# *VALEUR CLIENT : MODELISATION, THEORIE ET PRATIQUE EN ASSURANCE SANTÉ*

Mémoire CEA

**Avril 2018**

**Yann-Erlé LE ROUX**
yeleroux@gmail.com

A mes chéris,
Maïa, Milan & R.

# Remerciements

Merci à tous ceux qui m'ont aidé, Blaise, Charles, Fanny, Florence, Frédéric, Gaëlle, Jessica, Louis, Magalie, Martine, Mehdi, Nadège, Olivier, Pascal, Ronan, Stéphane, Sylvain, Yannick et tous mes collègues de travail qui ont cru au projet et qui m'ont soutenu dans la démarche. Je n'y serais jamais arrivé sans vous.

Un grand merci à Marie Piffault, pour la supervision de la rédaction de ce mémoire, et pour ses nombreux conseils en tous points très précieux.

Un grand merci à Marie-Hélène Péjoine, pour avoir détecté le potentiel du projet, pour m'avoir accordé le temps nécessaire à l'élaboration du modèle et pour son accompagnement sur la mise en place de la partie Marketing.

Un grand merci à Dana Stephane, qui m'a fait grandir, et qui m'a convaincu que chercher à progresser et continuer à se former sans cesse, n'était pas une option.

Enfin, un grand merci à Virginie Hauswald, qui me fait prendre de la hauteur et m'incite constamment à découvrir des nouvelles frontières.

# Résumé

Le contexte économique et réglementaire actuel génère une concurrence exacerbée entre assureurs. Cette situation les incite à mieux connaître leurs assurés afin de piloter plus finement leurs activités. Le contexte technologique présente une réelle opportunité d'obtenir enfin une vision 360° du client pour tenter notamment de répondre aux questions : Qu'est-ce qu'un bon client pour une entreprise d'assurance ? Pour aujourd'hui ou demain, est-ce la même réponse ?

La valeur client est un dispositif qui contribue au débat. Elle représente en effet une métrique prospective qui estime la rentabilité et le potentiel économique de chaque client d'un portefeuille d'assurance. Le modèle associé est un puissant outil marketing qui, placé au centre des processus opérationnels de l'entreprise, permet de définir des stratégies de relation client proportionnées à la marge dégagée par chaque client ou segment de clients. Il offre ainsi une possibilité d'allouer efficacement les ressources et de mieux prioriser les efforts afin de fidéliser les meilleurs profils, de conquérir de nouveaux marchés, et d'optimiser la communication et les produits.

Le présent mémoire décrit le processus qui nous conduit à construire une modélisation actuarielle de la sinistralité et de la durée de vie, futures probables tous les deux, des assurés d'un portefeuille individuel interprofessionnel, couvert par des garanties Santé et Prévoyance. Le modèle est bâti selon ce qu'on estime être les règles de l'art. Il est notamment fondé sur la théorie de la crédibilité et sur les méthodes de scoring modernes issues de l'école de la Data Science. L'ambition de ce mémoire est double puisque la pratique se joint à la théorie : il décrit une collection d'outils pertinents et efficaces, à portée opérationnelle immédiate auprès des directions en contact avec le client, le Développement Commercial et la Relation Client.

**Mots clés :** Valeur Client, Modélisation, Stratégie, Santé, Prévoyance, Marketing, Data Mining, Data Science, Crédibilité, Analyse prédictive, Scoring, Segmentation.

# Abstract

The current economic and regulatory environment generates an exacerbated competition among insurers. This situation prompts them to better understand the insured pool to steer their activity more precisely. The technological context offers a real opportunity to finally build a 360° customer view in order to answer the following questions: "What is a "good" customer for an insurance company? Will the answer be the same, now and in the future?

The customer lifetime value system contributes to this debate. It is a forward looking metric that estimates the profitability and economic potential of each client of an insurance portfolio. The associated model is a powerful marketing tool which, placed at the heart of the company business processes, helps define a client relationship strategy based on the margins generated by each customer or customer segment. It also allows to effectively allocate the available resources and better prioritise the effort to retain the best profiles, conquer new markets, optimise the communications and the products.

This paper describes the process that leads to build an actuarial modeling of loss ratio and life expectancy, both probable futures, of the insured in an interprofessional individual portfolio, covered with a health and life insurance. The model is built upon what we esteem is state of the art knowledge. In particular, it is based on the credibility theory and modern scoring methods as developed by the Data Science school of thought. The ambition of this paper is twofold given that theory meets practice: it presents a collection of relevant and effective tools, readily available to all departments working with the client, the business development and the client relationship units.

**Key words:** Customer Lifetime Value, Modeling, Strategy, Health, Life insurance, Marketing, Data Mining, Data Science, Credibility, Predictive analysis, Scoring, Classification.

# Sommaire

# Partie 1 : Prologue

*"Nous ne sommes qu'un maillon précieux d'une chaîne éternelle*
*dont une extrémité se perd dans l'inconnaissable*
*tandis que l'autre reste encore à forger."*

Robert Ardrey

*"- A quoi est due la chute d'Adam et Eve ?*
*- C'était une erreur de Genèse."*

Boris Vian

## 1.1. Genèse

Je suis entré dans le milieu de l'assurance il y a tout juste 10 ans lorsque j'ai intégré la Direction Technique de La Mutuelle Générale. Je n'y connaissais rien ou presque. Ma vision de l' « assurance » se limitait alors à des tâches administratives le plus souvent obligatoires et rébarbatives. « S/P » ne m'évoquait rien d'autre que l'empilement[1] sur une table à manger de la salière et du poivrier. « Solvabilité » raisonnait en moi comme un mot barbare. Et je n'avais qu'une idée très vague de ce qui se cachait derrière le mot « actuaire ». J'avais entendu une fois quelqu'un en parler ; cela ne m'avait pas, à l'époque, semblé particulièrement excitant. Bref j'entrais dans l'inconnue la plus totale.

Je l'avoue, j'ai en définitive, tout de suite accroché. De formation universitaire en sciences économiques, très orientée « traitement et analyse des données », la pratique de la science actuarielle représentait de fait une suite possible et logique de mon parcours professionnel. Après quelques années -heureuses- de formation et perfectionnement, j'ai ressenti le besoin de parfaire mes connaissances théoriques. Le CEA s'est pratiquement imposé à moi, comme une évidence. J'y suis allé, pas très serein, mais empli de bonne volonté et de désir d'apprendre.

Parallèlement, en cours de formation, la Direction Marketing LMG m'a offert l'opportunité de retrouver mes premières amours. Elle souhaitait monter en compétence sur la Connaissance Client en ouvrant un poste de Data Miner en son sein, en partie pour améliorer le ciblage des campagnes commerciales et marketing, et déployer une culture *Scoring* & *ROI*. Mon idée n'était en aucun cas d'abandonner l'actuariat mais plutôt de développer une polycompétence 'Actuariat-Marketing-Data'. De fait, je conservais durant ces trois années des liens étroits avec mon ancienne direction. Et j'incarnais, occasionnellement, –avec plaisir, et je crois, loyauté, transversalité et professionnalisme-, le chaînon manquant entre ces deux entités. Actuariat et Marketing, deux facettes essentielles de l'assurance. Une dualité aux relations parfois tendues en raison de leurs objectifs généralement opposés : Améliorer ou conserver l'équilibre technique *vs* Augmenter le nombre de ventes de contrats ou le chiffre d'affaires.

Le Data Mining, c'est un vaste sujet. Une traduction littérale pourrait être « exploration de données » ou encore « extraction de connaissances à partir de données ». Usama M. Fayyad, qui a notamment exercé les fonctions de vice-président et de responsable des données chez Yahoo, a proposé en 1996, la définition suivante : « Processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données ».

Autrement dit, on cherche, par un processus complexe, du fait du volume et de l'hétérogénéité des données, à découvrir des faits nouveaux, véritables, et éventuellement praticables à travers les systèmes d'information. L'analyse de données et les statistiques exploratoires existent depuis plus de 40 ans. Considérons-le Data Mining comme un prolongement de ces domaines, avec toutefois des différences notables :

- ajout de techniques issues de l'Intelligence Artificielle, comme le Machine Learning
- travail potentiel sur des données pas nécessairement structurées
- finalité de plus en plus orientée « Business »

[1] Lorsque cela est possible !

Au cours de la dernière décennie, nous avons assisté à un profond bouleversement des métiers de la donnée, une « rupture de paradigme »[2]. Les capacités de stockage et de traitement des machines ont littéralement explosé. Hasard ou coïncidence, au même moment, nous faisions face à l'apparition d'un déluge de nouvelles données, principalement issue de la démocratisation planétaire des outils Web et des services associés. En 15 ans, la production annuelle de données a été multipliée par 500 et 90% des données récoltées à ce jour ont moins de deux ans d'existence !

L'expression « *Big Data* » est née. Elle a fait le *buzz* pendant quelques mois avant de se démocratiser comme l'ont fait par la suite, le *Cloud*, l'*Internet of Things*, le *crowdfunding*, tous ces termes qui définissent les technologies et les usages qui façonnent le numérique et, plus largement, l'économie d'aujourd'hui et de demain. Le Data Mining y est étroitement lié. Certains y voient une potentielle 4ème révolution industrielle, tant les avancées potentielles et les domaines d'application sont nombreux. Dorénavant la « Data » infléchit le business en accélérant la transformation des entreprises.

La définition du Big Data par Gartner[3] et la règle dite des 3 V (Volume, Vélocité, Variété) se sont imposées : l'accent est mis sur le volume et la performance de recueil et de stockage des données. Aujourd'hui, on parle aussi de *Smart Data*. 2 V supplémentaires : Véracité et Valeur. On pointe davantage sur la qualité et l'intelligence du traitement des données.

Peu importe ces *buzzwords*, la réalité, c'est que nous sommes confrontés à une puissance technologique illimitée, à des nouveaux usages liés à un traitement digital et global des informations ainsi qu'à des exigences d'interaction accrues avec les utilisateurs. Les données sont désormais structurées, semi-structurées ou non structurées par l'intermédiaire des réseaux sociaux, des terminaux mobiles, des transactions Internet et des capteurs électroniques ; les individus confiant toujours plus de données, parfois à leur insu, tandis que les réglementations ne sont pas encore harmonisées.[4]

D'autre part, pour la première fois dans l'Histoire, nous disposons d'historiques de données suffisamment profonds qui permettent de dépister des modèles de comportements inédits. Nous sommes davantage en capacité de comprendre, agir, cibler, mesurer, décrire, diagnostiquer, prédire, prescrire, plus que jamais auparavant. C'est pourquoi, les choix stratégiques d'investissement doivent dorénavant se matérialiser nécessairement en prise de décision et en actes tangibles. Ce n'est plus un mythe, ni même un rêve, la Data représente très concrètement désormais, une source de valeur et un actif stratégique contribuant au développement d'une intelligence collective centrée autour du client.

A La Mutuelle Générale, nous avons tué dans l'œuf le projet « Big Data », en démythifiant le phénomène « Magie de l'or noir ». Mais nous sommes restés attentifs à l'évolution de cette révolution et nous nous positionnons aujourd'hui en tant qu'acteur agile et innovant, à la pointe du défi technologique qui s'offre à nous.

---

[2] Gilles Babinet, « digital champion » de la France auprès de la Commission européenne.

[3] «"Big data" is high-volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making » ce qui peut se traduire par « Les mégadonnées sont des informations de gros volume, de haute vitesse et de grande variété qui exigent des formes de traitement d'un bon rapport coût-performance et innovantes pour une compréhension accrue et la prise de décisions ».

[4] En attendant la mise en application du règlement européen sur la protection des données GDPR le 25 mai 2018.

En assurance, 5 fonctions de la chaîne de valeur sont particulièrement impactées par ce déluge :

- le marketing qui devient *data driven*,
- la lutte contre la fraude qui est désormais automatisable à grande échelle,
- la tarification dont les fondements se voient bouleversés,
- la gestion qui voit là une extraordinaire opportunité pour réduire ses coûts opérationnels et de structure,
- et la relation client qui peut enfin espérer devenir une expérience client de qualité.

Chacune de ses fonctions a fait et continue de faire l'objet de nombreux cas d'usage en interne, sans cesse challengés au cours d'ateliers transverses et collaboratifs.

C'est dans cet environnement singulier que je me situe au printemps 2015, au moment de choisir le sujet de mon mémoire validant le cursus CEA. Quatre contraintes s'imposent naturellement :

1. Le sujet doit traiter d'Actuariat, cela va de soi.
2. Le sujet doit être connecté avec mes nouvelles fonctions au Marketing.
3. Le sujet doit s'inscrire dans le contexte décrit ci-dessus et orienté vers les nouvelles technologies.
4. Enfin, il faut nécessairement que le sujet choisi apporte une valeur ajoutée substantielle à La Mutuelle Générale, tout en étant raccord avec la politique stratégique de l'entreprise.

Le thème retenu sera la **« Valeur Client »** : une modélisation en bonne et due forme, complète et au maximum de la faisabilité potentielle, guidée par les mécanismes théoriques les plus en phase avec la réalité, et accompagnée de travaux pratiques à finalités opérationnelles immédiates.

Depuis quelques années, on ressentait en interne un réel besoin en ce sens, un vide à combler. Nous avions, collectivement, le sentiment que LMG devait être mieux à même d'évaluer le potentiel de rentabilité de ses clients et prospects non affinitaires. Dans un contexte de marché extrêmement tendu en conquête, afin de nous distinguer efficacement de nos concurrents, il était nécessaire de nous équiper d'un outil capable de nous aider à mieux cibler nos investissements commerciaux, marketing et communication. Mais ce programme ne figurait sur aucune des feuilles de route des directions métiers.

En 2013, Marie Piffault, ma collègue de travail de l'époque à la DT, lance la première initiative en y consacrant son thème de mémoire CEA[5]. En particulier, la partie sur les lois de survie des assurés dans le portefeuille est unanimement saluée. La méthodologie et les résultats obtenus sont désormais intégrés dans la boîte à outils des actuaires. Son traitement du sujet est cependant résolument orienté « gestion du risque Santé ». Elle démontre en particulier que la Valeur Client (VC) doit être considéré comme une alternative sérieuse au ratio S/P, a minima un indicateur complémentaire plus complet et indispensable. Mais ce travail n'inclut pas de dimension « marketing » à finalité opérationnelle, ce qui justifie l'existence d'une deuxième approche.

[5] « La valeur client comme mesure de la rentabilité d'un portefeuille d'assurance santé individuelle ».

L'objectif de mes recherches reste toutefois similaire : **Comment peut-on doter l'entreprise d'une évaluation pertinente de la « valeur » d'un client ?**

Les approches classiques ne sont en effet plus appropriées : une méthode basée sur le bilan de l'entreprise permet certes d'obtenir une valeur économique de notre portefeuille mais ne dit rien du potentiel de croissance ni de la dynamique de survie des clients. Le compte de résultat apporte des informations précieuses sur la marge annuelle nette générée par l'ensemble des clients mais ne dit rien sur les perspectives dans la durée. En outre, il est très complexe de ventiler raisonnablement ces deux indicateurs tête à tête.

D'autre part, le ratio sinistre à prime[6] donne une représentation adéquate des équilibres techniques du portefeuille, mais lui non plus ne s'inscrit guère dans la prospective. Enfin, la durée de la relation client, mesure essentielle de la fidélité devenue un enjeu majeur pour les organismes d'assurance, procure des hypothèses sur la capacité à couvrir collectivement les coûts d'acquisition mais ne fournit en aucun cas d'indication sur la marge unitaire et son évolution au cours du temps.

Donc, seule la valeur client, concept qui consiste à mesurer la rentabilité d'un client dans la durée, apparaît comme la métrique appropriée pour définir quels sont nos clients les plus précieux, générateurs de marge, et quels sont ceux qui dégradent la rentabilité de l'entreprise. La VC permet aussi de répondre à la question : « **Quels sont les segments de clients vers lesquels doivent porter nos efforts de conquête, de fidélisation, et de communication ?** »

Par ailleurs, l'assurance fait aujourd'hui face aux « 4 cavaliers de l'apocalypse[7] qui sont la baisse des taux d'intérêt, l'avalanche réglementaire, la concurrence accrue des bancassureurs et la digitalisation ». En assurance santé individuelle en particulier, désormais les contrats sont de plus en plus harmonisés et la concurrence est de plus en plus sévère. De fait les marges se tassent inexorablement. **Dès lors la question se pose pour un assureur : « La mise en œuvre d'un contrat santé est-t-elle toujours rentable de nos jours ? Aurons-nous la garantie d'amortir les coûts d'acquisition investis ? Et si oui, dans quels délais ? ».** Le modèle de valeur client devrait être en mesure de répondre à ces interrogations.

Mon intuition : la valeur client doit représenter le sextant[8] de l'entreprise.
Ma conviction : le pilotage par la valeur client est la clé de la réussite.

Nous étudierons dans un premier temps les définitions et concepts clés de la VC (Partie II). Puis, nous examinerons attentivement les fondamentaux constitutifs de la mesure VC (Partie III). Ensuite, une attention particulière sera portée au domaine de l'analyse prédictive, substance primordiale du mémoire (partie IV). Nous aurons alors tous les éléments pour établir les calculs des différentes valeurs clients (Partie V). Enfin, nous terminerons par l'analyse des principaux résultats et nous présenterons les enseignements essentiels du modèle construit (Partie VI).

[6] Voir partie 5.2. Calcul de P/C
[7] Nicolas Gomart, Directeur Général de Matmut
[8] Guizouarn & Marescaux

# Partie 2 : Valeur Client : Définition et concepts

*« Réduit à l'essentiel, l'objectif du marketing est d'attirer et de fidéliser des clients rentables »*

Dr Philip Kotler & Bernard Dubois

## 2.1. Qu'est-ce que la valeur client ?

### 2.1.1. Un indicateur qui modélise le potentiel économique de chaque client

#### *Définition*

La valeur client (VC) est une métrique *forward looking*[9]. Elle se détermine par la somme actualisée des marges générées sur toute la durée de vie du client, en tenant compte de son historique de comportement et de ses données sociodémographiques.

#### *Formule générique*

La valeur du client est égale à la somme actualisée des flux qui seront générés par cet actif.

$$valeur\ client = \sum_{t=0}^{\infty} \frac{F_t}{(1+i)^t}$$

t : année considérée
i : taux d'actualisation
$F_t$ : flux de l'année t

#### *Horizon*

Dans la pratique, on travaille sur un horizon fini.

#### *Taux d'actualisation*

Les flux peuvent être actualisés à partir d'un taux de rendement interne, d'un objectif de croissance, d'un taux d'inflation estimé ou tout autre indicateur économique qui se justifierait.

#### *Modélisation des flux financiers*

Les flux financiers peuvent être déterminés de la façon suivante : $F_t = E(M_t) \times p_t$

$E(M_t)$ : espérance de marge en t
$p_t$ : probabilité de survie à t

#### *Formule précise*

$$VC = \sum_{t=0}^{T} \frac{[E(M'_t) - K_t] \times p_t}{(1+i)^t}$$

[9] Prospective.

Avec $E(M_t) = [E(M'_t) - K_t]$

$t$ : année considérée

$T$ : nombre d'années de la relation commerciale

$E(M'_t)$ : espérance de marge récurrente en t

$K_t$ : coûts récurrents, dont coûts marketings, en t

$p_t$ : probabilité de survie à t

$i$ : taux d'actualisation

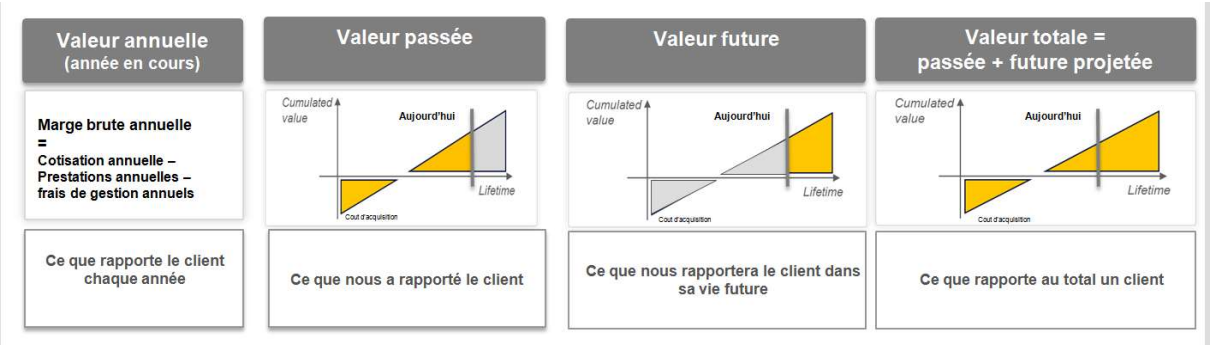
### *Espérance de marge*

Elle dépend de nombreux facteurs que nous pouvons supposer fixes ou variables dans le temps. Trois facteurs sont corrélés entre eux : le temps, l'âge et la génération (cohorte). La consommation dépend à la fois de notre environnement, de l'offre de soins, de notre état de santé lié en grande partie à l'âge et des habitudes de la génération. Si le client ne change pas de segment sur la période d'étude, la valeur portefeuille est alors la somme des valeurs de chaque cohorte.

*Source : Jean-Charles Guizouarn, Nicolas Marescaux « Assurance Santé – segmentation et compétitivité »*

## 2.1.2. Une modélisation mathématique appliquée à chaque client

La modélisation, objet du mémoire, prend en compte la richesse nette de frais et les coûts variables à partir des critères issus du Système d'Information de l'entreprise (front office, back office et enrichissement de données externes).




On distingue 4 valeurs clients « *in force* » :

VCP : valeur client passée
VCA : valeur client actuelle (année en cours)
VCF : valeur client future
VCT : la somme des trois précédentes moins les coûts d'acquisition

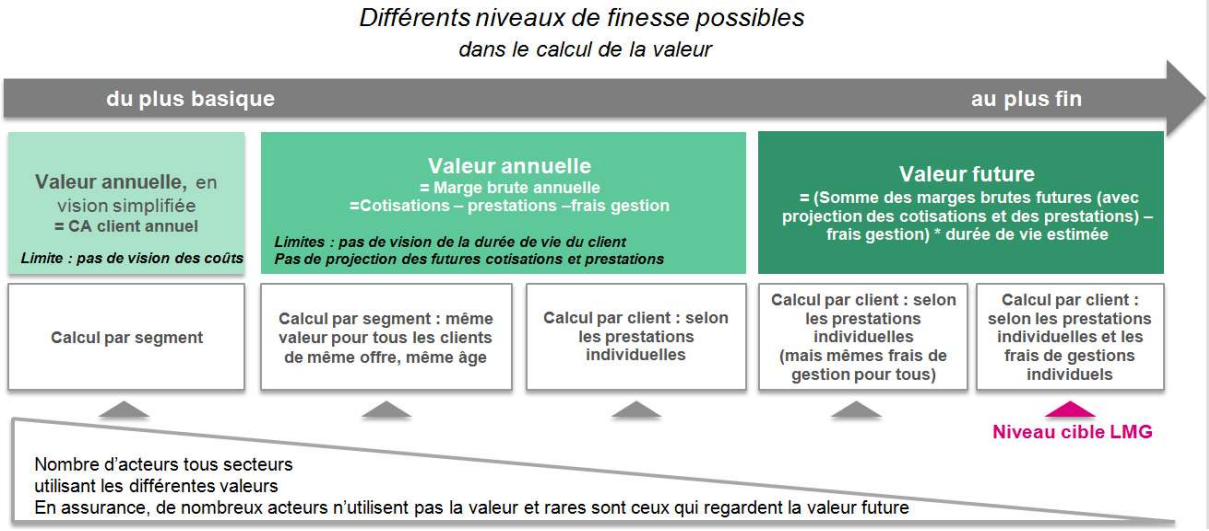
VCT = VCP + VCA + VCF – coûts d'acq.

En plus de la valeur client « *in force* » qui concerne le client avec contrat en cours ou contrat clos, nous définissons également :

- la valeur client « prospect » quand le contact, suspect qualifié, en est au stade du devis. Cette VC répond à la question « Quel prospect à contacter en premier ? »
- la valeur client « suspect » quand le contact, prospect non qualifié, est à l'aube de son parcours d'achat, qu'il n'a pas encore décidé si oui ou non, il allait se décider à acheter. Cette VC répond à la question « Quel effort marketing pour ce contact ? »

Ces deux types de valeur client sont hors du cadre du présent mémoire. Mais ils figurent dans le plan de déploiement de la VC au sein de l'entreprise à court/moyen terme.

Notre ambition : permettre une personnalisation grâce à un modèle de valeur individuelle




### 2.1.3. Un puissant outil marketing

Depuis toujours, les assureurs ont recherché une clientèle sélectionnée uniquement en fonction de leur niveau de risque. Les clients étaient exclusivement évalués en fonction de leur sinistralité potentielle ou constatée.

Depuis peu, on observe l'émergence d'une prise de conscience de l'intérêt de fidéliser les clients et plus particulièrement les « meilleurs clients » - et donc de définir qui sont les « meilleurs clients » -.

Les principaux assureurs traditionnels ont investi dans la modélisation et l'exploitation de la valeur client, avec pour objectifs de :

- Optimiser les dépenses commerciales et marketing
- Enrichir la base de connaissance clients
- Mener une politique de rétention
- Mettre en place un parcours client multicanal adapté au profil de chaque segment

C'est-à-dire, **au centre des processus opérationnels de l'entreprise, de définir des stratégies de relation client proportionnées à la marge dégagée par chaque client ou segment de clients.**

Aujourd'hui, les assureurs font face à de nouveaux challenges :

- Evolution de la réglementation (loi Hamon, ANI 2013, DDA 2017, GDPR[10] 2018, …)
- Standardisation de l'offre (contrat responsable, demain les 3 contrats types « Macron » ?)
- Volatilité des consommateurs
- Contexte durable de taux bas
- Apparition de nouveaux distributeurs (bancassureurs, Assurtech[11], …)
- Défis technologiques (Digital, Intelligence Artificielle, …)

L'impact sur le modèle économique de l'assurance est déjà amorcé : guerre des prix, baisse des marges, baisse des produits financiers et augmentation du churn.

Demain, les assureurs seront confrontés à une concurrence plus « violente » : les « GAFA ». Ces géants de l'Internet, probablement par le biais de partenariats bien sentis, pourraient développer des parts de marché significatives grâce à leur maîtrise de la relation et des données clients.

Face à ces défis, les assureurs ont programmé des investissements massifs dans le digital et le multicanal, pour bâtir avec leurs clients une relation dans la durée. Mais ces investissements nécessitent d'être orientés en fonction des segments de clients visés et de leur rentabilité attendue. Définir une stratégie pour demain implique donc pour les assureurs de :

- Doser leurs investissements en conquête
- Cibler des segments de clientèle rentables
- Adapter leurs offres
- Piloter de façon volontariste les canaux de distribution en fonction de la valeur de leurs clients et de leurs parcours

Il existe toutefois de nombreux freins au développement de la valeur client : La VC est une notion relativement récente en assurance (90's). Son développement est lent car il se heurte à de nombreux facteurs de résistance liés aux spécificités des entreprises d'assurance :

- La prépondérance de la rentabilité technique et financière : le S/P est l'outil de pilotage
- L'organisation des entreprises en silos : il est donc difficile d'obtenir une vision globale de l'assuré
- La vision à court terme du pilotage des compagnies d'assurance dont les 2 critères principaux sont le ratio combiné et le CA
- La perte d'informations sur le client lié au recours aux réseaux intermédiés
- Des Systèmes d'Informations décentralisés qui communiquent mal entre eux.

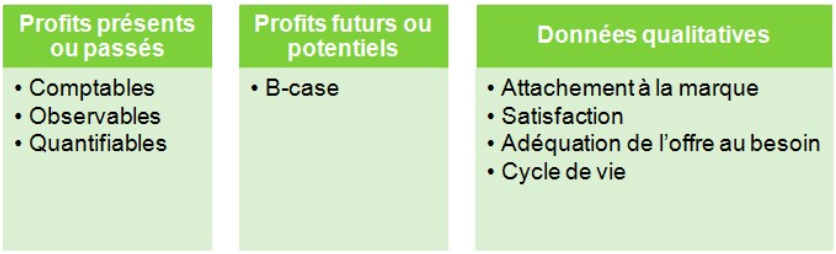
Néanmoins, ma conviction est que le pilotage par la valeur client est la clé de la réussite. Elle doit alimenter les décisions stratégiques marketing et commerciales : Quelles produits, pour quelles cibles, à quel moment, et surtout par quel canal ?

[10] http://www.eugdpr.org/
[11] Start-up disruptives comme *Oscar*, *Lemonade*, *Trov*, *Alan*, *Otherwise*, …

## 2.2. Les bénéfices d'un pilotage par la Valeur Client

### 2.2.1. Postulat de départ

La conquête d'un nouveau client représente non pas un coût qu'il faut réduire mais un investissement qui se traduit par un profit négatif à court terme et qu'il convient de rentabiliser.




Nous devons garder à l'esprit que l'objectif ultime d'une politique marketing fondée sur la valeur client demeure de maximiser non le CA mais la marge de l'entreprise. Loin de demeurer l'apanage des actuaires, la valeur client prouve ainsi qu'elle se situe bien au cœur de la stratégie des entreprises dès lors que l'on traite de politique commerciale, de marketing client, de service client ou de relation client omnicanale.

En assurance individuelle, la question qui se pose s'apparente donc à un problème d'optimisation sous contrainte. En effet, compte tenu de l'inversion du cycle de production et des coûts de conquête élevés, la problématique se résume ainsi : Augmenter la durée de vie des clients, notamment en détectant les clients à potentiel de multi-équipement, tout en trouvant un équilibre entre développement commercial et rentabilité technique.

Pour optimiser l'allocation des ressources, la valeur client semble l'indicateur le plus pertinent. Il doit prendre en compte à minima : la rentabilité des produits détenus par un client, leur probabilité de comportement futur et leur potentiel maximum.

Nous n'oublierons pas que la valeur client est un indicateur nécessaire mais pas suffisant. D'autres critères de segmentation devront également être étudiés, notamment à travers des études qualitatives et des « focus groupes », pour développer l'activité et fidéliser les clients.

Enfin, une démarche de valeur client s'inscrit dans un projet global entreprise. C'est en effet un excellent moyen de fédérer l'ensemble des collaborateurs de la société autour d'un projet porteur de sens, à condition de :

- Communiquer clairement
- Définir des objectifs atteignables
- Construire des indicateurs partagés par tous
- Etablir un modèle simple à comprendre
- Et surtout, mener ce projet en cohérence avec la stratégie de l'entreprise

Avant tout investissement dans ce domaine, il est donc indispensable de définir l'objectif principal visé par l'entreprise. La sophistication du ou des modèle(s) et la définition des indicateurs et des

facteurs clés de succès dépendront ensuite des données disponibles dans le Système d'Information de l'entreprise.

*Source : Annie Dillard, Est-il encore pertinent pour les assureurs d'investir dans la valeur client ? »*

### 2.2.2. Quelle place dans la stratégie ?

- La VC n'est pas un remède à la faible croissance d'un marché arrivé à maturité. C'est en revanche un facteur d'optimisation des investissements pour piloter la croissance des entreprises.
- La VC est une réponse adéquate face à la baisse de rentabilité annoncée car elle autorise une meilleure allocation des investissements commerciaux.
- La VC est un bon moyen de se défendre des attaques de la concurrence, grâce à un pilotage des investissements en conquête et en fidélisation sur des profils ciblés.
- La VC est un moyen d'identifier les clients présentant un risque de volatilité ; ce qui permet de mettre en place des programmes de rétention adaptés à chaque profil.
- La VC ne répond pas aux nouvelles attentes des clients en termes de personnalisation et service ; ce que permet en revanche la mise en place d'une relation digitale et/ou multicanale en fonction de la VC.
- Ni le pilotage par la VC, ni la mise en place de nouveaux modes de relation ne constituent une réponse suffisante. La menace d'une concurrence élargie à de nouveaux opérateurs extérieurs au secteur de l'assurance nécessite une réflexion prospective plus complète.

Dès lors, un temps passé de mode, le calcul de la valeur client dans l'assurance redevient donc une opportunité : le pilotage par la valeur client contribuera vigoureusement à la bataille que devront mener les assureurs pour conserver leurs parts de marché. En particulier la valeur client devrait être identifiée comme une excellente alternative au ratio combiné (ou S/P) pour piloter le risque d'une société d'assurance.

Grâce au développement de la relation multicanale, les modèles de valeur client (hier exclusivement fondés sur le « Small Data »), devraient demain être enrichis de nouvelles données (Big Data, Smart Data, DMP[12], Open Data …) de plus en plus qualitatives et comportementales, et de moins en moins structurées.

[12] Data Management Plateform

### 2.2.3. Les objectifs opérationnels

Rappel des enjeux :

- Maximiser la valeur des clients en parc en définissant une politique de fidélisation/rétention, différenciée sur les clients les plus rentables et permettant de retenir les meilleurs clients
- Optimiser la valeur des nouveaux clients pour une prospection plus efficace via l'Outil de Gestion de Campagne (OGC) et la DMP en identifiant les profils jumeaux[13] de nos meilleurs clients, permettant de cibler et recruter les meilleurs profils

**But ultime : Permettre une meilleure allocation des ressources répondant à des objectifs stratégiques et opérationnels en activant une stratégie proportionnée à la marge dégagée par chaque segment de client**

Pourquoi ? Pour qui ?

- Une utilisation opérationnelle pour les conseillers mutualistes (CM), les chargés de fidélisation (CFA) et la gestion de la relation client (GRC)
- Une vision détaillée pour définir une stratégie au niveau des directions Technique, Financière, Marketing stratégique, Relation Client et Développement

### 2.2.4. Les bénéfices pour l'entreprise

Une entreprise dotée d'un tel outil serait, en conséquence, enfin en mesure de conquérir les nouveaux clients dans une logique de rentabilité grâce à des politiques d'acquisition (promotions, mix canal) et des politiques d'offres adaptées, ainsi que piloter la fidélisation par l'intermédiaire d'un programme relationnel, d'une expérience client et d'une politique de rétention différenciés, tout en développant une stratégie de multi-équipement ajustée.

Nous bénéficierons alors de ce fait :

- D'une meilleure connaissance des clients
- De ciblages plus performants
- D'une allocation des ressources bonifiée
- D'une productivité améliorée
- D'une rentabilité supérieure

[13] Via la technique dite « look alike »

### 2.2.5. Les bénéfices pour les clients

Le pilotage par la valeur client conduira mécaniquement à davantage de personnalisation qui sera profitable à nos clients par le biais :

- D'une meilleure mutualisation
- De récompenses pour leur fidélité
- D'une stratégie d'accueil différenciée
- D'une baisse des sollicitations commerciales non choisie

### 2.2.6. Exploitation souhaitée par l'entreprise

Opérationnellement, il ne faut pas concevoir le résultat d'un modèle de valeur client comme un output unique d'une métrique issue d'un algorithme complexe ; mais plutôt la mise à disposition vers la Direction Marketing et les collaborateurs en contact avec le client, d'un ensemble d'informations utiles pour faciliter la relation assureur - assuré. Ces indications prennent la forme d'au moins 6 types de données qui sont complémentaires entre elles. Et chacune d'elles apportent des éléments essentiels pour les différents cas d'usage. Examinons quelques hypothèses possibles d'utilisation :

*1 / Valeur Client Totale par segment*

Pour la Direction Marketing :

**Politiques d'acquisition :** définition du mix-canal optimum, recherche de look-alike via DMP, politique de promotions par canal et/ou par segment de clientèle

**Politique d'offres :** bonus fidélité, …

Pour les conseillers mutualistes :

**Mise en œuvre du programme relationnel** segmenté selon le cycle de vie client…

*2 / Valeur Client Future par client*

Pour la Direction Marketing :

**Politique de fidélisation :** programmes de rétention préventive (action anti-churn), réactif (Stop-résil), multi-équipement, définition du programme relationnel segmenté selon le cycle de vie client…

**Ciblage et communications** personnalisées adressées directement aux clients, prospects ou via des opérations spécifiques adressées aux réseaux CM et GRC

*3 / Valeur Client Actuelle par client*

Pour les conseillers relation client :

**Traitement premium** et file prioritaire en Relations Adhérents - traitement prioritaire des segments de clientèle à haut ou très haut degré de solidarité

*4 / Risque pré-contentieux par client*

Pour la Direction Développement :

**Exclusion du ciblage** des campagnes fidélisation et multi-équipement

Pour les conseillers relation client :

**Actions de recouvrement**

*5 / Score de risque de churn par client*

Pour les directions Marketing et Relation Client :

**Ciblage et mise en action des opérations préventives** Anti-Churn et Stop-Résil

*6 / Potentiel de multi équipement par client (scores par produit)*

Pour la Direction Développement :

**Ciblage** campagne multi-équipement

Pour les conseillers relation client et mutualistes :

**Rebond commercial** sur appels entrants

**Actions** de multi-équipement

Le but *in fine* est donc quadruple : **Conquérir** (arbitrer entre plusieurs cibles de recrutement), **Attacher à la marque** (meilleure allocation des budgets fidélisation), **Retenir les clients** (meilleure allocation des budgets rétention), et **Multi-équiper / Augmenter les garanties** (meilleur ciblage). Dans cette optique, et afin de rendre chacun de ces cas d'usage le plus facile possible, nous envisageons non seulement de pousser ces indicateurs dans le SI mais aussi d'implémenter un algorithme d'« aide à la décision » ou « aide à l'action ».

Par exemple, pour s'adapter aux contraintes d'un appel entrant, l'information pourra être de préférence affichée sous forme d'actions à réaliser, paramétrée en fonction des 6 types de données clients évoqués plus haut.

Plus généralement, le CRM sera, certainement à très court terme, enrichi, pour chacun de nos clients en gestion directe, d'un bandeau qui pourrait prendre l'une des deux formes suivantes :


### 2.2.7. Lien entre fidélité et valeur client

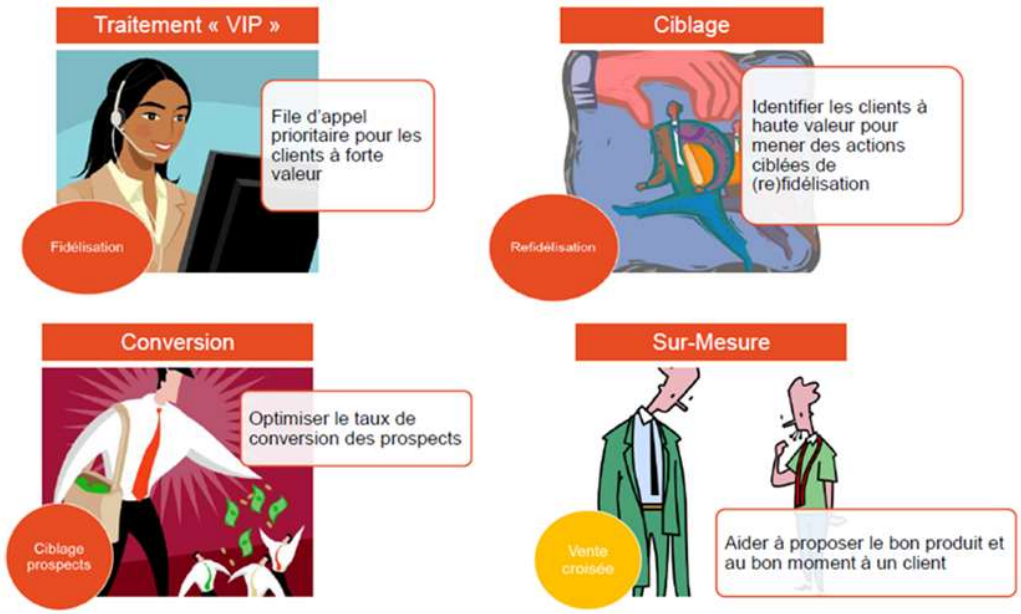
*« La fidélité est le désir de la part d'un client de continuer à faire du business avec un fournisseur donné dans le temps ».*[14]

Deux types de fidélité coexistent : la fidélité **comportementale** (cartes client, réductions, …) et la fidélité **attitudinale** (attachement du client à l'entreprise, à la marque). En présence de la seule fidélité comportementale, les consommateurs ont tendance à changer de fournisseur dès qu'une meilleure alternative se présente. Les deux types de fidélité doivent être cultivés mais c'est la 1ère qui est principalement touchée par les nouvelles réglementations. La 2nde est donc plus stratégique. Une minorité des clients génère la majorité des profits (Résultat bien connu de la loi de Pareto : 20% des clients sont à l'origine de 80% des profits). Il est donc indispensable d'identifier les clients à plus haute valeur pour les fidéliser, en priorisant l'investissement sur ceux-ci.

*Source : Ilaria Dalla Pozza & Lionel Texier – AssurMarketing 2015*
*« Pression tarifaire et enjeux de fidélisation soulevés par la loi Hamon : les bénéfices de la Customer Lifetime Value »*

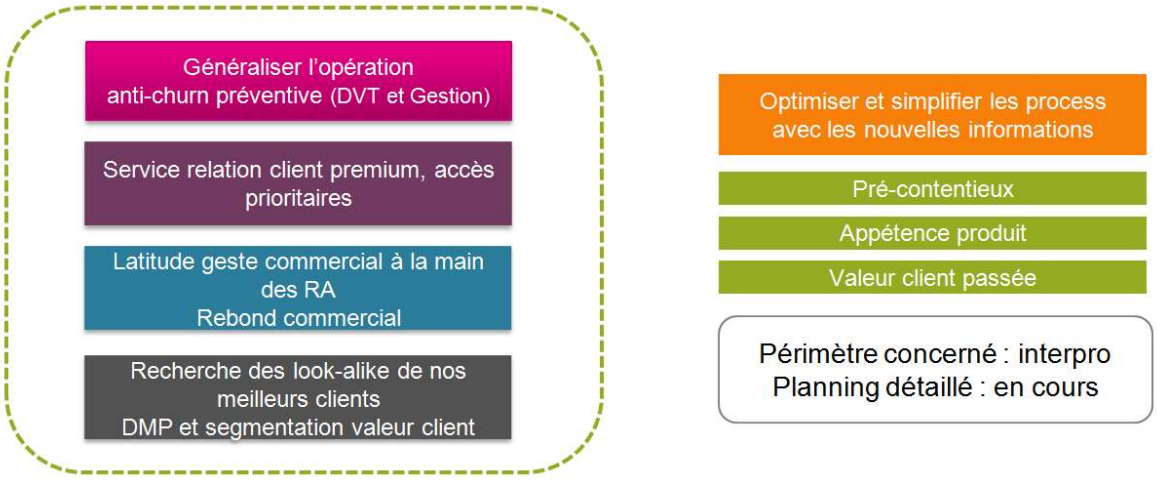
[14] Sargeant & West, 2001.

### 2.2.8. Quelques exemples d'utilisation de la Valeur Client en assurance




*Source : AssurMarketing*

### 2.2.9. Premières actions retenues



## 2.3. Conclusion partie II

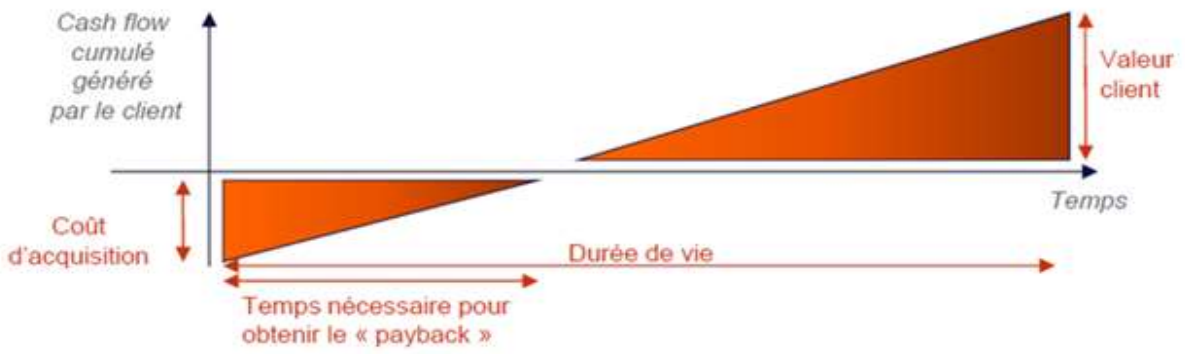
- **La valeur client est la somme des marges acquises et de la valeur actualisée de l'espérance des profits futurs d'un client.**
- **Elle évolue dans le temps et à mesure qu'on obtient de la nouvelle information.**
- **Elle aide à l'obtention d'une vision 360° du client.**
- **Elle permet une optimisation des budgets marketing acquisition et fidélisation sous contrainte de maximisation de la profitabilité client.**
- **Elle permet de définir et de mettre en œuvre une politique de fidélisation pertinente en priorisant les efforts à allouer pour chaque couple « segment de clients – canal ».**
- **Elle offre un cadre d'analyse particulièrement puissant pour obtenir un avantage compétitif dans un environnement de plus en plus concurrentiel.**
- **La valeur client permet de concilier les deux dimensions, profit et fidélité, en une seule mesure car elle est construite sur la base d'une espérance de marge et d'une probabilité de survie.**

# Partie 3 : Eléments constitutifs de la Valeur Client

*« Chaque bonne réalisation, grande ou petite,*
*connait ses périodes de corvée et de triomphe ;*
*un début, un combat et une victoire »*

Gandhi

Le calcul de la valeur client s'établit principalement autour de trois volets : la durée de couverture dans le portefeuille, la contribution annuelle à la marge de l'entreprise et les frais.




Pour qu'un client soit rentable pour l'assureur, la contribution cumulée doit, sur l'ensemble de la durée de vie, dépasser les coûts variables non récurrents.

## 3.1. Les scores

Le scoring est une technique qui permet d'affecter un score ou une note à un client ou prospect. En marketing, le score obtenu traduit généralement la probabilité qu'un individu réponde à une sollicitation ou appartienne à la cible recherchée. Il mesure donc l'affinité pour l'offre potentielle. C'est le score d'appétence.

Il existe d'autres types de score (liste non exhaustive) : score de risque, score de d'octroi de crédit, score de recouvrement, score d'attrition…

13/20 pour un devoir d'Histoire-Géo est une note donc également un score. Aux US, le Credit Score qui reflète la crédibilité du payeur va de 330 à 850. Certaines DMP (Data Management Plateform) affectent des scores sous forme d'arbre, selon le parcours d'un individu sur un ou plusieurs site(s) internet. Chaque nœud symbolise un critère ; une branche donne une note. Le score est la somme des notes de chaque branche. Il peut varier par exemple de -5 à +10.

Pour ma part, le score est une probabilité : Score = P $\in$ [0 ; 1]. Dans ce qui suit, nous nous intéressons principalement à trois types de score :

- le score d'appétence qui mesure la probabilité d'un client d'être intéressé par un produit de multi-équipement ;
- le score d'attrition, qui mesure la probabilité de quitter le portefeuille par démission volontaire ou départ vers un contrat collectif suite à un changement de situation professionnelle (ex : départ « ANI ») ;
- le score de risque de pré-contentieux qui mesure la probabilité pour un adhérent de rencontrer un incident de paiement entraînant un processus de contentieux.

