INSTITUT DES ACTUAIRES

Lyon 1

ISFA GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES
INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

Mémoire présenté le :

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

Par : Thomas LE HO

Titre : Reporting santé : comment le perfectionner à l'aide de la Data Science ?

Confidentialité : ☐ NON ■ OUI (Durée : ☐ 1 an ■ 2 ans)

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : Galea & Associés

Signature :

Directeur de mémoire en entreprise :

Nom : Léonard Fontaine

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : Segmentation de portefeuille, GLM, *Machine Learning*, Arbres de décision, modèles agrégés, assurance complémentaire, santé

Data Science, actuariat et santé. Voici les trois notions choisies si ce mémoire devait être résumé en quelques mots.

Les capacités technologiques augmentant à une vitesse vertigineuse, les outils informatiques possèdent aujourd'hui des capacités de stockage (de l'ordre de quelques tera-octects) inimaginables il y a encore dix ans. Les bases de données se multiplient et s'agrandissent : on parle alors de *Big Data*. Devant cette masse nouvelle et importante de données, les modèles actuarielles classiques deviennent alors perfectibles : la discipline de la *Data Science* est née. Celle-ci peut être définie comme un mélange entre les mathématiques, les statistiques et l'informatique et emploie des méthodes d'un nouveau genre. A l'origine évoquée dans des domaines tels que l'astronomie ou la génétique, ces notions peuvent désormais s'élargir à presque toutes sortes d'applications (politique, sport, ...). Dans le domaine de l'assurance et de l'actuariat, ces différentes notions ont fait leur apparition plus récemment au début des années 2010. Néanmoins, malgré leur relatif jeune âge, leurs possibilités d'utilisation sont potentiellement nombreuses. Tarification, suivi des risques, détection de phénomènes averses (fraude, résiliation, ...), ... Les actions éventuelles des *Data science* dans le domaine de l'assurance sont nombreuses.

Dans ce mémoire, nous nous intéressons à l'analyse de la consommation santé d'une population couverte par un régime complémentaire collectif. Pour ce faire, nous mettons tout d'abord en place une tarification reposant sur des méthodes actuarielles classiques afin d'analyser les risques sous-jacents. Or, une partie importante des risques en assurance santé est la problématique des gros consommants. En effet, une idée communément admise dans le milieu de l'assurance santé est que 20 % d'un portefeuille contribue à près de 80 % des dépenses engagées. Il peut donc être tout à fait intéressant, dans un second temps, d'exhiber des caractéristiques propres à ces gros consommants, à l'aide de méthodes *Data Science*.

Ce modèle étant le plus utilisé dans la tarification, un modèle linéaire généralisé (ou GLM) a été tout d'abord implémenté. Celui-ci est alors comparé aux autres modèles de type *Data Science* afin de cerner les apports et les limites de chacun.
Les différents modèles testés sont les arbres de décision de type CART, et deux méthodes d'agrégation d'arbres de décisions : les *Random Forest* et l'*eXtreme Gradient Boosting* (ou XGBoost). L'étude précise de ces différents modèles sur un poste de santé relatif aux consultations chez les médecins généralistes et spécialistes permet de choisir le modèle le plus approprié et de le généraliser aux autres postes et à l'étude des gros consommants.

Au-delà des buts évoqués précédemment, ce mémoire vise avant tout à améliorer le reporting santé. De plus, l'utilisation de techniques issues des *Data Science* doit permettre d'affiner l'analyse de la consommation médicale et d'identifier de potentielles dérives sur des groupes d'assurés présentant certaines caractéristiques.
Pour ce faire, de nombreuses variables, dont certaines d'origine externe, sont ajoutées à la base de données initiale afin d'exploiter au mieux le potentiel des modèles *Data Science*.
Enfin, un reporting santé un peu particulier, mêlant statistiques descriptives et résultats de modèles *Data Science*, est présenté.

L'analyse des risques effectuée permet de mettre en évidence certaines variables externes comme contributives dans l'explication de la consommation santé. De plus, l'étude des gros consommants met en avant de nouvelles variables qui n'étaient pas importantes dans l'analyse de la consommation santé.

Abstract

Keywords : Portofolio segmentation, GLM, Machine Learning, Decision trees, Aggregation methods, supplementary insurance, health

---

Data Science, actuary and health. If this master thesis should be sum up in a few words, these words would be choosen.

Technological capabilities increasing fastly, software tools have storage capacities (in the order of some tera-octets) unthinkable a decade ago. Databases become more numerous and wider : we speak about Big Data. With this new and significant data mass, classical statistics models have become unsuitable : Data Science branch was born. It can be described as a combination of mathematics, statistics and computer sciences and uses advanced methods. Whereas astronomy or genetics were the first fields of Big Data, this notion has almost applied to all kinds of application (politics, sport, ...). In the context of insurance and actuarial science, theses notions had been developped recently, in the early 2010s. Nevertheless, despite their early age, possibilities for using could be plentiful. Pricing, risks monitoring, averse behaviours detect (fraud, termination, ...). Potential actions of Data Science in insurance are numerous.

In this master thesis, we care about healthcare consumption analysis of a population covered by a compulsory collective contract. To that end, at first, we implement a pricing based on classical actuarial methods in order to analyze subjacent risks. Besides, an important part of risks in health insurance is large consumers issue. Indeed, a generally accepted idea in health insurance is the next one : 20 % of the portofolio amount to 80 % of expenditures. Then, it can be useful to extract specific features of large consumers, thanks to Data Science models.

As it is the most widely used model in pricing, a GLM model had firstly been implemented. This model is compared to other Data Science models in order to understand beneficits and limits.
The models tested are CART decision tree and two aggregated models : Random Forests and eXtreme Gradient Boosting (or XGBoost). The precise study of those models on medical consultations post enable the choice of the most accuracy model and its wider application to other health post and large consumers study.

Beyond goals raised precedently, this master thesis targets health reporting improvement. Moreover, Data Science technics use should permit a best analysis of health consumption and the identification of potential drifts.
In order to maximize Data Science models, some variables, including external variables, are added to the initial database.
Finally, a particular health reporting, mixing descriptive statistics and results of Data Science models, is submitted.

Risks analysis higlights some significant external variables in the explanation of health consumption. Moreover, large consumers study shows new variables which are not important in health consumption analysis.

Remerciements

Avant toute chose, je tiens à exprimer ma reconnaissance au cabinet d'Actuaires Conseil GALEA & Associés pour m'avoir accordé leur confiance et m'avoir accueilli ces deux dernières années. Les travaux proposés, la disponibilité de chacun ainsi que l'ambiance de travail, à la fois sérieuse et conviviale, m'ont permis de découvrir et de comprendre les enjeux du métier d'actuaire tout en m'épanouissant pleinement dans celui-ci.

Ce mémoire n'aurait pu voir le jour sans l'aide de plusieurs personnes. Je remercie particulièrement Norbert GAUTRON pour ses idées nombreuses et pertinentes, Florence CHIU et Léonard FONTAINE pour leur assistance quotidienne et leurs réponses à mes questions, ainsi que Mylène FAVRE-BEGUET et Perrine CAROLO pour leur grande connaissance de l'assurance santé.

La formation reçue à l'ISFA m'a évidemment été d'une grande aide tout au long de la réalisation de ce mémoire. Je remercie l'ensemble les professeurs que j'ai pu croisé durant ces trois années et tout spécialement Christian ROBERT pour ses remarques toujours justes.

Enfin, mes pensées vont à toutes les personnes qui ont participé de près ou de loin à la réalisation de ce mémoire.
Merci à ma famille pour leurs encouragements.
Merci à Juliette pour ses très nombreuses relectures et son soutien quotidien.

Table des matières

Introduction

De nos jours, une des problématiques majeures des assureurs est la connaissance de leur portefeuille d'assurés et de leurs comportements. Une bonne connaissance client permet de proposer des produits adaptés à ses assurés et ainsi faciliter les étapes de tarification ou de provisionnement associées et la gestion des risques sous-jacents.
Les modèles *Data Science* sont très souvent associés à la notion de *Big Data* et de ce fait à des bases de données de grandes dimensions. Quelles seraient donc les sources potentiellement à disposition des assureurs leur permettant d'enrichir leurs bases de données et par conséquent leur connaissance client ?
Nous pensons tout d'abord, et de façon instinctive, aux *open sources* disponibles sur internet, véritable mine d'or en ce qui concerne la donnée. Pourquoi ne pas utiliser à terme les données des réseaux sociaux ou celles de navigation des internautes ? Le terme de *Big Data* englobe également l'ensemble des données numériques. Ainsi, les données fournies par les objets tels que les GPS ou les montres connectées entrent parfaitement dans le cadre du *Big Data* et dans l'amélioration de la connaissance client.
En effet, de nombreuses applications existent déjà dans ce domaine. AXA et Allianz ont lancé, en 2015, une offre d'assurance automobile connectée, plus communément appelée *Pay how you drive* qui peut permettre de faire baisser sa prime d'assurance en cas de bonne conduite. De plus, selon une étude menée par Accenture auprès de ses clients assureurs français [3], plus de 30% d'entre eux possèdent ou sont en train de lancer une telle offre. Il est possible d'imaginer que l'assurance habitation, via des "box domotiques" qui alerteront en cas de sinistres l'assuré, l'assureur ou les secours, pourrait suivre le même chemin.

Les assureurs semblent donc réceptifs à l'introduction du *Big Data* dans leur quotidien. Les résultats d'une étude parue sur le site de l'argus de l'assurance en mai 2014 [2] viennent corroborer ces propos : pour près de 70 % des assureurs, le *Big Data* est une priorité à l'horizon 2018.
L'assurance santé ne semble pas, à première vue, faire partie des domaines d'assurance les plus propices à l'intégration du *Big Data* et à l'utilisation des données individuelles. En effet, les données santé ont toujours été très sensibles, du moins en France et seront encadrées par le Règlement Général sur la Protection de Données (RGPD) qui rentrera en vigueur en 2018. Selon des chiffres dévoilés durant le *hub Tday insurance* en 2016 [8], seulement 34 % des répondants seraient favorables à communiquer des données santé à un assureur. A titre de comparaison, plus de 70 % d'entre eux pourraient être intéressés par une offre d'assurance automobile connectée.
Une étude publiée par la DREES [12] montre que de nombreux autres pays ont commencé à mettre en place des modèles prédictifs afin de prévoir la survenance d'événements comme l'hospitalisation non programmée ou l'entrée dans un établissement d'hébergement pour personnes âgées dépendantes. Aux États-Unis ou au Royaume-Uni, ces modèles sont d'ailleurs utilisés afin d'allouer des ressources aux assureurs et aux organisations des soins.

Les problématiques autour de la santé sont pourtant nombreuses en France. Le vieillissement de la population, l'apparition de déserts médicaux sont autant de phénomènes qui poussent les acteurs de l'assurance santé à innover.
Le domaine de la santé a également été récemment touché par des réformes législatives. Si l'on prend exemple du poste dentaire, dans une annonce faite le 9 mars 2017, les prix des prothèses dentaires seront plafonnés à partir de 2018. Dans le même temps, la base de remboursement de la Sécurité Sociale des implants dentaires passera de $107,5$ € à 120 €. Le président Macron souhaite également instaurer le remboursement intégral des soins optiques et dentaires.
Ces nombreux changements vont évidemment avoir des répercussions chez les organismes complémentaires. De nouveaux paramètres vont rentrer en compte dans les travaux réalisés et la *Data*

Science peut apporter des réponses aux nouvelles problématiques posées.
Néanmoins, les assureurs pourraient être confrontés à la réticence des assurés à partager leurs données personnelles et devront respecter le prochain RGPD. A ce titre, il est stratégique pour l'assureur d'appréhender les données réellement contributives à ces analyses et ainsi concentrer ses efforts sur la collectes de ces dernières.

La réalisation de ce mémoire suit deux objectifs. Il vise avant tout à mieux comprendre la consommation en santé et à améliorer le reporting santé actuel. L'intégration de variables d'origine externe peut être un moyen de parfaire le reporting santé. Ajouter de nouvelles variables entraîne inévitablement une augmentation des dimensions de la base de données. L'utilisation de techniques issues des *Data Sciences*, adaptées à des bases de grandes dimensions, doit permettre de traiter les effets de ces variables additionnelles.

Deux études seront donc menées au cours de ce mémoire.
Dans un premier temps, la consommation des individus selon les différents postes en santé sera étudiée à l'aide de quatre modèles. Cette consommation sera modélisée via une méthode "Fréquence / Coût moyen". Néanmoins, pour s'astreindre des plafonds de garanties, ce ne sont pas les remboursements moyens de l'organisme complémentaire qui seront étudiés mais les frais réels moyens. Les résultats des différents modèles seront analysés de manière précise au niveau de la famille "Consultation visites", représentant les actes liés aux médecins généralistes et spécialistes et permettront de choisir le modèle le plus adapté à cette étude. Ce modèle sera ensuite généralisé aux autres postes de santé.
Cette tarification, qui n'est pas une tarification classique, s'apparente à une analyse des risques du portefeuille. La deuxième étude, effectuée à l'aide du modèle *Data science* sélectionné précédemment, s'intéressera aux caractéristiques des gros consommants du portefeuille, une des problématiques majeures de l'assurance santé. Un individu sera considéré comme gros consommant si ses dépenses font partie des 20 % des plus hautes dépenses engagées.
Les résultats de ces deux études seront incorporés dans un reporting santé présentant également des statistiques descriptives. Ce reporting sera également l'occasion de mettre en valeur l'apport des variables externes dans la compréhension de la consommation en santé.

En résumé

Le premier chapitre présente quelques éléments contextuels concernant les domaines de la santé et du *Big Data*.

La deuxième partie explicite les traitements effectués afin de rendre les bases de données initiales propres et utilisables dans le cadre de notre étude.

Le troisième chapitre pose les bases théoriques indispensables à la compréhension des modèles utilisés dans le cadre de notre modélisation de la consommation médicale.

Enfin, le quatrième et dernier chapitre est dédié à l'application des modèles théoriques et à la mise en place de la démarche décrite précédemment.

1 Éléments contextuels

Ce premier chapitre va permettre d'introduire quelques notions intéressantes concernant deux des principaux sujets abordés dans ce mémoire, à savoir la santé et la mise en place de modèles *Data Science*.

1.1 Panorama de la santé en France

Cette première partie vise à présenter de manière plus précise quelques informations relatives à la santé en France. Les chiffres ainsi que les évolutions législatives récentes permettront de mieux appréhender les nombreuses problématiques actuelles liées à la santé.

1.1.1 Le système de soins français

Présentation

Le système de soins français est composé de deux parties :

- Le régime obligatoire ;
- Le régime complémentaire.

Le **régime obligatoire** est la Sécurité Sociale, mise en place en 1945, qui intervient principalement dans le remboursement de Frais Médicaux (on parle d'Assurance Maladie), mais également dans d'autres domaines tels que la retraite, l'incapacité, l'invalidité ou le décès.

Le **régime complémentaire**, proposé par les assureurs, intervient en plus du régime obligatoire qui, dans la quasi-totalité des cas, n'est pas en mesure de rembourser l'ensemble des frais médicaux engagés par les personnes couvertes.

Pour bien saisir le fonctionnement du système de soins en France, il est nécessaire de définir les différentes variables intervenant dans les remboursements :

- Les **Frais Réels (FR)** : cette variable désigne le montant global dépensé par un individu pour un acte médical déterminé, comme par exemple le prix d'une consultation chez un généraliste ;
- La **base de Remboursement de la Sécurité Sociale (BRSS)** : pour un acte médical, elle correspond à un montant référence, exprimé en euros, remboursé totalement ou partiellement par la Sécurité Sociale. Le montant effectivement remboursé ou Remboursement Sécurité » Sociale (RSS) étant déterminé par un taux de remboursement appliqué sur la BRSS ;
- Le **ticket Modérateur (TM)** : la différence entre la BRSS et le RSS, *i.e.* la part du montant BRSS non remboursé par la Sécurité Sociale ;
- Le **montant remboursé par l'organisme complémentaire** : il dépend des niveaux des garanties souscrites par l'assuré et peut notamment comprendre la prise en charge du ticket modérateur ;
- Le **reste à Charge (RAC)** : c'est le montant restant à régler par l'assuré pour rembourser les frais réels de ses soins, après remboursement de la Sécurité Sociale et de son organisme complémentaire.

Un schéma sera néanmoins plus parlant que n'importe quelle explication manuscrite.

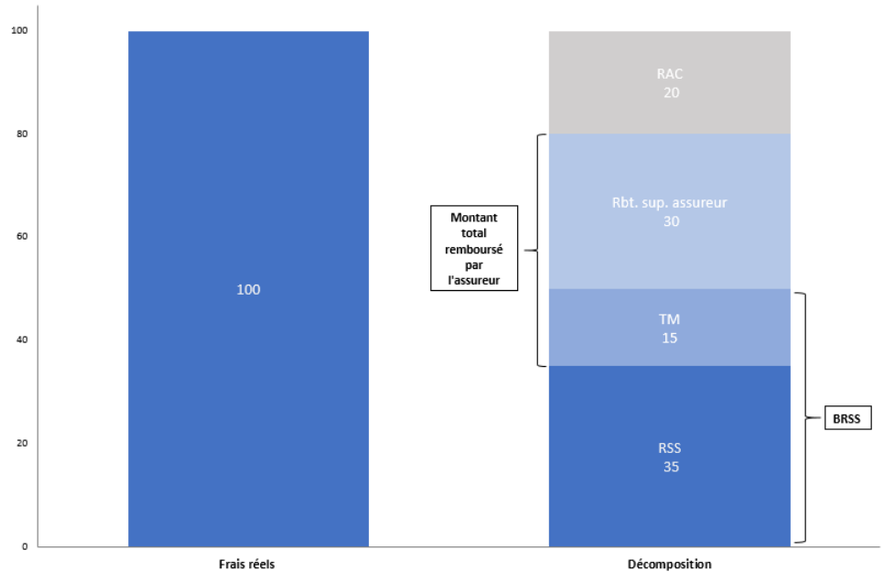


FIGURE 1.1 – Exemple de décomposition de dépenses de santé

N.B : Les chiffres ci-dessous sont mentionnés à titre purement illustratif.

La prise en charge des dépenses de santé

Quatre acteurs interviennent dans les dépenses de santé. Le tableau suivant récapitule la prise en charge de ces différents acteurs :

Acteurs	Prise en charge
Sécurité sociale	77 %
Etat, CMU-C	1,4 %
Organismes complémentaires	13,3 %
Ménages	8,3 %

FIGURE 1.2 – Prise en charge des dépenses de santé selon les différents acteurs

Source : DREES

La part globale de prise en charge de ces différents organismes évolue peu depuis de nombreuses années. Cependant, cette structure de prise en charge varie fortement d'un poste à l'autre. En effet, les niveaux de recouvrement de la Sécurité sociale présentent des disparités importantes selon le poste considéré.

Le graphique ci-après illustre ces variations parfois importantes :

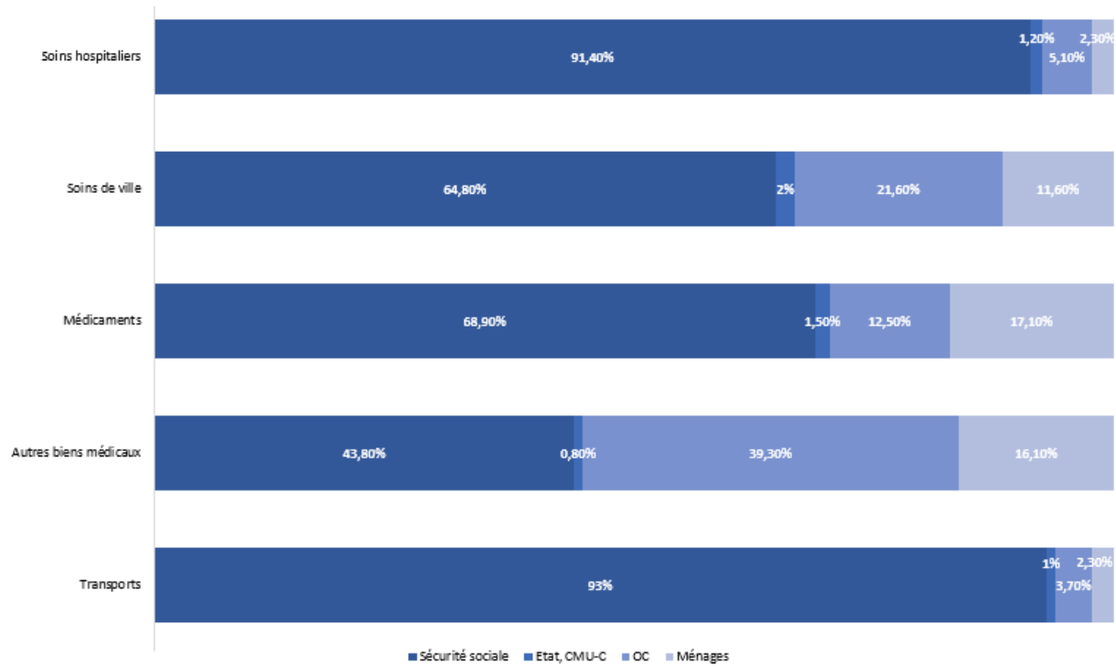


FIGURE 1.3 – Prise en charge des dépenses de santé selon les différents postes

Source : DREES

N.B : Les différents postes présentés dans le schéma ci-dessus sont les composantes de la Consommation de Soins et Biens Médicaux (CSBM), notions qui va être développé dans la partie suivante.

C'est donc pour le poste "Autres biens médicaux" que la contribution de la Sécurité sociale est la plus basse. En effet, cette composante contient le poste "Optique" dont la prise en charge par la Sécurité sociale est extrêmement faible (moins de 5 % du total des dépenses en 2016). Par conséquent, c'est également pour ce poste que la part des organismes complémentaires est la plus élevée. Ces organismes prennent en charge près des trois quarts des dépenses liées à l'optique.
Au contraire, les "Soins hospitaliers" sont particulièrement bien remboursés par la Sécurité sociale. En effet, la base de remboursement de la Sécurité sociale s'élève à 80 % des frais réels. De plus, de nombreux cas existent où le taux de prise en charge passe à 100 %, en cas d'hospitalisation de longue durée par exemple.

1.1.2 Étude des dépenses de santé

La très grande majorité des informations mentionnées dans ce paragraphe provient de l'étude intitulée "Les dépenses de santé en 2016 - Résultats des comptes de la santé" publiée par la Direction de la Recherche, des études et de l'Evaluation et des Statistiques (DREES). Ce document présente une estimation précise de la consommation finale relative à la santé en France et les financements correspondants et introduit deux agrégats comptables relatifs au domaine de santé, à savoir la Consommation de Soins et Biens Médicaux (CSBM) et la Dépense Courante de Santé (DCS).

La Consommation de Soins et Biens Médicaux

La CSBM synthétise la dépense de biens et de services médicaux consommés en vue de traiter une perturbation provisoire de l'état de santé.
Cette dépense inclut donc l'ensemble des biens médicaux et soins courants, y compris ceux des personnes pris en charge au titre des affections de longue durée (ALD) [1] ; elle exclut en revanche diverses composantes de la dépense relatives notamment à la gestion et au fonctionnement du système de santé. De plus, les dépenses de soins aux personnes handicapées et aux personnes âgées en institution sont exclues puisque seules les dépenses qui concourent au traitement d'un trouble temporaire de l'état de santé sont prises en compte.

1. Selon le site d'Ameli, les *ALD* "nécessitent un traitement prolongé d'une durée prévisible supérieure à six mois et une thérapie particulièrement coûteuse."

La CSBM présente cinq composantes :

— Les soins hospitaliers ;
— Les soins de ville (médecins, dentistes, auxiliaires médicaux, ...) ;
— Les transports de malades ;
— Les médicaments en ambulatoire ;
— Les autres biens médicaux (optique, prothèses, ...).

Voici la ventilation de la CSBM selon ses cinq composantes :

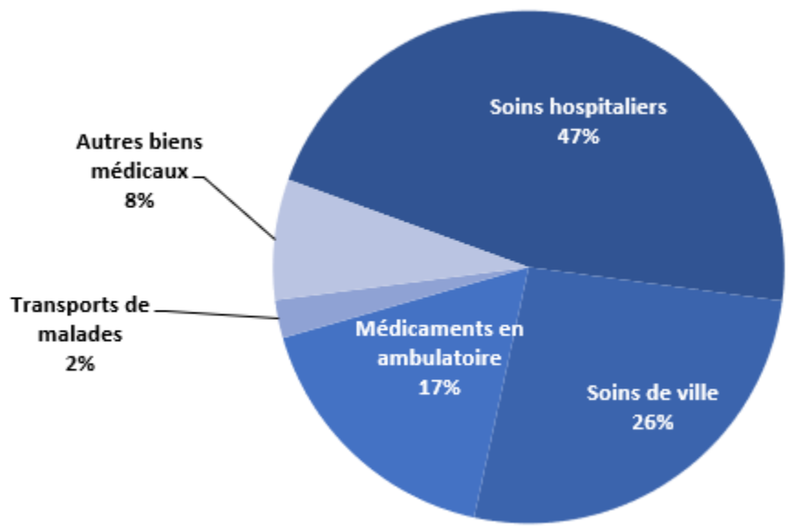


FIGURE 1.4 – Ventilation de la CSBM selon ses différentes composantes en 2016
Source : DREES

En 2016, la CSBM s'élève à 198,5 milliards d'euros. Ce montant représente environ 2 970 € par habitant et pèse pour près de 9 % dans le PIB. Celle-ci progresse en valeur de 2,3 % par rapport à 2015.

Tout au long de cette partie, les notions de variation en valeur, en volume ou encore en prix seront évoquées. Sans rentrer plus dans le détail, la variation en volume peut être définie comme la variation en valeur corrigée de l'effet de l'inflation (ou effet prix). Plus concrètement, en 2016, la CSBM augmente de 2,9 % en volume car les prix relatifs aux dépenses de santé diminuent de 0,6 % par rapport à 2015. Cette progression n'est pas surprenante puisque globalement similaire à celles observées depuis 2008.

Une analyse de la variation de la CSBM en fonction de ces différentes composantes peut être intéressante afin de mettre en évidence les différents éléments pouvant influer sur les dépenses de santé et de connaître plus en détail la composition des différents postes de santé.

Les soins hospitaliers

La croissance des dépenses relatives aux soins hospitaliers progresse de 2,2 % en valeur. Cette progression est portée par la hausse des volumes (+2,0 %), tandis que les prix sont restés plutôt stables (+0,2 %).
Dans le secteur public, qui représente environ 80 % des soins hospitaliers, la consommation progresse en valeur de 2,2 %, soit une augmentation moindre qu'au cours de la période 2010-2014 (+2,7 % en moyenne). Ce rythme moins élevé s'explique par le resserrement de l'Objectif National de Dépenses d'Assurance Maladie (ONDAM), qui peut être défini comme une limite de dépenses en matière de soins de ville et d'hospitalisation dispensés dans les établissements privés ou publics,

mais aussi dans les centres médico-sociaux.
La consommation de soins dans le secteur privé est en hausse de $2,1$ %, entièrement portée par l'augmentation des volumes ($+3,4$ %). En revanche, les prix ont fortement diminué ($-1,3$ %). Cette baisse des prix peut s'expliquer par les coupes tarifaires appliquées par les établissements privés ainsi qu'à la modération des dépassements d'honoraires due principalement à la réforme des contrats responsables qui limite la prise en charge des dépassements par les organismes complémentaires.

L'ensemble des soins ambulatoires, qui correspond aux dépenses liées aux quatre autres composantes de la CSBM, progresse quant à eux de $2,4$ % en valeur, soit un rythme nettement plus élevé par rapport à l'année 2015 ($+1,6$ %). Cette accélération concerne chacun des postes de dépenses en ambulatoire en 2016 à l'exception des médicaments qui sont stables.

Les soins de ville

La consommation des soins de ville augmente de $3,3$ % en valeur contre $2,3$ % en 2015. Les soins d'auxiliaires médicaux sont toujours autant prisés ($+4,4$ %). Cette hausse est entièrement portée par l'effet volume.
Les consommations des soins de médecins généralistes et spécialistes augmentent également de façon significative. De la même manière que pour les auxiliaires médicaux, c'est la hausse des volumes du fait d'un contexte épidémiologique plus important (forte épidémie de grippe en décembre) et de l'augmentation du nombre d'actes techniques (scanners, IRM, ...) pour les spécialistes qui explique cette augmentation. La consommation des soins dentaires augmente également de $2,9$ % en valeur. Cette hausse est principalement expliquée par la multiplication des actes chirurgicaux et orthodontiques.
Enfin, la consommation d'analyses et prélèvements en laboratoires progresse de $2,7$ %, portée essentiellement par l'augmentation du nombre d'actes.
L'augmentation de la consommation des soins de ville est donc dans la très grande majorité des cas expliquée par la hausse des volumes.

Les médicaments

Au contraire, la consommation de médicaments reste globalement stable. En effet, la diminution des prix ($-3,7$ %) compense la hausse des volumes ($+4,1$ %). Cette baisse des prix peut être grandement attribuée à la baisse des tarifs des médicaments remboursables et par «l'effet générique». En effet, les médicaments génériques, à coût moindre, sont de plus en plus utilisés.

Les autres biens médicaux

Enfin, la consommation des autres biens médicaux progresse de $3,6$ % en valeur. Néanmoins, deux tendances au sein de ce poste peuvent être distinguées.
La dépense d'optique médicale, qui augmentait de façon importante dans les années 2000, a nettement ralenti depuis quelques années, tant en prix qu'en volume. Deux principales raisons peuvent être à l'origine de cette tendance. Confrontés à une croissance très dynamique il y a encore une dizaine d'année, certains organismes complémentaires ont décidé de limiter leurs garanties. La réforme des contrats responsables, entrée en vigueur en 2015 et dans laquelle les dépenses d'optique ne sont remboursées que tous les deux ans, accentue ce phénomène.
Au contraire, la croissance de la consommation de prothèses reste très forte, de l'ordre de 6 % en valeur.

La Dépense Courante de Santé

La DCS est le deuxième agrégat comptable relatif aux dépenses de santé. Il est défini comme la somme de toutes les dépenses « courantes » engagées par les différents financeurs (détaillés dans le paragraphe précédent) pour la fonction santé.
La DCS comprend donc :

- La CSBM ;
- Les soins de longue durée prodigués aux personnes âgées et aux personnes handicapées ;
- Les indemnités journalières (en cas de maternité ou d'accident du travail par exemple) ;
- Les dépenses de prévention institutionnelles (vaccins, dépistages, lutte contre la pollution, ...) ;
- Les dépenses en faveur du système de soins (recherches et formations principalement) ;
- Les coûts de gestion.

La DCS est donc une notion plus large que la CSBM. Cette notion ne sera pas plus détaillée dans ce mémoire. En effet, la présentation de la CSBM suffit à introduire les notions nécessaires à la compréhension du cadre de la santé en France.

En résumé

Il existe donc deux principaux agrégats comptables en matière de santé :

- La **CSBM** qui synthétise la dépense de biens et de services médicaux consommés en vue de traiter une perturbation provisoire de l'état de santé.
  Valeur 2016 : 198,5 milliards d'euros
  Évolution par rapport à 2015 : $+2,3$ %
  Évolution des principales composantes de la CSBM :
  - Soins hospitaliers : $+2,2$ % ;
  - Soins de ville : $+3,3$ % ;
  - Médicaments : $+0,3$ % ;
  - Autres biens médicaux : $+3,6$ %.
- La **DCS** qui est définie comme la somme de toutes les dépenses « courantes » engagées par les différents financeurs dans le domaine de la santé. Cet agrégat, plus global, regroupe :
  - La CSBM ;
  - Les soins de longue durée prodigués aux personnes âgées et aux personnes handicapées ;
  - Les indemnités journalières ;
  - Les dépenses de prévention ;
  - Les dépenses en faveur du système de soins (recherches et formations principalement) ;
  - Les coûts de gestion.

1.1.3 Évolutions législatives récentes

Cette partie présente quatre évolutions réglementaires qui ont touché le domaine de l'assurance santé.

L'Accord National Interprofessionnel (ANI)

Cet accord, conclu le 11 janvier 2008, instaure le principe de portabilité : les anciens salariés peuvent, sous conditions, continuer de bénéficier des garanties santé de la mutuelle de leur ancienne entreprise pendant une certaine durée. Ce dispositif a été rendu obligatoire depuis le 1^{er} juin 2014. Sous conditions car le bénéficiaire doit satisfaire trois conditions pour bénéficier du maintien de ses garanties :

- La rupture de son contrat de travail ne doit pas être consécutive à un licenciement pour faute lourde ;
- Il doit bénéficier du droit à indemnisation auprès du régime d'assurance chômage à la suite de cette rupture ;
- Il doit également être bénéficiaire des garanties santé avant la rupture du contrat de travail.

La période de maintien des garanties est égale à la période d'indemnisation du chômage dans la limite de la durée du dernier contrat de travail sans pouvoir excéder un an. La portabilité cesse en cas de reprise d'une activité professionnelle, en cas de liquidation de la pension de retraite ou en cas de radiation des listes de Pôle emploi.

Les contrats responsables

Ce dispositif, contenu dans un décret issu de la Loi de Financement de la Sécurité Sociale 2014 paru au Journal Officiel le 19 novembre 2014 et entré en vigueur le 1^{er} avril 2015 [2], s'applique à l'ensemble des contrats d'assurance maladie complémentaire, qu'ils soient individuels ou collectifs. Comme mentionné dans une circulaire de la Direction de la Sécurité sociale, il a pour objectif *"en contrepartie du bénéfice d'aides fiscales et sociales (...), de renforcer les exigences imposées aux contrats d'assurance maladie complémentaires (collectifs ou individuels) tant en terme de qualité de niveau de couverture procuré qu'en terme de cohérence avec la politique d'accès aux soins menée par le Gouvernement".*

Ce dispositif s'articule autour de trois axes principaux :

- Fixer des planchers de prise en charge ;
- Réguler les dépassements d'honoraires ;
- Agir sur les prix de l'optique.

Les principales mesures de ce dispositif sont les suivantes :

- Prise en charge de manière illimitée du forfait journalier hospitalier y compris pour les séjours psychiatriques ;
- Prise en charge obligatoire du ticket modérateur à l'exception des cures thermales et des médicaments remboursés à 15 % et 30 % ;
- Si le contrat prévoit la prise en charge des dépassements d'honoraires, celle-ci doit être différenciée entre les praticiens signataires du Contrat d'Accès aux Soins (CAS) et ceux qui ne le sont pas. Le CAS, mis en place en 2012, contient une liste de règles qu'un médecin signataire s'engage à respecter. Plus précisément, celui-ci s'engage à recevoir plus de patients sans dépassements d'honoraires et à limiter la part de son activité sujette à ces dépassements. Sous ces conditions, il bénéficie d'avantages sur leurs cotisations sociales. La garantie des dépassements d'honoraires de médecins non signataires du CAS est soumise à un double plafond et ne peut donc excéder :
  - Le niveau des garanties des dépassements d'honoraires des médecins signataires du CAS minoré de 20 % de la Base de Remboursement de la Sécurité Sociale (BRSS) ;

2. Une période transitoire d'adaptation est prévue pour les contrats collectifs jusqu'au 31 décembre 2017.

- 100 % de la BRSS (porté de manière dérogatoire à 125% de la BRSS pour 2015 et 2016).

