Mémoire présenté pour l'obtention du
Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires
le : 04/07/2023

Par : MEGHAZI Celia

Titre : Prise en compte de la censure dans la liquidation des provisions pour sinistres à payer de la Responsabilité Civile des professions réglementées.

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

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

Membres présents du jury de l'Institut des Actuaires :

Frédéric PLANCHET

Christian FETTIG

Entreprise : **MMA - Covéa**

Nom : **Audrey ASSOUS-BENGUIGUI**
Signature :

Directeur de mémoire en entreprise :
Nom : **Pierre GOLHEN**
Signature :

Membres présents du jury de l'ISFA :

Anne EYRAUD

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

Signature du responsable entreprise

Signature du candidat

Résumé

En France, certaines professions sont soumises à l'obligation de souscrire une assurance de responsabilité civile professionnelle. Les professions réglementées du chiffre et du droit (avocats, experts-comptables, mandataires judiciaires,…) sont soumises à cette obligation. Les enjeux élevés associés à cette garantie impliquent un taux de judiciarisation des sinistres importants. Ces éléments font de la responsabilité civile professionnelle des professions du chiffre et du droit une branche longue avec une liquidation des sinistres de plusieurs années.

Les provisions techniques représentent une partie considérable du passif de l'assureur. D'un point de vue comptable, établir le montant juste de provisions est au cœur de la dualité entre rentabilité et solvabilité. La concurrence toujours plus soutenue, avec notamment l'entrée dans le marché des bancassureurs, ainsi que les crises financières ont encouragé la mise en place de la norme Solvabilité II. La provision pour sinistres à payer (PSAP) est, en assurance de dommages et de responsabilités, la provision technique la plus significative en termes de montant. Ces provisions visent à couvrir les dépenses imputables aux règlements des sinistres, mais également liées aux frais accessoires (rémunération d'experts, frais judiciaires,…). La comptabilité technique va s'atteler au calcul de ses PSAP via une approche « *Best Estimate* » (comme cela est imposé dans Solvabilité II), cependant un Actuariat produit peut également avoir un besoin de projeter à l'ultime ces PSAP pour bien mesurer que le niveau du risque est en cohérence avec les provisions positionnées dossier par dossier. En effet, la charge sinistre nette de recours à l'ultime est définit comme la somme des règlements déjà effectués et des PSAP auxquels on soustrait les recours encaissés. C'est dans cet objectif de tarification que s'inscrit ce mémoire.

Deux approches majeures régissent le provisionnement en assurance : la méthode agrégée et la méthode ligne à ligne. Les méthodes agrégées offrent une bonne stabilité, mais les méthodes ligne à ligne permettent de prendre en considération l'intégralité de l'information disponible sur chaque sinistre. C'est cette dernière méthode qui est choisie dans ce mémoire.

Ce mémoire s'articule autour de deux axes. Le premier axe vise à améliorer la méthode existante mettant en œuvre des modèles de régression en introduisant un modèle de *Machine Learning* : le *XG-Boost*. Cette première partie a pour objectif de démontrer l'intérêt de ces modèles de *Machine Learning* en appliquant la méthode de manière iso *i.e.* sur la base des sinistres clos uniquement. Cette méthode présente des limites : elle ne permet pas d'inclure les sinistres en cours dans la modélisation, qui sont pourtant porteurs d'informations non-négligeables. De plus, on remarque un lien direct entre l'ancienneté du sinistre et son montant final. La prise en compte des dossiers en cours depuis plusieurs années devient alors nécessaire pour avoir une vision exhaustive de la sinistralité. On introduit donc dans le second axe une méthode tenant compte des sinistres en cours et donc plus généralement permettant de gérer la censure à droite en modifiant l'algorithme classique des arbres CART à l'aide de poids calculés à l'aide de l'estimateur de Kaplan-Meier.

Mots-clés : Provisionnement ligne à ligne, PSAP, *Machine Learning*, Responsabilité Civile Professionnelle, Branche Longue, Censure, Tarification, *GLM*, *XG-Boost*, Régression logistique, Kaplan-Meier, *CART*.

Abstract

In France, several professions are subject to the obligation to underwrite a professional liability insurance. The regulated professions in the legal and accounting fields (lawyers, chartered accountants, legal representatives, etc.) are subject to this obligation. The high stakes associated with this guarantee imply a high rate of litigation of claims. These factors make professional liability coverage for the legal and accounting professions a long-term insurance with a claim's development period of several years.

Technical reserves represent a considerable part of the insurer's liabilities. From an accounting point of view, establishing the right amount of reserves is at the heart of the duality between profitability and solvency. The ever-increasing competition, with the entry of bank insurers, and the financial crises have encouraged the implementation of the Solvency II standard. Outstanding claims provisions are the most significant technical provisions in property and casualty insurance in terms of amount. These reserves are intended to cover expenses attributable to the settlement of claims but also related to incidental expenses (expert fees, legal fees, etc.). The accounting department will work on calculating its outstanding claims provisions via a "Best Estimate" approach (as required by Solvency II), but a Product Actuary may also need to project these provisions to ensure that the level of risk is consistent with the provisions set up for each case. Indeed, the net ultimate loss expense is defined as the sum of the settlements already made and the provisions from which we subtract the recoveries collected. It is within this pricing framework that this paper is written.

Two major approaches govern insurance reserving: the aggregate method and the individual method. Aggregate methods offer good stability, but individual methods allow us to consider all the information available on each claim. The latter method is chosen in this paper.

This thesis is articulated around two axes. The first axis aims at improving the existing method implementing regression models by introducing a Machine Learning model: the *XG-Boost*. The objective of this first part is to demonstrate the interest of these Machine Learning models by applying the method in an iso way, i.e., based on closed claims only. This method has its cons: it does not allow to use ongoing claims in the modelling; however these claims can carry really precious information. Furthermore, we observe a direct link between the lifespan of the claim and its final amount. Taking into account these claims is necessary to have the exhaustive vision of our sinistrality. We therefore introduce in the second axis a method that considers the opened claims and handles right censoring by adding to the classic CART algorithm Kaplan-Meier calculated weights.

Keywords: Individual reserving, PSAP, Machine Learning, Professional Liability, Long Term Care, Censoring, Underwriting, *GLM*, XG-Boost, Logistic regression, Kaplan-Meier, CART.

Remerciements

J'aimerais commencer par adresser mes remerciements à Audrey ASSOUS-BENGUIGUI, directrice de l'Actuariat Statistique et Performance Économique au sein de la Direction Développement Courtage & Marché Entreprises de MMA. Ses conseils sur le monde professionnel, son expertise ainsi que la confiance qu'elle a su m'accorder, tout au long des mois passés, ont considérablement contribué à la réussite de ce mémoire.

Je remercie également Erwan GALÈS, désormais Responsable de la Technique et du Pilotage Économique au sein de Covéa Affinity et anciennement manager de l'Actuariat Construction et Marchés Spécialisés ; il m'avait à l'époque convaincue de rejoindre l'ISFA lorsque je n'avais même pas envisagé l'idée.

Je tiens à remercier mon tuteur d'apprentissage Pierre GOLHEN. Sa compétence et sa connaissance des différents marchés de l'assurance IARD m'ont permis d'acquérir de nombreuses compétences. Il s'est rendu disponible pour chacune de mes interrogations en faisant toujours preuve de bienveillance. Sa bonne humeur, même par temps maussade, a été une source de motivation.

Je remercie également le reste de mon équipe : Hervé ANDRE pour son savoir des différents produits d'assurance entreprise et Kevin MORIN qui a toujours répondu présent lorsque je rencontrais des difficultés sur SAS.

Je saisis également cette occasion pour adresser mes remerciements à mon tuteur pédagogique M. Nicolas LEBOISNE qui s'est montré très disponible pour nos échanges au détour d'une visioconférence ou d'un retard de la SNCF.

Enfin, je remercie ma famille pour leur soutien inconditionnel qui m'a permis de réaliser les études que je voulais et par conséquent ce mémoire. Merci à ma mère pour les soirées consacrées à la relecture, et surtout à la correction, de ce mémoire.

Table des matières

Introduction

Avocats, experts comptables, commissaires aux comptes, médecins, taxis, agents généraux d'assurance ces professions partagent un même point commun : il s'agit de professions réglementées en France. Elles sont soumises à certaines conditions qui encadrent leur exercice. Certaines de ces professions peuvent être regroupées, en particulier les professions du chiffre et du droit, dans lesquelles on retrouve les experts comptables, les commissaires aux comptes, mais aussi les avocats, les greffiers, les mandataires judiciaires,… Lors de l'exercice de ces professions des erreurs peuvent être commises par le professionnel : signature de comptes présentant des incohérences, retard dans l'exécution d'une prestation ou encore défaut de conseil. Définie sur le même socle que la responsabilité civile, la responsabilité civile professionnelle (RC Pro), couvre le professionnel contre les dommages causés à un tiers dans l'exercice de ses fonctions. Le marché des professions du chiffre et du droit est un atout important au sein de la Direction Développement Courtage & Marché Entreprises du groupe MMA, leader sur ce marché en France.

Les sinistres de ces professions sont assujettis à de nombreuses problématiques : une judiciarisation accrue des dossiers due aux enjeux financièrement importants, des liquidations de sinistres longues dues aux procédures judiciaires, une gestion déléguée à des sociétés d'assurance ou à des courtiers, synonyme parfois d'asymétrie d'information. Tarifer le risque lié à ces professions nécessite donc de prendre en considération ces éléments notamment au moment de la détermination de la charge sinistre ultime. La charge nette de recours encaissés est définie comme suit :

$$Charge\ nette\ de\ recours = Règlements + Provisions\ pour\ sinistres\ à\ payer - Recours$$

Les règlements et les recours constituent la partie constatée du sinistre. C'est sur la troisième composante, les provisions pour sinistres à payer (ou PSAP), que réside l'incertitude. Bien que ces provisions soient établies par le service comptable au niveau marché et par le gestionnaire sinistre au niveau dossier, l'actuaire produit a besoin de les projeter à l'ultime afin de quantifier correctement le risque lors de tarification. Le provisionnement en assurance non-vie s'effectue via deux grandes méthodes : la méthode agrégée et la méthode ligne à ligne. Même si les méthodes agrégées offrent une bonne stabilité, on souhaite appliquer une méthode ligne à ligne afin de déterminer, à l'ultime, le montant de liquidation des PSAP. En effet, la diversité des sinistres sur ce marché nous pousse à considérer les sinistres de manière individuelle et à utiliser le maximum d'informations disponibles sur le dossier. La méthode existante utilisait des modèles de régression sur la base des sinistres clos. Je souhaite prouver dans un premier temps l'intérêt d'un modèle de *Machine Learning*, le *XG-Boost*, avant d'introduire un modèle CART permettant de tenir compte des sinistres en cours lors de la modélisation.

Une première partie sera consacrée à la définition du cadre de ce mémoire : le contexte des professions réglementées en France, la définition de la responsabilité civile professionnelle et des enjeux du provisionnement en assurance non-vie. La seconde partie s'intéressera à la constitution de la base de données et aux premières statistiques descriptives sur celle-ci. L'approche théorique de cette étude sera abordée dans la troisième partie qui présentera les modèles de régression, les arbres *CART*, la méthode de *Boosting* et la procédure *weighted CART*. Enfin, la quatrième et dernière partie, présenteront les résultats de la modélisation.

Section 1 : Présentation du contexte professionnel

1. Les professions réglementées

1.1. Qu'est-ce qu'une profession réglementée ?

1.1.1. Définition

Administrateur judiciaire, expert-comptable, avocat, commissaire aux comptes sont des professions bien différentes et pourtant elles ont toutes une chose en commun : ce sont, en France, des professions réglementées. Une activité réglementée est soumise à des conditions d'accès et/ou des conditions d'exercices. L'exercice de ces professions est encadré soit :

- Par des dispositions législatives : c'est le cas pour les administrateurs judiciaires ou encore les commissaires aux comptes. (cf. Livre VIII du Code de Commerce)

- Par des dispositions réglementaires : pour les notaires, il faut se référer au « Règlement National/Règlement Inter-cours » publié par le Conseil Supérieur du Notariat et approuvé par le Garde des Sceaux.[1]

La liste des professions réglementées en France n'est pas exhaustive, car leurs définitions ne sont ni restrictives ni consensuelles cependant, la Commission Européenne[2] en recense 264.

Les professions réglementées se décomposent souvent en deux groupes :

- Certaines activités disposent de conditions d'accès bien définies : un diplôme particulier et un numerus clausus (limitation du nombre de personnes admises à exercer la profession). On peut citer par exemple : les médecins, les avocats, les notaires, les commissaires de justice ou encore les agents immobiliers.

- Certaines activités ne font pas l'objet de conditions d'accès particulières, mais de conditions d'exercice strictes : par exemple une obligation de formation ou de tenue d'un registre. C'est le cas des experts judiciaires.[3]

[1]BPI France, Activités réglementées
[2]Regulated professions database
[3]Les professions réglementées, article internet du Journal du Net

1.1.2. Quels sont les différents types de professions réglementées ?

Voici un schéma[4] reprenant les différents types de professions réglementées que l'on retrouve en France :

Figure 1 : Les professions réglementées en France

1.1.3. Pourquoi réglementer certaines professions ?

Les professions réglementées sont le plus souvent des professions anciennes. Disposer d'une réglementation est alors un gage de qualité et de compétence du professionnel pour le client. Les conditions d'accès limitantes, permettent de réguler le nombre de professionnels et de ne recruter que les meilleurs profils.

Au cours de l'exercice de certaines professions, il est possible de mettre en danger la vie d'autrui, c'est le cas des professions médicales (une erreur médicale par exemple) : la réglementation agit dans l'intérêt de limiter ces accidents et de protéger le client ou le patient.

1.1.4. Contexte politique : la libéralisation des professions réglementées

Depuis 2014, le gouvernement se penche sur les professions réglementées, en cause : les salaires jugés trop élevés et le manque de concurrence. Le schéma ci-dessous présente les salaires moyens nets par mois[5] en France de quelques professions réglementées :

[4] Page Wikipédia sur les professions réglementées
[5] Salaires moyens source : hellowork

Figure 2 - Salaires moyens nets de quelques professions réglementées

Depuis les années 1960, de nombreux rapports se succèdent : Augier[6], Attali[7], Ferrand[8], dont l'avis est unanime : la modernisation de l'économie doit impliquer une libéralisation des professions réglementées. Favorisée par la loi Macron pour la croissance, l'activité et l'égalité des chances économiques[9], cette libéralisation pourra se manifester par une baisse des honoraires ou un allègement des règles régissant ces professions. Cette dernière option a notamment été encouragée par la Commission Européenne dans le cadre de mise en conformité pour des professions telles que les avocats, les notaires ou encore les comptables vis-à-vis des normes de concurrence européenne. En effet, selon la théorie économique, pour qu'un marché atteigne l'équilibre, il doit être en concurrence parfaite. Lorsque ce n'est pas le cas, comme sur le marché des professions réglementées, la concurrence est réduite et par conséquent, l'offre devient inférieure à la demande ce qui entraîne inexorablement des prix élevés. L'objectif de la loi Macron était d'alléger la réglementation de ces professions par le biais de mesures variées : suppression du numerus clausus pour les notaires, mais également pour les avocats à la Cour de cassation ou encore instauration de la liberté d'installation pour les professions du droit.

Le marché des professions réglementées est donc un marché qui va subir de fortes transformations que cela concerne les professions réglementées du domaine de la santé, du domaine du droit ou du domaine de la certification de comptes. Dans la suite de ce mémoire, nous ne traiterons que des professions réglementées du chiffre et du droit qui est l'un des marchés des professions réglementées chez MMA. Le contexte politique autour de ces professions peut être source d'évolution du risque : conditions d'accès facilitées (avec la fin du numerus clausus pour les notaires par exemple). Ce constat encourage la mise en place d'une méthode de provisionnement plus adaptée.

1.2. MMA et les Professionnels du Chiffres et du Droit

MMA est un assureur historique des professions de chiffres et de droit en France : certaines professions sont assurées depuis des dizaines d'années. Souvent, l'adhésion au contrat d'assurance se fait via une instance représentative de la profession :

- Pour les avocats, l'adhésion se fait via leur barreau respectif.
- Pour les commissaires aux comptes, elle se fait via la Compagnie Nationale des Commissaires aux comptes.

[6] Rapport Augier 2012
[7] Rapport Attali 2014
[8] Rapport Ferrand 2014
[9] Loi Macron pour la croissance, l'activité et l'égalité des chances économiques

Nous sommes donc sur des contrats de type « groupe » avec un souscripteur représentant jusqu'à plusieurs dizaines de milliers d'assurés (on parle de conglomérat.) MMA assure des professionnels français :

- du chiffre : experts-comptables, commissaires aux comptes,…
- du droit : avocats, administrateurs judiciaires,…

Certains de ces comptes sont assurés en coassurance avec d'autres assureurs.

Étant leader sur ce marché, MMA dispose d'un historique conséquent qui lui permet de connaître au mieux les données disponibles.

2. La Responsabilité Civile Professionnelle

2.1. Qu'est-ce que la Responsabilité Professionnelle ?

2.1.1. Définition et exemples

Quel que soit le secteur, l'exercice d'une activité professionnelle peut être à l'origine de dommages causés à des tiers. Les dommages peuvent être causés de deux façons différentes :

- Soit par le professionnel dans le cadre de l'exécution de son contrat : par une faute professionnelle, une négligence ou encore la détérioration d'un bien qui lui est confié. On parle alors de Responsabilité Civile Professionnelle ou RC Pro.
- Soit par l'exploitation de l'entreprise. Par exemple l'exploitation de ses locaux : un accident survient dans les locaux de l'entreprise ou à cause d'un mauvais entretien (une enseigne qui tomberait sur un passant). On parle alors de Responsabilité Civile d'Exploitation ou RC Exploitation.

Voici quelques exemples de situations qui engagent la RC Pro :

- Exemple 1 : lors de la vente d'un bien, un agent immobilier omet d'informer le futur acheteur d'un problème d'infiltration. Quelques mois après la vente, l'acheteur se rend compte du problème et décide de se retourner contre son agent immobilier en remettant en cause le prix de vente. L'agent immobilier a failli à sa mission de conseil envers son client, c'est donc une faute professionnelle : sa RC Pro est donc engagée.

- Exemple 2 : lors d'un contrôle fiscal, une entreprise est mise en redressement fiscal à la suite d'erreurs de comptabilité. L'expert-comptable peut avoir commis une faute qui va entraîner un préjudice, on parle alors de responsabilité civile professionnelle délictuelle.

- Exemple 3 : lors d'une action en justice un avocat a fait preuve de négligence au niveau des délais et voies de recours. Pour engager la responsabilité de son avocat, le client a réuni les trois conditions cumulatives : le fait dommageable, le lien de causalité et le préjudice et donc la RC Pro du professionnel est engagée.

Bien qu'elle semble être indispensable pour n'importe quelle entreprise, la RC Pro n'est obligatoire que pour certains corps de métiers et facultative, mais conseillée pour d'autres.

2.1.2. Cadre juridique de la Responsabilité Civile Professionnelle

La définition de la RC Pro est fondée sur le même socle juridique que celle de la Responsabilité Civile. C'est le Code Civil et en particulier l'article 1240 qui définit la notion de responsabilité civile au sens large : « *Tout-fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé à le réparer.* ». L'article 1242 du même ouvrage vient préciser cette première définition : « *On est responsable non seulement du dommage que l'on cause par son propre fait, mais encore de celui qui est causé par le fait des personnes dont on doit répondre, ou des choses que l'on a sous sa garde. ».* Pour avoir des exemples concrets de situations impliquant la RC Pro, on se référera à la partie 2.1.1. Les professions réglementées font l'objet d'une obligation de souscription à une garantie RC Pro.[10]

Cette obligation peut-être soit fixée par le Code des Assurances (c'est le cas notamment pour les professionnels du bâtiment.) soit par la réglementation de la profession en question. Pour les avocats par exemple, c'est dans la *Loi n° 71-1130 du 31 décembre 1971 portant réforme de certaines professions judiciaires et juridiques*[11] qui stipule que : *« Il doit être justifié, soit par le barreau, soit collectivement ou personnellement par les avocats, soit à la fois par le barreau et par les avocats, d'une assurance garantissant la responsabilité civile professionnelle de chaque avocat membre du barreau, en raison des négligences et fautes commises dans l'exercice de leurs fonctions. »*.

2.2. Typologie des sinistres

2.2.1. Une liquidation longue des sinistres

En France, on compte quotidiennement environ 30 000 sinistres : bris de glace, accidents de la circulation, fuites, inondations, incendies,… Ce sont des évènements qui, bien qu'ils soient malheureux, sont assez récurrents. Entre la déclaration de ces sinistres courants et leur indemnisation/réparation, la période qui s'écoule est très rapide (parfois quelques jours pour un sinistre auto). Sur ces sinistres, dont les dossiers sont clos rapidement, on parle de branche courte en opposition aux branches longues qui sont représentées par des garanties comme la Responsabilité Civile Décennale en Construction ou encore la RC Pro. Sur le marché des professions réglementées du chiffre et du droit la durée de vie moyenne est beaucoup plus importante : il faut compter plusieurs années avant que le sinistre ne soit clos. Cela s'explique notamment par une forte judiciarisation des dossiers : plus de la moitié sont résolus au cours d'une procédure judiciaire.