### 3.1.1. La durée de vie

La durée de vie passée d'un assuré est calculée à partir des informations de gestion contenues dans le Système d'Information de l'entreprise. Factuellement, pour les contrats clos : différence entre date de fin de couverture et date de début de couverture de l'assuré ; pour les contrats en cours : différence entre date du jour et date de début de couverture de l'assuré. Pour des raisons pratiques, cet indicateur est ramené à une durée annuelle, en divisant donc la durée de vie en jours par 365,25.

La durée de vie future d'un client s'entend comme l'espérance de la durée de la relation commerciale entre l'entreprise et le client. Cette espérance, aussi identifiée comme étant la probabilité de survie d'un client dans le portefeuille, ne peut être qu'estimée, par définition.

Habituellement, pour réaliser ces estimations, on utilise une modélisation par le biais des lois de survie. Notamment, l'estimateur non paramétrique Kaplan-Meier ou le modèle semi-paramétrique de Cox permettent d'obtenir des courbes de référence sur les probabilités de survie des assurés. Ces fonctions prennent le plus souvent la forme de **lois exponentielles** dont les paramètres dépendent des variables exogènes propres à chaque groupe d'assurés étudié. Toutefois, cet aspect a déjà été traité[15] par Marie Piffault dans le cadre de son mémoire actuariel.[16]

Une autre possibilité aurait consisté à utiliser une approche paramétrique proposée par Hardie et Fader en 2014, relayée en France par Lionel Texier de *Risk and Analysis* : modéliser le taux de rétention d'un assuré par un mélange **Bêta-Géométrique**.

Probabilité de résilier en année t :

$$P\ (T = t\ |\theta) =\ \theta\ (1-\theta)^{t-1} \text{ pour t = 1, 2 … T}$$

$\theta$ représente la probabilité annuelle de résilier, et suit une loi Bêta. Elle est propre à chaque assuré. Nous devrions donc plus rigoureusement écrire pour l'assuré i, $P\ (T_i = t\ |\theta_i)$.

La fonction de survie se définit ainsi :

$$S(t|\theta) =\ P(T > t|\theta) = (1-\theta)^t, \text{ t} \geq 1$$

Espérance mathématique :

$$E_\Theta = [P(T = t|\Theta = \theta)]$$

$$= \int_0^1 P(T = t|\Theta = \theta)\ g\big(\theta\big|(\gamma,\delta)\big)d\theta$$

$$Avec\ g\big(\theta\big|(\gamma,\delta)\big) = \frac{\theta^{\gamma-1}\ (1-\theta)^{\delta-1}}{B(\gamma,\delta)}$$

$$E\ (\Theta) = \frac{\gamma}{\gamma+\delta}$$

Les paramètres d'un tel modèle peuvent être estimés par le maximum de vraisemblance. On estime alors sur la base d'un jeu de données empiriques quelles sont les valeurs des paramètres qui maximisent la vraisemblance de ce modèle appliqué à nos données. Il s'agit ensuite de faire cette

[15] Et très bien traité !

[16] « La valeur client comme mesure de la rentabilité d'un portefeuille d'assurance santé individuelle ».

opération pour chaque segment de client (par exemple : niveau de garantie x tranche d'âge x canal d'entrée) pour calibrer le taux de rétention.

Une extension possible est de remplacer la loi Bêta-Géométrique par la **loi Bêta-Discrète-Weibull**, en substituant la loi discrétisée de la Weibull à la loi géométrique. Elle apporte une réponse à la limite de la loi Bêta-Géométrique : pas de point de discontinuité dans la courbure ni de changement de convexité. Cette modification revient à ajouter un exposant *c* au nombre de période dans la loi géométrique. Si $c$ est supérieur à 1, cela accroit le taux de résiliation avec le temps, si $c$ est inférieur à 1, cela le diminue. Ainsi les courbes de taux de rétention peuvent prendre des formes plus variées, avec un coude ou une inflexion.

La fonction de survie devient alors $S(t|\theta, c) = (1-\theta)^{t^c}, t \geq 1$

Je privilégie finalement une autre solution, davantage en adéquation avec l'expertise que j'ai développé ces deux dernières années : l'estimation de la survie dans le portefeuille à travers **un modèle de scoring**. Cette technologie a l'avantage d'être scalable, c'est-à-dire qu'elle présente la capacité à s'adapter facilement à des importantes variations d'ordre de grandeur des jeux de données. Ses capacités et ses performances ne seront pas de trop altérées lors du passage d'un prototype à un produit industrialisé.

Notons que la durée de vie future dépend d'au moins trois mesures à estimer séparément, qui recouvrent distinctement tous les cas de sortie de portefeuille : la probabilité de décès, la probabilité de démission volontaire et la probabilité de radiation.

La première mesure est obtenue par le biais des tables de mortalité LMG certifiées. La deuxième est déduite de la modélisation du risque de churn. Tandis que la troisième peut être estimée à partir de la modélisation du risque de passer en pré-contentieux. En effet, nous verrons par la suite que la probabilité de radiation et le risque de pré-contentieux sont quasi parfaitement liés.

De manière générique, la probabilité de survie annuelle s'obtient alors ainsi :

*P = (1 – taux de mortalité) x (1 – score de démission volontaire) x (1 – P [radiation])*
*avec P [radiation] = [0.65 x exp (-0.11 x Ancienneté) x score de pré-contentieux]*[17]

### 3.1.2. L'appétence aux produits de multi-équipement

Dans la même section, nous traitons également de l'appétence probable de chaque client à chacun des 4 produits prévoyance interprofessionnels : Prd Prev 1 (Garantie Accidents Corporels), Prd Prev 2 (Protection Juridique), Prd Prev 3 (Tempo Décès) et Prd Prev 4 (Assurance Décès Accidentels).

En effet, pour modéliser ces appétences probables, nous utilisons la même méthode du scoring, et les fondements théoriques sont les mêmes que pour les mesures constitutives de la survie.

[17] Voir paragraphe 6.1.2.

### 3.1.3. La méthodologie générale

Nous avons donc 6 scores à construire :

1. Appétence à la vente additionnelle Prd Prev 1
2. Appétence à la vente additionnelle Prd Prev 2
3. Appétence à la vente additionnelle Prd Prev 3
4. Appétence à la vente additionnelle Prd Prev 4
5. Risque de churn (ou démission volontaire)
6. Risque de pré-contentieux

L'idée générale consiste à challenger la méthode traditionnelle, la plus utilisée dans ce contexte, i.e. la régression logistique, avec plusieurs autres méthodes d'analyses prédictives[18]. Pour chaque score, nous mettons en place trois processus de validation, par comparaison des courbes ROC[19] (procédure AUC, pour *Area Under the Curve*) :

1. Validation temporelle (2016 vs 2015 vs 2014)
2. Validation spatiale
   [échantillon test (30% de la pop) vs échantillon d'apprentissage (70%)]
3. Validation type de modèle (Régression logistique vs Arbres de décision, Réseaux de neurones, GBM, …)

La population à scorer est la population assurée à fin de mois précédent (dernier mois entièrement chargé dans l'entrepôt de données), limitée aux assurés principaux (`CD_GRP_ASS = 'ASSPRI'`). Par ailleurs, on sélectionne les adhérents éligibles à l'apprentissage en N, c'est-à-dire ceux qui sont couverts à date. Enfin, on exclut les adhérents dont la souscription a concerné simultanément le produit « Santé » et le produit « Prévoyance » en question, car ils ne sont plus à multi-équiper.

On choisit pour modalités de référence, les modalités les plus fréquentes de toutes les dimensions en entrée puis on détermine les critères de sélection des dimensions discriminantes : procédure « *Forward* » et seuil d'entrée dans le modèle à 10%. La procédure « *Forward* », i.e. on part de rien et on ajoute dans le modèle, au fur et à mesure, les dimensions dont les p-value sont les plus petites, est la procédure de sélection la plus indiquée dans le cas où les dimensions en entrée sont nombreuses. En effet, les procédures « *Backward* », i.e. on part de tout et on enlève dans le modèle, au fur et à mesure, les dimensions dont les p-value sont les plus grandes, et « *Stepwise* », i.e. allers-retours « *Forward*, *Backward* », - on peut retrancher une variable à chaque étape si son pouvoir discriminant est contenu dans une combinaison des nouvelles variables -, très performantes lorsque les variables disponibles sont peu nombreuses, sont plus coûteuses en temps, ce qui altère la performance globale du modèle. D'autre part, le plus souvent, la p-value max est fixée à 0.05 mais ici nous souhaitons qu'un maximum de dimensions participe à la constitution du score, d'où une p-value max à 0.1.

[18] Voir Partie 4 : l'analyse prédictive
[19] Voir paragraphe 4.2.2.2. La courbe ROC et le critère AUC

Puis, nous étudions avec attention, les outputs traditionnels d'un modèle prédictif[20]: la significativité globale du modèle, le calcul des odds ratios, le cas échéant le test de Hosmer et Lemeshow, et bien sûr le calcul de l'aire sous la courbe.

Enfin, nous procédons à l'application du modèle sur l'échantillon à scorer, en intégrant deux conditions sur les bornes du score : il doit être au maximum égal à 0.5 car au-delà cela ne paraît pas crédible d'un point de vue métier, et s'il est inférieur à 1/10000 alors on considère qu'il vaut 0.

En dernier lieu, nous réalisons les contrôles de cohérence par rapport à l'attendu puis nous créons la table finale des scores : une ligne par client et 7 colonnes : l'identifiant et les 6 scores.

## 3.2. Les contributions périodiques

La contribution annuelle individuelle à la marge de l'entreprise est, pour une année donnée, la somme des cotisations HT acquises reçues de la part d'un adhérent, à laquelle on ôte les chargements, i.e. les frais de gestion estimés, et la somme des prestations payées (ou restant à payer) à l'adhérent.

### 3.2.1. La contribution passée

Elle est relativement simple à déterminer puisque tout est connu ou presque : il n'y a pratiquement plus d'aléa.

Rigoureusement parlant, pour une période de survenance passée donnée, il demeure toujours des prestations à payer pour les sinistres IBNR[21], c'est-à-dire les sinistres survenus mais pas encore enregistrés dans le système de gestion. Ces prestations prennent la forme de provisions, les PSAP[22], qui sont intégrés aux comptes techniques pour obtenir une bonne estimation de la charge totale des prestations. Elles sont inconnues mais l'étude des triangles de liquidation (ventilation des montants de sinistres par année-mois de survenance et année-mois de paiement) nous apprend qu'elles peuvent être estimées, par la loi des grands nombres, avec peu de variance.

C'est particulièrement le cas pour les risques dits courts, en santé notamment. En effet, l'étude de l'historique des paiements nous indique qu'au bout de six mois après la fin d'une période de survenance donnée, la quantité de prestations restant à payer devient marginale. En outre, un organisme de complémentaire santé, intervient le plus souvent après intervention de la Sécurité Sociale. Or la SS ne rembourse normalement plus les prestations survenues en année N et connues au-delà du 31/12/N+2. C'est le délai de forclusion. Si bien qu'à partir de N+3, les remboursements deviennent rares et ne concernent qu'en principe des régularisations de paiement.

Toutefois, autant, globalement pour un ensemble d'assurés, les PSAP sont facilement modélisables et intégrables aux comptes de budget et d'inventaire avec peu d'aléa, autant, individuellement, pour

[20] Voir Partie 4 – L'analyse prédictive
[21] *Incurred But Not Reported*
[22] Provisions Pour Sinistres A Payer

un assuré donné, il est beaucoup moins aisé d'estimer sa propre part de PSAP. Pour beaucoup d'entre eux, elle est en pratique nulle, tandis que pour un petit nombre, elle peut être conséquente - l'ensemble d'une hospitalisation à payer par exemple : des frais de séjours, des honoraires de chirurgiens et/ou d'anesthésistes, des Forfaits Journaliers Hospitaliers, des frais de chambre particulière, à rembourser peuvent rapidement atteindre des sommes représentants plusieurs dizaines de milliers d'euros -.

Nous pourrions envisager d'appliquer à chacun des assurés, le coefficient de PSAP calculé pour l'ensemble du groupe. Nous aurions alors une bonne estimation du volume global de prestations pour une année donnée, mais fausse tête à tête. Nous pensons qu'agir ainsi dégraderait la pertinence du modèle de valeur client. Nous préférons donc nous abstenir et nous contenter au niveau de la tête assurée des prestations payées à date. Finalement, ces prestations survenues mais encore inconnues finiront bien par être payées un jour, et comme le modèle est rejoué tous les mois, l'exécution régulière des requêtes viendra mettre à jour automatiquement la valeur client de chacun. Il faut prendre cette hypothèse comme un *lag* inéluctable.

En conséquence, la contribution passée se résume, pour l'interpro, à :

$\sum$ (cotisations acquises Santé nettes de chargements) - $\sum$ (prestations Santé payées)
+ $\sum$ (cotisations acquises Prévoyance nettes de chargements) - $\sum$ (prestations Prévoyance payées)
+ commission distribution * $\sum$ (cotisations acquises des produits partenaires),

calculée annuellement de date d'ouverture des droits à fin de mois précédent, l'année en cours étant donc incomplète[23],

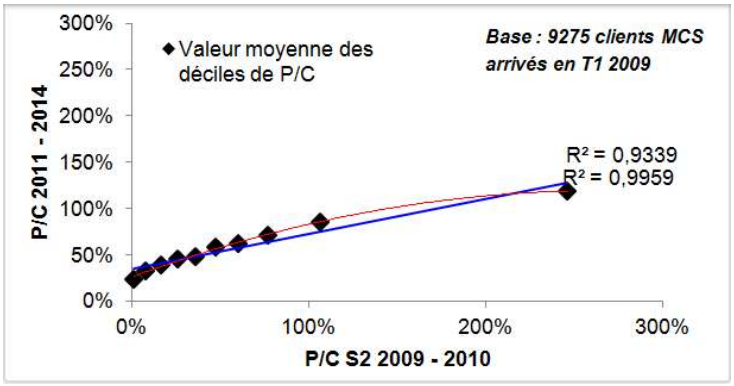
avec : les chargements supposés financer les frais de gestion
et commission de distribution = 17.5% depuis 2013, et toujours d'actualité en 2017.

### 3.2.2. La contribution future

#### *3.2.2.1 Des bons fondamentaux*

Contrairement à la contribution passée, la contribution future est par essence inconnue, il faut donc l'estimer. Elle n'est pas non plus totalement aléatoire car de nombreux constats nous conduisent à penser qu'elle dépend fortement d'éléments déjà connus. En particulier, pour les adhérents dont l'ancienneté est suffisante, on remarque qu'il existe une forte corrélation entre le P/C des 18 premiers mois de couverture et le P/C des 3 années suivantes, ceci valant statistiquement pour un groupe d'assurés et non pas pour une seule personne protégée.

[23] Sauf si l'on se trouve entre le 2ème mardi de janvier et le 2ème mardi de février. Explication au 5.2.1.




Le P/C des 18 premiers mois permet de situer avec une bonne fiabilité le client dans un segment de P/C long terme et donc de prendre en compte sa valeur individuelle.

Pour les adhésions récentes, on remarque de la même façon, que ceux-ci ont un comportement de consommation proche de celui des adhérents du même segment (niveau de garantie **X** tranche d'âge) lors de leurs débuts de période de couverture. Nous avons donc en stock des éléments suffisants pour nous permettre de modéliser au mieux la contribution future des assurés présents en parc à date. Toutefois, il est nécessaire de poser un cadre car en pratique de nombreux événements exogènes peuvent venir bouleverser le comportement de consommation des assurés. Nous établissons donc 4 hypothèses fortes concernant la projection des flux futurs du modèle :

- Une constance dans le futur, des caractéristiques socio-démographiques et administratives de chaque assuré
- Une stabilité de l'environnement économique et réglementaire du secteur de l'assurance santé
- Des frais d'administration, d'acquisition et de gestion, identiques pour l'ensemble des assurés.
- Des montants des prestations futures modélisés indirectement via la théorie de la crédibilité

#### *3.2.2.2 Evolution des cotisations*

En ce qui concerne le risque santé, d'une manière générale, l'évolution d'une année sur l'autre de la cotisation d'un assuré en individuel dépend de 5 éléments :

- L'évolution liée à l'âge (+2% environ)
- L'inflation médicale
- Un éventuel rattrapage du P/C
- L'évolution réglementaire (PLFSS[24])
- L'éventuelle évolution des garanties

Nous raisonnons à périmètre réglementaire et garanties stable, dans un environnement d'inflation faible à moyen/long terme et sans dérive du P/C. Donc, dans un premier temps, nous retenons l'hypothèse simplificatrice mais réaliste d'une évolution annuelle des cotisations de **+4%** pour tous les assurés. Cette hypothèse sera « *shockée* » en fin de modèle.

[24] Projet de Loi de Financement de la Sécurité Sociale

Concernant la prévoyance, les barèmes HT n'ont pas évolué depuis le lancement en raison de bons P/C et principalement de manque de recul. Mais la DT ne s'interdit pas de le faire, en cas de dérive du risque. Nous prenons l'hypothèse d'une évolution annuelle de **+1%** sur tous les produits.

#### *3.2.2.3 Projection du P/C : le modèle Bühlmann-Straub*

Le modèle Bühlmann-Straub est un des outils de la théorie de la crédibilité. Cette dernière puise son fondement dans la théorie de la fluctuation limitée[25]. A l'origine, des entreprises américaines, General Motors (une très grande entreprise) et Tucker (une petite) qui constatent que leurs primes d'expérience sont sensiblement inférieures aux primes respectives qui leur sont demandées.

Cette première approche de crédibilité est conçue comme un outil de tarification permettant d'envisager des primes différentes pour des risques a priori homogènes. En embarquant la sinistralité passée de chaque entreprise, le modèle prend en compte le phénomène d'inertie propre à chaque risque, et peut ainsi proposer une tarification adaptée et équitable pour un ensemble de contrats d'assurance.

Notion de prime de crédibilité :
En théorie de la crédibilité, la prime pure $\pi$ de l'assuré i se définit par :

$$\pi_i = z_i\, S_i + (1 - z_i)\, m_k$$

avec :
$S_i$ : la consommation moyenne passée de l'assuré
$m_k$ : la consommation moyenne passée du groupe homogène k auquel appartient l'assuré i
$z_i$ : le facteur de crédibilité

On parle de crédibilité totale lorsque z vaut 1. La prime pour la période suivante ne dépend alors plus qu'exclusivement de la sinistralité passée. $\pi_i = S_i$
Lorsque z vaut 0, la crédibilité est nulle (c'est le cas en général, lorsque l'on a affaire à un nouveau contrat) : la prime de l'assuré est égale à la moyenne des primes de son groupe. $\pi_i = m_k$
Lorsque $z \in ]0\ ;\ 1[$, on dit que la crédibilité est partielle, et la prime de l'assuré s'interprète donc comme une moyenne pondérée d'une prime individuelle et d'une prime collective.

En assurance santé, la réglementation interdit aux complémentaires de fixer la prime de l'assuré en fonction d'informations concernant l'état de santé, et en particulier de sa consommation passée. Nous n'utiliserons donc pas cet outil pour déterminer la prime future des assurés mais pour estimer la sinistralité future en calculant des « P/C individuels », vus comme le résultat d'une pondération par le facteur de crédibilité, des rapports sinistres à primes passés de chacun des assurés et du P/C de leurs groupes respectifs.
Généralement, je n'aime pas cette notion de « P/C individuel ». Le P/C est un indicateur qui prend tout son sens pour mesurer l'équilibre technique d'un groupe d'individus, lorsque l'on peut - et doit - mutualiser ces risques entre eux. Nous allons toutefois tordre ce principe dans le cadre de ce

[25] Mowbray [1914] puis Whitney [1918]

mémoire, car je n'ai pas trouvé d'outil plus adéquat que la théorie de la crédibilité dans ce contexte où il est nécessaire d'évaluer les risques séparément.

Traditionnellement en assurance non-vie, et en particulier en assurance santé, l'outil le plus communément utilisé pour modéliser la sinistralité, mutualisée, est le Modèle Linéaire Généralisé. Il est particulièrement puissant quand il s'agit d'évaluer le risque d'un groupe homogène, mais ici dans un environnement de valeur client, appliquer des montants de prestations moyens « écrase de trop » la spécificité d'un risque en ramenant sa sinistralité future probable trop rapidement à la moyenne estimée pour son groupe.

En 1967, Hans Bühlmann, se démarque de l'approche bayésienne alors en vigueur, modèle souvent optimal mais impraticable, en ne spécifiant plus de distribution a priori du risque, ce qui permet de prendre en considération un portefeuille de risques hétérogènes, et formalise un cadre d'analyse robuste pour renouveler la théorie de la crédibilité.

Dans le modèle de Bühlmann, on note $X_{ij} = (X_{i1}, ..., X_{in})$, i = 1 à I, le vecteur des observations associées au risque i et $\Theta_i$ son profil de risque ; j = 1 à n, figure l'année j. Les I risques d'un même groupe k, défini au préalable, par exemple par la combinaison "Niveau de garantie x Tranche d'âge" sont a priori considérés homogènes.

Les hypothèses du modèle sont les suivantes :

(H1) : Les variables aléatoires $X_{ij}$ sont, conditionnellement à $\Theta_i = \theta$, indépendantes et identiquement distribuées selon une loi $F_\theta$ avec les moments conditionnels :

$$\mu(\theta) = E\,[X_{ij}|\Theta_i = \theta],$$
$$\sigma^2\,(\theta) = Var\,[X_{ij}|\Theta_i = \theta]$$

(H2) : Les couples $(\Theta_1, X_1)$, ..., $(\Theta_I, X_I)$ sont indépendants et identiquement distribués.

Le modèle de Bühlmann recherche la meilleure approximation linéaire du P/C d'un assuré. Pour un portefeuille tel que défini précédemment et respectant les hypothèses (H1) et (H2), la meilleure approximation linéaire non homogène du P/C, $\mu(\theta_i)$ (ou de $X_{i,n+1}$) est :

$$\mu(\theta_i) = X_{i,n+1} = z_i\,X_i + (1 - z_i)\,m$$

avec :

$$z_i = \frac{n}{n + \frac{\sigma^2}{\tau^2}}$$

$z_i$ représente le facteur de crédibilité associé au risque i.

$\sigma^2$ et $\tau^2$ sont les paramètres de structure. La somme des deux donne la variance totale du portefeuille. En effet[26] :

$$\text{Var}\,[X_{ij}] = \text{Var}\,[E[X_{ij}|\Theta_i]] + E\,[\text{Var}[X_{ij}|\Theta_i]]$$
$$\text{Var}\,[X_{ij}] = \text{Var}\,[\mu(\theta_i)] + E\,[\sigma^2\,(\theta_i)]$$
$$\text{Var}\,[X_{ij}] = \tau^2 + \sigma^2$$

La variance intra, $\sigma^2$, moyenne des variances conditionnelles, quantifie la part de la variabilité intrinsèque du P/C, i.e., le risque interne au risque individuel ; tandis que la variance inter, $\tau^2$, variance des moyennes conditionnelles, mesure l'hétérogénéité du portefeuille.

Le rapport de ces deux paramètres $\kappa = \frac{\sigma^2}{\tau^2}$ est appelé coefficient de crédibilité.
Le facteur de crédibilité peut aussi s'écrire, en multipliant numérateur et dénominateur par $\frac{\tau^2}{n}$ si $n \neq 0$ :

$$z_i = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}$$

qui tend bien vers 0 quand $n$ s'approche de 0, et qui tend vers 1 lorsque n tend vers $+\infty$ ; ce qui cadre bien avec l'interprétation intuitive du facteur de crédibilité : toutes choses égales par ailleurs, la crédibilité progresse avec le nombre d'années de couverture au risque, et inversement.
Cas $n = 0$, alors $z = 0$ : sans historique, pas de crédibilité, cela va de soi.

D'après (H2), les risques sont indépendants, $\text{Cov}\,\mu(\Theta_i), X_h) = 0$ si i ≠ h. L'estimateur de crédibilité de $\mu(\Theta_i)$ ne dépend donc que des observations associées au risque i.

Le modèle comporte toutefois une hypothèse peu réaliste :

$$\sigma^2(\Theta_i) = \text{Var}\,[X_{ij}|\Theta_i]$$

Il semble logique que la variance conditionnelle devrait être décroissante avec l'exposition du risque i. En 1970, Erwin Straub apporte sa contribution. Il généralise ainsi le modèle de Bühlmann en relâchant l'hypothèse de variance constante par l'introduction de poids, représentant la couverture au risque de chaque assuré. Le modèle devient Bühlmann-Straub. En 1980, W. Jewell améliorera de nouveau, en développant un modèle hiérarchique pour les portefeuilles segmentés.

Dans le contexte de valeur client, ce deuxième archétype est plus adéquat. En effet, certains assurés arrivent en cours d'année, et d'autres quittent le portefeuille avant la prochaine échéance (décès, radiation pour non-paiement ou départ vers un contrat collectif). Toutefois les principes déjà évoqués demeurent : dans la suite du mémoire nous traiterons d'un portefeuille de I risques et nous noterons $X_{ij}$ le ratio de sinistres à primes de l'assuré i pendant l'année j et $\omega_{ij}$ le poids associé, c'est-

[26] D'après la formule de décomposition de la variance :
https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_la_variance_totale

à-dire la durée de couverture de l'assuré i au cours de l'année j. Ce modèle peut être considéré comme le véritable « cœur » du modèle macro de Valeur Client.

Les hypothèses du modèle de Bühlmann-Straub :

Ce sont les mêmes que le modèle précédent à ceci près que la variance est désormais pondérée par la durée d'exposition au risque :

$$\sigma^2(\Theta_i) = \omega_{ij}\, \mathrm{Var}\,[X_{ij}|\Theta_i]$$

Bien entendu, $\omega_{ij}$ peut être égal à 0 si l'assuré i n'était pas couvert au cours de l'année j.
Par ailleurs, l'indépendance conditionnelle des variables $X_{ij}$ peut être relâchée en exigeant seulement la non-corrélation conditionnelle, c'est-à-dire : $\mathrm{E}\,[\mathrm{Cov}\,(X_{ij}, X_{ih}\,|\Theta_i)] = 0$ pour tout j ≠ h.

L'estimateur homogène de crédibilité est alors donné par :

$$\hat{\mu}(\theta_i) = z_i\, X_i + (1 - z_i)\, \hat{\mu}_0$$

avec :

$$z_i = \frac{\omega_{i.}}{\omega_{i.} + \frac{\sigma^2}{\tau^2}}$$

$$\hat{\mu}_0 = \sum_{i=1}^{I} \frac{z_i}{z..} X_i$$

et

$$z. = \sum_{i=1}^{I} z_i$$

Pour estimer $\hat{\mu}_0$, l'intuition nous conduit à utiliser la moyenne observée :

$$\bar{X} = \sum_{i=1}^{I} \frac{\omega_{i.}}{\omega..} X_{i.}$$

mais nous lui préférons la moyenne pondérée par les facteurs de crédibilité.

Remarquons que le modèle de Bühlmann-Straub est construit de telle façon que si la prime homogène de crédibilité avait été utilisée dans la passé, il y aurait eu équilibre entre les ratios sinistres à primes estimés et les ratios constatés :

$$\sum_{ij} \omega_{ij}\ \hat{\mu}(\theta_i) = \sum_{ij} \omega_{ij}\ X_{ij}$$

Sous les hypothèses du modèle, l'erreur quadratique moyenne de l'estimateur de crédibilité est donnée par :

$$E\left[\left(\hat{\mu}(\theta_i) - \mu(\theta_i)\right)^2\right] = \tau^2 \left(1 - z_i\right) \left(1 + \frac{1-z_i}{z_.}\right)$$

Les estimateurs $\widehat{\sigma^2}$ et $\widehat{\tau^2}$ des paramètres de structure sont sans biais et convergents pour autant que $\sum_i \left(\frac{\omega_i}{\omega_.}\right)^2$ tende vers 0 quand le nombre de risques tend vers l'infini, ce qui est naturellement le cas.

$$\widehat{\sigma^2} = \frac{1}{I\,(n-1)} \sum_{i=1}^{I} \sum_{j=1}^{n} \omega_{ij}\, \left(X_{ij} - X_i\right)^2$$

$$\widehat{\tau^2} = \frac{\omega_.}{\omega^2_. - \sum_i \omega^2_{i.}} \left[\sum_{i=1}^{I} \omega_{i.}\, \left(X_i - \overline{X^2}\right) - (I - 1)\, \widehat{\sigma^2}\right]$$

Notons que l'estimateur $\widehat{\tau^2}$ peut donner une valeur négative ; en pratique, on utilisera donc $(\widehat{\tau^2})' = \max(\widehat{\tau^2}, 0)$[27].

Dans la pratique, nous remplacerons les paramètres de structure par leurs estimateurs ; si bien que le facteur de crédibilité individuel devient :

$$\hat{z}_i = \frac{\omega_{i.}}{\omega_{i.} + \frac{\hat{\sigma}^2}{\hat{\tau}^2}}$$

Et donc l'estimateur de crédibilité empirique est donné par :

$$\hat{\mu}(\theta_i)^{emp.} = \hat{z}_i\, X_i + (1 - \hat{z}_i)\, \hat{\mu}_0$$

Avec, pour un groupe k donné :

$$\hat{\mu}_{0k} = \frac{\sum \hat{z}_i\, X_i}{\sum \hat{z}_i}$$

Par construction des groupes, l'assuré i ne peut appartenir qu'à un groupe k et un seul. Les notations $X_{ik}$ et $\hat{z}_{ik}$ n'ont donc pas tellement de sens.

L'interprétation de l'estimateur de crédibilité empirique est donc la suivante :
Plus le risque interne au risque individuel est fort, plus le facteur de crédibilité est faible et donc moins le P/C « individuel » a de poids dans l'estimation du ratio P/C de l'année suivante.
Plus l'hétérogénéité du portefeuille est forte, plus le facteur de crédibilité est fort et donc moins le P/C « groupe » a de poids dans l'estimation du ratio P/C de l'année suivante.

Symétriquement, on retrouve on ne peut plus logiquement : plus le risque interne au risque individuel est faible, plus le P/C « individuel » a de poids ; et plus l'hétérogénéité du portefeuille est faible, plus le portefeuille est homogène et donc plus le P/C « groupe » a de poids dans l'estimation.

[27] Voir notamment l'article de Pierre Pétauton :
http://www.ressources-actuarielles.net/EXT/IA/sitebfa.nsf/0/70F52757568D8C39C1257B7B00325289/$FILE/25_Article4.pdf?OpenElement

Plus de détails, et notamment de nombreuses démonstrations, sur le cours en ligne de Pierre Thérond.[28]

**Récapitulatif de la démarche à suivre en pratique[29]:**

**1 - Détermination des poids $\omega_{ij}$**

**2 - Calcul des $X_i$ = somme des $X_{ij} * \omega_{ij} / \omega_{i\bullet}$**

**3 - Estimation de $\sigma^2$ par $\widehat{\sigma}^2$**

**4 - Estimation de $\tau^2$ par $\widehat{\tau}^2$**

**5 - Estimation des $z_i$ par $\hat{z}_i = \omega_{i\bullet} / (\omega_{i\bullet} + \widehat{\sigma}^2/\widehat{\tau}^2)$**

**6 - Estimation pour chaque groupe k des $\mu_{0k}$ par $\hat{\mu}_{0k}$**

**7 - Calcul des $\hat{\mu}(\theta_i)^{emp.} = \hat{z}_i X_i + (1 - \hat{z}_i) \hat{\mu}_0$**

#### *3.2.2.4 Montant des prestations*

En santé, elles sont simplement déduites des cotisations et du P/C, par assuré i, et par année j :

$$P_{ij} = P/C_{ij} * C_{ij}$$

En prévoyance (hors PJ), en l'absence d'un historique suffisant, nous retenons une hypothèse forte : les prestations sont estimées pour tous les adhérents à *x%* des cotisations HT. Le P/C cible à la tarification a été évalué à *y%* (hormis Garantie Décès à *z%*), volontairement légèrement sur-estimé, en application du principe de précaution face à des risques peu maîtrisés ; tandis que le P/C constaté entre 2014 et 2016 oscille de $x_1$% à $x_2$% selon l'année et le produit. L'hypothèse retenue est donc à mi-chemin environ de ces deux valeurs. Elle peut paraître trop simplificatrice mais d'une part, le poids de la prévoyance dans la marge totale est relativement faible (< 4%), et d'autre part, en pratique, certains, très peu, auront des sinistres tandis que la majorité n'en aura aucun.

Pour la PJ, les prestations valent 0 € pour tous les adhérents, le produit étant assuré par une filiale de la GMF.

[28] http://www.therond.fr/wp-content/uploads/cours/Credibilite.pdf

[29] Pierre Thérond

#### *3.2.2.5 Choix du taux d'actualisation*

*« Un euro sûrement aujourd'hui vaut plus qu'un euro peut-être demain »*[30]

Comme évoqué plus haut, les flux futurs doivent être nécessairement actualisés. Nous devons donc choisir un taux d'actualisation. Parmi les innombrables candidats, examinons en trois :

Le candidat académique : **le WACC**
(Weighted Average Cost of Capital i.e. le Coût Moyen Pondéré du Capital).[31]

Il désigne le taux de rentabilité annuel minimal exigé par les investisseurs en capital dans les entreprises. Autrement dit, cet indicateur mesure ce que l'entreprise doit à tous ceux qui ont apporté des capitaux.

$$WACC = [FP * r_E + D * r_D * (1 - t_{IS})] / (FP + D)$$

FP est le montant des fonds propres de l'assureur
$r_E$ est le taux de rendement exigé des actionnaires
D est le montant de la dette de l'assureur
$r_D$ est le taux de remboursement de la dette
$t_{IS}$ est le taux d'impôt sur les sociétés

En pratique, il est assez peu utilisé, principalement car le calcul suppose d'être en mesure d'estimer le coût des fonds propres, or celui-ci est rarement accessible à l'analyste financier. Une méthode consiste à l'estimer par la méthode du MEDAF basée sur le Bêta des capitaux propres de la société :

$$r_E = r_f + \beta_E * (r_m - r_f)$$

$r_f$ est le taux sans risque
$\beta_E$ est le coefficient Beta de l'assureur par rapport au marché de référence
$r_m$ est le taux de rendement du marché de référence

Mais cette méthode connaît elle-même ses limites, le bêta des fonds propres n'étant possible à déterminer que pour une société cotée et dont le cours de bourse est liquide.

D'autre part, pour une mutuelle comme LMG, il n'y a ni capital social, ni actionnaire à rémunérer ; ce candidat est donc inéligible et de facto éliminé.

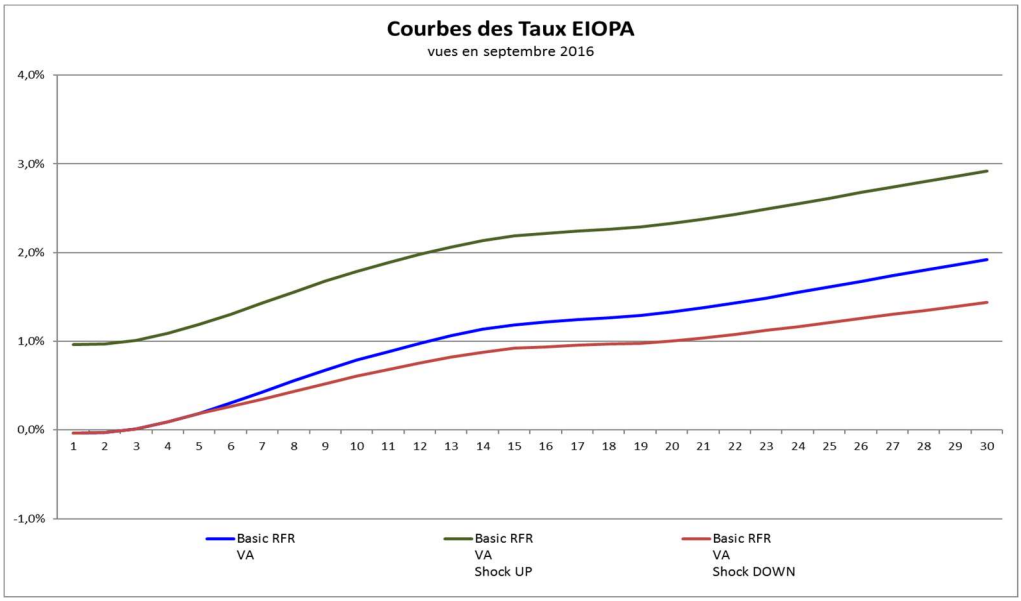
Le candidat potentiel : **le TME** (Taux Moyen d'emprunt d'Etat)

Le TME correspond au taux moyen de rendement des emprunts d'État et des obligations assimilables du Trésor (OAT) émises par l'État français, à taux fixe, et d'une durée supérieure à 7

[30] T. Béhar, 1er cours du CEA.
[31] https://fr.wikipedia.org/wiki/Co%C3%BBt_moyen_pond%C3%A9r%C3%A9_du_capital

ans. Il sert de référence aux banques et aux assurances pour déterminer le niveau des taux d'intérêts fixes. Le TME, qui de plus en plus remplace le TMO, est employé, notamment, pour le calcul du taux fixe d'un prêt à taux variable convertible en taux fixe ou calculer le taux des avances sur contrats d'assurance vie.
Calcul : moyenne (TEC 10) + 0.05% avec des bornes de sauts. Valeur mars 2017 : **1,10%** (source : Banque de France[32])

Le candidat retenu, que j'ai jugé le plus pertinent dans notre contexte en raison du caractère « court » du risque principal (la santé) de notre étude : **le taux Forward** de maturité 1 an calculé à partir de la courbe des taux EIOPA[33], un taux pour chaque année.



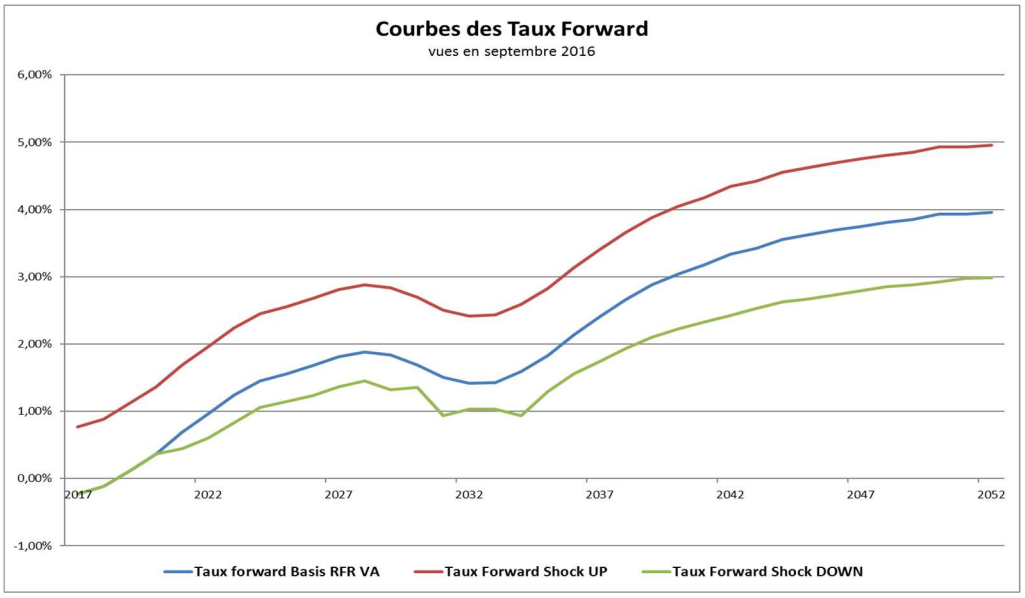

Formule générique : $$F(t_1, t_2) = \left[\frac{(1+r_2)^{d_2}}{(1+r_1)^{d_1}}\right]^{\left(\frac{1}{d_2-d_1}\right)} - 1$$

Avec $F(t_1, t_2)$ : taux Forward entre $t_1$ et $t_2$
$r_1$ : taux Eiopa d'échéance $t_1$
$r_2$ : taux Eiopa d'échéance $t_2$
$d_1$ : nb années entre la date initial et l'échéance $t_1$
$d_2$ : nb années entre la date initial et l'échéance $t_2$

Formule maturité 1 an, entre n et n+1 : $$F(t_n, t_{n+1}) = \left[\frac{(1+r_{n+1})^2}{(1+r_n)}\right] - 1$$

Valeur pour 2017 calculé en septembre 2016 : **-0,229 %**

[32] https://www.banque-france.fr/statistiques/taux-et-cours/les-indices-obligataires
[33] https://eiopa.europa.eu/regulation-supervision/insurance/solvency-ii-technical-information/risk-free-interest-rate-term-structures.


Courbes des Taux Forward
vues en septembre 2016
6,00%
5,00%
4,00%
3,00%
2,00%
1,00%
0,00%
-1,00%
2017
2022
2027
2032
2037
2042
2047
2052
Taux forward Basis RFR VA
Taux Forward Shock UP
Taux Forward Shock DOWN

## 3.3. Les coûts

Le traitement analytique des coûts est le volet du modèle le moins développé à ce jour. En effet, l'objectif de démoyenniser au maximum les coûts pour chacun des clients suppose de pouvoir s'appuyer sur une comptabilité analytique très large et particulièrement robuste. Or, historiquement, LMG ne disposait pas d'une culture très étendue des coûts de l'entreprise.

Au gré des nombreux changements réglementaires, et notamment du fait désormais de la prééminence de la directive européenne Solvabilité II, notre entreprise se dote peu à peu de l'arsenal des outils les plus adéquats et les sophistiquées en termes de gestion de l'ensemble des coûts d'une entreprise d'assurance.

Toutefois, aujourd'hui, nous ne sommes pas encore en mesure de proposer une démoyennisation, c'est-à-dire une individualisation, fiable et puissante des coûts telle que nous la souhaiterions en cible, et telle qu'elle apporterait une dimension nouvelle appropriée au modèle de Valeur Client.

Néanmoins, des travaux récemment initiés pourraient nous conduire à revoir au moins partiellement cette position. En attendant, nous utiliserons des agrégats généraux mais certifiés, à savoir :

**Frais de gestion** (pour calcul des cotisations nettes de FG) **: x% des cotisations HT** santé et prévoyance. Une alternative consistait à moduler par adhérent, le taux selon le montant de prestations (corrélé à 90% environ avec le nombre de contacts[34]), en intégrant une équation - trouvée empiriquement - de la forme :

$$\text{Taux de FG} = 4\% + 0.277\% * \text{Max}[(\text{Montant annuel de prestations} - 192.21) ; 0] / 5.28.$$

Par exemple, nous pourrions prendre l'hypothèse d'un taux de FG par client, au minimum de 4%, croissant selon le montant des prestations, et, en moyenne pondérée, égal à 8%. Il serait dans ce cas, à recalculer chaque année pour chaque adhérent, selon le montant des prestations estimées par Bühlmann-Straub. Cette hypothèse n'est finalement pas retenue dans un premier temps.

**Coût d'acquisition unitaire** (qui embarque arbitrairement le coût des actions de fidélisation non récurrentes) : sur la base d'une analyse d'experts, choix d'un coût unitaire à $\mathbf{y_1}$ **€** (cible 2016) pour les nouveaux adhérents. Le choix d'un coût modulé par canal d'entrée n'est pas retenu. En effet, la plupart des nouveaux adhérents sont cross-canaux ! $\mathbf{y_2}$ **€** pour les adhérents précédemment « enfants » sur le contrat de leurs parents (coût de la campagne de conversion des enfants en assurés principaux).

A noter, les **coûts fixes de structure** (comme le coût du réseau physique par exemple) ne sont pas intégrés au modèle (conformément à la pratique sur le marché), car ils ne dépendent pas directement du volume de clients.

[34] Téléphone entrant, courrier papier, e-mail, chat, réseaux sociaux, visite agence.

## 3.4. Conclusion partie III

- **Trois éléments essentiels composent le modèle de valeur client :**
  - **La durée de vie contractuelle entre le client et l'assureur,**
  - **La contribution du client à la marge de l'entreprise,**
  - **Les coûts associés au client, le plus possible démoyennisés.**

- **Ces 3 mesures sont pour parties connues – le passé enregistré dans les SI – et pour parties à évaluer – le futur estimé par les algorithmes -.**

- **Le modèle de Bühlmann-Straub me semble être la meilleure solution pour évaluer la sinistralité future. Il a pour avantage de tenir compte à la fois de la sinistralité probable propre au client, et de la sinistralité probable de son groupe homogène associé.**

- **Le volet « Coûts » est le moins développé à ce jour. Les prochains efforts d'amélioration du modèle devront nécessairement en grande partie se porter sur ce terrain. En particulier, il s'agira d'appréhender les différents coûts de façon différenciée selon les groupes d'assurés homogènes.**

# Partie 4 : L'analyse prédictive

« *Aimeriez-vous savoir pourquoi je suis dans*
*ce fauteuil et pourquoi je gagne tout ce fric ?*
*Je suis là pour une raison et une seule :*
*je dois deviner quelle sera la musique d'ici une semaine,*
*un mois ou un an. C'est tout. Rien de plus.*
*Et là, ce soir, j'ai bien peur de ne rien entendre du tout.*
*Rien que le silence…* »

Extrait du film Margin Call, 2012.

## 4.1. Généralités sur le Machine Learning et l'analyse prédictive

### 4.1.1. L'intelligence artificielle

Le principe du Machine Learning est de permettre à une machine d'apprendre par elle-même par l'intermédiaire d'un processus d'apprentissage supervisé ou non. Cette implémentation de processus et d'analyse rend possible la capacité d'automatisation d'une machine en rapport avec une problématique donnée, c'est-à-dire une capacité de raisonnement, d'apprentissage et de prise de décision. On parle également pour cela d'Intelligence Artificielle.

*« A condition de pas s'embarrasser de trop de scrupules sémantiques, la définition[35] de l'intelligence artificielle (IA) est aisée : c'est l'ensemble des disciplines techniques et scientifiques qui permettent de reproduire certains processus cognitifs humains comme l'apprentissage, l'intuition, l'auto-amélioration, la créativité, la planification des tâches ou encore la compréhension du langage naturel. Les plus optimistes y ajouteront pour leur part, en les plaçant dans un futur indéfini, des attributs propres à leur conscience comme la volonté et les émotions ».*[36]

L'intelligence artificielle trouve son fondement dans la pluridisciplinarité. Au moins 5 domaines des sciences sociales et techniques ont favorisé son essor :

- **la philosophie** – l'articulation entre l'esprit et la matière -,
- **l'économie** – comprendre les performances collectives des agents rationnels qui maximisent leur revenu ou leur bien-être -,
- **les mathématiques** – garantir la validité d'un raisonnement -,
- **les neurosciences** – étudier le fonctionnement du cerveau et les différentes fonctions cognitives -,
- **la cybernétique** – l'étude des mécanismes autorégulés-.