— Respect de planchers et plafonds de garanties fixés en fonction du type de verre, avec limitation à un équipement (2 Verres et 1 Monture) tous les deux ans (sauf si mineurs ou changement de correction pour lesquels la limitation est ramenée à un équipement par an).

Panier de soins de l'Accord National Interprofessionnel (ANI)

Ce dispositif, paru au Journal Officiel le 10 septembre 2014 et appliqué à partir du 1^{er} janvier 2016, s'applique uniquement aux contrats collectifs.

Le panier de soins ANI ne prévoit que des garanties minimales et aucun plafond. De par la parution du décret sur les contrats responsables, les contrats ANI sont soumis aux règles de plafonnement des contrats responsables.

Les principales modifications de ce dispositif sont :

— L'augmentation des planchers de prise en charge en optique ;

— La prise en charge des soins dentaires prothétiques et des soins d'orthopédie dento-faciale à hauteur de 125 % de la BRSS y compris Sécurité Sociale.

Loi de modernisation du système de santé

Cette loi, promulguée le 26 janvier 2016, a trois objectifs principaux :

— La prévention ;

— L'accès aux soins ;

— L'innovation.

Cette dernière loi diffère des dispositifs des contrats responsables ou du panier de soins ANI dans le sens où ce ne sont plus les assureurs ou les entreprises qui sont directement visés mais bien les ménages. De plus, la loi de modernisation du système de santé encourage plus que n'oblige les ménages à modifier leurs habitudes de consommation en matière de santé. De ce fait, la notion de prévention devient réellement au centre du débat.

En résumé

Le schéma ci-après permet de résumer les principaux apports des lois les plus récentes dans le domaine de la santé :

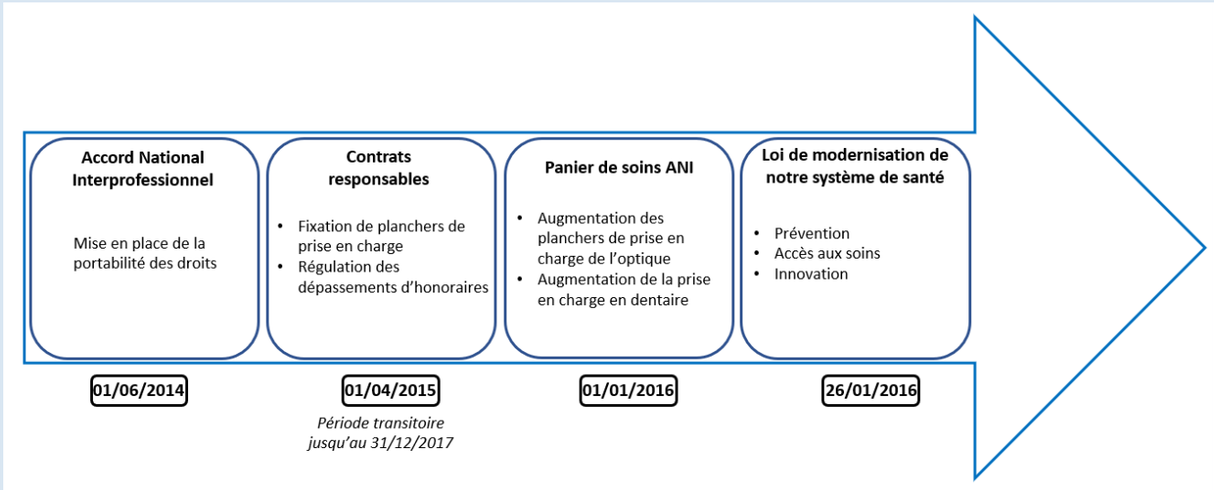


FIGURE 1.5 – Évolutions législatives en matière de santé

1.2 Problématiques liées à l'utilisation des bases de données

Cette deuxième partie a pour objectif de présenter comment les modèles *Data Science* ont pu se démocratiser. Cette multiplication des données disponibles a néanmoins amené un certain nombre de lois visant à protéger la vie privée des individus.

1.2.1 *Big Data*,*Data Science*, *Data Scientist* : quelques précisions

Les modèles *Data Science* supposent de posséder un gros volume de données. Le terme de *Big Data*[3], très présent dans de nombreux domaines ces dernières années, n'aurait pu exister sans l'augmentation des capacités de stockage de nos ordinateurs.
Il convient de dégager plusieurs étapes importantes dans l'évolution des capacités de stockage et de traitement des bases de données :

Date	Evénements
1950	Apparition des premiers ordinateurs. Les bases de données ne comportent que quelques centaines de lignes et un petit nombre de variables. Les bases pèsent alors que quelques octets.
1970	La puissance des ordinateurs permet de manipuler des bases de données en kilooctets. Les techniques d'analyse des données commencent à se développer.
1980	Les bases de données peuvent atteindre plusieurs mégaoctets. Les techniques de *Data Mining* (ou exploration des données) telles que les réseaux de neurones font leur apparition.
1990	On assiste à la première évolution majeure. Les données ne sont plus planifiées mais préalablement acquises et stockées. Les modèles de *Machine Learning* font leur apparition.
2000	Le nombre de variables explose. On manipule désormais des bases de données pouvant aller jusqu'à plusieurs téraoctets.
2010	La capacité de stockage ne cesse d'augmenter, le nombre de lignes explose également.

FIGURE 1.6 – Evolution de la taille des bases de données

Source : DREES

Les technologies actuelles sont donc capables de stocker et de traiter des bases de données de tailles très importantes. Néanmoins, une simple machine ne peut effectuer l'ensemble du travail. De par la connaissance des sujets abordés et la maîtrise des outils informatiques, l'intervention d'un expert est indispensable à la bonne réalisation d'un projet *Data Science*.

Les étapes indispensables à la bonne mise en place d'un projet *Data Science* sont décrites ci-après :

1. **Compréhension du périmètre de l'étude** : voir quelles sont les connaissances sur le sujet et bien cerner les besoins du commanditaire du projet ;
2. **Extraction de la base de données** sur laquelle le *Data Scientist* va pouvoir mettre en place les modèles ;
3. **Nettoyage de la base** (erreurs, fautes de saisies, valeurs aberrantes, ...) ;
4. **Exploration des données** : étude des distributions des variables, de leurs interactions, ... ;
5. **Transformation des données** : réduction de dimension (dans le cadre d'une *ACP* par exemple), découpage en classes, ... ;
6. **Choix des modèles**. Selon le but final de l'étude (interprétabilité, prédictibilité ; ...), les modèles mis en œuvre ne seront pas forcément les mêmes.
7. **Tests sur les différents modèles** : on vérifie notamment leur qualité d'ajustement et leur qualité de prévision ;
8. **Choix du modèle final** en se basant sur les tests qui viennent d'être réalisés.

3. Plusieurs définitions du *Big Data* existent. Néanmoins, celle affirmant que l'on parle de données massives lorsque leur taille dépasse la capacité de stockage d'un seul ordinateur est communément admise.

En résumé

Le plan de ce mémoire est construit dans le but de suivre cette démarche :

- Le premier chapitre peut être apparenté à la première étape décrite précédemment : il permet de rappeler quelques informations relatives à la santé et à la *Data Science* ;
- La deuxième partie explicitera les étapes 2 à 5 ;
- Le troisième chapitre, quant à lui, posera les bases théoriques indispensables à la compréhension des étapes 6 à 8 ;
- Le quatrième chapitre illustrera de manière pratique la mise en place des étapes 6 à 8.

1.2.2 Un cadre législatif contraignant

L'explosion du volume de données a donc eu lieu en quelques dizaines d'années. La question de la protection de données s'est donc également rapidement posée.

La "Loi Informatique et Libertés"

La France a légiféré très tôt sur le sujet puisque la "Loi Informatique et Libertés", qui réglemente le traitement des données personnelles, date du 6 janvier 1978. Cette loi faisait suite à la vive réaction suscitée par la révélation d'un projet gouvernemental, nommé SAFARI [4], qui visait à identifier par un unique numéro chaque citoyen français et de regrouper, via ce numéro, tous les fichiers de l'administration française.
Pour la première fois, un texte proposait les définitions d'une "donnée à caractère personnel" et d'un "traitement de données à caractère personnel".
Pour information, c'est également par cette loi que fut créée la Commission nationale de l'informatique et des libertés (CNIL), autorité indépendante chargée de veiller à ce que l'informatique ne porte atteinte ni aux droits de l'Homme, ni à la vie privée, ni aux libertés individuelles ou publiques.
Cette loi a depuis subi plusieurs modifications, la dernière fois en 2017, afin de protéger de la manière la plus optimale possible la vie privée des individus.

Le RGPD

L'Union Européenne a également planché sur ce sujet puisqu'une directive, parue en 1995, devenait le premier texte de référence de niveau européen en matière de protection des données personnelles. Néanmoins, l'Union Européenne semble passer à la vitesse supérieure depuis peu. Le Règlement Général sur la Protection de Données (*RGPD*), publié le 4 mai 2016 au Journal Officiel de l'Union Européenne, sera applicable à partir du 25 mai 2018 et constituera le nouveau texte de référence européen en matière de protection des données à caractère personnel.
Ce règlement vise à harmoniser les règles des différents Etats membres afin de rendre plus efficiente la protection des données individuelles.
Dans ce but, le RGPD s'articule autour de six principes ou notions clés :

- Le consentement : l'obtention du consentement explicite de la personne concernée est indispensable à l'utilisation de ses données personnelles. En outre, le responsable du traitement des données doit être en mesure de prouver l'obtention de ce consentement ;
- La transparence : les données doivent être adéquates et pertinentes à ce qui est nécessaire au regard d'une finalité déterminée et légitime ;
- La protection à la conception des données ;
- La rectification et l'effacement : notion de "droit à l'oubli" ;
- Sous-traitance : obligation de garantir le respect des dispositions par l'ensemble des sous-traitants ;

4. pour Système Automatisé pour les Fichiers Administratifs et le Répertoire des Individus

— Accès et portabilité : Possibilité de récupérer les données communiquées et de les transmettre à une autre plateforme.

Les deux derniers principes mentionnées ci-dessus sont d'ailleurs des principes introduits pour la première fois par le RGPD. De plus, les organismes concernés devront nommer un *Data Protection Officer* ou délégué à la protection des données, chargé d'informer et de conseiller le responsable de traitement ou le sous-traitant, mais aussi les employés. Il doit être indépendant et être impliqué correctement et en temps utile dans toutes les problématiques liées à la protection des données à caractère personnel.

La NPA 5

Enfin, l'institut des Actuaires a également adopté une nouvelle norme de pratique actuarielle relative à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel nommée "NPA 5" le 16 novembre 2017.

En résumé

Le schéma ci-dessous permet d'avoir une vue d'ensemble des différentes évolutions législatives en matière de protection des données à caractère personnel :

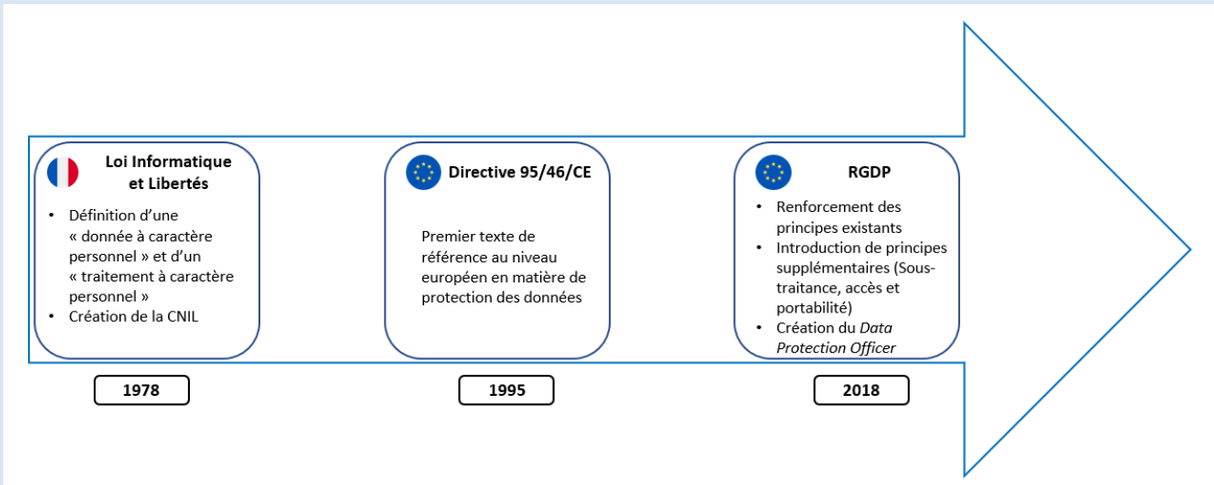


FIGURE 1.7 – Évolutions législatives en matière de protection des données à caractère personnel

Résumé de la démarche mise en place au cours du mémoire

Objectifs de l'étude

L'objectif principal de ce mémoire est de tenter de **mieux appréhender, à l'aide de modèles *Data Sciences*, la consommation en santé des salariés d'une entreprise et de leurs bénéficiaires**.

Axes d'étude

Dans l'encadré précédent, la notion de **consommation en santé** a été définie comme l'objectif principal de ce mémoire. De ce fait, cette consommation sera étudiée dans un premier temps afin de pouvoir analyser les risques du portefeuille. L'apport des variables externes sera notamment mis en évidence.
Afin d'identifier les caractéristiques des individus à risque, un **focus sera ensuite réalisé sur les gros consommants**. L'intérêt de cette deuxième étude sera d'exhiber des variables explicatives dans le comportement des gros consommants qui n'apparaissaient pas dans l'analyse de la consommation en santé.
Enfin, un **reporting santé** sera présenté, mêlant résultats des modèles *Data Science* et statistiques descriptives.

Modélisations envisagées

La notion de consommation en santé a été abordée sans plus précision. Mais comment celle-ci peut-être modélisée ?
Les techniques actuarielles usuelles de tarification santé reposent sur la méthode "Fréquence / Coût moyen". Ainsi, pour un acte médical donné, le montant de la prime pure est équivalent au produit de la fréquence par le remboursement complémentaire moyen associé. La fréquence correspond au nombre annuel d'actes d'un individu rapporté à son exposition dans l'année. Le coût moyen est estimé en ramenant la charge annuelle de consommation médicale au nombre d'actes de soins dans l'année.
Ce modèle est connu et est relativement simple d'utilisation. Sa principale limite provient de l'hypothèse d'indépendance entre les fréquences et les coûts moyens, il s'agit d'une hypothèse forte qui n'est pas toujours vérifiée en réalité.
Ici, tarifer un produit d'assurance n'est pas l'objectif premier : il s'agit de comprendre et d'expliquer la consommation en santé des adhérents. De ce fait, ce n'est pas le remboursement complémentaire moyen qui sera modélisé mais les frais réels engagés.
Par conséquent, **la consommation médicale correspondra au produit de la fréquence par les dépenses engagées moyennes**.
Enfin, **plutôt que de la calculer acte par acte, la consommation médicale sera ventilée en fonction des postes de soins qui seront définis ultérieurement**, dans le but d'avoir une vision macro.

Concernant les gros consommants, la variable étudiée sera une variable binaire construite de la manière suivante :

- 1 si les dépenses totales de l'individu sont supérieures au quantile à 80 % des dépenses du poste étudié ;
- 0 sinon.

Démarche globale

Le schéma ci-dessous permet de présenter de manière très succincte la démarche mise en place au cours de ce mémoire :

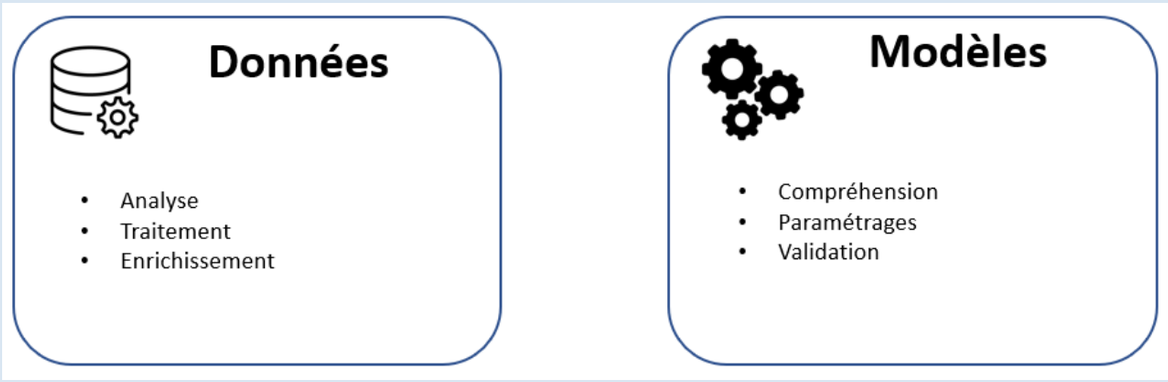


FIGURE 1.8 – Présentation de la démarche globale

Comme tout projet *Data Science*, deux phases principales peuvent être distinguées : l'une centrée sur les données et l'autre sur les modèles.
Ce schéma se veut comme une entrée en matière très rapide afin de cerner les deux enjeux principaux à la bonne réalisation de ce mémoire : avoir des données de qualité et utiliser de manière optimale l'ensemble des modèles testés.

2 Étude de la base de données

Dans tout projet de type *Data Science*, l'analyse et le traitement des données fait partie intégrante des travaux à réaliser. Cette étape n'est en aucun cas à négliger puisque des données non-adaptées à la problématique de l'étude ou de mauvaise qualité entraîneront de manière certaine des résultats peu probants. Cette partie vise donc à présenter l'ensemble des travaux réalisés en lien avec les données.

Le schéma suivant permet de récapituler les différentes étapes réalisées au cours de ce mémoire :

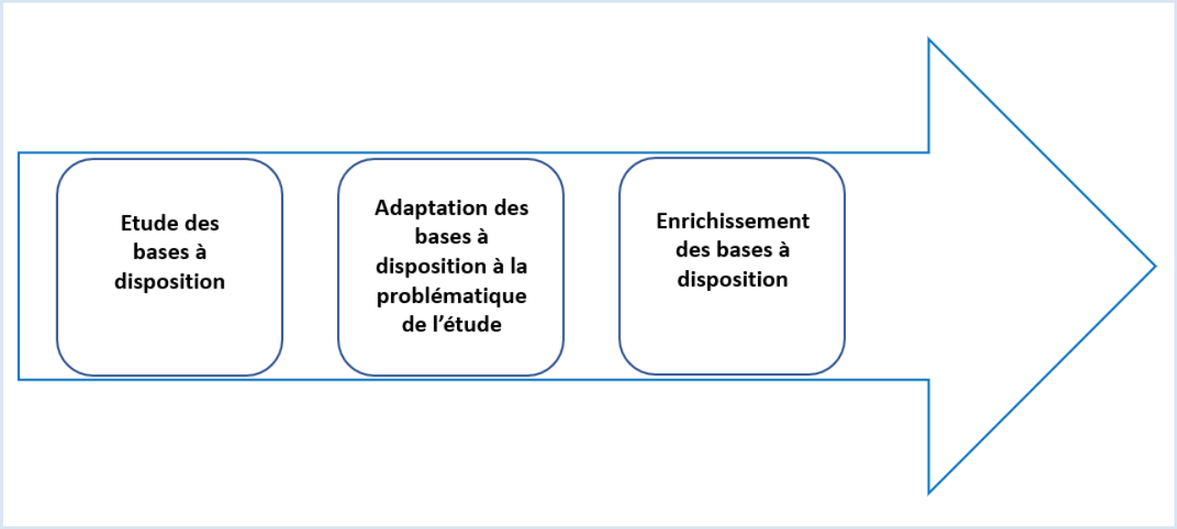


FIGURE 2.1 – Démarche de traitements des données

Tout au long de cette démarche, des contrôles de cohérence ont également été effectués. Les données récoltées étant globalement de bonne qualité, ces contrôles ne seront pas décrits outre mesure par la suite.

2.1 Etude des bases à disposition

Dans le cadre d'une étude santé, plusieurs données sont indispensables :

- Une **démographie détaillée de la population sous-risque**, notée base DEMO par la suite ;
- La **consommation médicale de la population sous-risque**, nommée base CONSO ;
- Les **taux de cotisation** ainsi que la **grille des garanties**.

2.1.1 Remarques préliminaires

Les données disponibles concernent un groupe possédant une dizaine de filiales et ayant une activité dans un secteur industriel.
Trois années d'étude, de 2013 à 2015, sont disponibles.

Il existe deux types de contrat en ce qui concerne les contrats santé, et plus généralement les contrats d'assurance :

- Le contrat individuel qui ne comprend que deux signataires : l'assureur et le souscripteur ;
- Le contrat collectif conclu entre l'assureur et une personne morale (en général une entreprise) au profit d'un groupe de personnes (en général les salariés de l'entreprise).

Concernant les contrats collectifs, plusieurs modes d'adhésion peuvent être rencontrés :

- A adhésion obligatoire : l'ensemble des salariés (ou une ou plusieurs catégories d'entre eux, définies de manière objective) y est obligatoirement affilié. Comme vu précédemment, les contrats collectifs obligatoires permettent, sous conditions, de bénéficier d'exonérations fiscales et sociales.
- A adhésion facultative : le contrat s'applique aux seuls salariés désirant bénéficier des garanties proposées. Ici, le risque d'antisélection [1] est élevé puisque la décision d'adhérer ou non est laissée aux salariés.
- Pour lutter contre ce phénomène, il existe des contrats à adhésion obligatoire et facultative (ou contrats à option) : ils permettent aux assurés de bénéficier de meilleures garanties sur certains postes contre le paiement d'une cotisation supplémentaire.

Le contrat étudié est un contrat collectif à adhésion obligatoire.

Les compositions de ces différentes bases vont maintenant être détaillées.

1. situation d'asymétrie d'informations où l'assuré possède plus d'informations sur son risque que l'assureur.

2.1.2 Les bases Démographie

Les bases DEMO disponibles concernent les exercices 2013 et 2014. Le nombre d'individus pour chacune de ces deux bases est présenté ci-dessous :

Exercice	Nombre d'individus
2013	6 087
2014	5 899

FIGURE 2.2 – Nombre de personnes par exercice au niveau des bases DEMO initiales

Ces deux bases possèdent la même structure et comptent toutes deux 62 variables.
Ces bases présentent les caractéristiques propres aux individus, comme par exemple :

- Identifiant ;
- Variables liées à leur lieu de résidence (ville, code postal, ...) ;
- Date de naissance ;
- Date de début du contrat et éventuelle date de résiliation ;
- Catégorie socioprofessionnelle (CSP) ...

Il est important de préciser que de nombreuses variables ne sont pas ou peu renseignées et ne sont donc pas exploitables en tant que telles. De plus, d'autres variables ne sont pas indispensables pour la suite de l'étude et ne seront donc pas utilisées.

2.1.3 Les bases Consommation

Les bases CONSO étudiées portent sur les exercices 2013, 2014 et 2015.
Chaque ligne représente un acte médical. De ce fait, le nombre d'actes qui ont été réalisés au cours d'une année peut être connu via ces bases. Dans le détail, cela donne :

Exercice	Nombre d'actes
2013	177 637
2014	148 256
2015	212 397

FIGURE 2.3 – Nombre de personnes par exercice au niveau des bases CONSO initiales

Une augmentation assez significative du nombre d'actes entre 2014 et 2015 peut dès lors être remarquée. Sans rentrer plus dans le détail et en ne tenant pour l'instant pas compte d'une éventuelle modification des effectifs, cette évolution peut être imputée à la mise en place d'honoraires de dispensation [2] chez les pharmaciens.

2. Depuis le 1er janvier 2015, les pharmaciens reçoivent de nouveaux honoraires destinés à valoriser leurs activités de conseil au moment de la dispensation des médicaments. Le dispositif prévoit deux catégories d'honoraires : un honoraire au conditionnement (à la boîte), et un honoraire d'ordonnance complexe pour les ordonnances comportant au moins 5 médicaments.

Elles présentent également une soixantaine de variables, un certain nombre d'entre elles étant redondantes vis-à-vis des bases DEMO.

Les éléments nouveaux présents dans ces bases concernent la caractérisation des actes médicaux. Voici une liste non-exhaustive des variables en lien avec l'acte médical :

- la date de l'acte ;
- la date de remboursement de l'acte ;
- la base de remboursement de la Sécurité Sociale ;
- les frais réels (ou dépenses engagées par l'assuré) ainsi que les remboursements de la Sécurité Sociale et de la Mutuelle associés ...

Enfin, trois variables (nommées par la suite FAMILLE, SOUS-FAMILLE et ACTE) permettent de classer le type d'acte réalisé, c'est-à-dire si telle ligne est reliée à une consultation chez un généraliste ou à une hospitalisation par exemple.
Chacune des trois variables classe de manière plus ou moins fine les actes selon un certain nombre de catégories. Le tableau suivant résume le nombre de modalités pour les trois variables citées précédemment :

Nom de la variable	Nombre de catégories
FAMILLE	8
SOUS-FAMILLE	24
ACTE	78

FIGURE 2.4 – Nombre de modalités des variables caractérisant le type d'acte

Pour se faire une idée générale de la classification des actes dans les bases initiales, voici l'intitulé des différentes catégories de la variable FAMILLE :

- Gros risque (actes liés au risque d'hospitalisation) ;
- Consultations Visites ;
- Soins courants (Auxiliaires médicaux, soins infirmiers, ... ;
- Pharmacie ;
- Dentaire ;
- Optique ;
- Cure thermale ;
- Divers.

2.1.4 Cotisations et garanties

Les cotisations ainsi que les garanties diffèrent plus ou moins selon la filiale à laquelle appartient l'individu.

Les cotisations sont toutes exprimées en pourcentage du Plafond Mensuel de la Sécurité Sociale (ou *PMSS*).
Certaines filiales prévoient de plus des structures de cotisation du type :

- Isolé/famille : la cotisation est différente pour les salariés souhaitant être assurés seuls et pour les salariés souhaitant assurer l'ensemble de leur famille ;
- Cotisant *AGIRC*/Non-cotisant *AGIRC*.

Les garanties proposées par la complémentaire interviennent après les remboursements de la Sécurité Sociale. Elles sont caractérisées sous plusieurs formes selon le type d'acte médical. Elles peuvent être exprimées en :

- Pourcentage des *FR* ;
- Pourcentage de la *BRSS* ;
- Pourcentage du *TM* ;
- Pourcentage du *PMSS* ...

2.2 Adaptation des bases à disposition à la problématique de l'étude

L'objectif de ce mémoire est de comprendre la consommation en santé de la population couverte. Pour cela, les bases de données initiales doivent être modifiées. En effet, **le format idéal et adapté à cette étude serait le suivant : chaque ligne correspondrait à un individu et présenterait ses caractéristiques (âge, sexe, ...) ainsi que sa consommation en santé sur les trois années d'étude**.

Le schéma suivant permet de récapituler les différentes étapes réalisées au cours de ce mémoire :

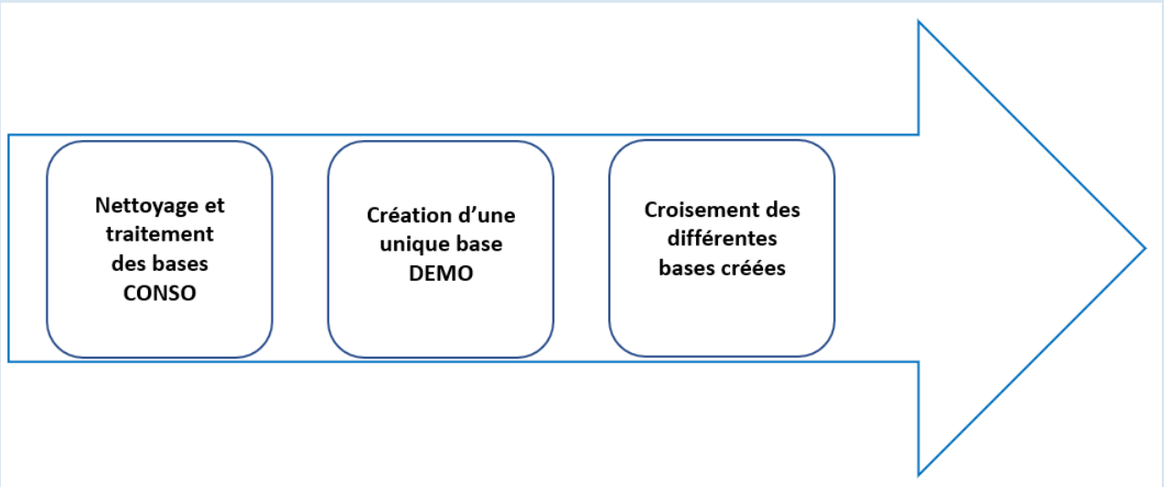


FIGURE 2.5 – Démarche de traitements des données

Tout au long de cette démarche, des contrôles de cohérence ont également été effectués. Les données récoltées étant globalement de bonne qualité, ces contrôles ne seront pas décrits outre mesure par la suite. Quelques anomalies peuvent néanmoins être citées comme par exemple le fait que certaines dates de souscription de bénéficiaires étaient antérieures à celles de l'assuré : la date de souscription de l'assuré a dans ce cas été prise en compte.

2.2.1 Nettoyage et traitement des bases CONSO

Le traitement des bases CONSO a été effectué en plusieurs étapes :

1. Découpage des bases CONSO ;
2. Modification de la répartition des actes ;
3. Attribution d'un identifiant caractérisant les individus ;
4. Agrégation des actes.

Découpage des bases CONSO

Les bases CONSO ont été traitées en séparant les actes selon ces modalités, afin d'obtenir des bases plus maniables et rassemblant des actes médicaux de même nature.
La variable FAMILLE [3], qui catégorise de la manière la plus grossière le type des actes, comporte huit modalités.

N.B. : La modalité "Cure thermale" ne présentant qu'un nombre infime d'actes, celle-ci sera supprimée dans la suite de l'étude.

Ayant à disposition trois années d'historique, 21 bases sont dorénavant à disposition selon le découpage suivant :

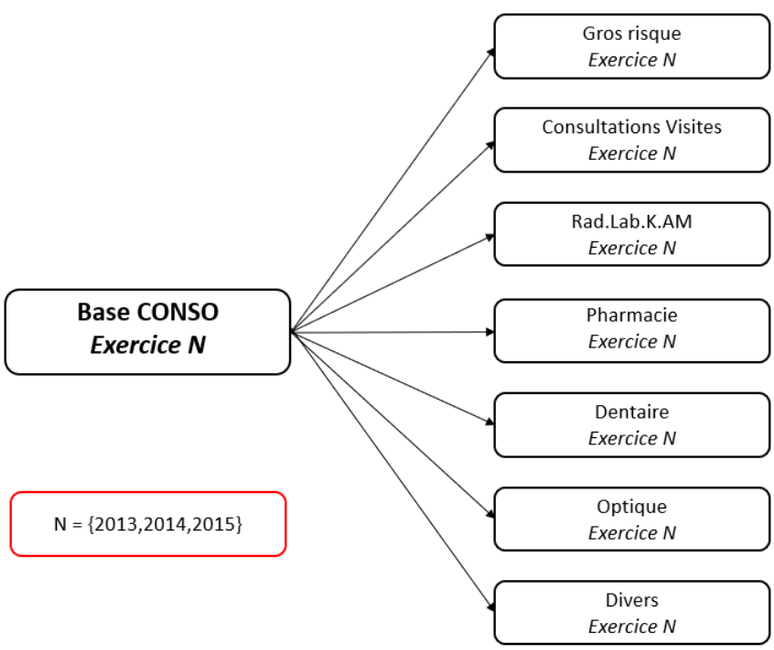


FIGURE 2.6 – Découpage des bases CONSO

Modification de la répartition des actes

Avant toute chose, il convient de mentionner que la répartition des actes a été modifiée au cours des différents traitements effectués sur les bases CONSO.
La segmentation de la variable ACTE étant trop fine [4], la répartition de la variable SOUS-FAMILLE (qui présente 24 modalités) a été retenue. Néanmoins, après l'étude des BRSS et des différents frais engagés, certaines catégories ont été renommées et quelques changements ont été effectués dans la répartition des actes.

3. c.f. Figure 2.4
4. De nombreuses catégories ne présentaient que quelques actes.

Le tableau suivant récapitule la nouvelle catégorisation des actes médicaux :

Famille	Sous-famille
Hospitalisation	Frais de séjour
	Transport ambulance
	Honoraires
	Ticket modérateur
Consultations Visites	Généralistes
	Spécialistes
Soins courants	Soins infirmiers
	Kinésithérapie
	Radiologie
	Laboratoire
	Acte en k
	Ortophonie
	Orthoptie
	Osthéopatie
	Podologie
Pharmacie	Pharmacie
Dentaire	Orthodontie
	Prothèses dentaires
	Soins dentaires
	Implants dentaires
Optique	Lunettes
	Lentilles
Divers	Appareillages

FIGURE 2.7 – Répartition des actes médicaux après traitements

N.B. : la sous-famille "Acte en k" comprend les actes de spécialités non-cités dans les autres sous-familles.

Attribution d'un identifiant

La première étape, qui concerne l'ensemble de ces sept familles d'actes, consiste à attribuer à chaque acte (ou ligne) une variable permettant de relier cet acte à un individu. En somme, cette variable jouera le rôle d'identifiant ou de clé lorsqu'il faudra lier les différentes bases CONSO et DEMO.
Au vu des variables à disposition au sein des bases CONSO, la clé permettant d'identifier les individus, nommée désormais *Clé individu*, est le résultat de la concaténation des trois variables suivantes :

- Numéro de Sécurité Sociale de l'adhérent ;
- Code Bénéficiaire : cette variable permet d'identifier si la personne consommante est un assuré (l'équivalent de l'adhérent), son conjoint ou un de ses enfants. Elle présente donc trois modalités :
  - Assuré ;
  - Conjoint ;
  - Enfant.
- Rang Bénéficiaire : cette variable vaut respectivement :
  - 0 si la variable Code Bénéficiaire prend la modalité "Assuré" ;
  - 1 si la variable Code Bénéficiaire prend la modalité "Conjoint" ;
  - > 1 si la variable Code Bénéficiaire prend la modalité "Enfant".

  Elle permet donc de distinguer les enfants. En effet, si un assuré a plusieurs enfants, la concaténation des deux premières variables ne permettrait pas de les individualiser.

Autrement dit, voici comment obtenir la *Clé individu* :

Clé individu = CONCATENER(
.....Numéro sécurité sociale ;
.....Code Bénéficiaire ;
.....Si Code Bénéficiaire = Enfant
.....Faire Rang Bénéficiaire
.....Sinon Ne rien faire)

Agrégation des actes

Une étude précise de chacune de ces bases a permis de mettre en lumière le fait que de nombreuses lignes faisaient référence au même acte.
Citons quelques exemples pour mieux illustrer notre propos :

— Concernant la famille "Consultations Visites", la majorité des actes liés à des consultations chez des spécialistes (cardiologue, neuropsychiatre, ...) étaient décomposés en deux voire trois lignes :

1. la première ligne correspondait au tarif de responsabilité ;
2. la deuxième ligne reflétait les différents dépassements d'honoraires.