Rappelons qu'en matière de responsabilité civile, trois conditions doivent être vérifiées pour que la RC Pro soit engagée : un fait dommageable, un préjudice et un lien de causalité entre le fait et le préjudice. Dans une partie importante des dossiers, il peut s'avérer particulièrement complexe d'établir les responsabilités et d'identifier ce lien de causalité entre le fait et le préjudice. Autre particularité, la jurisprudence : en plus d'être en évolution constante, celle-ci a une place très importante dans l'issue des sinistres en professions réglementées du chiffre et du droit. Il devient alors difficile de prédire l'issue d'un sinistre judiciarisé.

[10] La Responsabilité Civile du fait personnel
[11] Législation Avocats

Ces liquidations longues sont également expliquées par *l'article 4 du Code de procédure pénale* qui dispose que *« il est sursis au jugement de cette action tant qu'il n'a pas été prononcé définitivement sur l'action publique lorsque celle-ci a été mise en mouvement. »*[12]. Autrement dit, en cas de procédure pénale, l'assureur peut obtenir un sursis à statuer au civil dans l'attente de la résolution de l'affaire pénale. Le délai parfois très long de la justice pénale en France, compter plusieurs années, influence directement la durée de liquidation des dossiers.

2.2.2. Des sinistres sensibles aux évolutions réglementaires

Comme je l'ai déjà évoqué auparavant dans ce document, la réglementation et la déontologie ont une place considérable au sein des professions réglementées. Ainsi, les évolutions réglementaires de ces professions ont un impact direct sur les sinistres : la mise en place de nouvelles procédures judiciaires, ou d'évolutions en termes de fiscalité des entreprises, peut se matérialiser par une hausse des sinistres de la procédure concernée le temps que le professionnel s'adapte aux nouvelles mesures.

Sur ce risque de RC Pro des professions réglementées, une attention particulière doit donc être portée sur la réglementation et ses évolutions. En effet, les impacts de ces changements se matérialisent indéniablement sur la sinistralité, en particulier sur la fréquence des sinistres.

2.2.3. Sinistres et Gestion déléguée

Sur ce marché, les délégations de gestion sont très courantes et exigées par le courtier au moment de la souscription. Le fonctionnement de cette gestion déléguée diffère d'un courtier à l'autre. Que cela soit au niveau du délai de report, de l'appréciation du sinistre en premier avis (montant évalué lors de la connaissance du sinistre) ou encore du niveau de prudence associé. Des accords sur les forfaits d'ouverture des sinistres peuvent également exister entre le courtier et son client : un montant forfaitaire est alors affecté au sinistre en attendant d'avoir plus d'informations quant à l'issue de celui-ci. Les règles relatives au provisionnement restent toutefois fixées par l'assureur. Bien que les délégations de gestion soient encadrées et régulièrement auditées, elles peuvent engendrer un biais au niveau de la liquidation de ces sinistres puisque l'évaluation est à l'appréciation d'un tiers.

De plus, il est essentiel de pouvoir identifier certaines pratiques de gestion récurrentes chez les différents intermédiaires : cela peut soit donner des indications sur l'issue du sinistre, soit inciter à regarder de plus près des sinistres sous-évalués ou sur-évalués. Il est également intéressant de pouvoir identifier les évolutions de pratiques.

3. Enjeux de la RC Pro pour les Professions du Chiffre et du Droit

On relève 5 enjeux majeurs autour de cette garantie pour ce marché des professions réglementées du chiffre et du droit :

- La durée de gestion moyenne d'un sinistre : plusieurs années dues notamment à la judiciarisation importante des sinistres.

[12] Le sursis à statuer, Radier Associes, Septembre 2019.

- La judiciarisation importante des sinistres : en plus de participer à l'allongement de leur durée de gestion, cette judiciarisation rend plus difficile l'anticipation de l'issue du sinistre. En effet, celle-ci est très aléatoire, car elle dépend de la juridiction, de la jurisprudence, mais également du contexte actuel.
- La gestion déléguée : certains sinistres sont gérés par des partenaires externes (courtiers,…). Cet enjeu implique que l'on peut retrouver dans nos données des particularités de gestion avec notamment des pratiques qui évoluent dans le temps, mais également d'une profession à l'autre. (cf. 2.2.3)
- La temporalité : les sinistres des professions réglementées du chiffre et du droit sont sur une branche de développement longue et on dispose des différentes informations (montant du règlement, de la PSAP, des recours,…) à différentes visions.
- La multitude de professions : on compte une dizaine de professions environ. Contrairement à un produit IARD classique pour lequel les sinistres que peuvent subir les assurés sont similaires : dommage aux biens, incendie, vol, bris de glace,… Les sinistres sur le marché des professions réglementées du chiffre et du droit sont différents d'une profession à l'autre : ainsi, si un avocat peut se voir reprocher un manquement dans les délais ou pièces au cours d'une procédure judiciaire, un commissaire aux comptes va plutôt se voir reprocher la certification d'un bilan infidèle.

4. Le provisionnement en assurance non-vie

Le monde de l'assurance présente une différence majeure comparativement aux autres domaines économiques : son cycle de production est inversé, c'est-à-dire que l'assureur perçoit une prime avant même de savoir combien le risque lui coûtera à l'ultime. L'exercice de l'assureur (et de l'actuaire en particulier) est de calibrer cette prime afin qu'elle puisse absorber les sinistres à venir. Cette particularité impact immédiatement le bilan de l'assureur : en effet, il se voit dans l'obligation de constituer une part importante de son passif en provisions techniques afin de régler les sinistres à venir. Les provisions pour primes non acquises (PPNA), les provisions pour risques en cours (PRC) et les provisions pour sinistres à payer (PSAP) sont des exemples de provisions techniques d'assurance non-vie dont le but est de se couvrir :

- Des risques connus pour les PPNA et les PRC,
- Des risques à venir pour les PSAP.

Ce mémoire se concentrera uniquement sur les plus importantes de ses réserves techniques en assurance non-vie : les PSAP.

4.1. Principe du provisionnement

4.1.1. Exigences légales : article R331-6 du Code des Assurances

L'article R331-6 du Code des Assurances reprend les différentes provisions techniques à constituer et définit les PSAP comme la *« valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise »*[13]. Les provisions sont au cœur de la dualité entre solvabilité et rentabilité à laquelle sont exposés les assureurs :

- D'une part, la compagnie d'assurance doit garantir sa solvabilité en agissant de manière prudente au moment la tarification de ses produits, ou au moment de la constitution de ces provisions.
- D'autre part, comme toute autre entreprise, une compagnie d'assurance doit être rentable et elle doit donc mobiliser le montant optimal à provisionner afin de limiter le niveau des fonds propres et ainsi préserver sa rentabilité.

Cette dualité est constatée puisque les instances de réglementation ont tendance à favoriser une constitution importante de fonds propres tandis que leurs exigences en termes de fiscalité leur imposent un niveau de fonds propres raisonné (ceux-ci n'étant pas imposables).

4.1.2. Schéma explicatif du provisionnement ligne à ligne

Il existe 2 méthodes de provisionnement : ligne à ligne et agrégée. La méthode agrégée est l'approche de provisionnement classique mise en place par les assureurs. Elle repose sur l'agrégation des données en triangle puis la projection de ces données ainsi agrégées à l'ultime. Des méthodes comme Chain Ladder ou encore le modèle de Mack sont des exemples de méthodes agrégées. Une méthode plus récente commence à s'imposer : la méthode ligne à ligne. Cette méthode permet de prendre en compte les caractéristiques individuelles du sinistre, mais également d'exploiter toute nouvelle information que l'on obtiendrait au cours de la liquidation du sinistre et qui nous permettrait d'améliorer la prédiction. Voici un schéma explicatif du provisionnement ligne à ligne :

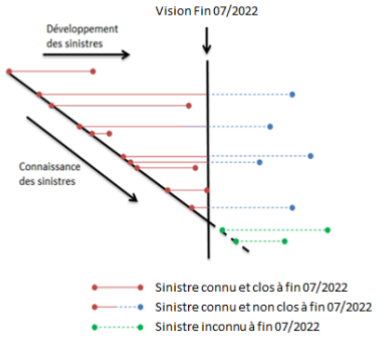


Figure 3 - Le provisionnement ligne à ligne

[13] Article R331-6 du Code des Assurances sur les provisions techniques

Ce schéma permet de décrire les 3 types de sinistres que l'on rencontre :

- Les sinistres clos arrivés à la fin de leur durée de vie (ligne rouge pleine)
- Les sinistres en cours dont le développement n'est pas clos (ligne rouge pleine et bleu pointillée) également appelés sinistres *RBNS (Reported But Not Settled).*
- Les sinistres inconnus à date (ligne verte pointillée) également appelés sinistres *IBNR* (*Incurred But Not Reported*)

L'essentiel de la charge sinistre ultime est porté par les sinistres clos et les sinistres en cours, mais il est nécessaire de comptabiliser les *IBNR* afin de ne pas sous-estimer la charge finale. L'estimation des *IBNR* ne sera pas abordée dans le cadre de ce mémoire.

4.2. Cadre de l'étude

4.2.1. Les provisions au sein de la tarification

Bien que la notion de provision soit souvent associée au domaine de la comptabilité, à juste titre, dans le cadre de cette étude, on aborde les provisions dans un intérêt de tarification. L'Actuariat Marchés Spécialisés de la marque MMA est un actuariat produit et non un actuariat prudentiel : il ne s'agit donc pas de positionner la provision comptable. Il devient important d'estimer le bon niveau des PSAP dans ce cadre lorsque l'on cherche à établir la charge ultime d'un contrat. Cette charge ultime peut être établie à des fins :

- De tarification : souscription d'un nouveau contrat
- De renouvellement : le marché des professions réglementées du chiffre et du droit est organisé par Grands Comptes. Il peut en exister plusieurs par profession et chacun des grands comptes est géré de manière individuelle au niveau de la souscription. La souscription de ces grands comptes se fait majoritairement via une *LTA (Long Term Agreement)* qui s'étend sur plusieurs années.
  Cela permet d'avoir un cycle de renouvellement des grands comptes à étudier chaque année et permet d'apporter une étude détaillée de chaque compte lors des discussions avec les différentes instances : actuariat, souscription, pilotage,…

Cette nécessité d'avoir une étude détaillée lors des différentes rencontres axées autour de la tarification du compte m'oriente déjà dans le choix de mes modèles :

- La méthode appliquée devra être une méthode ligne à ligne. En effet, les risques étant hétérogènes, il est intéressant d'utiliser les caractéristiques de chaque dossier pour estimer au mieux la liquidation ultime.
- Si jusqu'ici un modèle était entraîné par profession, on souhaite désormais partir sur une approche globale nous permettant d'utiliser l'intégralité de l'information disponible.

4.2.2. Calcul de la prime pure

Dans un objectif d'estimation de la charge sinistre, cette étude a pour objectif de déterminer le niveau le plus juste à l'ultime à affecter individuellement à chaque sinistre. L'approche classique de détermination de la prime pure est la suivante :

$$Prime\ Pure = Fréquence * Coût\ Moyen$$

Ici, mon objectif n'est pas de déterminer le nombre de sinistres puis la fréquence, mais d'estimer la charge sinistre ultime afin d'obtenir la prime pure globale du compte. On utilisera donc l'approche suivante :

$$Prime\ Pure = Charge\ Sinistre\ Totale\ Ultime$$

$$= Charge\ sinistres\ connus + Charge\ sinistres\ IBNR$$

La charge des sinistres connus peut être décomposée comme suit :

$$Charge\ sinistres\ connus = Réglements + PSAP - Recours$$

Les règlements sont les versements effectivement réalisés dans le cadre du dossier et constituent donc une partie constatée de la charge sinistre. La PSAP représente la provision associée au sinistre dans le cadre de nouveaux règlements potentiels qui pourraient survenir à la suite de l'évolution du dossier. Enfin, les recours représentent les sommes qui pourraient être remboursées par l'assureur adverse. On comprend ainsi que l'incertitude repose surtout sur le montant de la PSAP à l'ultime (les recours n'étant pas un phénomène récurrent sur ce marché). Il est nécessaire de prendre en compte le fait qu'un sinistre a de grandes chances que sa PSAP vaille 0€ au moment de la clôture du dossier. On fait donc évoluer la formule ainsi :

$$PSAP = \sum_{i=1}^{n} Probabilité\ de\ cloture\ \text{à}\ 0€\ sinistre_i * PSAP_{ultime}\ sinistre_i$$

Avec n le nombre total de sinistres. La modélisation de la PSAP s'articule donc en 2 étapes.

4.3. Méthode existante et problématique

Les travaux passés menés au sein de l'équipe avaient déjà démontré l'avantage de l'utilisation d'une méthode ligne à ligne au lieu d'une méthode par agrégation. La méthode utilisée jusqu'à présent reposait sur une application successive de modèles linéaires généralisés par profession. Puisque, comme on le voit sur l'histogramme ci-dessous, une masse importante des sinistres est liquidée à 0€ il fut nécessaire de raisonner en deux étapes :

- Une première régression logistique afin de déterminer si la PSAP sera liquidée à 0€ ou supérieure à 0€ : cette régression me permet d'obtenir la proportion de sinistres que l'on estime liquidés à 0€.
- Une seconde régression log-normale afin de déterminer le montant $PSAP_{ultime}$.

On obtient la charge totale en effectuant le produit de la probabilité que la PSAP soit liquidée à 0€ et le montant final estimé.

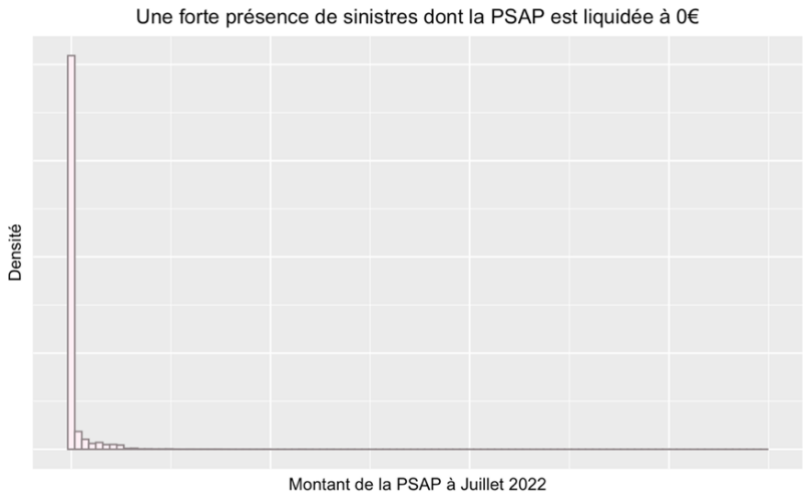


Figure 4 - Histogramme du montant de $PSAP_{ultime}$

Bien que cette méthode apporte des résultats satisfaisants, elle reste perfectible. D'abord, cela oblige à entraîner les modèles de régression sur la base des sinistres clos (c'est-à-dire ceux arrivés à la fin de leur liquidation) et ainsi ne pas tenir compte des sinistres en cours qui présentent des caractéristiques importantes, surtout sur les années récentes. Or, on sait que les sinistres importants ont une gestion plus longue ; ils représentent donc une part plus importante des dossiers en cours par rapport aux dossiers clos.

Ces deux typologies de sinistres ne sont donc pas correctement prises en compte lors de l'utilisation de cette méthode puisque l'on entraîne nos modèles sur des sinistres aux profils différents. Au regard des enjeux exposés et des problématiques associées à la méthode existante, traiter la méthode qui sera appliquée dans la suite devra :

- Pouvoir être appliquée sur la base de données complète des sinistres clos et en cours.
- Capter l'information liée aux déformations dans la liquidation.
- S'actualiser rapidement afin de prendre en compte les comportements des sociétés de gestion externes dans le traitement des sinistres.

Au regard des enjeux et des particularités du marché des professions réglementées du chiffre et du droit l'étude se décompose en deux parties :

- La première consistera à reproduire la méthode de régression mise en place sur SAS en l'améliorant (mise en place sur R d'un modèle globale) afin de pouvoir comparer l'existant avec nos autres modèles. On lancera un premier modèle de comparaison *XG-Boost* visant à améliorer la performance de cette méthode dont l'apprentissage ne s'appuie que sur les sinistres clos. Bien que celle-ci présente des points d'attention que j'ai soulevés, cette étape est nécessaire afin d'introduire le *Machine Learning*. En effet, pouvoir comparer les deux méthodes sur une base « iso » me permettra d'appuyer notre choix devant la souscription et la performance économique de la mise en place d'une méthode d'apprentissage automatique.
- La deuxième partie s'attelera à modifier la méthode en entrainant les modèles sur l'intégralité des sinistres et en intégrant des modèles permettant de gérer le phénomène de censure à droite observé sur les dossiers en cours.

Section 2 : Présentation et travaux autour de la base de données

5. Origine des données

Comme je l'ai évoqué précédemment, MMA est un assureur majeur et historique du marché des professions réglementées du chiffre et de droit, les données contrats et sinistres de la marque représentent donc une partie non-négligeable de l'information disponible sur le marché. La construction de la base de données est réalisée sous le logiciel *SAS*.

6. Construction des bases de données

On souhaite exploiter le maximum d'informations disponibles. Pour cela, on veut prendre en compte les différentes visions du sinistre. Je dispose de partitions des bases de données historiques me permettant de récupérer plusieurs variables tels que : le montant du règlement principal, le montant du règlement accessoire ou encore le montant de la PSAP à des visions de fin d'année 2021,2020, 2019,…, jusqu'à 1995 pour les plus anciens sinistres. Telle qu'elle est construite, la base de données possède de nombreuses colonnes, on va donc translater la base de données afin d'avoir les données en ligne ce qui facilitera par la suite leur traitement sous R pour la modélisation. Pour cela, je récupère les différentes visions et les différencie en introduisant la variable « id » qui va me permettre d'identifier la vision en question. Prenons l'exemple simple de 2 sinistres ouverts en 2020 :

- Un sinistre en cours
- Un sinistre clos

Dans la base globale des sinistres, on aura les lignes suivantes (pour les variables PSAP, Règlement et Règlements Accessoires par exemple) :

Base Sinistres

N° Contrat	N° Sinistre	Etat du Sinistre	Garantie	Montant PSAP N	Frais Accessoires N	Montant PSAP N-1	Frais Accessoires N-1	Montant PSAP N-2	Frais Accessoires N-2
1521	2217	Clos	RC Pro	0	500	600	500	1700	650
1432	543	En cours	RC Pro	3500	300	2980	0	2500	0

Figure 5 - Base sinistre initiale

Cette base de données me permet de créer les 2 bases qui serviront pour la première étape de cette étude :

- La « ***Base Liquidation*** » : comprenant tous les sinistres clos et pour lesquels on dispose du montant final de la PSAP. Cette base me servira pour la modélisation :

Base Liquidation

N° Contrat	N° Sinistre	Garantie	Id	Montant PSAP	Frais Accessoires
1521	2217	RC Pro	1	1700	650
1521	2217	RC Pro	2	600	500
1521	2217	RC Pro	3	0	500

Figure 6 - Base Liquidation « translatée »

- La « ***Base des Sinistres en cours*** » : comprenant les sinistres toujours en cours pour lesquels on doit prédire le montant final de la PSAP à partir des modèles entraînés sur la base de Liquidation.

Base Sinistres en cours

N° Contrat	N° Sinistre	Etat du Sinistre	Garantie	Montant PSAP N	Frais Accessoires N
1432	543	En cours	RC Pro	3500	300

Figure 7 - Base sinistres en cours

7. Présentation de la base de données

7.1. Les variables dites « clés de jointure »

Les variables suivantes sont les variables qui servent à la construction de la base de données, on retrouve notamment parmi elles :

- Le numéro de contrat
- Le numéro de sinistre (un contrat pouvant avoir plusieurs sinistres)
- La garantie sinistrée : RC Pro ou RC Exploitation essentiellement

Ces trois variables me permettront de récupérer l'intégralité des variables explicatives qui seront utilisées dans la suite.

7.2. Les variables explicatives

Contrairement à des sinistres IARD classiques, on dispose de peu d'informations sur les circonstances sinistres. En effet, j'ai déjà évoqué le caractère sensible de certains sinistres pour lesquels je ne dispose que de peu d'informations : par exemple, les circonstances du sinistre ne seront pas renseignées. On peut retrouver cependant des informations sur la procédure ou l'acte sinistré.

On décompose les variables explicatives en deux groupes :

- Les variables explicatives quantitatives :

- Les variables continues :
  - PSAP Nette de recours attendus : dernière vision de la PSAP avant liquidation.
  - Ancienneté : en jours écoulés depuis la date d'ouverture du sinistre.
  - Montant d'évaluation à l'origine.
  - Règlement principal : règlement imputable au sinistre directement
  - Règlement accessoire : règlement imputable aux frais de justice ou d'expertise engagés.

- Les variables binaires :
  - Forfait : 1 si le sinistre est au forfait ou 0 sinon.
  - Règlement principal : 1 si présence d'un règlement principal, 0 sinon.
  - Règlement accessoire : 1 si présence d'un règlement de frais accessoires, 0 sinon.

- Les variables explicatives qualitatives :
  - Nature du dommage : corporelle (l'enjeu est marginal pour cette garantie) ou matérielle.
  - La garantie sinistrée : RC Pro ou Autres.
  - La profession.
  - Les circonstances sinistres lorsque celle-ci est disponible : par exemple la procédure ou l'acte sinistré.
  - La nature juridique du sinistre : Amiable, Judiciaire ou Indéterminé.

7.3. Les variables à expliquer

Pour cette première modélisation, on décompose la prédiction du montant final en deux étapes :

- 1ère étape : modéliser la variable binaire « PSAP Liquidée à 0 » qui me permettra d'obtenir la probabilité que la PSAP soit liquidée à 0.
- 2ème étape : modéliser la variable continue « Montant $PSAP_{ultime}$ » qui me permettra d'obtenir le montant ultime de la PSAP.