L'IA s'est construite en un demi-siècle en émancipation de toutes ses disciplines, dans le but de concevoir des machines capables d'évoluer de manière autonome dans des environnements changeants. Il n'est pas aisé de dire si telle machine ou telle combinaison d'algorithmes est une IA. Pour y voir plus clair, les spécialistes ont défini 5 critères :

- **la profondeur** – mesure de performance spécifique au champ d'application pour lequel l'IA a été conçue -,
- **l'étendue** – spécialisation dans la résolution d'une catégorie bien précise de problèmes -
- **le mode d'apprentissage** – le Machine Learning, composante essentielle de nombreux systèmes intelligents, voir plus bas -,
- **l'autonomie** – recherche d'un système qui n'exigera en principe aucune intervention humaine -,
- **la conscience** – présence objective d'une subjectivité capable de ressentir, de se poser des questions et de vouloir -.

[35] D'autres définitions existent : voir Larousse, M.L. Minsky, D. Pastre ou J.L. Laurière
[36] P. Lemberger, J. Lepan, O. Reisse, in « Intelligence Artificielle : où en sommes-nous ? »

Quelques-uns des principaux usages de l'IA **aujourd'hui** :

- les systèmes de traduction automatique – **G**oogle traduction -,
- les assistants pour smartphones – Siri pour **A**pple -,
- la reconnaissance d'image – les photos de **F**acebook -,
- les systèmes de recommandation – type **A**mazon -,
- l'industrie du divertissement – les jeux vidéo en réseau -,
- la robotique domestique – Cutii, Buddy, Kompaï, Jibo, Zenbo, … -.

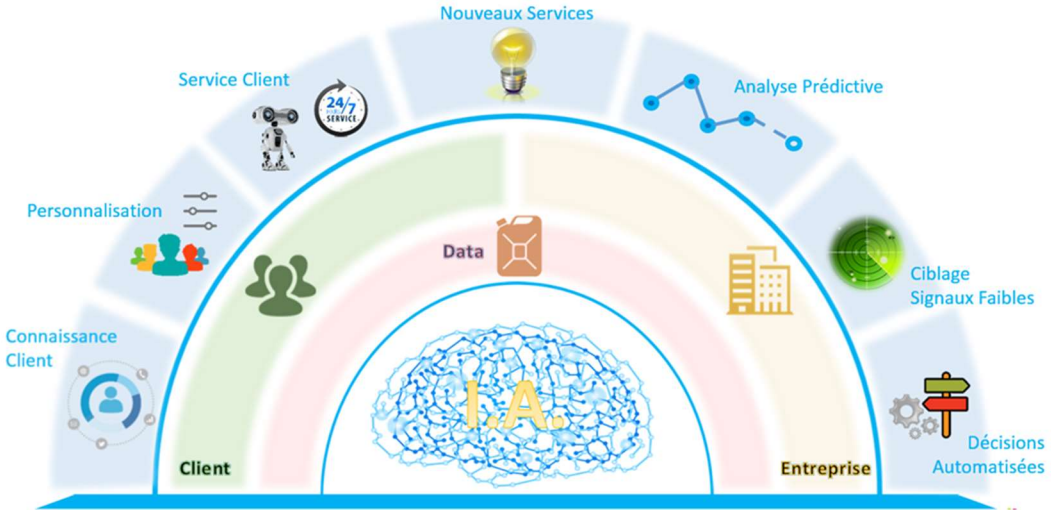
Quelques-uns des principaux usages de l'IA **demain** :

- les transports – Google car -,
- une nouvelle médecine – davantage préventive que curative -,
- un nouveau type d'Interface Homme-Machine
- de nouveaux moteurs de recherche plus puissants
- la tranquillité d'esprit – en nous débarrassant de l'agitation superflue -.

Après-demain, beaucoup de fantasmes : Verrons-nous de notre vivant des systèmes autonomes dotées de conscience ? Atteindrons-nous le point de singularité technologique, c'est-à-dire le moment où les machines pourront s'auto-améliorer mieux que nous ne saurions le faire nous-mêmes ? Faut-il s'inquiéter du mouvement transhumaniste ? Pour Stephen Hawking, le célèbre astrophysicien, *« l'IA sera soit la meilleure, soit la pire des choses jamais arrivées à l'humanité »*. Il précise que *« maîtriser l'IA sera l'un des plus grands événements dans l'histoire de notre civilisation »* … Vaste programme !

*Source : Weave, IA, où en sommes-nous ?*

Concernant le domaine de l'assurance, les champs d'application de l'IA sont très larges. D'une manière générale, l'IA a pour objectif de concevoir des machines capables d'évoluer dans un environnement hostile et changeant. Elle permet de **gagner en productivité** et de **créer de la valeur**, et elle est potentiellement profitable partout où il est envisageable d'**automatiser des tâches à faible valeur ajoutée**. Sans entrer dans le détail, quelques pistes que nous explorons à La Mutuelle Générale :



### 4.1.2. Data science et Machine Learning

Le « Big Data », c'est un peu comme le réchauffement climatique : à l'origine très partagées, les opinions sont dorénavant beaucoup moins data-sceptiques (respectivement climato-sceptiques). Peu d'experts en effet contestent que nous assistons depuis une dizaine d'années à une véritable « Révolution Data ». Une rupture de paradigme forte est apparue, déterminée par l'émergence de nouveaux algorithmes, la possibilité de traiter des opérations à distance avec le *cloud*, une capacité de stockage des informations considérablement étendue, et une nouvelle école : la *Data Science*[37]. Le tout donne des possibilités totalement démesurées par rapport à ce qui pouvait exister auparavant.

Alors que le « Big Data » peut s'interpréter comme la somme des technologies de gestion de données massives, la Data Science est le terme qui réunit l'ensemble des méthodes de traitement de données. Fondamentalement, cette discipline étudie la science de l'extraction de connaissance à partir d'ensembles de données. Elle s'appuie principalement sur trois domaines plus larges : les mathématiques, la statistique et l'informatique.

Le premier objectif de la Data Science est de produire des méthodes automatisées d'analyse de données volumineuses et de sources plus ou moins complexes, afin d'en extraire des informations potentiellement utiles pour les domaines métiers en assurance, comme l'actuariat, le marketing ou la relation client.

Pour cela, le *data scientist* s'appuie sur la fouille de données, les statistiques, diverses méthodes de référencement, l'apprentissage automatique et la visualisation de données[38]. Il s'intéresse donc au nettoyage, à l'exploration, à l'analyse et à la protection de bases de données capables de fonctionner avec d'autres systèmes d'informations de l'entreprise comme le *back office* et le *front office*. Il est désormais en capacité de répondre aux 4 questions suivantes tandis que son prédécesseur, le *data analyst* ne disposait d'outils que pour répondre aux 2 premières :

**Descriptif** : Que s'est-il passé ? **Diagnostic** : Pourquoi est-ce arrivé ?

**Prédictif** : Que va-t-il arriver ? **Prescriptif** : Comment faire mieux ?

On a beaucoup lu ou entendu *« Data is the new oil »*[39]. Il est vrai que dans les deux cas, on doit au préalable procéder à une extraction, les algorithmes représentant le raffinage. Je ne suis pas d'accord. La data n'a pas de prix unitaire dans l'absolu. Lorsque je vends un baril de pétrole, je ne possède plus ce baril. Certes j'ai reçu de l'argent en contrepartie. Si je partage mes données, je les conserve, je ne perds rien et j'en reçois des nouvelles en échange. Plutôt que de voir la *Data* comme une nouvelle déclinaison de l'or, je préfère percevoir ce bouleversement technologique comme l'émergence d'une nouvelle forme d'économie. Et en cela, l'axe 5 du plan stratégique[40] 2017-2020 de LMG est totalement raccord.

[37] Science des données en français
[38] Ou « Data viz »
[39] Cette devise fait écho à l'accroche publicitaire souvent citée par les éditeurs de logiciels : « Comment trouver un diamant dans un tas de charbon sans se salir les mains ? ».
[40] Voir 1.3.6. Stratégie de La Mutuelle Générale

Le *Machine Learning*, ou apprentissage automatique, intègre la conception, l'analyse, le développement et l'implémentation de processus permettant à une machine d'évoluer par elle-même, avec pour finalité d'essayer d'accomplir des tâches considérées difficiles par des moyens plus classiques. Elle matérialise la boîte à outils associée à la Data Science. On peut la scinder en trois familles :

- **l'apprentissage supervisé** (toutes formes de régression et de classement),
- **l'apprentissage non supervisé** (ou *clustering,* qui a pour finalité de diviser un ensemble d'éléments en groupes homogènes). L'objectif généralement poursuivi dans ce cas est la recherche d'une typologie ou d'une taxonomie des individus : comment regrouper ceux-ci en classes homogènes les plus dissemblables entre elles. Les représentants les plus connus sont :
    - la Classification Ascendante Hiérarchique[41]
    - la méthode dite des « k-means », un algorithme de réallocation dynamique
    - les cartes auto-organisatrices de Kohonen
- **l'apprentissage par renforcement** (apprendre à partir d'expériences ce qu'il convient de faire en différentes situations, de manière à optimiser au fil du temps, une récompense sous forme de gain quantitatif). Se rapproche dans certains cas de la théorie des jeux[42].

Là où le Data Mining visait à découvrir des relations précédemment inconnues dans les jeux de données analysés, le Machine Learning cherche à prédire des événements sur la base de relations connues déduites de données d'apprentissage. Le processus est le plus souvent itératif : apprentissage initial => prédictions du modèle retenu => boucle de retour pour améliorer les prédictions suivantes. De façon similaire à l'inférence bayésienne[43], le principe général réside dans le fait que les données parlent d'elles-mêmes, le data scientist ne prenant pas d'hypothèses a priori sur les modèles, lois ou probabilités sous-jacents, comme il est d'usage en statistique dite « classique ». Une probabilité s'interprète alors comme le passage à la limite de la fréquence d'un événement.

La difficulté principale réside dans le risque d'explosion combinatoire : dès que le nombre de descripteurs augmente, deux problèmes se posent : calculabilité (le nombre d'opérations est rapidement énorme) et fragmentation des données (beaucoup de « cases » à effectifs très faibles voire nuls). C'est ici que les algorithmes entrent en jeu : ils ont pour objectifs d'ajuster des modèles pour simplifier la complexité contenue dans les données brutes.

La théorie sous-jacente à ces méthodes est connue depuis longtemps. C'est l'explosion de la capacité de calculs des machines qui permet aujourd'hui leurs mises en œuvre. Les « nouveaux algorithmes » ne matérialisent en réalité que des relectures et approfondissements de procédés déjà existants mais inapplicables auparavant. Les noms changent mais les méthodes restent. Si bien que la Data Science et la Machine Learning pourraient tout aussi bien se percevoir comme une extension du Data Mining, une espèce de « Big Data Mining ».

[41] Voir Section 6.4.2. Classification Mixte
[42] https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_jeux
[43] https://sciencetonnante.wordpress.com/2012/10/15/linference-bayesienne-bayes-level-2/

### 4.1.3. L'analyse prédictive[44]

L'analyse prédictive, ou apprentissage supervisé, est une des trois familles que constitue le Machine Learning. Cette famille est elle-même composée de deux sous-familles : le **classement** (on dit aussi discrimination, ou « classification » dans la littérature anglo-saxonne), lorsque la variable à expliquer est **qualitative**, et la **régression** (ou prédiction), lorsque la variable à expliquer est **continue**.

Le classement consiste à placer tous les individus de la population ou d'un échantillon, dans une classe, parmi plusieurs classes prédéfinies, selon les caractéristiques de l'individu données par les valeurs des variables explicatives. Un modèle de classement permet donc d'affecter chaque individu à sa classe d'appartenance la plus probable.

La régression a pour objectif d'estimer la valeur d'une variable à expliquer (ou endogène), continue, en fonction des valeurs de descripteurs, c'est-à-dire des valeurs des variables explicatives (ou exogènes).

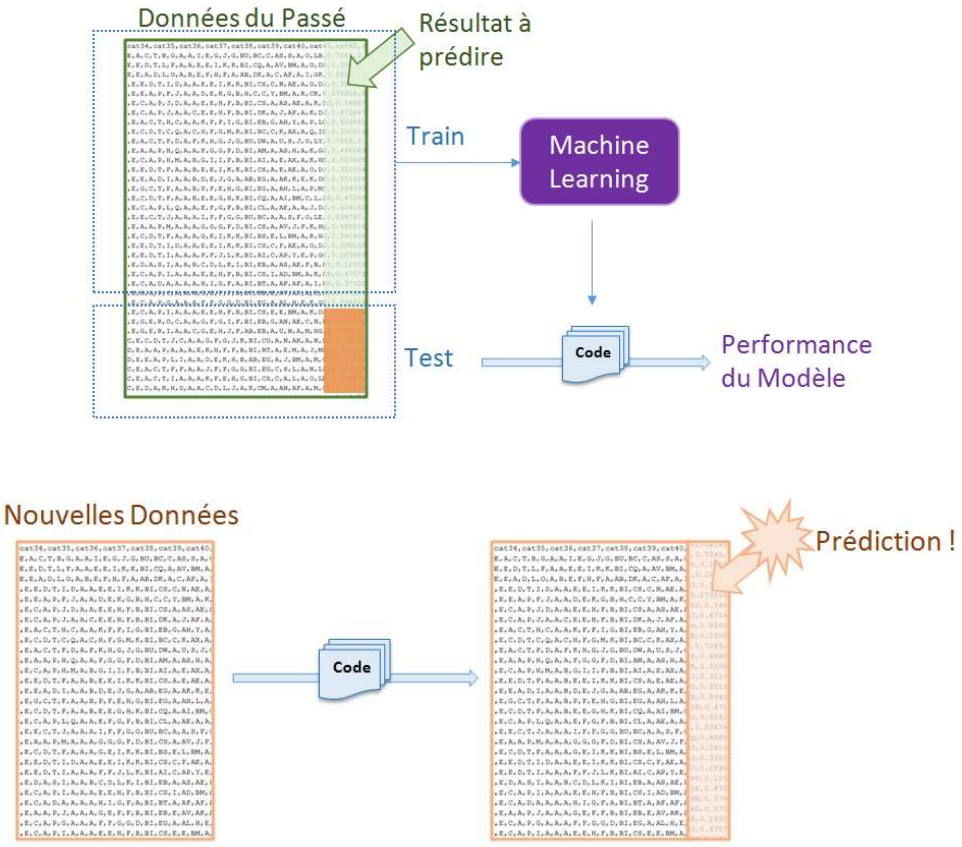
Nous supposons disposer d'un ensemble d'apprentissage constitué de données d'observations de type entrée-sortie $\{(x_1, y_1), \ldots (x_n, y_n)\}$, avec $x_i \in$ X quelconque et $y_i \in$ Y, pour tout i de 1 à n. En apprentissage supervisé, **l'objectif est donc de construire, à partir de cet échantillon d'apprentissage, un modèle qui va nous permettre de prévoir la sortie $y_{n+1}$ associée à une nouvelle entrée $x_{n+1}$.** Les observations 1, i, …, n, n+1, ne sont pas nécessairement ordonnées.

C'est la technique dite inductive-déductive : un processus d'apprentissage pour l'élaboration du modèle qui peut ensuite être appliqué à de nouvelles données pour en déduire un classement ou une prédiction. Plus précisément cette méthode se déroule en 4 étapes dont 3 sont fondamentales :

- l'apprentissage : construction d'un modèle sur un 1er échantillon d'individus tirés aléatoirement dans la population, pour lesquels on connaît le classement ou la prédiction
- le test : vérification du modèle sur un 2ème échantillon en comparant les classements ou valeurs prédits par le modèle avec les classements ou valeurs réels connus
- la validation (obligatoire pour certaines méthodes, facultative pour d'autres) : mesure de la performance du meilleur modèle sélectionné sur un 3ème échantillon, pour identifier la qualité des résultats et vérifier la stabilité du pouvoir prédictif
- l'application : exécution du modèle sur l'ensemble de la population à scorer pour déterminer le classement ou la valeur de la variable endogène de chaque individu

[44] Source : S. Tufféry, *Data Mining et statistique décisionnelle, Chapitre XI.*

**Schéma résumé du processus décrit ci-dessus :**




**Dilemme biais-variance et sur-apprentissage :**

D'une manière générale, plus la taille de l'échantillon d'apprentissage augmente, plus le taux d'erreur sur l'échantillon d'apprentissage augmente, et plus il diminue sur l'échantillon test. Il existe une propriété appelée « consistance » qui dit que ces deux taux convergent vers une même limite à partir d'une taille critique des deux échantillons.[45]

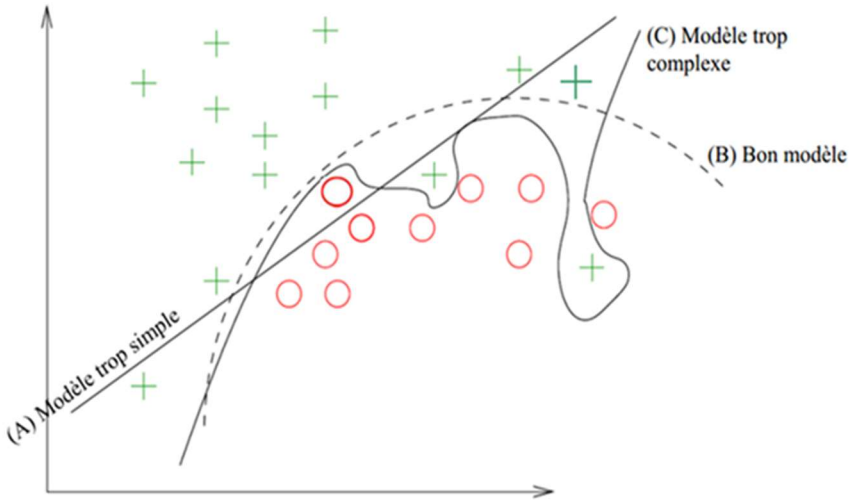
Si on définit $\mathbf{h}$ comme une mesure de complexité d'un modèle et $\mathbf{n}$ la taille de l'échantillon d'apprentissage, un théorème[46] nous apprend qu'un modèle possède un pouvoir de généralisation d'autant plus important que le rapport $\frac{h}{n}$ est faible ; mais sans que la complexité $\mathbf{h}$ soit trop faible car cela traduirait probablement un modèle mal ajusté et donc des taux d'erreurs élevés. Le modèle optimal réclame donc un arbitrage entre qualité d'ajustement et pouvoir de généralisation. Cet arbitrage est aussi appelé le « dilemme biais-variance »[47]. L'idée principale est que l'on cherche à minimiser à la fois l'erreur de modèle (le biais) et l'erreur due aux fluctuations de l'échantillon (la

[45] Pour plus de détails sur la convergence du risque empirique, et la vitesse de cette convergence, se référer aux travaux de Vladimir Vapnik, par exemple dans l'ouvrage de Gilbert Saporta citée dans la partie Bibliographie.
[46] Le théorème de Vapnik : file:///C:/Users/User/Downloads/fageotroyer.pdf
[47] https://fr.wikipedia.org/wiki/Dilemme_biais-variance

variance). Il faut donc nécessairement trouver un compromis entre qualité d'ajustement et variabilité de la prédiction sur les nouveaux cas.

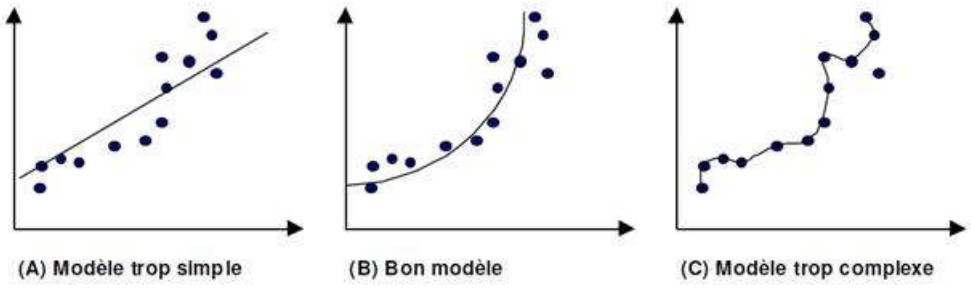
Le paragraphe précédent évoque ainsi la difficulté de généraliser les modèles trop complexes. En effet, poussés à l'extrême lors de la phase d'apprentissage, des modèles pourraient épouser exactement toutes les fluctuations dues à l'échantillon, générant un sur-ajustement, ce que nous ne souhaitons pas bien entendu. Les deux figures suivantes illustrent parfaitement ce phénomène.

Sur-apprentissage en classement :




*Source : Olivier Bousquet*

Sur-apprentissage en prédiction :




*Source : Stéphane Tufféry*

Les modèles (A) sont trop simplistes : ils présentent des biais trop importants. Ils sont facilement généralisables mais les taux d'erreurs sont trop forts, ce qui conduirait à de fausses interprétations.

Les modèles (C) à l'inverse sont trop complexes : les variances sur les prédictions seront trop fortes. Les taux d'erreurs en apprentissage seront certes très faibles mais aucune chance de réussir à généraliser sur les échantillons de test et de validation.

Seuls les modèles (B) sont acceptables. On admet des petites erreurs mais ils sont suffisamment complexes et conservent des bonnes capacités de généralisation. Ne pas oublier *in fine* qu'ils seront d'autant mieux admis qu'ils seront cohérents avec une expertise métier du sujet.

## 4.2. Méthodologie

### 4.2.1. Apprentissage supervisé : problème fondamental

L'ensemble des méthodes décrites ci-après appartient à la famille des techniques d'apprentissage spontané dites « Apprentissage supervisé », où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des cas déjà traités et avérés.

Dans ce chapitre, $Y\ \{y_1 \ldots y_n\}$ représentera la variable endogène (ou dépendante), à prédire, tandis que $X\ \{x_1 \ldots x_n\}$ symbolisera l'ensemble des variables, nominales ou continues, exogènes (ou indépendantes), qui prennent le rôle de descripteurs-prédicteurs. $\alpha\ \{\alpha_1 \ldots \alpha_n\}$ figurera le vecteur des paramètres associés aux prédicteurs.

On suppose que $Y$ et $X$ sont liés par une relation ou une liaison fonctionnelle sous-jacente, $f$ comme par exemple : $Y = f(X, \alpha) + \varepsilon$, où $\varepsilon$ est un bruit blanc centré sur 0 et de variance constante, connue ou facilement estimable. L'additivité du bruit est un parti pris. En cas d'erreur supposée multiplicative, une transformation logarithmique ramène au problème précédent. On travaille sur un échantillon $\Omega$, issu de la population totale $\Omega_{pop}$. L'objectif de tout apprentissage supervisé est de trouver une estimation $(\hat{f}, \hat{\alpha})$, c'est-à-dire une fonction $g = \hat{f}$ et des paramètres $\beta = \hat{\alpha}$ qui prédisent au mieux $Y$.

$$\hat{Y} = g\ (X, \beta) = \hat{f}\ (X, \hat{\alpha})$$

Une base de données d'apprentissage est un ensemble de couples entrée-sortie $(x_k, y_k)_{1 \leq k \leq n}$, avec $x_k \in X$ et $y_k \in Y$, que l'on suppose être choisis selon une loi sur $X \times Y$. Les méthodes d'apprentissage supervisé utilisent cette base de données pour déterminer la fonction de prédiction (ou « modèle ») $g$, qui à une nouvelle entrée $x$ associe une sortie $g\ (x, \beta)$. Le but d'un algorithme ainsi défini est donc de généraliser pour des entrées inconnues ce qu'il a pu déduire des données d'apprentissage.

Dans notre situation, $Y$ est binaire (sortant/non sortant, appétent/non appétent, ou risqué/pas risqué) mais $\hat{Y}$ peut prendre ces valeurs sur tout l'intervalle $[0\ ;\ 1]$, car $\hat{Y}$ est une probabilité, sans que $\hat{Y} < 0{,}5$ induise nécessairement $Y = 0$ ou $\hat{Y} > 0{,}5$ induise nécessairement $Y = 1$. En effet, dans le cadre de la discrimination binaire, nous considérons que la variable dépendante $Y$ ne prend que deux modalités : positif "+" ou négatif "-". Nous cherchons à prédire correctement les valeurs de $Y$, mais nous pouvons également vouloir quantifier la propension (la probabilité) d'un individu à être positif (ou négatif). Nous sommes donc à la frontière entre un problème de régression et un problème de classification. Mais ceci a peu d'importance, les méthodes dans les deux cas sont similaires.

### 4.2.2. Erreur de modèle et critères universels de performance

*« Attention, ne pas confondre erreur de modèle et modèle d'erreurs ! »*

#### *4.2.2.1 Erreur théorique et matrice de confusion*

Pour évaluer la qualité de la modélisation, et pour comparer plusieurs modèles entre eux, nous faisons appel à la qualité de la prédiction, c'est-à-dire la capacité du modèle à établir de bonnes estimations sur la population totale. Un moyen simple est de calculer l'erreur théorique :

$$\mathrm{ET} = \frac{1}{N} \sum \Delta \left[ Y,\ \hat{f}\ (X, \hat{\alpha}) \right]$$

Avec $\Delta \left[ \mathrm{Y}, \hat{f}\ (X, \hat{\alpha}) \right] = 0$ si $\mathrm{Y} = \hat{Y}$ , ou $= 1$ si $\mathrm{Y} \neq \hat{Y}$

Ce qui revient, à calculer la proportion de cas où la prédiction du modèle est discordante avec la vraie valeur de la variable endogène. Si toutes les prédictions sont correctes, l'erreur théorique vaut 0, et le modèle est hyper performant. Si l'erreur théorique est supérieure ou égale à ce qu'on obtiendrait aléatoirement, en laissant faire le hasard (0,5 dans le cas binaire), le modèle est inutile.

L'écart est dit « théorique » car il est illusoire de penser disposer de l'ensemble absolu de la population pour le calculer. Ce concept, bien qu'élémentaire, est essentiel. Il est le fondement de la matrice de confusion :

➢ **Matrice de confusion**

| | | Modèle | | |
|---|---|---|---|---|
| | | Souscr * | ~~Souscr *~~ | Total |
| Réalité | Souscr | 8 | 2 | 10 |
| | ~~Souscr~~ | 5 | 85 | 90 |
| | Total | 13 | 87 | 100 |

➢ **Taux de bons classements = 1 – taux d'erreur = (8 + 85) / 100 = 1 - (5 + 2) / 100 = 93%**

➢ **Modèle alternatif = modèle trivial**

| | | Modèle | | |
|---|---|---|---|---|
| | | Souscr * | ~~Souscr *~~ | Total |
| Réalité | Souscr | 0 | 10 | 10 |
| | ~~Souscr~~ | 0 | 90 | 90 |
| | Total | 0 | 100 | 100 |

➢ **93% > (0 + 90) / 100 = 90% => modèle OK !**

Si, d'autre part, le modèle obtenu n'est pas en capacité de surperformer par rapport au modèle trivial, c'est-à-dire celui qui se contente de prédire selon les totaux marginaux, la pertinence de ce modèle est douteuse.

Bien qu'indispensable, cette méthode atteint rapidement ses limites. Pour évaluer la performance des scores, nous préfèrerons nous concentrer sur des indicateurs plus élaborés et plus intéressants

que le taux d'erreur. Il en existe plusieurs, principalement fondés sur les mesures d'aire : la courbe de lift, l'indice de Gini, notamment. Nous nous contenterons dans le cadre de ce mémoire de la méthode retenue *in fine*, probablement la plus répandue : **la courbe ROC** et son indicateur associé **le critère AUC**.

#### *4.2.2.2 La courbe ROC et le critère AUC*

La courbe ROC est un outil graphique qui permet de comparer globalement le comportement de plusieurs modèles. Elle met en relation le taux de vrais positifs (TVP) et le taux de faux positifs (TFP) dans un nuage de points.

Soit $s$ un seuil de séparation du score ; $s \in [0 ; 1]$. Dans le cas standard, $s$ est le plus souvent fixé à 0,5. Soient $\pi$, un score et $\omega$, un individu quelconque.

Pour construire la courbe, nous définissons deux fonctions de $s$ :

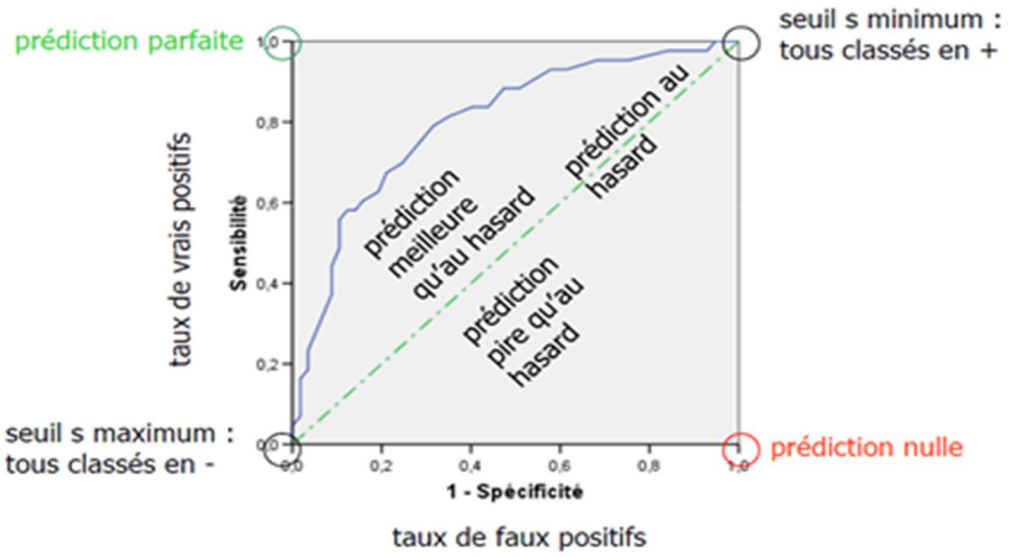
- La **sensibilité** $\alpha(s)$ qui correspond au TVP : probabilité de bien détecter un évènement au seuil $s$ i.e. $\alpha(s) = P\,[(\pi(\omega) \geq s \,/\, \omega$ est un évènement$]$
- La **spécificité** $\beta(s)$ : probabilité de bien détecter un non-évènement au seuil $s$ i.e. $\beta(s) = P\,[(\pi(\omega) < s \,/\, \omega$ est un non-évènement$]$
- On en déduit que TFP correspond à $1 - \beta(s) = P\,[(\pi(\omega) \geq s \,/\, \omega$ est un non-évènement$]$

La sensibilité mesure la performance d'un test lorsqu'il est utilisé sur les vrais positifs tandis que la spécificité mesure la performance d'un test lorsqu'il est utilisé sur les vrais négatifs. Les deux concepts donnent une évaluation de la validité inhérente à un test statistique de type dichotomique (oui/non, vrai/faux, appétent/non appétent, churner/non churner, …).

Attention, ces deux notions prisent séparément ne veulent rien dire. L'influence du seuil $s$ est essentielle. Sa valeur dépend grandement de l'utilisation que l'on fait du test : que cherche-t-on en priorité, peu de faux négatifs (sensibilité forte) ou peu de faux positifs (spécificité forte) ? Un modèle « sensible » prédit un maximum de cas (risque de détecter à tort) tandis qu'un modèle « spécifique » ne prédit que les probabilités fortes (risque à ne pas détecter).

Par exemple, dans le cas d'un dépistage, nous préférerons un test très sensible, puisque l'objectif est la détection du plus grand nombre possible d'individus « positifs ». A l'inverse, dans le cas du test de confirmation effectué sur les individus dont le test de dépistage s'est révélé positif, nous exigerons une spécificité très élevée, afin de minimiser le risque de détection à tort. On voit donc que sur un même sujet, nous pouvons être amenés à construire une stratégie de tests différents qui s'imbriquent les uns aux autres, mais où chaque étape répond à une problématique propre.

La courbe ROC est donc une courbe qui s'inscrit dans un carré de longueur 1 dont l'ordonné est la sensibilité, $\alpha(s)$, et l'abscisse, le complémentaire à 1 de la spécificité, $1 - \beta(s)$.

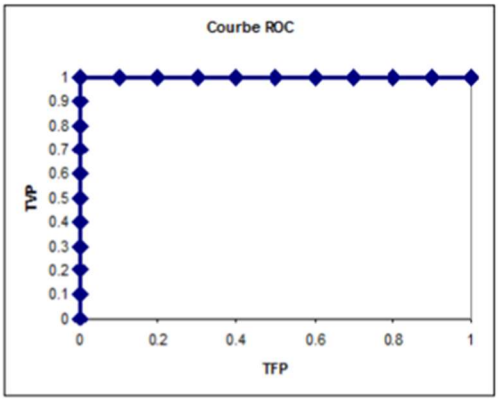



*Source : Stéphane Tufféry*

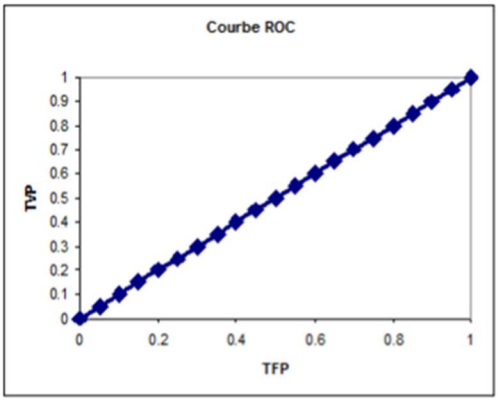
Interprétation : plus la courbe s'approche du coin nord-ouest, plus le modèle est performant, car il permet de capturer le maximum de vrais évènements tout en minimisant le nombre de faux évènements. Inversement, plus la courbe s'approche de la diagonale, moins le modèle est fiable.

Exemple : La courbe passe par le point ($1 - \beta(s) = 0{,}2$ et $\alpha(s) = 0{,}6$). Cela signifie que ce point correspond à un seuil $s$ qui est tel que, si on considère « positif » tous les individus dont le score est $\geq s$, le modèle a détecté : 20% de faux positifs et 60% de vrais positifs, c'est-à-dire que 20% des faux positifs et 60% des vrais positifs ont un score supérieur à $s$.

Cas des deux situations extrêmes :




[A] Discrimination parfaite. Tous les positifs sont situés devant les négatifs lorsque l'on trie le tableau selon un score décroisant.




[B] Pas de discrimination. Les positifs et les négatifs sont mélangés c.-à-d. présentent des scores en moyenne identiques.

*Source : Ricco Rakotomalala*

Tout se passe comme si chaque point de la courbe ROC correspondait à la matrice de confusion obtenue pour une certaine valeur du seuil. Tracer la courbe ROC est donc équivalent à généraliser la construction de toutes les matrices de confusion établies pour chacune des valeurs possibles du seuil.

Pour synthétiser numériquement la courbe ROC, on utilise l'indicateur AUC pour « Aire sous la courbe ». Cette aire s'interprète comme la probabilité que $\pi(\omega_E) > \pi(\omega_{NE})$ avec $\omega_E$ est un évènement et avec $\omega_{NE}$ est un non-évènement, c'est-à-dire la probabilité de placer un individu positif devant un individu négatif. Un modèle est donc d'autant plus performant que l'aire associée est proche de 1. Si l'aire vaut 0,5, ce modèle ne classe pas mieux que le hasard. Si l'aire est comprise entre 0,5 et 0,7, la discrimination est faible. Si elle est comprise entre 0,7 et 0,8, elle est acceptable. Entre 0,8 et 0,9, elle est excellente. Au-delà, elle est exceptionnelle.

Plusieurs méthodes existent pour calculer cette métrique. La plus simple qui revient à réaliser un test de Mann-Whitney,[48] consiste à s'intéresser aux nombres de paires concordantes et discordantes. Soient $n_1$ et $n_2$, respectivement le nombre d'évènements et de non-évènements dans l'échantillon. Il existe donc $n_1n_2$ paires constituées d'un $\omega_E$ et d'un $\omega_{NE}$. Parmi ces $n_1n_2$ paires, on a concordance si $\pi(\omega_E) > \pi(\omega_{NE})$, et discordance si $\pi(\omega_E) < \pi(\omega_{NE})$. Soient $nc$, le nombre de paires concordantes et $nd$, le nombre de paires discordantes. Le nombre $n_1n_2 - nc - nd$ figure donc les ex-æquo. Dans ce cas,

$$\mathbf{AUC \approx [nc + 50\% * (n_1n_2 - nc - nd)] / n_1n_2}$$

En pratique, on utilise une macro SAS, comme celle de M. Tufféry, qui convient parfaitement. La macro, applicable à toutes les méthodes, renvoie une estimation, dépendante de l'échantillon, une *p-value*, ainsi que des bornes pour un intervalle de confiance à 95%.

### 4.2.3. Panorama des méthodes d'analyse prédictive testées

Pour construire nos 6 scores, nous nous proposons d'étudier la performance et la qualité de 8 méthodes d'analyses prédictives en apprentissage supervisé. Cette liste est bien entendu loin d'être exhaustive, les méthodes prédictives étant nombreuses, en progrès constant et en perpétuel renouvellement.

- La régression logistique
- Les arbres de décision
- Le classifieur bayésien naïf
- L'analyse discriminante
- Les réseaux neuronaux
- Les machines à vecteur de support
- Les forêts aléatoires
- Le gradient boosting

Entrer dans le détail et les subtilités de ces huit méthodes serait certainement passionnant et enrichissant mais exigerait sans doute un mémoire entier dédié à ce sujet. Ce ne peut pas être le propos ici. Nous nous contenterons donc pour chacune d'entre elles, d'une présentation brève et concise mais je l'espère efficace. J'aiguille le lecteur intéressé par ces sujets vers les ouvrages de référence, le cas échéant.

[48] http://www.jybaudot.fr/Inferentielle/mannwhitney.html

## 4.3. Les méthodes d'analyse prédictive

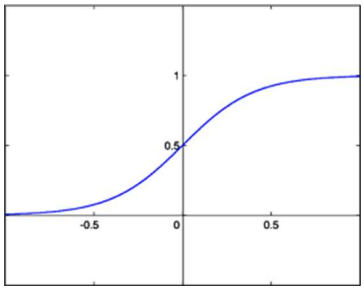
### 4.3.1. La régression logistique[49]

#### *4.3.1.1. Spécification du modèle*

La régression logistique est un modèle qui permet d'exprimer la relation entre une variable Y qualitative et des variables $X_j$ qui peuvent être quantitatives ou qualitatives. Dans le cas le plus répandu, celui qui nous intéresse plus particulièrement, la régression logistique binaire, Y a deux modalités : 1 ou « $Y^+$ » qui figure « événement = oui » et 0 ou « $Y^-$ » qui figure « événement = non ». Ce modèle permet de calculer la probabilité de survenue d'un événement quand la valeur des variables $X_j$ est connue : $P(Y = 1 / X_1, X_2, \dots X_K) = P(Y^+ / X_1, X_2, \dots X_K)$. Notons que **la régression logistique est un cas particulier du modèle linéaire général (MLG)**. Il reprend la plupart des méthodes de cette famille de modèles : estimation par maximum de vraisemblance, statistiques de test suivant asymptotiquement des lois du $\chi^2$, calcul des résidus, observations influentes, critère d'Akaike (AIC) pour la sélection du meilleur modèle…

Cas du modèle logistique simple, une seule variable X explicative : $\pi(x) = P(Y = 1 / X = x)$

$\pi$ ne peut être linéaire car Y ne prend que deux valeurs. Ainsi $\pi(x) = \beta_0 + \beta_1 x$ ne convient pas. On utilise alors la **fonction logistique qui a une forme sigmoïde**. Nous avons donc nécessairement $0 < \pi(x) < 1$, et cela permet de travailler avec des valeurs dans $]-\infty ; +\infty [$.




$$\pi(x) = \frac{\exp^{(\beta_0+\beta_1 x)}}{1+\exp^{(\beta_0+\beta_1 x)}}$$

ou l'équivalent en introduisant la fonction de lien Logit[50] :

$$\text{Logit}\,[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

Si X est binaire (sujet exposé $X = 1$, non exposé $X = 0$), on obtient ainsi :

$$P(Y = 1 / X = 1) = \frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}} \text{ et } P(Y = 1 / X = 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

[49] Ouvrage de référence en français : « Pratique de la régression logistique », Rakotomalala, Ricco, (2011). http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf
Présentation succincte très pédagogique : https://perso.univ-rennes1.fr/valerie.monbet/ExposesM2/2013/La%20re%CC%81gression%20logistique.pdf

[50] Il existe deux autres fonctions de transfert qui renvoient aux modèles Probit et Log-log.

#### *4.3.1.2. Odds et odds ratio*

Les **odds ratio** (OR) sont indissociables de la régression logistique. Considérons Z une variable qualitative à K modalités. On désigne ***odds***, la probabilité de voir se réaliser la j[ème] modalité plutôt que la k[ème], par le rapport : $\Omega_{jk} = \frac{\pi_j}{\pi_k}$

Où $\pi_j$ représente la probabilité d'apparition de la j[ème] modalité. Cette quantité est estimée par le rapport $\frac{n_j}{n_k}$ des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernoulli de paramètre π, l'*odds* est le rapport $\frac{\pi}{(1-\pi)}$, qui exprime la « cote », bien connue des parieurs. Supposons que le taux de chute soit de 20%. On dit que l'*odds* de l'échec (quitter le portefeuille) est $\frac{0,2}{0,8} = 0,25$, tandis que l'*odds* du succès (rester dans le portefeuille) est $\frac{0,8}{0,2} = 4$. On dit encore que la chance de succès est de 4 contre 1 tandis que celle de l'échec est de 1 contre 4.

Considérons à présent, deux variables $Z^1$ et $Z^2$. Les paramètres de la loi conjointe $Z^1 \times Z^2$ se placent dans une matrice : $\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$ où $\pi_{ij}$ est la probabilité d'occurrence de chaque combinaison. Sur la ligne 1, l'odds de la colonne 1 est $\Omega_1 = \frac{\pi_{11}}{\pi_{12}}$ tandis que sur la ligne 2, l'odds de la colonne 1 est $\Omega_1 = \frac{\pi_{21}}{\pi_{22}}$. On appelle ***odds ratio*** (ou rapport de cote), le rapport :

$$\boldsymbol{\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}}$$

Ce rapport prend la valeur 1 si les variables $Z^1$ et $Z^2$ sont indépendantes. Il est supérieur à 1 si les individus de la ligne 1 ont plus de chance de se situer en colonne 1 que les individus de la ligne 2. Il est inférieur à 1 dans le cas contraire.

Supposons que 25% des assurés du produit 1 sont appétents au multi-équipement prévoyance tandis que seuls 15% des assurés du produit 2 le sont. L'odds des assurés du produit 1 est 1/3 tandis que celui des assurés du produit 2 est 3/17. L'odds ratio vaut alors $\frac{1/3}{3/17} = 17/9 \approx 1,89$. La chance d'être appétent au multi-équipement prévoyance est près de deux fois supérieure pour les assurés du produit 1 que pour les assurés du produit 2. Notons que cette chance est bien différente de l'« intuition » 25%/15% = 5/3.

L'odds ratio est également défini pour deux lignes *(a, b)* et deux colonnes *(c, d)* quelconques d'une table de contingence croisant deux variables. L'odds ratio est le rapport : $\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}}$ qui est estimé par l'odds ratio empirique $\hat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}$.

Dans le cas du modèle logistique simple, $OR = \Theta = \frac{P(Y=1 \,/\, X=1)/\, P(Y=0 \,/\, X=1)}{P(Y=1 \,/\, X=0)/\, P(Y=0 \,/\, X=0)}$, il vient :

$$\boldsymbol{\Theta = e^{\beta_1}}$$

L'odds ratio vaut l'exponentiel du paramètre, ce qui est très commode en pratique.

#### 4.3.1.3. *Interprétation des coefficients*

Comme on vient de le voir, les paramètres ne sont pas interprétables tels quels. Il faut calculer les $OR = \exp(\beta_k)$ et les comparer à 1.

Si X est une **variable qualitative binaire**, alors $OR = \exp(\beta_1)$ permet de comparer les individus qui possèdent le caractère X avec ceux qui ne le possèdent pas. Si $OR > 1$, Y est plus fréquent pour les individus exposés à X. Si $OR < 1$, Y est plus fréquent pour les individus non exposés à X. Si $OR = 1$, Y est indépendant de X.

Si X est une **variable quantitative**, alors $\exp(\beta_1)$ donne l'OR quand X augmente d'une unité et $\beta_0$ permet d'obtenir la proportion d'individus pour lesquels $Y = 1$ quand X vaut 0.

Le Logit du modèle multiple, c'est-à-dire plusieurs variables explicatives qualitatives, binaires ou non, et quantitatives, s'écrit : $\mathbf{Logit}(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$

L'interprétation des coefficients est similaire au cas du modèle simple. On compare toujours l'OR à 1. Si ce dernier est significativement différent de 1, la variable X a une influence sur la survenue de l'événement Y. Dans le cas contraire, X n'a aucune influence et n'est pas conservée dans le modèle.

#### *4.3.1.4. Estimations et test du modèle*

Les paramètres du modèle logistique sont estimés par la **méthode du maximum de vraisemblance**[51]**.** La vraisemblance d'un échantillon de taille n $(y_1, y_2 \ldots y_n)$ est définie comme la probabilité d'observer cet échantillon.

Les n variables $Y_i$ sont supposées indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre $\delta$.

Pour tous les individus i de 1 à n, $P(Y_i = y_i) = {\delta_i}^{y_i}(1 - \delta_i)^{1-y_i}$

Ce qui donne la vraisemblance suivante :

$$L(\delta, y_1 \ldots y_n) = \prod_{i=1}^{n} \delta^{y_i}(1-\delta)^{1-y_i}$$

Les estimateurs $\hat{\delta}$ maximisent $L(\delta, y_1 \ldots y_n)$ et sont obtenus par des procédures numériques. Il n'existe en effet pas d'expression analytique.

La matrice de variance-covariance $V(\hat{\delta}) = \begin{bmatrix} V(\hat{\delta}_i) & \cdots & Cov(\hat{\delta}_i, \hat{\delta}_j) \\ \vdots & \ddots & \vdots \\ Cov(\hat{\delta}_i, \hat{\delta}_j) & \cdots & V(\hat{\delta}_j) \end{bmatrix}$ est estimée par la matrice $\left[\frac{-\partial^2 Log\, L(\hat{\delta})}{\partial \delta^2}\right]^{-1}_{\delta=\hat{\delta}}$

[51] Estimateur du maximum de vraisemblance : https://fr.wikipedia.org/wiki/Maximum_de_vraisemblance

1ère étape : Test de significativité globale du modèle.

On envisage deux modèles : l'un sans variable explicative, uniquement avec la constante $\beta_0$ ; l'autre avec toutes les variables disponibles. L'idée est de comprendre si les variables exogènes influencent simultanément le risque de survenue de l'événement. Pour cela, on réalise un test de vraisemblance.

$$H_0 : M1 \Rightarrow \text{Logit}\,[\pi(x)] = P\,(Y = 1 / X = x) = \beta_0$$

$$H_1 : M2 \Rightarrow \text{Logit}\,[\pi(x)] = P\,(Y = 1 / X = x) = \beta_0 + \sum_{j=1}^{K} \beta_j X_j$$

Le modèle M2 présente-t-il des meilleures qualités prédictives que M1 ? Soit V la statistique de test : $V = -2 \ln \left[ \frac{vraisemblance\ M1}{vraisemblance\ M2} \right]$ qui suit asymptotiquement une loi du $\chi^2$ à K degrés de liberté.