— Au niveau de la famille "Pharmacie", plusieurs lignes pouvaient correspondre à un même déplacement chez le pharmacien car chaque ligne était reliée à un seul médicament ...

Dans une mission de tarification classique, les différents actes ne sont pas regroupés puisqu'il faut ensuite étudier le coût moyen par acte. Ici, la démarche n'est pas la même puisque c'est la consommation médicale de manière plus globale qui est au centre de l'étude. Par conséquent, de nombreuses lignes ont dû être agrégées afin de mieux représenter le nombre d'actes de chaque individu.

Pour ce faire, une deuxième clé, nommée *Clé acte* a été créée afin d'identifier chaque acte. De la même manière que pour la *Clé individu*, cet identifiant a été créé à partir des variables présentes dans les bases CONSO :

Clé acte = CONCATENER(
.....Numéro d'enregistrement ;
.....Code Bénéficiaire ;
.....date acte) ;
.....Rang Bénéficiaire)

La variable "Numéro d'enregistrement" correspond au numéro de dossier lorsque l'acte médical est enregistré au sein de l'organisme complémentaire. Cette variable est directement liée à la date de remboursement de l'acte.
La *Clé acte* est utilisée au sein des 21 bases CONSO et sert à comptabiliser les actes réalisés par chaque individu identifié via la *Clé individu*.

Création des nouvelles bases CONSO

La démarche permettant de traiter les différentes bases CONSO va dorénavant être explicitée. Le schéma suivant illustre la logique globale de cette démarche.
N.B. : chacune des sept familles d'actes présente des spécificités de traitement qui ne seront pas abordées ici.

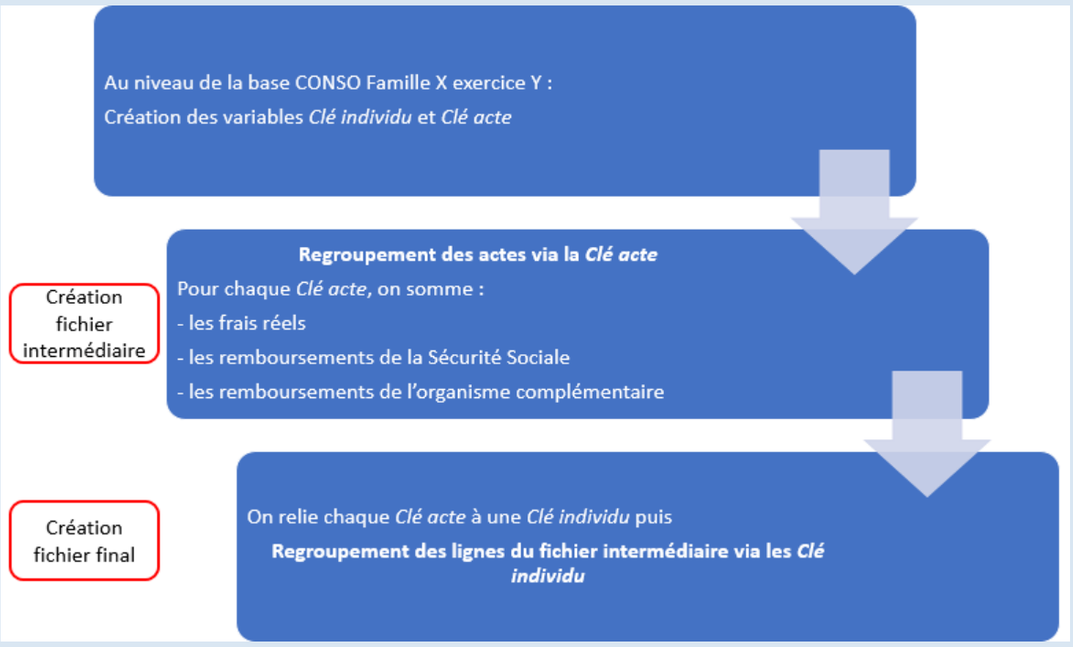


FIGURE 2.8 – Logique de traitement des bases CONSO

La dernière étape du schéma précédent permet d'introduire une variable "Nombre d'actes" qui prend en compte l'agrégation souhaitée de certaines lignes. **C'est cette variable qui correspondra à la variable à expliquer lors de la modélisation de la fréquence.**
A la fin de ce processus, nous obtenons donc un fichier final pour chacune des 21 bases CONSO présentant la forme suivante :

Clé individu	Nombres d'actes	Frais réels	Remb Sécurité Sociale	Rem Mutuelle
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

FIGURE 2.9 – Format des bases CONSO traitées

Les colonnes deux à cinq sont répétées autant de fois qu'il y a de sous-catégories dans la famille. Quatre colonnes, correspondant aux totaux des différentes sous-catégories de la famille d'actes, sont rajoutées à la fin de ces fichiers. Par exemple, le fichier final de la famille "Hospitalisation" comporte 21 colonnes :

- La première colonne correspond à la *Clé individu* ;
- Cette famille est composée de 4 sous-familles : il y a donc 16 colonnes supplémentaires ;
- Les 4 dernières colonnes représentent les différents totaux pour cette famille.

La structure de ce fichier permet donc d'avoir le détail des consommations de chaque individu pour chaque sous-famille d'acte. Les colonnes "Total" permettent d'avoir l'information agrégée au niveau des familles d'acte.

Les différentes bases CONSO sont maintenant utilisables dans le cadre de l'étude. Il s'agit dorénavant de traiter les bases DEMO.

2.2.2 Création d'une unique base démographie

Le but de cette partie est de construire une base DEMO qui regroupe l'ensemble des individus présents dans les différentes bases initiales. Avant toute chose, il est essentiel de rappeler que les bases DEMO à disposition concernent uniquement les exercices 2013 et 2014 tandis que trois bases CONSO, couvrant les exercices 2013 à 2015 sont disponibles.
L'information importante est l'absence de la base DEMO 2015, il faut donc la reconstruire.

Cette reconstruction de la base DEMO 2015 implique l'utilisation de la base CONSO 2015. Comme de nombreuses variables sont communes aux bases DEMO et CONSO, les bases DEMO 2013 et 2014 ne serviront que très partiellement dans la construction de cette nouvelle base.

Le schéma suivant permet d'illustrer la logique générale de création de la base DEMO :

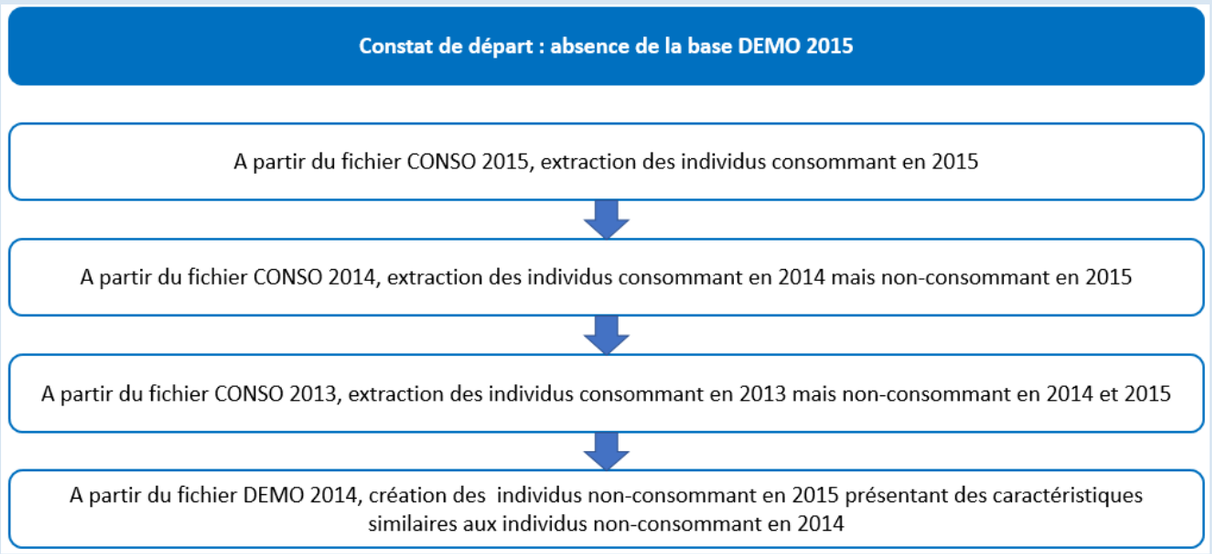


FIGURE 2.10 – Logique de construction de la base DEMO

L'extraction des différents individus s'effectue selon la *Clé individu* définie dans la partie précédente.

2.2.3 Croisements des différentes bases créées

A ce stade de l'étude, les bases à disposition sont les suivantes :

— 21 bases CONSO, chacune représentant la consommation d'une famille d'acte durant un exercice ;

— une base DEMO regroupant l'ensemble des individus présents dans les différentes bases initiales.

Il convient dorénavant de croiser ces différentes bases. Comme il l'a été déjà mentionné, la variable *Clé individu* va jouer le rôle de jointure afin de lier proprement les (nombreux) fichiers.

La base agrégée possède donc la structure suivante :

Base DEMO	Base CONSO famille x exercice 2013	Base CONSO famille x exercice 2014	Base CONSO famille x exercice 2015	Base CONSO famille y exercice 2013	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

FIGURE 2.11 – Composition de la base agrégée

La base finale comporte donc :

Nombre de lignes ou d'individus	7 771
Nombre de colonnes ou de variables	396

FIGURE 2.12 – Dimension de la base agrégée

2.3 Travail sur les variables

. Seules les variables spécifiques aux actes médicaux ont été présentées [5] précédemment. Cette section se concentre sur les variables spécifiques aux individus.
Étant au centre de la problématique de ce mémoire, celles-ci vont faire l'objet d'une description précise au sein de la partie suivante.
Plusieurs phases d'ajout de variables ont été mises en place au cours de ce mémoire. Le schéma suivant permet de résumer ces différentes étapes essentielles à la construction de la base finale :

Plusieurs phases d'ajout de variables ont été mises en place au cours de ce mémoire. Le schéma suivant permet de résumer ces différentes étapes essentielles à la construction de la base finale :

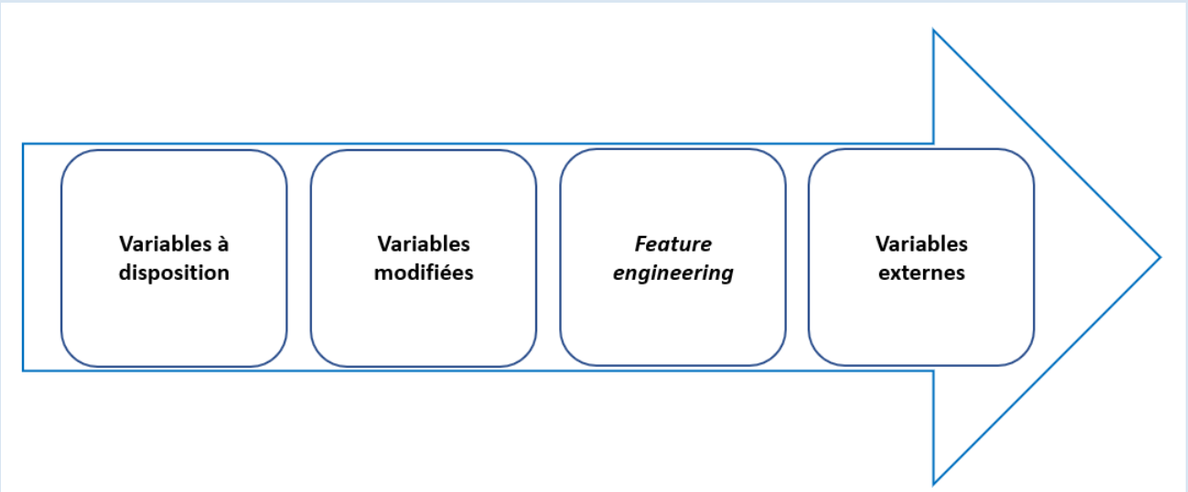


FIGURE 2.13 – Les différentes étapes d'ajout de variables

N.B. : Le Feature Engineering correspond à la création de nouvelles variables à partir de variables existantes.

2.3.1 Variables initiales

Les variables conservées à la suite de la phase de construction de la base de travail sont les suivantes :

- Le code bénéficiaire [6] ;
- Le nom des différentes filiales de la société ;
- La catégorie socio-professionnelle ;
- L'état de l'adhérent [7] ;
- Le numéro du contrat souscrit par la société à l'organisme complémentaire ;
- Les dates de souscription et de fin du contrat ;
- La date de naissance ;
- Le code postal ;
- L'ensemble des variables présentes dans les bases CONSO [8].

Il est à noter que l'ensemble de ces variables ne seront pas utilisées pour la modélisation de la consommation en santé des adhérents. Néanmoins, certaines de ces variables seront utiles à la création de nouvelles variables à la réalisation de statistiques qui permettront de mieux visualiser la base de données.

5. c.f. Figure 2.6
6. les modalités de cette variable sont décrites dans la section 2.2.1 lors de la création de l'identifiant caractérisant les individus
7. cette variable présente quatre modalités : Actif, Retraité, Congé maternité et licenciement
8. c.f. Figure 2.9

2.3.2 Traitements des variables initiales

Le tableau suivant permet de présenter les variables traitées au cours de l'étude ainsi que les variables créées à la suite de ces traitements :

Nom de la variable traitée	Nombre de la variable créée	Précisions
Date de souscription du contrat	Ancienneté	
Date de naissance	Age	
Etat de l'adhérent	Congé Maternité	Variable binaire prenant la valeur 1 si l'état de l'individu présente la modalité « Congé Maternité »
	Licenciement	Variable binaire prenant la valeur 1 si l'état de l'individu présente la modalité « Licenciement »
	Il est également à noter que les individus présentant la modalité « Retraité » ont été supprimés de cette étude. Ceux-ci représentent moins de 2 % des adhérents et présentent des options de garanties non accessibles.	

FIGURE 2.14 – Liste des variables modifiées

2.3.3 *Feature Engineering*

Quelques variables ont également été créées à l'aide des variables à disposition.

Le sexe

Cette dernière a été créée à l'aide des variables Numéro sécurité sociale et Code Bénéficiaire. Rappelons tout d'abord la composition d'un numéro de Sécurité Sociale :

Position	Signification
1	1 pour les hommes et 2 pour les femmes
2 et 3	Deux derniers chiffres de l'année de naissance
4 et 5	Mois de naissance
6 et 7	Département de naissance
8,9 et 10	Code officiel de la commune
11, 12 et 13	Numéro d'ordre de la naissance dans le mois et la commune
14 et 15	Clé de contrôle

FIGURE 2.15 – Décomposition d'un numéro de sécurité sociale

Le premier chiffre du numéro de Sécurité Sociale permet donc de connaître le sexe d'un individu. Dans le cas présent, ayant à disposition uniquement le numéro de sécurité sociale de l'assuré, un traitement en fonction de la variable Code Bénéficiaire a été effectué. En effet :

— Si Code Bénéficiaire = "Assuré", on prend le premier chiffre du numéro de sécurité sociale ;

— Si Code Bénéficiaire = "Conjoint", on regarde le premier chiffre du numéro de sécurité sociale et on affecte le sexe opposé ;

— Si Code Bénéficiaire = "Enfant", on garde cette modalité.

La variable Sexe présente donc trois modalités :

— Homme adulte ;

— Femme adulte ;

— Enfant.

Il peut néanmoins être intéressant de conserver ces trois modalités puisque les enfants n'ont pas forcément la même consommation médicale que les adultes (en orthodontie par exemple).

La composition du foyer de l'adhérent

Une variable "Situation familiale" était disponible au sein des bases DEMO initiales mais n'était malheureusement pas renseignée. De ce fait, les variables "Conjoint" et "Nombre d'enfants" ont été créées afin de renseigner la composition du foyer familial de celui-ci. Ces deux variables ne concernent que les individus dont la modalité de la variable "Code bénéficiaire" est "Assuré".
Ces deux variables ont été calculées à partir des variables Numéro sécurité sociale et Code Bénéficiaire. Plus précisément, voici une présentation sous forme d'algorithme de ces variables :

— Conjoint :
Si Code Bénéficiaire = "Assuré" ;

1. Recherche du numéro de sécurité sociale correspondant dans la base DEMO créée ;
2. Si la recherche aboutit et que Code Bénéficiaire = "Conjoint", alors la variable conjoint vaut 1 ;

— La construction de la variable "Nombre d'enfants" suit la même logique : la variable vaut simplement le nombre de fois où Code Bénéficiaire = "Enfant" lorsque l'on effectue la recherche du numéro de sécurité sociale.

Les variables présence

Via les dates de début et de résiliation de contrat retraitées, six variables représentant les présences pondérées et non-pondérées de l'individu selon les années 2013, 2014 et 2015. Ces variables seront importantes par la suite, que ce soit pour effectuer des statistiques descriptives au niveau de la base finale ou pour l'implémentation des modèles.

Les variables Consomme / Ne consomme pas

Dans le but d'étudier les évolutions des consommations de la population couverte sur les trois exercices, des variables "Consomme / Consomme pas" binaires ont été créées pour les sept familles d'acte et les trois exercices étudiés.
Concrètement et en prenant l'exemple de la famille "Gros risque", si un individu présente des dépenses en 2013, alors la valeur 1 sera affectée à la variable "Consomme / Consomme pas 2013" liée à la famille d'acte "Gros risque".

2.3.4 Variables externes

Pour enrichir la base de données, des données externes ont été récupérées sur internet. Elles proviennent principalement des données statistiques de l'INSEE relatives aux communes de France. Au total, **22 variables** ont été rattachées à la base de donnée. Elles concernent :

- La structure de la population de chaque commune (nombre d'habitants, proportion de mineurs, proportion de seniors, ...) ;
- Les équipements dans la commune (nombre d'équipements de santé, nombre d'équipements sportifs) ;
- L'emploi (proportion d'actifs, de chômeurs, de cadres, ...) ;
- Le loyer moyen de la commune ;
- La présence de professionnels de la santé dans la commune (médecins, dentistes, ophtalmologues, ...).

Ces données externes ont été traitées de la manière suivante :

1. La jointure des bases a été réalisée à l'aide d'une table de correspondance entre les codes communes INSEE et les codes postaux de la base de données.
2. Ces variables de type quantitatif sont transformées en variables de type qualitatif à l'aide d'étude statistiques (étude des quartiles, de la médiane, ...)

Par exemple, le nombre d'habitants de chaque commune a été rattaché à la base de données de l'étude grâce aux bases issues des différents recensements disponibles sur le site internet de l'INSEE. Cette variable, nommée "Zone" par la suite" est de type factorielle et est discrétisée de la manière suivante :

- Si le nombre d'habitants de la commune est inférieur à 2 000, alors la variable prendra la modalité "Rurale" ;
- Si le nombre d'habitants de la commune est compris entre 2 000 et 10 000 habitants, alors la variable prendra la modalité "Semi urbaine" ;
- Si le nombre d'habitants de la commune est compris entre 10 000 et 50 000 habitants, alors la variable prendra la modalité "Urbaine" ;
- Si le nombre d'habitants de la commune est supérieur à 50 000, alors la variable prendra la modalité "Très urbaine".

La liste exhaustive des variables externes est disponible en annexe. Comme il est indiqué en annexe, l'ensemble des variables externes sont liées à la commune de résidence de l'individu. Par la suite, une variable intitulée "Proportion seniors" est à interpréter dans le sens "Proportion de seniors dans la commune".

2.4 Statistiques descriptives

La base de données étant maintenant construite, quelques statistiques peuvent être introduites afin mieux se représenter les caractéristiques de la population sous-risque ainsi que ses modes de consommation médicale.

2.4.1 Vision globale du portefeuille

Cette première partie va permettre de mieux cerner les caractéristiques de la population sous-risque.

Statistiques démographiques

Il peut être intéressant de tout d'abord se focaliser sur le nombre de personnes présentes au sein du portefeuille étudié. L'analyse de la population est ventilée par catégorie de bénéficiaire et par exercice :

Exercice	Assurés	Conjoints	Enfants	Total Bénéficiaires	Coefficient familial
2013	2 611	1 785	2 465	6 861	2,63
2014	2 466	1 682	2 391	6 539	2,65
2015	2 378	1 654	2 478	6 510	2,74
Variation 2015/2014	-3,6%	-1,7%	3,6%	-0,4%	

FIGURE 2.16 – Démographie générale de la population sous-risque

N.B. :

— *le coefficient familial se calcule comme suit :*

$$coefficient\ familial = \frac{Nombre\ total\ bénéficiaires}{Nombre\ adhérents}$$

— *le terme de bénéficiaire désigne l'adhérent et ses ayant-droits.*

Il peut être également utile de présenter une démographie pondérée. Cette pondération est effectuée via les variables "Présence" créées au cours des traitements réalisés sur les bases de données initiales. Par exemple, si un individu est présent au 01/01/N mais résilie au 01/04/N, il ne comptera que pour 0.25 dans le tableau suivant :

Exercice	Assurés	Conjoints	Enfants	Total Bénéficiaires
2013	2 454	1 695	2 350	6 498
2014	2 284	1 571	2 263	6 118
2015	2 239	1 550	2 323	6 111
Variation 2015/2014	-2,0%	-1,4%	2,6%	-0,1%

FIGURE 2.17 – Démographie pondérée de la population sous-risque

La comparaison des démographies pondérée et non pondérée permettent de mettre en évidence les mouvements de population. Dans le cas présent, la population reste plutôt stable sur les trois exercices à disposition.
Dans la suite de cette partie, les statistiques présentées ne concernent donc que l'exercice 2015.

Présentation des caractéristiques du portefeuille

Dans ce paragraphe, plusieurs points d'attention sont réalisés sur les principales caractéristiques de la population sous-risque. Ceci permet de visualiser un premier "profil-type" des adhérents.

L'âge

Les consommations médicales variant fortement avec l'âge, il semble indispensable de présenter quelques statistiques sur l'âge des adhérents :

Age moyen	Assuré	Conjoint	Enfant	Total Bénéficiaires
2015	45,6	45,7	13,1	33,2

FIGURE 2.18 – Age moyen de la population sous-risque

Cependant, la présentation de seulement l'âge moyen par type de bénéficiaire ne peut être suffisante. Voici la pyramide des âges de la population composant le portefeuille :

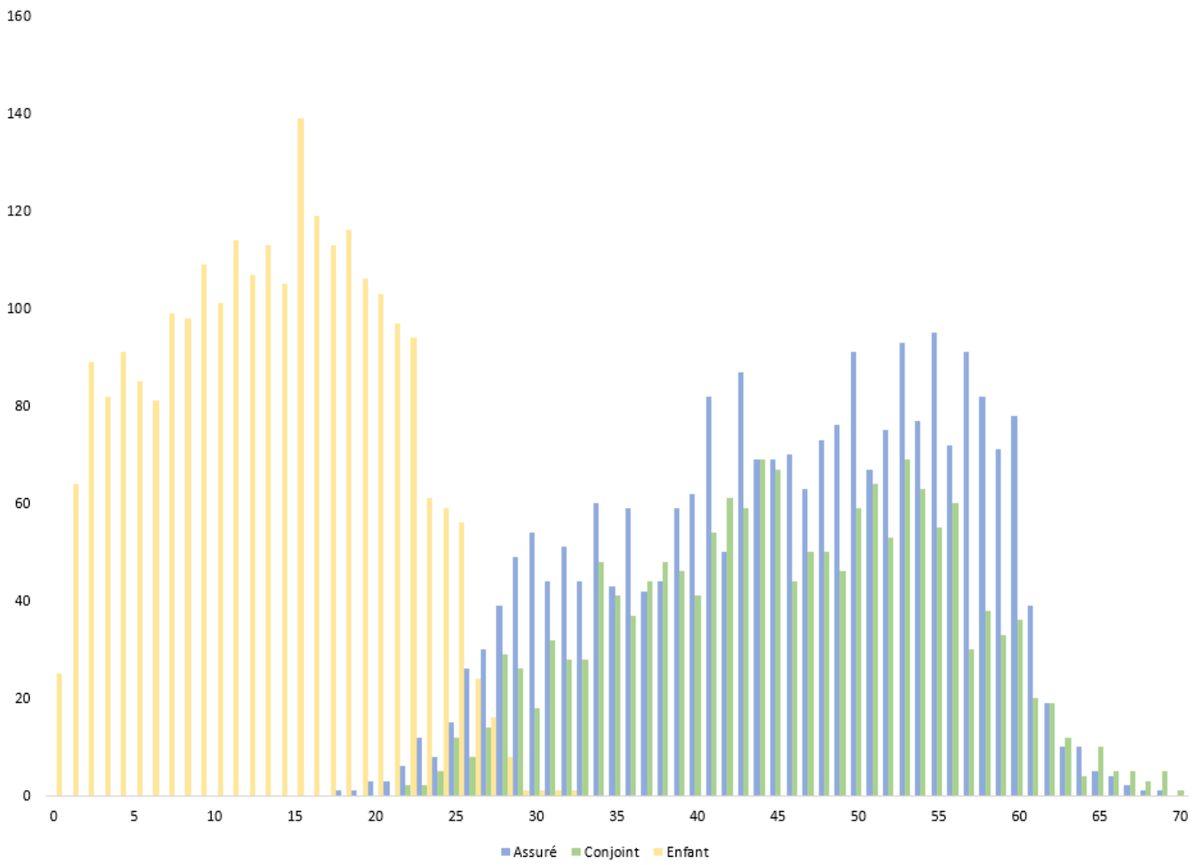


FIGURE 2.19 – Pyramide des âges sur l'exercice 2015

La majorité des assurés et de leur conjoint a donc un âge compris entre 40 et 60 ans. Cela semble logique puisque les données analysées concernent la complémentaire santé d'une entreprise.

Le sexe

Le sexe joue également un grand rôle dans l'étude des consommations médicale (on pense de manière logique au poste maternité en premier lieu). Voici donc la répartition des assurés (et non de l'ensemble du portefeuille) selon le sexe de la personne :

Assurés	Homme	Femme
2015	80%	20%

FIGURE 2.20 – Répartition des assurés selon le sexe

L'entreprise couverte par la mutuelle ayant une activité dans le secteur industriel, il n'est pas étonnant de constater une plus grande proportion d'hommes au sein des assurés.

La catégorie socio-professionnelle (CSP)

Il peut être logique de penser que la rémunération joue un rôle dans la consommation médicale. N'ayant pas à disposition cette information, la CSP peut être interprétée comme un indicateur de la rémunération de l'individu :

CSP	Cadre	Employé	Ouvrier	Technicien
2015	29%	31%	33%	7%

FIGURE 2.21 – Répartition des assurés selon leur catégorie socio-professionnelle

La population étudiée est donc composée à près de :

- 30 % de cadres ;
- 70 % de non-cadres.

2.4.2 Étude de la consommation

Les caractéristiques des individus composant le portefeuille étudié sont maintenant connues. Il est alors temps de s'intéresser aux consommations médicales.

Proportions de consommants

Le tableau suivant résume la part des individus ayant eu recours à au moins un acte médical durant l'exercice 2015 :

Exercice	Assurés	Conjoints	Enfants	Total Bénéficiaires
2015	91%	85%	80%	85%

FIGURE 2.22 – Proportion de consommants dans le portefeuille

La proportion des consommants est logiquement très élevée, la probabilité de réaliser au moins un acte médical au cours d'une année étant très forte. Au contraire d'une étude de résiliation par exemple où le signal (en l'occurrence la résiliation) est faible, l'étude d'un portefeuille santé permet donc d'avoir de nombreux cas positifs (i.e. des individus qui consomment).

Quelques statistiques sur les frais réels et les remboursements complémentaires

Les deux camemberts suivants permettent de visualiser la part des différents postes de santé dans les frais réels et les remboursements complémentaires :

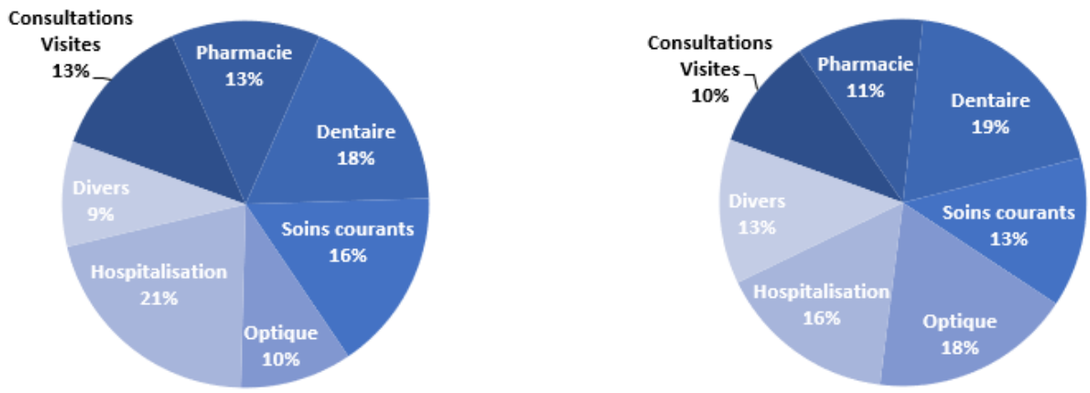


FIGURE 2.23 – Part des différents postes dans les frais réels (à gauche) et les remboursements complémentaires (à droite)

Une première remarque serait de dire que la répartition des actes reste globalement similaire lorsque l'on s'intéresse aux frais réels ou aux remboursements de la mutuelle. Cependant, il serait trop réducteur de s'arrêter à cette vision générale. Avec une étude plus approfondie, les remarques suivantes peuvent dès lors être effectuées :

- Le poids du poste "Hospitalisation" dans les frais réels est plus élevé de 5 points qu'au niveau des remboursements de la mutuelle ;
- Au contraire, si l'on s'intéresse au poste "Optique", la proportion de cette catégorie est prépondérante dans les remboursements de l'organisme complémentaire.

L'analyse des deux précédents graphiques a mis en évidence que la prise en charge des frais réels varie de manière significative selon les différents postes. Le graphique suivant vient appuyer ces propos :

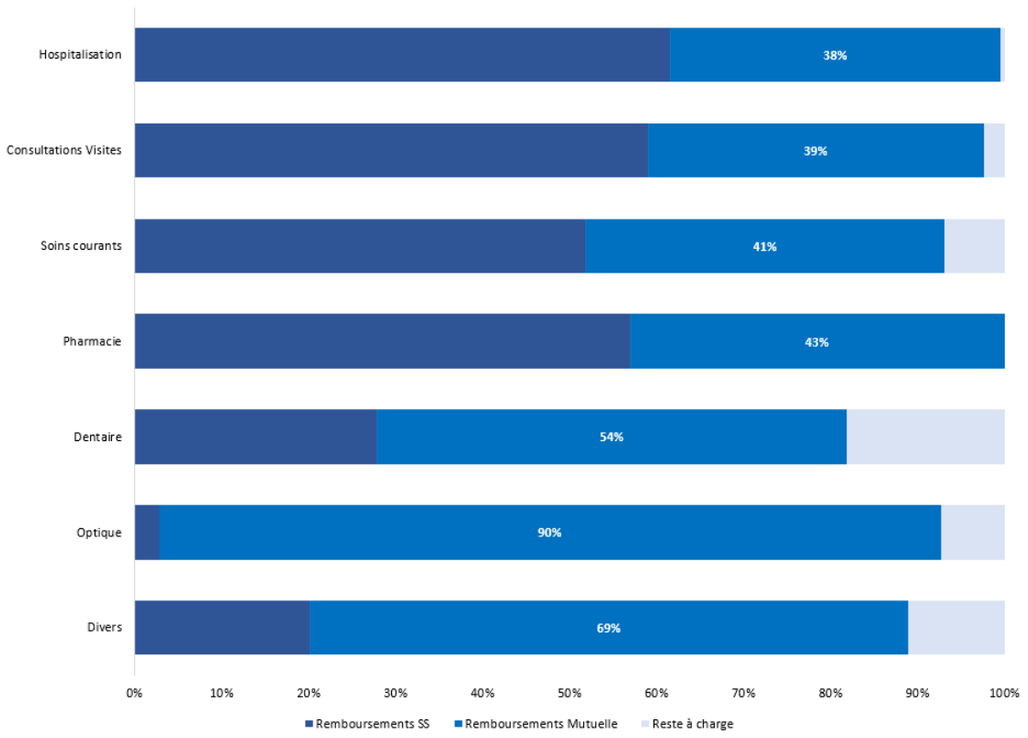


FIGURE 2.24 – Prise en charge des dépenses selon les postes en santé

Le graphique précédent montre que la prise en charge de l'organisme complémentaire diffère fortement selon les actes. Elle prend par exemple la quasi-totalité des sommes engagées en "Optique". On remarque également que les remboursements de la Sécurité Sociale représentent environ 60 % des dépenses engagées en "Hospitalisation". Or, le taux de remboursement classique de la Sécurité Sociale est de 80 %. Cette différence est principalement due aux Affectations de Longue Durée ou *ALD* qui sont intégralement prises en charge par la Sécurité Sociale. De ce fait, ces actes n'apparaissent pas dans les bases de données de l'organisme complémentaire. Cette absence d'informations peut évidemment être préjudiciable pour l'organisme complémentaire : celui-ci ne connaît pas entièrement les consommations de sa population sous-risque.

Évolution de la consommation médicale

Cette partie ne s'intéresse qu'à la consommation médicale totale et non poste par poste, afin d'avoir une vision d'ensemble des effets potentiels des variables à disposition.

L'âge

L'âge des bénéficiaires en santé est une information déterminante pour connaître le risque auquel l'organisme complémentaire est confronté. En effet, chaque âge correspond à des besoins particuliers en santé :

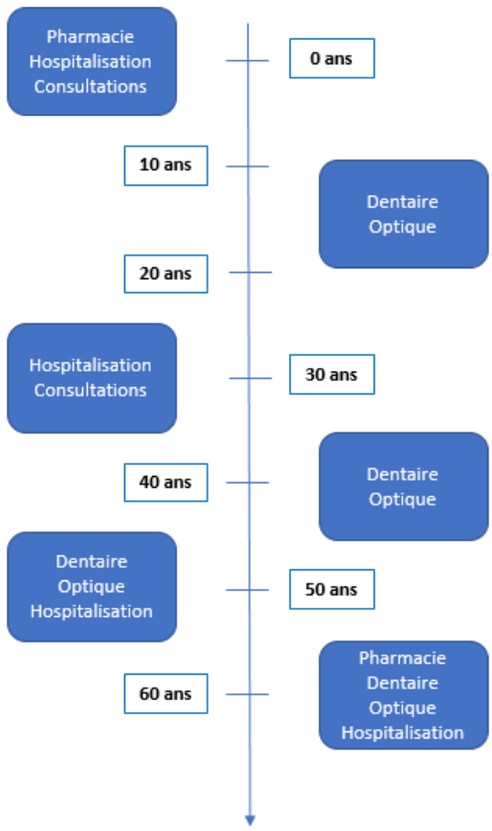


FIGURE 2.25 – Les âges clés de la santé

Les besoins en santé diffèrent donc en fonction de l'âge.
Jusqu'à environ 40 ans, les prestations en santé sont assez faibles mis à part pour la maternité chez les femmes autour de 30 ans. De ce fait, les assurés ayant entre 30 et 40 ans ont besoin de garanties en hospitalisation et en consultations de spécialistes (pour les nourrissons).
A partir de 40 ans, le risque santé devient plus important : les assurés commencent à avoir recours aux prothèses dentaires, ainsi qu'à des verres ayant des corrections élevées. Les enfants de ces salariés ont également des besoins en dentaire et en optique, avec l'orthodontie et la myopie des adolescents.