7.4. Autres variables

D'autres variables sont présentes dans nos bases, bien qu'elles ne soient pas utilisées lors de la modélisation, celles-ci servent de filtre lors de la présentation des résultats, on cite par exemple :

- Date de survenance du sinistre.
- Date d'ouverture du sinistre (date de connaissance du sinistre par la direction indemnisation).
- Date de clôture du sinistre.
- Exercice
- Nom du grand compte

8. Analyse et travail sur la base de données

8.1. Discrétisation des données continues

On se propose de discrétiser les variables continues suivantes : Ancienneté, $PSAP_{dernière\ vision}$ et Montant d'évaluation d'origine. Ces variables discrétisées me seront utiles lors de la modélisation par les modèles de régression afin d'observer les résultats suivant l'utilisation des variables continues ou discrétisées. La discrétisation a été obtenue en utilisant le package ***questionr*** disponible sur R. Par exemple la variable Ancienneté est discrétisée ainsi :

- Ancienneté :
  - Inférieure à 1 an
  - Entre 1 et 2 ans
  - Entre 2 et 3 ans
  - 4 ans et plus

On applique cette méthode aux deux autres variables citées ci-dessus.

8.2. Modification de variables

8.2.1. Traitement des valeurs aberrantes

L'historique dont dispose MMA sur ces données me permet d'avoir peu de valeurs aberrantes ou manquantes. Toutefois, pour la variable « Nature Juridique », on disposait de 3 modalités : Amiable, Judicaire et Indéterminé. Après un échange avec l'Indemnisation, il apparaît que les sinistres dont la Nature Judiciaire est indéterminée sont en réalité des sinistres « Amiables ». J'ai donc finalement décidé que les sinistres à la Nature Judiciaire « Indéterminé » seront inclus dans les sinistres « Amiables ».

8.2.2. Construction de nouvelles variables

Lors de cette étape, je cherche à créer de nouvelles variables à partir des variables explicatives dont je dispose. On possède comme information sur les sinistres la profession ainsi que les circonstances sinistres (lorsque celles-ci sont renseignées). Comme je l'ai déjà abordé dans la partie 2.1.1 suivant la profession les sinistres ne concernent pas la même procédure : en effet, un expert-comptable ne pourra avoir de sinistres liés à une erreur dans la procédure d'appel tout comme un agent immobilier ne possédera pas de sinistres liés à une erreur de déclaration fiscale. Fort de ces constats, il semble plutôt logique de concaténer la profession sinistrée avec la circonstance sinistre. Après s'être rapprochée de la souscription afin d'identifier les circonstances sinistres que l'on peut regrouper pour chaque profession, on introduit la variable explicative qualitative suivante : « Profession x Circonstance Sinistre ».

Pour les variables continues règlement principal et règlement accessoire, on introduit deux variables binaires :

- Règlement Principal : 1 s'il y a déjà eu un règlement principal, 0 sinon.
- Règlement Accessoire : 1 s'il y a déjà eu un règlement accessoire, 0 sinon.

Cela permettra d'évaluer leur effet quand ils sont introduits dans le modèle en tant que variables continues et en tant que variables binaires.

9. Statistiques Descriptives sur les bases de données

L'objectif de cette partie est à la fois d'illustrer les différents enjeux que j'ai pu soulever dans la partie 3, mais également de visualiser les données. ***Pour des raisons de confidentialité, les noms des professions ainsi que les différents montants ou volume des sinistres ne seront pas divulgués.***

9.1. Statistiques descriptives sur la base des sinistres

9.1.1. Une dizaine de professions assurées

Le portefeuille MMA est composé d'une dizaine de professions comme on peut l'observer sur le diagramme circulaire représentant le nombre de sinistres par professions ci-dessous :

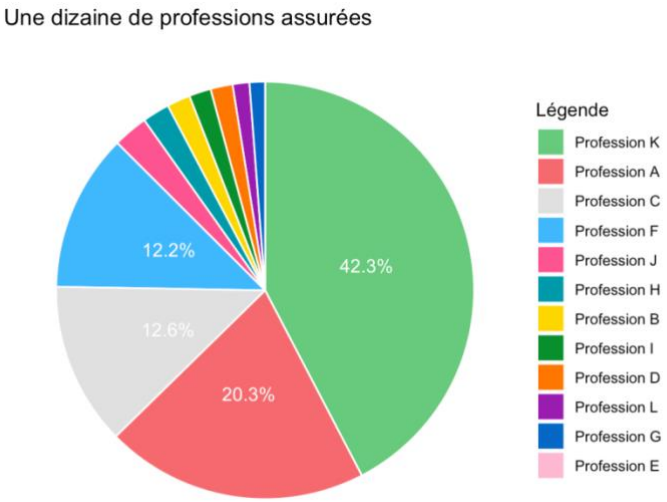


Figure 8 - Diagramme circulaire de la fréquence de sinistres par professions

Certaines professions sont plus représentées que d'autres avec notamment une profession représentant près de la moitié des sinistres. Trois autres professions se partagent quant à elle environ 40 % des sinistres en fréquence. Le huitième restant est reparti de manière plutôt équitable entre les 8 professions restantes.

9.1.2. Répartition selon la garantie

La partie 2.1 définissait la Responsabilité Professionnelle ainsi que les deux grandes garanties existantes en France :

- La RC Pro : assurant directement le professionnel lors de l'exécution de ses fonctions
- La RC Exploitation : assurant les locaux exploités par le professionnel lors de l'exécution de ses fonctions par exemple.

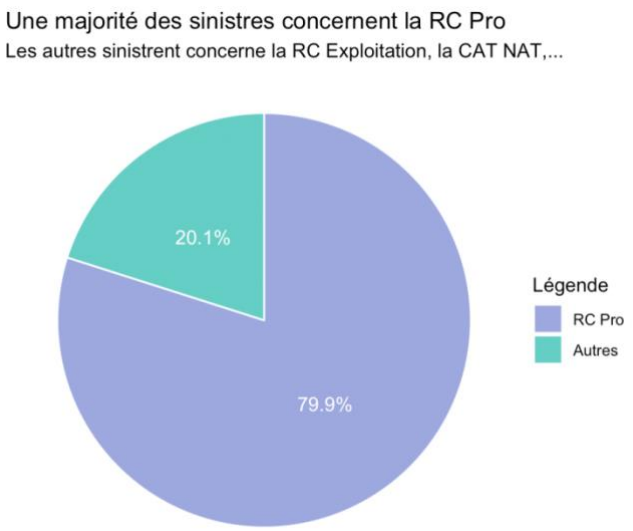


Figure 9 - Diagramme circulaire des garanties sinistrées

Comme on l'observe sur ce graphique, plus des ¾ des sinistres concernent la RC pro. Le reste des sinistres concernent les autres garanties en particulier la RC Exploitation, mais également des garanties complémentaires comme la CATNAT par exemple. En effet, certaines professions impliquent la conservation d'archives, si celles-ci se retrouvent endommagées, à cause d'une inondation par exemple, c'est la garantie CATNAT qui sera sinistrée.

9.1.3. La judiciarisation importante des sinistres

La judiciarisation des sinistres est la principale raison expliquant la liquidation longue des sinistres sur ce marché.

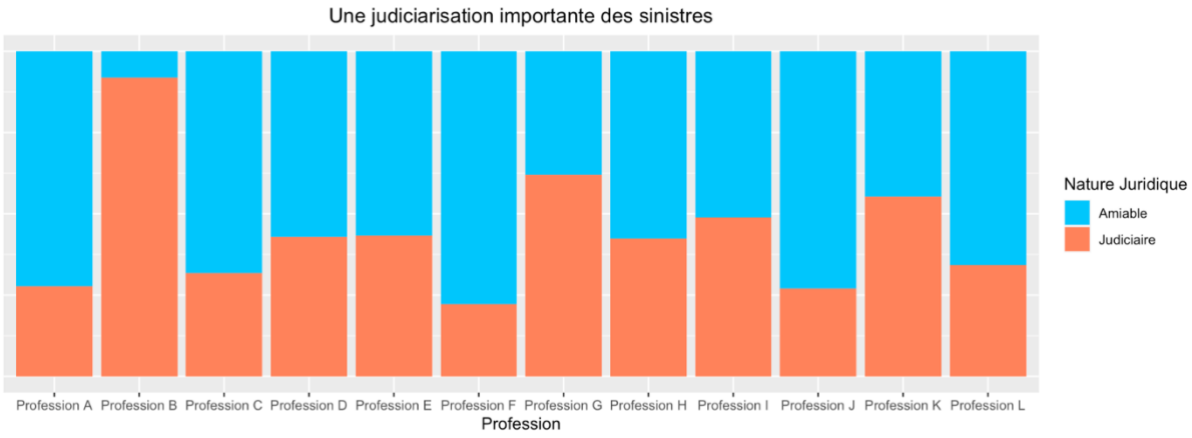


Figure 10 - Histogramme de la judiciarisation par profession

Comme nous pouvons le voir sur ce diagramme en bâtons, le pourcentage de sinistre judicaire varie d'une profession à l'autre. Il est normal d'observer ces différences ; en effet pour la profession présentant le taux le plus important de judiciarisation, un sinistre n'est ouvert qu'en cas de procédure judiciaire (pour la RC Pro). On comprend ainsi l'importance de la nature juridique dans la typologie des sinistres.

9.1.4. Sinistres clos et sinistres en cours

La représentation de la part des sinistres clos et des sinistres en cours nous apporte des informations sur la durée de clôture de ces sinistres et *a fortiori* sur leur durée de liquidation.

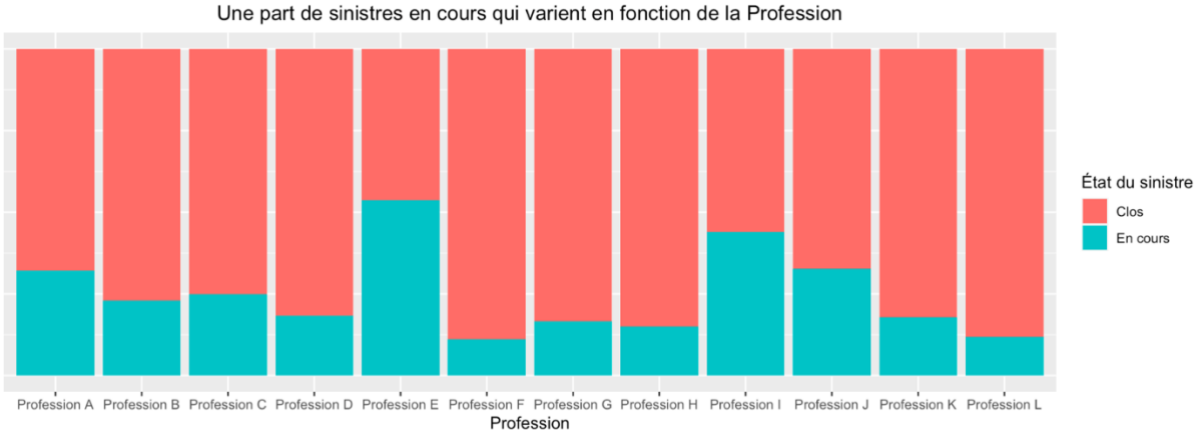


Figure 11 - Histogramme de l'état du sinistre par profession

Ce graphique nous indique par exemple que pour certaines professions, le taux de sinistres en cours oscille entre 30 % et 55 %. Constat flagrant : certaines professions fortement judiciarisées présentent une proportion plus importante de sinistres en cours. Les autres professions réussissent à maintenir un taux avoisinant les 20 % ce qui n'est pas négligeable comparé à d'autres marchés présentant des taux de sinistres en cours beaucoup plus faible (rappelons que je dispose d'un historique de sinistres de plus de 20 ans !).

9.1.5. Liquidation des sinistres

Les deux graphiques précédents ont tâché d'illustrer la liquidation relativement longue des sinistres sur le marché des professions de chiffres et de droit, mais voyons comment se comporte la variable en fonction de la profession. Le graphique *boxplot* ci-dessous représente le temps de développement en années par profession.

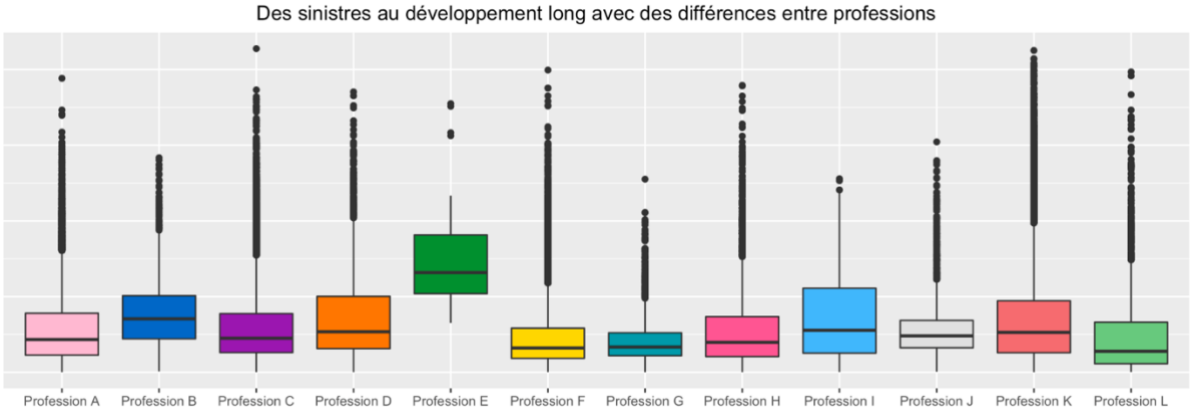


Figure 12 - Boxplot du temps de développement des sinistres

On peut cependant extraire quelques constats importants de ce graphique. Premièrement, la clôture du dossier survient 3 ans en moyenne après l'ouverture du sinistre.

Deuxièmement, d'après l'amplitude des boîtes, on conclut que les sinistres peuvent durer plusieurs années sans toutefois qu'il existe un écart considérable au sein d'une même profession. Troisième constat, une présence importante de valeurs jugées extrêmes par le graphique : encore une fois, cela illustre bien l'existence de sinistres très anciens, toujours en cours pour lesquels on devra déterminer un niveau de liquidation ultime. Enfin, la profession E qui s'illustre par son taux important de sinistres judicaires, mais également de sinistres en cours est la profession présentant les sinistres au temps de développement les plus longs.

9.1.6. Montant de la $PSAP_{ultime}$ (hors sinistres clos à 0€)

Toujours à l'aide d'un *boxplot* on souhaite observer les différences pouvant exister entre chaque profession par rapport au montant de la $PSAP_{ultime}$. Pour ce faire, on exclut les $PSAP_{ultime}$ liquidées à 0€ (représentant une part importante des $PSAP_{ultime}$ cf. 4.3). Ci-dessous, le graphique obtenu :

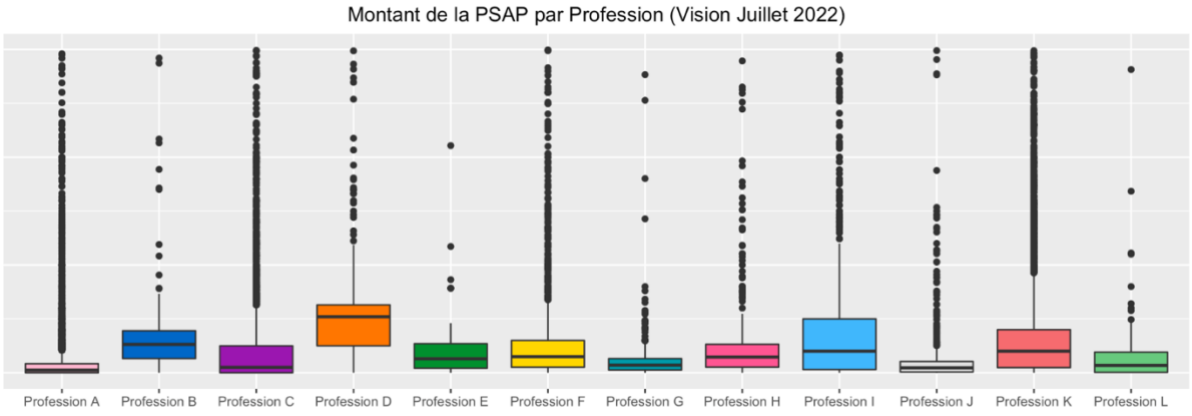


Figure 13 - Boxplot du montant de la PSAP

Les constats sont relativement similaires à ceux réalisés plus haut :

- Des différences de montant d'une profession à l'autre.
- Pour certaines professions l'amplitude de la boîte est relativement faible : les sinistres présentent donc peu de variance dans le montant de la $PSAP_{ultime}$. Cela peut être expliqué notamment par l'existence de forfaits d'ouverture.
- On observe toujours une présence importante de données jugées extrêmes. Elles sont caractéristiques de niveaux de $PSAP_{ultime}$ élevés et vont nécessiter la mise en place d'une méthode permettant d'établir le seuil de ses $PSAP_{ultime}$ graves.

Déterminer les bons niveaux de provisions a donc deux intérêts majeurs :

- Pour l'Actuariat : avoir une projection correcte du marché afin de prendre des décisions pertinentes lors des renouvellements.
- Pour le service comptable : éviter les rechargements excessifs sur d'exercices anciens pour lesquels on s'attend plutôt à une certaine stabilité.

9.2. Statistiques descriptives sur la base de liquidation

9.2.1. Liquidation du nombre de sinistres par vision de liquidation

Cette représentation permet d'afficher le triangle de liquidation du nombre de sinistres de manière graphique. On observe ainsi pour chaque exercice le nombre de sinistres supérieurs à 0€ par vision de liquidation. Premier constat sur ce graphique ; le nombre de sinistres décroît d'une vision à l'autre à partir de la première vision. Toutefois, remarquons que le nombre de sinistres supérieurs à 0€ augmente de façon importante au cours de la première année de liquidation, on observe un atterrissage du nombre de sinistres au fur et à mesure des visions de liquidation.

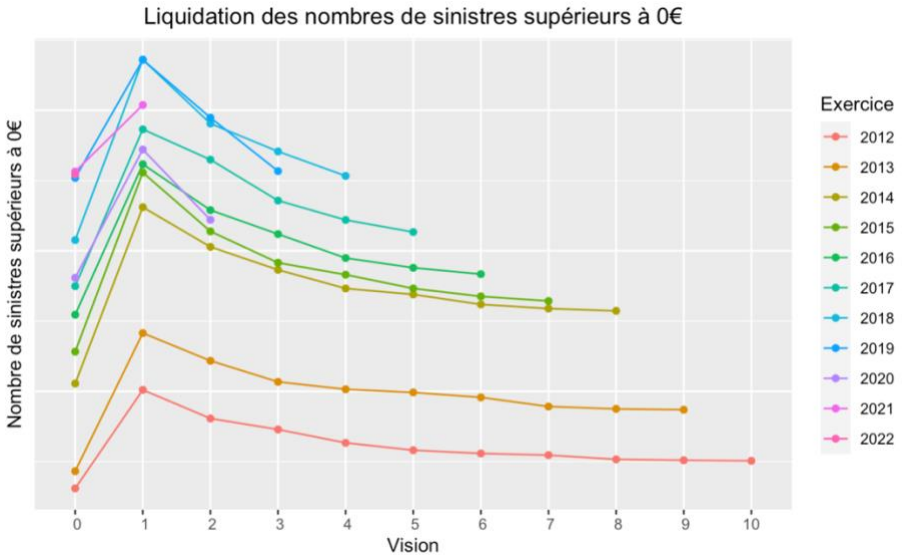


Figure 14 - Graphique de liquidation du nombre de sinistres supérieurs à 0€

9.2.2. Liquidation des PSAP par vision de liquidation

Après avoir observé la représentation graphique du triangle de fréquence, on se propose d'illustrer le triangle de montant de PSAP à l'aide du même type de graphique. Sur la figure ci-dessous, le constat sur le montant de la PSAP constatée diffère du constat fait sur la fréquence constatée. En effet, pour l'exercice 2016 (qui est un exercice relativement ancien : comprenez plus complet) on remarque que la vision de la PSAP lors de la 6ème année de liquidation est plus importante que le niveau de la PSAP en 5ème vision. De plus, lorsque l'on observe la liquidation en fréquence, on ne remarque pas une hausse au niveau du nombre de sinistre. Il apparaît donc que ce rechargement intervient sur un ou plusieurs dossiers déjà connus. Autrement dit, le montant de provisions attribué lors des différentes années de liquidation n'a pas été suffisant. Ce graphique illustre le problème rencontré : il est nécessaire d'affecter un montant correct afin de ne pas avoir à augmenter les PSAP d'une année à l'autre.

Hormis la sensibilité de ces rechargements au niveau comptables, ils doivent être pris en compte au niveau de l'Actuariat afin de ne pas sous-estimer la charge sinistre ultime et donc la prime pure du contrat. Ceci illustre notamment pourquoi il est nécessaire de fixer dès le départ le bon niveau de provision afin d'éviter ce genre de phénomènes de déformation de liquidation.

