consultations et Visites Majorations	12 843.32 €	16 676.51 €	2.11 €	2.74 €	29.8%
Dentaire	Implants dentaires	102 437.00 €	105 626.74 €	16.86 €	17.38 €	3.1%
	Inlays et Onlays	8 148.06 €	6 071.68 €	1.34 €	1.00 €	-25.5%
	Orthodontie Autre	23 147.38 €	30 236.24 €	3.81 €	4.98 €	30.6%
	Orthodontie TO90	198 339.04 €	228 503.56 €	32.64 €	37.61 €	15.2%
	Parodontologie non prise en charge	31 925.80 €	44 245.54 €	5.25 €	7.28 €	38.6%
	Soins dentaires	49 821.49 €	56 884.64 €	8.20 €	9.36 €	14.2%
	SPR 50	269 359.82 €	283 578.70 €	44.33 €	46.67 €	5.3%
	SPR 57	98 493.37 €	100 102.46 €	16.21 €	16.48 €	1.6%
	SPR Autre	150 919.00 €	133 060.18 €	24.84 €	21.90 €	-11.8%
Hospitalisation	Chambre particulière	122 092.12 €	84 386.03 €	20.09 €	13.89 €	-30.9%
	Forfait journalier	131 322.00 €	79 650.63 €	21.61 €	13.11 €	-39.3%
	Frais de séjour	203 710.48 €	115 831.86 €	33.53 €	19.06 €	-43.1%
	Honoraires CAS	1 460.11 €	2 496.04 €	0.24 €	0.41 €	70.9%
	Honoraires Hors CAS	61 799.87 €	83 330.08 €	10.17 €	13.72 €	34.8%
	Transport	9 333.97 €	8 983.61 €	1.54 €	1.48 €	-3.8%
Optique	Chirurgie optique	2 000.00 €	991.08 €	0.33 €	0.16 €	-50.4%
	Lentilles jetables	0.00 €	123.78 €	0.00 €	0.02 €	-
	Lentilles non prises en charge	71 645.00 €	5 100.10 €	11.79 €	0.84 €	-92.9%
	Lentilles prises en charge	1 497.23 €	1 185.17 €	0.25 €	0.20 €	-20.8%
	Montures Adulte	148 061.51 €	174 263.42 €	24.37 €	28.68 €	17.7%
	Montures Enfant	46 974.95 €	52 114.60 €	7.73 €	8.58 €	10.9%
	Verres complexes Adulte Hors réseau 10.37	82 640.65 €	75 698.16 €	13.60 €	12.46 €	-8.4%
	Verres complexes Adulte Hors réseau 6.86	2 074.08 €	2 518.51 €	0.34 €	0.41 €	21.4%
	Verres complexes Adulte Hors réseau 7.32	20 932.74 €	17 568.68 €	3.45 €	2.89 €	-16.1%
	Verres complexes Adulte Réseau 10.37	81 827.76 €	78 900.75 €	13.47 €	12.99 €	-3.6%
	Verres complexes Adulte Réseau 6.86	2 781.88 €	2 106.81 €	0.46 €	0.35 €	-24.3%
	Verres complexes Adulte Réseau 7.32	26 747.78 €	26 435.07 €	4.40 €	4.35 €	-1.2%
	Verres simples Adulte Hors réseau 2.29	11 408.63 €	8 765.15 €	1.88 €	1.44 €	-23.2%
	Verres simples Adulte Hors réseau 3.66	28 190.43 €	22 573.82 €	4.64 €	3.72 €	-19.9%
	Verres simples Adulte Réseau 2.29	11 780.57 €	9 655.61 €	1.94 €	1.59 €	-18.0%
	Verres simples Adulte Réseau 3.66	23 204.30 €	24 463.19 €	3.82 €	4.03 €	5.4%
	Verres simples Enfant Hors réseau 12.04	6 891.98 €	5 520.74 €	1.13 €	0.91 €	-19.9%
	Verres simples Enfant Hors réseau 14.94	18 203.22 €	14 138.96 €	3.00 €	2.33 €	-22.3%
	Verres simples Enfant Réseau 12.04	3 865.06 €	3 759.88 €	0.64 €	0.62 €	-2.7%
	Verres simples Enfant Réseau 14.94	14 008.66 €	12 737.99 €	2.31 €	2.10 €	-2.7%
Pharmacie	Pharmacie à 15%	31 978.95 €	26 030.90 €	5.26 €	4.28 €	-18.6%
	Pharmacie à 30%	79 878.87 €	55 137.47 €	13.15 €	9.07 €	-31.0%
	Pharmacie à 65%	208 560.67 €	191 955.20 €	34.33 €	31.59 €	-8.0%
	Vaccins non pris en charge	2 556.78 €	2 659.02 €	0.42 €	0.44 €	4.0%
Soins courants	Actes techniques médicaux CAS	12 506.47 €	12 262.53 €	2.06 €	2.02 €	-2.0%
	Actes techniques médicaux Hors CAS	83 670.33 €	77 106.93 €	13.77 €	12.69 €	-7.8%
	Analyses médicales	83 814.73 €	115 979.25 €	13.79 €	19.09 €	38.4%
	Auxiliaires médicaux	127 791.99 €	114 854.90 €	21.03 €	18.90 €	-10.1%
	Chiropractie	4 625.00 €	4 436.02 €	0.76 €	0.73 €	-4.1%
	Diététique	1 278.00 €	3 360.44 €	0.21 €	0.55 €	162.9%
	Majorations Soins courants	2 194.48 €	2 051.80 €	0.36 €	0.34 €	-6.5%
	Microkinesie	290.00 €	106.64 €	0.05 €	0.02 €	-63.2%
	Ostéopathie	63 000.46 €	70 339.58 €	10.37 €	11.58 €	11.6%
	Radiologie CAS	16 575.79 €	20 888.10 €	2.73 €	3.44 €	26.0%
	Radiologie Hors CAS	132 609.18 €	117 361.40 €	21.83 €	19.32 €	-11.5%
Autres	Autres sous postes	34 832.05 €	36 355.28 €	5.73 €	5.98 €	4.4%
Total		**3 431 065 €**	**3 341 381 €**	**564,71 €**	**549,95 €**	**-2,6%**

TABLES DES FIGURES ET DES TABLEAUX

Table des figures

Table des tableaux