Si $V > \chi^2_{1-\alpha}(K)$, on rejette $H_0$ avec une probabilité de se tromper de $\alpha$, et on conclut $H_1 : M2$ est meilleur que M1, la présence des variables exogènes disponibles améliore la prévision de Y.

Pour mémoire : $\alpha = P$ (rejeter $H_0$ / $H_0$ est vraie) et $1-\alpha$ est l'ordre du quantile de la loi du $\chi^2$ auquel est comparé la statistique de test.

2ème étape : significativité de chaque variable explicative

Il existe au moins deux méthodes pour tester l'apport d'une variable $X_j$ au modèle : le test de Wald et le rapport des vraisemblances.

**Le test de Wald** consiste à réaliser un test analogue au test de Student en régression usuelle.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Considérons la statistique w définie par $w = \frac{\hat{\beta}_j}{\hat{s}(\hat{\beta}_j)}$. $\hat{s}(\hat{\beta}_j)$ représente l'estimation de l'écart-type de l'estimateur de $\beta_j$. Sous l'hypothèse $H_0$, $w^2$ suit approximativement une loi du $\chi^2$ à $k-1$ degrés de liberté. On rejette $H_0$ si $w^2 > \chi^2_{1-\alpha}(k-1)$.

L'intervalle de confiance de $\beta_j$ s'obtient ainsi :

$$IC_{1-\alpha} = [\widehat{\beta_J} \pm u_{\alpha/2}.s(\widehat{\beta_J})]$$

avec $\alpha = 5\%$ et $IC_{1-\alpha} = IC_{95\%}$ le plus souvent ; et $u_{2.5\%} = 1{,}96$, le quantile d'ordre 2,5 % de la loi normale.

Si 1 appartient à l'intervalle de confiance, pas de relation entre X et Y. Au contraire, si l'IC ne contient pas 1, la variable X est significative et apporte de l'information au modèle.

Dans le cas du **test du rapport des vraisemblances**, l'apport de la variable X est mesuré à l'aide de la statistique : $V' = -2 \ln \left[ \frac{vraisemblance\ sans\ la\ variable\ X}{vraisemblance\ avec\ la\ variable\ X} \right]$ qui, sous l'hypothèse $H_0$, suit asymptotiquement une loi du $\chi^2$ à 1 degré de liberté.

#### *4.3.1.5. Adéquation du modèle*

En plus de la matrice de confusion, de la courbe ROC, et du critère AIC, déjà évoqués, la régression logistique propose un test supplémentaire pour déterminer la qualité d'ajustement du modèle aux données. C'est **le test de Hosmer et Lemeshow**[52]. Le principe consiste à évaluer la concordance entre les valeurs prédites et observées des observations regroupées en quantiles. Ce test dépend du nombre de groupes fixés a priori, et il est peu puissant en cas de mauvaise spécification.

En pratique, on regroupe les probabilités prédites par le modèle en dix groupes à l'aide des déciles. Pour chaque groupe, on observe l'écart entre les valeurs prédites et les valeurs observées. L'importance de la distance entre ces valeurs est évaluée grâce une statistique du $\chi^2$ à 8 degrés de liberté qui teste $H_0$ : la distance est faible *vs* $H_1$ : la distance est élevée.

#### *4.3.1.6. Conclusion*

La régression logistique est historiquement la méthode d'analyse prédictive la plus usitée lorsqu'il s'agit d'expliquer la survenue ou l'existence d'un phénomène. C'est particulièrement le cas dans l'industrie et en épidémiologie. Elle présente en effet de **nombreux avantages**, entre autres :

- Elle permet de tester la pertinence des variables discrètes, nominales ou ordinales, aussi bien que des variables continues, en classe ou non.
- Les conditions d'utilisation sont moins restrictives que pour les autres méthodes traditionnelles.
- Les modèles produits sont concis et facilement implémentables dans les logiciels.
- Elle propose de nombreux tests statistiques de significativité et des intervalles de confiance pour les paramètres.
- Enfin, elle offre la possibilité d'une sélection pas à pas des prédicteurs et fournit des modèles très souvent précis.

Toutes **ces qualités expliquent la popularité** de la régression logistique. N'omettons pas toutefois ses quelques inconvénients ; notamment :

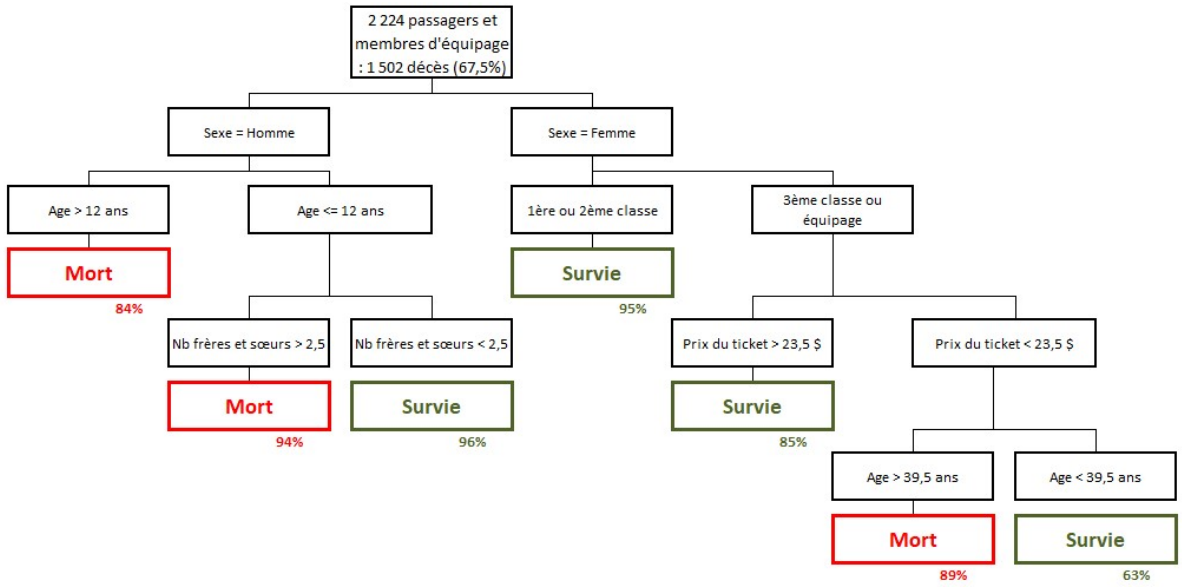
- L'exigence de non-colinéarité des variables explicatives.
- L'approximation numérique qui exige un calcul itératif coûteux en termes de temps de traitement.
- Et le fait que l'algorithme associé ne converge pas toujours vers une solution optimale, spécifiquement en cas de séparation totale des groupes d'individus.

[52] https://en.wikipedia.org/wiki/Hosmer%E2%80%93Lemeshow_test
Hosmer D.W., Lemeshow S., *Applied logistic regression*, Wiley Series in Probability and Mathematical Statistics, 2000

### 4.3.2. Les arbres de décision[53]

Les arbres de décision représentent **une des solutions les plus intuitives et populaires** d'apprentissage supervisé. La raison principale est que **le résultat produit est très visuel et facilement intelligible** par l'homme contrairement à d'autres approches qui conduisent à des prédicteurs typés « boîtes noires ». Les arbres fonctionnent selon **un enchaînement de décisions** que l'on prend à chaque étape, les « nœuds », tout au long des « branches », jusqu'à atteindre les extrémités, les « feuilles ». Ils conviennent aussi bien aux problèmes de classement que de régression.

Voici un exemple d'arbre de décision construit à partir d'un jeu de données très connu. Il s'agit de prédire la survie ou non des 2 224 passagers et membres d'équipage du Titanic en fonction de 6 variables (âge, sexe, classe, prix du ticket, port d'embarquement, nombre de membres dans la famille). Cet exercice a fait l'objet d'un challenge *Kaggle* en 2012.[54] Une solution possible parmi tant d'autres :




Fondamentalement, **un arbre de décision modélise une hiérarchie de tests** sur les valeurs d'un ensemble de variables appelées « attributs ». À l'issue de ces tests, le prédicteur produit, selon que la variable à expliquer soit quantitative ou nominale, une valeur numérique ou choisit un élément dans un ensemble discret de conclusions, le plus souvent binaire comme dans l'exemple ci-dessus.

**Chaque chemin partant de la racine vers les feuilles constitue une règle.** Dans notre exemple, on dénombre 7 règles. Pour construire l'arbre, il existe plusieurs types d'algorithme comme *CART, CHAID* ou *C5.0* mais la plupart des approches suivent le paradigme « diviser pour régner ». Ce qui induit **deux principes essentiels** :

---

[53] Ouvrage de référence : « Classification and Regression Trees », L. Breiman, J. Friedman, & col. (1984).
Tutoriels : https://www.rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-rakotomalala-33/rakotomalala-33-tutorial.pdf et http://sipina-arbres-de-decision.blogspot.fr/p/references.html

[54] https://www.kaggle.com/c/titanic

1. Les règles sont mutuellement exclusives, c'est-à-dire qu'un individu ne peut déclencher qu'une règle et une seule.
2. L'ensemble des règles couvre tout l'espace des valeurs possibles, c'est-à-dire que tout individu à classer va déclencher une de ces règles.

*Exemple de lecture* : Si Sexe = 'Femme' et si Classe = '1ère' ou '2ème' alors l'arbre prédit la survie de l'individu. Ce qui est vrai dans l'échantillon d'apprentissage à 95%.

On peut aussi imaginer lire l'arbre comme une cascade de règles imbriquées sous la forme :

Si Sexe = 'Femme' alors
    si Classe = '1ère' ou '2ème' alors décision = **'Survie'**
    sinon
        si Prix du ticket > 23,5 $ alors décision = **'Survie'**
        sinon
            si âge > 39,5 ans alors décision = **'Mort'**
            sinon décision = **'Survie'**
sinon
    si âge > 12 ans alors décision = **'Mort'**
    sinon
        si Nb frères et sœurs > 2,5 alors décision = **'Mort'**
        sinon décision = **'Survie'**

C'est un mode de lecture très pratique surtout si l'on a besoin d'implémenter ces règles sur un tableur.

Le **principe** de construction d'un arbre consiste à **déterminer une séquence de nœuds**. Au nœud initial, la racine, nous avons la totalité des individus de l'échantillon d'apprentissage. Si chaque nœud de l'arbre a au plus deux nœuds-fils, on dit que l'arbre est binaire. C'est de loin le cas le plus répandu. A chaque nœud, **l'algorithme choisit une variable** parmi les explicatives disponibles **et un seuil**, si la variable est quantitative, **ou un groupe de modalités**, si elle est nominale, **pour diviser l'échantillon en plusieurs[55] sous-groupes homogènes**. La variable X* choisie est la première disponible **la plus corrélée** avec la variable à expliquer Y. On peut[56] utiliser pour quantifier cette liaison, la grandeur du $\chi^2$ calculée sur le tableau de contingence X* x Y. Le $\chi^2$ augmente mécaniquement avec la taille de la population n, le nombre de lignes K et de colonnes L. Si bien que l'on préférera la statistique définie par :

$$t^2_{X*,Y} = \frac{\chi^2_{X*,Y}}{n\sqrt{(K-1)(L-1)}}$$

Dans le cas d'une prédiction, au lieu de discrétiser la variable quantitative pour obtenir un tableau de contingence, on peut utiliser le test de Fisher[57]. Dans cette éventualité, le critère obtenu doit faire en sorte que la variable à expliquer ait une variance plus faible dans les nœuds-fils que dans le nœud-père, et sa moyenne doit être la plus distincte possible d'un nœud-fils à l'autre.

[55] Deux sous-groupes dans le cas d'un arbre binaire.
[56] D'autres méthodes existent.
[57] Pour mémoire : http://www.jybaudot.fr/Inferentielle/testf.html

Nous devons ensuite **déterminer si ce nœud est terminal**, il deviendrait alors une feuille, ou non. Si c'est le cas, il faut **affecter à cette feuille, une classe ou une valeur** de la variable à expliquer. **Sinon, le processus est réitéré** jusqu'à n'obtenir que des branches terminées par des feuilles. D'une manière générale, la croissance de l'arbre s'arrête à un nœud lorsque celui-ci est considéré **homogène**, c'est-à-dire qu'il n'existe plus de partition pertinente ou bien si les effectifs restants à discriminer sont jugés trop faibles. Concernant l'affectation, si Y est continue, la valeur retenue est **la moyenne des observations associées à cette feuille**. Si Y est discrète, la classe affectée est **la plus représentée ou la plus probable** au sens bayésien si les probabilités a priori sont connues. En cas d'ex-æquo, on choisira la classe la moins coûteuse si les coûts de mauvais classement sont donnés.

Parfois, les modèles obtenus sont si complexes qu'ils peuvent générer des prévisions très instables car trop dépendantes des échantillons qui ont permis leur estimation. On essaye alors de rechercher des modèles plus parcimonieux et par conséquence plus robustes. Il faut dans ce cas procéder à un **élagage de l'arbre**. Il existe un algorithme pas à pas qui identifie le sous-arbre optimal parmi ceux obtenus après élagage successif des feuilles. Mais on peut aussi décider de se ranger à dires d'expert pour nous conduire à l'arbre qui sera jugé le plus efficace ou le plus fiable dans le cadre d'un déploiement métier opérationnel.

Un autre danger serait de mettre par mégarde dans les variables explicatives, une variable qui est directement corrélée à la variable à expliquer. Par exemple, dans le cas de la prédiction du churn, la date de fin de contrat du client. Une possibilité pour éviter ce piège est d'observer la variable à expliquer sur une période de temps qui ne recouvre pas la période d'observation des variables explicatives.

Les arbres de décision sont une méthode qui présente de nombreux avantages : lisibilité, intervention possible du praticien ou d'un expert, traduction éventuelle vers une base de règle, sélection automatique des variables discriminantes, adapté au classement et à la régression, robuste face aux données aberrantes ou manquantes, et traitement relativement rapide. Malgré tout, ils font face à deux **défauts majeurs** :

- empiriquement, les arbres affichent des **résultats moins performants en général** par rapport aux autres méthodes, en particulier dans le cadre d'un scoring
- la structure des arbres est généralement instable : une petite variation de l'échantillon d'apprentissage peut générer des transformations importantes.

### 4.3.3. Le classifieur Bayésien naïf[58]

Le classifieur bayésien naïf **trouve son fondement dans le théorème de Bayes** qui est un des plus importants de la théorie des probabilités, et aussi le théorème fondamental de la statistique bayésienne. Considérons deux événements A et B, ce théorème lie la probabilité conditionnelle de A sachant B et la probabilité conditionnelle de B sachant A par la relation :

$$P(A/B) = P(B/A)\frac{P(A)}{P(B)}$$

Si par ailleurs, un ensemble de $A_n$ événements définit une partition, alors on peut s'écrire :

$$P(B) = \sum_{k=1}^{n} P(B/A_k)P(A_k)$$

Le théorème de Bayes se réécrit alors, pour tout i :

$$P(A_i/B) = P(B/A_i)\frac{P(Ai)}{\sum_{k=1}^{n} P(B/A_k)P(A_k)}$$

Un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. On pourrait ainsi spécifier le modèle probabiliste sous-jacent comme étant un « **modèle à caractéristiques statistiquement indépendantes ».**

Pour résoudre ce modèle, nous cherchons à estimer la probabilité conditionnelle

$$P(Y = y_k/X) = \frac{P(Y = y_k)\ P(X/Y = y_k)}{P(X)}$$

Avec $X = (X_1 \dots X_n)$, le vecteur des variables explicatives que l'on suppose pour le moment toutes qualitatives.

En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de Y et les valeurs des dimensions X sont connues. Le dénominateur est donc en réalité constant.

Déterminer la conclusion revient donc à chercher le maximum : $y_{k^*} = \arg max_k\ P(Y = y_k/X)$.
C'est-à-dire : $y_{k^*} = \arg max_k\ P(Y = y_k)\ P(X/Y = y_k)$

$P(Y = y_k)$ est facilement estimable par les fréquences $\frac{n_k}{n}$ mais l'estimation de $P(X/Y = y_k)$ est impossible par les fréquences, le tableau serait trop grand et rempli de zéros. Toutefois $P(X/Y = y_k)$ est soumis à la loi de probabilité à plusieurs variables et peut être factorisé en utilisant la définition de la probabilité conditionnelle. Nous intégrons en outre l'hypothèse naïve : **les descripteurs sont deux à deux indépendants, conditionnellement aux valeurs prises par Y.**

[58] Références : https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne ou blog de Ricco Rakotomalala

$$P(X_i/Y = y_k, X_j) = P(X_i/Y = y_k), \text{ pour tout i différent de j}$$

Par conséquent,

$$P(X/Y = y_k) = \prod_{i=1}^{n} P(X_i/Y = y_k)$$

Et donc,

$$P(Y = y_k/X) = \frac{1}{C} P(Y = y_k) \prod_{i=1}^{n} P(X_i/Y = y_k)$$

Avec C, une constante qui dépend uniquement des descripteurs, donc totalement connue et n le nombre de descripteurs.

Nous avons ainsi **simplifié le problème.** S'il existe k classes pour Y et si le modèle pour chaque fonction $P(X_i/Y = y_k)$ peut être exprimé selon r paramètres, alors le modèle bayésien naïf correspondant dépend de (k-1) + n r k paramètres. Dans le cas répandu d'une classification binaire (k=2) discriminée par des booléens (r=1), le nombre total de paramètres à estimer du modèle ainsi décrit n'est plus que de 2n+1. Ces paramètres sont estimés par la **méthode du maximum de vraisemblance des probabilités**, ce qui implique que l'on peut utiliser le prédicteur bayésien naïf sans finalement se soucier de la théorie bayésienne.

Dans le cas de variables explicatives continues, on se ramène au cas précédent par une mise en classe des prédicteurs quantitatifs. C'est l'approche à privilégier si on a affaire à un mélange de prédicteurs quali et quanti. Il existe une procédure très utile et facile d'utilisation sur SPAD qui renvoie les bornes pertinentes conditionnellement à la variable à expliquer. Si les prédicteurs sont tous quantitatifs, on peut aussi passer par une démarche paramétrique en posant des hypothèses de distribution pour les probabilités conditionnelles.

Travailler à partir d'un produit de probabilités engendre malheureusement au moins deux problèmes : si une des probabilités est nulle, le modèle ne peut pas estimer la probabilité finale, c'est le phénomène bien connu de « séparation complète », présent également en régression logistique. Le 2ème concerne la lisibilité du modèle. Il n'est souvent pas suffisamment explicite, ce qui explique pourquoi il est peu utilisé opérationnellement pour des opérations marketing par exemple. Il existe toutefois une astuce, celle de passer par le logarithme. En effet, chercher

$$y_{k^*} = \arg max_k \, P(Y = y_k) \prod_{i=1}^{n} P(X_i/Y = y_k)$$

est équivalent à chercher

$$y_{k^*} = \arg max_k \, [\ln P(Y = y_k) + \sum_{i=1}^{n} \ln(P(X_i/Y = y_k))]$$

La recherche de l'optimum est un peu plus fastidieuse mais nous avons ainsi levé les deux écueils précédents.

Pour finir, la sélection et l'évaluation de la pertinence des variables ainsi que la recherche de parcimonie en éliminant les variables non pertinentes et en supprimant les redondances se déroulent de façon plus ou moins similaire à ce qui a déjà été évoqué.

Malgré les hypothèses d'indépendance très simplistes et leurs fortes dépendances vis-à-vis du modèle théorique, le classifieur bayésien naïf est parfois étonnamment performant et obtient un certain succès auprès des chercheurs et des praticiens. Il fait notamment preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes.

Le modèle améliore sensiblement les performances des modèles bayésiens « purs », il est relativement robuste et particulièrement pertinent dans le cas de traitement des très grandes bases de données, l'hypothèse d'indépendance permettant de s'affranchir du calcul de la matrice de variance-covariance. Enfin, il affiche une belle incrémentalité, c'est-à-dire que chaque dimension supplémentaire apporte une amélioration sans altérer la performance globale. En revanche, il souffre le plus souvent, tout de même un peu par rapport aux performances des nouveaux algorithmes de type 'agrégation de modèles'.

### 4.3.4. L'analyse discriminante[59]

L'analyse discriminante (AD) de Fisher (1936) a été la principale méthode de classement avant de se faire concurrencer par la régression logistique à la fin des 50's. **L'analyse discriminante prédictive (ADP) est un prolongement de l'analyse factorielle discriminante (AFD)**. Tandis que la première est adaptée aux problèmes de régression (trouver des règles d'affectation des individus à leur groupe, c'est-à-dire des règles de prédiction des modalités de Y à partir des valeurs des $X_i$), la seconde est uniquement descriptive et ne s'occupe que des problèmes de classement (trouver une représentation des individus qui sépare le mieux les groupes, c'est-à-dire une représentation des liaisons entre les Y et les $X_i$).

Les analyses discriminantes sont des méthodes applicables uniquement sur des **variables explicatives continues**, qui permettent d'expliquer les modalités d'une variable endogène qualitative. Elle est en cela **le symétrique de l'ANOVA**[60], où la variable à expliquer est numérique et les variables à expliquer sont nominales.

**Un prolongement** inventé par Gilbert Saporta sous le nom de **Méthode Disqual**[61] (DIScrimination sur variables QUALitatives) peut toutefois traiter aussi des variables qualitatives transformées. Le principe est très malin : il consiste à réaliser une AFCM[62] sur le tableau disjonctif complet, puis à récupérer les coordonnées (continues) des individus sur les axes factoriels les plus discriminants et enfin à injecter ces coordonnées en entrée d'une analyse discriminante linéaire classique.

Comme la méthode précédente, **l'ADP s'appuie sur l'approche bayésienne** puisqu'elle se fonde également sur le théorème de Bayes (voir 4.3.3. Bayésien naïf). L'idée générale est donc d'essayer de calculer $\mathrm{P}\,(\mathrm{Y} = \mathrm{y_k}/\mathrm{X})$. Trois moyens peuvent nous y conduire :

- une méthode non paramétrique (pas d'hypothèse sur la fonction de densité mais une approche par le rapport fréquence / volume) : fastidieux et certainement malhabile
- une approche semi-paramétrique dans laquelle on écrit $\mathrm{P}(\mathrm{Y} = \mathrm{y_k}/\mathrm{X}) = \frac{e^{\alpha x+\beta}}{1+e^{\alpha x+\beta}}$
- **une méthode paramétrique** qui suppose la multinormalité de $\mathrm{X}/\mathrm{Y}$ et l'homoscédasticité[63] des variances conditionnelles, où comme au chapitre précédent, pour la même raison, on cherche en pratique à estimer $\mathrm{P}\,(\mathrm{X}\,/\,\mathrm{Y} = \mathrm{y_k})$

Dans ce dernier cas, le plus répandu, l'**hypothèse de normalité multidimensionnelle** donne pour la distribution des nuages de points conditionnels :

$$f_k(X) = \mathrm{P}\,(\mathrm{X}\,/\,\mathrm{Y} = \mathrm{y_k}) = \frac{1}{|W_k|^{1/2}\,(2\pi)^{j/2}}\, e^{-\frac{1}{2}t(X-\mu_k){W_k}^{-1}(X-\mu_k)}$$

---

[59] Ouvrage de référence : Trevor Hastie, Robert Tibshirani, Jerome H. Friedman : The Elements of Statistical Learning (2009).
[60] ANalysis Of VAriance : https://fr.wikipedia.org/wiki/Analyse_de_la_variance
[61] Gilbert Saporta : Liaisons entre plusieurs ensembles de variables et codage de données qualitatives, (1975).
[62] Voir 6.4.1. Analyse Factorielle des Correspondances Multiples
[63] La variance des erreurs d'une régression est supposée la même pour chaque observation, par opposition à l'hétéroscédasticité où la variance de l'erreur des variables est différente.

Où $|W_k|$ représente le déterminant de la matrice de variance-covariance conditionnellement à $y_k$.

L'objectif est désormais de déterminer le maximum de la probabilité a posteriori d'affectation. Nous pouvons alors négliger tout ce qui ne dépend pas de k.

En appliquant le logarithme à la relation de Bayes, nous obtenons le score discriminant proportionnel à $\ln [P (Y = y_k/X)]$ : [64]

$$D (Y = y_k, X) = 2 \ln [P (Y= y_k)] - \ln |W_k| - {}^t(X - \mu_k) W_k^{-1}(X - \mu_k)$$

La règle d'affectation devient donc : $Y(\omega)=\arg max_k D (Y = y_k, X(\omega))$.

Si l'on développe complètement le score discriminant D, nous constatons qu'il s'exprime en fonction du carré et du produit croisé entre les variables prédictives. On parle alors d'analyse discriminante quadratique.

La seconde hypothèse concerne l'**homoscédasticité** : les matrices de variances-covariances sont identiques d'un groupe à l'autre. Géométriquement, cela veut dire que les nuages de points ont la même forme dans l'espace de représentation. La matrice de variance-covariance estimée est dans ce cas la matrice de variance covariance intra-classes calculée à l'aide de l'expression suivante :

$$W = \frac{1}{n - K} \sum_{k=0}^{K} n_k W_k$$

De nouveau, nous pouvons évacuer du score discriminant tout ce qui ne dépend plus de k, il vient :

$$D (Y = y_k, X) = 2 \ln [P (Y= y_k)] - {}^t(X - \mu_k) W_k^{-1}(X - \mu_k)$$

**Les 2 hypothèses**, relativement fortes, **ont permis de simplifier le problème**. Il nous reste ainsi **K fonctions linéaires discriminantes** qui sont de la forme :

$$D (y_k, X) = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

Ce sont **les fonctions de Fisher** qui sont particulièrement intéressantes puisque cela revient à attribuer une note à chacune des modalités. On peut ainsi déterminer le sens des causalités dans le classement et évaluer le rôle significatif des variables dans la prédiction.

En plus des mesures classiques basées sur les calculs d'aires, deux indicateurs sont spécifiquement utilisés en AD pour évaluer la qualité du modèle :

- Le coefficient de détermination $R^2$, qui représente le rapport de la variance expliquée par le modèle sur la variance totale, plus il est proche de 1, meilleur est le modèle. On lui préfèrera le $R^2$ ajusté $= 1 - \frac{(1-R^2)(n-1)}{n-p-1}$ car $R^2$ est trop optimiste (il croît mécaniquement avec le nombre de variables), avec n = nombre d'observations et p = nombre de descripteurs.

[64] Source : https://fr.wikipedia.org/wiki/Analyse_discriminante_lin%C3%A9aire

- **Le lambda de Wilks.**

Cet indicateur représente le rapport $\lambda = \det(W) / \det(V)$, c'est-à-dire le rapport entre le déterminant de la matrice des covariances intraclasse sur celui de la matrice des covariances totale. Il varie donc entre 0 et 1. Plus il est bas, meilleur est le modèle. S'il est proche de 1, tous les barycentres (ou « centroïdes ») sont égaux, ce qui n'apporte pas d'information. Le lambda de Wilks permet de répondre aux deux questions qui se posent au sujet des variables explicatives : Ces variables permettent-elles de discriminer les classes ? Pourrait-on les discriminer avec moins de variables ?

L'analyse discriminante est une méthode qui fournit des résultats sous une forme très pratique : les coefficients des modalités sont comparables puisqu'il s'agit d'indicatrices. Il n'y a pas d'effet d'ordre de grandeur, les coefficients sont naturellement normalisés.

Elle présente d'autres avantages : **les calculs sont très rapides et les prédictions explicites, précises et robustes**. En outre, elle permet des **généralisations vastes et profondes**, et il est facile d'intégrer les coûts d'erreur de classement.

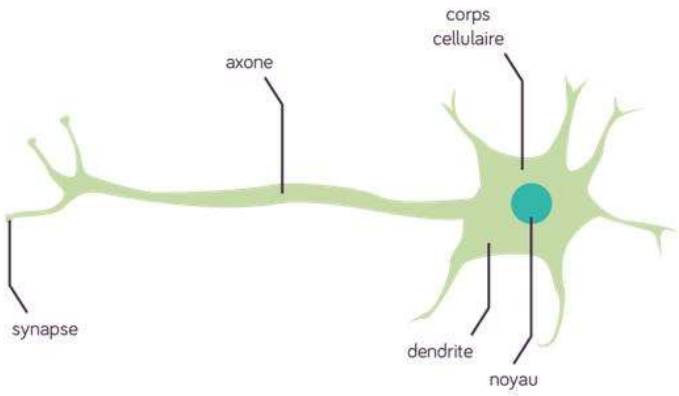
En revanche, l'optimalité des algorithmes n'est atteinte que sous vérification des hypothèses d'homoscédasticité, de multinormalité et d'indépendance linéaire des variables explicatives. D'autre part, l'AD est très sensible aux points extrêmes, et ne propose pas de test statistique de significativité des coefficients.

Pour faire face au cas fréquent d'hétéroscédasticité des variances, une idée consiste à **segmenter la population au préalable** afin de travailler sur des populations homogènes. On construit alors un modèle par segment avant de tout synthétiser.

### 4.3.5. Les réseaux de neurones[65]

Un réseau de neurones (RN) est une modélisation qui s'inspire fortement du fonctionnement du cerveau et du système nerveux en simulant au plus près des connaissances actuelles, les neurones, ces « unités de calcul » que nous avons par milliards en chacun de nous. Les RN sont probablement la méthode la plus difficile à vulgariser. Au-delà de la complexité inhérente aux algorithmes RN, la raison vient principalement du vocabulaire. Or celui-ci est calqué sur le vocabulaire des réseaux de neurones biologiques. Il faut donc avant de démarrer, bien maîtriser ce dernier.

Schéma d'un neurone biologique et description (très) simplifiée :




Les neurones sont des cellules qui permettent le fonctionnement du système nerveux. Ils assurent la transmission d'un signal électrique (potentiel d'action) que l'on nomme « influx nerveux ». Ils sont composés :

- d'un **corps** qui contient le noyau, bloqué et donc incapable de se diviser, et le cytoplasme,
- d'un **axone**, unique, long prolongement, qui se termine en se ramifiant par des **synapses**,
- de **dendrites**, des extensions nombreuses, courtes et très ramifiées.

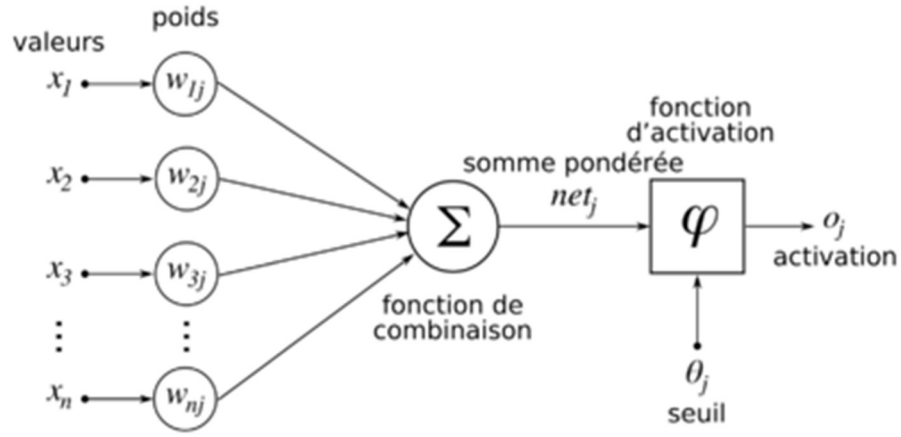
Le corps traduit le signal interne entre les dendrites et l'axone via un « cône d'action ». Les dendrites conduisent le signal des synapses vers le noyau (réception) tandis que l'axone conduit le signal du noyau vers les synapses (transmission). Les synapses sont des relais qui assurent la transmission de l'influx nerveux entre les axones et les dendrites à travers des neurotransmetteurs. Ceux-ci sont réceptionnés et convertis en signal qui est dépolarisé (effet inhibiteur) ou hyperpolarisé (effet excitateur) par des récepteurs post-synaptiques présents sur les dendrites.

On peut donc résumer le chemin d'un influx nerveux de la façon suivante : Le neurone n°1 transmet un signal par l'axone n°1. Il est polarisé puis réceptionné par les dendrites du neurone n°2 via les neurotransmetteurs émis par les synapses du n°1. Le signal est remonté par les dendrites au corps n°2. Le cône d'action n°2 teste la polarité (traduction) et envoie la réponse appropriée à la base de l'axone n°2. Ce dernier transmet l'information au neurone n°3. Etc.

[65] Ouvrage de référence : Dreyfus, Martinez & col. « Réseaux de neurones : Méthodologie et applications » (2002).

**Un neurone formel**, ou un nœud, **est une représentation** mathématique et informatique d'un neurone biologique. Il possède généralement plusieurs entrées (réf. les dendrites) et une sortie (réf. la base de l'axone). Les actions excitatrices et inhibitrices des synapses sont représentées par des coefficients numériques (les poids) et la fonction d'activation figure le cône d'action.

Sous sa forme la plus simple, le premier modèle proposé par W. McCulloch et W. Pitts en 1943, **un neurone formel binaire**, i.e. la sortie vaut 0 ou 1, **calcule une somme pondérée de ses entrées puis applique une fonction d'activation à seuil**, c'est-à-dire un test : si la somme pondérée dépasse une certaine valeur, la sortie du neurone est 1, sinon elle vaut 0.




Cas d'un neurone seul :

On note $(x_i)$, $i \in [1;k]$, les $k$ entrées, et $w_i$, le poids (coefficient) lié à l'information $x_i$. La $i^{ème}$ information qui parvient au neurone sera donc en fait $w_i * x_i$. On intègre un poids supplémentaire $w_0$ qui va représenter le coefficient de biais et qui est associé à l'information $x_0 = -1$.

Le neurone ne va pas traiter séparément chacune des $k$ informations. Il considère uniquement la somme pondérée que l'on nomme *sump*.

$$sump = \sum_{i=0}^{k} w_i\, x_i = \left(\sum_{i=1}^{k} w_i\, x_i\right) - w_0$$

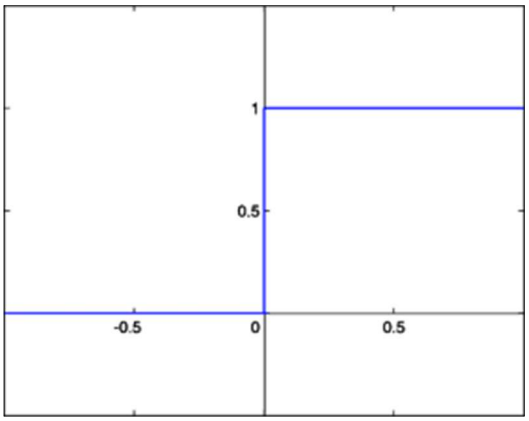
Cette donnée passe par la fonction d'activation, ou fonction de transfert, $g$, qui renvoie un réel compris entre 0 et 1. Ce réel est la sortie du neurone et est notée $a$. Lorsque $a$ est proche de 1, on dit que l'unité de traitement (le neurone) est active tandis que lorsque $a$ est proche de 0, on dit que l'unité est inactive. Si la fonction d'activation est linéaire, le réseau de neurone se réduit ainsi à une simple régression multi-linéaire, i.e. un cas sans bénéfice par rapport à l'existant par ailleurs. Pour cette raison, on privilégiera des fonctions d'activation non linéaires. En résumé :

$$a = g\,(sump) = g\left[\left(\sum_{i=1}^{k} w_i\, x_i\right) - w_0\right]$$

Les deux fonctions de transfert les plus utilisées sont :

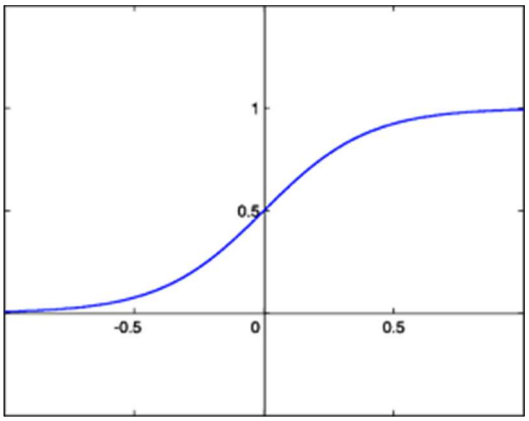
la fonction de Heaviside

$\forall x \in \mathbb{R}, g(x) = 1 \text{ si } x \geq 0, 0 \text{ sinon}$




la fonction sigmoïde

$\forall x \in \mathbb{R}, g(x) = \frac{1}{1+e^{-x}}$




La fonction sigmoïde présente un double avantage : elle est dérivable deux fois, ce qui est utile pour la suite, et elle renvoie potentiellement toute valeur comprise entre 0 et 1 contrairement à la fonction de Heaviside qui ne renvoie que 0 ou 1. Les deux fonctions possèdent un seuil en $x = 0$, qui vaut 1 dans le premier cas et ½ dans le second. Le retour de la fonction d'activation est comparé à ce seuil : s'il est inférieur, le neurone renvoie 0 ; s'il est supérieur, il renvoie 1.

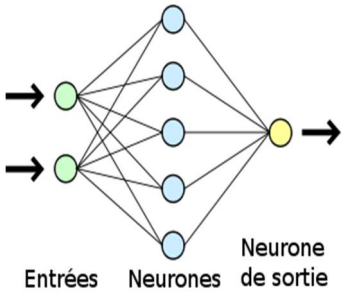
Notons X, le vecteur des informations en entrée et W, le vecteur des poids.

$$W \,.\, X = 0 \iff \text{sump} = 0 \iff \left(\sum_{i=1}^{k} w_i \, x_i\right) = w_0$$

La fonction d'activation atteint un seuil lorsque le produit des deux vecteurs vaut 0, c'est-à-dire lorsque la somme pondérée des entrées vaut le coefficient de biais. On dit que le neurone est actif quand $\text{sump} \geq 0$ et inactif quand $\text{sump} < 0$. En conséquence, puisque les fonctions de transfert sont croissantes, le neurone est actif lorsque $a = g\,(sump)$ est supérieur ou égal au seuil, c'est-à-dire à $g(0)$, et inactif dans le cas contraire.

Dans le cas où un hyperplan, espace de dimension k-1, est en capacité de séparer parfaitement les points appartenant à deux classes différentes, un seul neurone suffit. Dans le cas contraire, le plus fréquent de loin, il faut faire appel à un réseau de neurones.

Le perceptron monocouche :



Le perceptron est un réseau de *p* neurones qui possède $n$ informations en entrée. Chacune des *n* informations est connectée aux *p* neurones.

$w_{i,j}$ est le poids reliant l'information $x_i$ au neurone j ; $w_{0,j}$, $sump_j$ et $a_j$ sont respectivement le seuil, la donnée en entrée, et l'activation du $j^{ème}$ neurone. On a donc l'équation suivante :

$$\forall j \in [1 ; p],\ a_j = g(sump_j) = g[(\sum_{i=1}^{n} w_{i,j}\, x_i) - w_{0,j}]$$

Propriété importante des perceptrons : quelles que soient les informations en entrée, il existe au moins un algorithme d'apprentissage qui permet d'adapter ses poids de façon à obtenir pour les entrées données, le classement parfait. Si l'échantillon d'apprentissage est assez vaste, cette propriété garantit sur l'échantillon test, des résultats convenables généralement. Mais gare au sur-apprentissage !

Il existe de nombreux types d'algorithmes. Retenons deux des plus simples : l'algorithme d'apprentissage par descente du gradient et l'algorithme de Widrow-Hoff.

L'idée consiste à définir une fonction d'erreur qui minimise l'erreur quadratique E, plutôt que de classifier correctement tous les exemples.

$$E = \frac{1}{2} \sum_{k=1}^{N} (y_k - s_k)^2$$

Avec N, nombre d'exemples, $y_k$, la sortie attendue et $s_k$ la sortie obtenue, pour le $k^{ème}$ exemple. L'algorithme démarre avec des poids tirés au hasard. L'erreur est nulle si le réseau ne se trompe sur aucun exemple, ce qui n'arrive jamais. Choisissons $\alpha$, un nombre réel que l'on appelle le taux d'apprentissage.

Soient $w_i$ l'ancien poids, $w_{i2}$ le nouveau poids et $x_i$ l'information correspondante. Les poids sont modifiés de la façon suivante :

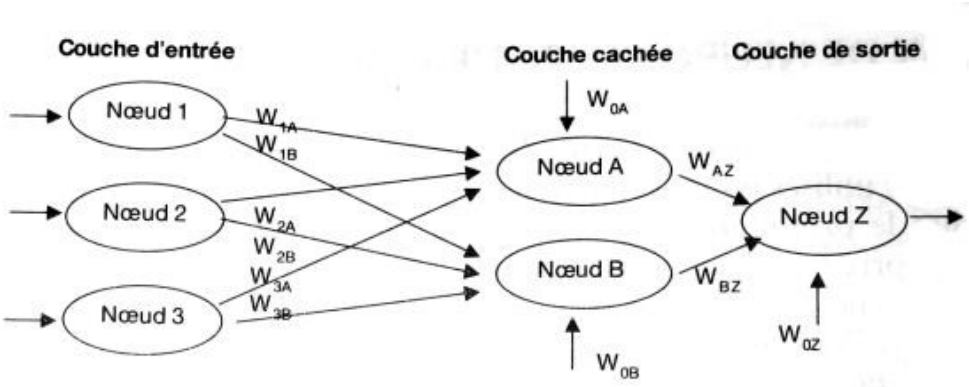
$$w_{i2} = w_i + \alpha * (y_k - s_k) * x_i$$

Pour faire très simple, l'algorithme de descente du gradient modifie les poids synaptiques après le traitement de tous les exemples tandis que celui de Widrow-Hoff les modifie un à un après chaque exemple. Le second est donc plus efficace et également plus simple à appliquer.

Les cas où le perceptron monocouche suffit à séparer linéairement les observations sont encore trop rares. Bien souvent, il est nécessaire de complexifier l'espace de représentation, ce que permettent les réseaux multicouches.

<u>Le perceptron multicouche :</u>

Le fonctionnement du perceptron multicouche est plus complexe mais grossièrement il s'agit d'une généralisation du modèle précédent. Dans ce cas, les neurones de la première couche reçoivent toutes les informations entrées, ceux de la deuxième reçoivent toutes les sorties des neurones de la première couche, et ainsi de suite jusqu'au neurone de sortie qui reçoit celles de la dernière couche. Les couches intermédiaires entre la couche d'entrée et la couche de sortie sont appelées « couches cachées ».




Généralement 3 à 4 couches, c'est-à-dire 1 ou 2 couches cachées, suffisent à résoudre un très grand nombre de problèmes de fonctions complexes. Les couches cachées permettent en effet beaucoup plus d'interactions. Le fonctionnement de chaque nœud demeure toutefois le même : choix de la fonction d'activation, évaluation de la sortie d'un neurone, correction des erreurs par modification itérative des poids.

L'algorithme d'apprentissage le plus commun se nomme la « rétro-propagation du gradient ». C'est un algorithme récursif inspiré de celui de Widrow-Hoff qui a pour but de minimiser l'erreur quadratique moyenne commise par l'ensemble du réseau. Mais comme dans ce cas, on ne connaît pas les outputs espérés des couches cachées, il faut propager la responsabilité des erreurs de la dernière couche à la première, dans le sens contraire de l'exécution du réseau, d'où le nom.

*Source : Blogs de Alp Mestan*[66] *et d'Emmanuel Istace*[67]

L'avantage principal des RN réside dans leur puissance de modélisation. Ils peuvent en effet approcher n'importe quelle fonction suffisamment régulière, ce qui permet de traiter les problèmes les plus complexes et les cas où les données sont difficiles à appréhender par les autres méthodes.

Ils sont très prisés des chercheurs car la recherche fondamentale sur ce thème est loin d'être terminée. Une des forces est en effet sa grande adaptabilité. On peut notamment y accoler les stratégies modernes *bagging* et *boosting* pour améliorer encore les performances.

Les RN souffrent toutefois de nombreux inconvénients ; le principal étant l'effet « boîte noire ». Les processus qui conduisent à une prédiction ou un classement sont totalement illisibles et impossibles à expliciter, ce qui peut générer de la perplexité voir du désarroi de la part des fonctions métiers. En outre, ils présentent des risques de sur-apprentissage supérieurs aux autres méthodes et un risque de convergence vers une solution globalement non optimale.

Une extension des RN : **les cartes auto-organisatrices de Kohonen** qui sont utilisées pour cartographier un espace réel, c'est-à-dire pour étudier la répartition de données dans un espace à grande dimension.

[66] http://alp.developpez.com/tutoriels/intelligence-artificielle/reseaux-de-neurones/

[67] https://istacee.wordpress.com/2010/03/05/vulgarisation-des-reseaux-neuronaux/

### 4.3.6. Support Vector Machine[68]

Les *Support Vector Machines* (machines à vecteurs de support ou séparateurs à vaste marge, en français) sont un ensemble de méthodes d'apprentissage supervisé basées sur **une séparation des données par des marges ou hyperplans.** L'idée générale consiste à déterminer un séparateur linéaire qui divise l'espace en deux en passant par un espace de dimension supérieure : l'espace des exemples positifs et l'espace des exemples négatifs.

Les SVM répondent à des problématiques de classement comme de régression. Ils sont parfois comparés aux réseaux de neurones (RN) car ils présentent des performances théoriquement proches. D'aucuns ont prétendu que l'on pouvait interpréter les SVM comme une généralisation des réseaux neuronaux car le choix d'un noyau particulier pouvait faire penser à un cas de RN. C'est au mieux un raisonnement bancal car les fondements des deux méthodes sont très différents, ceux du SVM étant plus proche des méthodes statistiques classiques. S'il fallait absolument présenter les SVM **comme une généralisation**, ce serait davantage celle **de l'analyse discriminante linéaire.**

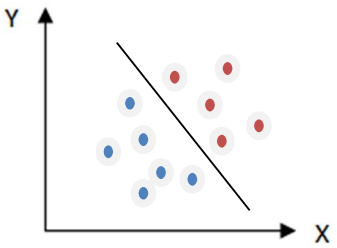
Les SVM reposent sur deux notions clés très anciennes, la marge maximale et la fonction noyau, ainsi que sur le théorème de Vapnik-Chervonenkis[69] (70's). Mais ce n'est qu'à la fin du XX$^{ème}$ siècle que l'assemblage de ces trois idées a permis la rédaction de l'article fondateur.[70]

Il existe, selon la configuration des données, deux cas distincts de SVM : le cas rare où les données peuvent être parfaitement séparées par une fonction linéaire, et celui plus fréquent en pratique, où cela n'est pas complètement possible.

Cas linéairement séparable | Cas non linéairement séparable

Intéressons-nous à la première éventualité, assez théorique mais pédagogique pour bien fixer les idées, dans le cas d'une recherche de classement en deux classes, sur un plan de données.