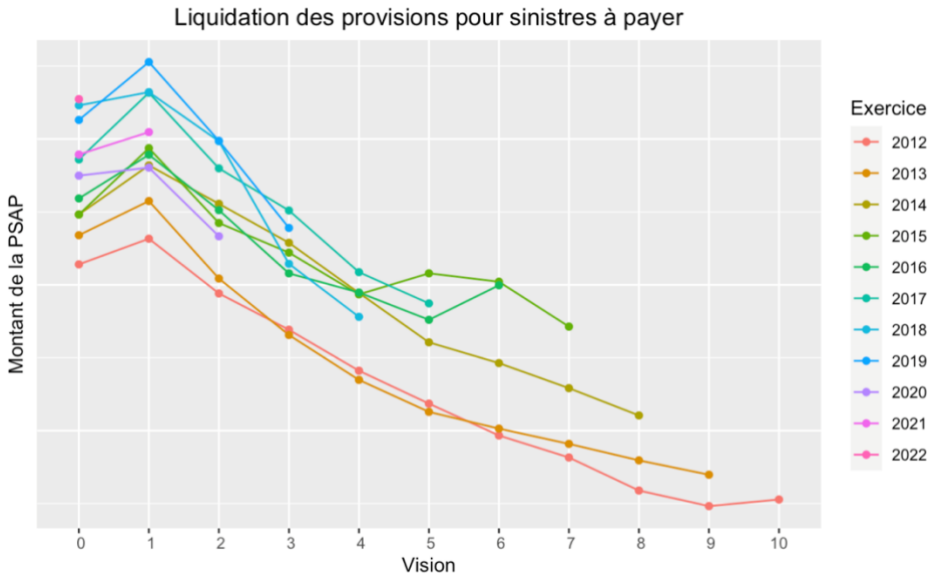


Figure 15 - Graphique de liquidation du montant des PSAP

Ce graphique nous permet également de caractériser la liquidation longue de ces provisions.

9.3. Statistiques descriptives sur la base des sinistres en cours

Dans la partie 4.3 qui traitait des problèmes rencontrés avec la méthode actuelle, j'avais évoqué le fait de n'entraîner les modèles que sur la base des sinistres clos (sinistres pour lesquels on connaît la valeur de $PSAP_{ultime}$). En faisant ce choix, on omet des sinistres en cours qui sont potentiellement ouverts depuis longtemps, judiciarisés, dont l'issue peut être défavorable et nécessiter le rechargement de la provision d'un exercice particulièrement ancien. On souhaite donc illustrer cette problématique en représentant l'ancienneté (en années) du sinistre sur la base des sinistres clos et sur la base des sinistres en cours :

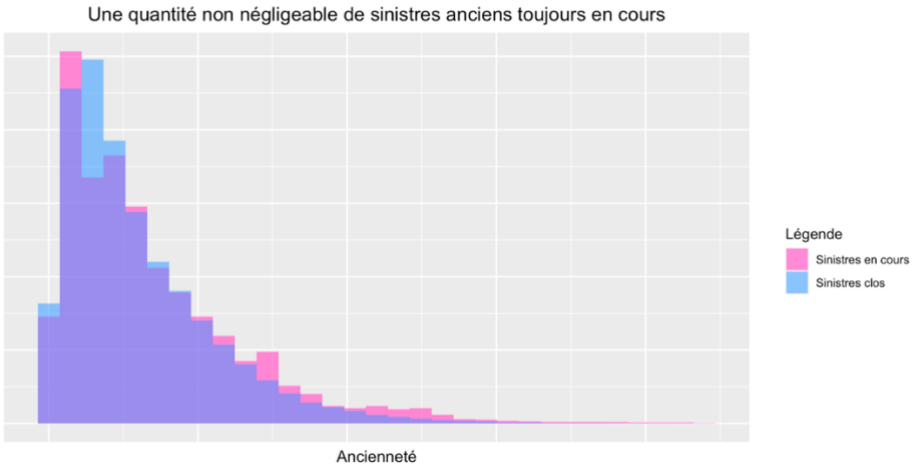


Figure 16 - Histogramme de l'ancienneté

L'histogramme ci-dessus confirme le constat fait précédemment : si la base des sinistres en cours présente une majorité de sinistres relativement nouveaux, on remarque tout de même une quantité de sinistres très anciens. Ceux-ci sont également plus représentés dans la base des sinistres en cours que dans la base des sinistres clos.

Section 3 : Approche théorique

10. Modèles de régression

10.1. Rappel sur les modèles linéaire généralisés

10.1.1. Contexte historique

On doit la formulation des modèles linéaires généralisés à Nelder et al. (1972)[14]. Les *GLM* sont une version souple de la régression linéaire, on les utilise lorsque la variable à expliquer ne vérifie plus les hypothèses de la régression linéaire :

- Hypothèse de normalité de la variable à expliquer ou des résidus
- Hypothèse d'homoscédasticité (c'est-à-dire que la variance ne dépend pas de la moyenne)

De manière générale lorsque la variable cible Y n'est pas une variable continue non bornée, mais une donnée de comptage ou une donnée binaire.

10.1.2. Les 3 propriétés importantes des *GLM*

La première propriété d'un *GLM* réside dans le fait que la distribution de la variable à expliquer doit être connue et appartenir à la famille exponentielle. Parmi les distributions de la famille exponentielle les plus connues, on retrouve : la loi Normale, la loi Exponentielle, la loi de Poisson, la loi Gamma, la loi du χ^2.

La seconde propriété des *GLM* porte sur la droite de l'équation, celle-ci doit être linéaire, c'est-à-dire du type :

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon$$

Avec $\theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon$ le prédicteur linéaire et Y une variable cible.

La troisième et dernière propriété concerne la relation entre la variable à expliquer et les données. Une fonction de lien doit lier les données et la variable à expliquer, parmi les fonctions suivantes : la fonction identité, la fonction log ou la fonction $logit$.

10.1.3. La variable à expliquer Y

La loi de probabilité de la variable Y doit appartenir à la famille exponentielle c'est-à-dire que l'on doit pouvoir l'écrire sous la forme :

$$f(y_i, \theta_i, \varphi, \omega_i) = exp^{\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)}\omega_i + c(y_i, \varphi, \omega_i)}$$ [15]

Les fonctions $a, b\ ou\ c$ diffèrent selon le type de loi exponentielle.

---

[14] Nelder et al. (1972)

[15] *GLM*, Michel Tenenhaus

- θ_i est appelé le paramètre canonique, il est défini comme une fonction de l'espérance. C'est un paramètre inconnu.
- φ est le paramètre de dispersion, celui-ci est supposé connu. Dans le cas où il serait inconnu, on l'estimera préalablement le plus souvent comme suit : $a(\varphi_0) = \varphi_0$.
- ω_i sont des poids.

10.1.4. La fonction de lien

La fonction de lien permet de s'assurer que les valeurs prédites par le modèle respectent les valeurs réelles de la variable Y. Par exemple pour des données de comptage, on doit s'assurer que le modèle prédit des valeurs appartenant à $\mathbb{R}^+$. Si plusieurs fonctions peuvent être appliquées, on choisira celle qui minimise l'erreur quadratique moyenne entre les variables réelles et les variables prédites.

De manière générale :

- Pour des variables continues sur $\mathbb{R}$, on choisira la fonction de lien identité :

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon$$

- Pour des variables discrètes sur $\mathbb{N}$, on choisira la fonction de lien log :

$$\log(Y) = \theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon$$

$$\Leftrightarrow Y = e^{\theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon}$$

- Pour une variable cible $Y \in [0\,;1]$, on choisira la fonction de lien $logit$:

$$\log\left(\frac{y}{1-y}\right) = \theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon$$

$$\Leftrightarrow \quad y = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X_1 + \cdots + \theta_n X_n + \varepsilon)}}$$

10.1.5. Le maximum de vraisemblance

Tandis que la régression linéaire simple fonde l'estimation de ses paramètres sur la minimisation des moindres carrés ordinaires, les modèles linéaires généralisés quant à eux cherchent à maximiser la vraisemblance notée $\mathcal{L}(\Theta)$:

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} f(y_i | X = x_i; \theta_i)$$

Avec $f(y_i | X = x_i; \theta_i) = \begin{cases} f_{\theta_{1,\ldots,\theta_n}}(y_i | X = x_i),\ pour\ les\ v.a.\ continues \\ \mathbb{P}_{\theta_{1,\ldots,\theta_n}}(y_i | X = x_i),\ pour\ les\ v.a.\ discrètes \end{cases}$

10.1.6. La déviance et le critère d'Akaike (*AIC*)

On veut estimer un modèle dont les résultats dévient le moins possible des valeurs observées de Y. On va minimiser l'écart entre la vraisemblance du modèle et la vraisemblance de l'échantillon observé, c'est-à-dire :

$$Déviance = -2(\log(L_{modèle}) - \log(L_{échantillon}))$$

On pourra ajuster un *GLM* en minimisant sa déviance.

Il est rare de lancer un unique *GLM*, on lance plusieurs *GLM* construits différemment puis on les compare. On va utiliser pour cela le critère d'Akaike (*AIC*). Le but est de satisfaire le principe de parcimonie, c'est-à-dire le principe philosophique qui affirme que *« les hypothèses suffisantes les plus simples doivent être préférées »*[16]. Autrement dit, on va pénaliser les modèles trop complexes en augmentant la déviance du modèle d'un terme $2(k+1)$ multiple du nombre de paramètres k. Le critère d'*AIC* s'écrit alors :

$$AIC = -2\log(L_{modèle}) + 2(k+1)Déviance$$

Le modèle qui aura l'AIC le plus faible sera le plus performant. Pour cela, on procède comme suit : on calcule l'$AIC_{modèle}$ et l'$AIC_{échantillon}$ afin d'obtenir la quantité :

$$\Delta AIC = AIC_{modèle} - AIC_{échantillon}$$

Si $\Delta AIC \leq 2$ alors le modèle est considéré comme le plus performant.

10.2. La régression logistique pour modéliser la probabilité de clôture à 0€

10.2.1. Constat pratique

Identification des facteurs liés à une maladie, détection d'individus risqués lors de l'octroi d'un crédit, ciblage d'une clientèle lors de la mise en vente d'un produit[17], sont des phénomènes appartenant à des secteurs bien différents, mais qui ont en commun le fait de pouvoir être modélisés à l'aide d'une régression logistique. Le choix d'un modèle logistique pour prédire la probabilité de clôture à 0€ de la $PSAP_{ultime}$ s'inscrit alors dans la continuité des exemples cités. Les modèles de régression logistique, également appelés modèles $logit$, sont utilisés lors de la prédiction de variables qualitatives ou quantitatives discrètes. La suite de notre étude s'attardera particulièrement sur la régression logistique d'une variable binaire. Voici la répartition de notre variable binaire cible, notée Y, et qui est définie comme suit :

$$Y = \begin{cases} 1, si\ la\ PSAP\ est\ liquidée\ \text{à}\ 0€ \\ 0, sinon \end{cases}$$

[16] Rasoir d'Ockham, principe de parcimonie
[17] Régression Logistique, IBM.com

On pourrait alors tracer une courbe linéaire qui passerait par les données (0 ou 1) :

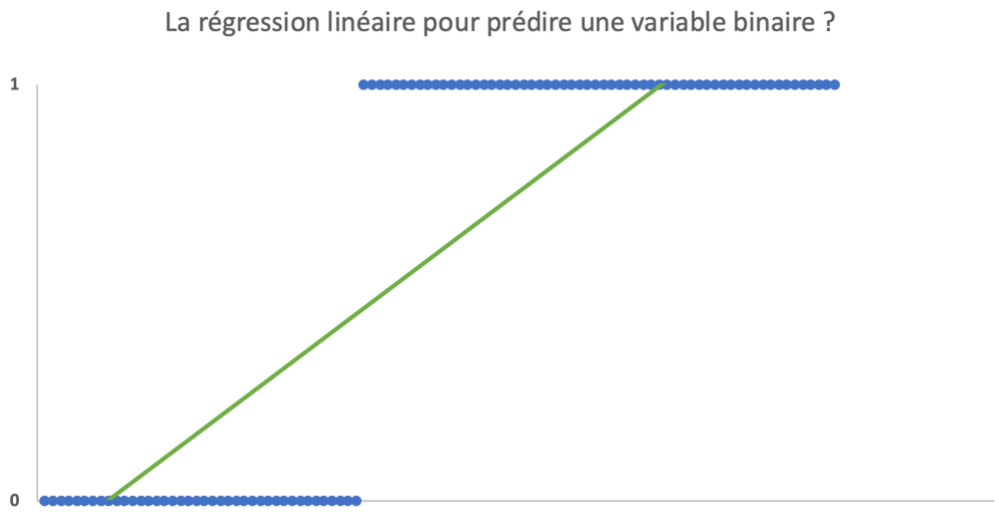


Figure 17 - Régression linéaire ajustée sur une variable binaire

Cependant, on sait que la variable Y ne prend que les valeurs 0 ou 1 ainsi $E[Y] = \begin{cases} 1 \\ 0 \end{cases}$[18].

Il apparaît que la fonction logistique est celle qui permet de représenter le mieux ce type de données, en tenant compte des conditions sur l'espérance. Son expression est la suivante :

$$x \longrightarrow \frac{1}{1 + e^{-x}}$$

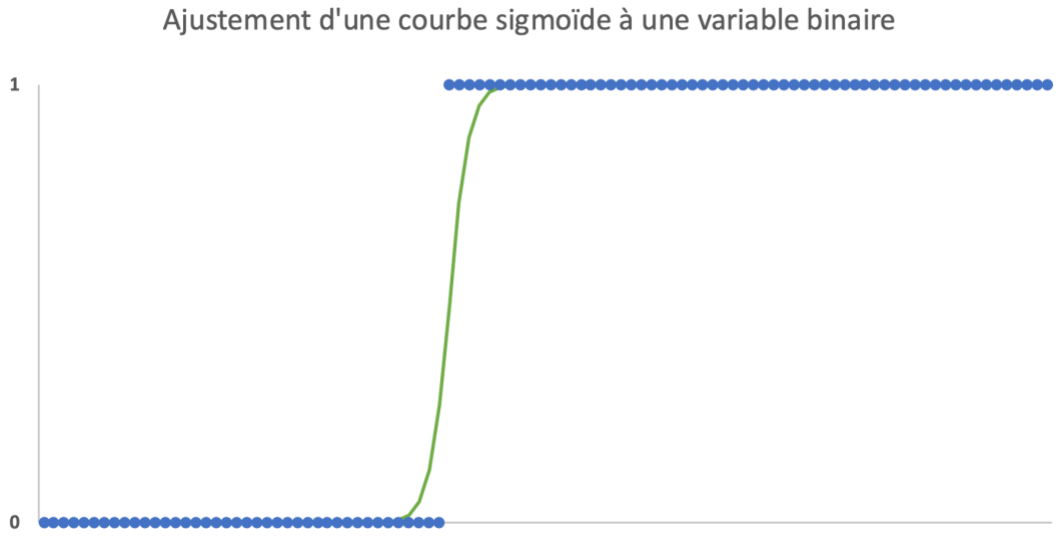


Figure 18 - Ajustement d'une courbe sigmoïde sur une variable binaire

[18] La régression logistique, Sonia NEJI et Anne-Hélène JIGOREL, 2013.

10.2.2. Cadre théorique général

Cette première approche empirique sur les données me permet de présenter le fondement de la régression logistique. Voyons désormais la formalisation mathématique. Premièrement la régression logistique n'est qu'un cas particulier des modèles linéaires généralisés, ainsi, elle profite des différentes propriétés de ce cadre théorique global, notamment :

- Un ensemble de variables explicatives qui peuvent être quantitatives ou qualitatives.
- L'estimation des coefficients par maximum de vraisemblance.
- L'utilisation de quantités comme la déviance, l'*AIC* ou le *BIC* afin de comparer les différents modèles.

On considère toujours une variable cible Y prenant deux valeurs : 0 ou 1 ainsi qu'un vecteur $X = (X_1, X_2, \dots, X_n)$ de n variables explicatives quantitatives ($PSAP_{dernière\ vision}$ ou l'ancienneté) ou qualitatives (Profession, nature juridique ou encore nature du dommage). On dispose également de n observations indépendantes.

L'objectif de tout modèle de régression est d'exprimer l'espérance conditionnelle de la variable cible Y en fonction des p variables explicatives. L'espérance conditionnelle de Y sachant X s'écrit :

$$E[Y|X = x] = 1 * P[Y = 1|X = x] + 0 * P[Y = 0|X = x] = P[Y = 1|X = x] = p(x)^{19}$$

Or on sait que la fonction de lien utilisée dans le cadre d'une répression logistique est la fonction $logit$ dont l'expression est la suivante :

$$logit(x) = log\left(\frac{x}{1-x}\right)\ pour\ tout\ x > 1$$

D'autres fonctions de lien existent, on peut citer par exemple la fonction $probit$, cependant dans le cadre de notre étude la fonction $logit$ suffit amplement pour les usages que l'on veut en faire.

Finalement, on souhaite trouver la combinaison de paramètres $\Theta = (\theta_0, \theta_2, \dots, \theta_n)$ optimale tel que :

$$log\left(\frac{p(x)}{1-p(x)}\right) = X\Theta = \theta_0 + \sum_{i=1}^{n} \theta_i X_i$$

En résolvant l'équation on obtient :

$$p(x) = \frac{1}{1 + e^{-\theta_0 - \sum_{i=1}^{n} \theta_i X_i}}$$

Par la suite, l'objectif de la régression est d'estimer les paramètres Θ optimaux par maximum de vraisemblance.

[19] Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage, Yufei LUO

10.2.3. Estimation des paramètres Θ

Par définition la vraisemblance d'un échantillon de n observations i.i.d. est la quantité suivante :

$$\mathcal{L}(\Theta) = f(y_1|X = x_1; \Theta) \dots f(y_n|X = x_n; \Theta)^{19}$$

C'est-à-dire,

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} f(y_i|X = x_i; \Theta) \Leftrightarrow \mathcal{L}(\Theta) = \prod_{i=1}^{n} \mathbb{P}(y_i|X = x_i; \Theta)$$

On s'aide des notations vues précédemment : une observation y_i est la réalisation d'une variable de Bernoulli qui prend comme valeur 1 avec une probabilité $\mathbb{P}(y_i|X = x_i) = p(x_i)$. On réécrit la vraisemblance ainsi :

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} p(x_i)^{\,y^i} (1 - p(x_i))^{\,1-y^i}$$

En remplaçant $p(x_i)$ par l'expression obtenue dans la section 10.2.2 on obtient :

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\theta_0 - \sum_{i=1}^{n} \theta_i X_i}}\right)^{y^i} \left(1 - \frac{1}{1 + e^{-\theta_0 - \sum_{i=1}^{n} \theta_i X_i}}\right)^{1-y_i}$$

Il n'existe pas de solution directe de cette équation, ainsi pour maximiser la log-vraisemblance, on utilisera des méthodes numériques itératives à l'image de la méthode de Newton-Raphson.

10.2.4. *Odds* et *Odds-ratio*

Une des particularités propres à la régression linéaire est la possibilité de calculer les *odds* et les *odds-ratios*, en français rapport de côtes, et ainsi quantifier l'effet d'un facteur. En notant $P[Y = 1|X = x] = p(x)$ on peut écrire l'*odds*[19] comme :

$$Odds = \frac{p(x)}{1 - p(x)}$$

A partir des *odds*, on peut calculer l'*odds-ratio*[19] d'une variable explicative X en fonction de x modalités comme suit :

$$Odds - ratio = \frac{Odds_{x+1}}{Odds_x} = \frac{\frac{p(x+1)}{1 - p(x+1)}}{\frac{p(x)}{1 - p(x)}}$$

Avec :

- $Odds_{x+1}$: la quantité *odds* calculée pour la variable X en prenant en compte $x + 1$ modalités.
- $Odds_x$: la quantité *odds* calculée pour la variable X en prenant en compte x modalités.

10.3. Avantages et points d'attention des modèles de régression

10.3.1. Avantages

Les modèles de régression présentent de nombreux avantages[19] tant sur leur mise en œuvre que sur l'exploitation de leurs résultats. On peut citer par exemple :

- L'approche paramétrique de ces modèles qui permet par la suite d'obtenir un coefficient par modalité et ainsi de quantifier l'effet de chaque variable, de chaque modalité sur le résultat final de la régression. Cette particularité est un réel atout notamment pour communiquer les résultats, cela permet d'apporter une certaine transparence sur le modèle et de vérifier des constats logiques sur les données.
- La sélection des variables, pas à pas, est un autre avantage des modèles de régression. La méthode est directement implémentée sur les différents langages de programmation et propose 3 manières d'effectuer la sélection : *forward*, *backward* et *stepwise*. La sélection *forward* débute avec un modèle simpliste ne comportant que l'*intercept,* puis à chaque itération, une variable est ajoutée jusqu'à épuisement des variables significatives. En opposition, on retrouve la sélection *backward* qui part d'un modèle complet, comprenant toutes les variables explicatives disponibles, puis à chaque itération on supprime la variable qui n'est pas significative jusqu'à épuisement des variables non significatives. La méthode *stepwise* quant à elle propose d'utiliser les deux méthodes citées juste avant. La fonction qui permet d'effectuer cette sélection sous R est la fonction ***stepAIC*** qui base sa sélection sur l'*AIC* global du modèle. La méthode par défaut dans la fonction ***stepAIC*** de R est la méthode *backward*. Évidemment, la sélection automatique des variables ne se substitue pas à un regroupement fin des modalités.