A partir de 50 ans, les risques déjà mentionnés dans le paragraphe précédent augmentent. Le risque Hospitalisation fait également son apparition.
Pour les âges plus élevés, les risques Dentaires et Optique ont tendance à diminuer mais sont remplacés par le risque Hospitalisation.

Une représentation de la consommation moyenne par âge peut donc être pertinente.
La consommation moyenne est calculée en rapportant, pour chaque âge, les frais réels à l'effectif total :

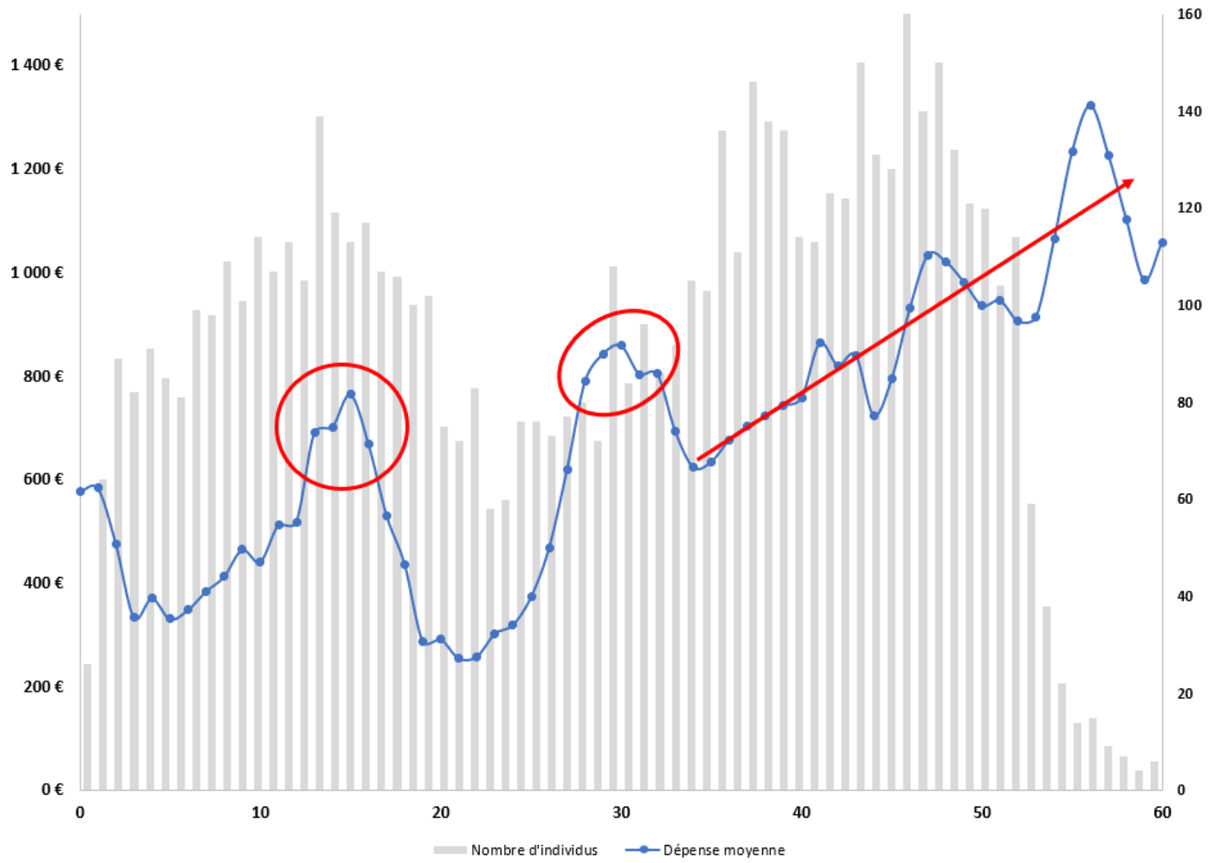


FIGURE 2.26 – Consommation moyenne par âge

Les pics mentionnés plus haut sont bien présents :

- La consommation à l'adolescence correspond aux dépenses engagées en orthodontie et en optique ;
- Le consommation autour des 30 ans correspond à la maternité ;
- On remarque une tendance à la hausse pour les individus âgés de plus de 40 ans.

La catégorie socio-professionnelle et l'ancienneté

Les graphiques suivants permettent d'avoir une première idée des effets de certaines variables sur la consommation médicale totale :

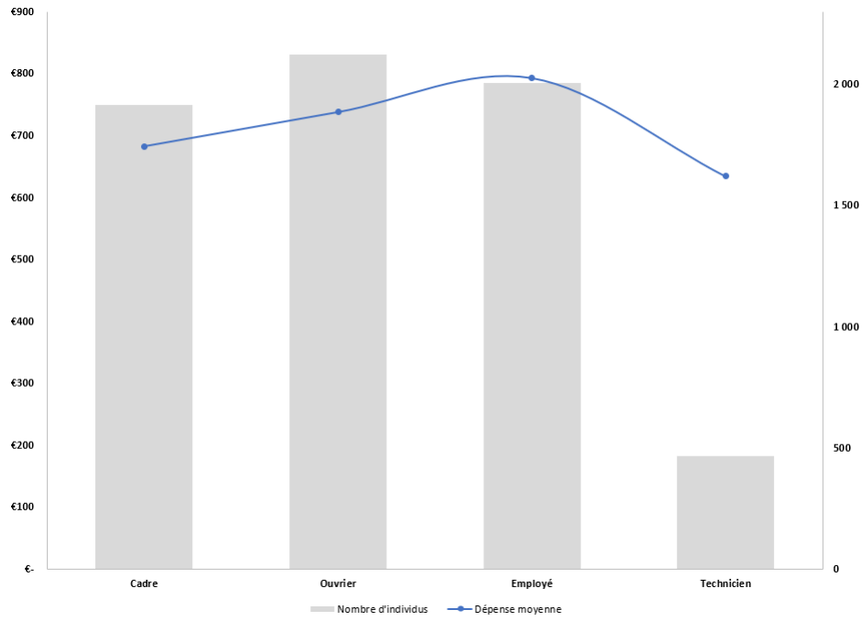


FIGURE 2.27 – Consommation moyenne selon la CSP

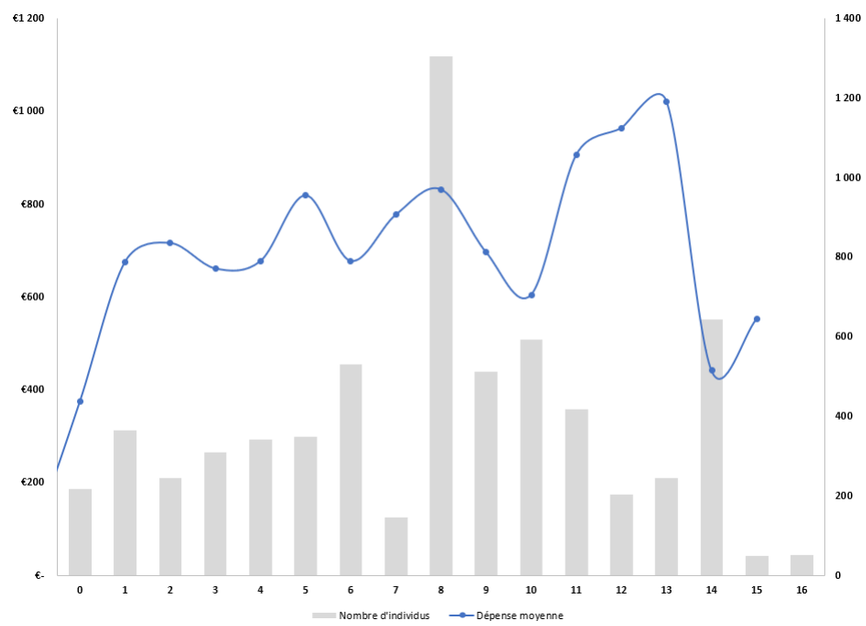


FIGURE 2.28 – Consommation moyenne selon l'ancienneté

La CSP ne semble donc pas jouer un rôle déterminant dans la consommation médicale. Au contraire, la consommation médicale a tendance à augmenter avec l'ancienneté du contrat.

2.4.3 Focus sur les gros consommants

Une idée couramment répandue en assurance santé est que 20 % d'un portefeuille santé représentent près de 80 % des remboursements complémentaires.
Le graphique suivant représente la fonction de répartition des frais réels en fonction de la proportion du portefeuille :

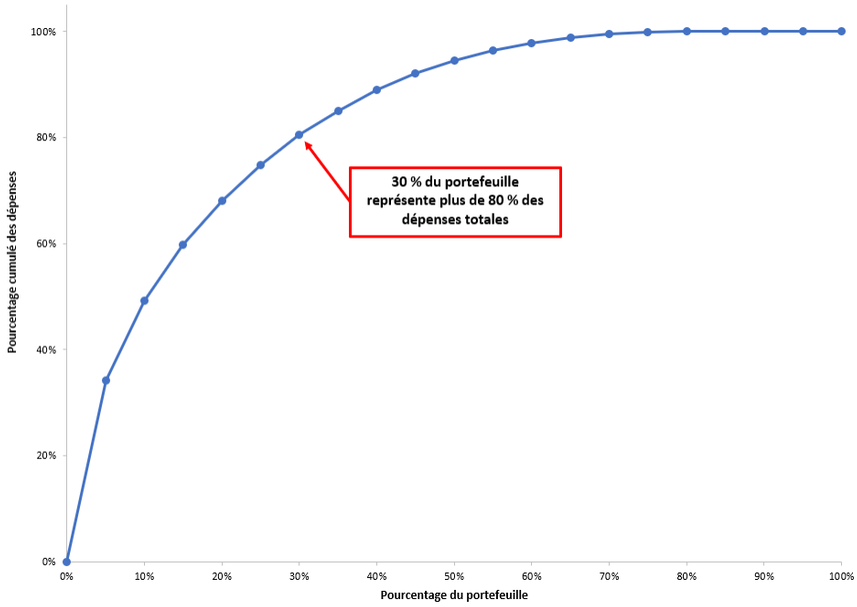


FIGURE 2.29 – Fonction de répartition des dépenses en fonction du nombre d'individus

Dans le cas présent, 30 % du portefeuille représente près de 80 % des frais réels.
Les statistiques présentées par la suite ne concernent que les 20 % du portefeuille consommant le plus. Pour chaque statistique, une comparaison est effectuée entre les chiffres des gros consommants et ceux de l'ensemble du portefeuille :

Population	Assurés	Conjoints	Enfants
Gros consommants	44%	32%	24%
Portefeuille	37%	25%	38%

Exercice	Cadre	Employé	Ouvrier	Technicien
Gros consommants	29%	33%	31%	6%
Portefeuille	29%	31%	33%	7%

Age moyen	Assuré	Conjoint	Enfant	Total
Gros consommants	47,5	46,9	13,1	39,1
Portefeuille	45,6	45,7	13,1	33,2

Assurés	Homme	Femme
Gros consommants	72%	28%
Portefeuille	80%	20%

FIGURE 2.30 – Caractéristiques des gros consommants du portefeuille

Les principales enseignements tirés des statistiques précédentes sont les suivants :

- La proportion de femmes est plus importante chez les gros consommants que sur l'ensemble du portefeuille ;
- Les gros consommants sont globalement plus âgés que la moyenne du portefeuille ;
- Les enfants sont minoritaires chez les gros consommants.

3 Présentation théorique des modèles utilisés au cours de l'étude

Ce chapitre va définir les principaux points théoriques des modèles mis en place au cours de ce mémoire, permettant ainsi de mieux comprendre leur utilisation et leur interprétation.

3.1 Introduction sur la mise en place des modèles

3.1.1 Typologie des modèles

Apprentissage supervisé / non supervisé

L'**apprentissage supervisé** est la forme la plus répandue de *Machine Learning*. Elle présuppose que l'on dispose d'un ensemble d'exemples où sont distingués des variables prédictives ou explicatives et une variable cible ou à expliquer. On parle également de données étiquetées.
L'objectif durant la phase d'apprentissage est de généraliser l'association observée entre les variables explicatives et la variable cible.

L'**apprentissage non supervisé** ne présuppose aucun étiquetage préalable des données d'apprentissage.
L'objectif dans ce cas est que le modèle parvienne de lui-même à regrouper en catégories les données fournies en exemple. L'apprentissage non supervisé peut donc s'apparenter à une cartographie de l'espace des variables.

L'ensemble des modèles testés au cours de ce mémoire sont des modèles à apprentissage supervisé.

Par souci de clarté, les notations suivantes sont définies dès maintenant :

- Y : variable cible, variable à expliquer, variable réponse ou variable d'intérêt ;
- $\hat{Y}$: prédiction par le modèle testé de la variable cible ;
- X : l'ensemble des variables explicatives.

Régression / Classification

On distingue communément la variable cible en deux catégories :

- Les variables quantitatives qui sont des variables continues, par exemple lorsque la variable à expliquer appartient à l'ensemble des réels.
- Les variables qualitatives qui sont définies en catégories ou classes.

La distinction entre ces catégories de variables est indispensable. En effet, l'analyse n'est pas le même si la variable à expliquer est quantitative ou qualitative.

On parle de **régression** lorsque Y est une variable quantitative. Le modèle qui correspond à une régression peut s'écrire de la façon suivante :

$$Y = f(X) + \epsilon$$

avec :

$$\epsilon \textit{ un bruit aléatoire}$$
$$f \textit{ fonction de régression a priori inconnue}$$

L'objectif de ce genre de problème est de caractériser f grâce aux observations du couple (X, Y). En supposant qu'ϵ soit centré conditionnellement à X, il existe une unique fonction f qui satisfasse l'égalité suivante :

$$f(X) = \mathbb{E}[Y|X]$$

Une méthode de régression répond donc au problème suivant : "Étant donné un échantillon de variables potentiellement explicatives, quelle est la relation entre ces variables qui va représenter de façon optimale la variable cible ?".

On parle de **classification** lorsque Y est une variable qualitative ou discrète. On cherche donc ici à trouver le groupe d'appartenance de chaque observation de la variable à expliquer. Un problème de classification peut donc être formulé en ces termes : "Sachant qu'il existe un lien certain entre les variables explicatives et la variable à expliquer, à quel groupe appartient cette observation ?".

Dans le cadre de ce mémoire, deux grandeurs sont à modéliser :

- La fréquence ;
- La consommation moyenne sur l'exercice.

Il semble évident que **la modélisation de la charge totale répond à une problématique de régression**, cette variable étant continue.
Le cas de la modélisation de la fréquence est moins simple. A première vue, un problème de classification peut sembler adapté puisque la variable à expliquer est discrète. Néanmoins, le but de cette étude est de prédire un comportement moyen de consommation sachant les caractéristiques individuel. Il s'agit bien d'approcher l'expression $\mathbb{E}[Y|X]$.
La modélisation de la fréquence est donc également un problème de régression.

Enfin, **l'étude des caractéristiques des gros consommants s'apparente à un problème de classification**. En effet, le but recherché ici est bien de trouver si tel ou tel individu appartient à la classe "Gros consommants".

3.1.2 Démarche générale de mise en place des modèles

Lors des différentes modélisations réalisées au cours de ce mémoire, certaines étapes indispensables sont communes à chacun des modèles présentés dans ce chapitre. Le schéma suivant permet de résumer cette démarche générale qui aboutit au choix du meilleur modèle possible, quelque soit le modèle sous-jacent :

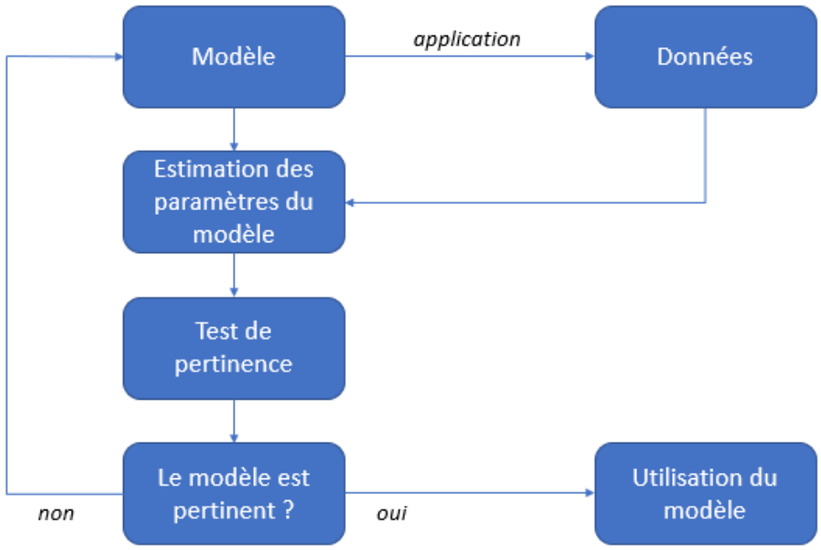


FIGURE 3.1 – Démarche générale d'implémentation d'un modèle

En effet, ce chapitre vise avant toute chose à appréhender de manière théorique les modèles utilisés. La compréhension de ces modèles permettra par la suite de mieux cerner les paramètres importants propres à chaque modèle et ainsi accélérer leur implémentation.

Cette partie introductive est aussi l'occasion de présenter les tests de pertinence utilisés au cours de ce mémoire.

Pour l'étude des gros consommants

Le critère de pertinence retenu est la **courbe ROC** (ou Receiver Operating Characteristic). Cet indicateur permet de classer les observations selon un seuil qui donne la prévision retenue. En effet, dans le cadre d'une variable à expliquer binaire, les prédictions sont des "scores" ou une probabilité et on a besoin de prédictions qui prennent les valeurs 0 ou 1.
De ce fait, on définit un seuil d'acceptation tel que :

— $\hat{Y} = 1$ si $P\left(\hat{Y} = 1 | X = x\right) > s$;

— $\hat{Y} = 0$ si $P\left(\hat{Y} = 1 | X = x\right) < s$.

On définit alors les valeurs suivantes :

— VP est le nombre de vrais positifs, i.e les observations classées positivement (score obtenu supérieur au seuil défini) et qui le sont effectivement ;

— FN est le nombre de faux négatifs, i.e les observations classées négatives (score obtenu inférieur au seuil défini) alors qu'elles sont positives ;

— FP est le nombre de faux positifs, i.e les observations classées positives alors qu'elles sont négatives ;

— VN est le nombre de vrais négatifs, i.e les observations classées négatives et qui le sont effectivement ;

— i est le nombre total d'observations.

On se sert de ces valeurs pour calculer les ratios suivants :

— Le taux d'erreur égal à $\frac{FN+FP}{i}$ et qui représente le taux d'observations mal classées ;

— Le taux de vrais positifs (ou sensibilité [1]) égal à $\frac{VP}{VP+FN}$ indique la capacité qu'a le modèle à retrouver les individus positifs ;

— Le taux de faux positifs (égal à 1 - spécificité) [2] égal à $\frac{FP}{FP+VN}$.

La courbe ROC est la représentation graphique du taux de vrai positif en fonction du taux de faux positif en faisant varier le seuil. L'aire sous la courbe (**AUC**) est ainsi la probabilité qu'un modèle classe une observation positive choisie aléatoirement au-dessus d'une observation négative. Plus l'aire sous cette courbe se rapproche de 1, plus le modèle est pertinent. Une représentation graphique est néanmoins plus pertinente :

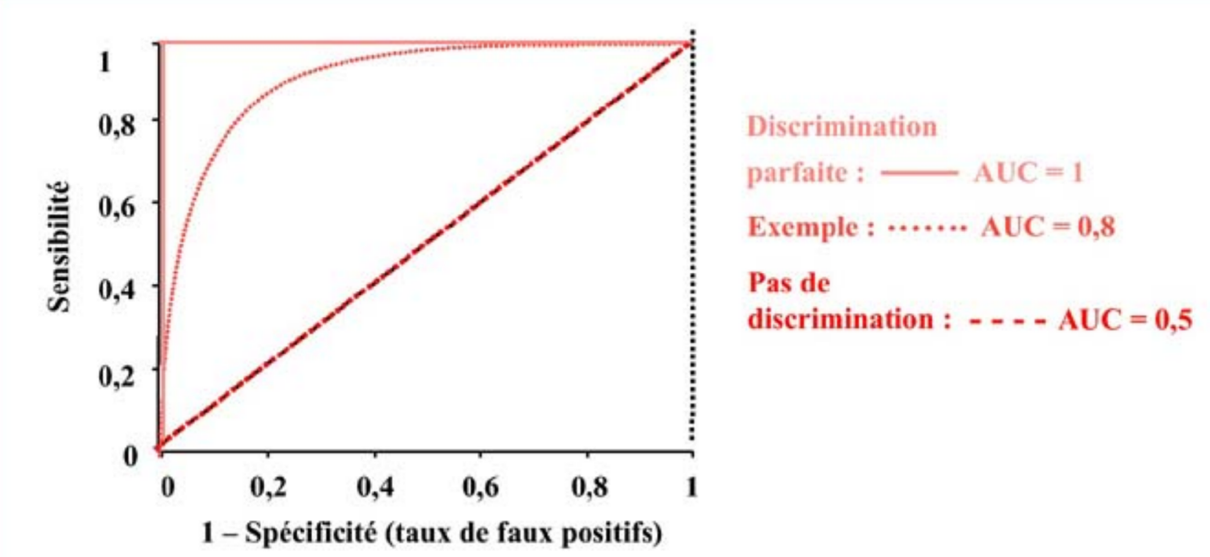


FIGURE 3.2 – Courbes ROC selon différentes valeurs de l'AUC

Source : jle.com

<u>Pour l'étude de la consommation médicale</u>

Ici, la notion de pertinence retenue est la *Mean Square Error* (MSE), qui mesure l'écart entre les prédictions du modèle et les valeurs observées :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

1. proportion à prédire l'événement.
2. la spécifité correspond à la capacité du modèle à prédire le non-événement.

3.1.3 Quelques notions importantes

Le compromis biais / variance

Il est important de préciser qu'il existe deux types d'erreur principaux dans le domaine du *Machine Learning* :

- **Le biais** qui provient d'hypothèses erronées dans l'algorithme d'apprentissage. Un modèle trop simple pourrait ne pas capter certaines relations entre les variables explicatives et la variable à expliquer. On parle alors de sous-apprentissage.
- **La variance** correspond à la sensibilité du modèle aux petites fluctuations de l'échantillon d'apprentissage. Une variance élevée est signe de surapprentissage.

Un schéma sera néanmoins plus parlant :

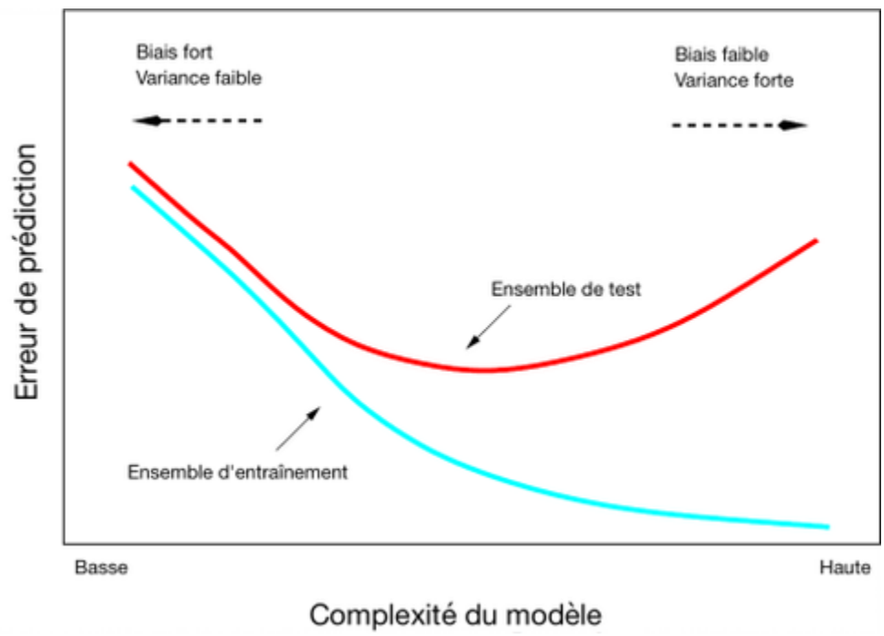


FIGURE 3.3 – Compromis biais / variance

Source : quantmetry.com

Le surapprentissage

Le surapprentissage caractérise une situation dans laquelle un modèle reproduit avec un très (trop) grande exactitude les données d'apprentissage. Ce modèle devient dès lors impossible à généraliser à d'autres observations que celles des données d'apprentissage.

Bases apprentissage / test

Afin d'éviter le sur-apprentissage, il est nécessaire de découper la base de données en deux sous-ensembles :

- La **base d'apprentissage** qui, comme son nom l'indique, permet au modèle d'"apprendre" des données. Plus concrètement, c'est sur cette base que les paramètres sont calibrés et optimisés.
- La **base de test** qui permet de mettre en oeuvre le modèle construit sur la base d'apprentissage et vérifier son adéquation ou pertinence sur des données indépendantes de la base d'apprentissage.

Validation croisée

La validation croisée consiste à découper son jeu de données en k jeux de données de taille identique. Dans le cas où l'on partitionne la base de données en dix *folds*[3], le modèle est alors lancé 10 fois. A chaque fois, le modèle utilisera neuf des dix *folds* comme base d'apprentissage et le deuxième comme base de test.
La validation croisée permet à la fois :

— D'éviter ou du moins de réduire le phénomène de sur-apprentissage comme un plus grand nombre de données est utilisé comme base d'apprentissage ;

— De s'intéresser à la variance des erreurs de validation croisée. Une variance élevée indique que de faibles changements dans les données entraînent des prédictions très différentes : le modèle est donc instable.

Le schéma suivant permet de résumer cette notion :

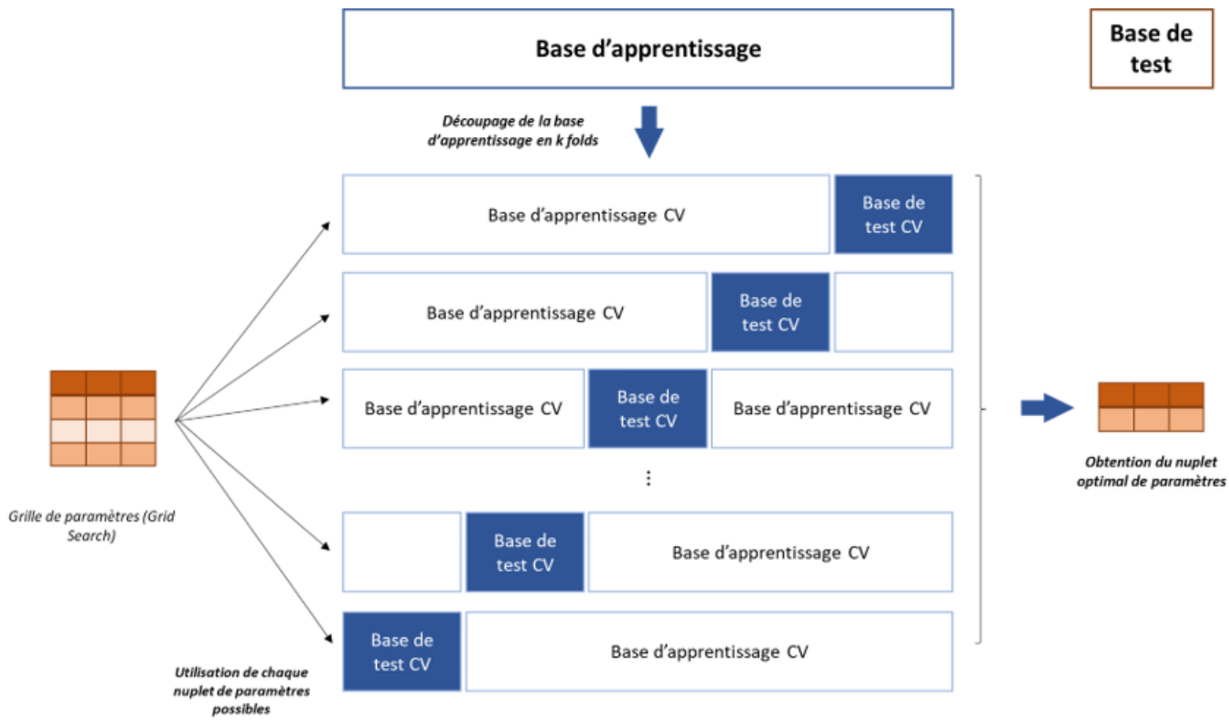


FIGURE 3.4 – Hyper-paramétrage et validation croisée

3. c'est le terme communément employé pour désigner les parties issues du découpage de la base initiale

3.2 Les modèles linéaires généralisés (GLM)

3.2.1 Remarques préliminaires

Comme leur nom l'indique, les modèles linéaires généralisés, définis pour la première fois par NELDER et MC CULLAGH [1983] [20], sont une extension des modèles linéaires : **ils permettent de modéliser une relation non-linéaire entre la variable à expliquer et les variables explicatives**.
Or, en assurance, de nombreux phénomènes tels que la probabilité d'avoir un sinistre ou celle de résiliation d'un contrat, ne peuvent être modélisés par une simple régression linéaire. L'utilisation de GLM prend donc tout son sens.

Les GLM permettant de modéliser plusieurs phénomènes ainsi que plusieurs relations entre les variables, le schéma suivant permet également de résumer de manière non-exhaustive les **choix de régressions possibles** dans le cadre d'un GLM :

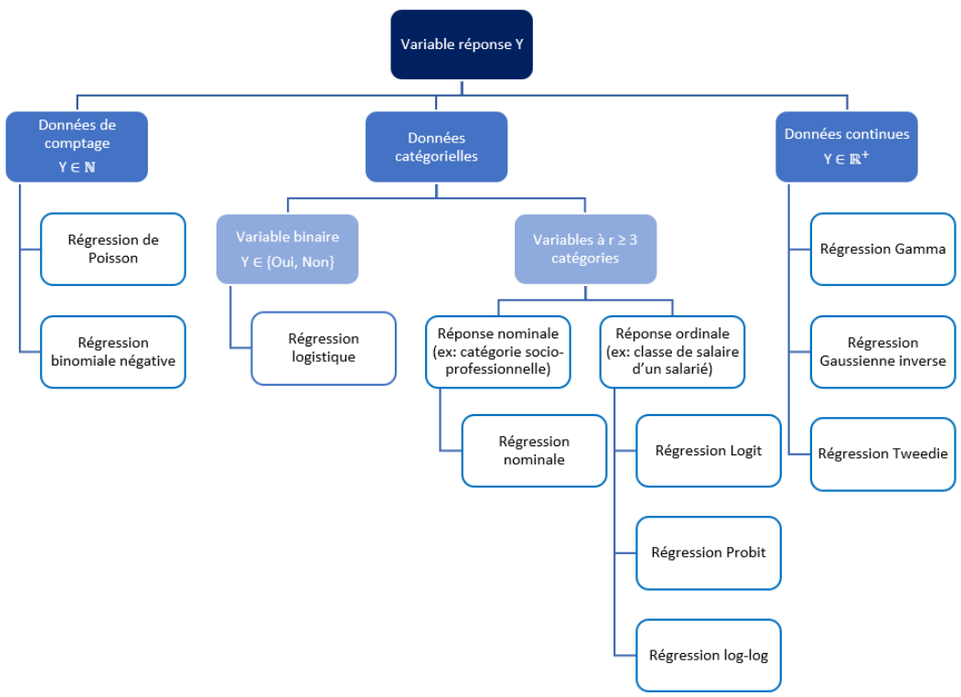


FIGURE 3.5 – Schéma classique de choix de régression pour un GLM

3.2.2 Comment mettre en place un GLM ?

Le modèle

Le GLM se distingue en trois composantes.

La composante aléatoire

C'est la variable à expliquer $Y = (Y_1, \ldots, Y_n)'$ avec n le nombre d'individus.
Sa densité appartient à la famille exponentielle, c'est-à-dire qu'elle peut s'écrire sous la forme :

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

avec :

$$\begin{aligned}
&\theta \in \mathbb{R} = \textit{paramètre canonique ou de la moyenne}\\
&\phi \in \mathbb{R} = \textit{paramètre de dispersion},\\
&\quad a\ \textit{fonction définie sur}\ \mathbb{R}\ \textit{non nulle}\\
&\quad b\ \textit{fonction définie sur}\ \mathbb{R}\ \textit{deux fois dérivable}\\
&\quad c\ \textit{fonction définie sur}\ \mathbb{R}^2
\end{aligned}$$

La plupart des lois communes appartiennent à la famille exponentielle, telles que les lois Normale, Poisson, Gamma ou encore la loi Binomiale négative.

La composante déterministe

Soit $i = 1, \ldots, n$.
Pour chaque Y_i, on dispose de la valeur d'un *p-uplet* (si on dispose de p variables explicatives) $(X_{1i}, \ldots, X_{pi})'$. Les vecteurs $X_j = (X_{j1}, \ldots, X_{jn})'$ pour $j = 1, \ldots, p$ sont les vecteurs explicatifs.

La fonction lien

C'est une fonction g définie sur $\mathbb{R}$, strictement monotone et inversible, telle que :

$$g_n\left(\underbrace{\mathbb{E}[Y]}_{\mu}\right) = \underbrace{\beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p}_{\eta = score} = X \cdot \beta$$

avec $g_n : \mathbb{R}^n \to \mathbb{R}^n, (x_1, \ldots, x_n) \to (g(x_1), \ldots, g(x_n))$
On a avec nos notations :

$$g\left(\underbrace{\mathbb{E}[Y_i]}_{\mu_i}\right) = \underbrace{\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_p \cdot x_{ip}}_{\eta = score} = x_i' \cdot \beta = \eta_i$$

Chacune des lois de probabilités appartenant à la famille exponentielle possède une fonction de lien particulière, dite canonique, et définie par $\theta = \eta$.
Le lien canonique est tel que $g(\mu_i) = \theta_i$, or on sait que $\mu_i = b'(\theta_i)$. On a donc formellement $g^{-1} = b'$.
Voici un tableau récapitulatif de quelques lois de probabilité appartenant à la famille exponentielle ainsi que leur fonction de lien canonique :

Loi	Nom du lien	Fonction de lien
Normale	Lien identité	$g(\mu) = \mu$
Poisson	Lien Log	$g(\mu) = \ln(\mu)$
Gamma	Lien inverse	$g(\mu) = \frac{1}{\mu}$
Binomial	Lien Logit	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$

FIGURE 3.6 – Fonctions de liens canoniques associées aux principales lois de probabilité de la famille exponentielle

Sélection des variables

Cette phase, qui permet de trouver les variables qui expliciteront le mieux la variable réponse Y est primordiale. Le nombre de modèles étant fini, il est possible d'effectuer une recherche exhaustive en explorant l'ensemble des combinaisons de variables et en choisissant celle qui optimise un critère défini. Ceci dit, devant un grand nombre de variables explicatives ($2^p - 1$ avec p variables explicatives), la recherche exhaustive peut laisser place à une recherche pas à pas.
Il existe trois méthodes de sélection pas à pas :

- **La méthode ascendante (*forward selection*)** consiste en l'ajout à chaque étape de la variable qui conduit à l'optimisation du critère de choix. Si aucune variable ne permet l'optimisation du critère de choix ou que toutes les variables sont intégrées, alors on arrête.
- **La méthode descendante (*backward selection*)** consiste à la suppression à chaque étape de la variable qui conduit à l'optimisation du critère de choix. Si aucun retrait ne permet d'optimiser le critère ou que toutes les variables ont été retirées, alors on arrête.
- **La méthode progressive (*stepwise selection*)** consiste en une méthode de sélection ascendante dans laquelle on intercale entre deux étapes une étape d'élimination (ceci permet d'éliminer des variables introduites en amont et n'ayant plus d'effet après introduction de nouvelles variables).