Considérons un plan (X, Y) et la projection d'observations sur ce plan.




[68] Ouvrage de référence : Abe S., « Support Vector Machines for Pattern Classification », (2010).
[69] https://fr.wikipedia.org/wiki/Th%C3%A9orie_de_Vapnik-Chervonenkis
[70] B.E. Boser, I.M. Guyon, V.N. Vapnik, « A Training Algorithm for Optimal Margin Classifiers », 1992.

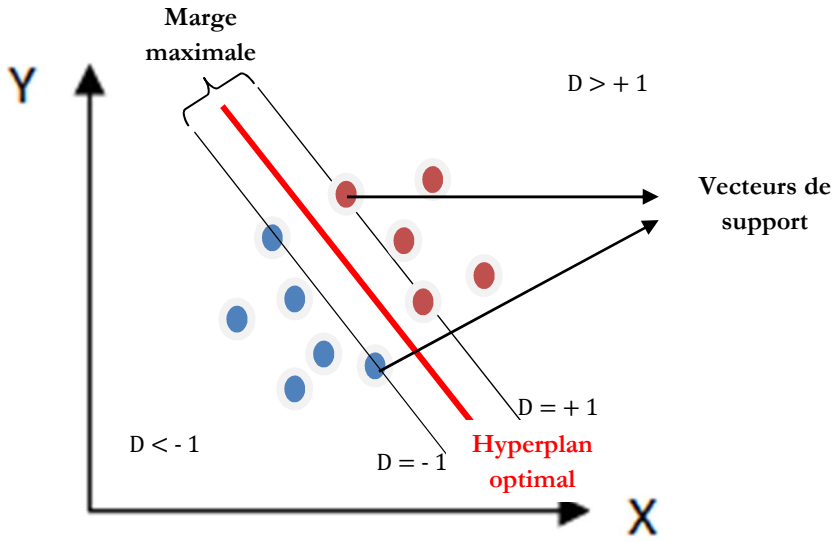
Le but est donc de séparer les points bleus (de la classe 1) des points rouges (de la classe 2), par une droite d'équation $\beta_0 + \beta_1 x + \beta_2 y = 0$, où x et y sont les coordonnées des individus sur les axes et $\beta_0, \beta_1$ et $\beta_2$ sont des paramètres à estimer. Un algorithme de SVM affecte un point à une classe notée +1 ou -1, selon que ce point se trouve d'un côté ou l'autre de la droite. Si $k$ est la classe, on essaye de chercher $\beta_0, \beta_1$ et $\beta_2$ tels que :

$k = +1$ si $\beta_0 + \beta_1 x + \beta_2 y > 0$ et $k = -1$ si $\beta_0 + \beta_1 x + \beta_2 y < 0$.

Ces deux assertions peuvent ainsi se résumer par $k\,(\beta_0 + \beta_1 x + \beta_2 y) > 0$. Il existe bien entendu une **infinité de solutions** ($\beta_0, \beta_1$ et $\beta_2$) possibles car théoriquement un nombre infini de droites est en capacité de séparer les « bleus » des « rouges ». **Les algorithmes SVM résolvent ce problème en calculant la distance minimale existant entre la droite de séparation et les données. Cette distance s'appelle la « marge maximale ».** Le problème revient donc à essayer trouver $\beta_0, \beta_1$ et $\beta_2$ tels que :

$$k\,(\beta_0 + \beta_1 x + \beta_2 y) > M$$

avec M, le plus grand réel positif dépendant des paramètres β.




Avec $D = k\,(\beta_0 + \beta_1 x + \beta_2 y)$

La marge maximale est définie par l'écart entre les droites d'équation $k\,(\beta_0 + \beta_1 x + \beta_2 y) = -1$ et $k\,(\beta_0 + \beta_1 x + \beta_2 y) = +1$. **Les observations qui sont exactement sur les frontières de la marge sont les points supports (« *support vectors* »).** La droite en gras rouge représente l'hyperplan optimal, c'est-à-dire la séparation parfaite, qui ne dépend que des points supports, les points les plus proches ; ce qui est favorable à la robustesse des SVM, notamment par rapport à l'analyse discriminante pour laquelle les points éloignés exercent une influence sur la solution. Vapnik démontre un résultat intéressant : **le pouvoir de généralisation d'un SVM est d'autant plus grand que le nombre de points supports est petit.**

Le modèle générique est très sensible à l'ajout d'un point supplémentaire. Donc pour conserver le maximum de stabilité, on accepte quelques erreurs, c'est-à-dire quelques points mal placés. On introduit alors le terme d'erreur $\varepsilon$. L'inéquation cible devient :

$$k\,(\beta_0 + \beta_1 x + \beta_2 y) > M\,(1 - \varepsilon)$$

Si $\varepsilon > 0$, le point (x, y) se situe du mauvais côté de la marge. Si $\varepsilon > 1$, il se situe du mauvais côté de la droite. $\varepsilon$ doit donc être relativement petit afin que le nombre d'individus mal classés ne soit pas top important.

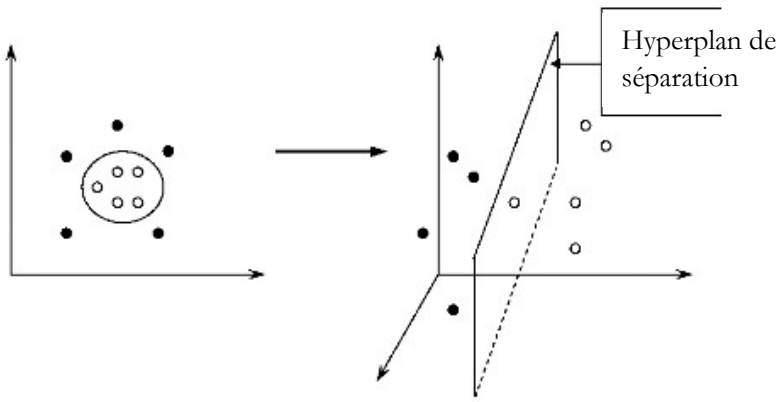
Il peut être aussi souhaitable d'**intégrer des termes polynomiaux**. En effet, il y a des chances que la frontière séparant les données ne soit pas une droite. L'inéquation pourrait alors par exemple prendre la forme suivante :

$$k\,(\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2) > M\,(1 - \varepsilon)$$

**Les méthodes SVM généralisent ce principe en introduisant des fonctions beaucoup plus complexes et flexibles que l'on appelle « noyaux »**[71].

Source : Optimind Winter

Voici par exemple un cas de noyau radial où l'on saisit bien l'idée générale : séparer les classes par un hyperplan en transformant l'espace des données par un changement de dimension des données en entrée. Le lecteur intéressé par ces développements trouvera facilement sur Internet de nombreux exemples plus complexes et plus détaillés.




Les trois avantages principaux des SVM sont leur **faculté à modéliser les phénomènes non linéaires** grâce au choix d'un noyau approprié, la **robustesse de leurs prédicteurs** et leur pouvoir de généralisation. En revanche, ils font face à des inconvénients certains. En premier lieu l'effet **« boîte noire »** : les modèles sont très peu lisibles et par conséquent difficiles à « vendre » aux fonctions métiers. D'autre part et surtout, **les temps de traitement des SVM sont extrêmement longs** et nécessitent des calculateurs très puissants, comme des machines uniquement dédiées aux algorithmes de Data Science ; ce qui génère le plus souvent un découragement pour qui n'y aurait pas accès. Cela explique aussi pourquoi ils sont finalement assez peu utilisés en pratique par les non-experts.

[71] https://fr.wikipedia.org/wiki/Noyau_(statistiques)

### 4.3.7. Random Forest[72]

Les *Random Forest* - ou forêts aléatoires - sont un cas particulier du ***bagging***. Le terme « *bagging* » vient de « ***b****ootstrap »* et « ***agg****regat****ing*** *»*. Il signifie en effet **une combinaison d'algorithme**, i.e. un méta-algorithme, issue de ces deux méthodes.

On appelle « échantillons bootstrap », ces échantillons obtenus par tirage aléatoire avec remise de $n$ individus parmi $n$. Dans un échantillon *bootstrap*, un individu peut être tiré plusieurs fois ou ne pas être tiré. La probabilité qu'un individu donné soit tiré est égale à $1 - (1 - \frac{1}{n})^n$, et elle tend vers $1 - \frac{1}{e} \approx 0{,}632$ quand $n$ tend vers $+\infty$.[73]

L'agrégation est une méthode qui consiste à **produire plusieurs bootstraps et à agréger les résultats** obtenus. L'union fait la force ! Dans le cas d'une variable à expliquer **Y nominale**, on procède à **un vote**, chaque résultat de bootstrap représentant une voix. Si **Y est quantitative**, chacun des bootstraps $T_1 \ldots T_M$, prédit une valeur $y_1 \ldots y_M$. La conclusion de l'agrégation est alors **la moyenne des** $y_i$, $\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$

**Le bagging corrige plusieurs défauts** des estimateurs ordinaires, notamment leur instabilité (de petites modifications dans l'échantillon d'apprentissage peuvent entraîner des arbres très différents) et leur tendance au sur-apprentissage. La contrepartie à payer est une **perte de lisibilité** : les prédictions d'une forêt d'arbres « baggés » ne sont plus le fruit d'une seule association de règles, mais sont issues d'un consensus de raisonnements potentiellement divergents. L'interprétation est rendue beaucoup plus difficile. L'analyse théorique du bagging demeure incomplète à ce jour : plusieurs arguments aident à comprendre son impact et suggèrent des conditions dans lesquelles il améliore les prédictions, mais l'élaboration d'un modèle dans lequel cet impact est compris et mesurable reste un sujet de recherche.

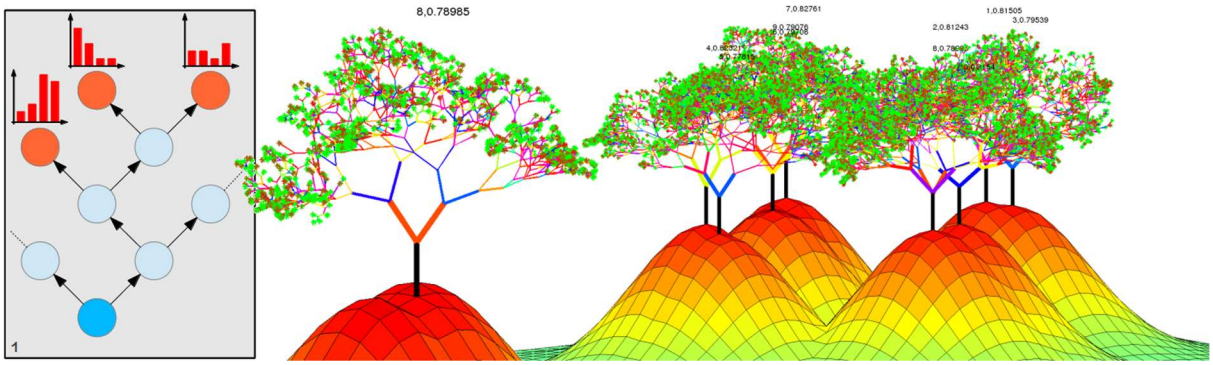
**La méthode des forêts aléatoires est un bagging d'arbres de décision qui intègre une partie aléatoire supplémentaire dans la modélisation** : on tire au sort un sous ensemble des variables explicatives disponibles à chaque nœud de l'arbre de décision. Elle a été développée par Leo Breiman et Adèle Cutler en 2001.

L'algorithme est un **processus itératif dans lequel des phénomènes aléatoires interviennent** : A chaque itération, un échantillon bootstrap, de taille identique à celle de l'échantillon initial, est tiré dans l'échantillon d'apprentissage. Avec cet échantillon bootstrap, on construit un arbre. Une deuxième partie aléatoire intervient à chaque nœud de l'arbre. Les K variables candidates à la segmentation de l'espace sont sélectionnées parmi toutes les variables de l'échantillon d'apprentissage avec une probabilité uniforme. La meilleure coupure, sélectionnée parmi toutes celles admissibles, est recherchée parmi ces K variables et non plus parmi toutes les variables disponibles. L'introduction de cet aléa supplémentaire permet de faire entrer dans le modèle des

[72] Ouvrage de référence : « Bagging predictors », Breiman, Leo, (1996).
[73] Démonstration par l'application d'un Développement Limité

variables qui ne pourraient pas l'être si toutes les variables étaient candidates. Ainsi, des variables moins discriminantes mais apportant quand même de l'information contribuent également à la construction du modèle.

Dans un contexte de classement, **l'étape finale d'agrégation consiste à faire voter les arbres à la majorité**. On retrouve le compromis biais - variance. Le compromis à réaliser dépend du nombre d'échantillons bootstrap tirés et du nombre de variables candidates à chaque nœud.




Il existe différents modèles de forêts aléatoires qui se différencient soit au niveau de la phase de sélection des données, soit au niveau de la construction de l'arbre. D'une manière générale, ils sont **faciles à mettre en place**, sont **plus efficaces** que les arbres de décision « ordinaires », et leurs algorithmes associés **convergent plus rapidement**, et d'autant plus dans des situations hautement multidimensionnelles. Ils sont en outre très robustes face aux données aberrantes et données extrêmes. En revanche, leur inconvénient majeur est qu'ils nécessitent des **capacités de calcul et de stockage relativement importants**. C'est pourquoi ils fonctionnent plutôt bien avec des infrastructures de type Hadoop[74].

[74] https://fr.wikipedia.org/wiki/Hadoop

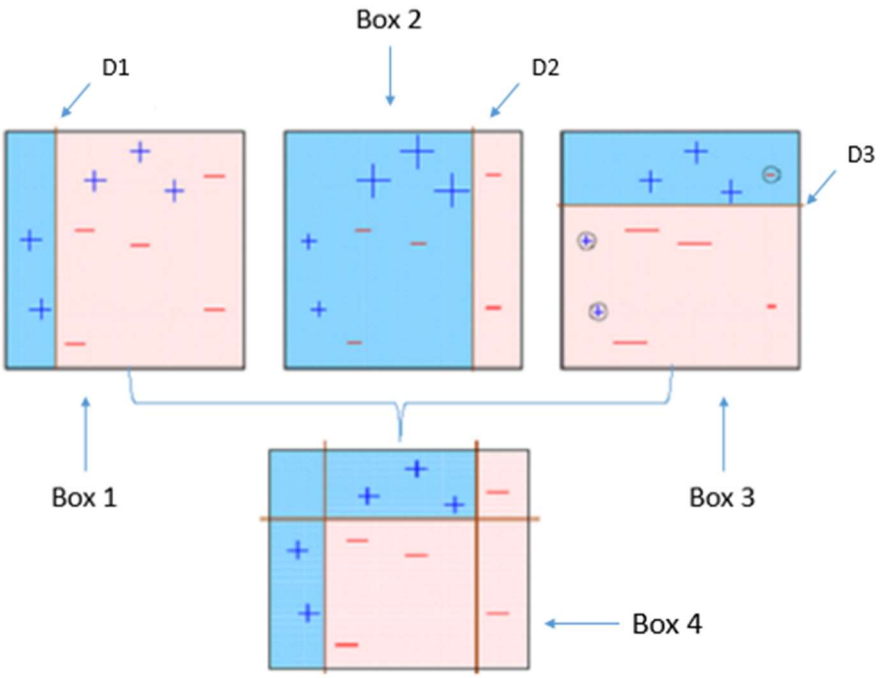
### 4.3.8. Gradient Boosting Machine[75]

Le ***boosting*** est une autre famille d'agrégation de modèles. L'idée principale est d'**appliquer de nombreuses fois le même algorithme** de classement ou de prédiction **sur différentes versions de l'échantillon initial** d'apprentissage. Ces variantes sont modifiées à chaque étape pour tenir compte des erreurs de l'étape précédente. Le modèle est ainsi amplifié ou « boosté ». In fine, on essaye de **combiner ces modèles pour obtenir un modèle final très performant.**

Le plus souvent, les modifications apportées à chaque étape consistent à **surpondérer les observations mal classées ou mal ajustées à l'étape précédente**. On force ainsi l'apprentissage à se focaliser sur les observations les plus difficiles à prédire. Parallèlement, les différents classifieurs sont aussi pondérés de manière à ce qu'à chaque prédiction, les classifieurs ayant prédit correctement auront un poids plus fort que ceux dont la prédiction est incorrecte. Les mêmes mécanismes de lutte contre le sur-apprentissage, vus auparavant, sont actionnés.

Une des méthodes parmi les plus répandues est ***AdaBoost***.[76] Cette méthode est l'une des premières versions du *boosting* ; et une des plus simples à comprendre. Le principe est de donner plus d'importance aux valeurs difficiles à prédire, en boostant les classifieurs qui réussissent quand d'autres ont échoué, à l'aide d'un paramètre de mise à jour adaptatif. L'algorithme s'appuie ainsi sur les classifieurs calculés et cherche à leur affecter les bons poids vis à vis de leurs performances réciproques.

**Mécanisme adaptatif du boosting :**




[75] Ouvrage de référence : Freund, Y. and Schapire, R. E : A decision-theoretic generalization of on-line learning and an application to boosting, (1997).
[76] https://fr.wikipedia.org/wiki/AdaBoost

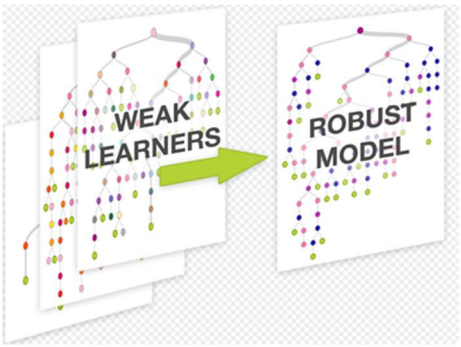
Nous décidons toutefois de *challenger* les méthodes traditionnelles par le **GBM** ou *Gradient Boosting Machine*. Il est assez difficile de vulgariser mathématiquement ces algorithmes. Très rapidement, on introduit des concepts complexes qui nécessitent des développements relativement consistants. Nous nous contentons donc d'exposer brièvement les idées générales.

Le plan est encore une fois d'**agréger plusieurs classifieurs créés de façon itérative**. Ces « mini-classifieurs » ou « classifieurs faibles » sont généralement des fonctions simples dont chaque paramètre est le critère de séparation des branches d'un arbre. **Le super-classifieur final est une pondération de ces mini-classifieurs.** Une approche pour construire ce super-classifieur est de :

1. Prendre une pondération quelconque de mini-classifieurs et former son super-classifieur.
2. Calculer l'erreur induite par ce super-classifieur, et chercher le mini-classifieur qui s'approche le plus de cette erreur, ce qui revient à le chercher dans l'espace des paramètres.
3. Retrancher le mini-classifieur au super-classifieur tout en optimisant son poids par rapport à une fonction de perte.
4. Répéter le processus itérativement un grand nombre de fois jusqu'à convergence.

*Source : Octo.com*

Le classifieur final du gradient boosting est donc obtenu par les poids de pondération des différents mini-classifieurs, ainsi que par les paramètres des fonctions utilisées.




L'opérateur **gradient** est un outil qui transforme un champ scalaire, défini par une valeur précise, en un champ vectoriel, défini par 3 éléments, un sens, une direction et une norme. Il s'oppose à l'opérateur **divergence** qui est l'outil qui transforme un champ vectoriel en champ scalaire.

La résolution du problème consiste à explorer un espace de fonctions simples par une descente de gradient sur l'erreur. En d'autres termes, **le GBM construit une séquence de modèles de sorte que chaque étape,** chaque modèle ajouté à la combinaison, **apparaisse comme un pas vers une meilleure solution.** La principale innovation du GBM par rapports aux algorithmes de boosting primitifs est que ce pas est franchi dans la direction du gradient de la fonction perte, lui-même approché par un arbre de régression.

Comme pour les forêts aléatoires ou toute autre méthode d'agrégation de modèles, la propriété d'interprétabilité des arbres est perdue avec le GBM. Néanmoins, il est possible de **calculer des**

**critères d'importance relative des variables** ayant servi à construire le modèle final. Dans la majorité des cas, **les algorithmes de boosting arrivent à diminuer à la fois la variance et le biais.** En conséquence, leurs capacités prédictives et leurs performances globales sont d'une manière générale très bonnes. **C'est ce qui rend les GBM si populaires.**

D'autre part, lorsque le nombre de variables en entrée est très important et que seules quelques-unes d'entre elles s'avèrent pertinentes, le boosting permet, davantage que les méthodes de bagging, de détecter quelles sont-elles. En revanche ces méthodes nécessitent le réglage et donc l'estimation d'un plus grand nombre de paramètres comme le choix de la fonction de perte, de la profondeur des arbres ou des coefficients de régularisation.

## 4.4. Conclusion partie IV

- **Les méthodes traditionnelles – régression logistique, arbre de décision, analyse discriminante et classifieur bayésien naïf -, conservent de solides arguments et de bonnes capacités de prédiction. Elles doivent continuer à être testées.**

- **Les nouveaux algorithmes d'agrégation de modèles offerts par la Data Science apportent une réelle plus-value.**

- **Ils sont basés sur des stratégies aléatoires (bagging) ou adaptatives (boosting) qui permettent d'améliorer l'ajustement par une combinaison ou une agrégation d'un grand nombre de modèles tout en évitant le sur-ajustement.**

- **Principaux avantages : ils sont plus faciles d'utilisation pour intégrer des nouvelles variables et pour modéliser des nouveaux supports (textes, images, vidéos, données Web).**

- **Le *stacking* est un procédé qui consiste à appliquer un algorithme de machine learning à des classifieurs générés par d'autres algorithmes de machine learning. Il s'agit, en quelque sorte de prédire quels sont les meilleurs classifieurs et de les pondérer. Pour certains data scientists, on est ici proche du domaine de l'art !**

- **Je conserve un petit faible pour la « vieille » régression logistique, rarement battue tout compte fait. Parmi les nouveaux algorithmes, j'avoue avoir une affection pour les forêts aléatoires et leurs dimensions démocratiques.**

- **Quoi qu'il en soit, un actuaire-data scientist ne peut raisonner de la sorte. Pour chaque problème, il est absolument nécessaire de tester a minima 5 ou 6 méthodes, et toutes choses égales par ailleurs, de retenir la plus performante, sans émotion.**

- ***Data scientist = « Statistician + Programmer + Coach + Storyteller + Artist »***[77] **J'adhère totalement.**

[77] Shlomo Aragmon.

# Partie 5 : Calcul de Valeurs Clients

*« Un modèle est une représentation*
*d'une réalité matérielle ou immatérielle,*
*un modèle n'est pas cette réalité.*
*Modéliser est donc l'acte de représenter*
*nos concepts et les objets de notre réalité. ».*

Jérémie Grodziski

## 5.1. Données nécessaires et Périmètre

### 5.1.1. Les sources de données

Plusieurs sources de données sont nécessaires pour construire un modèle puissant de valeur client :

- Clients/Prospects en provenance du CRM (SalesForce)
- Socio-démographiques et administratives : en provenance de l'outil de gestion des contrats (Activ'Infinite)
- Actuarielles : en provenance du système de tarification et de la base technique (DT)
- Comptables et Financières : en provenance du contrôle de gestion (DFC)
- Digitales : en provenance des sites web et réseaux sociaux (DMP)
- Third Party (éventuellement) : achat de données de ciblage publicitaire ou marketing Internet qui sont fournies à l'annonceur par une société tierce autre que l'éditeur
- + Open Data, si et seulement si on estime que des enrichissements de base apportent de la réelle valeur ajoutée

### 5.1.2. Le périmètre

Dans un premier temps, nous choisissons de nous concentrer sur les deux principaux portefeuilles interprofessionnels assurés et gérés par LMG. Les extensions possibles par la suite sont :

- les retraités affinitaires (portefeuille historique)
- les autres portefeuilles interpro
- les TNS
- le petit pro
- les grands comptes
- les PME

L'étude des flux financiers démarre en **2007**, première année complète dans le SI actuel. « Activ'Infinite » ou « AI », l'outil de gestion (back office) actuellement en vigueur à LMG est opérationnel depuis le 1er juillet 2006, date à laquelle il a remplacé le précédent système « Basic » en vue du basculement au 1er janvier 2007, des salariés de droits privé La Poste, du portefeuille historique, le Statutaire, vers le Contrat Collectif La Poste (CCLP), devenu de facto, le plus gros contrat groupe de France.

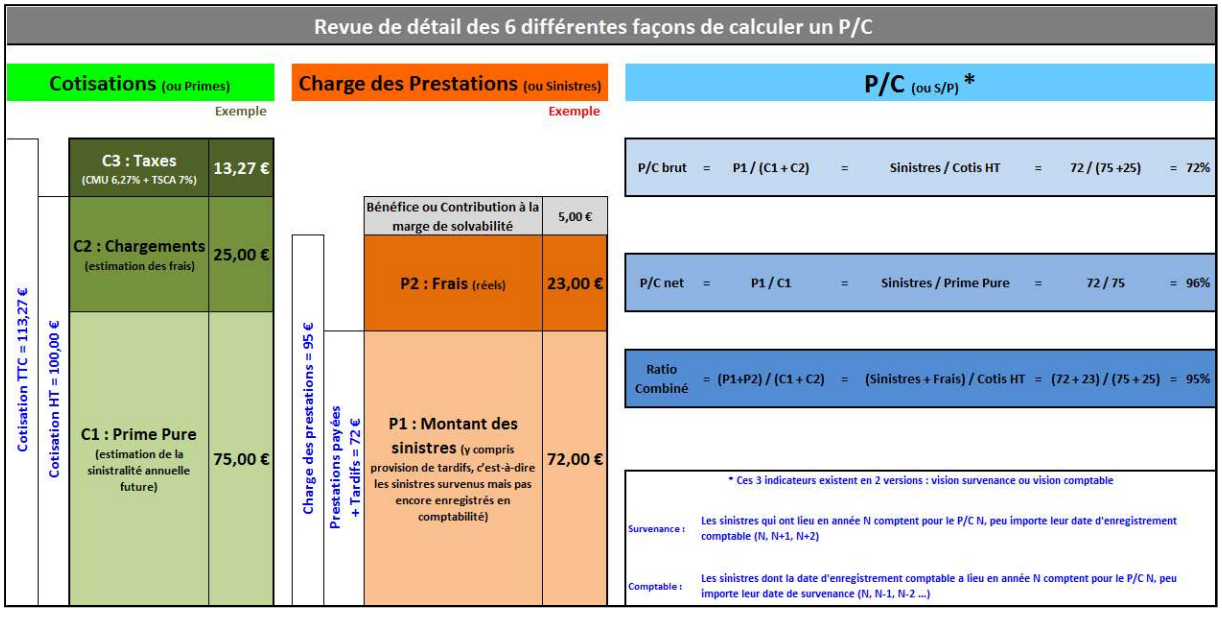
## 5.2. Calculs de P/C

Le P/C (ou S/P, les deux appellations ont cours, le S/P étant plus utilisé dans les sociétés d'assurance et le P/C en protection sociale) est le ratio clé en assurance, traditionnellement le principal indicateur pour le pilotage du risque.

Il représente le rapport entre Prestations et Cotisations (alt. : Sinistres et Primes). Calculé par l'assureur, il permet de vérifier la rentabilité d'une police d'assurance. En effet, l'activité assurantielle étant par son essence même ce que l'on appelle un cycle de production inversée (c'est-à-dire une activité dont on ne connait le prix de revient qu'après l'avoir vendue, au contraire d'une activité classique où on connait le prix de revient avant la vente), le P/C est l'indicateur qui permet de savoir si le volume des cotisations émises et acquises, encaissées ou non, a permis de couvrir (P/C net < 100%) le volume des prestations, payées ou restant à payer, ou si l'activité est déficitaire (P/C net > 100%).

Les prestations s'entendent le plus souvent prestations payées plus provisions, et dans la plupart des cas, le calcul se fait avec les cotisations brutes (HT ! bien entendu) c'est à dire en incluant les chargements, i.e. la part des cotisations supposée financer les frais de l'organisme d'assurance. Les éventuels produits financiers ne sont pas pris en compte dans le calcul. Cela explique aussi pourquoi un assureur avec un ratio combiné de 100% ou très légèrement supérieur, parvient tout de même à s'autofinancer[78] : les produits financiers peuvent compenser le besoin de fonds propres appelé marge de solvabilité.

On distingue plusieurs P/C :




Dans la suite, sauf mention, le P/C s'entend comme P/C brut en survenance.

Du fait de la mutualisation des risques, un P/C n'a communément de sens qu'au niveau d'un groupe d'individus, par exemple un groupe défini par son niveau de garantie et sa tranche d'âge. Dans

[78] De moins en moins vrai malheureusement en raison du contexte de taux bas.

notre modèle, nous allons toutefois prendre le parti de calculer des P/C dits « individuels ». Ils s'interprètent de la même façon : le rapport sinistre à prime sur une année (ou une fraction d'année) mais pour une seule tête assurée. Il faut faire attention à ne pas sur-interpréter ce ratio.

## 5.3. Valeurs Clients

### 5.3.1. Valeur Client Passée

Constitution d'une table qui contient l'ensemble des caractéristiques "Population" de chaque assuré ainsi que ses prestations et cotisations 2008 à 20XX + le "P/C individuel" + le P/C de son groupe + le facteur de crédibilité calculé par Bühlmann-Straub.

**VCP = VC 2008-2016 =**
**∑ (cotisations santé/prévoyance, nettes de FG) - ∑ (prestations santé/prévoyance)**

Symétriquement à l'actualisation, se pose la question de la capitalisation des marges. Nous prenons le parti de ne pas le faire.

### 5.3.2. Valeur Client Actuelle

La valeur client actuelle correspond à la valeur client de l'année en cours. Elle peut se décomposer en VCAP (Valeur Client Actuelle Passée) qui correspond à la marge unitaire acquise du 1er janvier au dernier jour du mois précédent, connue, et VCAF (Valeur Client Actuelle Future) qui correspond à la marge unitaire probable du 1er jour du mois en cours au 31 décembre, inconnue donc à estimer.

Quelques remarques :

- P/C cible = 75% pour les nouveaux arrivants
- La contribution à la marge de chaque produit est pondérée par le temps de présence sur l'année de chaque produit
- (12 - dernier_mois_num) / 12 : calcule la part de l'année entre dernier jour du dernier mois chargé en base et le 31 décembre
- Contribution à la marge Santé = Primes nettes de Frais de Gestion - P/C * Prime HT
  = Prime HT * (1 - Tx_FG) - P/C * Prime HT
  = Prime HT * (1 - Tx_FG - P/C)

### 5.3.3. Valeur Client Future

Projection dans le temps à Horizon 20 ans (parti pris) des assurés principaux et de leurs ayants-droit. Nous n'incorporons pas dans ce modèle les affaires nouvelles futures, le calcul de valeur client n'intégrant pas le New Business Margin (NBM)[79].

Nous devons nécessairement projeter les ayants-droit pour tenir compte de l'évolution de la composition familiale dans le temps et notamment de la sortie à 100% des enfants de 25 ans (à 20 ans sur le Statutaire). Les valeurs clients calculées par personne protégée seront ensuite additionnées pour obtenir la valeur client par assuré principal donc par client.

Quelques remarques :

- Si date fin de contrat > 31/12/2017, eff17 = 1
  sinon eff17 = (date de fin de contrat – 01/01/17) / 365
- Le modèle est joué pour tout i de XX à XX+19, et j = i +1.
- Hypothèse (forte) : taux de churn en prévoyance = taux de churn en santé !
- **VCF =** (Somme des marges annuelles santé nettes de FG + somme des marges annuelles prévoyance hors PJ nettes de FG * scores d'appétence cumulés + 17.5% cotisations PJ * score d'appétence PJ cumulé) **actualisées** * probabilités annuelles de survie
- Contrairement aux cotisations santé, les cotisations prévoyance ne sont pas "annualisées" pour l'année en cours. En effet, elles embarquent déjà le caractère "durée de couverture dans l'année". C'est pourquoi le facteur de pondération vaut "1" et pas "PREV_p&an".

### 5.3.4. Valeur Client Totale

**VCT = VCP + VCA + VCF - coût d'acquisition**
Limpide, ne nécessite pas de commentaire supplémentaire.

[79] http://institutionnel.generali.fr/taux-de-marge-sur-affaires-nouvelles

## 5.4. Conclusion partie V

- **Les sources de données d'un modèle de valeur client sont nombreuses et en perpétuel renouvellement.**
- **La partie 'Data Management' du modèle est très chronophage mais essentielle comme dans tout projet 'Data'.**
- **Lorsque tout est bien en place et toutes les hypothèses sont posées, la partie 'Projection', le « cœur du modèle », ne représente pas la principale difficulté.**
- **La Valeur Client Totale est une somme de valeur client connue, sans aléa, dont les éléments sont enregistrés dans le système d'information, et d'une valeur client estimée par le modèle.**

# Partie 6 : Résultats et extensions

*« Il n'est désir plus naturel que le désir de connaissance.*
*Nous essayons tous les moyens qui nous y peuvent mener.*
*Quand la raison nous manque, nous y employons l'expérience*
*qui est un moyen plus faible et moins digne ;*
*mais la vérité est chose si grande, que nous ne devons*
*dédaigner aucune entremise qui nous y conduise ».*

Montaigne

## 6.1. Résultats du scoring

### 6.1.1. Jeu de données utilisée

Plus de 150 variables constituent le jeu de données global du modèle de valeur client. Mais seule une partie d'entre elles est éligible pour le scoring.

Famille et nombre de variables retenues :

| | |
|---|---|
| Socio-démographiques (âge, sexe, situation familiale, code postal, …) | 15 |
| Produits et Garanties techniques (niveau de garantie, multi-équipement, …) | 8 |
| Vie du contrat (ancienneté, pré-contentieux, …) | 7 |
| Souscription (Canal d'entrée, jour d'ouverture des droits, type de vente, …) | 6 |
| Open Data Insee (caractérisation des IRIS[80]) | 6 |
| Interaction client (Nb demandes, réclamations, envoi courrier, …) | 5 |
| Données de contacts (téléphone, email, espace adhérent, newsletter) | 4 |

Soit **51 facteurs explicatifs potentiels** auxquels on ajoute 1 identifiant assuré, 1 variable pour l'échantillonnage et la variable à expliquer ou endogène ; c'est-à-dire un total pour le dataset 'Scoring' de **54 variables**. Parmi elles, 8 sont continues et 45 sont discrètes dont 7 qualitatives ordinales.

Les variables éliminées du dataset 'Scoring' sont :
- les autres identifiants,
- toutes les dates dont les dates de début et fin de couverture de chaque produit,
- les variables actuarielles comme les prestations par poste de soins et les cotisations,
- les détails techniques des contrats,
- les interactions détaillées,
- les durées de couverture annuelles,
- et enfin, l'open data très détaillé ou dont la complétude fait défaut.

Nombre d'observations retenues : **113 877 assurés principaux** interprofessionnels couverts au 31 janvier 2018 sur les deux principaux contrats.

Lorsque la méthode testée requiert un échantillon de validation ou d'élagage (Réseau de neurones ou Arbre de décision) la population est distribuée ainsi : Apprentissage 60% - Test 20% - Validation 20%. Lorsque la méthode ne nécessite pas processus de validation, l'échantillonnage alloue 70% des individus à l'apprentissage et 30% au test.

[80] https://fr.wikipedia.org/wiki/%C3%8Elots_regroup%C3%A9s_pour_l%27information_statistique

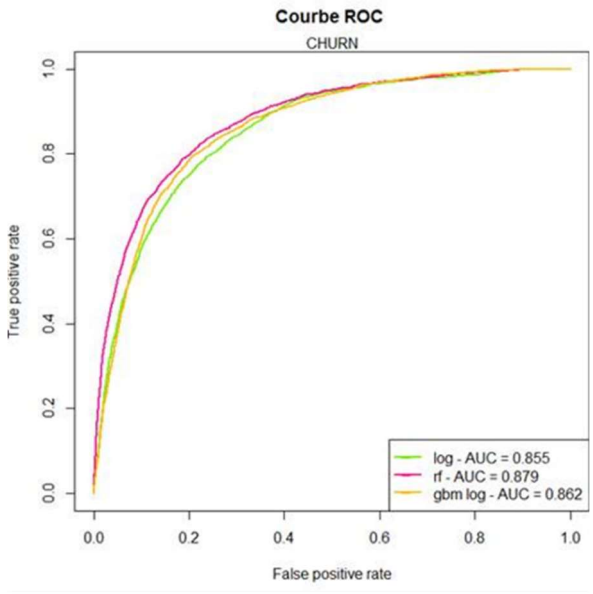
### 6.1.2. Progiciels utilisés

Les 7 premières méthodes ont été en premier lieu testées sur **SPAD**. La régression logistique fut retenue pour le modèle prototype sur le logiciel **SAS** car elle est la plus facile à implémenter sur notre version sans module complémentaire et en moyenne la plus performante. Puis nous avons challengé cette dernière avec les forêts aléatoires et le gradient boosting sur **R** et **Python**.

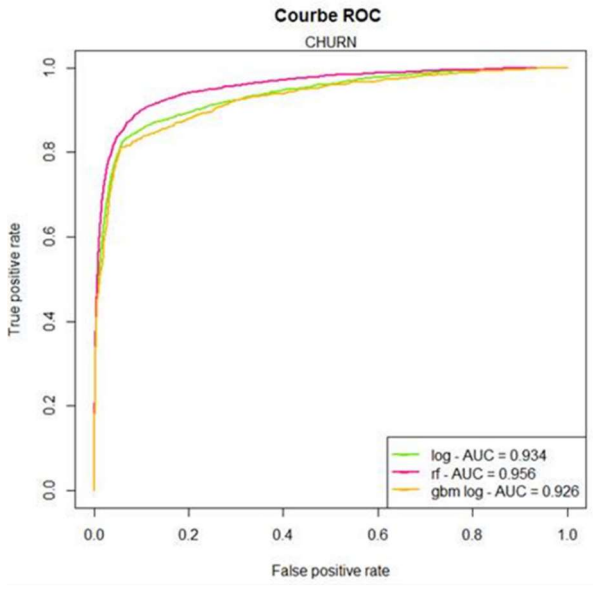
### 6.1.3. Discrétisation des variables explicatives continues

La majorité des méthodes d'analyses prédictives admet aussi bien les variables continues que les variables discrètes comme facteurs explicatifs potentiels. Toutefois le pouvoir discriminant des premières est souvent moindre ou a minima pas totalement exploité.

L'exemple suivant sur la prédiction de la démission volontaire est particulièrement frappant :




**Sans** discrétisation des variables continues




**Avec** discrétisation des variables continues

Initialement, j'ai intégré toutes les dimensions à leur état naturel. L'aire sous la courbe des trois meilleures méthodes se situe entre 0.85 et 0.88, ce qui représente déjà un résultat satisfaisant. Mais quel que soit le score ou la méthode étudiés, **aucune variable continue ne ressort en tant que facteur discriminant**.

Se pose alors la question de la discrétisation des variables continues comme l'âge, la durée de vie passée, le nombre d'enfants ou le nombre de contrats. On peut essayer de construire des nouvelles dimensions en créant des modalités « intelligentes », i.e. pertinentes d'un point de vue métier. Par exemple, pour l'âge, on pourrait choisir des classes traditionnelles comme celles souvent utilisées par l'INSEE : '1. < 25 ans', '2. 25 - 34 ans', '3. 35 – 44 ans', etc. On peut à l'inverse laisser un algorithme préconiser quels seraient les meilleurs découpages.

Il existe sur SPAD, une procédure particulièrement bien adaptée. Il s'agit de l'instruction « Mise en classes et regroupement de modalités, en mode supervisé ». Elle permet d'identifier et d'appliquer automatiquement un schéma efficace de création et/ou de regroupement de modalités afin de prévoir une variable catégorielle à partir d'une variable cible, qui est le plus souvent la variable à expliquer d'un modèle d'analyse prédictive.

Pour l'âge, la procédure donne comme meilleur regroupement les modalités suivantes : <23 ans ; [23-24 ans] ; [25-30 ans] ; [31-40 ans] ; [41-48 ans] ; [49-59 ans] ; [60-70 ans] ; >=71 ans

Autre exemple peu intuitif, le regroupement pour l'ancienneté :
< 2 ans ; [2-4 ans] ; [5-7 ans] ; >=8 ans

Finalement, en remplaçant les 8 variables continues par les 8 nouvelles dimensions obtenues par la procédure de recodage supervisé, la capacité prédictive de tous les modèles s'améliorent nettement : les AUC se situent désormais entre 0.92 et 0.96, ce qui témoigne d'une excellente modélisation de la variable à prédire.

### 6.1.4. Scoring de la démission volontaire

Dans cette partie, nous analysons en détail le scoring de la démission volontaire (ou équivalent, la probabilité de churn). Il représente en effet le score le plus important du modèle de valeur client, celui qui a le plus d'incidence. Les autres scores fonctionnent exactement de la même manière.

- **Distribution de la variable à expliquer par échantillons :**

| | Ech. Apprentissage | Ech. Test | Total |
|---|---|---|---|
| **Non sortants 2017** | 73 631 | 31 664 | 105 295 |
| **Sortants 2017** | 6 010 | 2 572 | 8 582 |
| **Total** | 79 641 | 34 236 | 113 877 |
| **Répartition** | 69,9% | 30,1% | 100,0% |
| **Taux de démission volontaire 2017** | 7,55% | 7,51% | 7,54% |

La variable à expliquer prend deux valeurs : 0 = non sortant et 1 = sortant (la modalité cible à modéliser). Les sortants 2017 sont bien répartis dans les deux échantillons : le taux de démission volontaire 2017 avoisine dans les deux cas 7,5%.

- **Détails de la régression logistique :**

*a. Ajustement du modèle et Test du rapport de vraisemblance*

| Indicateurs | Constante (intercept) | Modèle |
|---|---|---|
| Critère d'Akaike | 42 618 | 19 659 |
| Critère BIC | 42 627 | 20 412 |
| Déviance | 42 616 | 19 497 |
| R2 de Cox et Snell | 0,252 | |
| Coefficient de Nagelkerke | 0,608 | |
| R² ajusté | 0,705 | |

| | |
|---|---|
| Rapport de vraisemblance | 23 118 |
| Ddl | 80 |
| p-valeur | 0,000 |

L'ajustement du modèle est obtenu après 4 itérations. L'ensemble des indicateurs démontre la validité du modèle. Le $R^2$ est bon : 70.5% de la variance du nuage de points des données brutes est conservé par la modèle, ce qui est un très bon résultat compte tenu du grand nombre de facteurs explicatifs potentiels. Enfin le test du maximum de vraisemblance renvoie une p-valeur inférieure à 1‰. L'hypothèse H0 est rejetée : la variable endogène ne peut pas être expliquer uniquement par la constante. Il y a bien des dimensions discriminantes qui apportent de l'information.

*b. Sélection des facteurs discriminants*

La sélection des effets est effectuée pas à pas (procédure *stepwise*). Le modèle est stable après 23 étapes et retient 19 variables explicatives. Deux variables en effet sont entrées puis ressorties : la Direction des Ventes et le regroupement de départements selon les données Insee.

Les 19 dimensions conservées :

1. **Le nombre de bénéficiaires** mis en classe (s'il est supérieur à 1, la probabilité de démission est plus forte).
2. **La répartition par âge de l'Iris** du client
3. , 12, et 15 **La durée de vie passée** mise en classe (plus elle est importante, plus forte est la probabilité de rester), **l'année d'arrivée dans le portefeuille (= cohorte)**, ainsi que **l'ancienneté** mise en classe.
4. **Le Régime SS** (les adhérents dont le régime obligatoire est géré par LMG ont plus de chance de rester)
5. **La composition familiale** (les clients couverts avec leur conjoint partent moins souvent à l'inverse des familles avec 3 enfants ou plus).
6. , 13. et 16. **Les interactions avec la Relation Client** (Envoi de courriers, demandes ou réclamations) influent positivement sur le risque de sortie du portefeuille.
7. **L'âge du client** mis en classe (les 25-40 ans ont une plus grande probabilité de partir ; à l'inverse après 60 ans, la proba diminue fortement).
8. , 17. **La région et le département d'habitation** : (d'une manière générale, les assurés de l'Ouest de la France sont plus volatiles).

9. Le fait d'être un client **ancien enfant d'adhérents affinitaires** protège contre le risque de churn
10. , 14 et 19. Les clients multi-équipés en prévoyance ont tendance à moins partir par démission volontaire.
11. Viennent enfin : le fait de posséder ou non un **espace adhérent**, et le **jour de début d'ouverture des droits** (1er janvier *vs.* tous les autres),

Pour chacune de ses dimensions, on fait calculer par la machine, l'intervalle de confiance de l'odd ratio de chaque modalité. Ne sont conservées in fine par le modèle, parmi les dimensions discriminantes, uniquement les modalités dont l'intervalle de confiance de l'odd ratio ne contient pas 1, c'est-à-dire les cas où l'estimateur est significativement différent de ln (1), i.e. différent de 0.

*c.* *Test de Hosmer et Lemeshow*

**Partition pour les tests de Hosmer et de Lemeshow**

| | | dem_add = 0 | | dem_add = 1 | |
|---|---|---|---|---|---|
| Groupe | Total | Observé | Attendu | Observé | Attendu |
| 1 | 1973 | 965 | 973.93 | 1008 | 999.07 |
| 2 | 1976 | 1875 | 1864.64 | 101 | 111.36 |
| 3 | 1973 | 1950 | 1927.65 | 23 | 45.35 |
| 4 | 1973 | 1951 | 1944.56 | 22 | 28.44 |
| 5 | 1973 | 1953 | 1956.32 | 20 | 16.68 |
| 6 | 1968 | 1951 | 1958.35 | 17 | 9.65 |
| 7 | 1974 | 1958 | 1968.85 | 16 | 5.15 |
| 8 | 1976 | 1968 | 1974.08 | 8 | 1.92 |
| 9 | 1979 | 1976 | 1978.64 | 3 | 0.36 |
| 10 | 1967 | 1967 | 1966.97 | 0 | 0.03 |

**Test d'adéquation de Hosmer et de Lemeshow**

| Khi-2 | DDL | Pr > Khi-2 |
|---|---|---|
| 81.8651 | 8 | <.0001 |

Curieusement, malgré la bonne qualité apparente de l'ajustement, le test de Hosmer et Lemeshow rejette l'hypothèse nulle de distance faible entre les valeurs prédites et les valeurs observées. Toutefois, en observant le tableau des données qui sert à construire la statistique de test, on remarque que l'écart relatif entre observé et attendu lorsque la modalité de l'endogène est 'non sortant' (= dem_add = 0), est max pour le 3ème décile (1.2%), et parallèlement, l'écart absolu max est 22 individus, ce qui reste relativement acceptable. Le rejet de H0 tient davantage aux limites du test[81] et au fait que dès qu'on traite d'un nombre important d'individus, l'utilisation de la statistique du $\chi^2$ est suspecte.

[81] http://stats.stackexchange.com/questions/169438/evaluating-logistic-regression-and-interpretation-of-hosmer-lemeshow-goodness-of

*d. Matrices de confusions*

Les matrices de confusion sont construites au seuil de probabilité de 50% par défaut. C'est-à-dire que le modèle va considérer sortant un individu dont la probabilité de démission volontaire est supérieure ou égale à 50%.