10.3.2. Points d'attention

J'ai cité deux avantages importants des modèles de régression, voyons désormais quels sont les points d'attention sur ces modèles :

- Premièrement, les données doivent être pré-traitées : découpage des variables continues en classes afin d'assurer la robustesse du modèle, regroupement de modalités pour les variables qualitatives, étude des corrélations entre variables explicatives quantitatives et qualitatives ou encore traitement des valeurs manquantes. Cette étape peut être longue, mais elle est déterminante afin de s'assurer d'avoir des modèles pertinents.
- Deuxièmement, bien que l'approche paramétrique ait été citée comme un avantage, elle peut également être citée parmi les points d'attention. En effet, cette approche suggère que l'on doit trouver la loi adéquate à la modélisation. Trouver la loi selon laquelle sont distribuées les observations n'est pas toujours évident, particulièrement pour les variables continues (par exemple : le coût des sinistres).

11. Arbre CART

La première étape de notre étude consistait à convaincre que l'utilisation d'algorithme de *Machine Learning* permettrait d'améliorer les prédictions sur la base d'une modélisation similaire à ce qui était fait auparavant. L'algorithme choisi est celui de *Gradient Boosting*. Connu pour ses performances et sa rapidité qui sont très appréciées lors des différentes compétitions de *Machine Learning*, l'algorithme de *Boosting* permet la résolution de problèmes de classification et de régression.

Il est nécessaire de préciser que le *Gradient Boosting* n'est qu'une méthode de réalisation des modèles qui est imputée à un algorithme qui peut être : une régression linéaire, un modèle linéaire généralisé ou encore les arbres *CART*. Dans la suite de cette étude, il a été choisi d'appliquer le *Gradient Boosting* aux arbres *CART* afin de proposer une nouvelle manière de modéliser nos variables cibles.

11.1. Cadre théorique

11.1.1. Définition

Les arbres de régression et de classification (en anglais ***C**lassification **A**nd **R**egression **T**rees ou **CART***) ont été introduits par le statisticien Breiman et al.[20], au cours de travaux au sein de l'Université de Californie. Les arbres de classification et de régression se distinguent par la nature de la variable à expliquer. Si cette variable est qualitative, on utilisera alors un arbre de classification et si la variable d'intérêt est quantitative, on utilisera un arbre de régression.

D'un point de vue mathématique, on dispose de m observations $n_1, n_2, \dots, n_m$ d'une variable Y et des valeurs de p caractères $X_1, X_2, \dots, X_p$. Pour tout $i \in \{1, \dots, m\}$, les valeurs associées à n_i sont notées $x_{1,i}, \dots, x_{p,i}$. En résumé voici comment les variables s'articulent entre elles :

	$\boldsymbol{X_1}$	…	$\boldsymbol{X_p}$	$\boldsymbol{Y}$
n_1	$x_{1,1}$	…	$x_{p,1}$	$\boldsymbol{y_1}$
…	…	…	…	…
n_m	$x_{1,\mathrm{m}}$	…	$x_{p,\mathrm{m}}$	$\boldsymbol{y_m}$

L'objectif des arbres de régression est de prédire la variable Y pour un individu n_i en connaissant les valeurs des covariables $X_1, X_2, \dots, X_p$.

11.1.2. Une approche théorique des arbres de régression

Voici un schéma reprenant le fonctionnement d'un arbre *CART* :

[20] Breiman et al. (1984)

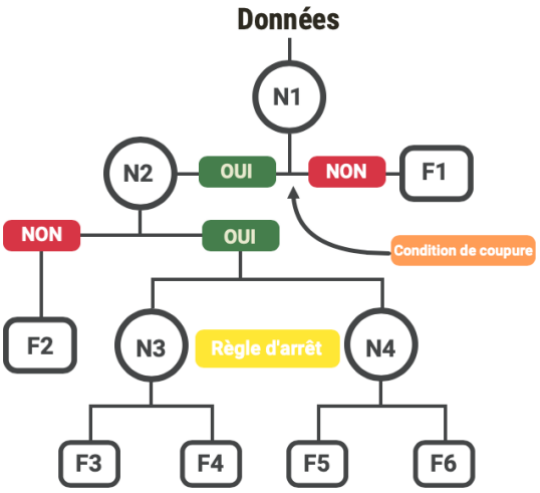


Figure 19 - Schéma du fonctionnement d'un CART

Le schéma suivant est un exemple d'arbre de décision. Lors de la première étape, on segmente les données initiales, ou racine, en deux groupes à l'aide la meilleure variable explicative (ou *split variable* en anglais) permettant cette division. Ces groupes sont appelés nœuds. L'arbre maximal pouvant être obtenu contient une observation par groupe ; cela se produit lorsque l'on ne spécifie pas de profondeur à l'arbre. Une fois l'arbre maximal obtenu, ou l'arbre avec la profondeur souhaitée fixée en amont, on parle de nœuds finaux ou de feuilles[21]. L'algorithme *CART* va pénaliser l'erreur d'ajustement d'un sous-arbre élagué à l'arbre maximal auquel on ajoute une pénalité sur le nombre de feuilles modulo un certain coefficient. Toujours pour respecter le principe de parcimonie, les arbres ayant le plus de feuilles et donc le plus de complexité seront pénalisés davantage.[22]

11.1.3. Les arbres de régression

Comme je l'ai expliqué un arbre de régression *CART* est un arbre d'aide à la décision quant à la valeur d'une variable Y à partir de la valeur des variables $X_1, X_2, \ldots, X_p$. La construction de cet arbre est fondée sur le partitionnement récursif des individus à partir des variables $X_1, X_2, \ldots, X_p$. Pour partitionner, il est nécessaire de couper l'arbre afin de créer des nœuds caractérisés par :

- La qualité des divisions
- Les règles d'arrêt
- Les conditions de coupure

Considérons que les caractères $X_1, X_2, \ldots, X_p$ sont numériques. Soit $i \in \{1, \ldots, p\}$ *et* $c \in \mathbb{R}$. Posons pour un nœud donné :

- $\overline{y_{inf}}$ la moyenne des valeurs de Y tel que $X_i < c$.
- $\overline{y_{sup}}$ la moyenne des valeurs de Y tel que $X_i \geq c$.
- $SS_{inf}(i, c)$ la somme des carrés des écarts entre les valeurs Y et $\overline{y_{inf}}$.
- $SS_{sup}(i, c)$ la somme des carrés des écarts entre les valeurs Y et $\overline{y_{sup}}$.

A partir de ces quantités on peut définir l'erreur globale commise lors du partitionnement des individus par rapport à c qui est donnée par :

[21] Machine Learning : Du *GLM* à l'arbre de CART en passant par le Random Forest, Periclès Group
[22] GENUER R., POGGI J.-M. [2017]

$$Err(i,c) = SS_{sup}(i,c) + SS_{inf}(i,c)$$[23]

L'objectif est de minimiser cette erreur à chaque coupure. On obtient donc comme condition de coupure : $X_{i*} \geq c_i{}^*$ où i^* et c^* minimisent l'erreur $Err(i,c)$. On appellera X_{i*} le caractère de coupure le plus déterminant de tous les caractères pour ce nœud et c_i* la valeur de seuil. Afin d'optimiser la condition de coupure l'arbre fonctionne par apprentissage automatique en calculant l'indice d'amélioration[22] I défini par :

$$I = 1 - \frac{Err(i^*, c^*)}{E}$$

11.1.4. Mesure de la qualité des nœuds à l'aide des fonctions d'hétérogénéité

La construction de l'arbre de classification passe par une phase d'apprentissage dont l'objectif est de minimiser l'hétérogénéité dans les classes. L'introduction d'un critère de division nous amène à introduire les fonctions d'hétérogénéité. Ces fonctions sont positives et prennent la valeur de 0 dans le cadre d'un nœud homogène (i.e. pour ces observations présentes dans le nœud, on observe la même valeur de Y) et, a contrario, maximale dans le cas d'une dispersion importante observée sur la variable cible. Ces fonctions diffèrent selon la nature de la variable d'intérêt[24].

11.1.4.1. Variables d'intérêt qualitatives

Soit une variable d'intérêt qualitative à m modalités notées $\{m_1, \ldots, m_k\}$. En pratique, on choisit généralement comme fonction d'hétérogénéité lors de l'estimation d'une variable qualitative :

- L'indice de Gini :

$$-2\sum_{i=1}^{m} |N| p_N^i \log\left(p_N^i\right)$$

Avec p_N^i est la proportion de la modalité i de Y dans le nœud N.

- La fonction d'entropie :

$$\sum_{i=1}^{m} p_N^i \left(1 - p_N^i\right)$$

On peut également considérer la règle de Bayes comme critère de maximisation de la décroissance de l'hétérogénéité dans le cadre d'une variable Y qualitative. Pour cela, il suffit de connaître les coûts de mauvais classement et de minimiser l'erreur de Bayes.

11.1.4.2. Variable d'intérêt quantitative

La fonction d'hétérogénéité pour une variable continue correspond à la variance intra-classe. Pour un nœud N, la variance du nœud est donnée par :

---

[23] CHESNEAU C. [2020]
[24] Université de Toulouse, « Arbre binaire de décision »

$$V_N = \frac{1}{|N|}\sum_{i \in N}(y_i - \overline{y_N})^2$$

Avec $|N|$ le nombre d'observations présentes dans le nœud N.

Plus précisément, lors du découpage de l'échantillon initial en deux nœuds « fils » N_G (à gauche) et N_D(à droite) l'arbre *CART* cherche à établir la variable ainsi que la règle qui maximisera la décroissance d'hétérogénéité entre les nœuds N_G et N_D. Autrement dit, on cherchera à minimiser la quantité suivante :

$$V_{intra} = \frac{1}{n}\sum_{i \in N_G}\left(y_i - \overline{y_{N_G}}\right)^2 + \frac{1}{n}\sum_{i \in N_D}\left(y_i - \overline{y_{N_D}}\right)^2$$

11.1.5. Algorithme récursif de création de nœud

Lors de la division d'un nœud, on souhaite maximiser la décroissance d'hétérogénéité. On va donc essayer toutes les divisions d possibles pour obtenir les deux nœuds fils N_G et N_D afin de trouver la division d^* maximisant la décroissance d'hétérogénéité. Finalement, cette démarche se résume en une procédure itérative qui est la suivante :

- **Étape 1 :** on détermine l'ensembles des divisions possibles qui à partir des variables explicatives $(X_1, X_2, \dots, X_p)$ nous permet d'obtenir les deux nœuds fils N_G et N_D.
- **Étape 2 :** on calcule les valeurs de la fonction d'hétérogénéité.
- **Étape 3 :** on retient la division d^* qui maximise la décroissance de l'hétérogénéité.
- **Étape 4 :** segmentation

On divise le nœud initial en deux nœuds fils N_G et N_D. On réitère ensuite la procédure en l'appliquant à chaque nœud fils et ainsi de suite.

Si la variable d'intérêt est qualitative ordinale (avec m modalités) ou quantitative alors le nombre de divisions binaires admissibles (*i.e. sans aucun nœuds vide)* est de $m - 1$. Toutefois, pour une variable d'intérêt qualitative non-ordinale à m modalités, le nombre de divisions admissibles est de $2^{m-1} - 1$.

11.1.6. Les règles d'arrêt

Une dernière question se pose : quand est-ce que l'arbre se termine ? On pourrait aller jusqu'à obtenir une observation par feuille, mais cela induirait une prédiction peu pertinente. Voici quelles sont en pratique les règles d'arrêt les plus utilisées, on peut en combiner plusieurs parmi :

- La profondeur de l'arbre, c'est-à-dire le nombre de niveaux maximums souhaités.
- Le nombre minimum d'individus présents pour permettre la formation du nœud.
- Le nombre minimum d'observations à avoir dans chaque feuille finale.
- La valeur du paramètre de complexité notée cp. Ce paramètre correspond à l'amélioration minimale du modèle nécessaire lors de la coupure de chaque nœud.

11.2. L'optimisation des arbres

L'optimisation des arbres est une étape nécessaire afin d'obtenir les prédictions les plus précises possibles. Cette étape s'appelle le *pruning*, ou élagage en français, et elle permet de réduire le risque d'*overfitting* (c'est-à-dire de sur-apprentissage) en vérifiant l'apport de chaque nœud dans la prédiction. Les nœuds dont l'utilité n'est pas avérée sont supprimés transformant ainsi les données en feuille. L'importance de cette étape réside dans le fait que les arbres de régression sont l'un des algorithmes de *Machine Learning* les plus sujets au sur-apprentissage.

Contrairement à ce que l'on pourrait penser, lorsqu'un modèle est sur-ajusté sur les données, c'est-à-dire lorsque ces prédictions reflètent exactement la réalité, cela n'est pas un avantage. En effet, dès lors qu'une observation future est un peu différente ou tout simplement si l'on ajoute de nouvelles observations le modèle ne saura pas les prédire correctement. Le meilleur moyen de réduire le sur-apprentissage reste l'élagage des arbres, il existe deux techniques permettant d'effectuer cette opération : le *pre-pruning* et le *post-pruning*.

11.2.1. Le *pruning* ou *post-pruning*

Cette technique est sûrement la plus employée, une fois l'arbre *CART* créé, on va supprimer certaines branches, autrement dit élaguer l'arbre, afin de lutter contre le sur-apprentissage. Pour élaguer un arbre, on peut considérer deux approches :

- Minimisation de l'erreur : on élague l'arbre jusqu'à ce que l'erreur de *cross-validation* soit la plus petite possible. La *cross-validation* est un procédé qui permet de *sampler* l'échantillon de données en deux paquets en construisant l'arbre avec une partie des données et en testant sur l'autre partie son pouvoir de prédiction. Dans Breiman et al.[20], le *pruning* est réalisé en calculant l'Error-Complexity, cette quantité est définie ainsi :

$$EC_\alpha = \sum_{i=1}^{n} \left(\varphi(X_i) - \hat{T}\ (X_i)\right)^2 + \frac{\alpha K(R)}{n}$$

  Avec α un paramètre de complexité choisit par validation croisée, $\frac{K(R)}{n}$ la quantité d'observations obtenues dans le sous-ensemble partitionné à l'aide des règles R et $\sum_{i=1}^{n} \left(\varphi(X_i) - \hat{T}\ (X_i)\right)^2$ l'erreur de prédiction commise par l'arbre T.

- L'arbre le plus petit : on élague l'arbre en considérant un niveau supplémentaire par rapport à l'arbre élagué avec l'approche précédente. Cela permet d'obtenir un arbre légèrement plus important, et donc plus intéressant à interpréter, pour une augmentation de l'erreur raisonnable.

11.2.2. Le *pre-pruning* ou *early-stopping*

Cette méthode d'élagage survient au moment de la formation de l'arbre de décision : avant que l'arbre ne produise des feuilles avec peu d'observations, cela afin d'éviter que la prédiction soit trop proche de la réalité et augmente ainsi le sur-apprentissage.

A chaque division l'erreur de cross-validation est calculée : si celle-ci ne décroît pas de manière satisfaisante alors la division n'est pas réalisée et l'arbre s'arrête au niveau précédent cette division. Cette méthode reste toutefois peu utilisée, car elle peut induire du sous-apprentissage. A l'instar du sur-apprentissage, le sous-apprentissage arrive lorsque l'arbre n'est pas assez entraîné sur les données et donc ne prédit pas correctement la variable Y. Les méthodes d'*early-stopping* et de *pruning* peuvent être utilisées ensemble afin d'optimiser le coût du calcul.

11.3. Avantage des arbres de régression et de classification

Les arbres *CART* disposent de nombreux avantages :

- Ces arbres sont faciles à interpréter et à comprendre même pour un public n'ayant pas un bagage statistique important. Ces arbres sont donc beaucoup plus pratiques pour la communication des résultats.
- Les variables $X_1, X_2, \ldots, X_p$ peuvent être de toute sorte : variables catégorielles ou variables numériques.
- Les variables n'ont pas besoin d'être préparées ou d'être traitées en amont. En effet, les arbres de régression sont tout à fait capables de gérer les données manquantes ou les corrélations entres les variables explicatives, grâce à l'apprentissage automatique.
- Cette méthode est performante sur des jeux de données importants notamment dans le cadre du *Big Data*.

Comparativement aux modèles de régression linéaire ou non-linéaire classiques, comme les modèles *GLM*, les arbres *CART* présentent d'autres intérêts particuliers notamment :

- Une approche simplifiée et directe : toutes les valeurs issues de la modélisation sont directement disponibles et visibles. Notamment le seuil déterminant pour la classification de la variable ou encore la valeur prédite de la variable à expliquer. Pour les régressions linéaires, il faut encore observer le modèle, choisir les variables à conserver et enfin prédire la variable Y.
- Dans le cas des modèles de régression ; la droite d'équation doit être linéaire ce qui n'est pas le cas avec les arbres *CART*. En effet, la structure liant les variables importe peu : elle peut être linéaire ou non.
- Il n'y a aucune hypothèse mathématique derrière les modèles *CART* comme cela peut être le cas pour les modèles de régression (normalité des résidus par exemple).
- Les arbres *CART* sont autonomes dans le fait de gérer les corrélations. Dans les modèles *GLM* par exemple, l'étude des corrélations est une partie à ne pas négliger qui peut avoir un impact important sur les résultats du modèle, ce qui n'est pas le cas avec les arbres *CART*.

Pour conclure, les arbres *CART* présentent de nombreux avantages techniques mais également opérationnels : ils sont rapides et faciles à mettre en place, mais également adaptés à la communication et compréhensibles par n'importe quel public. Toutefois, il ne faut pas négliger la logique sous-jacente aux arbres *CART* que je détaillerai dans la suite.

12. Modèle *XG-Boost*

12.1.1. Ensemble Learning

Comme je l'ai évoqué dans l'introduction de la partie 11 le *Gradient Boosting*, et plus généralement le *Boosting*, est une méthode d'entraînement des modèles dont le principe est de combiner les prédictions de plusieurs modèles faibles afin d'obtenir une prédiction plus forte. Autrement dit, on combine plusieurs *weak learners* en associant à chaque prédiction un poids en fonction de sa pertinence par rapport à la variable cible. Cela me permettra d'obtenir une meilleure prédiction en combinant tous les *weak learners*. Cette méthode permet d'obtenir une prédiction finale plus robuste et plus précise.

On dénombre deux grandes méthodes d'*Ensemble Learning* : le *bagging* et le *boosting*.[25] Ce qui différencie la méthode de *bagging*, comparativement à la méthode de *boosting,* est son approche : parallèle pour le *bagging* et plutôt séquentielle pour le *boosting*.

12.1.1.1. Le *bagging* : Définition

Le *bagging* appartient aux méthodes dites d'*Ensemble Learning* dont l'objectif est d'entraîner un nombre important de modèles appelés *weak learners.* Pour obtenir ces modèles dans le cadre d'une approche parallèle, il est nécessaire de les entraîner sur des données indépendantes. Pour créer une multitude d'échantillons de données indépendants les uns des autres, mais issus d'une même base initiale on a recours au *bootstrap*. La méthode de *bootstrap* va créer un nouvel échantillon en choisissant aléatoirement chaque observation de l'échantillon initial : une même observation pouvant être choisie plusieurs fois. On entraîne ensuite un modèle sur chacun des échantillons obtenus. Pour combiner ensuite les *weak learners* entraînés sur chaque échantillon, la technique diffère selon la nature de la variable d'intérêt :

- Si la variable d'intérêt est quantitative alors la prédiction finale sera obtenue en effectuant la moyenne des *weak learners*.
- Si la variable d'intérêt est qualitative, alors la prédiction finale sera obtenue en choisissant la modalité prédite est la plus représentée entre les modèles.

Le schéma simplifié ci-dessous résume la méthode de *bagging*, on note Y la variable à expliquer et X un ensemble de représentations d'une variable explicative :

[25] XG-Boost, blent.ia

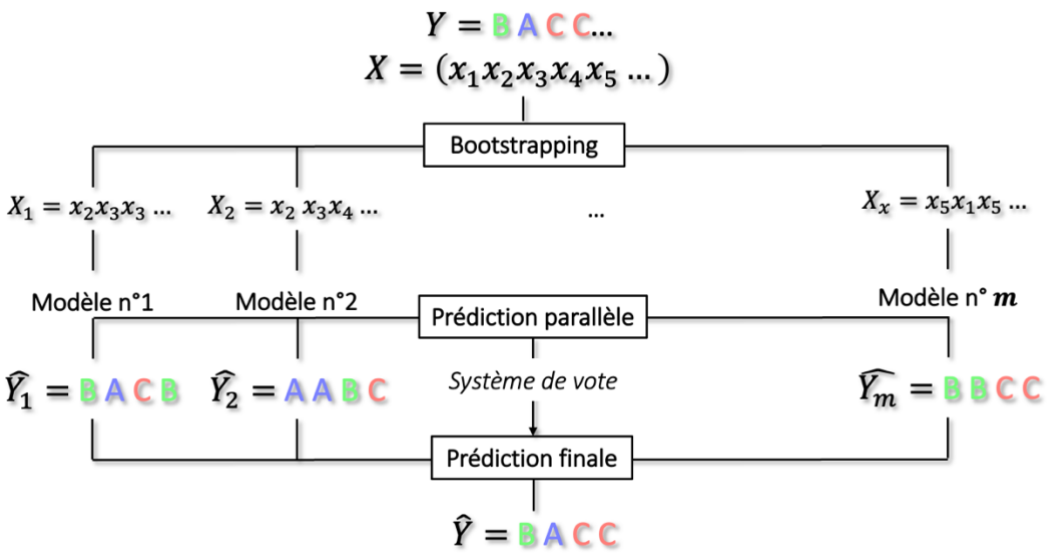


Figure 20 - Schéma explicatif du bagging

La méthode du *bagging* présente deux avantages :

- Avoir recours au *bootstrapping* crée des échantillons tous différents, mais ayant des données en commun. Ainsi, bien que l'on entraîne chaque modèle sur un échantillon différent, ceux-ci présentent de fortes similitudes. Cette particularité tend à augmenter le biais entrainant par construction une diminution de la variance.
- Comme on le voit sur le schéma simpliste ci-dessus la prédiction finale est obtenue en observant dans la prédiction de chaque modèle quelle est la modalité la plus représentée. Ainsi, même si aucun des modèles n'a réussi à prédire correctement la variable Y la prédiction finale obtenue $\hat{Y}$ est tout de même la bonne.