Il convient maintenant d'évoquer les critères de sélection utilisés ; deux d'entre eux sont généralement utilisés :

- **L'Akaike Information Criterion (AIC)** défini de la manière suivante : $AIC = -2\ln(L(\beta)) + 2N$ où :
  - N est le nombre de variables du modèle ;
  - L est la vraisemblance du modèle.

  L'AIC prend donc en compte à la fois la qualité de l'ajustement via la fonction de vraisemblance (qui dépend des coefficents du GLM) et de la complexité du modèle (via le nombre de variable).
- **Le Bayesian Information Criterion (BIC)** défini de la manière suivante : $BIC = -2\ln(L(\beta)) + 2N\ln(n)$ où n est le nombre d'observations. Néanmoins, le BIC n'est pas forcément adapté à des bases de données de petites tailles. En effet, celui-ci a tendance à choisir des modèles trop simples du fait de sa forte pénalisation.

Ces deux critères doivent être minimisés. Un modèle sera considéré comme meilleur s'il présente un critère plus petit par rapport aux autres modèles testés.

Estimation des paramètres

Afin d'estimer les paramètres β_i du GLM, on utilise la méthode du **maximum de vraisemblance**.
Néanmoins, si la fonction de lien utilisée n'est pas la fonction canonique, il n'existe pas de solution analytique pour résoudre le système final. En revanche, on peut utiliser **l'algorithme de Newton-Raphson** afin d'obtenir une approximation de la solution.

Test de significativité

Cette étape s'apparente à la recherche d'un modèle optimal au regard d'un critère choisi.

Les critères d'ajustement du modèle sont les suivants :

- **La déviance** qui se base sur le modèle saturé [4] : $D = 2(\ln(L_{sat}) - \ln(L))$.
- **La statistique de Pearson** : $\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(Y_i)}$

4. modèle pour lequel le nombre de variables est égal aux nombres d'observations

Le modèle est jugé de mauvaise qualité si ces critères sont supérieurs au quantile d'une loi du χ^2 à $n-p-1$ degrés de libertés.

S'intéresser à la significativité d'un modèle revient à tester les hypothèses suivantes :

$$\begin{cases} H_0 : \forall i, \beta_i = 0 \\ H_1 : \exists i, \beta_i \neq 0 \end{cases}$$

Un modèle est significatif si l'hypothèse H_0 est rejetée, *i.e* si la $p-value$ de la statistique du test est inférieure au seuil de significativité choisi.

Il existe trois tests permettant de choisir entre ces deux hypothèses. Le tableau suivant permet de résumer les principales informations sur deux de ces trois tests [5] :

Critère	Loi sous H_0	Statistique
Test du maximum de vraisemblance	χ^2	$G = -2\ln\left(\frac{vraisemblance\ sans\ la\ variable X_i}{vraisemblance\ avec\ la\ variable X_i}\right)$
Test de Wald	$N(0,1)$	$W = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$

FIGURE 3.7 – Deux tests permettant de vérifier la significativité globale d'un modèle *GLM*

3.2.3 Conclusion

Les modèles GLM sont relativement anciens et sont donc maîtrisés depuis longtemps dans le milieu de l'actuariat. Néanmoins, ces modèles souffrent de quelques défauts. Citons par exemple :

- **Le fait que le modèle soit paramétrique** : fixer une loi pour la variable d'intérêt est évidemment restrictif ;
- **La modélisation des interactions entre les variables**, bien que possible, relève souvent de l'avis d'expert, au même titre que leur sélection en amont. D'autant qu'une variable éliminée par une méthode basée sur l'AIC peut se révéler intéressante si couplée à une autre variable. Les modèles sont parfois longs à s'exécuter et l'on ne peut se permettre de tester toutes les interactions envisageables pour conserver le meilleur modèle.

De ce fait, les modèles non-paramétriques peuvent avoir tous leurs intérêts afin de combler les inconvénients des modèles GLM.

5. Le troisième test est le test du Score qui suit également une loi du χ^2 sous H_0

3.3 Une méthode d'arbre de décision : la méthode CART

3.3.1 Généralités

La technique des arbres de décision est fondée sur la classification de la variable réponse Y par une suite de tests sur les variables explicatives. Ces tests sont organisés de façon hiérarchique, dans le sens où la réponse à un test influence les réponses suivantes.

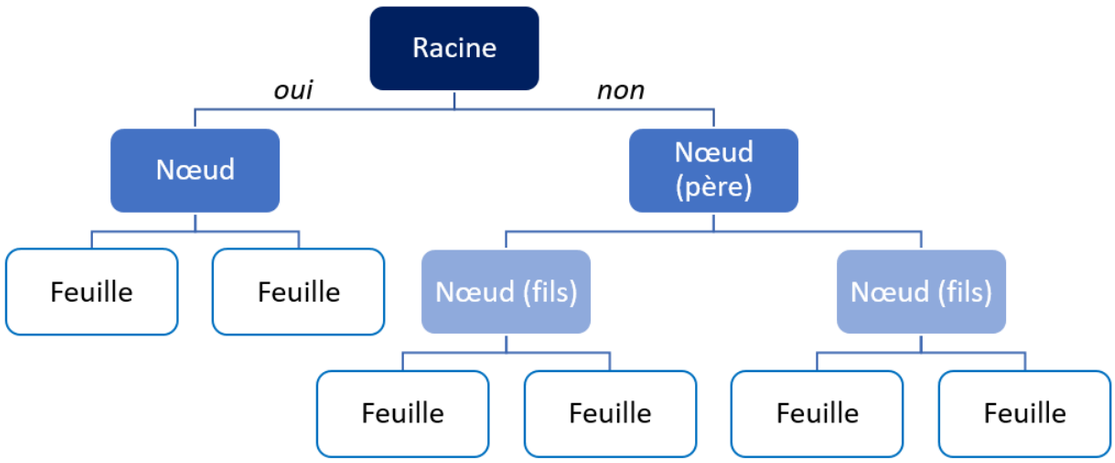


FIGURE 3.8 – Illustration de la structure d'un arbre de décision

La **racine** comporte l'ensemble de la population étudiée. Une **feuille** de l'arbre est associée à une classe d'individu et un **noeud** est associé à un test (une forme de sélection). Selon la réponse au test, on se déplacera vers tel ou tel fils du noeud. Le processus de classification est donc récursif, jusqu'à la rencontre d'une feuille (ou noeud terminal).
Ce type de structure forme un arbre de décision.

Les arbres de décision constituent une famille majeure de méthodes de *Data Science*.
Leur objectif est de répartir un ensemble d'individus en groupes les plus homogènes possibles du point de vue de la variable à expliquer, à partir de données portant sur les variables explicatives.

La logique d'implémentation des arbres de décision diffère fortement de celle des GLM. Celle-ci permet d'ailleurs de corriger un certain nombre de limites des GLM repérées un peu plus haut :

- Avant toute chose, **cette méthode est non-paramétrique** et il est donc inutile de définir une loi de probabilité pour la variable Y.
- **Les arbres ne sont pas monotones**, c'est-à-dire qu'une même variable explicative peut avoir différents effets, par exemple s'il existe une valeur seuil en dessous de laquelle cette variable a un effet positif sur la variable d'intérêt et au-dessus un effet négatif, un noeud pourrait être construit en cette valeur seuil pour partager la population en deux groupes plus homogènes et partageant le même effet.
- **Les interactions entre variables explicatives** peuvent être modélisées dans un arbre de décision. Ce sont même ces différentes interactions qui forment les différents noeuds.

Il existe un certain nombre de techniques d'arbres de décision, parmi lesquelles la méthode *CHAID* (*Chi-Squared Automatic Interaction Detector*) publiée par KASS [1980] [16], *CART* (*Classification and Regression Trees*) développée par BREIMAN *et al.* [1984] [17], ou encore la méthode *C4.5*, publiée par QUILAN [1992] [21].
Ayant utilisé au cours des différentes applications la méthode *CART*, c'est cette méthode qui sera développée par la suite.

Il convient enfin de distinguer deux types d'arbres en fonction de la nature de la variable d'intérêt :

— On parle d'**arbres de régression** si la variable réponse est quantitative ;

— L'arbre de décision est nommé **arbre de classification** si la variable réponse est qualitative.

Le principe de construction de l'arbre reste néanmoins identique.

3.3.2 La construction d'un arbre de décision

Deux phases principales sont nécessaires à l'élaboration d'un arbre :

— Comment construire l'arbre maximal ?

— Comment choisir l'arbre optimal ?

Comment construire l'arbre maximal ?

Comme mentionné précédemment, la segmentation d'un arbre est réalisée par des choix successifs au niveau de chaque noeud, selon une règle de division binaire, de variables explicatives contributives dans l'explication de la variable d'intérêt, afin d'assurer l'homogénéité des groupes créés.

Parmi tous les partages possibles explorés sur l'ensemble des variables explicatives, l'algorithme CART retient celui qui maximise la quantité suivante, appelée **gain d'information** :

$$\Delta = I_{noeud\ père} - (I_{premier\ noeud\ fils} + I_{deuxime\ noeud\ fils})$$

avec :

$$I = fonction\ d'impureté$$

Le choix du partage qui maximise la gain d'information revient donc à rechercher la séparation qui minimise l'hétérogénéité intra-classes des groupes ainsi créés.

Afin d'effectuer cette séparation , **une fonction d'hétérogénéité ou fonction d'impureté** doit donc être définie. Cette fonction croit avec l'hétérogénéité des valeurs prises par la variable d'intérêt au sein d'un même noeud.
La forme de cette fonction d'impureté diffère selon la nature de la variable à expliquer :

— Dans le cas de la classification (variable d'intérêt qualitative), deux fonctions d'impureté sont couramment utilisées :

- L'indice de Gini :

$$I_{Gini} = \sum_{i \neq j} \mathbb{P}(i|t)\, \mathbb{P}(j|t)$$

  où $\mathbb{P}(i|t)$ désigne la proportion d'éléments de la classe i affectés au noeud t

- La fonction d'entropie :

$$I_{entropie} = \sum_{i \neq j} -\mathbb{P}(i|t) \ln(\mathbb{P}(i|t))$$

— Dans le cas de la régression (variable d'intérêt quantitative), c'est la variance intra-noeud qui est utilisée.

Voici un schéma récapitulant l'algorithme (récursif) de création d'un noeud :

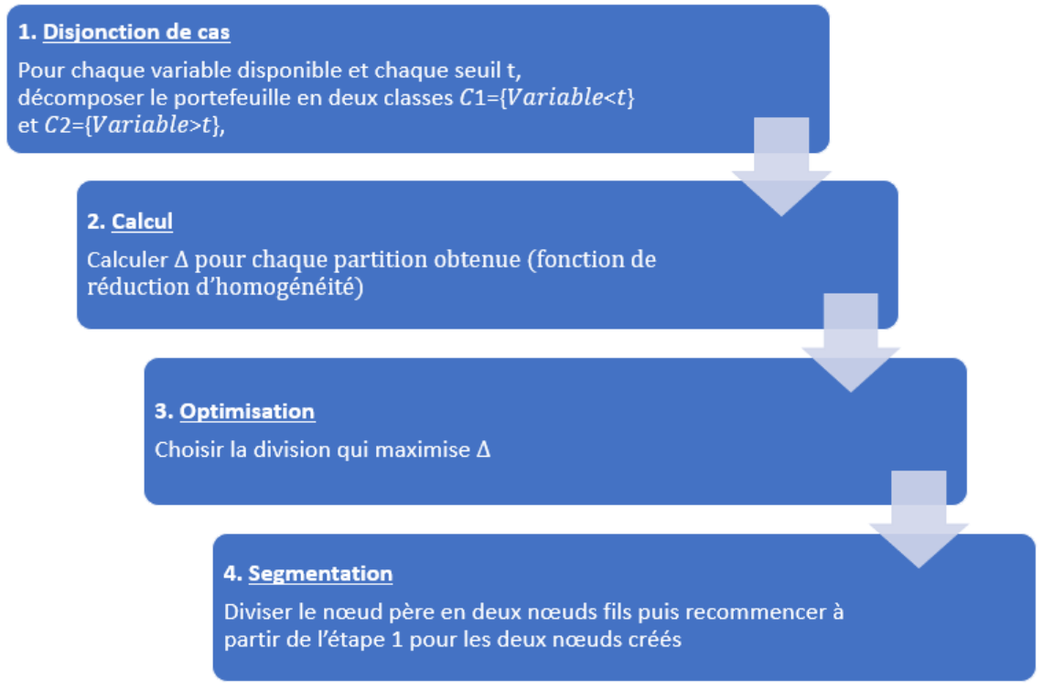


FIGURE 3.9 – Algorithme de création d'un arbre

Une fois les segments de l'arbre construits grâce aux règles définies plus tôt et à la maximisation de la pureté de chaque noeud, la répartition des individus peut être effectuée, un individu ne pouvant être affecté qu'à un seul noeud.

Comment obtenir l'arbre optimal ?

L'arbre maximal est maintenant construit. Néanmoins, cette méthode, si elle s'arrêtait à ce stade, comporterait un inconvénient majeur. En effet, deux individus n'étant jamais à 100 % identiques, cet algorithme ne cesserait que lorsqu'un seul individu se trouverait au niveau des noeuds terminaux ou feuilles (on parle d'arbre saturé) : l'apprentissage serait alors parfait mais cet arbre ne pourrait n'être en aucun cas transposable à d'autres données dans le cadre de prédiction. Il y a **sur-apprentissage**.
L'arbre obtenu à la phase précédente peut s'avérer très lourd à lire (de très nombreux noeud peuvent être présents).
Il faut alors définir un critère d'arrêt afin de contrôler la taille de l'arbre. Deux solutions principales existent :

- **Une optimisation "a priori"** : avant de commencer la construction de l'arbre, on impose des critères d'arborescence (nombre d'individus minimum par feuille, nombre de partitionnement maximum, ...) ;
- **Une optimisation "a posteriori"** : Cette phase opère donc dans le sens inverse de la précédente. Elle part de l'arbre de complexité maximale, puis supprime les noeuds les uns après les autres jusqu'à revenir à une unique classe couvrant tout le portefeuille.
  L'idée est de minimiser la quantité $c(T) + \alpha |T|$ où :
  - $c(T)$ la somme des variances de chacune des classes [6] ;
  - α coût de complexité de l'arbre (coût en terme d'erreur de l'addition d'un noeud dans le modèle) ;
  - $|T|$ le nombre de feuilles de l'arbre T.

  On cherche donc à réduire le facteur de complexité de l'arbre final en choisissant celui qui comporte le moins de branches terminales (*i.e.* le plus simple). On fait un compromis entre le coût d'erreur de classement en apprentissage et la complexité de l'arbre.

6. aussi appelé coût d'erreur

3.3.3 Conclusion

Comme vu précédemment, la méthode *CART* et plus généralement les arbres de décision permettent de résoudre de nombreux défauts présents dans les modèles *GLM*. De plus, les résultats sont aisément lisibles et interprétables graphiquement.
Néanmoins, ces méthodes présentent elles-aussi quelques inconvénients, en particulier leur manque de robustesse.

Sensibilité aux optimums locaux

C'est là l'une des plus importantes faiblesses des arbres de décisions. L'algorithme étudie les variables explicatives successivement. Les noeuds sont construits de façon enchaînée et un critère choisi pour figurer à un emplacement de l'arbre n'est plus réétudié par la suite. Cela sous-entend que modifier en amont la construction d'un critère fort peut remettre en question l'intégralité de l'arborescence.

Sur-apprentissage

Le problème du sur-apprentissage, mentionné lors de l'introduction à cette partie, peut particulièrement s'appliquer aux méthodes d'arbres de décision si la phase d'élagage n'est pas réalisée correctement. L'arbre s'imprègne alors de toutes les caractéristiques des données d'apprentissage, en particulier les caractéristiques des données aberrantes. Il les considère ensuite comme des comportements normaux et la généralisation à d'autres bases devient donc biaisée.

Les nombreux avantages des méthodes d'arbres de décision en font cependant un outil incontournable. De plus, les défauts qui viennent d'être cités peuvent être en partie corrigés par des modèles agrégés.

3.4 Quelques modèles agrégés

Trois modèles agrégés d'arbres de décisions vont maintenant être évoqués. L'idée essentielle de ces différentes méthodes est de pallier à l'instabilité des arbres de décisions en générant un grand nombre d'arbres puis de les réunir en une unique estimation. Cette procédure permet de réduire considérablement la variance des estimateurs.

3.4.1 *Bagging*

Le *Bagging*, contraction de *bootstrap aggregation* et proposé par BREIMAN [1994] [6] **construit une multitude d'arbres par *bootstrap* au niveau des individus** [7] : chaque échantillon est issu d'un tirage aléatoire avec remise sur la base d'apprentissage et est de même taille que celle-ci. Pour chaque tirage, on construit un arbre de taille maximale [8]. On estime enfin la variable d'intérêt en regroupant les différents résultats obtenus suivant les arbres construits :

— Si la variable d'intérêt est discrète, on examine pour chaque arbre dans quel profil l'individu a été classé et on retient celui où l'individu apparaît le plus souvent ;

— Dans le cas d'une variable à expliquer continue, on effectue la moyenne des différentes estimations.

Un rapide calcul permet de montrer que la variance de l'estimateur décroît avec le nombre de tirages.

Voici un schéma récapitulant la logique de cet algorithme :

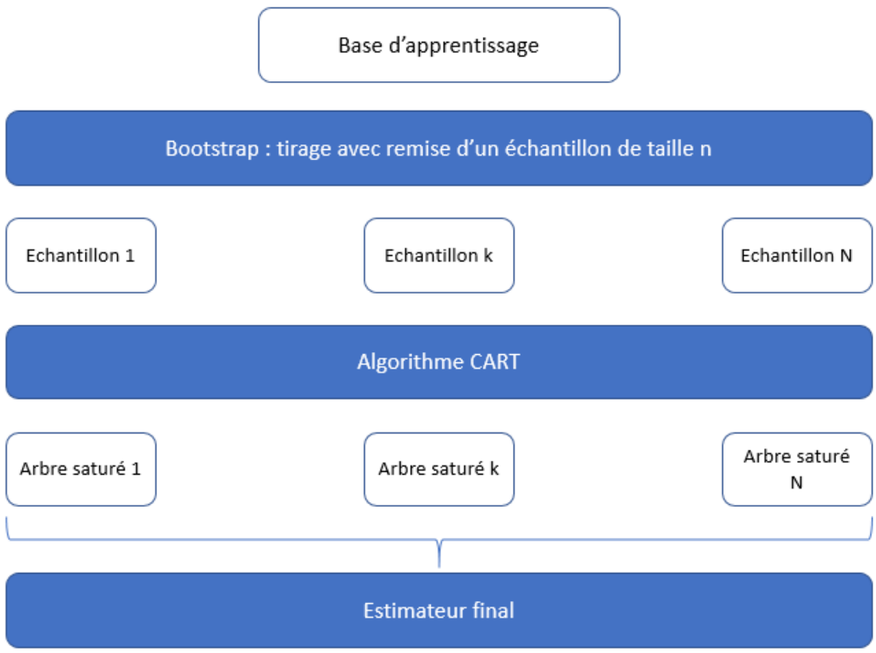


FIGURE 3.10 – Algorithme du *Bagging*

Cette méthode, plus robuste, perd cependant en lisibilité : on obtient une moyenne de résultats et non plus un simple arbre.

7. au niveau des lignes
8. sans élagage

3.4.2 Forêts aléatoires ou *Random Forest*

La méthode des forêts aléatoires est similaire au *Bagging* dans la manière d'agréger les modèles. Néanmoins, celle-ci peut être vue comme une amélioration du *Bagging*. Développée par BREIMAN [2001] [7], **la méthode des forêts aléatoires rajoute de l'aléa dans la sélection des variables explicatives au niveau de chaque noeud de l'arbre**.
Ajouter cet aléa dans la construction des arbres permet de les rendre plus « indépendants » et de réduire la variance de l'estimation. De plus, dans le cas où il y a de nombreuses variables explicatives, et particulièrement des variables explicatives avec un grand nombre de modalités, la sélection aléatoire d'un sous-groupe de variables explicatives permet d'utiliser des variables qui n'auraient peut-être pas été sélectionnées si elles avaient été confrontées à toutes les variables.

Le nombre par défaut de variables explicatives tirées aléatoirement dépend du type d'arbre. Si l'on suppose que l'on possède p variables explicatives :

— $\sqrt{p}$ pour un arbre de classification ;
— $\frac{p}{3}$ pour un arbre de régression.

De nouveau, un schéma permettra de mieux résumer la logique de cet algorithme :

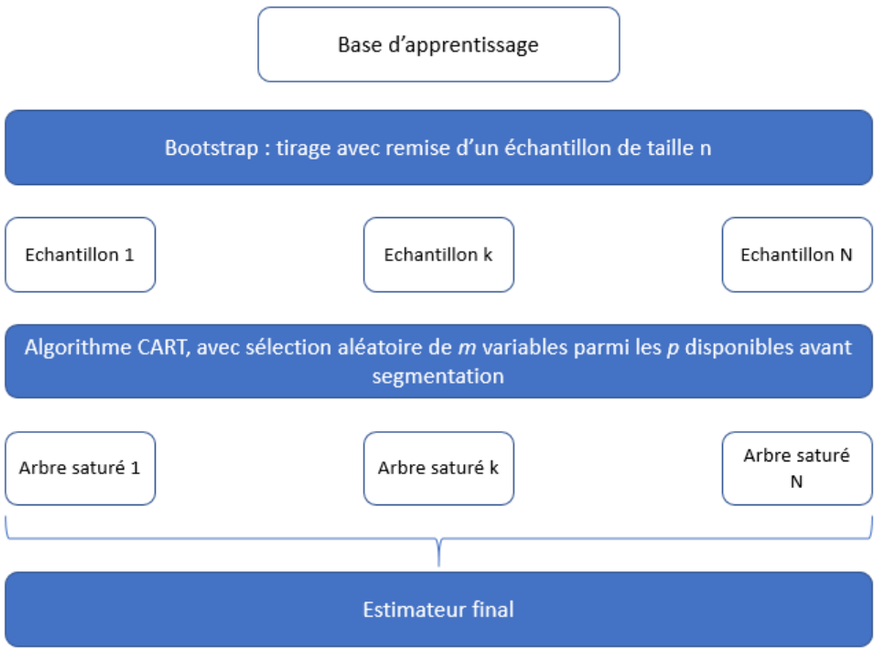


FIGURE 3.11 – Algorithme du *Random Forest*

3.4.3 *Boosting*

Le *Bagging* et les forêts aléatoires sont des méthodes que l'on pourrait qualifier de parallèles. En effet, ces deux méthodes construisent les arbres séparément les uns des autres.
Le *Boosting* s'inscrit dans une autre logique : il s'agit d'un algorithme adaptatif. On construit ici les arbres les uns après les autres, tel que chaque nouvel arbre corrige les défauts du précédent.

Le Boosting a cela de commun avec le Bagging qu'il promet une meilleure qualité de prédiction en agrégeant plusieurs modèles. Cependant, comme vu précédemment, plusieurs points le distinguent du *Bagging* :

— Utilisation de l'échantillon d'apprentissage entier et non d'échantillons bootstrapés ;
— Ajustement au fur et à mesure des poids des individus. A chaque itération, l'algorithme de Boosting se concentre sur les individus les plus difficiles à classer. Il augmente leurs poids

(proportionnellement à son défaut d'ajustement) et ne réagit pas par rapport au reste des individus : leur ajustement ne se fera qu'à l'itération suivante sans que les individus ayant été préalablement classés ne le soient à nouveau ;

— Le Boosting permet de réduire le biais d'un modèle. Dans le cas du Bagging, la moyenne de la somme des espérances des échantillons bootstrapés est égale à l'espérance d'un seul modèle : le biais du modèle final est inchangé contrairement au cas du Boosting ;

— Le modèle agrégé est basé sur une pondération de chaque modèle par sa qualité d'ajustement.

Le premier modèle nommé *adaBoost* a été développé par FREUND et SHAPIRE [1996] [23], mais ne s'appliquait qu'à une variable binaire. C'est le modèle qui a été décrit dans les paragraphes précédents.
Il existe depuis de nombreux algorithmes de *Boosting* qui diffèrent selon :

— **La pondération** (c'est-à-dire la façon de renforcer l'importance) **des observations mal estimées lors de l'itération précédente** ;

— **L'agrégation (ou plutôt la pondération) des modèles successifs** ;

— **La fonction perte**, qui peut être choisie plus ou moins robuste aux valeurs atypiques, **pour mesurer l'erreur d'ajustement** ...

Les plus fameuses d'entre elles sont :

— L'***Arcing (adaptively resample and combine)*** ou version aléatoire de l'*adaBoost*. Ainsi plutôt que de jouer sur les pondérations, à chaque itération, un nouvel échantillon est tiré avec remise, comme pour le bootstrap, mais selon des probabilités inversement proportionnelles à la qualité d'ajustement de l'itération précédente. La présence des observations difficiles à ajuster est ainsi renforcée pour que le modèle y consacre plus d'attention ;

— Le ***Gradient Boosting***, qui présente le même principe de base que l'*adaBoost*, c'est-à-dire construire une suite de modèles de sorte qu'à chaque étape, chaque nouveau modèle apparaisse comme un pas vers une meilleure solution. Dans le *Gradient Boosting*, ce pas est franchi dans la direction du gradient de la fonction perte ;

— L'***eXtreme Gradient Boosting*** qui est une adaptation du *Gradient Boosting*. Ce modèle comporte un certain nombre de paramètres. Il est donc à la fois robuste et complexe. Néanmoins, ce modèle possède un très fort pouvoir prédictif et est donc particulièrement utilisé.

4 Application de ces modèles à notre base de données

Cette partie permet de passer de la théorie à la pratique et de mettre en oeuvre les différents modèles expliqués dans la partie précédente. La base construite lors du deuxième chapitre sert de support à l'application de ces modèles.

Pour chacun des modèles, après avoir analysé leurs spécificités, un focus sera fait sur l'importance des variables (c'est-à-dire les variables ayant eu une importance dans la construction du modèle) ainsi que les résultats des tests de pertinence effectués.

Avant de passer à la présentation des sorties et spécificités des différents modèles, quelques informations utiles à la compréhension des données utilisées sont décrites ci-dessous.

4.1 Remarques préliminaires sur les données utilisées lors des modélisations

4.1.1 Création des variables d'intérêt

Comme déjà mentionné au cours de ce mémoire, **deux modélisations distinctes sont réalisées**. Dans un premier temps, une **tarification** est mise en place. Cette tarification repose sur une méthode "Fréquence /Coût moyen".

Ces deux variables sont construites de la façon suivante :

- **Fréquence** : la variable "Nombre d'actes" créée au moment des traitements effectués sur les bases CONSO [a] est utilisée lors de la modélisation ;
- **Coût moyen** : le coût moyen est égal au rapport de la somme des frais réels annuels sur la fréquence.

a. C.f. partie 2.2.1

Cette tarification ne repose donc pas sur les remboursements moyens de l'organisme complémentaire. En effet, cette tarification peut s'apparenter à une "tarification des risques". L'étude des frais réels permet de plus de s'émanciper des plafonds des garanties qui limitent donc le montant des remboursements de l'organisme complémentaire. L'étude des **gros consommants** est donc plus justifiable en s'intéressant aux frais réels.

Dans ce mémoire, les gros consommants sont apparentés au 20 % des individus consommant le plus, au sens du montant des frais réels totaux annuels. La variable "Gros consommants" est donc une variable binaire :

- 1 si les dépenses totales de l'individu sont supérieures au quantile à 80 % des dépenses du poste étudié ;
- 0 sinon.

4.1.2 Variables explicatives à disposition

Lors du chapitre présentant la construction de la base, de nombreuses variables sont mentionnées.

Néanmoins, l'ensemble des variables utilisées lors de la création de la base n'ont pas été conservées lors des différentes étapes de modélisation.

Voici la liste des variables explicatives utilisées lors des différentes modélisations :

- Le code bénéficiaire (Assuré / Conjoint / Enfant) ;
- Le croisement Adulte / Enfant : cette variable prend 3 modalités (Homme adulte / Femme adulte / Enfant) ;
- L'âge ;
- L'ancienneté, c'est-à-dire depuis combien de temps l'individu est affilié au contrat ;

Les variables "Age" et "Ancienneté" sont initialement des variables continues. Celles-ci sont discrétisées lors de la mise en place des modèles GLM. Par contre, ces variables sont utilisées telles quelles pour les autres modèles.

- La catégorie socio-professionnelle (Cadre / Employé / Ouvrier / Technicien) ;
- Congé maternité : si la personne a été en congé maternité durant l'exercice 2015 ;
- Licenciement : si l'individu a été licencié durant l'exercice 2015 ;
- Présence 2015 : combien de temps l'individu a été affilié au contrat sur l'année 2015 ;
- L'ensemble des variables externes (liste exhaustive disponible en annexe).

Comme le laisse supposer la description des variables ci-dessus, **les modèles sont implémentés sur l'exercice 2015**. Le tableau suivant résume les dimensions de la base :

Nombre de lignes ou d'individus	Nombre de colonnes ou de variables
6 510	30

FIGURE 4.1 – Dimension de la base utilisée lors des modélisations

Les remarques suivantes permettent de justifier les choix de suppression ou de conservation de variables :

- Certaines variables utilisées lors de la construction de la base et n'apportant pas d'informations supplémentaires sur les assurés, telles que le numéro de sécurité sociale ou le numéro d'enregistrement de l'acte médical, ont été supprimées ;
- La variable "Code postal" a été supprimée de l'étude. En effet, l'information contenue dans cette variable est prise en compte dans la variable externe "Zone" ;
- La tarification a été effectuée au niveau des postes de santé et non au niveau de chaque sous-famille [1] dans le but d'avoir une vision macro. Les variables spécifiques aux sous-familles ont donc été supprimées. Pour chaque poste de soins, seules les variables "Nombre d'actes" et "Dépense total" ont été conservées et servent à la construction des variables d'intérêt ;
- Les variables "Conjoint" et "Nombre d'enfants" spécifiques aux adhérents n'ont pas été prises en compte dans la suite de l'étude. Celles-ci ont néanmoins permis d'avoir une idée générale de la situation familiale des adhérents ;
- Enfin, il avait été décidé dans un premier temps de prendre en compte l'ensemble des variables "Consomme / Ne Consomme pas" dans les diverses modélisations dans le but d'amener une information sur l'évolution de la consommation des adhérents. Néanmoins, ces variables avaient tendance à écraser les autres variables d'une part et amenait un très fort sur- apprentissage d'autre part. Celles- ci ont donc été supprimées.
- La modélisation se focalise sur les individus présents sur l'exercice 2015. La variable "Présence 2015" est conservée et sera utilisée comme poids dans les modèles. Les autres variables "Présence" sont supprimées.

1. c.f. la Figure 2.7. résumant le découpage des actes médicaux

4.1.3 Séparation des données entre base d'apprentissage et base de test

Comme il l'a été mentionné en introduction du troisième chapitre, il est nécessaire de découper la base de données en deux sous-ensembles, à savoir en base d'apprentissage et en base de test.

Dans ce mémoire, et pour l'ensemble des modélisations effectuées, 80 % des observations composeront la base d'apprentissage et 20 % la base de test.

Les résultats des différents modèles vont maintenant être présentés. Dans un premier temps, les sorties et spécificités de chaque modèle sont détaillés pour la tarification du poste "Consultations Visites". Seuls les résultats importants sont exposés pour les autres postes. Enfin, l'étude des gros consommants" est effectuée.
L'ensemble de ces résultats permettent la réalisation d'un reporting santé, qui mélange statistiques descriptives et sorties de modèles *Data sciences*.

4.2 Modélisation de la fréquence au niveau du poste "Consultations Visites"

La partie suivante va permettre d'appliquer de façon concrète les modèles dont la théorie a été présentée précédemment.
Quatre modèles sont donc mis en place, à savoir :

- Le modèle linéaire généralisé ;
- Un modèle d'arbre de décision : le modèle CART ;
- Deux modèles d'arbres agrégés : les *Random Forest* et l'*eXtreme Gradient Boosting*

4.2.1 GLM

Avant toute chose, il est important de préciser que la mise en place des GLM n'est pas le but principal de ce mémoire. Ce modèle est simplement une base de comparaison afin de souligner les apports et limites des autres modèles testés.

Démarche de mise en place lors de la modélisation par GLM

1. **Étude des corrélations** ;
2. **Traitements des variables** :
   - Discrétisation des variables continues ;
   - Regroupements de modalités ...
3. **Choix de la fonction lien** ;
4. **Sélection des variables** :
   L'algorithme utilisé repose sur le critère AIC du modèle et la méthode descendante ou "*backward*".
   Un test du χ^2 de significativité des variables, basé sur la déviance du modèle, a été réalisé. La déviance étant un indicateur basé sur la vraisemblance (et donc la qualité d'ajustement) du modèle, plus la déviance est faible, plus le modèle est ajusté correctement.
5. **Importance des variables** :
   L'importance des variables peut être déduite de l'étape de sélection des variables par rapport à la diminution de l'AIC.
6. **Test de pertinence** :
   Calcul de la MSE sur la base de test.
7. **Analyse des coefficients de la régression**.

Etude des corrélations

De fortes corrélations sont observées entre certaines variables explicatives d'origine externes. En effet, certaines variables externes sont très liées. Ceci s'explique par le fait que l'ensemble de ces variables ont été rattachées à l'aide de la même variable : le code postal. Néanmoins, il a été choisi de conserver l'ensemble des variables externes à ce stade ; l'étape de sélection des variables permettant déjà d'éliminer un certain nombre de variables.
De plus, les variables "Code Bénéficiaire" et "Sexe" sont également corrélées de manière significative. La variable "Code bénéficiaire", peut-être moins interprétable, a donc été supprimée.

Traitements des variables continues

Les variables quantitatives "Age" et "Ancienneté" ont été découpées en classes, respectivement :

- Par tranches de 10 ans pour l'âge.