**Effectifs (éch. Test)**

| Démission volontaire en 2017 | Classé 0 - non | Classé 1 - oui | Total |
|---|---|---|---|
| 0 - non | 31 197 | 467 | 31 664 |
| 1 - oui | 1 120 | 1 452 | 2 572 |
| Total | 32 317 | 1 919 | 34 236 |

**Répartition (éch. Test)**

| Démission volontaire en 2017 | Classé 0 - non | Classé 1 - oui | Total |
|---|---|---|---|
| 0 - non | 91% | 1,4% | 92% |
| 1 - oui | 3,3% | 4,2% | 7,5% |
| Total | 94% | 5,6% | 100% |

Le taux d'individus mal classés par le modèle atteint alors 4.64% (1.36% + 3.27%), ce qui peut sembler faible donc acceptable. Toutefois en analysant plus finement (tableau suivant), on se rend compte que le modèle classe bien surtout les non sortants (1.5% de mal classés). Par contre il classe très mal les sortants (seulement 56.5% de bien classés).

**Matrice de classement (éch. Test)**

| Démission volontaire en 2017 | Bien classé | Mal classé |
|---|---|---|
| 0 - non | 98,5% | 1,5% |
| 1 - oui | 56,5% | 43,5% |
| Total | 95,4% | 4,6% |

Le choix d'un seuil de probabilité à 50% est plutôt intuitif. Si j'essaye de prédire la valeur future d'une variable à deux modalités 'Noir' et 'Blanc', et que le modèle me revoie une probabilité d'obtenir 'Noir' supérieure à 50% alors j'ai tendance à conclure 'Noir'. Pour autant, ce principe n'est pas nécessairement pertinent pour un modèle dont l'incidence de la modalité cible n'est pas très élevé, comme dans notre cas (Rappel : taux de démission volontaire 2017 ≈ 7.5%).

C'est ici qu'il est intéressant d'analyser les notions de sensibilité et spécificité[82]. Notamment dans notre cas, nous souhaitons un meilleur équilibre entre ces deux indicateurs. Bien identifier les sortants nous importe au moins autant que réduire le taux d'erreur dans l'absolu.

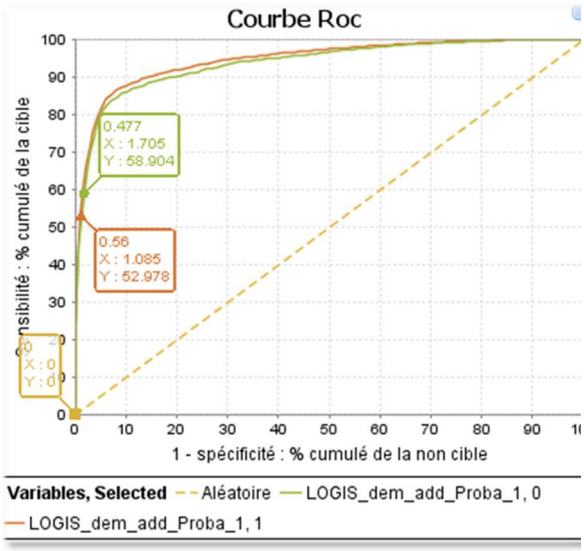
[82] Voir paragraphe *4.2.2.2 La courbe ROC et le critère AUC – exemple du dépistage et confirmation du test*

Au seuil de probabilité 50%, on obtient les indicateurs suivants :

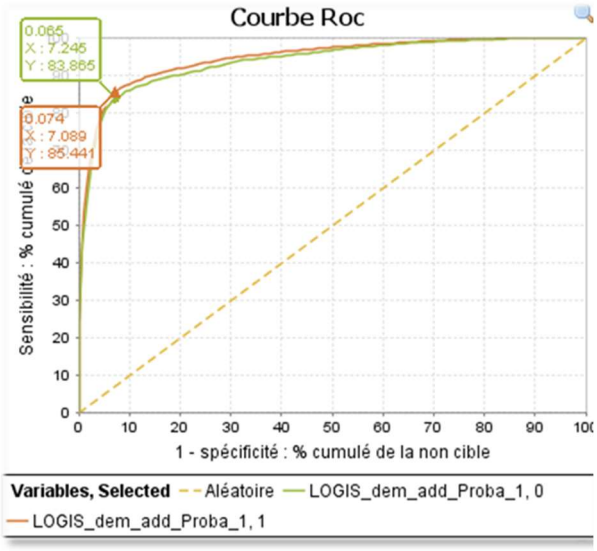
| Indicateur | Ech. Apprentissage | Ech. Test |
|---|---|---|
| Sensibilité (taux de vrais positifs) | 49% | 47% |
| Spécificité (taux de vrais négatifs) | 99% | 99% |
| Précision | 85% | 83% |
| Taux d'erreur | 4,50% | 4,64% |

Comme on l'a vu la sensibilité n'est pas assez élevé. L'idée désormais, c'est de **trouver un autre seuil de probabilité qui permet d'augmenter la sensibilité** pour détecter davantage de vrais cas positifs. En cela la courbe ROC nous apporte un solide éclairage.

### e. *Courbes ROC, Aires sous la Courbe (AUC) et Sur-apprentissage*




Seuil de probabilité **par défaut** : 50%




Seuil de probabilité qui **approche le coin nord-ouest** : autour de 7%

Nous n'observons pas de décrochage de la courbe ROC de Test *versus* celle d'Apprentissage. On ne constate donc pas de sur-apprentissage du modèle retenu.

SPAD permet de trouver empiriquement un seuil de probabilité répondant davantage à notre problématique. Il suffit de déplacer le curseur sur les courbes et identifier le seuil de probabilité qui s'approche au maximum du coin nord-ouest du graphique. C'est le seuil qui permet l'arbitrage optimum entre sensibilité et spécificité. **Je décide de retenir le seuil à 7%.** Remarquons que ce seuil empirique est proche du taux de démission volontaire constaté sur la base d'étude (7.5%). Par expérience, c'est un résultat qu'on retrouve souvent.

Au seuil de probabilité 7%, les indicateurs deviennent :

| Indicateur | Ech. Apprentissage | Ech. Test |
|---|---|---|
| Sensibilité (taux de vrais positifs) | 59% | 56% |
| Spécificité (taux de vrais négatifs) | 98% | 99% |
| Précision | 76% | 76% |
| Taux d'erreur | 4,61% | 4,72% |

Même si le taux de vrai positif reste relativement peu élevé, nous avons tout de même sensiblement amélioré. L'inconvénient c'est que mécaniquement cela vient diminuer la précision, c'est-à-dire la capacité du modèle à bien identifier les vrais positifs parmi les prédits positifs. Et par conséquence cela dégrade également le taux d'erreur du modèle. Mais ce dernier reste encore largement acceptable.

- **Autres méthodes traditionnelles :**

L'ensemble de la méthodologie décrite sur ces cinq dernières pages est reproduit sur les cinq autres méthodes dites « classiques » : Arbre de décision, Bayésien naïf, Analyse discriminante, Machine à vecteurs de supports et Réseau de neurones. Aucune d'entre elles ne parvient à dépasser voir à égaliser les performances de la Régression logistique.

L'arbre de décision présente une AUC et un taux d'erreur acceptables mais il ne retient que quatre variables explicatives, ce qui ne permet pas de caractériser des probabilités individuelles pertinentes pour la suite du modèle de valeur client (trop de valeurs à 0).

Le modèle de Bayes retient un nombre de dimensions discriminantes beaucoup plus important (37) mais le taux de mal classés est presque le double (8,8%).

L'analyse discriminante DISQUAL est encore pire puisque le taux d'erreur atteint 11,7% en test, et la précision s'écroule (37,2%).

Le réseau de neurones et le SVM affichent une AUC assez bonne mais les probabilités déduites des modèles sont inutilisables : plus de 80% d'entre elles sont évaluées à 0 (non sortant presque certain) ou à 1 (sortant presque certain), ce qui ne cadre pas du tout avec l'intuition « métier » du problème et ne nous aide pas du tout pour la suite de la modélisation de la Valeur Client.

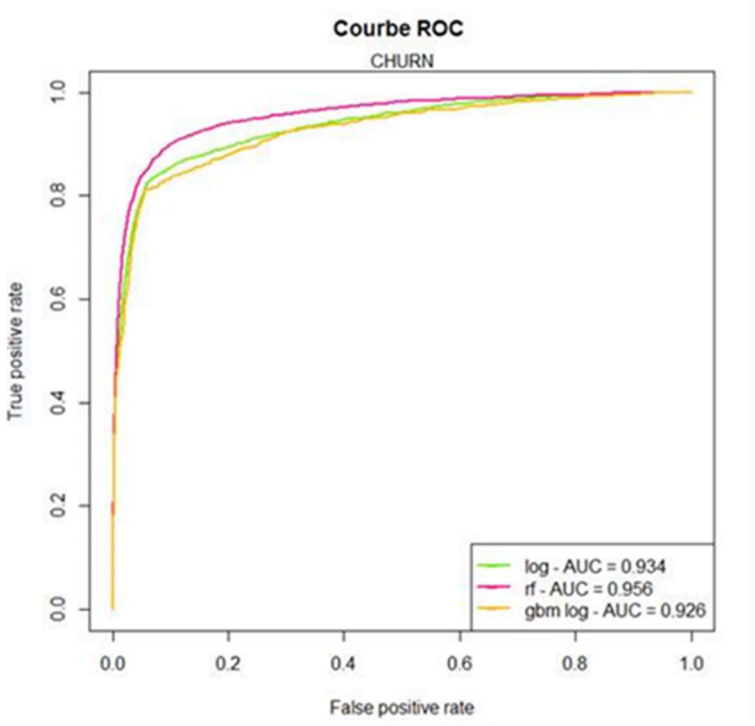
Ces cinq méthodes ne sont pas adaptées à notre problématique et/ou nécessitent un paramétrage spécifique qui n'est pas disponible avec les outils en notre possession. Elles sont donc éliminées du scope et nous ne poursuivons pas leur expérimentation. Par souci de cohérence, je les ai tout de même testés sur les autres scores. Les indicateurs AUC sont présentés dans la partie suivante mais les conclusions observées sur le score de démission volontaire sont identiques voir dégradées sur les autres scores.

- **Méthodes modernes de Data Science :**

Contrairement aux méthodes précédentes, les algorithmes modernes issus de la Data Science, apporte une réelle plus-value et challengent correctement la Régression Logistique. Elles ont le triple avantage de présenter des AUC plus consistants, des taux d'erreurs acceptables et des probabilités individuelles déduites réutilisables.

Dans ces cas là aussi, nous devons choisir un seuil de probabilité différent du seuil par défaut pour améliorer la sensibilité du modèle. Par souci de cohérence et pour comparer pertinemment, je conserve le seuil à 7%. Ce n'est pas sans inconvénient : le taux d'erreur progresse fortement, surtout avec les forêts aléatoires, et les taux de précision chutent. Mais à l'inverse, nous obtenons un meilleur équilibre taux de vrais positifs *vs* taux de vrais négatifs.

| | Indicateur | Seuil de Proba = 50% | Seuil de Proba = 7% |
|---|---|---|---|
| Random Forest | Sensibilité (taux de vrais positifs) | 66% | 90% |
| | Spécificité (taux de vrais négatifs) | 98% | 90% |
| | Précision | 78% | 43% |
| | Taux d'erreur | 4,0% | 9,8% |
| Gradient Boosting | Sensibilité (taux de vrais positifs) | 52% | 81% |
| | Spécificité (taux de vrais négatifs) | 98% | 95% |
| | Précision | 73% | 54% |
| | Taux d'erreur | 5,0% | 6,5% |




La courbe ROC des forêts aléatoires est remarquable : l'aire sous la courbe de ce modèle atteint 0.956, et ceci sans sur-apprentissage apparent.

**Les dimensions retenues en décroissant de leur pouvoir explicatif :**

| Random Forest | Gradient Boosting Machine |
|---|---|
| Nb bénéficiaires mis en classes | Nb bénéficiaires mis en classes |
| Région d'habitation | Type d'Iris |
| Tranche d'âge du client | Composition de ménages de l'Iris |
| Ancienneté mise en classe | Nb envoi de courrier mis en classe |
| CSP Max de l'Iris | Taux de chômage de l'Iris |
| Canal d'entrée recodé | Durée de vie passée mise en classe |
| Répartition par tranche d'âge de l'Iris | Répartition par tranche d'âge de l'Iris |
| *Nb envoi de courrier mis en classe* | CSP Max de l'Iris |
| *Direction des Ventes* | *Régime Sécurité Sociale* |
| *Type d'Iris* | *Composition familiale* |
| *Régime Sécurité Sociale* | *Couvert en prévoyance par le contrat Statutaire* |
| *Niveau de garantie* | *Tranche d'âge du client* |
| *Situation familiale* | *Espace Adhérent ouvert* |

Les dimensions en italique sont statistiquement discriminantes mais leur apport dans l'explication du modèle est marginal. Elles sont toutefois utiles car elles viennent affiner les probabilités individuelles de démission volontaire.

### 6.1.5. Comparaison des méthodes testées

Comme on l'a vu sur la partie précédente, pour le score de démission volontaire (churn), la méthode des forêts aléatoires (5 000 arbres générés) présente nettement le meilleur AUC. La régression logistique fait jeu égal avec le GBM.

Concernant l'appétence à la GAC, la méthode des forêts aléatoires l'algorithme des **forêts aléatoires** arrive de nouveau légèrement en tête, devançant de peu le **gradient boosting**. Le bayésien naïf fonctionne également très bien.

| | Aire sous la courbe de l'échantillon test | | | | | |
|---|---|---|---|---|---|---|
| **Méthodes** | Appétence à Kimono | Appétence à Helpeo | Appétence à MCT | Appétence à Poséo | Risque de pré-contentieux | Proba de churn |
| **Régression logistique** | 0,787 | 0,769 | 0,929 | 0,821 | **0,773** | 0,934 |
| **Arbre de décision** | 0,731 | 0,651 | 0,844 | 0,845 | 0,745 | 0,904 |
| **Analyse discriminante** | 0,708 | 0,641 | 0,739 | 0,790 | 0,737 | 0,857 |
| **Bayésien naïf** | 0,810 | 0,786 | 0,926 | 0,848 | 0,760 | 0,906 |
| **Réseau de neurones** | 0,729 | 0,757 | **0,940** | 0,811 | 0,768 | 0,854 |
| **SVM** | Non aboutis | | | | | 0,880 |
| **Random Forest** | **0,841** | 0,910 | 0,890 | 0,690 | Non aboutis | **0,956** |
| **Gradient Boosting (Adaboost)** | 0,829 | 0,866 | 0,864 | **0,870** | | Non abouti |
| **Gradient Boosting (Log)** | 0,830 | **0,927** | 0,927 | 0,838 | | 0,926 |

Je n'ai pas poursuivi l'expérimentation avec la méthode SVM car elle est extrêmement lourde (104 h de traitement sur mon PC pour modéliser la probabilité de démission volontaire !!) et par ailleurs, soit l'algorithme ne converge pas, soit les résultats sont non pertinents. A retester éventuellement sur une machine super performante dédiée au Machine Learning et avec éventuellement un paramétrage spécifique adéquat.

**Les algorithmes modernes de l'école *'Data Science'* sont plus performants en moyenne.** Ils ne convergent toutefois pas pour modéliser le risque de pré-contentieux (une seule variable discriminante). La régression logistique affiche également des performances intéressantes, ce qui la qualifie pour développer le modèle prototype.

| | Résultats de la régression logistique | | | | | |
|---|---|---|---|---|---|---|
| Indicateurs | Appétence à Kimono | Appétence à Helpeo | Appétence à MCT | Appétence à Poséo | Risque de pré-contentieux | Proba de churn |
| $R^2$ ajusté | 0,178 | 0,177 | 0,194 | 0,197 | 0,115 | 0,705 |
| Proba du test Hosmer-Lemeshow | 0,249 | 0,252 | 0,706 | 0,448 | 0,152 | 0,205 |
| AUC Pop. Totale | 0,804 | 0,790 | 0,954 | 0,828 | 0,774 | 0,943 |
| AUC Ech. apprentissage | 0,819 | 0,801 | 0,966 | 0,833 | 0,774 | 0,946 |
| AUC Ech. test | **0,787** | **0,769** | **0,929** | **0,821** | **0,773** | **0,934** |

Toutes les AUC des échantillons de test sont largement convenables. L'AUC minimale est atteinte pour la modélisation de l'appétence à la Protection Juridique avec 0.769, ce qui correspond à une discrimination acceptable. Les écarts avec les AUC des échantillons d'apprentissage sont relativement faibles, on n'observe donc pas de sur-apprentissage.

### 6.1.6. Stratégie de calcul des scores retenue

Lors de la phase d'industrialisation opérationnelle et de généralisation à tous les clients en gestion directe LMG, nous avons développé le modèle sur **Python**, ce qui permet de substituer la régression logistique par une méthode moderne. Nous avons convenu de la stratégie suivante, qui nous semble à ce stade la plus pertinente :

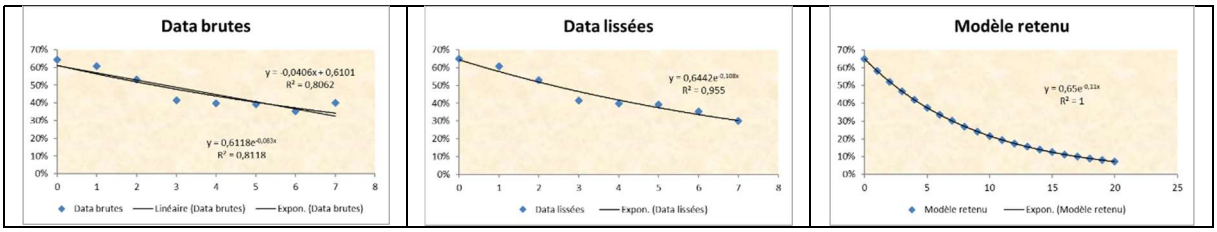
Attention ! « Score » ou « Probabilité » ? :

Lorsque l'on utilise les méthodes de scoring pour estimer des probabilités d'événements, il faut faire preuve de prudence. Travaille-t-on des simples scores ou bien des vraies probabilités ? Il est parfois nécessaire de calibrer les scores obtenus pour en faire des probabilités « praticables ». En effet, le scoring dont la fonction principale est de classer et déterminer des positions relatives, peut générer des « probabilités » dont l'espérance mathématique n'est plus tout à fait alignée avec celle qui est observée sur l'échantillon d'apprentissage. Si par exemple, j'avais obtenu une moyenne des scores de churn sur la population à scorer très différente du taux de démission volontaire constaté dans un passé proche (8,8% en 2016), il aurait fallu procéder à un ajustement linéaire de tous les scores de façon à travailler la suite du modèle avec une moyenne proche de l'expérience. Ce ne fut pas nécessaire dans ce cas précis, car l'espérance du score était à peu près égal à l'attendu.

### 6.1.7. Lien entre radiation et pré-contentieux

La modélisation de la probabilité de pré-contentieux donne de meilleurs résultats que la modélisation de la probabilité de radiation. Or, nous constatons empiriquement que ces deux grandeurs semblent liées. En effet, lors de la 1ère année de souscription, environ 65% des dossiers en pré-contentieux aboutissent à une radiation du client ; et cette proportion décroît avec le temps. On se propose donc de modéliser le rapport probabilité de radiation / probabilité de pré-contentieux, selon l'ancienneté dans le portefeuille, afin de substituer l'une par l'autre dans l'équation de la probabilité de survie.[83]

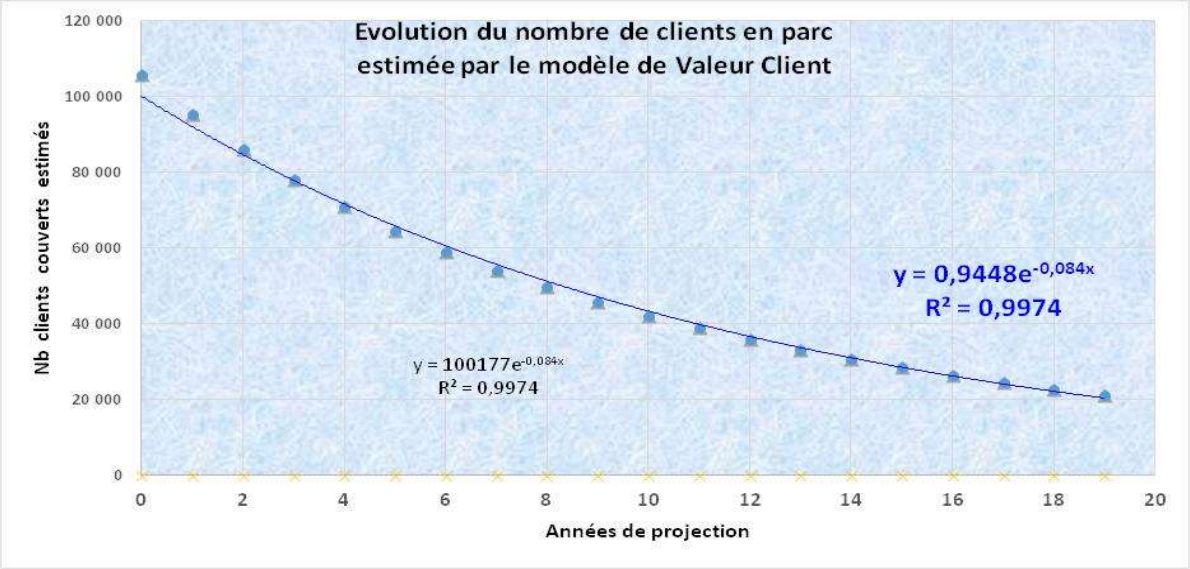
A partir des données brutes, une modélisation par une courbe de tendance exponentielle dans Excel fait à peine mieux que la régression linéaire ($R^2$ = 81,1% vs 80,6%). En lissant légèrement les données par la méthode de Whittaker-Henderson[84], le pouvoir explicatif à travers une fonction exponentielle, s'améliore très nettement ($R^2$ = 95,5%). Nous retenons *in fine* sur l'ensemble de la période, le modèle : $\mathrm{P(radiation)} = 0{,}65\, e^{-0{,}11 * Ancienneté} \times \mathrm{P(pré\text{-}contentieux)}$.




[83] Voir fin du paragraphe 3.1.1.
[84] http://www.nber.org/chapters/c9366.pdf ou http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance6.pdf?OpenElement, partie 3.3.

### 6.1.8. Evolution des effectifs couverts

*6.1.3.1 Loi de survie empirique pour le risque Santé*




**106 034 adhérents sont couverts** par un contrat Santé interpro au 31 mars 2017. Le modèle prévoit qu'au bout de 8 ans, on aura perdu la moitié de nos effectifs assurés aujourd'hui. En pratique, le portefeuille n'étant pas en « run off », nous aurons entre temps enregistré des affaires nouvelles qui viendront compenser ces départs probables.

**L'adéquation** entre la fonction de survie empirique obtenue et la courbe de tendance donnée par Excel **est excellente**, le coefficient de détermination valant en effet **99,7%.** Notons qu'un test du $\chi^2$ [85] rejette ici largement l'hypothèse H0 d'adéquation mais ce test n'est pas du tout adapté dans le cas d'effectifs très importants.[86]

L'équation de la courbe de tendance est de la forme : $y = \alpha\, e^{-\beta x}$, donc *in fine*, la fonction de survie empirique est une exponentielle. Comme quoi, on a beau essayer s'émanciper en testant de nouvelles méthodes, on est rattrapé par la théorie ! (Cf. 3.1.1.) C'est en fait plutôt rassurant.

[85] http://gandalfmagicien.free.fr/psycho/Licence%203/Premier%20Semestre/CM%20Maths/chapitre5.pdf
[86] http://www.lmpt.univ-tours.fr/~gallardo/Stat2008-2.pdf : voir proposition 19

#### *6.1.3.2 Formule de la probabilité cumulée d'appétence à un produit de multi-équipement prévoyance*

Chaque client non couvert par un contrat de multi-équipement a chaque année une probabilité $p$ de souscrire à un contrat de prévoyance donné. Si cette probabilité $p$ était constante dans le temps pour un assuré, la probabilité cumulée d'appétence serait :

$$1 - (1 - p)^n = \sum_{k=0}^{n-1} p^1 (1 - p)^k$$

#### *6.1.3.3 Impacts réglementaires*

A propos de l'utilisation du score de churn pour estimer la valeur client, il est absolument nécessaire de faire attention à l'interprétation des résultats obtenus par le modèle. Nous avons dit dans la partie 3 que nous raisonnions à environnements réglementaire et économique stables. C'est l'usage dans ce type de problématique et c'est une hypothèse assez confortable pour le technicien. Toutefois, la réalité est tout autre : nous naviguons en effet actuellement dans un environnement très chahuté :

- L'ANI 2013 qui impose à toutes les entreprises de couvrir ses salariés par un contrat collectif en santé a vigoureusement modifié l'équilibre individuel *vs* collectif.
- Le dispositif ACS[87] réformé en 2015 a également contribué à des transferts massifs de contrats individuels interpro vers des contrats de complémentaire santé dédiés.
- Pour la population des jeunes retraités, désormais identifiée comme prioritaire chez de nombreux assureurs, la concurrence est exacerbée, les offres sont de plus en plus alléchantes, ce qui implique des nouveaux mouvements entrées-sorties.

Le score de churn est estimé en fonction du passé mais des impacts réglementaires forts peuvent profondément changer la dynamique. Ainsi le taux de churn est manifestement altéré par des facteurs extérieurs dont l'impact est non négligeable ces dernières années. Et il faut bien le garder à l'esprit sans quoi cela pourrait nous conduire à des erreurs d'interprétation.

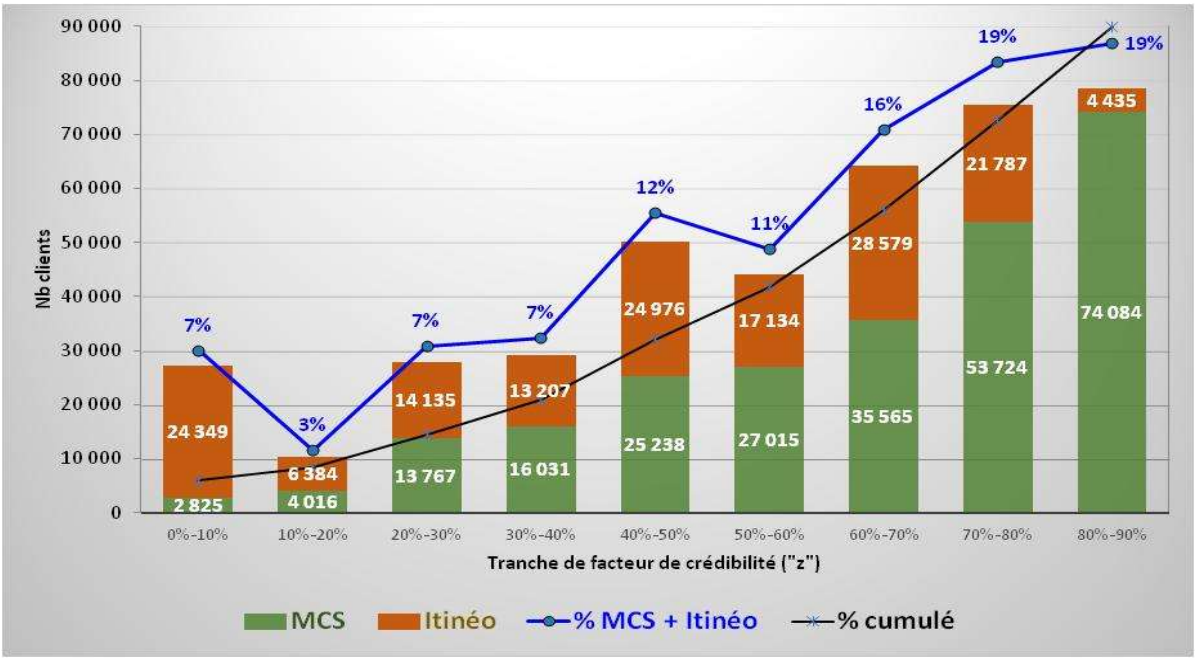
[87] ACS : Aide à la Complémentaire Santé - https://www.info-acs.fr/

## 6.2. Premiers résultats de VC

### 6.2.1. Les principaux enseignements des travaux préparatoires

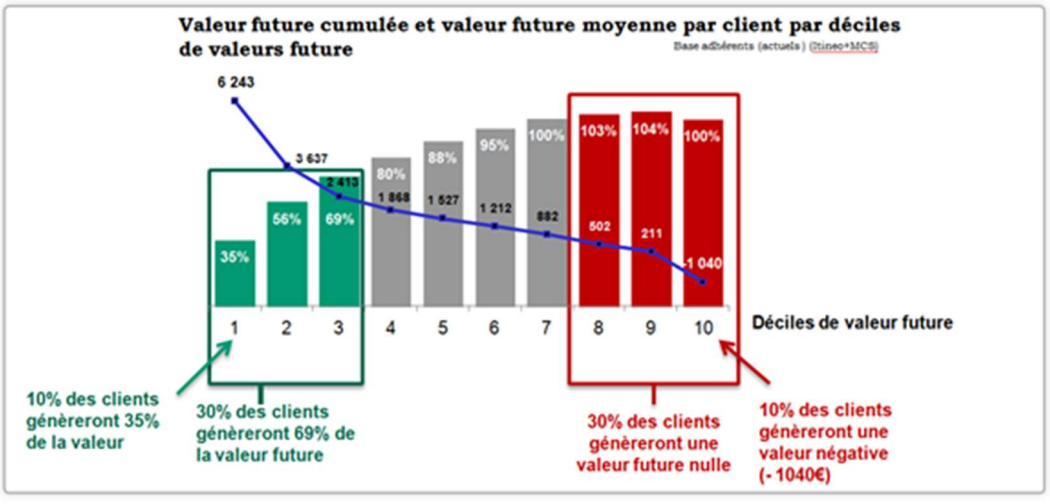

D'une manière générale, la valeur client annuelle moyenne s'améliore au cours de la projection, ce qui est concordant avec l'intuition. En effet, le plus souvent **un portefeuille se solvabilise avec le temps** : la part des opportunistes et des mauvais payeurs se réduit au fil des acquisitions et radiations. C'est encore plus vrai lorsque le portefeuille passe en *« Run Off »*.

### 6.2.2. Répartition du facteur de crédibilité



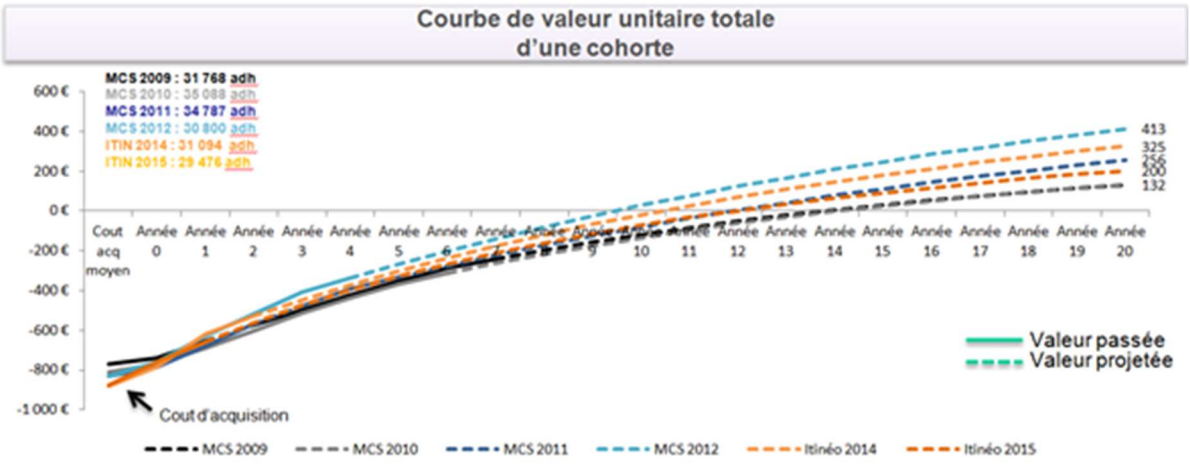

Le facteur de crédibilité est par construction majoré à 90% (parti-pris). Au moins 10% du P/C futur d'un assuré est expliqué par le P/C de son groupe. Plus de 50% de la population possède un facteur de crédibilité ($z$) supérieur à 60%. $\mathbf{M_E(z)_{TOT} = 63{,}4\%}$ ; $\mathbf{\bar{z}_{TOT} = 57{,}3\%}$. Où l'on voit ici l'importance de l'ancienneté dans le calcul : les assurés MCS, produit commercialisé de 2008 à 2013, ont majoritairement un « z » fort tandis que les « z » faibles sont principalement le fait des assurés Itinéo, plus récents. Les indicateurs de tendance centrale viennent confirmer cette perception : $\mathbf{\bar{z}_{MCS} = 65\%}$ ; $\mathbf{\bar{z}_{ITI} = 45\%}$ ; $\mathbf{M_E(z)_{MCS} = 71\%}$ ; $\mathbf{M_E(z)_{ITI} = 50\%}$ ; Les assurés Itinéo avec « z » fort sont essentiellement des transferts de MCS ou des ex-enfants d'adhérents affinitaires.

### 6.2.3. Répartition de la valeur




Une forte dispersion de la valeur future qui justifie une approche segmentée de la fidélisation.

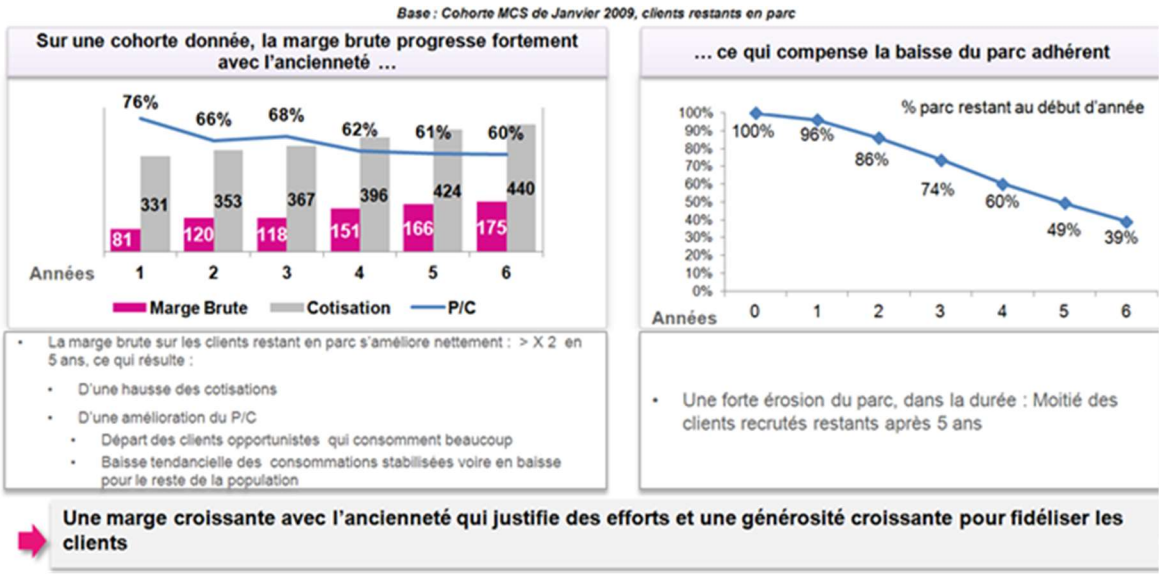
### 6.2.4. VC par cohorte



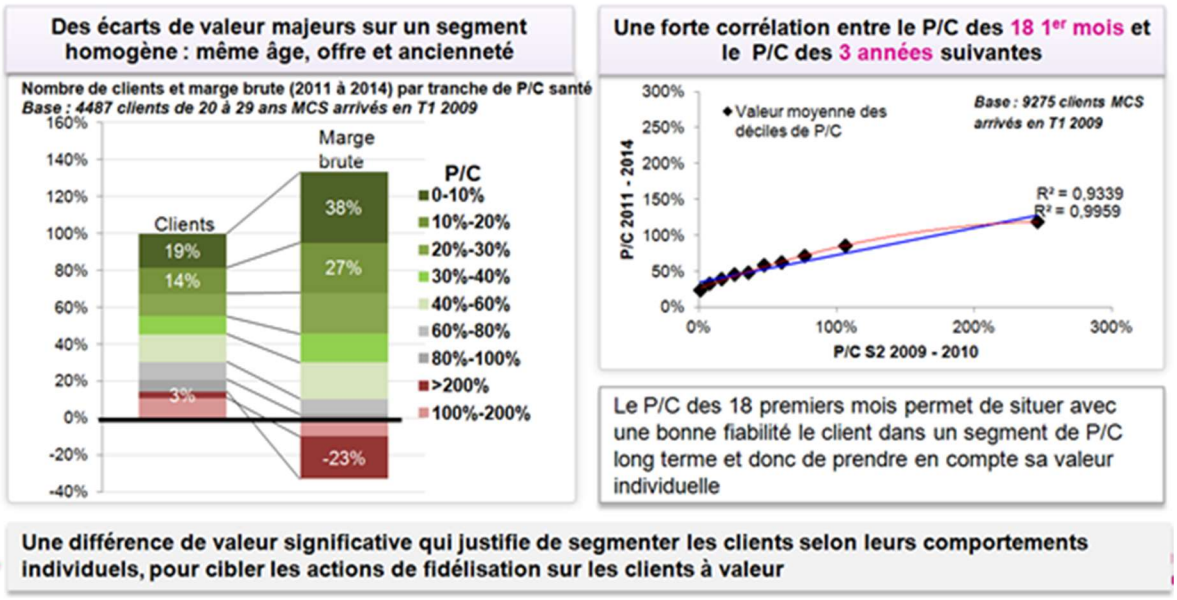

On observe des écarts de valeur importants entre les cohortes. Notamment la valeur des cohortes du produit 1 s'améliore avec le temps. L'explication vient d'une évolution de la politique commerciale entre 2008 et 2012. L'effet modulaire du produit 1 était pleinement exploité les premières années. Mais cela engendrait trop d'anti-sélection et l'équilibre technique n'était pas atteint sur les combinaisons de garanties hétérogènes. Graduellement, la Direction Générale a incité les commerciaux à proposer davantage les garanties monoblocs. Et la combinaison bas de gamme est sortie progressivement du catalogue.

Les coûts d'acquisition moyens varient d'une année à l'autre autour de 800 €, en fonction de la part des nouveaux arrivants (1 050 €) et celle des anciens enfants d'adhérents affinitaires (50 €). Selon les cohortes, l'amortissement des coûts d'acquisition est atteint entre 9 ans et 14 ans !

### 6.2.5. VC par ancienneté

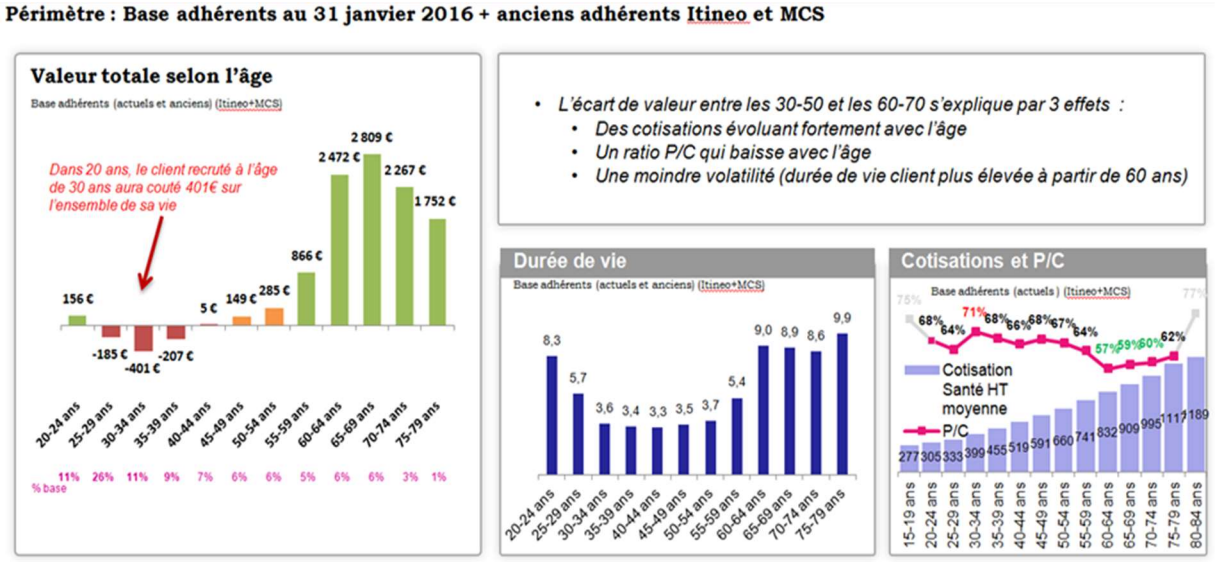


### 6.2.6. VC selon P/C



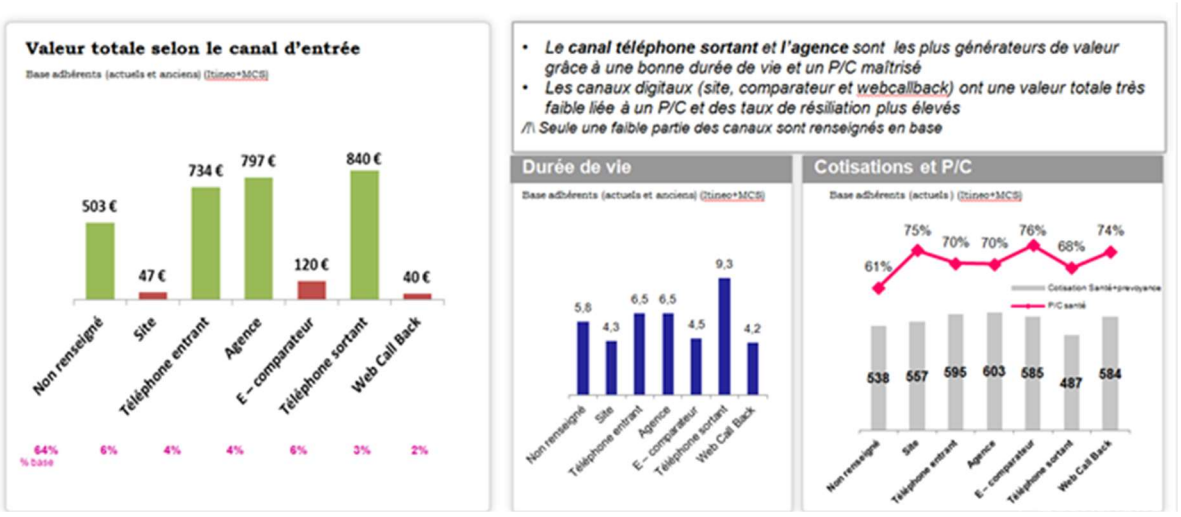

Il y a un lien évident entre le P/C (vision sur une année de survenance) et la valeur client (sur 20 ans au moins) mais celui-ci n'explique pas tout. Notons que moins de 15% des adhérents dégradent près de 35% de la valeur totale.

### 6.2.7. VC par tranche d'âge




On s'en doutait a priori. L'analyse confirme l'intuition : **l'âge est un driver majeur de la valeur client**. Désormais, nous sommes en droit de nous poser la question : est-il encore justifié de continuer à consacrer tant d'investissement financier, en temps et en énergie pour chercher à conquérir autant de clients de 25 à 40 ans ? La réponse est semble-t-il dans la question. L'effet ANI 2013 et le basculement progressif de tous les salariés de l'individuel vers les contrats collectifs amplifie le phénomène…

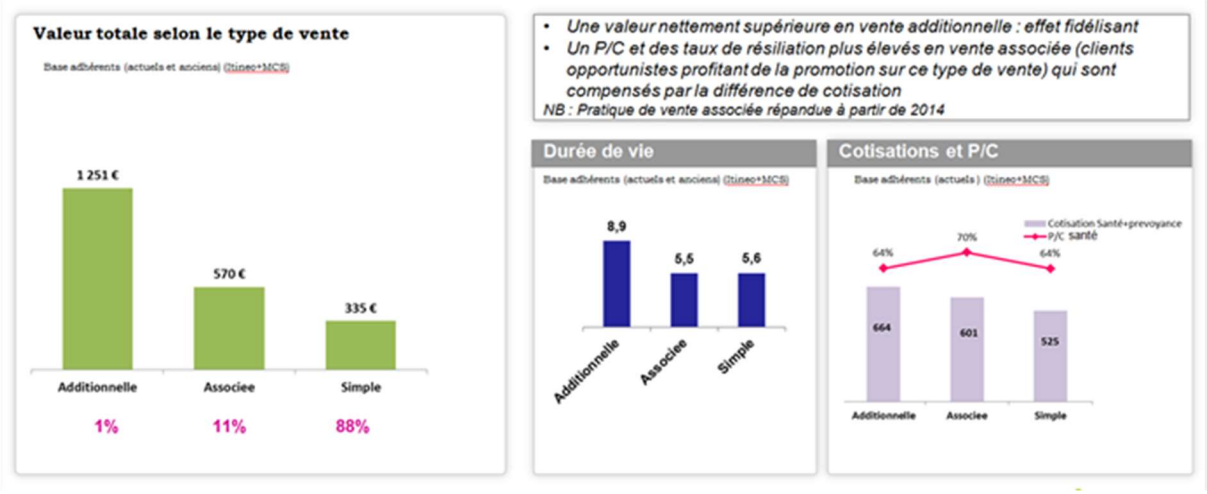
### 6.2.8. VC par canal d'entrée




Malgré le caractère multi-canal de la majorité des clients (recherche sur les comparateurs ou appel entrants puis contractualisation en téléphone sortant ou en agence), le canal d'entrée renseigné en base est révélateur de la valeur. Et les résultats sont conformes à l'intuition. Modifier la politique d'acquisition ? C'est en cours !