Un des algorithmes les plus connus faisant appel au *bagging* est l'algorithme *Random Forests*.

12.1.1.2. Le boosting : Définition

Contrairement à la méthode du *bagging* dont les modèles sont entraînés simultanément et indépendamment les uns des autres, la méthode du *boosting* va entraîner les modèles de façon itérative en créant une forte dépendance entre eux. Voici un schéma explicatif du fonctionnement de la méthode de boosting[25] :

- La 1ère étape consiste à entraîner un premier modèle sur les données. Chaque donnée dispose du même poids au départ et à l'issue de la prédiction. Les données mal prédites se verront attribuer un poids plus important :

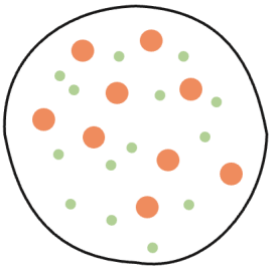


Figure 21 - Échantillon de données

- La 2ème étape consiste à construire un deuxième modèle dont l'objectif est de corriger les erreurs effectuées par le premier modèle en tenant compte de la pondération associée à l'étape 1. On répète ensuite cette étape autant de fois que nécessaire : soit jusqu'à ce que toutes les observations soient classées correctement, soit si l'on atteint le nombre maximal d'arbres à entraîner. Chaque nouveau modèle est entraîné sur la base de la prédiction du modèle précédent : c'est l'approche séquentielle.

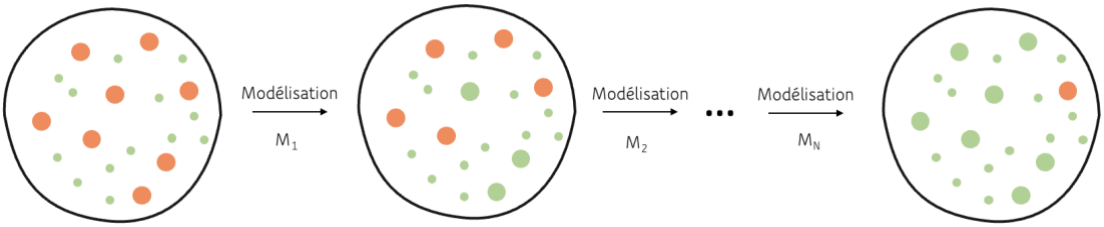


Figure 22 - Schéma du fonctionnement de la méthode de Boosting

- La 3ème et dernière étape, consiste à sommer tous les résultats des modèles pondérées de leur poids afin d'obtenir la prédiction finale :

$$Prédiction\ finale = \sum_{i=1}^{N} \omega M_i$$

Malgré leurs différences le *bagging* et le *boosting* présentent des similitudes notamment un échantillonnage de la base initiale en plusieurs bases d'apprentissage, une agrégation des prédictions de chaque modèle pour obtenir la prédiction finale, mais également une réduction de la variance. Ces similarités peuvent questionner quant à la méthode à choisir dans le cadre d'un problème. Tout dépend de l'objectif final :

- Si le modèle initial présente du sur-apprentissage, il sera préférable d'avoir recours au *bagging*
- Si le modèle initial ne présente pas des prédictions satisfaisantes, l'algorithme de *boosting* sera plus adapté étant donné son action sur les erreurs du modèle. Il existe différents types d'algorithmes de *boosting*, on peut citer par exemple : *l'Adaboost*, le *Gradient Boosting* et le *XG-Boost*.

12.1.2. Les différents algorithmes de *boosting*

12.1.2.1. L'Adaboost

Introduit en 1997 dans Freund et al.[26], l'Adaptive *Boosting* (plus connu sous le nom *d'Adaboost*) est le premier modèle de *boosting* qui a été introduit. L'idée générale du *boosting* repose sur la minimisation d'une fonction de perte qui, dans le cadre de *l'Adaboost*, est une fonction de perte exponentielle. En effet, dans Freund et al.[26] voici comment est décrite l'approche mathématique de l'algorithme *Adaboost* :

- **Étape n°1** : Pour un échantillon E de taille N, on calcule les poids initiaux $\omega_i^m = \frac{1}{N}$.
- **Étape n°2** : Les modèles *Adaboost* ne sont pas entraînés directement sur l'échantillon initial E, mais sur une base d'apprentissage obtenue depuis E. La présence des poids ω_i^m implique que chaque observation à la même probabilité d'être choisie dans la base d'apprentissage. Cet échantillonnage est effectué pour tout m allant de 1 à M. On obtient ainsi pour l'itération m une base d'apprentissage contenant x_i observations.
- **Étape n°3** : Entraîner un modèle pour chaque itération m noté f_m. f peut être une régression linéaire, un *Support Vector Machine (SVM)* ou un arbre de décision.
- **Étape n°4** : Pour chaque itération m, l'algorithme va fonctionner de manière séquentielle en ajustant les poids ω_i^m en fonction des erreurs commises par le modèle précédent. Ainsi, comme nous l'avons décrit dans la partie 12.1.1.2 le poids des observations x_i mal prédites sera plus important et celles des observations x_i correctement prédites sera réduit. La mise à jour de ces poids est effectuée en calculant premièrement, pour l'itération m, la quantité suivante :

$$\varepsilon_m = \frac{\sum_{y_i \neq f_m(x_i)} \omega_i^m}{\sum_{y_i} \omega_i^m}$$

Avec y_i la valeur observée de la variable d'intérêt y.

A l'aide de ε_m on peut ensuite calculer α_m représentant la confiance accordée à la prédiction de f_m :

$$\alpha_m = \frac{1}{2} ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$$

Plus ε_m est faible plus la quantité α_m sera importante.

Les poids obtenus pour l'itération $m+1$ sont décrits comme :

$$\omega_i^{m+1} = \omega_i^m e^{-\alpha_m y f_m(x)}$$

Précisons que l'algorithme *Adaboost* ne fonctionne pas de manière binaire avec une variable d'intérêt y qui prendrait comme valeur $\{0,1\}$ mais avec une variable y prenant comme valeur $\{-1,1\}$.

[26] Freund et al. "A Short Introduction to Boosting", 1999.

A fortiori, les prédictions de l'algorithme représentées par $f_m(x_i)$ prendront comme valeur $\{-1,1\}$. Cette particularité à toute son importance lors de la mise à jour des poids ω_i^{m+1}en effet :

- Si $f_m(x_i)$ prédit correctement y_i alors $f_m(x_i) = 1$ et $y_i = 1$ ou $f_m(x_i) = -1$ et $y_i = -1$. Finalement, on aura : $yf_m(x) = 1$. Dans ce cas le coefficient $-\alpha_m y f_m(x)$ sera négatif et *a fortiori* le poids ω_i^m sera réduit de manière exponentielle à la prochaine itération.
- Si $f_m(x_i)$ ne prédit pas correctement y_i alors $f_m(x_i) = 1$ et $y_i = -1$ ou $f_m(x_i) = -1$ et $y_i = 1$. Finalement, on aura : $yf_m(x) = -1$. Dans ce cas le coefficient $-\alpha_m y f_m(x)$ sera positif et *a fortiori* le poids ω_i^m sera augmenté de manière exponentielle à la prochaine itération.

- **Étape n°5** : les prédictions finales notées $F(x)$ sont obtenues en calculant :

$$F(x) = sign\left(\sum_{m=1}^{M} \alpha_m f_m(x)\right)$$

12.1.2.2. Le Gradient *Boosting*

Lorsque les erreurs sont minimisées à l'aide de l'algorithme de descente de gradient, on parle de *Gradient Boosting*. L'algorithme de descente de gradient[27] repose sur la minimisation d'une fonction de coût convexe (par exemple l'erreur quadratique moyenne). La formule générale d'un algorithme de descente est la suivante :

$$y_{t+1} = y_t - \lambda \Delta y_t$$

Avec λ le taux d'apprentissage et Δy_t la direction de descente. On applique cette formule à une fonction f convexe et on obtient :

$$f(x_{t+1}) = f(x_t) - \lambda \Delta f(x_t)$$

La direction de descente devient donc l'opposé du gradient : -Δf. Le gradient d'une fonction indique la direction où la pente est la plus lente, prendre l'opposé du gradient revient donc à prendre la direction où la pente est la plus raide.

Voici le programme simplifié du fonctionnement de l'algorithme de descente de gradient :

- **Étape 1** : on initialise l'algorithme à un point x_0 et on choisit un taux d'apprentissage λ. Bien que cette première étape semble être la plus simple, elle est en réalité cruciale pour la suite de l'algorithme. Des méthodes existent afin de choisir correctement le taux d'apprentissage et le point d'initialisation.
- **Étape 2** : Calcul de $f(x_0)$.
- **Étape 3** : On calcule la nouvelle coordonnée $f(x_1)$ à partir de la formule $f(x_{t+1}) = f(x_t) - \lambda \Delta f(x_t)$.
- **Étape 4** : On réitère l'opération jusqu'à ce que la condition d'arrêt soit vérifiée i.e. jusqu'à ce que $|f(x_{t+1}) - f(x_t)| \leq \varepsilon$ avec un ε généralement choisi très petit.

[27] Descente de gradient, datascientest.com

12.1.2.3. Le *XG-Boost*

Le modèle ***eXtreme Gradient BOOSTing*** diffère du *Gradient Boosting* de différentes façons :

- Premièrement, c'est un algorithme qui se veut plus rapide et entraînable sur plusieurs cœurs.
- Deuxièmement, il propose des régularisations en norme L^1 et L^2.

Enfin, il propose d'utiliser les dérivées partielles du second ordre comme approximation afin de recueillir le maximum d'informations sur la direction gradient.

La formule qui fonde le *XG-Boost* est la suivante :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)^{26}$$

C'est-à-dire qu'à chaque itération t on ajoute f_t afin de minimiser$\mathcal{L}^{(t)}$. Comme cité plus haut, on peut également faire intervenir les dérivées partielles du second ordre :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + \delta_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) f_t(x_i) + \frac{1}{2}\delta^{\,2}_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) f_t^{\,2}(x_i)) + \Omega(f_t)^{28}$$

On implémentera le modèle *XG-Boost* via le package ***XG-Boost*** disponible sur R.

Après l'algorithme *XG-Boost* d'autres algorithmes de *boosting* de gradient seront introduits :

- Le premier en 2016 sous le nom de *Light GBM*[29]. Cet algorithme, mis au point par Microsoft, diffère principalement de l'algorithme *XG-Boost* par l'introduction du *Gradient-based One-Side Sampling* (ou *GOSS*). La méthode *GOSS* permet d'échantillonner les données en conservant les observations présentant les gradients les plus importants, i.e. les erreurs les plus importantes, et de ne choisir aléatoirement que des observations parmi les plus faibles gradients. L'idée derrière est simple, les observations qui présentent les erreurs les plus importantes seront les plus déterminantes au moment du *split*. Cette méthode permet d'entraîner les modèles sur des bases d'apprentissages plus faibles et donc de gagner en temps de calcul sans dégradation de la performance. D'autres différences existent entre le *Light GBM* et le *XG-Boost* que nous ne détaillerons pas dans ce document, mais des articles comme Ke et al.[29] (2017) proposent une présentation plus complète de l'algorithme.
- En 2017, c'est l'algorithme *CatBoost* qui, comme son nom l'indique, permet le traitement des variables catégorielles sans passer par une transformation en *dummies*. L'algorithme permet également de prendre en compte des variables explicatives composées de texte ou groupe de mots. Autre particularité, les arbres prédits par l'algorithme *CatBoost* sont symétriques : autrement dit à chaque profondeur atteinte, c'est la même condition qui est utilisée sur l'ensemble du niveau. A contrario pour les modèles *XG-Boost* et *Light GBM* l'arbre peut faire appel à différentes *splitting* variables sur un même niveau. Les arbres symétriques garantissent un meilleur contrôle du sur-apprentissage et une prédiction plus rapide.

[28] CHEN T., GUESTRIN C. (2016)
[29] Ke et al. (2017)

12.1.3. *XG-Boost* : les avantages

L'algorithme *XG-Boost* présente de nombreux avantages :

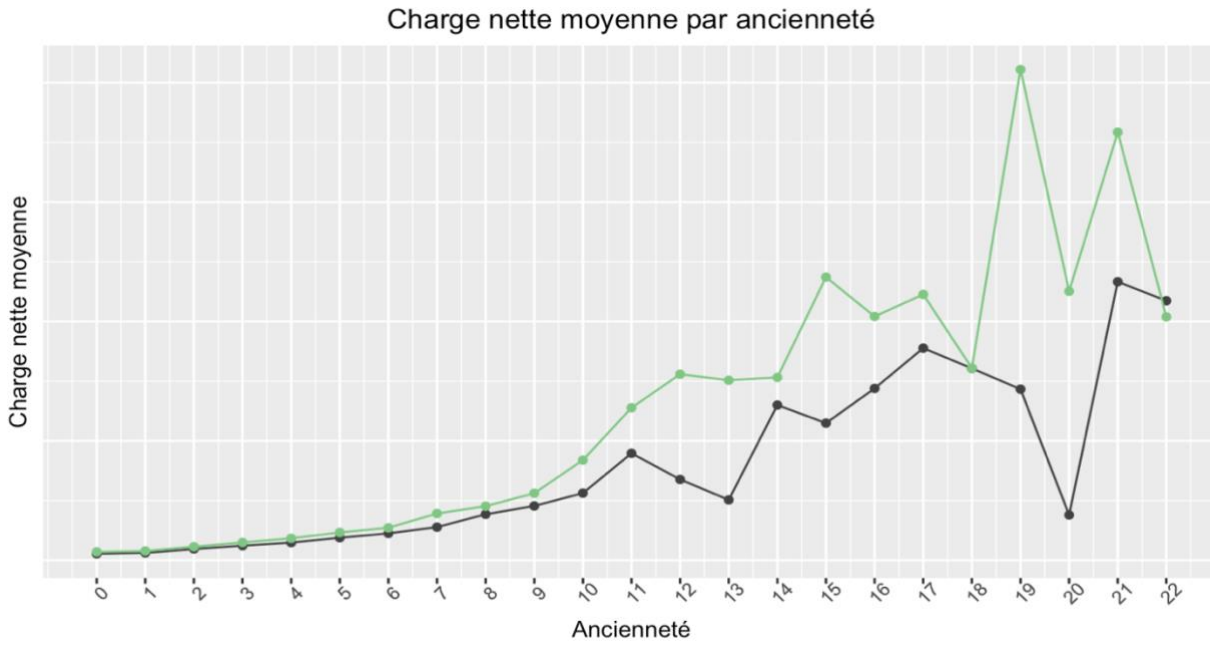
- La possibilité d'appliquer une régularisation de Lasso L^1 ou une régularisation de Ridge L^2. Ces régularisations, directement implémentées dans l'algorithme *XG-Boost,* peuvent se révéler particulièrement utiles dans le cas de sur-apprentissage.
- Le modèle *XG-Boost* permet également d'entraîner les modèles en parallèle de façon à utiliser le maximum de cœurs disponibles sur la machine.
- Comme les autres algorithmes de *Machine Learning*, le *XG-Boost* permet le traitement des valeurs manquantes. Pour cela, l'algorithme va affecter premièrement la valeur manquante au nœud de droite N_D, puis au nœud de gauche N_G, et observe dans quel cas l'erreur de prédiction est réduite. Ce mécanisme est appliqué à la fois en cas de présence de valeurs manquantes dans la base d'apprentissage ou dans la base de test.
- La validation croisée : comme nous le verrons dans la partie pour optimiser le premier paramètre qui est le nombre optimal d'itérations *nrounds*, nous procédons par validation croisée en choisissant au départ un nombre important d'itérations. La validation croisée permet ainsi d'optimiser ce premier paramètre de manière efficace et rapide. A contrario, la méthode de *Gradient Boosting* nécessite de mettre en place un *grid search* coûteux en temps d'exécution même pour optimiser ce seul paramètre.
- Le *post-pruning* : lors de l'entraînement de l'algorithme *XG-Boost,* on spécifie la valeur *maxdepth*, le maximum de profondeur atteint par l'arme. Le modèle *XG-Boost* entraînera chaque arbre jusqu'au maximum que nous aurons implémenté puis procédera à l'élagage de chaque arbre afin d'obtenir les arbres optimaux. Cet élagage est effectué tant que l'erreur de prédiction ne s'améliore pas. Le *Gradient Boosting* fonctionne de manière différente : il construit un arbre jusqu'à ce que la fonction de perte soit négative. Ainsi le *Gradient Boosting* peut s'arrêter prématurément au niveau d'un *split* présentant une perte négative tandis que le *split* suivant aurait apporté une perte positive qui aurait compensé cette perte négative.

Enfin, comparé à des techniques de *Machine Learning* plus populaires de nos jours comme les réseaux de neurones, le *XG-Boost* à l'avantage d'être le sujet de nombreux papiers de recherche. Ces recherches ont permis au fur et à mesure de construire des méthodes pour contrôler l'erreur du modèle *XG-Boost* ou encore pour implémenter des régularisations permettant de contrôler *l'overfitting*. Les réseaux de neurones quant à eux sont encore des « boîtes noires » pour lesquels il est difficile d'exprimer l'erreur du modèle et de mettre en place des mécanismes pour les contrôler. L'interprétation des résultats n'est également pas directe, a contrario un modèle *XG-Boost* appliqué à des arbres de régression *CART* nous permet de bénéficier du caractère intuitif des arbres, allié à une optimisation performante et que l'on peut contrôler.

13. Procédure « *weighted CART* »

13.1. Motivations

Bien qu'elle soit performante et une des méthodes les plus réputées pour résoudre des problèmes de classification ou de régression à l'aide de *Machine Learning*, le *XG-Boost* ne permet pas de tenir compte d'observations éventuellement censurées. En effet, cette méthode faisant intervenir des arbres *CART,* telle que présentée, ne permet pas de tenir compte des sinistres en cours dans la modélisation. Et pourtant, comme le montre l'histogramme disponible dans la partie 9.3, les observations censurées présentent des durées de liquidation importante : les sinistres anciens sont sur-représentés dans la base des sinistres en cours. Mais pourquoi la prise en compte de ces observations censurées devient-elle nécessaire pour estimer la provision ultime ? Simplement car la durée de vie d'un sinistre à un effet certain sur son niveau ultime et donc sur sa provision. Les sinistres les plus anciens coûtent plus cher que les sinistres récents. Cette affirmation se vérifie dans nos données lorsque l'on observe l'évolution de la charge nette moyenne par l'ancienneté des sinistres (en année) :



Toutes ces raisons nous poussent à considérer une méthode de provisionnement ligne à ligne permettant de tenir compte de cette censure à droite qui est récurrente dans ce genre de problème. Le choix des arbres *CART* n'est pas anodin, faire appel à des techniques de *Machine Learning* permet d'introduire des relations non-linéaires entre les variables explicatives et la variable cible. D'autres études faisant appel à des modèles de *Machine Learning* pour résoudre des problématiques de provisionnement existent, on peut citer par exemple les travaux de Wüthrich[30] proposant une méthode de provision ligne à ligne faisant intervenir des arbres *CART* ou encore les travaux de Kuo[31] avec la mise en place de réseaux de neurones pour prédire la provision ultime des sinistres *RBNS*.

[30] Wüthrich (2016)
[31] Kuo (2019)

Finalement, c'est sur la méthode weighted *CART* (définie par Lopez et al.) que les travaux se poursuivent. Plusieurs papiers de recherches abordent cette méthode :

- La procédure des arbres *CART* pondérés tenant compte de la censure à droite sont introduits dans Lopez et al.[32] : le cadre théorique est décliné avec notamment les preuves de consistance de l'approche.
- En 2019, Lopez et al.[33] propose une méthode permettant, à l'aide d'arbres *CART*, d'estimer la durée de vie résiduelle des observations censurées.
- En 2020 la méthode[34] est élargie afin de tenir compte des problèmes de troncature à gauche, autrement dit les sinistres tardifs, en plus de la gestion des observations censurées.

Nous déclinerons dans la suite les différentes composantes théoriques auxquelles la méthode fait appel. La théorie autour des arbres *CART* a été définie dans la partie 11.