— Par tranches de 5 ans pour l'ancienneté.

Par la suite, des modalités de certaines variables seront regroupées afin de les rendre significatives dans le modèle GLM. Les tranches de la variable "Ancienneté" restent inchangées. Cependant, la variable "Âge" ne présente plus que quatre modalités :

— Inférieur à 10 ans ;

— Compris entre 10 et 20 ans ;

— Compris entre 20 et 50 ans ;

— Supérieur à 50 ans.

Choix de la loi et de la fonction de lien

Pour modéliser la fréquence, la loi de Poisson avec une fonction de lien Log est la plus adaptée.

Sélection des variables

La méthode "backward" conserve finalement 17 variables dont 13 variables externes. Néanmoins, ces variables ne ressortent pas toutes comme significatives selon le test de significativité présenté précédemment.

La sortie R de ce test est présentée ci-dessous :

Variable	Degrés de liberté	Déviance	AIC	*p-value*
Age	2	22404	33 239	<2 e-16
Proportion propriétaires	2	22 424	33 262	2,3 e-10
Ancienneté	2	22 454	33 292	<2 e-16
Proportion seniors	2	22 578	33 415	<2 e-16
Sexe	2	22 957	33 794	<2 e-16

FIGURE 4.2 – Sélection des variables par la méthode "backward" pour la modélisation de la fréquence

Ce test présente une *p-value* inférieure à 5 % pour ces cinq variables sélectionnées par la méthode "backward". Elles sont donc significatives.

Importance des variables

L'importance des variables peut être déduite de l'étape précédente par rapport à la diminution de l'AIC :

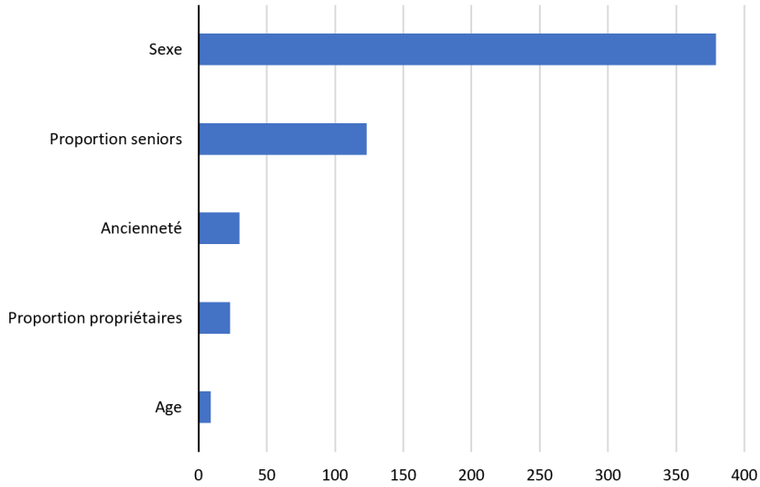


FIGURE 4.3 – Importance des variables dans la modélisation de la fréquence par le GLM

Deux variables externes , la proportion de seniors dans la population et la proportion de maisons apparaissent dans les variables importantes. Néanmoins, la variable "Sexe" joue un rôle très prépondérant dans la modélisation de la fréquence du poste "Consultations Visites".

Pertinence du modèle

Comme indiqué précédemment, le critère de pertinence retenu pour la modélisation de la fréquence est la MSE.
Cette valeur est calculée sur la base de test. Les résultats sont présentés dans le tableau suivant :

MSE base test
22,6

FIGURE 4.4 – MSE base test pour la modélisation de la fréquence par GLM

Cette valeur présentée dans le tableau ci-dessus sera comparée avec la MSE des autres modèles testés. Le modèle retenu sera celui avec la MSE la plus faible.

Analyse des résultats

Le logiciel *R* fournit les coefficients de la régression sous la forme suivante :
N.B. : La sortie R présente les coefficients pour l'ensemble des variables explicatives. Ici, par souci de clarté et de parcimonie, seules les lignes correspondantes aux cinq variables principales sont présentées.

	Estimate	*Std Error*	*Z-value*	*p-value*
(Intercept)	0,30	0,02	15,82	< 2 e-16
Age inférieur à 10	1,54	0,06	26,44	< 2 e-16
Age inférieur à 20	1,25	0,06	21,83	< 2 e-16
Age supérieur à 50	0,06	0,02	3,06	2,2 e-3
Sexe Enfant	-1,38	0,06	-24,42	< 2 e-16
Sexe Femme adulte	0,31	0,02	16,88	< 2 e-16
Ancienneté inférieure à 5 ans	-0,25	0,02	-11,89	< 2 e-16
Ancienneté supérieure à 15 ans	-0,04	0,02	-2,16	3 e-2
Proportion propriétaires faible	-0,11	0,02	-6,30	< 2 e-16
Proportion seniors faible	0,08	0,02	4,92	< 2 e-16
Proportion seniors élevée	-0,78	0,023	-23,87	< 2 e-16

FIGURE 4.5 – Coefficients de régression de la probabilité de consommer

Les coefficients de la régression sont détaillés au niveau de la colonne *"Estimate"*.
Les colonnes *"Z-value"* et *"p-value"* se réfèrent à la théorie des tests. Ce sont respectivement la valeur de la statistique et de la p-valeur associée au test de nullité des coefficients. Ce test est fondé sur l'hypothèse suivante :

$$H_0 : \beta = 0$$

Autrement dit, H_0 peut s'interpréter comme : "la modalité n'apporte rien par rapport à un modèle composé des mêmes autres modalités et dont elle serait absente".
Si la p-valeur est supérieure à 5 %, cela signifie que l'hypothèse H_0 est acceptée avec un risque d'erreur de 5 % et que la modalité en question n'a pas d'influence significative.
Ici toutes les *p-values* présentent des valeurs inférieures à 5 %, elles sont donc significatives.

Un des avantages du GLM est de fournir ces coefficients par modalité, ce qui permet de rapidement voir les effets principaux de chaque variable.
La première ligne nommée "Intercept" (communément noté β_0) correspond à l'assuré type, qui se réfère aux modalités les plus fréquentes. Ici, cet individu de référence correspond à un assuré disposant des caractéristiques suivantes :

- Age entre 20 et 50 ans ;
- Sexe Homme adulte ;
- Ancienneté entre 5 et 10 ans ;
- Proportion de maisons dans la commune normale par rapport à la moyenne nationale ;
- Proportion de seniors dans la commune normale par rapport à la moyenne nationale ;

L'étude des coefficients permet donc prédire qu'en moyenne, un individu présentant ces caractéristiques se rend chez un généraliste ou spécialiste avec une fréquence de :

$$\exp(\beta_0) = \exp(0,30) = 1,35$$

En moyenne, ce type d'individu se rend $1,35$ fois par an chez le médecin.
Les autres probabilités sont ensuite obtenues par les valeurs des β_i. Par exemple, la fréquence moyenne sur l'exercice 2015 d'une femme présentant les mêmes caractéristiques de l'individu type est la suivante :

$$\exp(\beta_0 + \beta_{Femme}) = \exp(0,30 + 0,31) = 1,84$$

Ce résultat peut paraître cohérent dans le sens où, selon une étude de l'INSEE parue en 2010, *"tout au long de leur vie, les femmes sont plus attentives à leur état de santé et plus proches du système de soins que les hommes : elles sont plus nombreuses à déclarer consulter des médecins généralistes ou spécialistes et à recourir à la prévention"*[2].

2. Santé et recours aux soins des femmes et des hommes, Premiers résultats de l'enquête Handicap-Santé 2008

4.2.2 CART

La constuction d'arbre CART a été effectuée à l'aide de la librairie *rpart* du logiciel *R*.
Au contraire du GLM, les arbres CART ne nécessite pas d'étude de corrélations ou de transformations des variables explicatives. Le modèle CART est donc lancé avec l'ensemble des variables explicatives [3].

Démarche mise en place lors de la modélisation par CART

1. **Construction de l'arbre maximal** ;
2. **Élagage de l'arbre maximal** :
   La fonction *rpart* permettant de lancer les modèles CART sous le logiciel *R* offre de multiples paramètres sur lesquels jouer afin d'obtenir le modèle optimal. Il a été ici fait le choix de s'intéresser à trois paramètres :
   - ***xval*** permettant d'effectuer une validation croisée..
   - ***minbucket*** : le nombre minimal d'observations observées dans chaque feuille.
   - ***cp*** ou critère de complexité. Ce paramètre s'apparente au α présenté au sein du paragraphe sur l'optimisation "a posteriori" de l'arbre. Ce paramètre est donc le coefficient de pénalisation de l'arbre. Plus le *cp* est important, moins l'arbre possède de feuille.
3. **Visualisation de l'arbre optimal** ;
4. **Importance des variables** :
   L'arbre construit permet de visualiser les variables actives, c'est-à-dire celles qui participent directement à la procédure de construction et donc à la discrimination et à la prévision correspondante. Néanmoins certaines variables explicatives n'apparaissant pas dans l'arbre élaboré ont pu être pour plusieurs nœuds concurrentes des variables actives. La connaissance de ces variables peut permettre d'établir une hiérarchie de l'ensemble des variables explicatives.
5. **Test de pertinence** :
   Calcul de la MSE sur la base de test.

Construction de l'arbre maximal

Comme indiqué dans la partie présentant les modèles, la première phase consiste en la construction d'un arbre dit maximal, c'est-à dire le plus complet et donc le plus complexe possible. Cet arbre comporte un nombre très important de feuilles et ne sera pas présenté ici.

Elagage de l'arbre maximal

Les valeurs des différents paramètres entrant en compte dans l'élagage de l'arbre sont les suivantes :

- ***xval*** : ce paramètre a été fixé à 5 (une valeur trop élevée pourrait entraîner un sur-apprentissage). La base de départ est donc découpée en 5 parties ; le modèle apprenant à chaque fois sur 4 d'entre elles et la cinquième servant à la validation.
- ***minbucket*** : plusieurs paramètres ont été testés a posteriori. La valeur retenue au final est 100. Cette valeur évite le sur-apprentissage et permet l'observation d'un nombre acceptable de feuilles.

3. Seule la variable "Présence" est supprimée du modèle. Celle-ci joue néanmoins le rôle de "poids" dans le modèle par la suite.

La démarche permettant d'obtenir la valeur du critère de complexité ou **cp** va maintenant être explicitée.
Dans un premier temps, ce critère est fixé à zéro, ceci permettant de construire l'arbre maximal. Ensuite, le logiciel *R* permet d'observer la décroissance de l'erreur relative en fonction du critère de complexité :

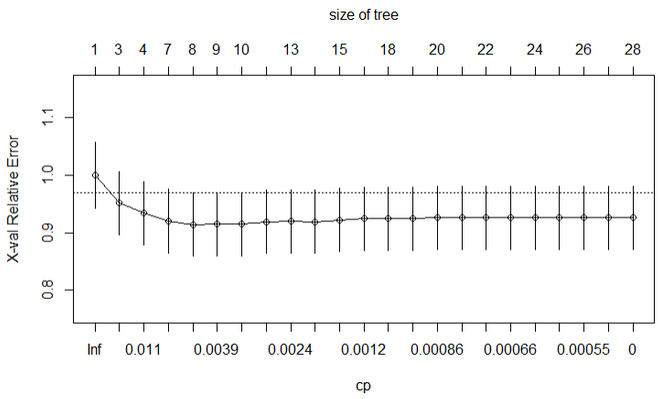


FIGURE 4.6 – Décroissance de l'erreur relative en fonction du critère de complexité

L'axe situé au-dessus du graphique permet également d'indiquer la taille de l'arbre correspondant aux différentes valeurs du critère de complexité.
L'erreur relative s'arrête de décroître assez rapidement du fait de la valeur du paramètre *minbucket*, fixé avant l'élagage. Pour reprendre les termes utilisés lors de la présentation des techniques propres à l'élagage d'un arbre de décision, le fait de fixer en amont la valeur du *minbucket* est une illustration de l'optimisation "a priori".
Au contraire, la démarche d'obtention du critère de complexité est une optimisation "a posteriori". Cette valeur peut être déduite en utilisant la règle de Breiman qui recommande un critère de complexité inférieur à :

$$min(erreur\ relative) + std(erreur\ relative)$$

Dans le cas présent, cette valeur a été fixée à $0,013$.
L'arbre optimal comporte ainsi 3 feuilles.
Néanmoins, seules deux variables interviennent dans un arbre CART comportant 3 feuilles. Il peut donc être intéressant de fixer une autre valeur pour le critère de complexité, inférieure à celle obtenue avec la règle de Breiman afin de visualiser l'action d'autres variables explicatives. L'arbre présenté dans la partie suivante est obtenu avec **cp = 0,002**.

Visualisation de l'arbre optimal

Le principal avantage des arbres de décision est leur côté interprétable, avec une lecture facile de la segmentation effectuée. De ce fait, voici la représentation de l'arbre obtenu avec une valeur du critère de complexité égale à $0,002$:

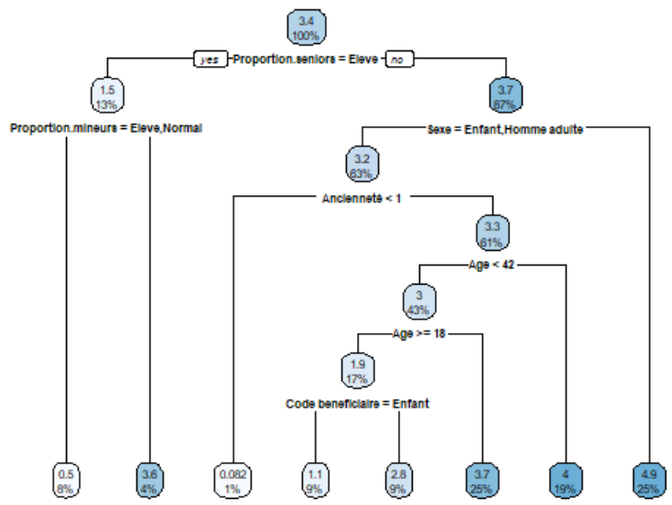


FIGURE 4.7 – Arbre pour la modélisation de la fréquence

Le schéma suivant permet de bien lire et interpréter cet arbre :

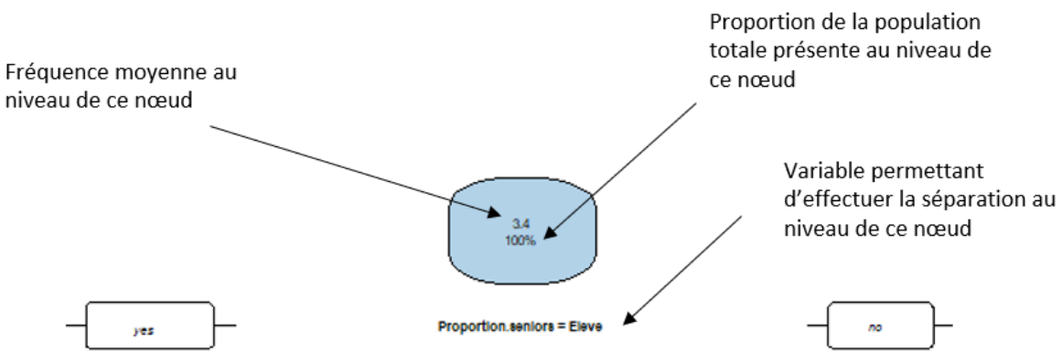


FIGURE 4.8 – Aide à la lecture de l'arbre

Importance des variables

Le graphe suivant permet de visualiser les variables importantes. La valeur présentée correspond un indice d'importance sur une échelle de 0 à 100, plus la valeur attribuée est élevée, plus la variable a été importante dans la construction du modèle. Le schéma suivant permet de présenter les variables importantes :

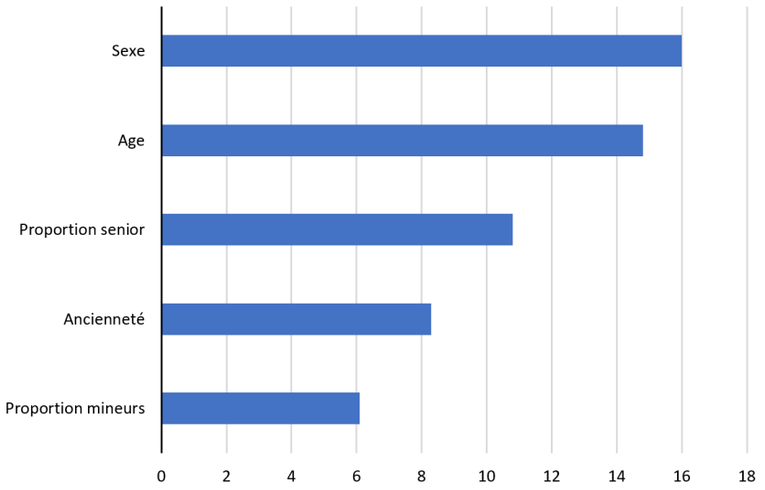


FIGURE 4.9 – Importance des variables dans la modélisation de la fréquence par CART

Certaines variables sont redondantes par rapport au GLM. Néanmoins, la variable "Sexe" n'écrase pas les effets des autres variables dans ce modèle. De plus, une nouvelle variable externe est ici significative : la proportion de mineurs dans la population.

Pertinence du modèle

Comme pour le GLM, la valeur de la MSE sur la base test est présentée :

MSE CART	MSE GLM
23,5	22,6

FIGURE 4.10 – MSE base test pour la modélisation de la fréquence par CART

Le résultat augmente vis-à-vis du GLM : le modèle semble donc moins adapté pour la modélisation de la fréquence. Les modèles d'agrégation d'arbres de décision vont maintenant être implementés afin de voir si la pertinence peut cette fois être augmentée.

4.2.3 *Random Forest*

L'agrégation d'arbres de décision via la méthode *Random Forest* a été effectuée à l'aide de la librairie *randomForest* du logiciel *R*.

Démarche mise en place lors de la modélisation par Random Forest

La fonction *randomForest* du package *R* du même nom offre également de nombreux paramètres sur lesquels jouer afin d'obtenir la meilleure forêt aléatoire possible. Il a été ici fait le choix de s'intéresser à trois paramètres :

- ***nodesize*** : le nombre minimal d'observations observées dans chaque feuille.
- ***ntree*** : le nombre d'arbres comportant la forêt.
- ***mtry*** : le nombre de variables tirées aléatoirement à chaque noeud de chaque arbre.

La démarche mise en place a donc été la suivante :

1. **Fixation du paramètre *ntree* à une valeur élevée** ;
2. **Calcul du paramètre *mtry* par hyper-paramétrage et validation croisée** :
   Pour ceci, la fonction *train* du package *caret* a été utilisée.
3. Une fois le paramètre *mtry* fixé, **calcul du paramètre *ntree* à l'aide de la décroissance de la courbe d'erreur** ;
4. **Importance des variables** :
   La fonction *randomForest* permet de visualiser l'importance des variables selon 2 critères :
   - *IncNodePurity* qui est la mesure de décroissance de la pureté ;
   - *% IncMSE* qui représente l'effet des variables explicatives sur la précision du modèle.

   Dans ce mémoire, la grandeur retenue est la *% IncMSE*. Le calcul est réalisé de la manière suivante :
   Le modèle évalue l'erreur de prédiction avec un positionnement initial des variables explicatives. Les valeurs de ces dernières sont ensuite permutées. L'erreur de prédiction est à nouveau évaluée. Ce calcul est effectué pour les différents arbres de la forêt. Pour chaque variable, les différences entre les erreurs avant et après permutations, ainsi que la somme des carrés des erreurs, sont normalisées (par le nombre d'arbres) et divisées par l'écart-type des différences si cet écart-type est non nul. Il est utile de précisER que le calul de cet erreur de précision, appelée également "erreur OOB" s'effectue sur les échantillons "*Out-Of-Bag*". En effet, chaque arbre étant construit sur un échantillon bootstrap, il n'est pas construit sur l'ensemble de la base d'apprentissage. Pour chaque arbre construit, l'erreur calculée s'effectue sur les observations qui n'ont pas été utilisées lors de la construction de l'arbre. L'erreur finale correspond à la moyenne de ces erreurs.
5. **Test de pertinence** :
   Calcul de la MSE sur la base de test.

Les trois premières étapes permettent de fixer les paramètres du modèle. Dans la suite, ces trois étapes seront donc regroupées en un unique paragraphe.

Fixation des paramètres

Dans un premier temps, le nombre d'arbres utilisé est fixé arbitrairement à 500. Cette valeur va alors être utilisée dans le calcul de la valeur du paramètre *mtry*.

Le paramètre *mtry* a été calculé via un hyper-paramétrage.
Concrètement, une grille de paramètre est définie (ici cette "grille" ne comporte qu'une seule ligne puisqu'un seul paramètre est testé). Ensuite, le modèle est lancé pour chaque combinaison des paramètres. Les paramètres conservés correspondent au modèle présentant les meilleurs résultats selon une métrique donnée (ici la *Root Mean Square Error*). Il est à noter que la fonction *train* du package *caret* propose une option de "contrôle" qui permet d'effectuer une validation croisée (ici sur la base d'un découpage en 5 parties ou *folds*).

La fonction *randomForest* fixe par défaut le nombre de variables sélectionnées à chaque étape comme égal à la racine carrée du nombre de variables explicatives dans le cas d'une régression. Dans le cas présent, le modèle comporte 31 variables explicatives. De ce fait, les valeurs testées sont les suivantes :

Valeurs testées pour le paramètre *mtry*
5, 6, 7, 8, 9 et 10

FIGURE 4.11 – Valeurs testées pour *mtry*

Via donc le critère de la RMSE, **la valeur retenue pour le paramètre *mtry* est 5**.

Maintenant que le paramètre *mtry* est fixé, le nombre d'arbres peut être optimisé. Pour cela, il est intéressant d'observer la courbe de décroissance de l'erreur OOB en fonction du paramètre *ntree* :

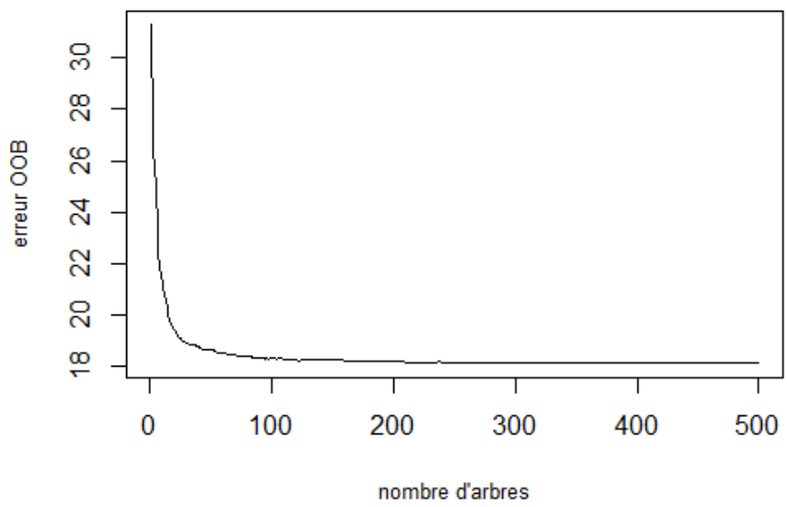


FIGURE 4.12 – *Random Forest : décroissance de l'erreur OOB en fonction du nombre d'arbres*

L'erreur OOB décroît rapidement avant de se stabiliser aux alentours de $\boldsymbol{ntree} = \mathbf{100}$. Cette valeur est donc conservée.

Concernant la valeur de ***nodesize***, après plusieurs essais, la valeur retenue est 5.

Importance des variables

La décroissance moyenne de la précision (*% IncMSE*) est présentée pour les cinq variables les plus importantes dans le graphique suivant :

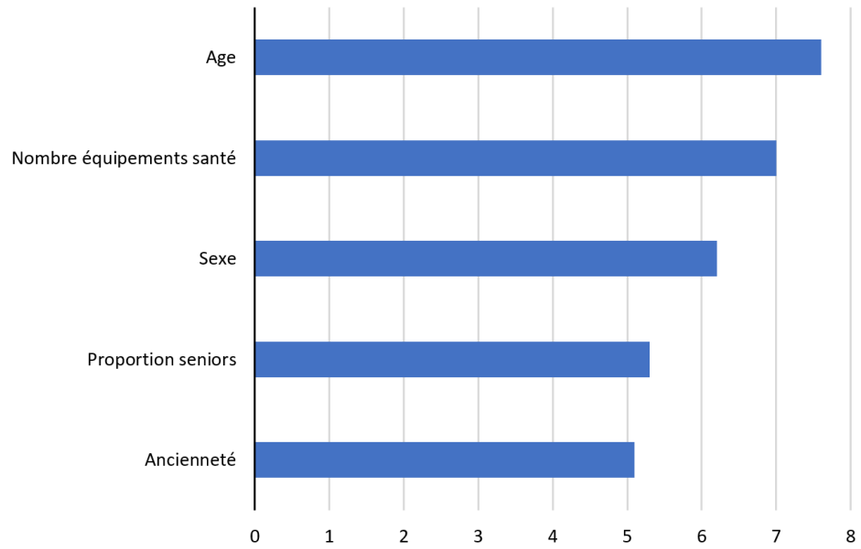


FIGURE 4.13 – Importance des variables dans la modélisation de la fréquence par *Random Forest*

Les variables Âge, Ancienneté et Sexe sont de nouveau considérées comme importantes. Une nouvelle variable externe fait son apparition : le nombre d'équipements de santé.

Pertinence du modèle

De nouveau, la valeur de la MSE sur la base de test est indiquée :

MSE *Random Forest*	MSE GLM
22,1	22,6

FIGURE 4.14 – MSE base test pour la modélisation de la fréquence par *Random Forest*

Cette fois-ci, la valeur augmente par rapport au GLM : le modèle devient plus pertinent.
L'algorithme du *Random Forest* construit les arbres indépendemment avant de les agréger. Le modèle de l'*eXtreme Gradient Boosting* permet lui de construire les arbres les uns après les autres, l'arbre suivant apprenant des précédents. Cette modification entraîne-t-elle une amélioration des performances ?

4.2.4 *eXtreme Gradient Boosting* (XGBoost)

L'agrégation d'arbres de décision via la méthode XGBoost a été effectuée à l'aide de la librairie *xgboost* du logiciel *R*.

Démarche mise en place lors de la modélisation par XGBoost

Du fait de sa complexité, l'algorithme XGBoost possède de multiples paramètres à calibrer. Seuls ceux utilisés au cours du mémoire seront présentés ici :

- ***nrounds*** : le nombre d'arbres implémentés ;
- $\boldsymbol{\eta}$: le taux d'apprentissage. C'est le taux auquel le modèle apprend à chaque étape de l'algorithme. Sa valeur par défaut est $0,3$ et ce paramètre doit être compris entre 0 et 1.
- ***maxdepth*** : la profondeur de l'arbre. L'augmentation de ce paramètre permet de prendre en considération les effets, parfois combinés, de plus de variables, mais augmente aussi le risque de sur-apprentissage.
- $\boldsymbol{\gamma}$: la régularité du modèle. Plus sa valeur sera grande, plus le modèle sera lissé. Si la séparation d'un noeud en deux branches réduit la fonction de perte d'une quantité inférieure à γ, alors le noeud n'est pas développé.
- ***colsample bytree*** : la proportion de variables prise en compte dans la construction de chaque arbre ;
- ***subsample*** : le pourcentage d'observations prises en compte dans chaque arbre ;
- ***min child weight*** : ce paramètre correspond au nombre minimum d'observations présentes dans chaque noeud pour poursuivre le développement de l'arbre.

La démarche mise en place a été globalement identique à celle des forêts aléatoires si ce n'est la première étape :

1. **Dummisation de la base** ;
2. **Calibration des paramètres par hyper-paramétrage et validation croisée** ;
3. **Importance des variables** ;
4. **Test de pertinence**.

Dummisation de la base de données

Avant toute chose, il est nécessaire de rendre binaires les variables catégorielles : cette phase est indispensable pour l'application de l'algorithme XGBoost. Un premier traitement est effectué en ce sens sur la base de données.

Calibrage des paramètres

Dans un premier temps, le paramètre *nround* a été fixé à 100. De nouveau à l'aide de la fonction *train* de la librairie *caret*, plusieurs combinaisons des paramètres cités précédemment ont été testées :

Valeurs testées	
η	0,0001 ; 0,001 ; 0,01 ; 0,1 ; 0,3
max depth	2 ; 4 ; 6 ; 8
γ	0 ; 1
colsample bytree	0,8
subsample	0,75
min child weight	0

FIGURE 4.15 – Valeurs testées pour les paramètres du XGBoost

Par ailleurs, la validation croisée est réalisée sur la base d'un découpage en 5 parties. Les graphiques suivants présentent les *Root Mean Square Errors* (RMSE) suivant les différentes valeurs des paramètres testés :

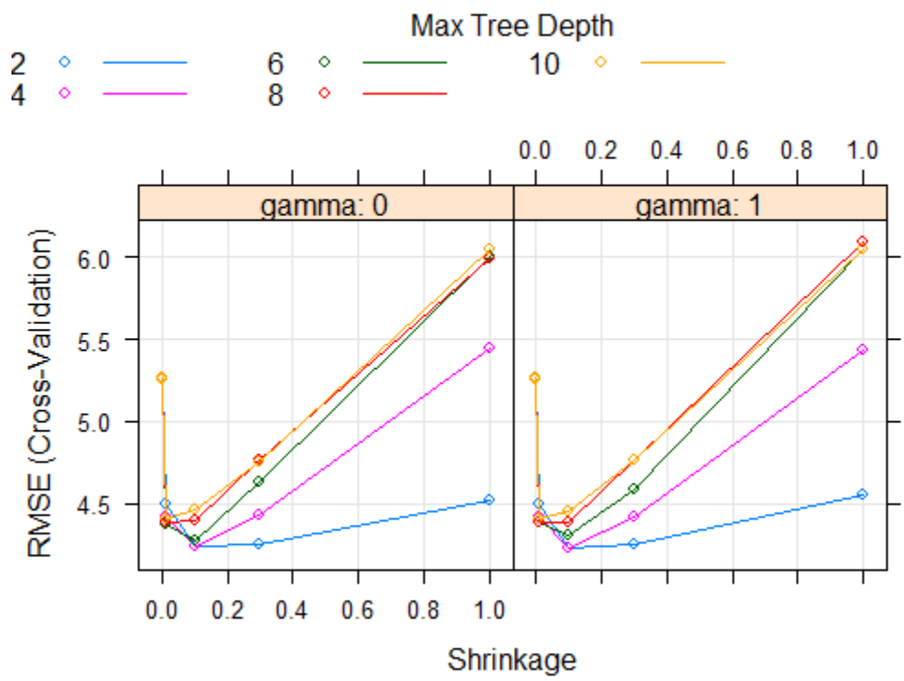


FIGURE 4.16 – RMSE en fonction des différents paramètres XGBoost

Les paramètres qui minimisent la valeur de la RMSE sont les suivants :

Valeurs optimales	
η	0,1
max depth	2
γ	1
colsample bytree	0,8
subsample	0,75
min child weight	0

FIGURE 4.17 – Paramètres optimaux du XGBoost

De la même façon que pour les forêts aléatoires, le nombre d'itérations (paramètre *nrounds*) est fixé dans un second temps à l'aide d'une deuxième validation croisée. La métrique ici retenue est la RMSE sur la base de test.
Le graphique suivant permet de visualiser cette évolution :

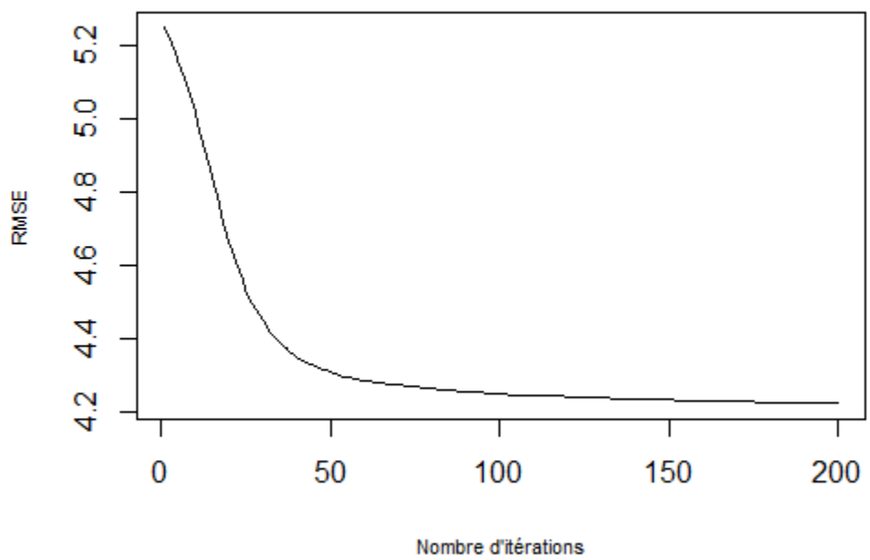


FIGURE 4.18 – Évolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations

La valeur de la RMSE ne cesse de diminuer. Néanmoins, cette diminution est d'abord très forte au départ et a tendance à ralentir. **La valeur du paramètre *nrounds* est donc fixé à 100 afin d'éviter le sur-apprentissage.**

Importance des variables

Dans le cas de l'algorithme XGBoost, l'importance des variables est mesurée en terme de "gain". Plus précisément, l'idée sous-jacente est de mesurer l'augmentation de la précision du modèle avant et après l'ajout d'une division dans le modèle. Le résultat final de ce gain correspond au gain moyen pour chaque variable explicative.
Le tableau suivant permet de visualiser de nouveau les cinq variables les plus importantes :

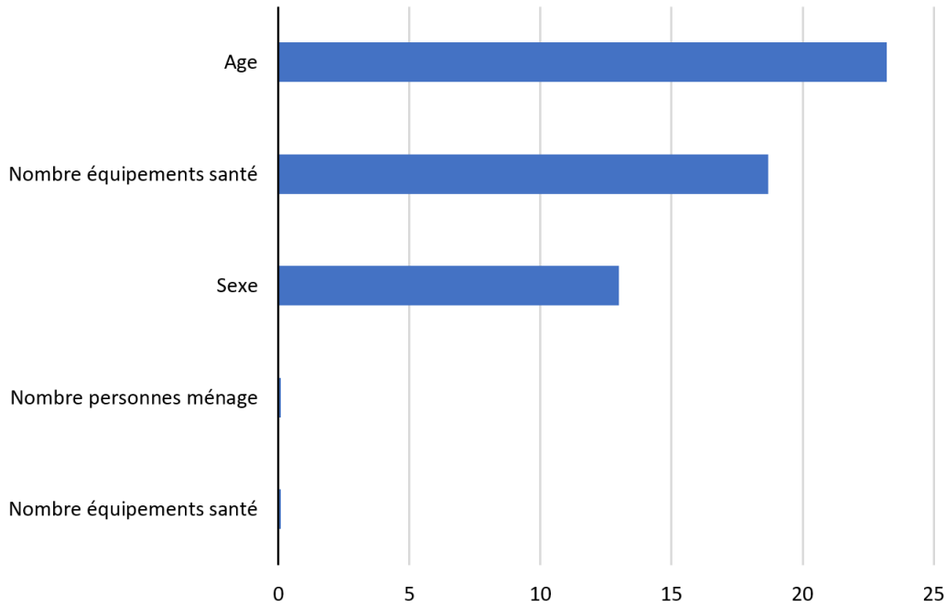


FIGURE 4.19 – Importance des variables dans la modélisation de la fréquence par XGBoost

Le modèle XGBoost retient globalement 3 variables dont une externe : le nombre d'équipements de santé dans la commune.