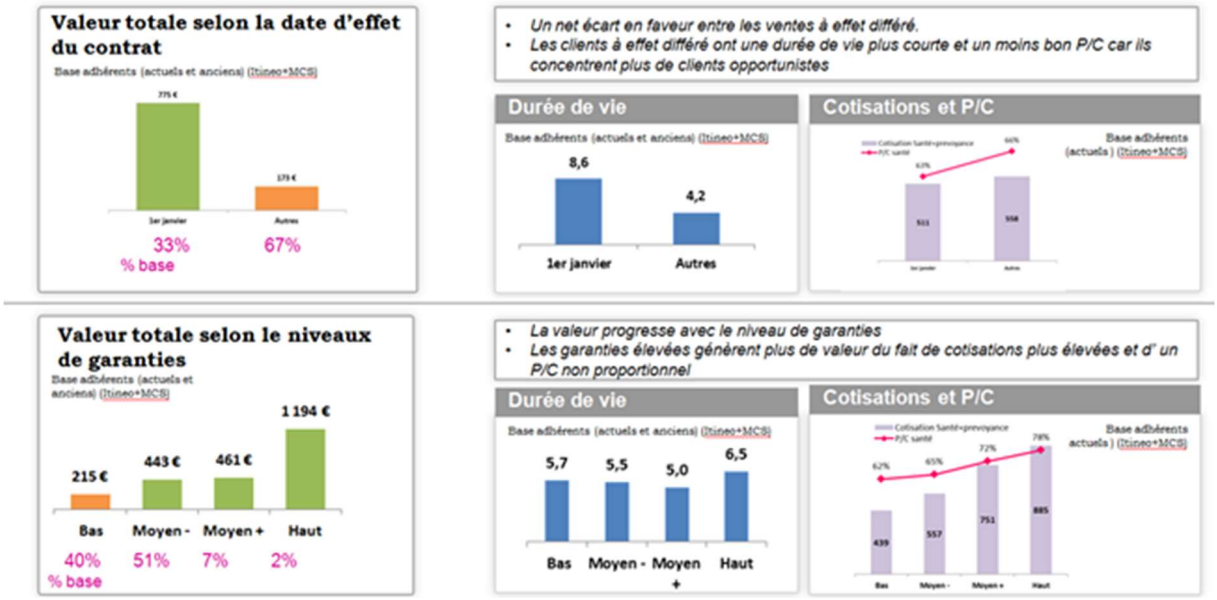
Le taux de non renseigné est très important. Désormais les conseillers mutualistes sont sensibilisés à la complétude et la qualité des données (canal d'entrée comme d'autres) enregistrées dans le CRM. LMG envisage même des « *incentives* » voir des challenges. Révolution ? Adaptation !

### 6.2.9. VC par type de vente




La vente additionnelle est fortement créatrice de valeur et la vente associée demeure meilleure que la vente simple malgré des P/C moins bons. Multi-équiper génère de la valeur ! Mais la manière de proposer une offre est essentielle : à court terme, on peut se satisfaire des ventes « en pack » qui génèrent du CA immédiat mais ne garantit en rien la satisfaction et la pérennité des contrats ; tandis que sur la durée, répondre à un réel besoin de couverture du client, est toujours profitable, pour le client et pour l'entreprise.

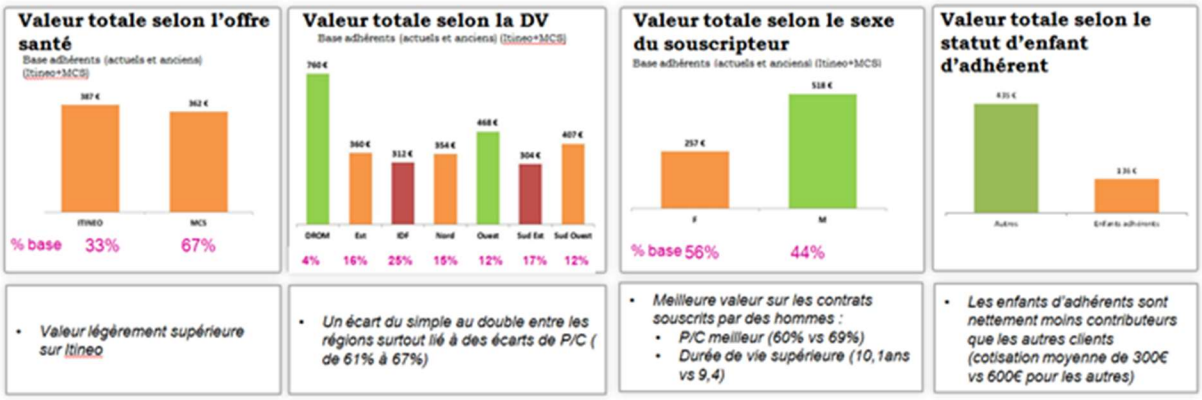
### 6.2.10. VC par type de souscription




La souscription à effet différé, principalement au 1er janvier, regroupe des clients qui jouent davantage le jeu de l'assurance contrairement aux clients qui souscrivent à effet immédiat ou en cours d'année, plus optimisateurs et moins solvables. Cela se confirme nettement sur les 3 indicateurs.

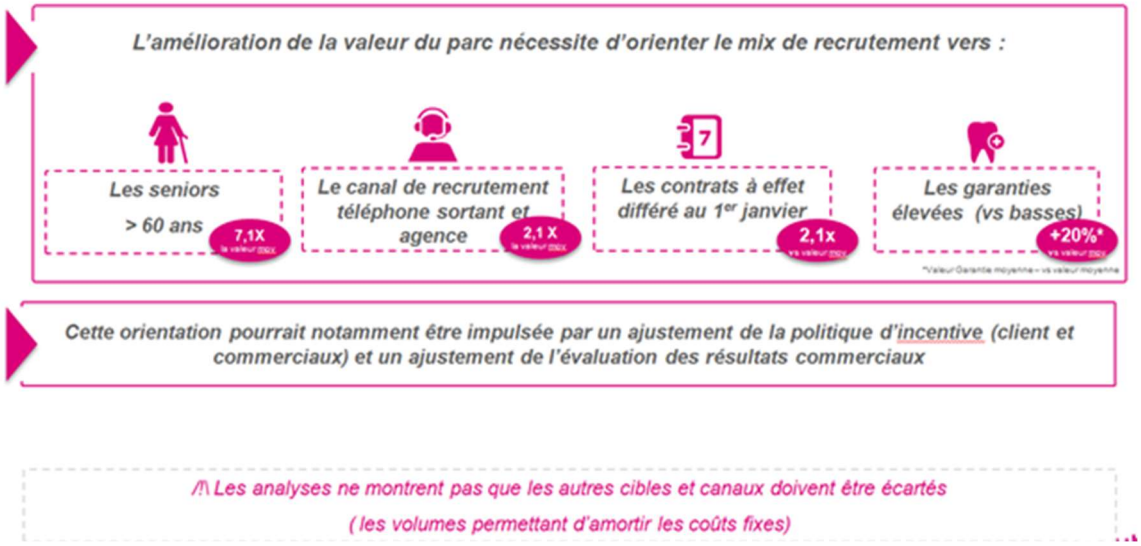
Les niveaux de garantie élevés sont largement plus générateurs de valeur, en dépit de ratios de sinistralités supérieurs. L'effet volume joue à plein.

### 6.2.11. VC selon autres dimensions

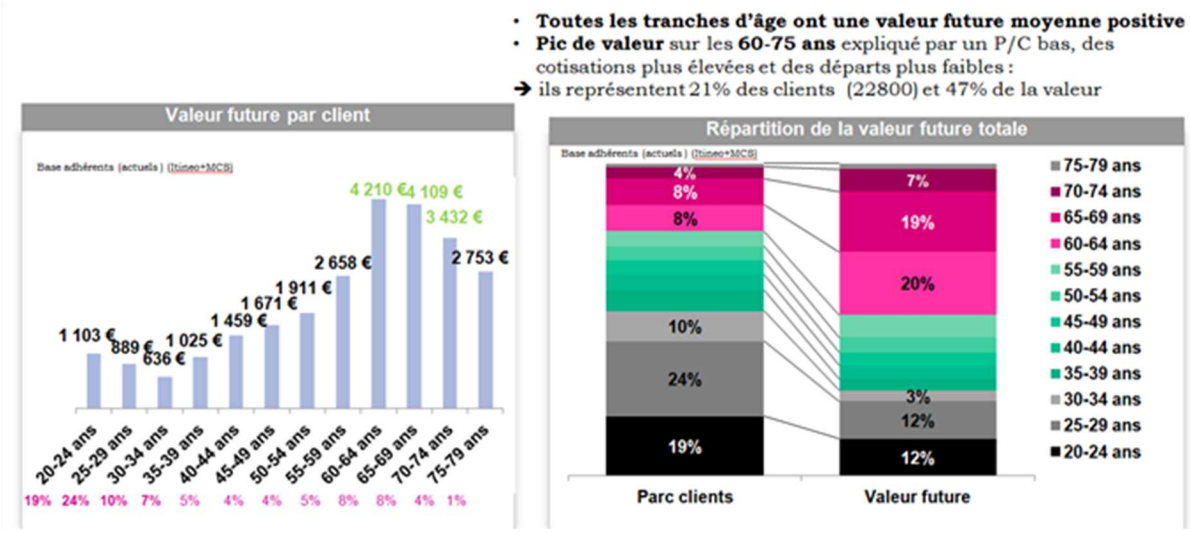



- Produit santé n°2 : moins de clients avec des garanties bas de gamme
- La valeur est corrélée négativement avec l'offre de soins. Une piste ?
- Malgré des coûts d'acquisition très faibles, les enfants d'adhérents génèrent en moyenne une valeur nettement inférieure à la moyenne. Ils sont surtout plus jeunes donc avec un levier volume ténu. Et par ailleurs, ils profitent à vie d'une remise de cotisation de 10%...

### 6.2.12. Politique d'acquisition proposée

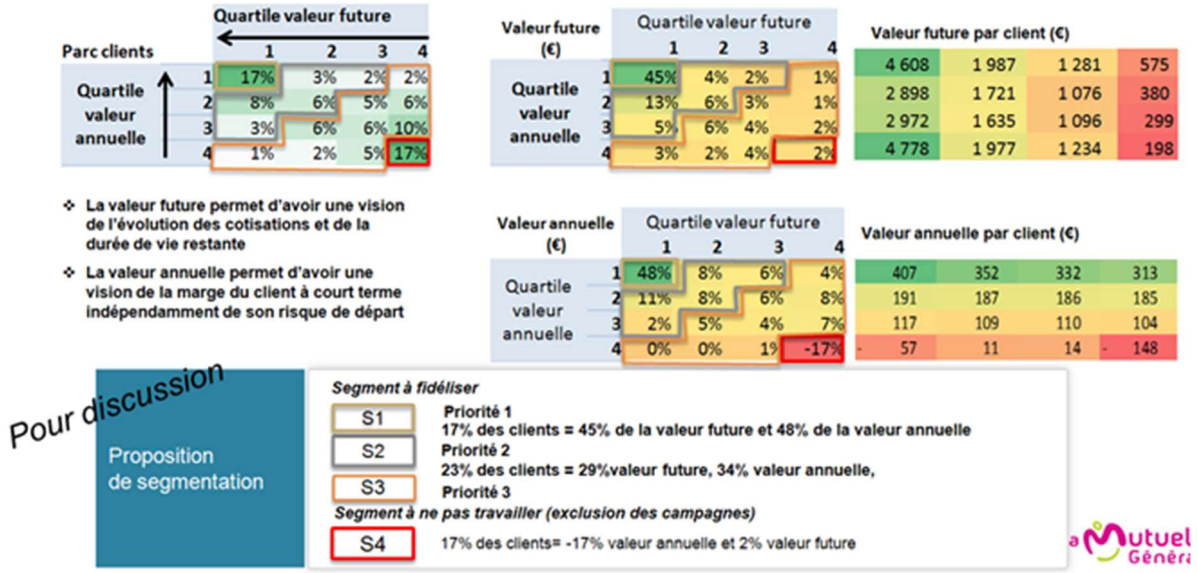



### 6.2.13. Valeur future par tranche d'âge




Une fois enregistré le coût d'acquisition, tout segment du portefeuille est fidélisable et à fidéliser. La sélection, si elle doit avoir lieu, se situe nécessairement en amont, en période de définition de la stratégie de conquête. Par la suite, les efforts de cocooning, de rétention, et de développement de l'expérience client sont fructueux pour tous les clients. Cela n'empêche pas, à un degré plus fin, LMG d'adopter une stratégie différenciée au cas par cas.

### 6.2.14. Segmentation selon quartile de valeur




Proposition de segmentation intéressante mais finalement non retenue, car non suffisamment orientée opérationnelle, et très peu lisible par les conseillers mutualistes.

**<u>Caractéristiques des 4 segments définis :</u>**

| | Nb clients | Valeur future moyenne | Valeur annuelle moyenne | P/C 2016 | Cotisation Santé 2016 | Cotisation Prevoyance | Nombre de contrats | Durée de vie estimée |
|---|---|---|---|---|---|---|---|---|
| S1 | 19K | 4 608 € | 407 € | 44 % | 861 € | 16 € | 1,17 | 11 |
| S2 | 25K | 2 291 € | 220 € | 47 % | 529 € | 11 € | 1,17 | 10 |
| S3 | 48K | 997 € | 117 € | 61 % | 410 € | 12 € | 1,22 | 8 |
| S4 | 18K | 198 € | -148 € | 120 % | 456 € | 13 € | 1,21 | 6 |

| | Age 2016 | Ancienneté | Bénéficiaires | Segmentation familiale S0 | S1 | S2 | S3 | S4 | S5 | S6 | Espace adhérent | Email | Téléphone |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 62 ans | 3,8 ans | 1,4 | 1% | 5% | 7 % | 2% | 7% | 64% | 15% | 28% | 55% | 94% |
| S2 | 42 ans | 3,9 ans | 1,4 | 14 % | 38% | 11 % | 3% | 7% | 22% | 6% | 30% | 59% | 92% |
| S3 | 33 ans | 2,9 ans | 1,3 | 29% | 43 % | 12 % | 3% | 5 % | 7% | 2% | 32% | 66% | 89 % |
| S4 | 34 ans | 2,6 ans | 1,3 | 19% | 51% | 12 % | 3% | 6 % | 7% | 2% | 40% | 75% | 88 % |

### 6.2.15. Modèle Linéaire Généralisé[88]

Les résultats précédents sont intéressants à analyser en tant que tels car ils donnent une bonne idée de la répartition de la valeur selon les modalités des variables étudiées. Mais, on se rend rapidement compte qu'une interprétation fine des effets est impossible en raison des corrélations inéluctables entre dimensions ; si bien qu'il est parfois difficile d'en tirer des enseignements pertinents.

L'âge et le canal d'entrée, par exemple, sont liés. Les jeunes ont davantage propension à utiliser les canaux digitaux tandis que les plus âgés choisiront plus souvent des canaux traditionnels (cliché). Or, entre 30 et 70 ans, la VCT est croissante avec l'âge. Analyser la VC selon le seul canal d'entrée peut donc conduire à des erreurs d'interprétation. C'est comme si nous décidions d'étudier la liaison entre la taille d'un individu et la longueur de ses cheveux sans tenir compte du sexe. Il s'agit du phénomène bien connu de « variable cachée ».

En conséquence, nous devons essayer de **neutraliser ces corrélations** afin d'obtenir, des évaluations des effets de chacune des dimensions sur la valeur, « toutes choses égales par ailleurs » ; ce qui pourrait permettre d'identifier des aspects moins intuitifs que l'analyse non décorellée. Pour ce faire, le meilleur moyen est de s'appuyer sur un **Modèle Linéaire Généralisé.**

Les MLG sont très utilisés en assurance non-vie pour modéliser les fréquences (loi de Poisson), et les coûts annuels de sinistres (lois gamma) dans le but d'affiner des tarifications. Mais ils peuvent aussi servir à modéliser des variables binaires ou multinomiales. Dans notre cas, un MLG permet d'étudier la liaison entre la variable dépendante (VCT) et des variables explicatives prédéfinies. Ces modèles ont trois composantes :

- La composante aléatoire :

On construit un échantillon $(Y_1 \dots Y_n)$ de taille $n$, dont les variables aléatoires sont supposées indépendantes entre elles, de la variable à expliquer $Y\ (= VCT)$ qui suit une loi de probabilité appartenant à la famille des lois exponentielles.

- La composante déterministe :

Vue sous la forme d'une combinaison linéaire $\alpha_0 + \sum \alpha_i X_i$, appelée aussi « prédicteur linéaire » Les $X_i$ sont des variables explicatives qualitatives et quantitatives :

  - Caractéristiques Socio-démographiques du client : âge, sexe, segment/composition familiale, ex-enfant, région (DV)
  - Dimensions relatives au contrat ou à la souscription : ancienneté, type de vente, jour de souscription, canal d'entrée, niveau de garantie, multi-équipement, VCP …

- Le lien entre composante déterministe et aléatoire :

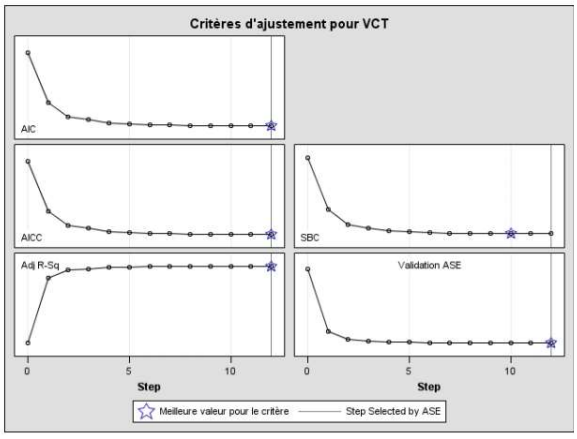
[88] Présentation succincte du MLG : http://maths.cnam.fr/IMG/pdf/Presentation_MODGEN_02_2007.pdf
Présentation plus fournie : S. Tufféry : « Data Mining et statistique décisionnelle, 4ème édition » - Chapitre 13.

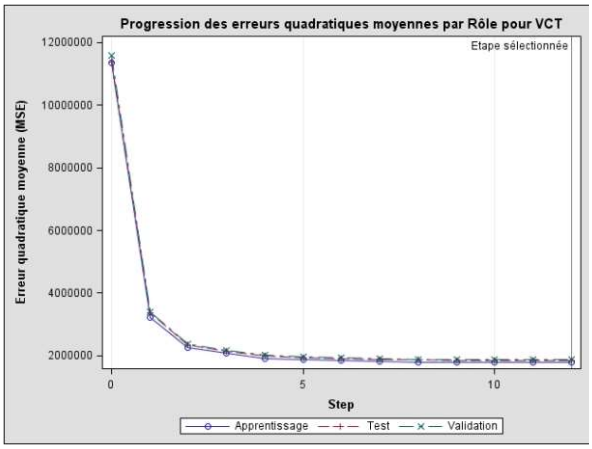
On suppose qu'il existe une liaison entre l'espérance de la variable à expliquer et les $k$ variables explicatives $X_i$, $i = 1 \dots k$, de la forme : $g(E(Y)) = \alpha_0 + \sum \alpha_i X_i$. La fonction $g$ est la fonction de lien du modèle.

Deux procédures SAS ont été testées : `GLMSELECT` et `GENMOD`. Elles permettent toutes les deux d'établir la significativité globale du modèle, de sélectionner les dimensions les plus déterminantes et d'identifier les modalités les plus significatives. La première méthode propose aussi une cross-validation, c'est-à-dire une élaboration du modèle sur un échantillon d'apprentissage, une optimisation des paramètres sur un échantillon de validation et enfin une validation sur l'échantillon de test. Nous vérifions que les deux procédures renvoient globalement les mêmes résultats et nous retenons `GLMSELECT` associée au processus `stepwise` pour sélectionner les dimensions les plus influentes et à l'option `lasso` pour ne retenir que les modalités significatives.

**Résultats de la modélisation :**

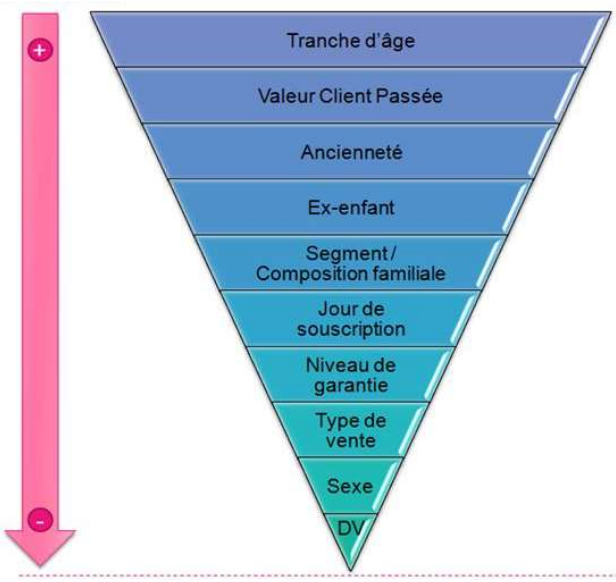
1. Tous les effets ont une p-value < 0.05, cela signifie que toutes les dimensions sélectionnées apportent de l'information pour expliquer la valeur client totale.
2. Le modèle converge rapidement : au bout de 5 itérations, les critères d'ajustement (AIC, AICC, SBC, $R^2$ adj., et validation ASE) atteignent une valeur proche de leur valeur finale.
3. D'autre part, on ne constate pas de sur-apprentissage : les erreurs quadratiques moyennes des échantillons 'validation' et 'test' ne divergent pas par rapport à ceux de l'échantillon 'apprentissage'.
4. Enfin, le coefficient de détermination ajusté[89] qui mesure l'adéquation entre un modèle et les données observées qui ont permis de l'établir, et qui se définit aussi comme la part de variance expliquée dans la variance totale, est très bon. $R^2$ ajusté = 0,843. C'est-à-dire que 84,3% de la dispersion est expliquée par le modèle, ce qui est très satisfaisant.




[89] Ajusté car cet indicateur tient compte du nombre de variables. En effet, le principal défaut du $R^2$ non ajusté est de croître avec le nombre de variables explicatives. Or, on sait qu'un excès de variables produit des modèles peu robustes. C'est pourquoi on s'intéresse davantage au $R^2$ ajusté qu'au $R^2$.

**L'importance décroissante des variables :**




**Les modalités les plus significatives :**

| INFLUENCES NEGATIVES | |
|---|---|
| VCP | • Valeur Client Passée < 500 €<br>• En particulier ceux qui ont une VCP < 5 000 € |
| Tranche d'âge | • Les clients actuels âgées de 20 à 39 ans |
| Ex-enfant | • Clients non présents dans nos bases auparavant en tant qu'« enfant d'adhérent » |
| Sexe | • Féminin |
| Type de vente | • Le client a choisi le seul produit « Santé » lors de la souscription |
| Niveau de garantie | • Les garanties bas de gamme |
| Ancienneté | • Les clients avec 7-8 ans d'ancienneté |
| Composition familiale | • Les familles |

| INFLUENCES POSITIVES | |
|---|---|
| VCP | • Valeur Client Passée ≥ 1 000 €<br>• En particulier ceux qui ont une VCP > 4 000 € |
| Tranche d'âge | • Les clients actuels âgées de 50 à 74 ans<br>• Il y a une forte contribution des 60-69 ans |
| Ancienneté | • Moins d'un an d'ancienneté |
| Jour de souscription | • Le 1er janvier |
| Composition familiale | • Solos/duos seniors |
| Niveau de garantie | • Les garanties moyen de gamme + et haute |
| Type de vente | • Vente additionnelle et vente associée |
| DV | • DROM |

Nous retrouvons globalement les résultats préétablis aux paragraphes précédents, et qui sont conformes à nos intuitions initiales. Mais les conclusions du MLG sont plus appropriées car les effets sont décorrelés entre eux, et le modèle permet de quantifier l'apport de chacune des modalités.

## 6.3. Analyses de sensibilité

*« L'analyse de sensibilité globale (AS) permet d'analyser un modèle mathématique en étudiant l'impact de la variabilité des facteurs d'entrée du modèle sur la variable de sortie. Déterminant les entrées responsables de cette variabilité à l'aide d'indices de sensibilité, l'AS permet de prendre les mesures nécessaires pour diminuer la variance de la sortie si celle-ci est synonyme d'imprécision, ou encore d'alléger le modèle en fixant les entrées dont la variabilité n'influe pas la variable de sortie. […] ».*[90]

Pour mesurer la sensibilité du modèle, nous nous proposons de faire appel à deux outils complémentaires l'un de l'autre : **l'élasticité**, indicateur très répandu en science économique, particulièrement en micro-économie, et la **méthode Monte-Carlo**, qui utilise des procédés aléatoires pour estimer la dispersion de la variable de sortie (la VCF) suite à une variabilité subie des facteurs en entrée.

Dans le 1er cas, nous allons proposer **deux chocs**, un positif (ex : +1%) et un négatif (ex : -1%) sur les 17 inputs en entrée pour mesurer la sensibilité au modèle de chaque hypothèse.

Dans le 2nd, nous automatiserons l'intégralité du modèle afin de réaliser une **simulation Monte-Carlo** : 1000 tirages aléatoires de chaque input entre une borne « Min » et une borne « Max ». Nous obtiendrons alors pour chacune des années jusqu'à l'horizon, des estimations de la tendance centrale et de l'intervalle de confiance associé. En effet, la 25ème et la 975ème valeur, respectivement $Q_{2,5}$ et $Q_{97,5}$ pourront être assimilées aux bornes de l'intervalle à 95% construit empiriquement.

Liste des paramètres à « *shocker* » :

| N° | Paramètres | Valeur Centrale | *"Shock down"* | *"Shock up"* | Borne min | Borne Max |
|---|---|---|---|---|---|---|
| 1 | Horizon | 20 ans | - 1 an | + 1 an | | |
| 2 | Facteur de crédibilité | z | - 1 pt | + 1 pt | z-10% | z+10% |
| 3 | PsurC_2017 | P/C | - 1 pt | + 1 pt | P/C-10% | P/C+10% |
| 4 | Score churn | sc_chu | - 1 pt | + 1 pt | sc_chu-3% | sc_chu+3% |
| 5 | Score Pré-contentieux | sc_pre | - 1 pt | + 1 pt | sc_pre-3% | sc_pre+3% |
| 6 | Appétence Helpeo | sc_hel | - 1 pt | + 1 pt | 0% | sc_hel+2% |
| 7 | Appétence Kimono | sc_kim | - 1 pt | + 1 pt | 0% | sc_kim+2% |
| 8 | Appétence MCT | sc_mct | - 1 pt | + 1 pt | 0% | sc_mct+2% |
| 9 | Appétence Poseo | sc_pos | - 1 pt | + 1 pt | 0% | sc_pos+2% |
| 10 | Taux actualisation | tx_actua | - 1 pt | + 1 pt | Courbe EIOPA *Down & Up* | |
| 11 | Taux de mortalité | tx_mortalite | - 1 pt | + 1 pt | tx_mortalite-2% | tx_mortalite+2% |
| 12 | Age | age | - 1 an | + 1 an | max(age-1,0) | age+1 |
| 13 | Ancienneté | anci | - 1 an | + 1 an | max(anci-1,0) | anci+1 |
| 14 | Frais de gestion | tx_FG = 8% | - 1 pt | + 1 pt | 6% | 10% |
| 15 | Coût d'acquisition | 1 050 € / 50 € | - 1% | + 1% | 950 € / 0 € | 1 200 € / 100 € |
| 16 | Evol. Cotisations santé | 4% | - 1 pt | + 1 pt | 0% | 8% |
| 17 | Evol. Cotisations prev. | 1% | - 1 pt | + 1 pt | -3% | 5% |

[90] Début du résumé de « Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique » par Julien Jacques, Université Lille 1.

Une autre façon de procéder, probablement plus adéquate pour mesurer l'impact d'une variation de taux, consisterait à calculer la dérivée d'une fonction dépendante du taux $r$ examiné. Considérons par exemple : $\mathrm{VCF}(r)$, la valeur client future fonction de $r$.

Sa dérivée est par définition $\mathrm{VCF'(r)} = \frac{\mathrm{VCF}(r)}{dr}$. La sensibilité $\mathrm{S}(r)$ s'obtient donc par la relation $\mathrm{S}(r) = -\frac{\mathrm{VCF'(r)}}{\mathrm{VCF(r)}}$. Le problème est que cela suppose de connaître précisément la relation liant VCF à $r$, ce qui n'est pas le cas à ce stade. Mais on pourrait essayer de l'estimer.

### 6.3.1. Elasticité

#### *6.3.1.1. Définition*

L'élasticité mesure la variation relative d'une grandeur provoquée par la variation relative d'une autre grandeur. Dans notre cas, nous cherchons à évaluer l'élasticité de la valeur client par rapport à un paramètre X, c'est-à-dire la variation relative de la valeur client provoquée par une variation relative du paramètre X. Ce qui s'écrit :

$$e(VCF, X) = \frac{\frac{\Delta VCF}{VCF}}{\frac{\Delta X}{X}} = \frac{\Delta VCF}{\Delta X} \times \frac{X}{VCF}$$

On en déduit la variation relative de la VCF :

$$\frac{\Delta VCF}{VCF} = e(VCF, X) \times \frac{\Delta X}{X}$$

Ainsi définie, l'élasticité s'apparente à une tentative pour mesurer la sensibilité du modèle. En effet, ce calcul permet de répondre à la question : Comment se comporte la variable endogène (la VCF) lorsque je fais subir une petite variation à un des paramètres en entrée ? En inversant le problème, elle peut aussi servir à répondre à la question : De combien dois-je faire varier le paramètre X afin d'obtenir une variation de l'endogène de y% ?

En pratique :

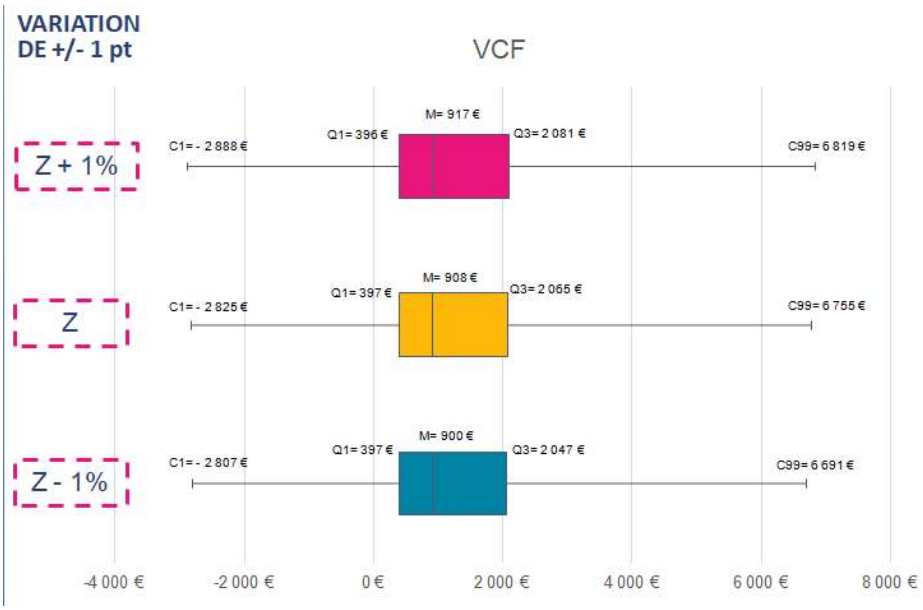
- Comme la VCP est par construction totalement déterminée, mesurer l'élasticité de la VCT ne s'avèrerait pas très pertinent. Mieux vaut la mesurer uniquement sur la partie qui est estimée ou indéterminée avec certitude, c'est-à-dire la VCF.
- L'élasticité se calcule sur la VCF moyenne du portefeuille.
- On utilise des faibles variations des paramètres X, le plus souvent ± 1% (100 pts de base).
- Préalablement au calcul de l'élasticité, nous réalisons des représentations graphiques de la variation de VCF provoquées par les « *shocks* » « *up* » et « *down* », sous forme de boîtes à moustaches.

Interprétation des résultats possibles :

- Si l'élasticité est nulle, $e(VCF, X) = 0$, une variation du paramètre X n'a aucune incidence sur la valeur client future.
- Si l'élasticité est positive, $e(VCF, X) > 0$, le paramètre X et la valeur client future évoluent dans le même sens.
- Si l'élasticité est négative, $e(VCF, X) < 0$, le paramètre X et la valeur client future évoluent en sens contraire.

#### *6.3.1.2. Principaux résultats*

**Facteur de crédibilité**

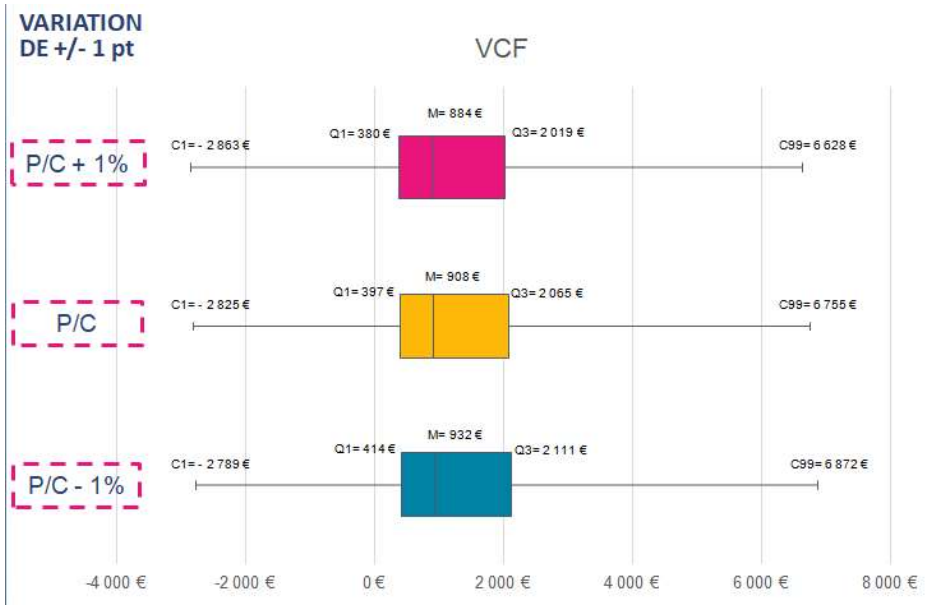



$$\text{e}\ (\overline{VCF}, \text{Z}) = [(1328\text{-}1319)/1319] \ / \ (0.01/0.595) = 0.42$$

$$\frac{\Delta VCF}{VCF} = 0.42 \times (0.01/0.595) = 0.007$$

L'impact sur la VCF d'une légère variation du facteur de crédibilité est assez faible. Une augmentation du facteur de crédibilité de 1 point entraîne une hausse moyenne de la VCF de 0,7%.

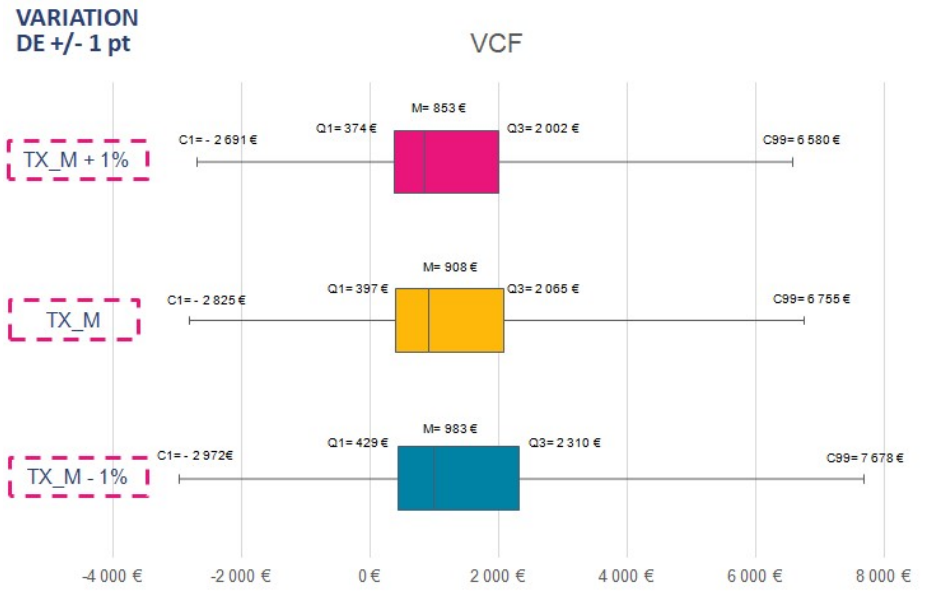
**P/C**




$$e\ (\overline{VCF}, \text{P/C}) = [(1284\text{-}1319)/1319] \ / \ (0.01/0.698) = \text{-}1.85$$

L'impact sur la VCF d'une légère variation du P/C est très fort. Une amélioration du P/C de 1 point entraîne une hausse moyenne de la VCF de 2,6%.

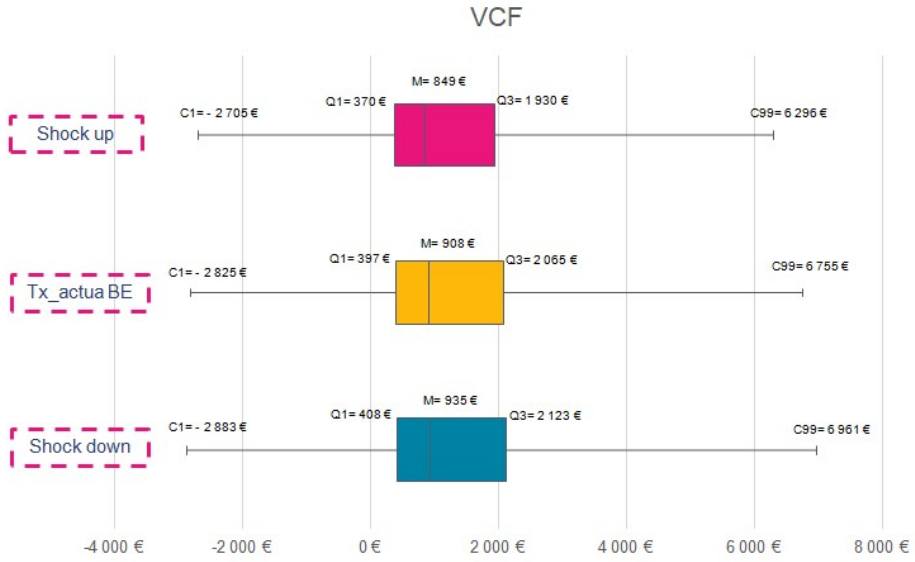
**Taux de mortalité**




$$e\ (\overline{VCF}, \text{Tx mortalité}) = [(1273\text{-}1319)/1319] \ / \ (0.01/0.026) = \text{-}0.09$$

Une baisse du taux de mortalité de 0,1 point entraîne une hausse moyenne de la VCF de 0,35 %.

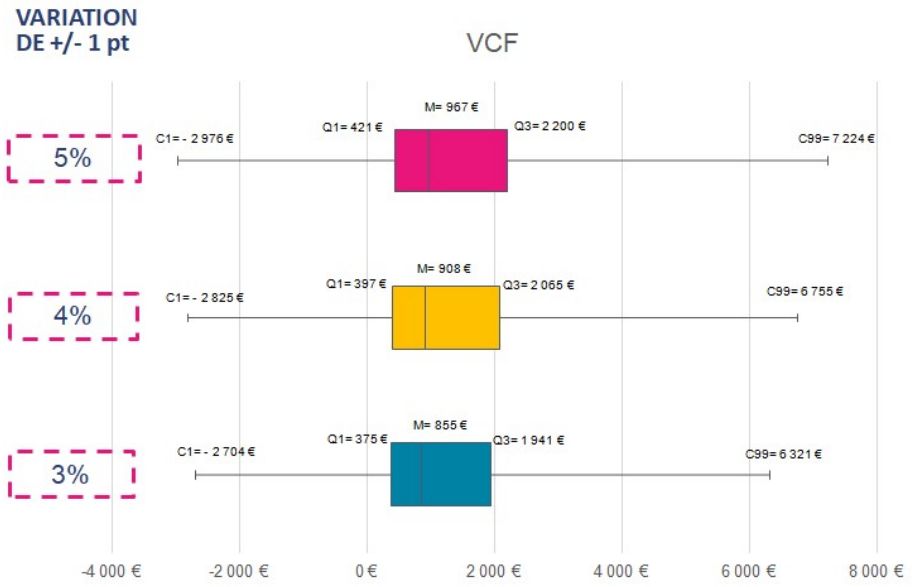
**Taux d'actualisation**




e ($\overline{VCF}$, Tx actualisation) = [(1227-1319)/1319] / (*shock up/ best estimate*) = -0.08

Un *shock up* sur le taux d'actualisation entraîne une baisse moyenne de la VCF de 6,9 %.

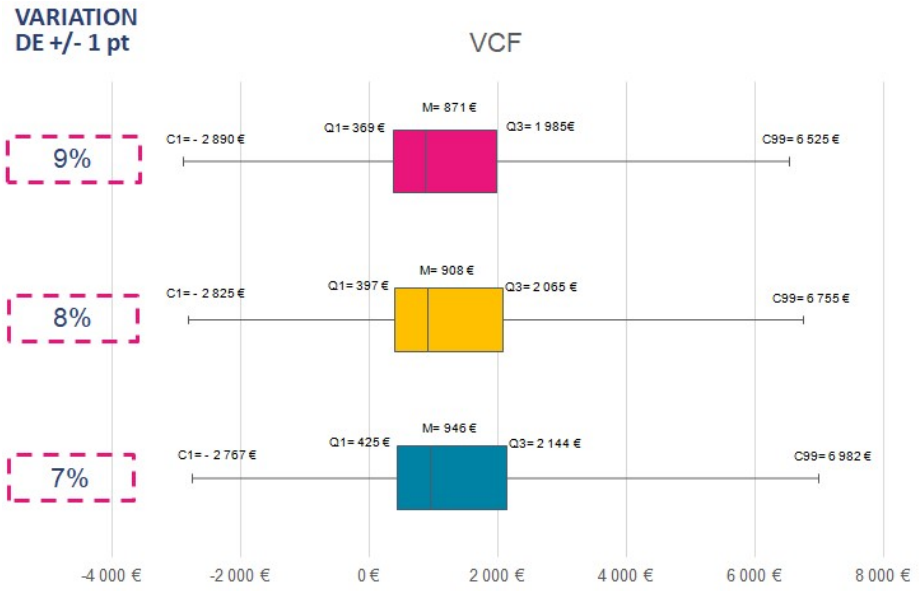
**Evolution des cotisations santé**




e ($\overline{VCF}$, Evol. Cotisations santé) = [(1410-1319) /1319] / (0.01/0.04) = 0.26

L'élasticité est assez faible mais revaloriser chaque année les primes de 5% au lieu de 4% entraînerait une hausse moyenne de la VCF de 6,4%.

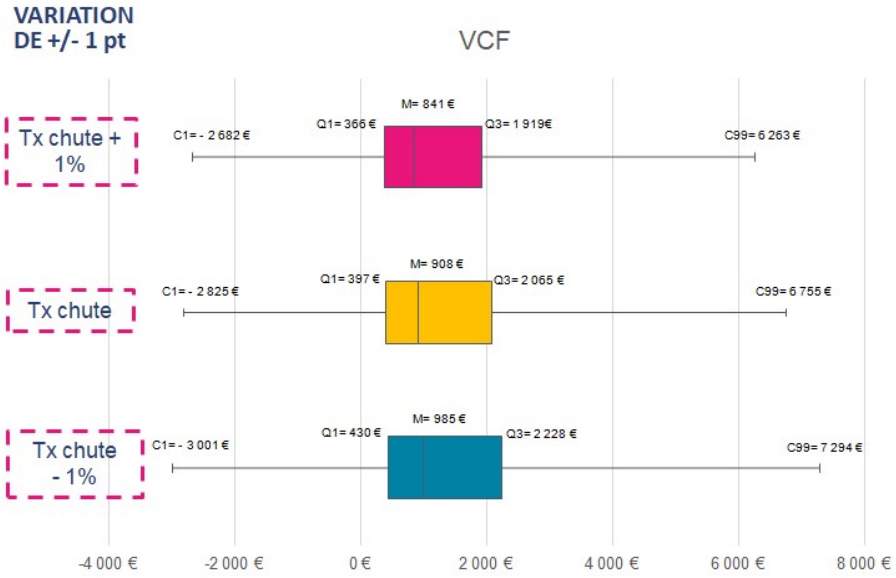
**Taux de frais de gestion**




$$e\ (\overline{VCF}, \text{Evol. Cotisations santé}) = [(1258\text{-}1319)\ /1319] \ /\ (0.01/0.08) = \text{-}0.39$$

Comme pour l'évolution des cotisations, l'élasticité apparaît assez faible mais tout de même gagner un point sur les de frais de gestion permettrait de gagner 4,8% sur la VCF moyenne.

**Taux de chute**




$$e\ (\overline{VCF}, \text{Evol. Cotisations santé}) = [(1220\text{-}1319)\ /1319] \ /\ (0.01/0.088) = \text{-}0.66$$

$$\frac{\Delta VCF}{VCF} = -0.66 \times (-0.01/0.088) = 0.075$$

L'élasticité Taux de chute – Valeur Client Future est assez forte. Réduire de 1 point les démissions volontaires, permettrait de générer une hausse de la VCF de 7,5%, ce qui est considérable.

**Bilan sur les paramètres principaux :**


| Paramètre | Elasticité |
|---|---|
| P/C | -1,85 |
| Taux de churn | -0,66 |
| Facteur de crédibilité | 0,42 |
| Taux de frais de gestion | -0,39 |
| Evol. des cotisations | 0,26 |
| Taux de mortalité | -0,09 |
| Taux d'actualisation | -0,08 |

### 6.3.2. Simulation façon « Monte-Carlo »

#### *6.3.2.1. Définition du plan*

Habituellement, les méthodes de simulation Monte-Carlo sont des algorithmes de calcul qui consistent à générer des échantillons aléatoires, à effectuer des calculs sur chacun de ces échantillons, puis à agréger les résultats obtenus.[91]

D'un point de vue pratique, le *bootstrap* non-paramétrique est utilisé pour produire des perturbations dans les données et mesurer la stabilité des estimateurs à l'égard de ces perturbations.

Dans notre cas précis, ce ne sont pas les individus que nous allons tirer au hasard, mais des valeurs de paramètres situées dans un intervalle prédéfini. La méthode n'est donc pas stricto sensu une *« Monte–Carlo »* mais elle est « inspirée » par la simulation *« Monte–Carlo »*. Parallèlement, les 1 000 échantillons obtenus ne sont pas réellement des purs échantillons *bootstrap*. En effet, *in fine*, après le tirage aléatoire, nous aurons 1 000 échantillons composés de 16 paramètres[92] dont la valeur aura été tirée aléatoirement entre une borne « Min » et une borne « Max ». Pour chacun de ces échantillons, la population étudiée ne varie pas : il s'agit de l'ensemble des personnes couvertes par un contrat santé interpro à fin de mois précédent l'exécution du modèle. En conséquence, même si on a un peu tordu le procédé initial, nous pouvons aisément convenir que le principe du *bootstrap* demeure.