13.2. Observations censurées : Définition

Reprenons les notations de Lopez et al.[32], et notons le vecteur (M, T, X) où $M \in \mathbb{R}^p$ représente le montant ultime des sinistres, $T \in \mathbb{R}^+$ la durée de vie du sinistre et $X = \left(X^{(1)}, X^{(2)}, \dots, X^{(d)}\right) \in \mathbb{R}^d$ un vecteur de variables explicatives permettant éventuellement d'expliquer les niveaux de Met T. La présence de sinistres en cours nous prive de l'observation complète des variables Met T. Les variables explicatives quant à elle sont toujours observées peu importe l'état du sinistre. Soit $C \in \mathbb{R}^+$une variable de censure. L'introduction de cette variable nous permet d'introduire trois nouvelles variables $Y \in \mathbb{R}^+, \delta \in (0,1)$ et $N \in \mathbb{R}^p$ tel que :

$$Y = \inf(T, C)$$
$$\delta = \mathbf{1}_{T \leq C}$$
$$N = \delta M$$

La valeur Y prendracomme valeur la durée de vie du sinistre pour les sinistres clos, et pour les sinistres en cours la durée de vie à date. Le coefficient δ prendra la valeur 0 pour les observations censurées : *a fortiori* la variable N prendra comme valeur 0 pour les sinistres en cours. Les variables (Y, N) sont observables pour toutes les observations contrairement aux variables (T, M). Même si cette construction suppose que M est le dernier paiement effectué sur le dossier, lorsque nous disposons d'informations sur les paiements de manière régulière, ceux-ci peuvent être intégrés par le biais de variables explicatives. J'utiliserai dans la suite la même construction, mais au lieu de considérer le montant final du sinistre M, je considère le montant final de la provision que l'on notera P. Le lien entre le montant final M et la provision P à l'aide de la formule suivante :

$$M = Règlements + P - Recours$$

[32] Lopez et al. (2016)
[33] Lopez et al. (2019)
[34] Lopez et al. (2020)

13.3. Poids de Kaplan-Meier

13.3.1. Estimateur de Kaplan-Meier

La volonté de prendre en compte les observations censurées pour prédire le montant final de la provision nous oblige à introduire des poids de Kaplan-Meier. La procédure de Kaplan-Meier permet d'estimer des modèles de durée à l'événement en tenant compte d'observations censurées. Dans le cadre de notre étude, la durée de survie est donc égale au temps qui s'écoule avant un événement qui est représenté par la clôture du sinistre. Mis au point par Edward L. Kaplan et Paul Meier[35], cet estimateur permet d'estimer la probabilité conditionnelle de survie pour chaque temps donné malgré la présence d'observations censurées. Par construction, s'il n'y a aucune censure observée dans l'échantillon alors l'estimateur de Kaplan-Meier sera identique à la fonction de survie.

La probabilité que le sinistre ne soit pas clos à un temps t_i peut s'écrire :

$$S(t_i) = \mathbb{P}(T > t_i)$$

En appliquant la formule de Bayes qui stipule que :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A).\mathbb{P}(A)}{\mathbb{P}(B)}$$

on obtient alors pour tout $t_i > t_{i-1} > \cdots > t_2 > t_1$:

$$\mathbb{P}(T > t_{i-1}|T > t_i) = \frac{\mathbb{P}(T > t_i|T > t_{i-1}).\mathbb{P}(T > t_{i-1})}{\mathbb{P}(T > t_i)}$$
$$\mathbb{P}(T > t_{i-1}|T > t_i).\mathbb{P}(T > t_i) = \mathbb{P}(T > t_i|T > t_{i-1}).\mathbb{P}(T > t_{i-1})$$

Or $\mathbb{P}(T > t_{i-1}|T > t_i) = 1$ car $t_i > t_{i-1}$, on a donc :

$$\mathbb{P}(T > t_i) = \mathbb{P}(T > t_i|T > t_{i-1})\mathbb{P}(T > t_{i-1})$$
$$\Leftrightarrow S(t_i) = \mathbb{P}(T > t_i|T > t_{i-1})S(t_{i-1})$$

En appliquant ensuite la formule de Bayes à $S(t_{i-1})$ on obtient :

$$S(t_i) = \mathbb{P}(T > t_i|T > t_{i-1})\mathbb{P}(T > t_{i-1}|T > t_{i-2})S(t_{i-2})$$
$$S(t_i) = \mathbb{P}(T > t_i|T > t_{i-1}) \dots \mathbb{P}(T > t_2|T > t_1)S(t_1)$$

avec T la durée de vie du sinistre au moment de sa clôture.

On note les durées de conditionnement choisies t_m qui nous permettent de réécrire le problème comme l'estimation des quantités :

$$\begin{cases} p_m = \mathbb{P}(T \geq t_{m+1}|T > t_m) \\ q_m = \mathbb{P}(T < t_{m+1}|T > t_m) \end{cases}$$

[35] Kaplan, Meier (1958)

La probabilité p_m représente la probabilité que le sinistre ne soit pas clos au temps t_{m+1}sachant qu'il n'était pas clos au temps t_m. A contrario, la probabilité q_m, complémentaire de la probabilité p_m, représente la probabilité que le sinistre soit clos au temps t_{m+1}sachant que le sinistre n'était pas clos au temps t_m. On peut alors réécrire la fonction $S(t)$ comme étant :

$$S(t) = p_0.p_1 \dots p_{t-1} = (1-q_0).(1-q_1) \dots (1-q_{t-1}) = \prod_{m=0}^{t-1}(1-q_m)$$

Les probabilités q_m sont ensuites estimées par maximum de vraisemblance. Pour un échantillon contenant n observations, le nombre de sinistres clos suit une loi Binomiale $\mathcal{B}(n, q_m)$ dont la vraisemblance est donnée par :

$$\mathcal{L}(q_1; \dots; q_m) = \prod_{m=1}^{j} \binom{k_m}{n_m} . i_m^{k_m} . (1-q_m)^{n_m - k_m}$$

Avec n_m le nombre de sinistres totaux juste avant le temps t_m et k_m le nombre de sinistres clos au temps t_m. En passant par une transformation log et en égalisant la dérivée du premier ordre à 0, comme il est coutume dans la maximisation de la vraisemblance, on obtient comme estimateur de q_m pour tout $m \in \{1, \dots, j\}$:

$$\widehat{q_m} = \frac{k_m}{n_m}$$

La fonction de survie $S(t)$ peut ensuite être estimée par :

$$\widehat{S(t)} = \prod_{m=0}^{t-1}\left(1 - \frac{k_m}{n_m}\right)$$

Finalement l'estimateur de Kaplan-Meier est donné par :

$$\widehat{q_m} = 1 - p_m = 1 - \widehat{S(t)} = 1 - \prod_{m=0}^{t-1}\left(1 - \frac{k_m}{n_m}\right)$$

13.3.2. Fonction de répartition de (M, T, X)

Comme il est stipulé dans Lopez et al.[32], on cherche à estimer la fonction de répartition suivante :

$$\mathbb{E}[\phi(M, \mathrm{T}, \mathrm{X})]$$

Or, nous n'observons pas les variables M et T pour les observations censurées. Nous devons donc trouver un estimateur de cette fonction de répartition prenant en compte la censure. L'introduction des poids de Kaplan-Meier est motivée par le fait que :

$$\mathbb{E}\left[\frac{\delta\phi(N, Y, X)}{(1 - G(Y-))} \mid X\right] = \mathbb{E}[\phi(M, \mathrm{T}, \mathrm{X})]$$

avec $G(t) = \mathbb{P}(C \leq t)$. Pour que cette égalité soit vraie, on doit formuler une hypothèse stipulant que la variable de censure C est indépendante de (M, T, X).

Les poids de Kaplan-Meier optimaux pour une observation i sont donc donnés par $\omega_i^* = \delta i\, n^{-1}[1 - G(Y_i -)]^{-1}$. Comme, dans la plupart des cas nous ne connaissons pas la fonction G, nous allons l'approcher à l'aide de l'estimateur de Kaplan-Meier. On obtient alors :

$$\tilde{G}(t) = 1 - \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^{n} 1_{Y_j \geq Y_i}}\right)$$

Par construction, les poids de Kaplan-Meier seront nuls pour une observation censurée. Pour les observations complètes, plus la période d'observation est importante, plus le poids de Kaplan-Meier sera important. Ainsi, cela nous permettra de compenser la faible proportion de sinistres pour lesquels on observe une liquidation complète sur une période prolongée en leur accordant un poids plus important dans la prédiction finale.

13.4. Procédure *weighted CART*

Nous reprenons ici les notations de Lopez et al.[33]. L'intégration des poids de Kaplan-Meier dans la procédure *CART* entraîne nécessairement une modification de celle-ci permettant la prise en compte de ces poids. La procédure *CART* cherche à estimer $\pi(z) = \mathbb{E}[\varphi(P)|Z = z]$ avec $Z = (T, X)$. Notons $R_j(z)$ la règle permettant de séparer l'échantillon initial en deux nœuds fils avec $R_j(z) = \begin{cases} 1 \\ 0 \end{cases}$. Pour la régression *CART* classique décrite dans le papier de Breiman et al.[20] et comme nous l'avons exprimé dans la partie 11.2.1, le critère de sélection de la meilleure division repose sur la minimisation de la variance intra-nœuds (dans le cadre d'une variable d'intérêt quantitative). La minimisation de l'erreur quadratique décrite dans Breiman et al.[20] est pertinente aux observations complètes, mais dans le cadre d'observations censurées, cette quantité n'est plus adaptée. Pour tenir compte de la présence de données censurées, on échange l'erreur quadratique avec une erreur quadratique pondérée de nos poids de Kaplan-Meier. La prédiction finale de la fonction de répartition φ issue de la procédure *CART* pour les règles $R = (R_1, \dots, R_K)$ est notée :

$$\hat{\pi}^R(Z) = \sum_{j=1}^{K} \hat{\pi}_j R_j(Z) \;\; o\grave{u}\; \hat{\pi}_j = \frac{\sum_{i=1}^{n} \boldsymbol{\omega_{i,n}} \varphi(N_i) R_j(Z_i)}{\sum_{i=1}^{n} \boldsymbol{\omega_{i,n}} R_j(Z_i)}$$

L'arbre obtenu pour $j = K$ est l'arbre maximal, mais comme cela a été décrit dans la partie 11.2, certaines techniques existent afin d'élaguer l'arbre maximal pour ne conserver que les branches apportant de l'information. Là encore l'approche de *pruning* doit être adaptée à la présence d'observations censurées à l'aide de l'introduction des poids de Kaplan-Meier afin que la quantité à minimiser soit :

$$\sum_{i=1}^{n} \boldsymbol{\omega_i} \left(\varphi(N_i, T_i, X_i) - \hat{\pi}^R(X_i)\right)^2 + \frac{\alpha K(R)}{n}$$

Avec α un paramètre de complexité choisi par validation croisée et $\sum_{i=1}^{n} \boldsymbol{\omega_i} \left(\varphi(N_i, T_i, X_i) - \hat{\pi}^R(X_i)\right)^2$ l'erreur quadratique pondérée à l'aide des poids de Kaplan-Meier.

Lopez et al.[32] (2016) démontrera que cet estimateur est consistant.

Section 4 : Résultats de la modélisation

Pour des raisons de confidentialité, les noms des professions ainsi que les différents montants ou volume des sinistres ne seront pas divulgués.

14. Préparation des données

Cette étape de *data processing* a été réalisée lors de la constitution de la base de données (cf. Analyse et travail sur la base de données). La discrétisation des variables continue comme la présence ou non d'un règlement principal et d'un règlement accessoire, l'ouverture du sinistre au forfait ou encore la transformation des variables continues (Ancienneté, $PSAP_{dernière\ vision}$, …) en variables catégorielles sont des exemples de modifications faites sur les données afin de faciliter leur traitement.

15. Travaux préalables à la modélisation

15.1. Analyse des corrélations

L'étude des corrélations est une étape nécessaire lorsque l'on souhaite utiliser des modèles de régression. En Statistiques, la corrélation est une mesure de l'intensité de liaison qui peut exister entre deux ou plusieurs variables. On peut mesurer cette corrélation à l'aide de différents coefficients, mais dans cette étude, c'est le V de Cramer qui est utilisé. Mis au point par Harald Cramer en 1946, le V de Cramer, ou Phi de Cramer, mesure l'intensité entre plusieurs variables qualitatives nominales. Sur R, la fonction ***catcor*** me permet de calculer le V de Cramer, puis la fonction ***corrplot*** du package ***ggplot2*** me permet d'obtenir une matrice de corrélation plus visuelle. On compare les données qualitatives d'abord, puis les données quantitatives. Ci-dessous la matrice de corrélation obtenue :

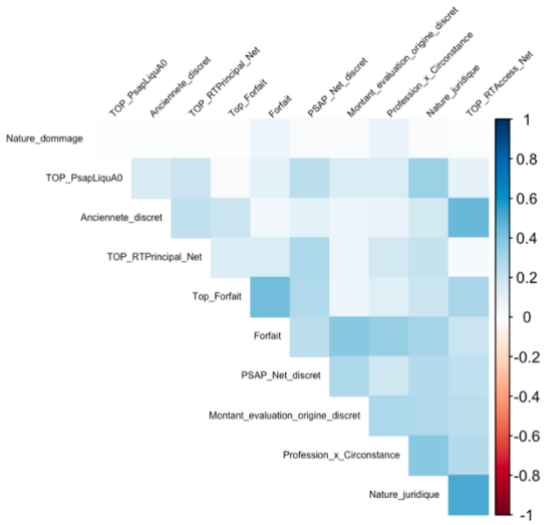


Figure 23 - Matrice de corrélation des variables qualitatives

Voici une liste exhaustive des constats que l'on peut conclure de cette matrice :

- La nature juridique du sinistre est assez corrélée avec le fait qu'il y ait eu ou non des règlements accessoires : cela parait logique puisque les frais liés à la judiciarisation du sinistre sont inscrits en frais accessoires.
- L'ancienneté du sinistre est également une variable corrélée avec le Top montant accessoires. Plus un sinistre est ouvert depuis longtemps, plus les chances qu'il soit judiciarisé sont importantes.
- La variable « Profession x Circonstance » est fortement corrélée avec la variable forfait : on peut trouver une explication dans le fait que les forfaits dépendent de la profession mise en cause dans le sinistre.

15.2. Transformation des variables catégorielles en *dummies*

On peut segmenter nos variables explicatives (cf. Présentation de la base de données) en deux parties :

- Les variables quantitatives : Montant d'évaluation, Ancienneté,…
- Les variables qualitatives : Profession x Circonstances, Nature juridique,…

Cette diversité de type de variables est très utile lors de la modélisation ; elle me permettra d'utiliser l'ensemble de l'information disponible afin d'expliquer nos variables cibles. Cependant, une interrogation se pose quant au traitement de ces variables dans les différents modèles. La solution réside dans la mise en place de variables muettes, en anglais *dummies*, prenant généralement comme valeur 0 ou 1. Soit une variable qualitative possédant m modalités, cette transformation me permet de créer pour chaque variable qualitative $m-1$ variables. Prenons par exemple la variable « Nature Juridique » qui comprend deux modalités : « Amiable » ou « Judiciaire ».

Elle sera remplacée par une variable muette « Nature Juridique Amiable » qui prendra comme valeurs :

- 1 si la nature juridique du sinistre est amiable,
- 0 si la nature juridique du sinistre est judiciaire.

Précisions tout de même que cette transformation est faite de manière automatique par le logiciel R lorsque des variables qualitatives sont intégrées un modèle de régression (la fonction ***GLM*** disponible sur R fait appel à la fonction ***model.matrix*** qui permet de prendre correctement en compte les variables qualitatives dans le modèle). Cependant, c'est pour le modèle *XG-Boost* que la transformation devient nécessaire. En effet, si le modèle gère très bien les données numériques, sa gestion des variables catégorielles est toujours en phase d'expérimentation et ne propose que des options limitées. On procède donc à la transformation des variables qualitatives nominales uniquement, car les variables qualitatives ordinales présentent un ordre naturel qui ne nécessite pas cette transformation. Voici un schéma récapitulatif :

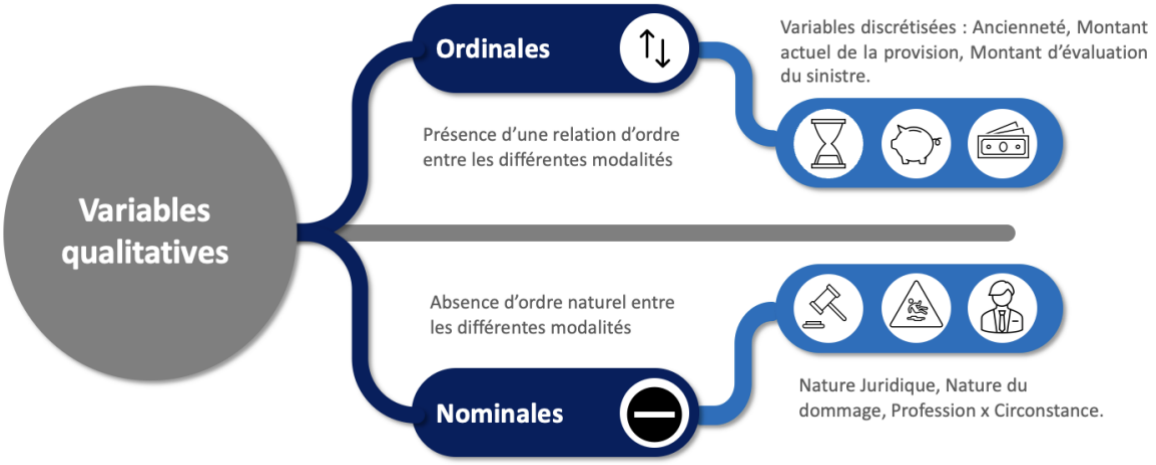


Figure 24 - Variables qualitatives ordinales et nominales

Pour conclure, un modèle de *boosting* permet de traiter ces variables catégorielles, c'est le modèle *CatBoost* (cf. 12.1.2.3) qui intègre à chaque variable qualitative un index permettant ensuite de procéder à une transformation de manière automatique. Toutefois, le modèle *CatBoost* se révèle être plus lent que le modèle *XG-Boost* classique.

16. Base d'apprentissage et base de test

Étape qu'on ne présente plus et qui est nécessaire lors d'une modélisation afin de s'assurer de la performance du modèle sur des données qu'il n'a jamais encore rencontrées. Pour cela, on va séparer notre base de données en deux bases :

- Une base d'apprentissage qui contiendra 70 % de la base initiale
- Une base de test qui représentera les 30 % restants.

Ainsi, on pourra apprécier la performance du modèle de manière correcte, c'est-à-dire sur la base de test. Les observations sont choisies de manière aléatoire sur R.

17. Modélisation de la probabilité de liquidation à 0€

À la suite de l'étape 3 d'échantillonnage en base d'apprentissage et en base de test, on vérifie que la variable cible binaire est représentée de manière similaire dans les deux échantillons. On constate que la part des PSAP liquidées à 0€ dans la base représente une partie correcte des observations ; les données ne sont pas déséquilibrées et ne nécessitent donc pas un traitement particulier.

17.1. La régression logistique

17.1.1. Variables continues ou Variables discrètes ?

Dans la partie *Discrétisation des données continues* j'avais transformé trois variables explicatives continues en variable discrètes : ancienneté, montant d'évaluation d'origine et $PSAP_{dernière\ vision}$.

Cette transformation permet d'envisager des effets non-linéaires de la variable explicative sur la variable cible (puisqu'avec une variable continue, on aurait seulement un effet linéaire). Cela facilite également la communication des résultats. Le seul point d'attention est la perte d'information : on passe d'une variable prenant une infinité de valeurs à une variable ne pouvant prendre que trois ou quatre modalités. Pour se convaincre de la meilleure stratégie au regard de la modélisation, on se propose d'entraîner deux modèles :

- Le premier contenant les variables discrètes.
- Le second contenant les variables continues.

On compare les modèles obtenus à l'aide du critère *AIC* que j'avais défini dans le paragraphe La déviance et le critère d'Akaike (*AIC*). Il apparaît que le modèle contenant les variables discrétisées présente un *AIC* plus faible que le modèle contenant les variables continues. On conserve donc ce modèle avec les variables ancienneté, montant d'évaluation d'origine et $PSAP_{dernière\ vision}$ discrétisées.

17.1.2. Sélection des variables explicatives

J'avais cité comme avantage des modèles de régression (cf. Avantages et points d'attention des modèles de régression) la possibilité de mettre en place une sélection *backward* des variables afin de ne conserver que les variables explicatives ayant un effet significatif sur la variable cible. On applique la fonction ***stepAIC*** disponible sur R sur le modèle complet contenant l'intégralité des variables explicatives et finalement, on constate que :

- Dans les deux modèles la variable représentant la nature du dommage n'a pas un impact significatif. Elle ne sera donc pas retenue comme variable explicative du modèle. Cela s'explique notamment par le fait que les sinistres corporels représentent une partie négligeable des sinistres en RC Pro des professions réglementées du chiffre et du droit.
- La variable présence d'un montant accessoire n'a pas d'impact significatif. Cette variable ne sera donc pas retenue dans la suite du modèle comportant les variables discrétisées.
  On avait déjà fait un constat similaire lors de l'étude des corrélations : en effet, la présence d'un montant accessoire est généralement liée à une instruction judiciaire (frais avocat).
- Les modalités des variables catégorielles qui ne sont pas significatives sont agrégées entre elles afin de voir si les résultats s'améliorent. C'est le cas notamment pour la variable Tranche de $PSAP_{dernière\ vision}$ (PSAP Nette dans la base de données) pour la modalité « 1 – inférieur à x_1€ » et « 2- Entre x_2 et x_3€ » qui devient « 1 - Inférieur à x_3€ »

Après avoir obtenu le modèle final contenant les variables explicatives significatives, interprétons les résultats du modèle afin de voir si ceux-ci sont pertinents.