Pertinence du modèle

Comme pour les autres modèles, la MSE calculée sur la base de test est présentée dans le tableau ci-dessous :

MSE *XGBoost*	MSE GLM
23,2	22,6

FIGURE 4.20 – MSE base test pour la modélisation de la fréquence par XGBoost

Le modèle XGBoost n'est donc pas le modèle le plus pertinent pour la modélisation de la fréquence dans le poste "Consultations Visites".

4.2.5 Conclusion

Pertinence des modèles

Le tableau suivant permet de récapituler les valeurs de la MSE sur la base de test :

Modèle	MSE base test
GLM	22,6
CART	23,5
Random Forest	22,1
XGB	23,2

FIGURE 4.21 – Récapitulatif des tests de pertinence pour la fréquence

Le meilleur modèle pour la modélisation de la fréquence est donc le *Random Forest*.

Importance des variables

Le tableau suivant permet de résumer, pour chaque modèle testé, les variables ressortant comme importantes dans la modélisation de la fréquence :

GLM	CART	*Random Forest*	XGBoost
Sexe	Sexe	Age	Age
Proportion seniors	Age	Nombre équipements santé	Nombre équipements santé
Ancienneté	Proportion seniors	Sexe	Sexe
Proportion propriétaires	Ancienneté	Proportion seniors	Nombre personnes ménage
Age	Nombre équipements santé	Ancienneté	Ancienneté

FIGURE 4.22 – Tableau récapitulant les variables les plus importantes dans la modélisation de la fréquence

Les variables internes ressortant comme significatives pour la fréquence de consommation du poste "Consultations Visites" sont similaires pour l'ensemble des modèles testés. Ces variables sont les suivantes :

- L'âge ;
- L'ancienneté ;
- Le sexe.

L'information importante est le fait que plusieurs variables externes [a] sont considérées comme significatives dans chaque modèle.
Les variables externes les plus redondantes sont la proportion de seniors dans la commune et le nombre d'équipements de santé dans la commune.

a. case bleutée dans le tableau précédent

4.3 Modélisation des dépenses moyennes au niveau du poste "Consultations Visites"

La modélisation de la consommation moyenne nécessite de ne s'intéresser qu'aux individus ayant consommés dans l'année. De ce fait, le nombre d'observations à disposition est moindre que lors de la modélisation de la fréquence. De la même façon que dans la partie précédente, les résultats seront détaillés selon les différents modèles testés.

4.3.1 GLM

La démarche de mise en place du GLM est similaire à celle implémentée lors de la modélisation de la fréquence. Néanmoins, **un test d'adéquation a tout d'abord été réalisé afin de choisir la loi de distribution de la charge moyenne utilisée lors du GLM**.
La distribution de la charge moyenne a donc été comparée aux lois gamma et log-normale. Les *p-value* associées à ces lois étant toutes les deux significatives, le test ne nous permet pas de trancher.
Le choix a donc été d'opter pour **une loi gamma avec un lien log**.

Traitements des variables continues

Les traitements réalisés sont identiques à la partie précédente.

Choix de la loi et de la fonction de lien

Comme indiqué dans l'encadré précédent, une loi gamma avec une fonction de lien log a été retenue.

Etude des corrélations

Les corrélations entre les variables sont globalement similaires à celles observées lors de la modélisation de la fréquence. La variable "Code bénéficiaire" a donc également été supprimée.

Sélection des variables

L'algorithme conserve 16 variables dont 11 d'origine externe.

Néanmoins, le test de significativité ne retient que cinq variables qui sont les suivantes :

Variable	Degrés de liberté	Déviance	AIC	*p-value*
Proportion famille	1	282,26	23 851	<2 e-16
Zone	3	282,90	23 852	4,5 e-4
Ancienneté	2	282,79	23 853	2,0 e-10
Sexe	2	283,24	23 856	<2 e-16
Loyer moyen	2	283,87	23 861	<2 e-16

FIGURE 4.23 – Sélection des variables par la méthode descendante pour la modélisation des dépenses moyennes

Importance des variables

L'importance des variables peut être déduite de l'étape précédente par rapport à la diminution de l'AIC :

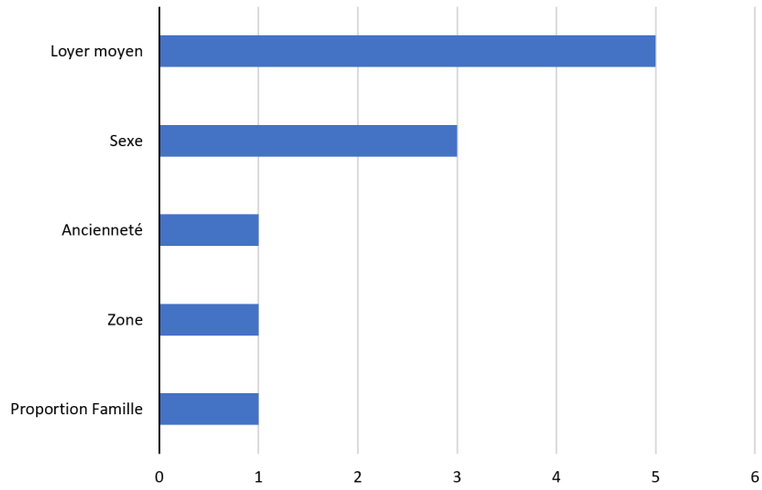


FIGURE 4.24 – Importance des variables dans la modélisation des dépenses moyennes par GLM

Les variables "Loyer moyen" et "Zone" sont donc pertinentes ici. Ces variables représentant le lieu d'habitation des assurés, le fait que ces variables externes ressortent comme importantes peut être interprété de la façon suivante : l'accès aux soins (ici les consultations chez les généralistes et spécialistes) ne coûte pas le même prix selon la zone géographique.

Pertinence du modèle

Comme indiqué précédemment, le critère de pertinence retenu pour la modélisation des dépenses moyennes est la MSE.
La valeur calculée pour le GLM est la suivante :

MSE CART	MSE GLM
59,6	82,9

FIGURE 4.25 – MSE base test pour la modélisation des dépenses moyennes par le GLM

Cette valeur sera également comparée à celles évaluées par les autres modèles testés.

Analyse des résultats

Les coefficients de la régression sont les suivants :

	Estimate	*Std Error*	*Z-value*	*p-value*
(Intercept)	3,21	0,02	102,15	< 2 e -16
Sexe Enfant	0,02	0,02	1,08	0,28
Sexe Femme adulte	0,08	0,02	4,57	< 2 e -16
Proportion Famille supérieure à 70 %	0,04	0,02	2,62	< 2 e -16
Ancienneté inférieure à 5 ans	0,12	0,02	6,87	< 2 e -16
Ancienneté supérieure à 15 ans	0,02	0,02	0,93	0,35
Loyer moyen faible	-0,17	0,02	-8,40	< 2 e -16
Loyer moyen normal	-0,20	0,02	-11,71	< 2 e -16
Zone Très urbaine	0,26	0,03	9,21	< 2 e -16
Zone Urbaine	0,03	0,02	1,73	0,08

FIGURE 4.26 – Coefficients de régression des dépenses moyennes

Ici, l'individu de référence correspond à un assuré disposant des caractéristiques suivantes :

— Sexe Homme adulte ;
— Proportion de familles dans la commune inférieure à 70 % ;
— Ancienneté comprise entre 5 et 10 ans ;
— Loyer moyen élevé par rapport à la moyenne nationale ;
— Zone rurale.

4.3.2 CART

La démarche de mise en place de la méthode CART pour la modélisation de la consommation totale est en tout point identique à celle utilisée pour la modélisation de la fréquence.

Construction de l'arbre maximal

L'arbre maximal a de nouveau été construit dans un premier temps.

Elagage de l'arbre maximal

Les trois paramètres utiles à l'obtention de l'arbre optimal sont de nouveau ***xval, minbucket et cp***. Les valeurs de ces paramètres sont les suivantes :

— les valeurs des paramètres ***xval*** et ***minbucket*** ont de nouveau été fixées à **5** et **100** ;
— Selon la règle de Breiman, la valeur du critère de complexité ***cp*** vaut $\mathbf{0,07}$.

Visualisation de l'arbre optimal

Cependant, **l'arbre optimal construit avec la valeur précédente du critère de complexité ne comporte que 2 feuilles**.

Il peut donc être intéressant de fixer une autre valeur pour le critère de complexité, inférieure à celle obtenue avec la règle de Breiman afin de visualiser l'action d'autres variables explicatives. Voici par exemple l'arbre obtenu avec **cp = 0,002** :

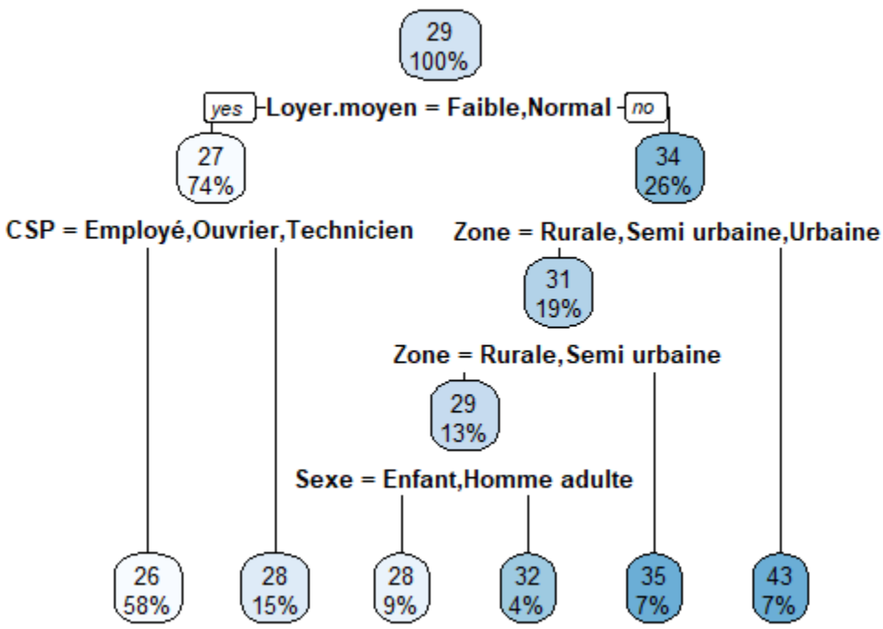


FIGURE 4.27 – Arbre obtenu après variation du critère de complexité pour la modélisation des dépenses moyennes

L'arbre obtenu comporte 6 feuilles et d'autres variables rentrent ici en compte comme le sexe ou le lieu d'habitation de l'adhérent. Tous les résultats présentés dans les parties suivantes concernent cet arbre.

Importance des variables

Le graphe suivant permet de visualiser les variables importantes :

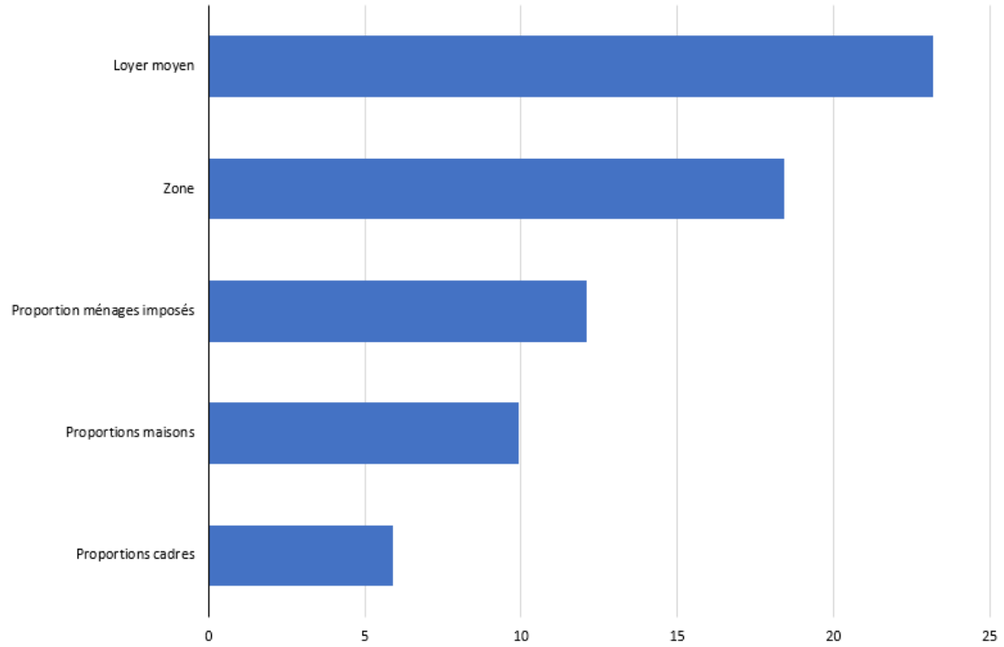


FIGURE 4.28 – Importance des variables dans la modélisation des dépenses moyennes par CART

Les cinq variables les plus importantes sont d'origine externe. Elles font également toutes référence à la "richesse" de l'environnement de l'individu.

Pertinence du modèle

La MSE calculée pour ce modèle est égale à :

MSE CART	MSE GLM
59,6	82,9

FIGURE 4.29 – MSE base test pour la modélisation des dépenses moyennes par CART

Cette valeur est nettement inférieure à celle calculée par le GLM. Le modèle CART est donc plus pertinent en ce qui concerne la modélisation de la charge moyenne au niveau du poste "Consultations Visites".

4.3.3 *Random Forest*

La démarche de mise en place de la méthode *Random Forest* pour la modélisation des dépenses moyennes est en tout point identique à celle utilisée pour la modélisation de la fréquence.

Fixation des paramètres

Les trois paramètres calibrés durant ce mémoire sont les suivant :

- ***ntree*** ;
- ***mtry*** ;
- ***nodesize***.

Dans un premier temps, le nombre d'arbres utilisé est une nouvelle fois fixé arbitrairement à 500.

Via la fonction *train* du package *caret*, plusieurs valeurs du paramètre *mtry* sont testées :

Valeurs testées pour le paramètre *mtry*
5, 6, 7, 8, 9 et 10

FIGURE 4.30 – Valeurs testées pour *mtry*

Via de nouveau le critère de la RMSE, **la valeur retenue pour le paramètre *mtry* est 5**.

Maintenant que le paramètre *mtry* est fixé, le nombre d'arbres peut être optimisé par observation de la courbe de décroissance de l'erreur OOB en fonction du paramètre *ntree*.
De la même manière que pour la modélisation de la probabilité de consommer, l'erreur OOB décroît rapidement avant de se stabiliser aux alentours de $\boldsymbol{ntree = 100}$. Cette valeur est donc conservée.

Concernant la valeur de ***nodesize***, la valeur retenue est 10.

Importance des variables

Les cinq variables les plus importantes sont présentées pour la décroissance moyenne de la précision (*% IncMSE*) :

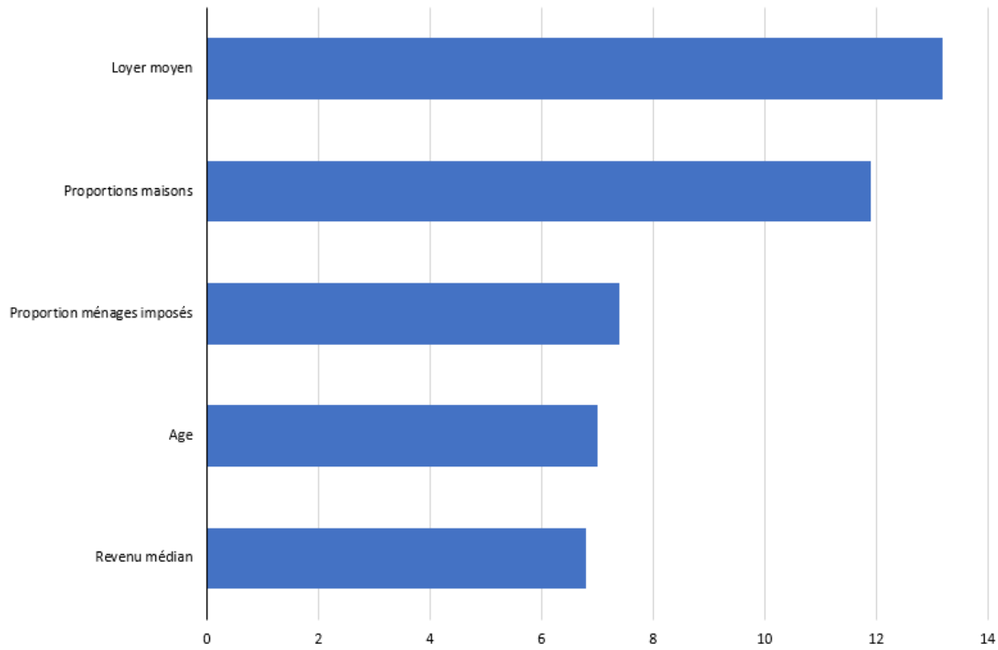


FIGURE 4.31 – Importance des variables dans la modélisation des dépenses moyennes par *Random Forest*

Les variables considérées comme significatives sont globalement identiques aux variables ressorties par le modèle CART. L'environnement de l'individu joue un grand rôle dans l'explication des dépenses moyennes.

Pertinence du modèle

La MSE calculée pour le modèle est la suivante :

MSE *Random Forest*	MSE GLM
56,3	82,9

FIGURE 4.32 – MSE base test pour la modélisation des dépenses moyennes par le *Random Forest*

4.3.4 XGBoost

La démarche de mise en place de la méthode XGBoost pour la modélisation des dépenses moyennes est en tout point identique à celle utilisée pour la modélisation de la fréquence.

Dummisation de la base de données

Cette étape est répétée pour la modélisation des frais réels moyens sur le poste "Consultations Visites".

Calibrage des paramètres

Les mêmes combinaisons de paramètres que pour la modélisation de la probabilité de consommer sont testées, à savoir :

Valeurs testées	
η	0,0001 ; 0,001 ; 0,01 ; 0,1 ; 0,3
max depth	2 ; 4 ; 6 ; 8
γ	0 ; 1
colsample bytree	0,8
subsample	0,75
min child weight	0

FIGURE 4.33 – Valeurs testées pour les paramètres du XGBoost

Par ailleurs, la validation croisée est réalisée sur la base d'un découpage en 5 parties. Selon donc le critère de la RMSE, les paramètres optimaux sont les suivants :

Valeurs optimales	
η	0,1
max depth	2
γ	0
colsample bytree	0,8
subsample	0,75
min child weight	0

FIGURE 4.34 – Paramètres optimaux du XGBoost

De nouveau, le nombre d'itération (paramètre *nrounds*) est fixé dans un second temps à l'aide d'une deuxième validation croisée. La métrique ici retenue est la mesure la RMSE sur la base d'apprentissage.

Le graphique suivant permet de visualiser cette évolution :

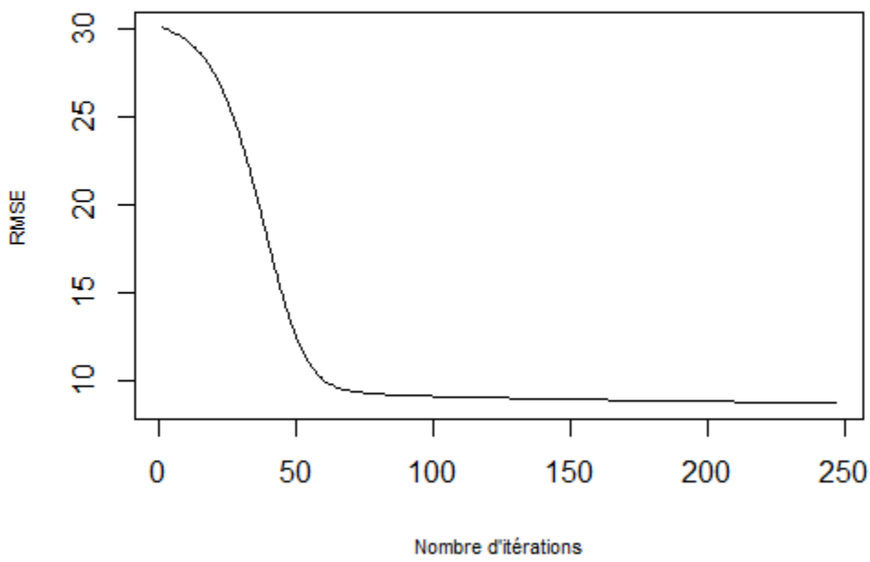


FIGURE 4.35 – Évolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations

La valeur de la RMSE diminue continuellement. Néanmoins, cette valeur a tendance à se stabiliser assez rapidement. **La valeur du paramètre *nrounds* est donc fixé à 100 afin d'éviter le sur-apprentissage.**

Importance des variables

Le tableau suivant récapitule les cinq variables les plus importantes selon l'algorithme XGBoost :

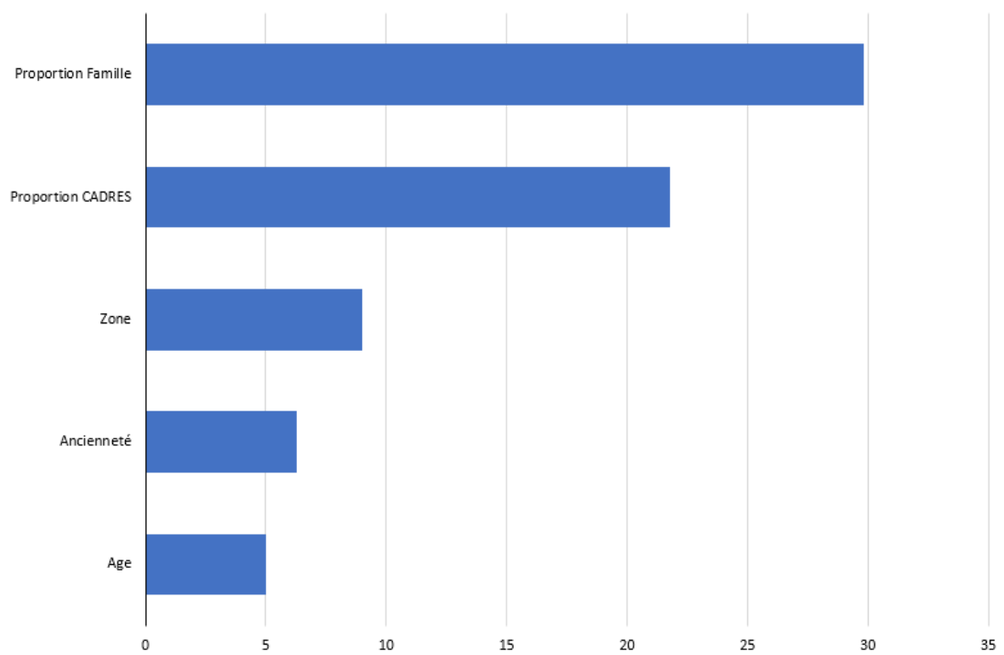


FIGURE 4.36 – Importance des variables dans la modélisation des dépenses moyennes par XGBoost

De nouveau, des variables externes liées au niveau de vie de l'individu (proportion cadres) ressortent comme significatives.

Pertinence du modèle

La MSE sur la base test s'élève à :

MSE *XGBoost*	MSE GLM
54,5	82,9

FIGURE 4.37 – MSE base test pour la modélisation de la charge moyenne par XGBoost

Le modèle XGBoost est donc le plus pertinent en ce qui concerne la modélisation des dépenses moyennes.

4.3.5 Conclusion

Pertinence du modèle

Le tableau suivant permet de récapituler l'ensemble des valeurs des MSE pour les différents modèles testés :

Modèles	MSE
GLM	82,9
CART	59,6
Random Forest	56,3
XGBoost	54,5

FIGURE 4.38 – Récapitulatif des tests de pertinence pour la modélisation des dépenses moyennes

Le modèle le plus pertinent au sens de la MSE est donc le XGBoost. On remarque également que le GLM est bien moins approprié que les autres modèles dans le cadre de la modélisation du coût moyen.

Importance des variables

Le tableau suivant permet de résumer, pour chaque modèle testé, les variables ressortant comme importantes dans la modélisation de la fréquence :

GLM	CART	*Random Forest*	XGBoost
Proportion famille	Loyer moyen	Loyer moyen	Proportion famille
Zone	Zone	Proportion maisons	Proportion cadres
Ancienneté	Proportion ménages imposés	Proportion ménages imposés	Zone
Sexe	Proportion maisons	Age	Ancienneté
Loyer moyen	Proportion cadres	Revenu médian	Age

FIGURE 4.39 – Tableau récapitulant les variables les plus importantes dans la modélisation des dépenses moyennes

De nouveau, plusieurs variables externes sont considérées comme significatives dans chaque modèle.
Ces variables externes sont la plupart du temps liées à l'environnement économique et fiscal de l'individu. En effet, des variables comme la proportion de cadres, le revenu médian ou la proportion de ménages imposés sont considérées comme importantes dans la plupart des modèles.
L'âge, le sexe et l'ancienneté sont également considérés comme des variables significatives pour la quasi totalité des modèles.

4.4 Résultats généraux pour la famille "Consultations Visites"

4.4.1 Récapitulatif des différents tests de pertinence

Le tableau suivant permet de résumer les résultats des tests de pertinence effectués lors des modélisation de la fréquence et des dépenses moyennes :

Modèle	Fréquence	Dépenses moyennes
	MSE base test	MSE base test
GLM	22,6	82,9
CART	23,5	59,6
Random Forest	22,1	56,3
XGBoost	23,2	54,5

FIGURE 4.40 – Récapitulatif des tests de pertinence

Le code couleur utilisé est le suivant :

	Modèle le plus pertinent
	2ème modèle plus pertinent

La première conclusion que l'on peut tirer est le fait que les tests de pertinence réalisés lors des modélisations de la fréquence et des dépenses moyennes sélectionnent deux modèles différents.
De plus, il apparaît que le modèle *Random Forest* se classe à chaque fois parmi les deux modèles les plus pertinents.

4.4.2 Comparaison des différents modèles testés

Jusqu'à présent, les modèles implémentés n'ont été comparés que via les tests de pertinence. Néanmoins, une multitude d'autres critères rentrent en compte dans le choix du modèle à retenir. Le tableau suivant permet de résumer les avantages et inconvénients de modèles testés :

Algorithme	Explication de l'algorithme	Vitesse d'apprentissage	Interprétabilité des résultats	Pouvoir prédictif
GLM	★★★	★★★★★	★★★★	★
CART	★★★★	★★★★★	★★★★★	★★
Random Forest	★★	★	★★	★★★
XGBoost	★	★	★★	★★★★

FIGURE 4.41 – Synthèse des différents modèles

Une des première chose à prendre en compte lors de la mise en place d'un modèle est **la capacité de l'utilisateur à maîtriser l'algorithme** derrière le modèle. Au vue des présentations théoriques des modèles effectuées au cours du troisième chapitre du mémoire, l'algorithme CART semble être le plus à même d'être expliqué de manière claire et concise.

Concernant la **vitesse d'apprentissage** du modèle, le *Random Forest* et le *XGBoost* étant des modèles agrégés, celle-ci s'en retrouve forcément ralenti. Néanmoins, la vitesse d'apprentissage n'est pas le seul élément à prendre en compte lors de la comparaison du temps de mise en place de chaque modèle. La calibration des paramètres et autres traitements spécifiques prennent un certain temps.
Voici donc les principaux enseignements tirés à ce sujet au cours de la réalisation de ce mémoire :

- Le modèle GLM est quasi-instantané en terme d'exécution. Néanmoins, l'étude des variables en amont ainsi que le traitement de ces variables afin d'obtenir le modèle optimal peut demander un certain temps.
- Les arbres CART sont les modèles les plus rapide à mettre en place. Que ce soit pour les étapes de calibration ou d'exécution, une durée assez restreinte passée sur ces étapes est suffisante afin d'obtenir des premiers résultats.
- De la même manière que pour les arbres CART, l'étape de calibration des modèles *Random Forest* est assez rapide. Néanmoins, l'exécution de ces modèles, même sur une base de quelques milliers de lignes comme celle utilisée au cours de cette étude, demande plusieurs minutes.
- Devant le nombre important de paramètres à disposition, l'étape de calibration des modèles XGBoost demande un temps conséquent. De même, la phase d'exécution, comme tout modèle agrégé, est également importante.

L'interprétabilité des résultats est également moins aisée avec des modèles agrégés puisque le résultat obtenu est une agrégation d'estimateurs et non plus un simple estimateur. Les modèles CART, avec une lecture simplifiée de la segmentation effectuée grâce à la visualisation de l'arbre, possède les résultats les plus faciles à interpréter.

Enfin, le **pouvoir prédictif** des modèles est évidemment à prendre en compte lors de la comparaison de différents modèles. Dans le cadre de ce mémoire, les modèles ayant le plus grand pouvoir prédictif sont, sans surprise, les modèles agrégés. Ceux-ci apprennent de façon optimisée des données tout en évitant le surapprentissage via leurs multitudes de paramètres.

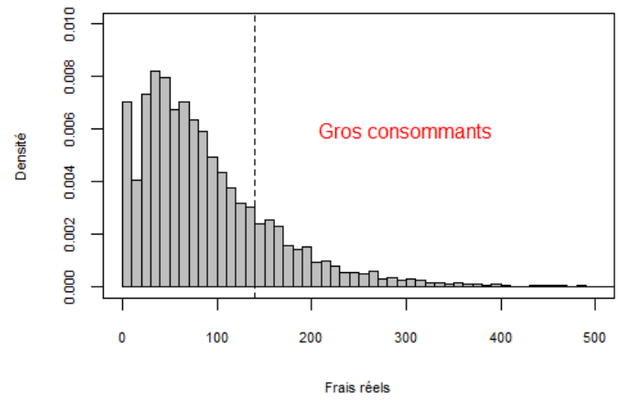
Dans le cadre de ce mémoire, il a été fait le choix de s'intéresser à la fois à la complexité des modèles et à leur capacité prédictive. A la vue des résultats, **c'est le modèle *Random Forest* qui a été conservé.**
Il est à noter que ce choix est propre à la problématique de ce mémoire et que d'autres modèles auraient pu être choisis dans d'autres cas.
Par conséquent, la modélisation de la consommation au sein des autres postes se fera exclusivement via le modèle *Random Forest*.

4.5 Présentation des résultats pour les autres postes de soins

Maintenant que les sorties des différents modèles ont été analysées via le poste "Consultations Visites", seules l'importance des variables et ainsi que la représentation graphique de la consommation seront présentées pour les autres postes. **Ces postes ont été modélisés à l'aide de modèles *Random Forest*.**

Le poste Pharmacie

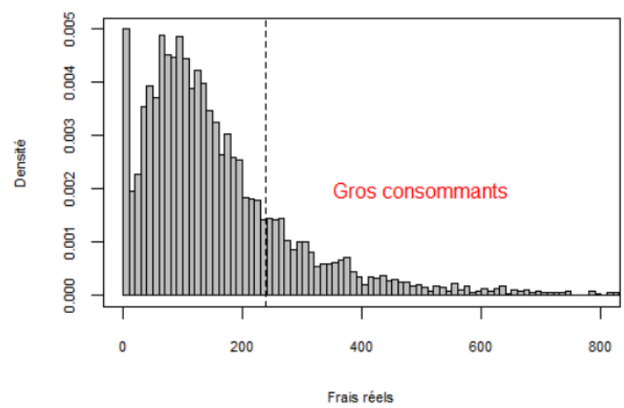
Pharmacie	
Fréquence	**Dépenses moyennes**
Ancienneté	Nombre médecins
Proportion propriétaires	Nombre pharmacies
Proportion maisons	Age
Revenu médian	Proportion chômeurs
Loyer moyen	Sexe



Un peu au contraire du poste "Consultations Visites", ce sont les variables liées à l'environnement économique et à la richesse des individus (proportion propriétaires, loyer moyen, revenu médian, ...) qui sont importantes dans la modélisation de la fréquence pour le poste "Pharmacie". Ceci peut s'expliquer par le fait les personnes plus aisées peuvent se rendre plus facilement en pharmacie pour acheter des médicaments sans ordonnance.
En ce qui concerne les variables importantes pour les dépenses moyennes, le fait que les variables "Nombre de médecins" et "Nombre de pharmacies" soient significatives peut vouloir dire que la concurrence peut jouer dans les prix pratiqués par les pharmaciens.

Le poste Soins courants

Soins courants	
Fréquence	**Dépenses moyennes**
Ancienneté	Nombre équipements santé
Proportions seniors	Age
Nombre médecins	Proportion maisons
Proportion mineurs	Zone
Nombre équipements santé	Loyer moyen

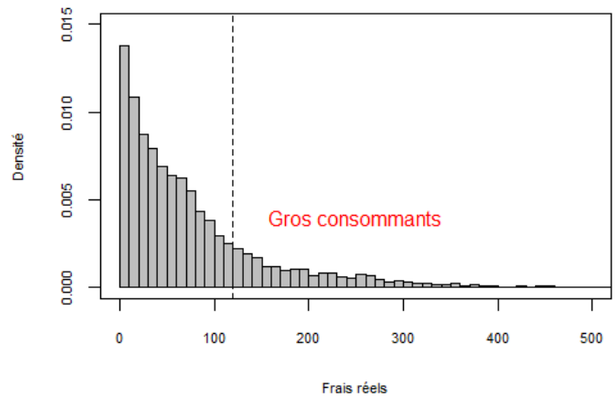


Deux variables externes concernant la répartition démographique de la population (proportion seniors et proportion mineurs) sont significatives dans la modélisation de la fréquence. Ceci est logique puisque ce poste regroupe quelques activités médicales spécifiques aux enfants comme l'orthophonie ou aux seniors comme les soins infirmiers [a].
Le nombre d'équipements de santé intervient à la fois dans l'explication de la fréquence et des dépenses moyennes au sein du poste "Soins courants". La présence d'équipements de santé adaptés peut donc inciter les individus à avoir recours à ce type d'actes médicaux.

a. C.f. Figure 2.7 pour la liste exhaustive.

Le poste Optique

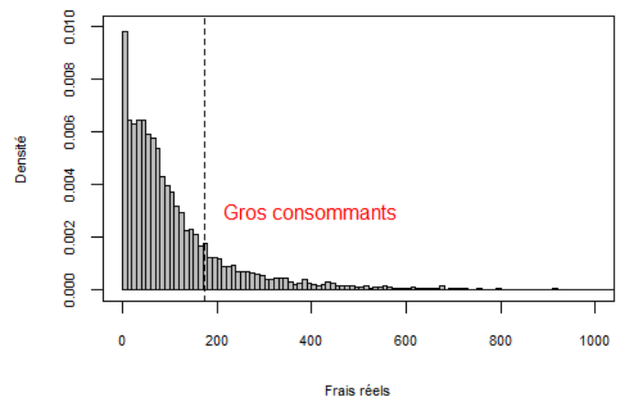
Optique	
Fréquence	**Dépenses moyennes**
Age	Age
Sexe	Sexe
Proportion maisons	Ancienneté
Nombre équipements santé	Nombre équipements santé
CSP	Zone



Que ce soit lors de la modélisation de la fréquence ou des dépenses moyennes pour le poste "Optique", l'âge apparaît comme un critère déterminant. Cela n'est évident pas étonnant : l'optique est prépondérant durant l'enfance et l'adolescence lorsque les problèmes de vue sont diagnostiqués. Concernant la modélisation des dépenses moyennes, on remarque de nouveau la présence d'une variable externe reliée à la l'environnement de l'assuré : son lieu d'habitation identifiée par la variable "Zone".