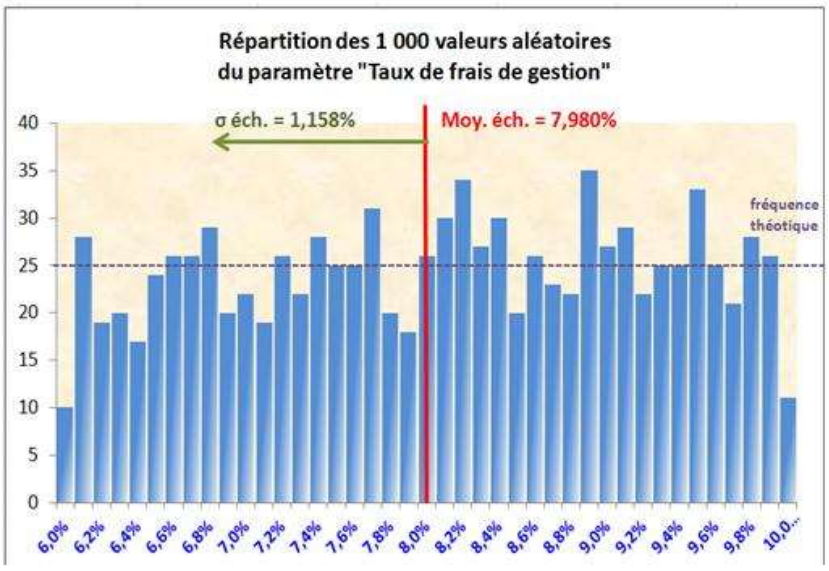
[91] C'est la méthode *bootstrap*, déjà vu en 4.3.7. Random Forest

[92] 16 et pas 17 car dans ce plan, l'horizon ne change pas. Nous restons sur une durée d'expérience de 20 ans.

*6.3.2.2. Résultats*

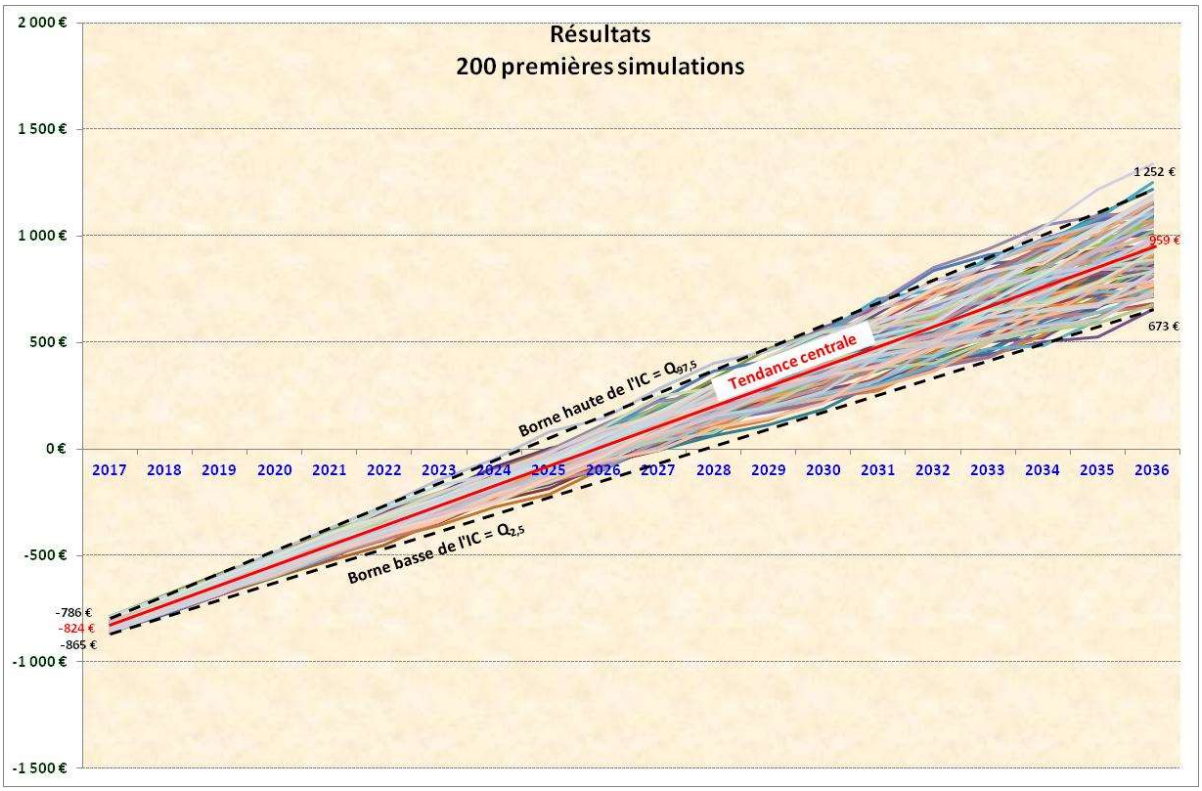
Examinons dans un premier temps la distribution des paramètres pour un tirage aléatoire entre bornes [a ; b] : logiquement, ils suivent tous une loi uniforme centrée sur la moyenne de l'échantillon, qui est très proche de la valeur initiale du paramètre, et de variance $\frac{(b-a)^2}{12}$, soit un écart-type $\frac{(b-a)}{2\sqrt{3}}$.

Un exemple : les taux de frais de gestion. La valeur initiale est de 8%. Le tirage aléatoire s'effectue dans l'intervalle [6% ; 10%], l'écart type attendu est de $\frac{(10\%-6\%)}{2\sqrt{3}} \approx 1{,}155\%$. En arrondissant les valeurs trouvées à 0,1%, voici une représentation graphique :




Le résultat obtenu est très satisfaisant. Une rapide vérification montre que c'est le cas pour l'ensemble des paramètres testés.

Simulation sur 1 000 échantillons de 16 paramètres tirés au sort entre leurs bornes Min et Max :




A horizon 2036, la moyenne de valeur client future[93] sur l'ensemble de nos clients en parc à fin mars 2017, est proche de **960 €**. L'issue de la simulation indique que cette tendance centrale, selon les propriétés de notre plan, pourrait varier dans une fourchette de ± 30%, en excluant les 5% des cas extrêmes. En effet, nous obtenons $\mathbf{Q_{2,5} = 673\ €}$ et $\mathbf{Q_{97,5} = 1\ 252\ €}$.

$$\overline{\boldsymbol{VCF}} = \mathbf{959\ €}\ \boldsymbol{et}\ \boldsymbol{IC_{95\%}(VCF)} = [\mathbf{673\ €}\,;\mathbf{1\ 252\ €}]$$

En toute honnêteté, je ne suis pas totalement satisfait de cette simulation, il faut la retravailler.

[93] Rappel, arbitrairement pour cette partie, il s'agit de VCF – coût d'acquisition

## 6.4. Segmentation

Pour la majorité des cas d'usage du modèle de valeur client, nous ne souhaitons pas adresser nos clients en fonction de leur valeur client propre. D'une part, d'un point de vue business, une telle pratique ne serait pas nécessairement pertinente. D'autre part, dans certains cas, une approche trop individualisée ne serait pas GDPR & CNIL compatible. Enfin, l'éthique et les codes de déontologie, notamment celui de l'Institut des Actuaires[94], nous protègent contre des procédés qui peuvent paraître parfois, au mieux, hasardeux.

**Une segmentation de notre portefeuille client est donc nécessaire.** Celle-ci est logiquement fondée sur les résultats du modèle de valeur client, mais elle doit être en outre, inévitablement confrontée aux connaissances « métiers » et en particulier celles de l'équipe 'Marketing Stratégique' qui valide la pertinence d'une telle approche.

Nous procédons en premier lieu à une analyse factorielle dans le but d'obtenir un résumé des dimensions qui constituent la base d'études en sortie, c'est-à-dire l'ensemble des outputs du modèle de VC (les 4 valeurs clients obtenus + les scores) ainsi que les principales dimensions qui caractérisent les individus. Puis dans un second à une classification, c'est-à-dire une partition des clients fondée sur les axes factoriels de l'AFCM, visant à constituer des groupes d'individus les plus homogènes possibles, et les plus hétérogènes possibles vis-à-vis des autres groupes. Enfin, nous donnons des noms aux groupes retenus afin de les rendre les plus opérationnels possibles pour les équipes Marketing & Développement.

### 6.4.1. Analyse Factorielle des Correspondances Multiples[95]

Les analyses factorielles, ACP, AFC, AFCM, sont des méthodes de traitement d'un jeu de données (individus **x** dimensions) qui consiste à transformer les variables d'origine pour obtenir de nouvelles variables décorrelées les unes des autres. Elles peuvent être mises en œuvre pour :

- Décrire et visualiser de manière très synthétique un grand jeu de données
- Résumer l'information en éliminant le « bruit statistique », et en ne retenant que les effets principaux
- Décorréler un ensemble de variables
- Hiérarchiser l'information présente dans un grand tableau

L'AFCM est une généralisation de l'AFC. C'est une méthode très commode car elle permet en discrétisant les variables continues, de mixer des données quantitatives et qualitatives. En pratique, elle consiste à réaliser une AFC sur un tableau disjonctif complet, c'est-à-dire, un tableau (individus **x** modalités des dimensions) ne contenant que des éléments 1 (l'individu est associé à cette modalité) ou 0 (l'individu n'est pas associé à cette modalité).

[94] http://www.institutdesactuaires.com/global/gene/link.php?doc_id=138&fg=1
[95] https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-afcm.pdf

Préalablement, pour qu'une variable ne concentre une trop grande inertie à elle seule, on ne doit pas oublier d'équilibrer le découpage des variables en modalités, et d'éviter les modalités à faibles effectifs, en les regroupant.

Processus :

- Export de la table finale enrichie de toutes les dimensions utiles à l'analyse et tous les indicateurs calculés ou estimés par le modèle
- Import sur SPAD et utilisation des procédures adaptées
- Interprétation et description des axes factoriels

Les résultats bruts de l'AFCM n'apportent pas d'information opérationnelle significative. On s'en sert uniquement comme phase exploratoire et étape intermédiaire pour la suite. En effet, ces résultats permettent de représenter le nuage de points dans l'espace en le résumant à des combinaisons linéaires de variables décrivant au mieux l'information d'origine. Seules sont conservées les coordonnées des individus projetés sur les axes factoriels.

### 6.4.2. Classification mixte

Il existe principalement deux approches de classification non supervisée[96] : la Classification Ascendante Hiérarchique (CAH) et les agrégations autour des centres mobiles (ou K-means). On n'entre pas davantage dans le détail. La page *Wikistat* référencée est bien construite et très pédagogique. Les deux méthodes ont des avantages et inconvénients. Il existe une solution qui permet de bénéficier des avantages des deux systèmes : la classification mixte.

Principe : à partir des coordonnées sur les axes factoriels, produire un premier niveau de regroupements par des nuées dynamiques en utilisant un grand nombre de centres initiaux, puis utiliser l'algorithme de CAH sur les classes obtenues à l'étape précédente.

Processus :

- Classification mixte (Procédure SEMIS) sur coordonnées factorielles de l'AFCM
- Recherche automatique des 3 meilleures partitions
- Sélection de la meilleure partition au sens « métier »
- Contrôle stabilité des groupes sur plusieurs années projetées.

### 6.4.3. *Naming* groupes de la partition retenue

Processus :

- Qualification de chacun des groupes retenus

[96] http://wikistat.fr/pdf/st-m-explo-classif.pdf

- Vérification de la cohérence « métier »
- Contrôle par rapport aux politiques et stratégies définies
- Association des groupes avec pictogrammes / smileys / jauges / recommandations / … pour utilisation par les équipes commerciales.
- Comparaison de la segmentation statistique retenue avec une segmentation « intuitive » de type « RFM[97] », très utilisée en Marketing

Il y a au moins deux niveaux de lecture pour une partition donnée : niveau macro, par exemple A, B, C, etc., et un niveau plus fin (A1/A2/A3/B1/B2 …). En effet, il ne suffit pas de savoir que le client λ appartienne au groupe A – clients à forte valeur ; il faut aussi savoir pour quelles raisons. C'est pourquoi nous affinons la 1ère partition retenue, avec par exemple :

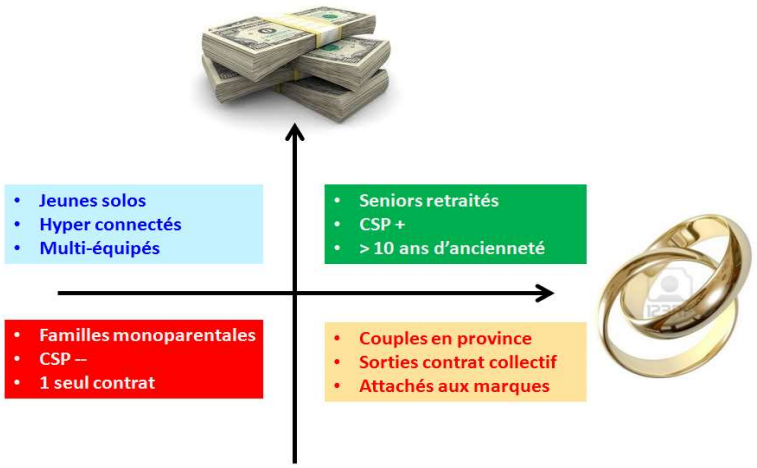
A1 – clients à forte valeur ; contribution forte à la marge
A2 – clients à forte valeur ; fidélité forte à la marque
A3 – clients à forte valeur ; gros potentiel de multi-équipement
Etc.

En toute transparence, cette partie n'est, à ce jour, pas totalement aboutie. C'est pourquoi nous nous contentons de donner des pistes de noms ainsi qu'un exemple trivial fondé sur des données fictives.

Les groupes potentiels : Ambassadeurs / Fortes valeurs / Clients occasionnels / Faibles valeurs / Inactifs / Opportunistes / Prescripteurs / …

### 6.4.4. Un exemple trivial

*Exemple fondé sur des données fictives*




[97] « RFM » pour « Récence, Fréquence, Montants » à compléter par la méthode « NBA » (Next Best Action).

## 6.5. Conclusion partie VI

- **La régression logistique est la plus performante parmi les méthodes traditionnelles. Et elle ne sous-performe pas tellement par rapport aux nouveaux algorithmes de type bagging et boosting.**

- **C'est la méthode qui est retenue pour l'ensemble du *Scoring* car elle est plus facile à mettre en œuvre sur la version de SAS dont je dispose.**

- **Le passage du modèle de valeur client sur Python remet en cause cette conclusion, et devrait nous conduire à choisir un algorithme d'agrégation, éventuellement une méthode non présentée dans la partie IV.**

- **D'une manière générale, les premiers résultats viennent confirmer nos intuitions de départ. Comme attendu, l'âge et l'ancienneté sont deux des principaux *drivers* de la valeur client.**

- **En neutralisant les effets majeurs de la consommation de soins (âge, niveau de garantie, ancienneté) par un MLG, les conclusions sur les autres dimensions demeurent globalement similaires.**

- **L'élasticité et la simulation façon « Monte-Carlo » sont deux bonnes méthodes pour mesurer la sensibilité du modèle à une légère variation d'un ou plusieurs des inputs.**

- **Une segmentation des clients sur la base des résultats obtenus tout au long de ce mémoire, est nécessaire pour les finalités marketing opérationnelles et pour rester CNIL et GDPR-compatible.**

# Partie 7 : Conclusion

*« Ce n'est pas parce qu'un modèle est sophistiqué qu'il est juste »*

Éric Lombard –
Directeur Général de la Cour Des Comptes

*"Cette nuit, en regardant le ciel,*
*je suis arrivé à la conclusion*
*qu'il y a beaucoup plus d'étoiles*
*qu'on en a besoin."*

Quino via Mafalda

La valeur client est un indicateur de la qualité d'un client d'un portefeuille d'assurance. Elle mesure la rentabilité et le potentiel économique de cet assuré en se fondant sur ses données socio-démographiques, actuarielles, financières et au travers de l'ensemble des points de contacts du client avec la marque, et notamment la relation client. Elle se compose d'une partie connue sans aléa, enregistrée dans les systèmes d'information de l'entreprise, et d'une partie prospective, inconnue, à estimer par des algorithmes prédictifs. Cette évaluation est modélisée et évolue dans le temps à mesure que l'on obtient des nouvelles informations. Elle est parfaitement adaptée au contexte de l'assurance non-vie en général, et aux risques Santé et Prévoyance en particulier.

Le modèle de valeur client est un puissant outil marketing qui, placé au centre des processus opérationnels de l'entreprise, permet de définir des stratégies de relation client proportionnées à la marge dégagée par chaque client ou segment de clients. Il offre ainsi une possibilité d'allouer efficacement les ressources et de mieux prioriser les efforts afin de fidéliser les meilleurs profils, de conquérir de nouveaux marchés et d'optimiser la communication et les produits.

Trois ensembles d'éléments sont absolument essentiels :

- **la contribution périodique**, marges nettes annuelles acquises passées et valeur actualisée de l'espérance des profits futurs, estimée à travers le modèle de crédibilité Bühlmann-Straub.
- **la durée de vie contractuelle** entre l'assuré et l'assureur, observée depuis la date de début du contrat, et estimée en fonction des probabilités de chute et d'appétence aux produits de multi-équipement
- **les frais**, c'est-à-dire les coûts d'acquisition, d'administration et de gestion

Le modèle développé dans le cadre du mémoire s'appuie en grande partie sur l'analyse prédictive, notamment les méthodes de machine learning. Les techniques statistiques traditionnelles de Data Mining sont confrontées aux algorithmes modernes de scoring issus de la Data Science, dans le but d'optimiser le pouvoir prédictif du modèle. Les 5 outils testés, SAS, SPAD, R, Python et Tanagra renvoient à peu de choses près les mêmes résultats, ce qui est plutôt rassurant.

Conformément à l'intuition, l'âge de l'assuré et son ancienneté dans le portefeuille, sont les deux *drivers* majeurs de la valeur client. La composition familiale, le jour d'ouverture des droits, le niveau des garanties, le canal d'entrée, le type de vente, le sexe et le lieu d'habitation sont les autres facteurs discriminants. L'ensemble de ces variables explicatives alimente une segmentation fine des clients, pertinente d'un point de vue métier et rendue nécessaire par la réglementation.

Le modèle démontre aussi que développer une activité d'assurance santé en B2C est toujours pertinente de nos jours. Malgré les contraintes réglementaires, la concurrence accrue, l'harmonisation des contrats et les coûts d'acquisition élevés, un organisme d'assurance est en mesure d'atteindre le *payback point*, c'est-à-dire le moment où l'assureur rentabilise ses investissements. Selon le modèle développé dans le cadre de ce mémoire, ce moment peut être atteint entre 9 et 14 ans selon la cohorte considérée.

L'existence de ce modèle et ses applications opérationnelles garantissent un niveau de solidarité élevé entre les assurés qui jouent le jeu de la mutualisation des risques car elles luttent indirectement contre l'opportunisme, la fraude, l'anti-sélection, le non-paiement des primes et l'abus d'utilisation des outils de la relation client.

* * *

Nous disposons donc désormais d'un modèle de valeur client qui contribue à l'obtention d'une vision 360° du client et qui offre un cadre d'analyse puissant pour tenter d'obtenir un avantage comparatif dans un environnement de plus en plus concurrentiel.

Il y a encore quelques années, le client n'existait pas. Les membres d'un organe de l'économie sociale et solidaire étaient des adhérents que l'on ne cherchait pas à distinguer les uns des autres. Les tensions apparues sur le marché ont incité les actuaires et les marketers à orienter leurs travaux selon des segments de clients. Aujourd'hui, nous entrons dans la phase ultime de la relation client. Le paradigme originel « satisfaction – fidélité – profits » ne tient plus. L'hyper personnalisation est une réponse qui est rendue possible grâce à la récolte systématique, l'analyse et l'exploitation en quasi temps réel d'une multitude d'informations concernant individuellement le client et son contexte.

En 2004, Guizouarn & Marescaux terminent leur ouvrage par le résumé suivant : *« Nous pouvons faire bouillir ce livre, pour le réduire à 3 mots : santé, segmentation et valorisation. Et à une idée : connaître le capital client est une clé de la compétitivité, de la rentabilité et de la solvabilité de l'assurance santé de demain ».* Treize ans plus tard, cette synthèse demeure on ne peut plus d'actualité.

* * *

Comme de nombreux acteurs sur le marché récemment, La Mutuelle Générale a décidé de placer le client au cœur de ses préoccupations et au centre de son *business model*. Il ne faut pas seulement l'annoncer, il faut aussi le mettre en pratique et de façon pragmatique. Le modèle décrit dans ce mémoire y contribue grandement car il permet de piloter cette nouvelle stratégie. Avec la valeur client, la profitabilité devient l'origine des processus. En intégrant la segmentation et la valorisation du capital client dans l'ensemble de l'entreprise, nous disposons de deux moyens supplémentaires pour répondre à l'objectif principal. En bénéfice collatéral, nous devrions logiquement observer une réduction considérable des coûts marketing et de gestion.

L'utilisation de la Valeur Client à La Mutuelle Générale est désormais quotidienne : Plan de fidélisation, Actions de rétention, Cocooning, Plan d'Action Commercial, Ciblage des campagnes marketing, Traitement des réclamations, Priorisation des mails entrants, toutes ces opérations embarquent dorénavant le système de valeur client. Elles existaient déjà auparavant mais aujourd'hui on priorise les clients à forte valeur potentielle. Nous n'avons absolument pas l'intention d'essayer de retenir ou conquérir les clients/prospects qui ne jouent pas le jeu de l'assurance. Nous serions alors en défaut de solidarité vis-à-vis de nos adhérents historiques.

Le modèle alimente directement les directions Marketing stratégique, Développement et Relation Client mais il peut aussi aider la Direction Technique, le Contrôle de Gestion, la Gestion Des Risques, notamment pour alimenter les modèles ALM[98] et les divers Business Plans de l'entreprise. Il remplit en ce sens, amplement sa fonction d'indicateur transverse.

[98] ALM pour Asset Liability Management, Gestion Actif Passif en français.

* * *

Il n'existe pas de modèle unique de Valeur Client. Différentes approches sont possibles. Comme dans toute modélisation, le concepteur a un rôle important notamment à travers ses partis-pris de programmation. J'assume pleinement cette responsabilité. Un autre actuaire avec une sensibilité légèrement différente pourrait obtenir des résultats modérément distincts. Quoi qu'il en soit, plusieurs principes sont invariants :

- La valeur passée ne suffit pas pour prédire les gains futurs. Un modèle prédictif construit dans les règles de l'art n'est pas une option, il est indispensable pour évaluer le potentiel probable de chaque client.
- On doit essayer de transformer un prospect en client si et seulement si sa VCT est supérieure à 0, c'est-à-dire, si sa VCF dépasse son coût d'acquisition estimé.
- La qualité de la modélisation est primordiale mais il est surtout essentiel de conserver un esprit critique sur son travail et de bien en connaître les limites.
- La qualité de la donnée est fondamentale. L'adage « *garbage in, garbage out* » est dans notre cas particulièrement approprié.

* * *

**Les limites du modèle et les pistes d'amélioration :**

1. Théoriquement, le modèle est supposé embarquer la somme actualisée des flux à l'infini. Nous avons choisi un horizon temporel de 20 ans car il fallait a minima projeter les flux assez longtemps pour se donner la possibilité d'atteindre le *payback point* qui comme on l'a vu peut aller jusqu'à 14 ans. Aller au-delà de 20 ans nous semble un peu aléatoire compte tenu du contexte de changements réglementaires fréquents et de l'environnement politique, social et économique incertain. D'autre part, au bout de 20 ans, un portefeuille interpro a perdu, en espérance mathématique, près de 80% de son stock de personnes protégées à t0.

2. La marge unitaire représente l'élément fondamental du modèle, mais ne serait-il pas trop réducteur ? Elle ne couvre sans doute pas suffisamment l'ensemble des caractéristiques de chaque client : la satisfaction, la prescription, le comportement de consommation, la prévention, les injustices de la vie (le patrimoine génétique) … Peut-être serons-nous amenés à segmenter différemment en fonction des nouvelles données que l'on va capter notamment à travers les objets connectés ?

3. Actuellement, les agrégats de frais dont nous disposons ne permettent pas de démoyenniser suffisamment. Nous effectuons des réformes aux niveaux de la comptabilité analytique et du contrôle de gestion spécifiquement mais ce travail de longue haleine n'aboutit pas à des résultats encore suffisants pour piloter les coûts dans les règles de l'art. La conséquence directe de cette vue globalisée des frais est notre incapacité à allouer de façon pertinente les frais propres aux différents groupes de clients homogènes en termes de coûts.

Notamment :
- les coûts d'acquisition ne sont pas différenciés selon le canal d'entrée.
- les frais de gestion des sinistres sont identiques pour tous quels que soient la fréquence et la nature des sinistres (Tiers Payant *vs* Hors Tiers Payant, Hospitalisation *vs* Pharmacie, etc.)
- les frais de gestion de la relation client appliqués sont indépendants du nombre d'appels téléphoniques, de réclamations ou de visites en agence.

De ce fait, nous surestimons la valeur des clients qui proviennent de leads internet (payés chers), pour lesquels les forces commerciales ont investi beaucoup de temps, qui ont beaucoup de sinistres, spécialement HTP, ou ceux qui abusent des outils de relation adhérent. Symétriquement, à l'inverse, nous sous-estimons la valeur des clients qui sont venus d'eux-mêmes à LMG, avec 0 sinistre, ou qui ne contacte jamais l'assureur.

Si les frais réels ou a minima estimés proches de la réalité étaient appliqués, la combinaison de plusieurs de ces éléments, pourrait générer des écarts importants avec le modèle actuel. Pour pallier à cet effet, la première initiative, en cours de construction, consiste à créer un score d'utilisation des services de relation client. Dans un avenir proche, nous réactiverons aussi l'étude des fréquence et nature des prestations, ainsi que la construction d'hypothèses de variabilité des coûts d'acquisition selon le point d'entrée. L'idée générale s'appuiera sur le fait de conserver une moyenne pondérée des frais à $x\%$ des primes + $y$ €; les valeurs unitaires se distribuant dans l'intervalle [4% ; 100%] des cotisations.

4. La corrélation des risques Santé et Prévoyance aurait probablement mérité un traitement approfondi. Le problème est qu'il n'est pas aisé de dire en théorie, si intrinsèquement le fait de posséder un produit prévoyance génère davantage de sinistres en santé du fait de l'anti-sélection notamment (des sinistres en prévoyance engendrent des sinistres en santé, corrélation positive) ou non (logique de prévention, corrélation négative). Il aurait alors fallu s'appuyer sur une étude très poussée des prestations Santé et Prévoyance. Ça n'est pas possible sur le périmètre interprofessionnel : l'historique est trop court et il n'y a pas suffisamment de sinistres enregistrés. Une telle étude serait davantage pertinente sur les assurés Statutaire, le portefeuille historique, pour lesquels la Prévoyance est en inclusion, depuis de nombreuses années. Nous atteindrions alors facilement, et le volume, et la masse critiques.

5. La valeur client c'est aussi un chemin vers l'évaluation de la valeur globale de l'entreprise. En effet, en quelque sorte l'*Embedded Value*[99] peut s'apparenter à la valeur économique du portefeuille de contrats existants, c'est-à-dire pratiquement ce qu'essaye de représenter la somme de la VC sur l'ensemble des portefeuilles. Comprendre comment l'un peut conduire pleinement à l'autre est une piste à creuser sérieusement.

6. Le modèle à date est encore largement perfectible. Nous allons continuer à explorer et tester continuellement de nouvelles méthodes de Data Science. Rien n'est jamais gravé dans

[99] L'Embedded Value correspond à la valeur intrinsèque d'une compagnie d'assurance sans prise en compte d'un Goodwill qui représente la différence entre l'actif net du bilan d'une entreprise et sa valeur de marché.

le marbre. Nous allons aussi davantage enrichir le modèle par l'ajout de données, notamment celles issues de l'Open Data. La GDPR, le *self care*, le développement de la prévention et des services sont des leviers, des opportunités qui vont eux aussi générer des nouvelles données que nous allons intégrer au fur et à mesure. Plus particulièrement, l'apport du digital est fondamental. Nous aurons accès à des données plus larges et en plus grande quantité qui complèteront les données classiques récoltées pour la tarification. Enfin, nous allons creuser encore davantage via du *feature engineering,* c'est-à-dire retravailler et recombiner les variables que l'on a déjà pour en générer des nouvelles. A nous de tirer le profit maximal de ces nouvelles informations.

7. D'une manière générale, nous allons continuer à optimiser la performance du modèle et à progresser nous-mêmes. Notre cible, c'est la projection des flux, l'analyse des sentiments du client et le calcul de la valeur client **en temps réel** ; tout ceci avec audace mais sans gourmandise et sans excès de confiance … en adoptant la politique des petits pas.

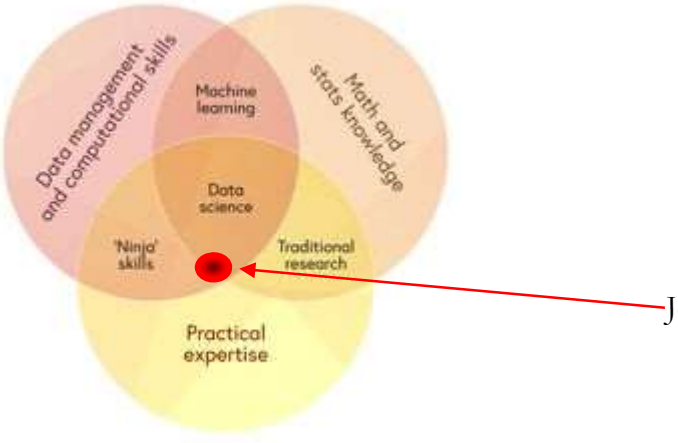
* * *

J'ai, successivement dans ma carrière, occupé les postes de « data manager », « data analyst », « chargé d'études statistiques », « statisticien », « chargé d'études actuarielles », « data miner », et enfin « responsable de la stratégie data ». Aujourd'hui on me demande parfois quelles sont les différences principales entre ces fonctions. Je réponds le plus souvent que tous ces métiers ont beaucoup plus de points communs que de dissemblances. Je trouve que plutôt qu'épiloguer sur les nuances, il est plus aisé d'expliquer que nous autres les professionnels du traitement de la donnée, nous formons avant tout la famille des « Experts Data ». C'est plus simple et c'est très clair.

On me demande aussi si je me considère aujourd'hui davantage « actuaire » ou « data scientist ». Je réponds, certes avec un peu de malice, qu'on me pardonne, et essentiellement pour éluder la question : « Ni l'un, ni l'autre, et les deux à la fois. *First of all, I'm a* …


»

Par ailleurs, j'aime beaucoup le diagramme de Venn suivant auquel j'adhère totalement :




Avec *« Practical expertise »* = « Pratique de la science actuarielle » pour notre part.

Je crois beaucoup aux polycompétences. Je suis aussi absolument convaincu que nous devons continuer à nous former tout au long de la vie et à développer ces polycompétences. Je n'ai aucun problème par rapport au fait que mon métier ne puisse être clairement défini. Actuaire + Data Scientist, cela me convient très bien. C'est tout de même la combinaison du « *best job* 2015 » selon l'agence de Ressources Humaines Careercast[100] et du « job le plus sexy du 21ème siècle » d'après la Harvard Business Review[101] d'Octobre 2012. Même si cela reste réducteur malgré tout.

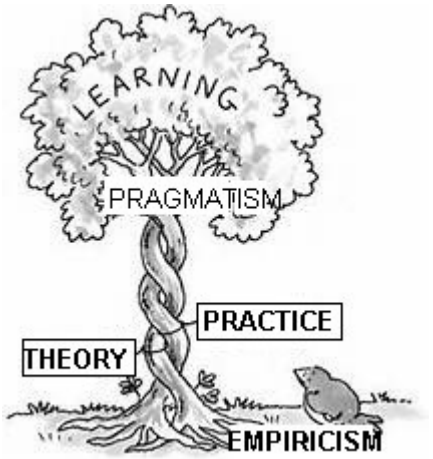
Pour autant, aucune raison de s'arrêter en si bon chemin. Il existe encore une multitude de savoirs et savoir-faire différents à acquérir pour dominer l'ensemble de la chaîne de valeur assurantielle. Rien qu'au niveau de la « donnée », des nouveaux domaines apparaissent qui méritent de s'y intéresser. Une liste non exhaustive :

Data Visualisation / Data Discovery / Marketing Automation / Data Monétisation / Stratégie Data / Gouvernance Data / Data Market Place / Data Agrégation / …

Ce mémoire marque aussi la fin d'un cycle personnel et professionnel. J'assure depuis peu au sein de mon entreprise, la fonction de responsable de la Stratégie Data. A ce titre, j'ai pour double mission de garantir la transversalité des projets Data et de contribuer à impulser une culture *data driven* à La Mutuelle Générale, un programme particulièrement excitant. Mon rôle change sensiblement. Il m'impose davantage de manager les profils « experts Data », accompagner les nouveaux talents, superviser les projets, et incarner une posture de transmission des savoirs. C'est une fonction que j'assume avec plaisir et dévouement.

Que ce soit pour la démarche valeur client ou à propos de ma vie professionnelle, j'essaye de conserver la même ligne de conduite : approche par petit pas, à la recherche de l'efficience, avec agilité, lucidité et pragmatisme…

[100] http://www.careercast.com/jobs-rated/best-jobs-2015
[101] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century


LEARNING
PRAGMATISM
PRACTICE
THEORY
EMPIRICISM

# Bibliographie

## Bibliographie principale

### Sites internet « Stats »

Dans cette première double liste, les sites qui m'ont directement inspiré pour le mémoire. Les autres sites consultés ponctuellement sont en bibliographie secondaire.

1. **Philippe BESSE,** Université Toulouse III Paul Sabatier **/ WIKISTAT** : www.wikistat.fr

Un cours complet sur la science des données. Synthétique mais dense et pédagogique. Régulièrement dans les premiers liens lors de recherches sur Google. Et c'est bien mérité.

2. **Ricco RAKOTOMALALA**, Laboratoire Eric, Université Lumière Lyon 2 : https://eric.univ-lyon2.fr/~ricco/cours/index.html

Mon professeur de Data Mining en Master qui m'a profondément marqué et qui a très nettement orienté mes choix professionnels. Ricco met à disposition avec passion, tout son savoir à travers des cours très pédagogiques. C'est un régal de le suivre dans les méandres de la data science.

3. **Stéphane TUFFERY, Université** Rennes 1 : data.mining.free.fr http://blogperso.univ-rennes1.fr/stephane.tuffery/

Les compléments indispensables à la « bible » mentionnée plus bas.

4. **Pierre THEROND** : http://www.therond.fr/wp-content/uploads/cours/Credibilite.pdf

Cours ISFA sur la théorie de la crédibilité. Le seul détaillé et consistant, disponible en langue française.

5. **SAS, aide en ligne** : https://support.sas.com/en/support-home.html

Costaud, vaste et parfaitement complet. En cherchant bien, on trouve toujours comment se sortir d'un problème simple ou complexe.

6. **COURSERA :** https://fr.coursera.org/

Le site qui centralise les meilleurs MOOC data science. Pistez notamment l'introduction à la Data Science de l'université John Hopkins[102] et celui sur Python de l'université du Michigan[103]. Fondamentaux et Précieux.

[102] https://fr.coursera.org/specializations/jhu-data-science
[103] https://fr.coursera.org/learn/python

## Sites internet « Institutionnels »

1. **Wikipedia** : https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

« L'encyclopédie libre », très décriée mais ne jouons pas les hypocrites : tout le monde s'en sert ! Et cela reste quoi qu'on en dise, un excellent exemple de travail collaboratif.

2. **Institut des actuaires** : http://www.institutdesactuaires.com

3. **Ressources actuarielles :** http://www.ressources-actuarielles.net/

Des articles captivants écrits par les experts les plus éminents. Et la liste des mémoires actuariels, indispensable pour la recherche collective pratique et fondamentale.

4. **Argus de l'assurance :** http://www.argusdelassurance.com/

5. **EIOPA** https://eiopa.europa.eu/

6. **Banque de France :** https://www.banque-france.fr/

On y trouve notamment des informations précieuses sur les principaux indicateurs statistiques français et étrangers.

7. **GDPR :** http://www.eugdpr.org/

8. **DDA :** http://www.argusdelassurance.com/directive/directive-sur-la-distribution-d-assurance-dda/

9. **INSEE** : https://www.insee.fr/fr/accueil

10. **Open data** : http://www.data.gouv.fr/fr/

Partage et utilisation des données publiques. Déjà un must.

11. **Vernimmen** : https://www.vernimmen.net

Un site de référence pour les professionnels et étudiants en Finance

## Livres

Une liste volontairement limitée. J'y ai inscrit uniquement les ouvrages qui m'ont accompagné pleinement sur toute la période de conception du modèle et de rédaction du mémoire. Les autres références importantes sont dans la seconde liste.

1. **Alain TOSETTI,** *Assurance : Comptabilité, Réglementation, Actuariat,* ECONOMICA, 2011.

Plus ou moins le programme Assurance de 1ère année du CEA. Une excellente introduction à l'actuariat que je recommande systématiquement pour qui souhaite développer sa culture de la science actuarielle. Je m'y réfère régulièrement pour vérifier une formule ou revoir une notion.

Chapitre4, Annexe3, Crédibilité et tarification, page 133 : une introduction au modèle de Bühlmann. Complétée et détaillée en cours CEA par Thomas Béhar.

2. **Jean-Charles GUIZOUARN et Nicolas MARESCAUX,** *Assurance Santé : Segmentation et Compétitivité,* ECONOMICA, 2004.

Un ouvrage généraliste sur l'assurance santé très complet. Très agréable à lire. Un peu daté probablement mais l'essentiel des sujets couverts est encore d'actualité aujourd'hui.

**Voici les 4 ouvrages de référence** qui ne devraient pas, selon moi, quitter le bureau d'un statisticien. Je n'ose pas écrire « data scientist » car ceux-ci se servent majoritairement sur le net ☺ :

3. **Stéphane TUFFERY**, *Data Mining et statistique décisionnelle,* 4ème édition, Editions TECHNIP, 2012.

Indispensable pour tout data miner. C'est une véritable bible : plus de 800 pages dédiées à la science des données. Je m'y réfère systématiquement pour améliorer mes connaissances sur les sujets que je ne domine pas ou pour vérifier les conditions d'utilisation d'une méthode. S'il ne devait en rester qu'un !

4. **Gilbert SAPORTA,** *Probabilités, analyse des données et statistique,* Editions Technip, 3ème édition, 2011.

Mon 1er livre de stat - l'édition de 1990 -, il y a 20 ans ! Cet ouvrage m'a ouvert les yeux et a définitivement affirmé mon intérêt pour la science statistique. Essentiel !

5. **Ludovic LEBART, Marie PIRON et Alain MORINEAU,** *Statistique exploratoire multidimensionnelle : Visualisations et inférences en fouille de données,* DUNOD, 4ème éd, 2006.

Un grand classique désormais. Le livre de référence sur les analyses factorielles « à la française ». Il traite également de nombreuses méthodes d'analyses prédictives. Fondamental !

6. **Ricco RAKOTOMALALA,** *Pratique de la Régression Logistique,* LYON II, 2011.

Le seul écrit complet en langue française totalement dédié à la régression logistique. A lire absolument pour qui veut dominer le sujet. Très détaillé et pédagogique.

## Publications

Excepté les deux dernières références, encore confidentielles en mai 2017 et que j'ai hâte de pouvoir parcourir, l'ensemble de ces travaux m'a abondamment alimenté pour la rédaction. Ce sont tous des épreuves d'une très grande qualité qu'il ne faut pas hésiter à consulter pour en connaître davantage sur leurs expertises respectives.


1. **Annie DILLARD** *« Est-il encore pertinent pour les assureurs d'investir dans la valeur client ? »,* Thèse MBA ENASS, 2014.

2. **Marie PIFFAULT** *« La valeur client comme mesure de la rentabilité d'un portefeuille d'assurance santé individuelle »,* Mémoire CEA, 2015.

3. **Ilaria DALLA POZZA & Lionel TEXIER,** AssurMarketing, *Pression tarifaire et enjeux de fidélisation soulevés par la loi Hamon : les bénéfices de la Customer Lifetime Value*, 2015.

4. **Etienne PLANTE-DUBE** *« Mesures de Valeur Client »*, Desjardins Groupe d'assurances générales, 2010.

5. **Mélanie MASSIAS,** *« La valeur client, enjeu majeur en IARD »,* L'argus de l'assurances n°7446, février 2016.

6. **Yohan WASMES BENQUE,** *« Du Big Data au Smart Data, la nouvelle ère du digital, Quel Business Model pour aider les entreprises françaises à s'internationaliser ? »*, Thèse HEC Paris, 2016.

7. **Kezhan SHI,** *« Data Science et Assurance »*, Caritat, 2016.

8. **Julien JACQUES**, *« Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique »,* Université Lille 1, 2011.

9. **Dr V. KUMAR, Ilaria DALLA POZZA, & al.,** *« Reversing the Logic: The Path to Profitability through Relationship Marketing »,* Journal of Interactive Marketing, 2009.

10. **COHERIS**, Support de formation *« Connaissance Client : Scoring, Segmentation et Recommandation »,* Coheris, 2015.

11. **DELOITTE,** *« Le risque modèle : les nouveaux challenges de l'industrie d'assurance »,* juin 2016.

12. **SOLUCOM,** *« Big Data : Une mine d'or pour l'assurance »,* octobre 2015.

13. **INSTITUT MONTAIGNE,** *« Big Data et objets connectés »,* avril 2015.

14. **WEAVE,** Livre blanc *« Intelligence Artificielle, où en sommes-nous ? »,* 2017.

15. **DATA MARKETING PARIS**, *« The Data Wikis »,* 2016.

16. **D. HENNOM,** « *Création d'un indicateur de valeur client en assurance non vie* », Mémoire CEA - AVIVA, confidentiel jusqu'au 09/11/2018.

17. **S. COUAILLAC,** « *Création d'un indicateur de valeur client en assurance non vie* », Mémoire CEA - BPCE, confidentiel jusqu'au 16/09/2017.

## Bibliographie secondaire

### Livres

Cette liste contient des ouvrages que j'ai consultés ponctuellement ou qui sont des références dans leur domaine respectif.


1. **Saporta, G.,** *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives,* Thèse de l'université Paris VI, 1975.

2. **Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J.,** *Classification and Regression Trees,* Springer, 1984.

3. **Breiman, L.,** *Bagging predictors*, Springer, 1996.

4. **Fayyad U., Piatetsky-Shapiro G., Smyth P.,** (Eds.) R. U., *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

5. **Feldblum S.,** *Asset Share Pricing for Property-Casualty Insurers*, 1996.

6. **Freund, Y., Schapire, R. E.,** *A decision-theoretic generalization of on-line learning and an application to boosting,* Journal of Computer ans System Sciences, 1997.

7. **Hosmer, D. W., Lemeshow, S.,** *Applied Logistic Regression*, Wiley, 2nd edition, 2000.

8. **Fayyad U., Grinstein G., Wierse A.,** *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, 2001.

9. **Dreyfus, G., Martinez, J.M., Samuelides, M., Gordon, M. B., Badran, F., Thiria, S., Hérault, L.,** *Réseaux de neurones : Méthodologie et applications*, Eyrolles, 2002

10. **Nakache J.P., Confais, J.,** *Statistique explicative appliquée*, Editions Technip, 2003.

11. **Bühlmann H., Gisler A.,** *A course in Credibility Theory and its Applications*, Springer, 2005.

12. **Denuit M., Charpentier A.,** *Mathématiques de l'assurance non-vie*, volume II, Economica, 2005.

13. **Dr Kumar V.,** *The power of CLV: Managing Customer Lifetime Value at IBM*, 2006.

14. **Donkers, B., Verhoef, P., de Jong, M.,** *Modeling CLV: a test of competing models in the insurance industry,* Quant Market Econ, 2006.

15. **Hastie, T., Tibshirani R., Friedman J.,** *The elements of statistical learning: data mining, inference and prediction*, Springer, 2nd edition, 2009.

16. **Lecoutre, J.P.,** *Statistique et probabilités*, Dunod, 2009.

17. **Wu, X., Dr Kumar V.,** *The top ten algorithms in data mining*, Chapman & Hall/CRC, 2009.

18. **Tufféry, S.,** *Etude de cas et statistique décisionnelle,* Editions Technip, 2009.

19. **Hastie,T., Tibshirani, R., Jerome H. Friedman, J. H.,** *The Elements of Statistical Learning,* Springer Verlag, 2nd edition, 2009.

20. **Myers, R. H., Montgomery, D. C., & al.,** *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley-Interscience, 2nd edition, 2010.

21. **Abe S.,** *Support Vector Machines for Pattern Classification,* Springer, 2010

22. **Decourt, O.,** *SAS l'essentiel : SAS v8 et SAS v9, SAS Enterprise Guide, langages SAS, SQL et macro*, Dunod, 2011.

23. **Jacques, J.,** *Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique,* Université Lille 1, 2011.

24. **Dupuis, M., Berthelé, E.,** *Big Data dans l'assurance,* Optimind Winter, 2015.

La meilleure publication lue sur le sujet. Je conseille vivement.

25. **Cardon, D.,** *A quoi rêvent les algorithmes*, Seuil, 2015.

## Sites internet

1. **Société Française de Statistique** : www.sfds.asso.fr

2. **Etalab** : https://www.etalab.gouv.fr/

Blog de la mission gouvernemental sur les données.

3. **Generali France** : http://institutionnel.generali.fr/

4. **National Bureau of Economic Research :** http://nber.org/

5. **Arthur Charpentier,** Université Rennes 1 : http://freakonometrics.hypotheses.org/

6. **Gilbert Saporta,** CNAM **:** http://cedric.cnam.fr/~saporta/

7. **CNAM Maths & Stats** : http://maths.cnam.fr/

8. **Site du logiciel R :** https://www.r-project.org/

Une institution désormais.

9. **Analyse-R :** http://larmarange.github.io/analyse-R/

Nombreux cas pratiques bien illustrés sur R & R studio

10. **Python :** https://www.python.org/

Hier le challenger, demain le leader de la data science en Open Source ?

11. **Gianluca Bontempi :** http://www.ulb.ac.be/di/map/gbonte/Welcome.html

On y trouve notamment un bon cours sur les modèles prédictifs et les méthodes d'agrégation de modèle Bagging et Boosting.

12. **OCTO** : http://blog.octo.com/category/big-data/

Nombreux article à la pointe de l'actualité de Data Science

13. **Jérémy Greze :** Jereze.com

Blog orienté Data et Analytics

14. **Kaggle** : https://www.kaggle.com/

Une plateforme d'échange sur la data science et de nombreux challenges. Le site international de référence en Data Science.

15. **Datascience.net :** https://www.datascience.net/fr/home/

Le petit frère du précédent en langue française.

16. **JY Baudot :** http://www.jybaudot.fr/

Concepts et techniques organisationnelles, descriptives, prédictives et prévisionnelles pour l'entreprise, la finance et l'économie (et fondements mathématiques).

17. **David Louapre :** https://sciencetonnante.wordpress.com/

La science côté fun !

18. **Laboratoire Mathématique et Physique Théorique** de l'université de Tours : http://www.lmpt.univ-tours.fr/

19. **Valérie Monbet,** Université Rennes 1 : https://perso.univ-rennes1.fr/valerie.monbet/enseignement.html

20. **Alp Mestan :** http://alp.developpez.com/#page_articles

21. **Emmanuel Istace** : https://istacee.wordpress.com/

22. **Institut d'électronique et d'informatique Gaspard-Monge** de l'Université Paris-Est Marne-la-Vallée **:** http://igm.u-pem.fr/