Le poste Dentaire

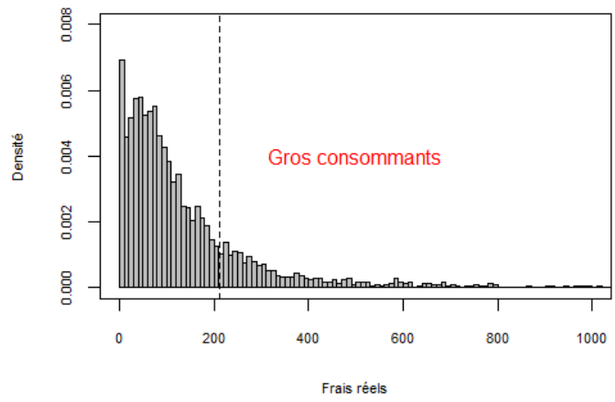
Dentaire	
Fréquence	**Dépenses moyennes**
Age	Nombre dentistes
CSP	Nombre équipements santé
Nombre dentistes	Proportion ménages imposés
Nombre équipements santé	Ancienneté
Proportion activité	CSP



De la même façon que pour le poste Optique, l'âge est très prépondérant dans l'explication de la consommation dentaire. En effet, les besoins en dentaire ne sont pas les mêmes à l'adolescence (orthodontie) que vers 30 ou 40 ans où les soins dentaires forment la majorité des actes au sein du poste "Dentaire". Les présences des variables externes "Nombre dentistes" et "Nombre équipements santé" montrent que la consommation dentaire pourrait dépendre de l'équipement sanitaire de la commune.

Le poste Hospitalisation

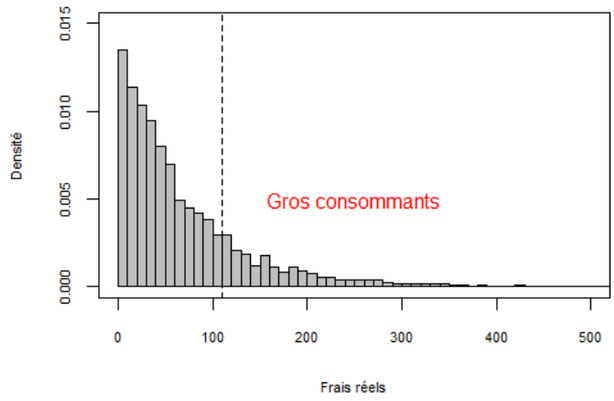
Hospitalisation	
Fréquence	**Dépenses moyennes**
Age	Age
Sexe	Ancienneté
Nombre pharmaciens	Sexe
Nombre médecins	CSP
Nombre dentistes	Loyer moyen



La variable interne "Âge" est prépondérante dans les modélisations de la fréquence et des dépenses moyennes. Mis à part la variable "Loyer moyen" dans la modélisation des dépenses moyennes, la présence des autres variables externes sont difficilement explicables.

Le poste Divers

Divers	
Fréquence	**Dépenses moyennes**
Age	Age
Proportions seniors	Ancienneté
Ancienneté	Proportion seniors
Proportion actifs	Proportion maisons
Sexe	Sexe



Ce poste regroupe principalement les prothèses auditives. De fait, les personnes plus âgées sont concernées par ce type de consommation médicale. Par conséquent, l'âge joue un rôle important dans l'explication de la consommation au sein du poste "Divers". La présence de la variable "Proportion seniors" vient appuyer cette dernière remarque.

4.6 Focus sur les gros consommants

La modélisation des gros consommants a été réalisée de la manière suivante :

- Création d'une variable binaire pour la variable à expliquer :
  - 1 pour les 20 % des individus qui consomment le plus ;
  - 0 sinon.
- Modélisation des gros consommants à l'aide de modèles *Random Forest*.

Ici, le problème est bien un problème de classification : on cherche à savoir si oui ou non un individu appartient à la classe des gros consommateurs. De ce fait, le critère de pertinence retenu a été celui de l'AUC. Les modèles testés présentent un AUC aux alentours de 70 %. Ceux-ci peuvent bien entendu être améliorés mais permettent tout de même de présenter de premiers résultats sur l'étude des gros consommants.
Les variables présentées sont les plus contributives à la classification des gros consommants selon le modèle *Random Forest*.

Les résultats de l'étude des gros consommants présentés concernent uniquement les postes Optique et Dentaire. En effet, ces deux postes présentent plusieurs intérêts. Tout d'abord, ceux-ci sont au coeur de l'actualité (projet de zéro reste à charge en optique et pour les prothèses dentaires notamment). De plus, comme il a été vu au cours de ce mémoire, les parts des organismes complémentaires dans le remboursements des dépenses en dentaire et optique sont particulièrement importantes. Il peut donc être intéressant d'étudier les gros consommants afin de mettre en place une politique de prévention et de ce fait, faire baisser les dépenses associées.

Le but de cette partie serait d'exhiber des variables explicatives du comportement des gros consommants qui n'interviennent pas dans la tarification décrites en amont. De ce fait, les tableaux ci-dessous présentent les variables contributives aux modèles mis en place lors de l'étude des gros consommants et les comparent aux variables considérées comme importantes lors de l'étape de tarification.

Le poste Dentaire

Fréquence		Dépenses moyennes		Gros consommants	
Age	10,2			Age	16,9
CSP	5,4	CSP	7,3		
Nombre dentistes	5,1	Nombre dentistes	13,8		
Nombre équipements santé	4,7	Nombre équipements santé	8,4	Nombre équipements santé	6,3
Proportion activité	4,6				
		Ancienneté	7,4	Ancienneté	7,3
				Sexe	6
		Proportion ménages imposés	7,9		
				Proportion seniors	8,7

FIGURE 4.42 – Récapitulatif des variables importantes pour les différentes études du poste Dentaire

Le tableau précédent permet de mettre en évidence deux variables qui sont considérées comme contributives uniquement dans l'étude relative aux gros consommants : **le sexe et la proportion de seniors dans la commune**.

Le poste Optique

Fréquence		Dépenses moyennes		Gros consommants	
Age	14,3	Age	38,9	Age	19,6
Sexe	7,3	Sexe	7,9	Sexe	11,9
Proportion maison	5,2			Proportion maison	8,2
Nombre équipements santé	4,6	Nombre équipements santé	3		
CSP	4,4		2,8		
		Zone	2,8	Zone	6,4
		Ancienneté	3,3		
				Proportion seniors	5

FIGURE 4.43 – Récapitulatif des variables importantes pour les différentes études du poste Optique

De nouveau, la **variable externe "Proportion de seniors"** est spécifique à l'étude des caractéristiques des gros consommants.

L'étude des caractéristiques des gros consommants permet donc de mettre en évidence certaines variables qui ne ressortaient pas comme significatives dans l'étude de la consommation en santé.

4.7 Reporting santé

Les sorties des modèles *Random Forest* pour l'ensemble des postes de santé ont été présentées dans la partie précédente. La consommation médicale a donc été analysée et les variables contributives à l'explication de cette consommation ont été mises en évidence.
Il s'agit dorénavant de présenter un reporting santé en s'appuyant sur ces résultats.

Les deux pages suivantes s'efforcent de présenter un reporting mêlant statistiques descriptives et sorties des modèles *Data Science*.

La première page est entièrement constituée de statistiques descriptives. En haut, une vue d'ensemble est présentée, ce qui permet d'introduire quelques informations démographiques et financières sur le portefeuille. Dans un second temps, la pyramide des âges donne une idée plus précise de la répartition du portefeuille. Enfin, la répartition géographique des individus ainsi que l'évolution de la consommation médicale totale en fonction de l'âge est rappelée afin de compléter la vue globale du portefeuille.

La page suivante est composée d'idées potentielles de présentation des sorties des modèles *Data Science*.

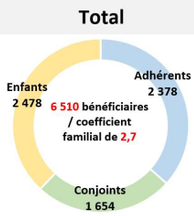
La première partie de la deuxième page de ce reporting expose le montant, par poste de santé et agrégé, de la consommation médicale modélisée à l'aide des algorithmes *Random Forest*. En d'autres termes, **les modèles *Random Forest* permettent de prédire le panier de soins d'un individu appartenant à la population sous-risque étudiée**. Des encadrés additionnels présentent les principales caractéristiques d'un individu moyen du portefeuille ainsi que quelques indicateurs.

La deuxième partie permet de faire un focus sur les gros consommants du portefeuille au niveau des postes optique et dentaire. Le but de cette partie est de présenter les caractéristiques afférentes aux gros consommants et de mettre en évidence de potentielles différences dans ces caractéristiques vis-à-vis de celles de l'ensemble du portefeuille. Ces différences sont indiquées par un encadré autour de ces variables.

Reporting Santé - Statistiques descriptives

Vue ensemble portefeuille

Adhérent	Conjoint	Enfants	Total
46 ans	46 ans 0,69 conjoints	13 ans 1,04 enfants	6 510 bénéficiaires / coefficient familial de 2,7 85 % de consommants
80% 20%			



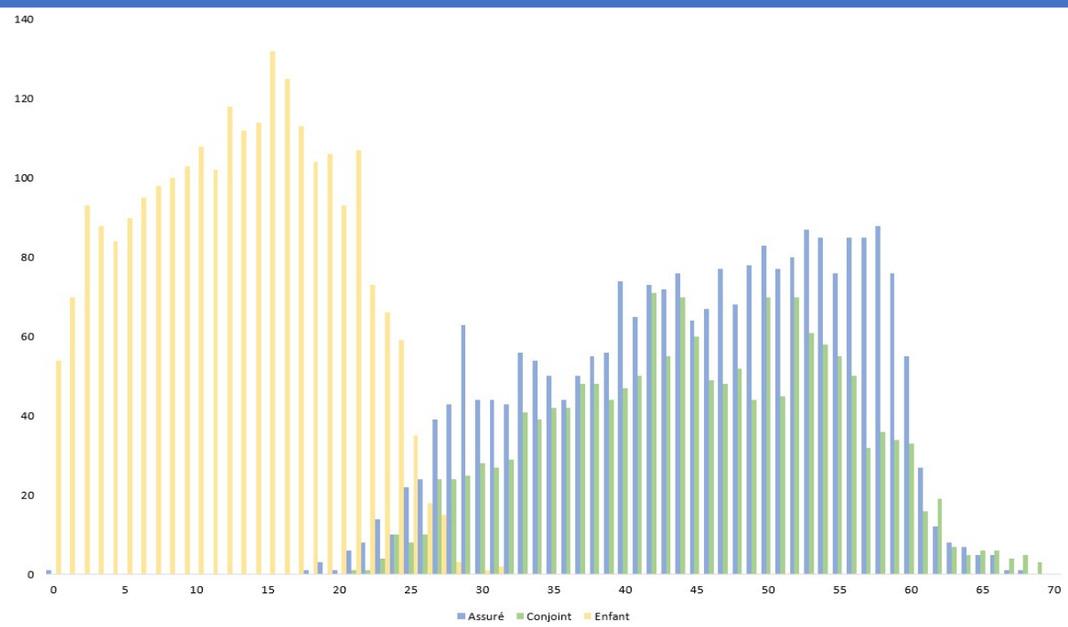
Dépenses totales **4,8 M €**

Remboursements Sécurité sociale 2,1 M€

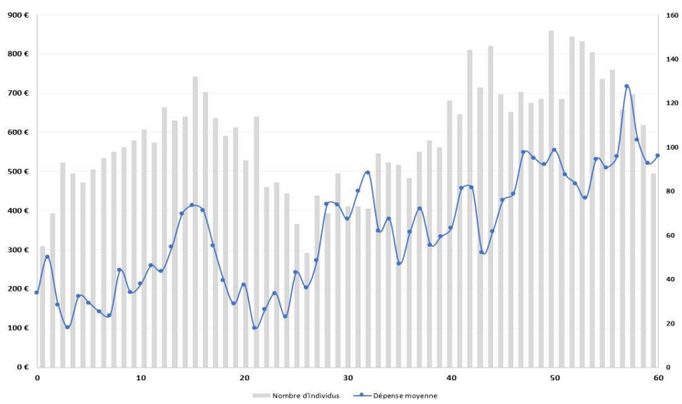
Remboursements Mutuelle 2,4 M€

Reste à charge Ménages 0,3 M €

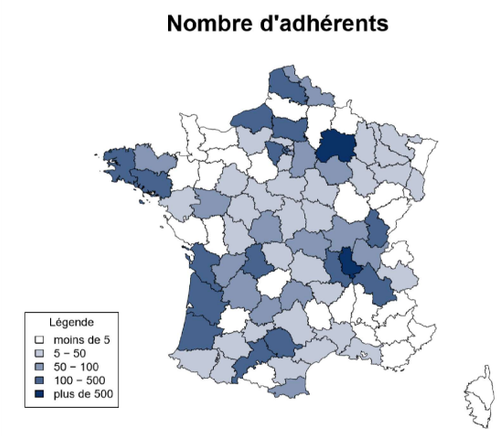
Pyramide des âges



Analyse de la consommation en fonction de l'âge

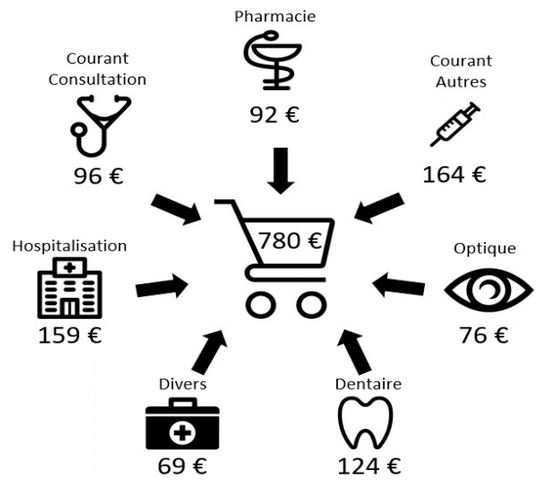


Répartition géographique des adhérents



Reporting Santé - Résultats des modèles prédictifs

Consommation moyenne annuelle - Ensemble portefeille



Indicateurs - Ensemble du portefeuille

85 % de consommants

Reste à charge moyen des ménages :
47 €

Caractéristiques individu moyen - Ensemble du portefeuille

Non cadre	68 %
Age :	33,2 ans
Ancienneté :	7,8 ans

Focus Gros consommants sur le poste Optique

Variables discriminantes

- Sexe
- Age
- Lieu d'habitation
- Nombre équipements santé
- **Proportion seniors**

Statistiques d'un consommant en optique

Dépenses
427 €

Prise en charge organisme complémentaire
384 €

Caractéristiques

Non cadre	66 %
Age :	42,7 ans
Ancienneté :	8,0 ans

Focus Gros consommants sur le poste Dentaire

Variables discriminantes

- **Sexe**
- Age
- Ancienneté
- Proportion maisons
- **Proportion seniors**

Statistiques d'un consommant en dentaire

Dépenses
496 €

Prise en charge organisme complémentaire
269 €

Caractéristiques

Non cadre	68 %
Age :	37,4 ans
Ancienneté :	8,5 ans

Conclusion

Nous avons, au cours de ce mémoire, étudié la consommation médicale d'une population couverte par un contrat collectif à adhésion obligatoire à l'aide de modèles *Data Science*.
Par conséquent, il était indispensable de rappeler quelques informations et chiffres sur le contexte économique et législatif dans le domaine de la santé afin de cerner les (multiples) problématiques associées.

Comme tout projet de *Data Science*, deux phases de travail bien distinctes sont nécessaires :

- Une première centrée sur les données ;
- Une deuxième étape de mise en place et calibrage des modèles.

La première étape de travail de la donnée découle de la problématique de l'étude. Nous souhaitons étudier la consommation médicale de manière individuelle. En effet, les bases de données initialement à notre disposition ne présentaient pas ce format : il a donc fallu les adapter à notre cadre d'étude.
De plus, afin de mettre en évidence les apports des modèles *Data Science*, les bases initiales ont été enrichies de différentes manières. Certaines variables ont été créées à partir de variables existantes : on parle de *feature engineering*, alors que d'autres variables ajoutées d'origine externe ont été ajoutées.

Une fois les données adaptées à notre problématique, le temps est venu de choisir sous quelle forme modéliser la consommation médicale. Dans ce mémoire, nous avons adopté une modélisation "Fréquence / Coût" où les coûts ne sont non pas les remboursements de l'organisme complémentaire mais les frais réels engagés.
La consommation médicale a été modélisée à l'aide de quatre modèles et du logiciel R :

- Les modèles linéaires généralisés (GLM) ;
- une méthode d'arbre de décisions : la méthode CART ;
- Deux modèles d'agrégation : le Random Forest et l'*eXtreme Gradient Boosting*.

L'étude approfondie des sorties des différents modèles au niveau du poste relatif aux consultations chez les médecins généralistes et spécialistes ainsi que les résultats des tests de pertinence ont entraîné la sélection du modèle *Random Forest* comme le modèle le mieux adapté à notre étude.

L'étude des résultats de ces modèles ont également permis de mettre en évidence que de nombreuses variables externes rentrent en compte dans l'explication de la consommation médicale. Certaines variables internes comme l'âge, le sexe ou la catégorie socio-professionnelle sont importantes dans l'étude de la consommation médicale. Néanmoins, des variables relatives à l'environnement économique ou fiscal (proportion de ménages imposés dans la commune, proportion de cadres dans la commune, ...) de l'assuré jouent également un rôle non-négligeable dans l'explication de la consommation médicale.

Par ailleurs, afin d'enrichir nos travaux, nous avons choisi pour cela de nous intéresser aux caractéristiques des individus consommant le plus au sein des postes dentaire et optique. Cette étude, parfois difficile avec les méthodes actuarielles classiques, a de nouveau été réalisée à l'aide de modèles *Data science*. Au sein du portefeuille étudié, 30 % des individus représentent près de 80 % des dépenses engagées. L'étude des caractéristiques de ces individus prend donc tout son sens.

Ces deux études ont été enfin incorporées au sein d'un reporting santé. Celui-ci peut être enrichi mais son intérêt réside dans le fait suivant : les résultats de modèles prédictifs trouvent toute leur place en santé et permettent d'exhiber des variables externes explicatives dans la consommation en santé.

Si l'on réfère à l'introduction de ce mémoire, nous nous interrogions sur la manière dont l'assureur pourrait améliorer la connaissance de son portefeuille. Nous citions par exemple les *open source* ou les réseaux sociaux. Ce mémoire tend à montrer qu'en effet, la recherche sur des sources accessibles et connues de tous (INSEE) de données et d'informations supplémentaires peuvent être un premier pas pour une meilleure connaissance client.
En outre, l'utilisation de ces informations disponibles serait un moyen de contourner la réticence de certains assurés à communiquer certaines données personnelles relatives à la santé.

Il existe bien évidemment des pistes d'amélioration à ce mémoire. Il pourrait être intéressant de poursuivre l'étude en analysant les données des années précédentes, ceci dans le but d'intégrer une variable "temporelle" à notre travail. Une étude de l'évolution de la consommation santé serait un vrai plus et contribuerait à enrichir ces travaux. De plus, il existe de nombreux autres modèles de *Data Science* qui mériteraient que l'on s'attarde dessus. Enfin, il existe une multitude de sources potentielles de données. Nous avons pris ici le parti de ne s'intéresser qu'à des données facilement accessibles. La prochaine mise en place de la RGPD va néanmoins entraîner des contraintes d'utilisation des données personnelles.
Néanmoins, cette étude nous aura permis de nous familiariser avec les domaines de la santé et des *Data Science*, de mettre en place une démarche précise et d'aboutir à des résultats et des modélisations de la consommation en santé.

Table des figures

Bibliographie

[1] Introduction au modèle linéaire général. math.univ-toulouse.fr.

[2] Seuls 10% des assureurs ont une stratégie *Big Data*. argusdelassurance.com.

[3] *Company accenture survey consumers buy insurance web giants*. accenture.com.

[4] Vivre en couple. insee.fr.

[5] R. Bellina. Méthodes d'apprentissage appliquées à la tarification non-vie. Master's thesis, ISFA, 2014.

[6] L. Breiman. *Bagging Predictors*. Department of Statistics University of California, Berkeley, 1994.

[7] L. Breiman. *RANDOM FORESTS*. Department of Statistics University of California, Berkeley, 2001.

[8] M. Baudry C. Robert. Introduction à la data science et applications à l'assurance. 2016.

[9] DREES. Les contrats les plus souscrits auprès des organismes complémentaires santé en 2010.

[10] DREES. Les dépenses de santé en 2015 Édition 2016. pages 4–25.

[11] DREES. La complémentaire santé : Acteurs, bénéficiaires, garanties. page 46, 2016.

[12] DREES. L'essor des modèles prédictifs dans les systèmes de santé internationaux. pages 4–25, 2017.

[13] M. Lutz E. Biernat. *Data science : fondamentaux et étude de cas*. Eyrolles, 2016.

[14] J. H. Friedman. *Greedy function approximation : a gradient boosting machine*. IMS 1999 Reitz Lecture, 1999.

[15] S. Jamal. Construction du taux de rachat structurel en epargne : approximation non-linéaire et agrégation de modèles. Master's thesis, ISFA, 2016.

[16] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics*, pages 119–127, 1980.

[17] C. J. Stone R.A. Olshen L. Breiman, J. Friedman. *Classification and Regression Trees*. Taylor and Francis, 1984.

[18] M. J. Lagnaoui. Machine learning et nouvelle classe de problèmes actuariels : une illustration par l'étude de cas. Master's thesis, EURIA, 2015.

[19] B. Lantz. *Machine Learning with R Second Edition*. Packt Publishing, 2015.

[20] J.A Mc Cullagh P. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

[21] J. R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[22] J. Friedman T. Hastie, R. Tibshirani. *The Elements of Statistical Learnings Second Edition*. Springer, 2008.

[23] R. Schapire Y. Freund. A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, pages 119–139, 1997.

Annexes

Liste ALD 30

Les Affectations de Longue Durée (ALD) ont été mentionnées plusieurs fois au cours de ce mémoire. Les ALD sont des maladies dont la gravité et/ou le caractère chronique nécessitent un traitement prolongé et une thérapeutique particulièrement coûteuse. Voici la liste exhaustive de ces affectations :

- Accident vasculaire cérébral invalidant ;
- Insuffisances médullaires et autres cytopénies chroniques ;
- Artériopathies chroniques avec manifestations ischémiques ;
- Bilharziose compliquée ;
- Insuffisance cardiaque grave, troubles du rythme graves, cardiopathies valvulaires graves, cardiopathies congénitales graves ;
- Maladies chroniques actives du foie et cirrhoses ;
- Déficit immunitaire primitif grave nécessitant un traitement ; prolongé, infection par le virus de l'immuno-déficience humaine (VIH) ;
- Diabète de type 1 et diabète de type 2 ;
- Formes graves des affections neurologiques et musculaires (dont myopathie), épilepsie grave ;
- Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères ;
- Hémophilies et affections constitutionnelles de l'hémostase graves ;
- Maladie coronaire ;
- Insuffisance respiratoire chronique grave ;
- Maladie d'Alzheimer et autres démences ;
- Maladie de Parkinson ;
- Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé ;
- Mucoviscidose ;
- Néphropathie chronique grave et syndrome néphrotique primitif ;
- Paraplégie ;
- Vascularites, lupus érythémateux systémique, sclérodermie systémique ;
- Polyarthrite rhumatoïde évolutive ;
- Affections psychiatriques de longue durée ;
- Rectocolite hémorragique et maladie de Crohn évolutives ;
- Sclérose en plaques ;
- Scoliose idiopathique structurale évolutive (dont l'angle est égal ou supérieur à 25 degrés) jusqu'à maturation rachidienne ;
- Spondylarthrite grave ;
- Suites de transplantation d'organe ;
- Tuberculose active, lèpre ;
- Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique.

Exemple de découpage des actes médicaux dans les bases initiales

Comme indiqué au cours de la partie 2.1.3, trois variables (FAMILLE, SOUS-FAMILLE et ACTE) permettent de classer les actes médicaux.
Voici un exemple de cette classification pour la catégorie Dentaire de la variable FAMILLE :

FAMILLE	SOUS-FAMILLE	ACTE
Dentaire	Orthodontie	Orthodontie
	Soins dentaires	Radio dentaire
		Honoraires chirurgien-dentiste
		Parodontose
		Inlay-Onlay
		Soins dentaires
		Soins et honoraires non-conventionnés
	Prothèses dentaires acceptées	Prothèses dentaires
		Prothèses SPR 57
		Prothèses SPR 67
		Couronnes
		Réparations prothèses et couronnes
	Prothèses dentaires refusées	Prothèses dentaires non-conventionnées
		Implants dentaires
		Moignons implants

Quelques statistiques additionnelles sur le portefeuille étudiée

Les deux tableaux suivants permettent de présenter les coûts moyen par consommant et par bénéficiaire selon les différents postes et sous-postes étudiés :

Poste	Frais réels moyens	Remboursements moyens Sécurité Sociale	Remboursements moyens Mutuelle	Reste à charge moyens
Forfait journalier	152,84 €	1,77 €	150,88 €	0,19 €
Frais de séjour	1 126,56 €	803,98 €	318,93 €	3,65 €
Transport ambulance	507,54 €	340,71 €	166,72 €	0,12 €
Honoraires	172,13 €	92,40 €	78,50 €	1,22 €
Ticket modérateur	21,05 €	- €	21,00 €	0,05 €
Hospitalisation	**463,16 €**	**284,89 €**	**176,29 €**	**1,98 €**
Généralistes	102,14 €	66,54 €	34,91 €	0,69 €
Spécialistes	93,12 €	45,95 €	43,48 €	3,69 €
Consultations Visites	**143,98 €**	**85,79 €**	**55,55 €**	**2,64 €**
Soins infirmiers	24,70 €	14,91 €	9,70 €	0,09 €
Kinésithérapie	325,66 €	182,11 €	141,52 €	2,03 €
Radiologie	85,17 €	53,72 €	28,87 €	2,59 €
Laboratoire	86,35 €	52,13 €	34,15 €	0,07 €
Acte en k	74,75 €	47,45 €	26,51 €	0,80 €
Ortophonie	626,73 €	376,04 €	250,69 €	- €
Orthoptie	77,79 €	46,16 €	31,62 €	0,00 €
Osthéopatie	86,18 €	- €	50,06 €	36,12 €
Podologie	103,48 €	31,56 €	42,00 €	29,92 €
Soins courants	**226,30 €**	**117,21 €**	**93,57 €**	**15,52 €**
Pharmacie	**141,29 €**	**80,43 €**	**60,82 €**	**0,04 €**
Implants dentaires	1 534,94 €	8,90 €	702,96 €	823,08 €
Orthodontie	759,94 €	225,41 €	468,57 €	65,96 €
Prothèses dentaires	1 313,05 €	226,05 €	790,38 €	296,63 €
Soins dentaires	107,57 €	69,72 €	32,25 €	5,60 €
Dentaire	**496,51 €**	**137,96 €**	**268,74 €**	**89,82 €**
Lunettes	475,93 €	13,87 €	424,06 €	38,00 €
Lentilles	249,10 €	4,13 €	238,55 €	6,42 €
Optique	**427,31 €**	**11,78 €**	**384,30 €**	**31,23 €**
Divers	**279,35 €**	**56,24 €**	**192,11 €**	**31,00 €**

FIGURE 4.44 – Coût moyen observé par consommant

Poste	Frais réels moyens	Remboursements moyens Sécurité Sociale	Remboursements moyens Mutuelle	Reste à charge moyens
Forfait journalier	6,78 €	0,08 €	6,70 €	0,01 €
Frais de séjour	84,45 €	60,27 €	23,91 €	0,27 €
Transport ambulance	8,81 €	5,91 €	2,89 €	0,00 €
Honoraires	51,19 €	27,48 €	23,35 €	0,36 €
Ticket modérateur	1,16 €	- €	1,16 €	0,00 €
Hospitalisation	**152,40 €**	**93,74 €**	**58,01 €**	**0,65 €**
Généralistes	61,00 €	39,74 €	20,85 €	0,41 €
Spécialistes	33,13 €	16,35 €	15,47 €	1,31 €
Consultations Visites	**94,13 €**	**56,09 €**	**36,32 €**	**1,72 €**
Soins infirmiers	6,65 €	4,01 €	2,61 €	0,02 €
Kinésithérapie	27,91 €	15,61 €	12,13 €	0,17 €
Radiologie	20,24 €	12,77 €	6,86 €	0,61 €
Laboratoire	25,85 €	15,61 €	10,23 €	0,02 €
Acte en k	6,96 €	4,42 €	2,47 €	0,07 €
Ortophonie	9,24 €	5,55 €	3,70 €	- €
Orthoptie	1,41 €	0,84 €	0,57 €	0,00 €
Osthéopatie	12,64 €	- €	7,34 €	5,30 €
Podologie	6,34 €	1,93 €	2,57 €	1,83 €
Soins courants	**117,25 €**	**60,73 €**	**48,48 €**	**8,04 €**
Pharmacie	**96,19 €**	**54,76 €**	**41,41 €**	**0,03 €**
Implants dentaires	8,25 €	0,05 €	3,78 €	4,43 €
Orthodontie	30,93 €	9,18 €	19,07 €	2,69 €
Prothèses dentaires	67,77 €	11,67 €	40,79 €	15,31 €
Soins dentaires	23,84 €	15,45 €	7,15 €	1,24 €
Dentaire	**130,80 €**	**36,34 €**	**70,80 €**	**23,66 €**
Lunettes	63,24 €	1,84 €	56,35 €	5,05 €
Lentilles	9,03 €	0,15 €	8,65 €	0,23 €
Optique	**72,27 €**	**1,99 €**	**64,99 €**	**5,28 €**
Divers	**67,84 €**	**13,66 €**	**46,65 €**	**7,53 €**

FIGURE 4.45 – Coût moyen observé par bénéficiaire

Description des variables externes

La quasi-totalité des variables externes ont été extraites sur le site internet de l'INSEE et sont relatives à la commune de résidence de l'individu. L'extraction a été possible grâce à la présence du code postal dans les données initiales. Celui a ainsi pu être rattaché au code INSEE et les informations désirées ont pu être importées.
Voici la liste exhaustive des variables externes utilisées au cours de ce mémoire :

- Nombre Équipements santé : nombre d'établissements de court, moyen et long séjours, de maternités, de pharmacies, de laboratoires d'analyses médicales, ... dans la commune en 2016 pour 1 000 habitants ;
- Nombre Pharmacies : nombre de pharmacies dans la commune en 2016 pour 1 000 habitants ;
- Nombre Praticiens : nombre de fonctions médicales et paramédicales dans la commune en 2016 pour 1 000 habitants ;
- Nombre Médecins : nombre de médecins généralistes dans la commune en 2016 pour 1 000$habitants$;
- Nombre Ophtalmo : nombre d'ophtalmologues dans la commune en 2016 pour 1 000$habitants$;
- Nombre Dentistes : nombre de dentistes dans la commune en 2016 pour 1 000$habitants$;
- Nombre Équipements sportifs : nombre d'équipements sportifs (terrains de football, courts de tennis, pistes d'athlétisme, ...) dans la commune en 2016 pour 1 000$habitants$;

N.B. : Le nombre d'habitants pour chaque commune a été extrait de la base de données issue du recensement effectué en 2014*.*

- Proportion Activité : proportion d'individus ayant un emploi de 15 à 64 ans de la commune en 2014 ;
- Proportion Chômeurs : proportion de chômeurs de 15 à 64 ans (à mettre en rapport avec le nombre de personnes actives de 15 à 64 ans) dans la commune en 2014 ;
- Proportion Cadres : proportion de cadres dans la population active de 15 à 64 ans de la commune en 2014 ;
- Proportion Ouvriers : proportion d'ouvriers dans la population active de 15 à 64 ans de la commune en 2014 ;
- Proportion CDI : proportion de de personnes ayant un contrat à durée indéterminée dans la population active de 15 à 64 ans de la commune en 2014 ;
- Proportion allant à pied : proportion d'actifs de 15 à 64 ans se rendant à pied à leur lieu de travail dans la commune en 2014 ;
- Proportion Ménages imposés : proportion des ménages imposés dans la commune en 2014 ;
- Médiane revenus : revenu médian de la commune en 2014 ;

N.B. : le revenu évoqué ici correspond au revenu déclaré ou revenu fiscal, c'est-à-dire aux ressources mentionnées sur la déclaration des revenus. Cette notion est à dissocier du revenu disponible qui comprend les revenus d'activité, les revenus du patrimoine, les revenus de transferts et les prestations sociales[4].

- Nombre moyen personnes ménages : nombre moyen de personnes par ménage dans la commune en 2014 ;
- Proportion famille : part des ménages composée d'une famille (au moins deux personnes) dans la commune en 2014 ;
- Proportion mineurs : proportion de mineurs dans la commune en 2014 ;
- Proportion seniors : proportion de seniors (individus âgés de plus de 65 ans) dans la commune en 2014 ;
- Loyer moyen : prix moyen du mètre carré dans le département en 2017 ;
- Zone : nombre de résidents dans la commune en 2014.

Ces variables, à l'origine numériques, ont été catégorisées afin de leur donner plus de sens. Ces regroupements ont été réalisés à l'aide d'études statistiques sur le portefeuille étudié. Des modalités "Faible", "Moyen" ou "Élevé" ont ainsi été créées pour la majorité des variables.

4. Source : INSEE

Effet de l'agrégation d'arbre sur la variance des estimations d'un modèle *Bagging*

Après la mise en place d'un modèle *Bagging*, une suite d'arbres, et non plus un simple arbre, est obtenue. En notant cette suite $(\phi_k)_{k=1,\dots,N}$, l'arbre final est défini par $\phi = \frac{1}{N}\sum_{k=1}^{N}\phi_k$.

En supposant de plus que les arbres ϕ_k sont deux à deux corrélés par ρ et qu'ils possèdent une variance σ^2, alors :

$$\begin{aligned} V[\phi] &= V\left[\frac{1}{N}\sum_{k=1}^{N}\phi_k\right] \\ &= \frac{1}{N^2}V\left[\sum_{k=1}^{N}\phi_k\right] \\ &= \frac{1}{N^2}\sum_{k=1}^{N}\left(\sigma^2 + \sum_{k'=1,k'\neq k}^{N}\rho\sigma^2\right) \\ &= \frac{\sigma^2}{N^2}\sum_{k=1}^{N}(1+(N-1)\rho) \\ &= \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{N} \end{aligned}$$

La variance décroît donc en fonction du nombre d'itérations.