17.1.3. Résultats du modèle logistique

➢ Variable Importance

Pour apprécier l'importance de chaque variable dans le modèle, on utilise la fonction ***varImp*** sur R. Pour les modèles de régression l'importance de chaque variable est obtenue à partir de la statistique du test de Wald[36] qui me permet de déterminer la significativité du modèle. Ainsi plus la statistique est grande, plus la variable est importante. Voici le graphique obtenu (seules les 14 variables les plus importantes ont été affichées) :

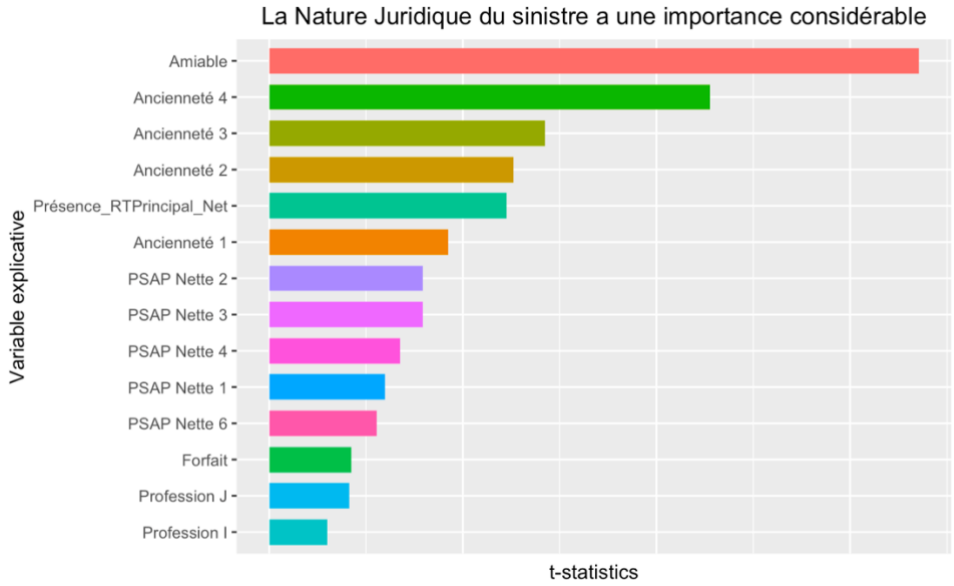


Figure 25 - Importance des variables dans la régression logistique

Ce graphique nous apprend que la variable contenant la nature juridique du sinistre est la variable la plus importante du modèle expliquant la liquidation des $PSAP_{ultime}$ à 0€. L'ancienneté, de manière décroissante, a également un impact important sur notre variable binaire cible (Ancienneté 4 = 4ème modalité de la variable Ancienneté).

➢ Coefficients du modèle

Après avoir observé l'importance de chacune des variables dans le modèle, on se demande désormais qu'elles sont les effets des variables explicatives sur la variable cible, toute choses égales par ailleurs. Les constats effectués au niveau Actuariat sont ensuite partagés avec les autres services en lien avec le marché des professions réglementées du chiffre et du droit : la Souscription, l'Indemnisation ainsi que le Pilotage économique.

Chacune de ses équipes à une vision différente du risque, pouvoir ainsi partager les résultats permet de confirmer certains constats, d'apporter de nouveaux points d'attention ou encore de challenger le modèle à la suite de certaines remarques.

Voici un schéma reprenant quelques constats qui ont été faits lors de l'analyse des coefficients du modèle :

[36] Test de Wald, stringfixer.com

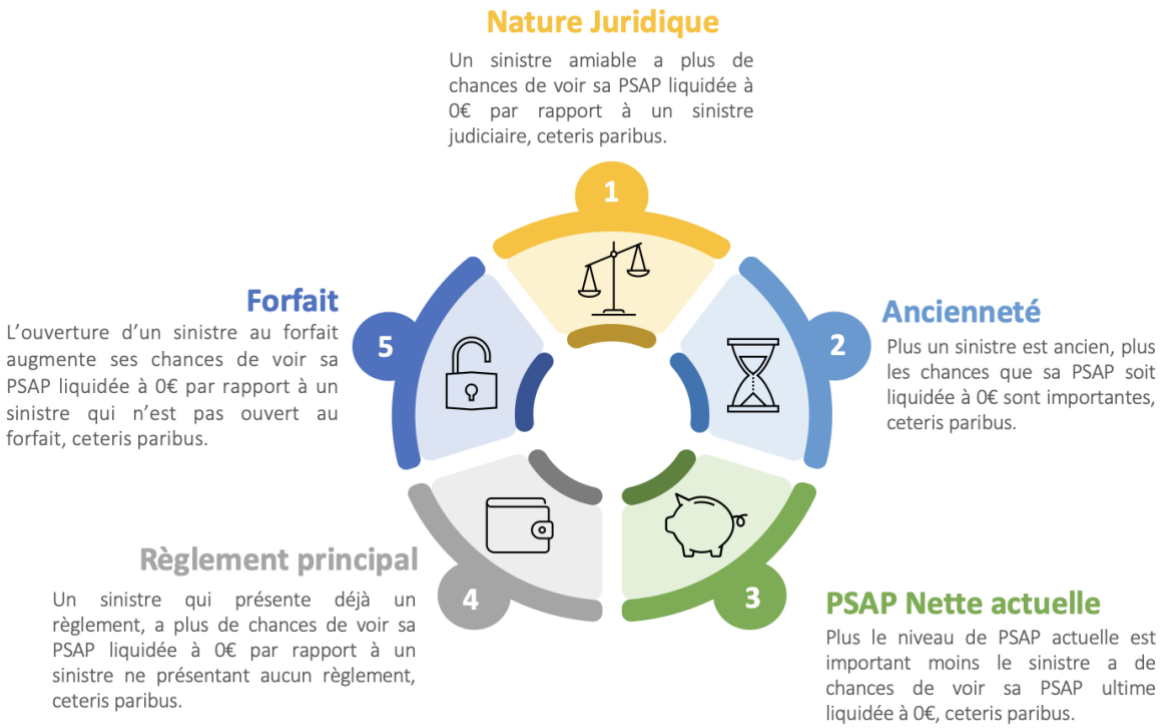


Figure 26 - Effet des variables dans le modèle logistique

À la suite du partage de ces constats avec les autres services il apparaît que :

- Effectivement un sinistre, une fois judiciarisé, à plus de chances d'avoir une $PSAP_{ultime}$ liquidée à un montant supérieur à 0€, car l'Indemnisation a remarqué que les décisions de justice sont de moins en moins favorables à l'assureur.
- Le constat sur l'ancienneté a d'abord surpris, mais finalement, il apparaît assez logique : un sinistre ancien est révélateur d'une procédure judiciaire ou d'un accord amiable plutôt difficile. Cela peut être révélateur d'un dossier dans lequel on a du mal à identifier les responsabilités de chacun (cf. 9.1.5) et dont l'issue peut nous être favorable.
  (rappelons que la base de modélisation ne comprend que les sinistres clos : les sinistres anciens toujours en cours présentent des caractéristiques différentes)
- Un sinistre est ouvert au forfait lorsque les informations sur le sinistre ne sont pas suffisantes pour établir une évaluation au réel.
  En fonction de l'évolution du dossier le montant est adapté par la suite. Ainsi les sinistres très importants ne sont que rarement ouverts au forfait : dès l'ouverture, on dispose des informations nécessaires pour lui affecter une évaluation au réel à la hauteur des dommages.

Les résultats du modèle sur les autres variables étaient plutôt en accord avec les résultats attendus.

17.1.4. Résidus du modèle de régression

L'analyse des résidus est une étape importante lors de l'utilisation de modèles de régression. En effet, ceux-ci sont calculés comme les écarts entre la prédiction et la valeur cible, et ont pour objectif

de matérialiser la partie non expliquée par le modèle. Voici les résidus obtenus pour la modélisation logistique de la probabilité de clôture à 0€ de la provision :

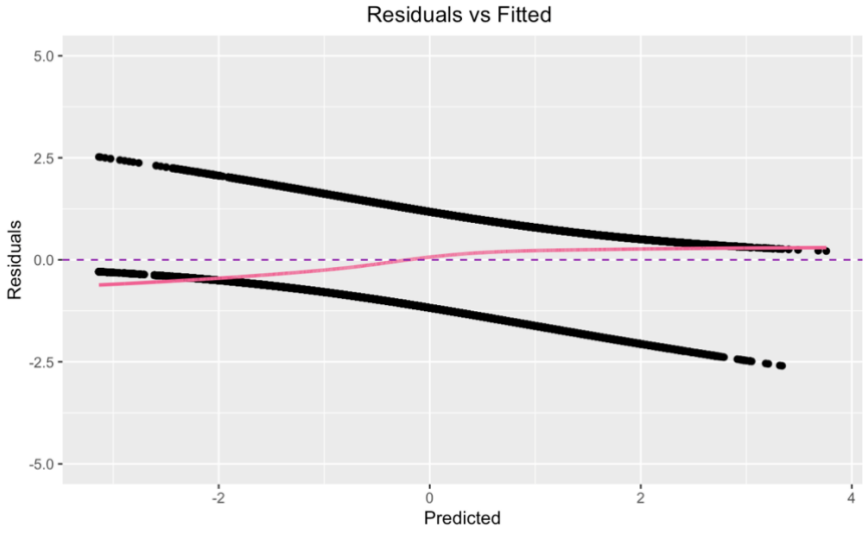


Figure 27 - Résidus du modèle logistique

Les résidus du modèle sont représentés par les deux courbes noires. Habituellement, les résidus prennent la forme d'un nuage de points centrés autour de 0. La forme particulière des résidus est due au type de régression employé : la régression logistique. Pour rappel, la variable cible est une variable binaire valant 0 ou 1, le modèle quant à lui prédit une probabilité qui est, par définition, comprise entre 0 et 1 :

- Si la valeur cible était 0, la valeur prédite sera quant à elle forcément supérieure ou égale à 0 et, *a fortiori*, les résidus seront négatifs : ils sont donc représentés par la courbe négative sur le graphique.
- Si la valeur cible était 1 la valeur prédite sera quant à elle forcément inférieure ou égale à 1 et, *a fortiori*, les résidus seront positifs : ils sont donc représentés par la courbe positive sur le graphique.

Si l'on se base uniquement sur ces deux courbes, difficile de pouvoir apprécier les résidus du modèle, il faut au préalable passer par une régression. Cette méthode est décrite dans Charpentier[37] (2013) et fait appel à une régression non-paramétrique permettant de produire des courbes lissées sur un ensemble de points, la régression LOWESS. Cette régression est représentée par la courbe rose sur le graphique : plus cette courbe est droite plus le modèle est performant. La droite obtenue pour notre modèle de régression logistique n'est pas une ligne horizontale bien qu'elle soit relativement plate. Prenons toutefois ce graphique avec une certaine mesure, en effet, on pourrait complexifier le modèle, mais nous avons fait le choix de ne garder que les variables participant à la réduction de l'*AIC* (cf. partie 17.1.2). La recherche de simplicité dans la modélisation peut entraîner des résidus plus importants et ainsi une courbe et non une droite : compte tenu du travail de sélection des variables faite sur ce modèle, l'aspect des résidus est positif.

[37] Charpentier, 2013.

17.2. Le modèle *XG-Boost*

17.2.1. Traitement des données

Pour appliquer le modèle *XG-Boost*, les données ont besoin d'être retraitées. On va notamment :

1. A partir de la base d'apprentissage : ne conserver que les variables qui serviront dans le modèle : explicatives et à expliquer (on ne conserve pas les variables non-explicatives date d'ouverture, date de clôture,…)
2. On isole la variable cible des variables à expliquer en créant deux matrices.
3. On crée ensuite notre base d'apprentissage adaptée au *XG-Boost* via la fonction R ***xgb.DMatrix***

On réitère ces étapes sur la base de test. Je peux désormais appliquer le modèle *XG-Boost* à nos données.

17.2.2. Modèle global et optimisation de *nrounds*

On lance un premier modèle global *XG-Boost* en mettant la majorité des paramètres à leur valeur par défaut. On choisit les valeurs de trois paramètres :

- *booster* = "*gbtree*" : on souhaite utiliser des arbres afin de prédire la variable, mais une version « *gblinear* » existe aussi.
- *objective* = "*binary*:*logistic*" : on précise que l'on travaille sur une variable binaire.
- *metrics* = "*logloss*" : la métrique utilisée pour les problèmes de classification.

Il reste un paramètre qui n'a pas été fixé, le nombre d'itérations de l'algorithme : ***nrounds***. ***nrounds*** représente le nombre d'arbres de décisions présents dans le modèle final. Afin d'optimiser ce paramètre, on va procéder par validation croisée à l'aide de la fonction ***xgb.cv***. La validation croisée est une méthode d'échantillonnage permettant d'évaluer des quantités sur différents sous-ensembles de la base initiale. En précisant le paramètre ***nfold*** on peut fixer le nombre de sous-ensembles ainsi créés.

On fixe au départ un nombre important d'itérations (admettons 1500) et on va pouvoir ainsi calculer sur la base d'échantillon « test » l'erreur *logloss* en fonction de chaque itération afin d'identifier pour quelle itération l'erreur est la plus faible. Voici le graphique obtenu :

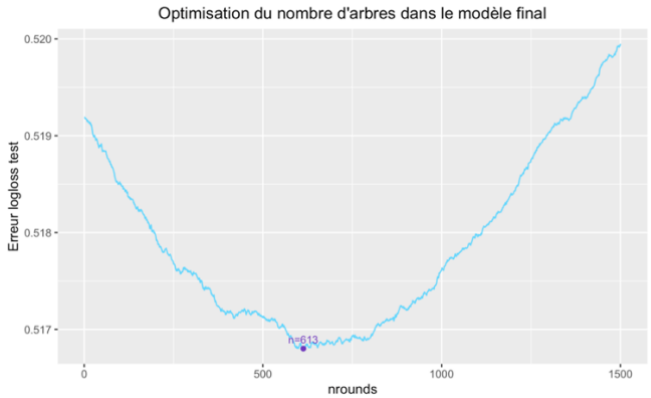


Figure 28 - Optimisation du nombre d'itérations

Le paramètre ***nrounds*** minimise l'erreur ***logloss*** lorsqu'il prend la valeur 613 (cette valeur est modifiée.) : notre modèle final comportera donc 613 arbres ! Il reste une multitude d'hyperparamètres à optimiser : *eta, colsample_bytree, min_child_weight*,... Après avoir lancé un *gridsearch* on observe finalement que l'apport sur la précision de la prédiction finale est faible. On décide donc de n'optimiser que le paramètre *nrounds*.

17.2.3. Importance des variables

Le modèle de *XG-Boost*, et plus globalement les modèles de *Machine Learning*, ont la réputation d'être des « boîtes noires ». Autrement dit, il peut être parfois compliqué d'analyser les résultats du modèle variable par variable comme on a l'habitude de le faire avec un modèle de régression classique. Cependant, il existe quelques alternatives afin d'avoir un peu plus de visibilité sur le modèle : extraire quelques arbres de décision modélisés par le modèle par exemple. On se propose plutôt d'observer l'importance des variables, de la même fonction que pour la régression logistique, à l'aide de la fonction ***xgb.importance***. Le calcul, cependant, diffère de celui utilisé pour les modèles de régression. En effet, dans la partie 11.1.4 j'avais décrit les différentes fonctions utilisées pour minimiser l'hétérogénéité entre les classes à avoir : Indice de Gini, entropie ou encore erreur de Bayes. Un arbre crée un nœud lorsque l'on maximise la perte associée à la minimisation de cette hétérogénéité. L'importance de chaque variable est alors calculée en effectuant la moyenne sur tous les arbres du modèle de la perte associée à la minimisation de cette hétérogénéité pour cette même variable[38]. On obtient ainsi le diagramme suivant :

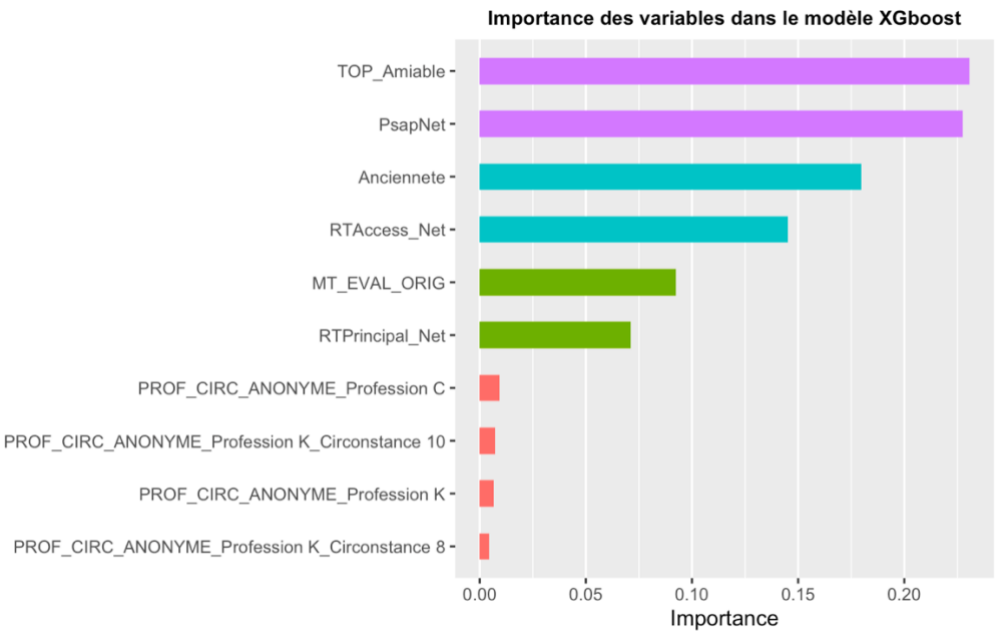


Figure 29 - Importance des variables dans le modèle XG-Boost

Finalement la variable Amiable ressort comme étant la variable la plus explicative du modèle (j'avais fait le même constat sur la régression logistique). La $PSAP_{dernière\ vision}$ ainsi que l'Ancienneté viennent compléter le podium ; on les retrouvait également dans la régression logistique. (en ordre inverse)

[38] XG-Boost importance des variables, Josiah Parry, Décembre 2018

17.2.4. Quid du sur-apprentissage ?

Si l'on doit citer un point d'attention majeur des modèles de *Machine Learning,* c'est le sur-apprentissage. Le sur-apprentissage (cf. 11.2), ou *overfitting* en anglais, se produit lorsque l'algorithme s'adapte trop sur les données d'apprentissage en intégrant les moindres variations inexpliquées et aléatoires dans les données. Le modèle va donc afficher de très bons résultats sur la base d'apprentissage. Même si de prime abord, on pense que cela est un avantage, lorsque le modèle rencontrera de nouvelles données, il affichera de bien moins bonnes prédictions. Plusieurs méthodes peuvent être utilisées afin de se rendre compte de l'existence de sur-apprentissage dans un modèle de *Machine Learning*. Après avoir comparé la précision du modèle obtenue sur la base d'apprentissage et celle obtenue sur la base de test, on observe une précision légèrement inférieure sur la base de test. Un écart trop important soulèverait des doutes quant à la présence d'*overfitting*. On se propose également d'observer l'erreur logloss obtenue par validation croisée sur la base test et sur la base d'apprentissage :

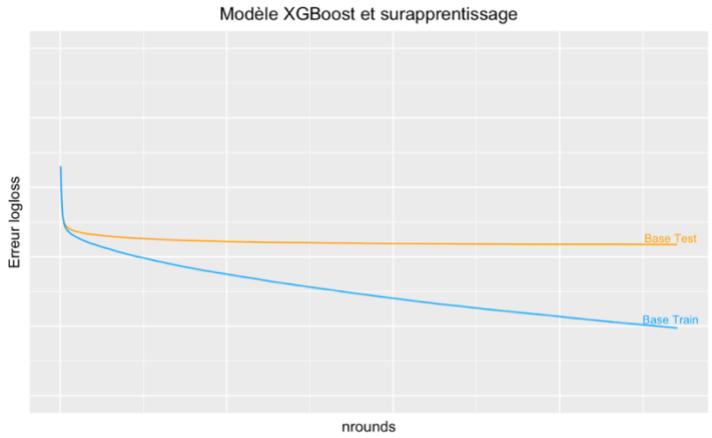


Figure 30 - Graphique erreur logloss sur la base d'apprentissage et sur la base de test

Les conclusions sont plutôt positives : on voit que l'erreur sur la base de test décroît puis se stabilise, mais n'augmente pas à nouveau. Dans le cas où l'erreur augmenterait à partir d'une certaine itération, on aurait été face à du sur-apprentissage et on aurait dû employer des méthodes comme l'*early-stopping* qui est définit dans la partie 11.2.2.

17.3. Matrice de confusion des modèles

Après avoir lancé nos deux modélisations, on se retrouve avec une prédiction de la probabilité que la $PSAP_{ultime}$ soit liquidée à 0€.

Voyons désormais comment, à l'aide de la probabilité prédite, revenir à une variable binaire prenant 0 ou 1 comme valeur afin de comparer les résultats entre eux avec les résultats effectivement observés. Cette démarche a pour unique but de comparer les modèles, nous conserverons la probabilité prédite par la suite.

Pour cela rien de plus simple, on fixe un seuil tel que :

- Si la probabilité est supérieure au seuil alors $\hat{Y}_{modèle} = 1$
- Si la probabilité est inférieure au seuil alors $\hat{Y}_{modèle} = 0$

On peut choisir un seuil personnalisé de deux manières :

- Scénario 1 : maximiser la précision avec un seuil de 0,47.
- Scénario 2 : fixer une spécificité à 80 % avec un seuil de 0,4.

La procédure menant au choix de ces seuils est décrite en Annexe 2. Voici les matrices de confusion obtenues pour la régression logistique, (un coefficient aléatoire a été appliqué aux valeurs) :

Scénario 1 : Maximisation de la précision		Valeurs de références	
		0 (= Liquidée > à 0€)	1 (= Liquidée à 0€)
Valeurs prédites	0 (= Liquidée > à 0€)	39 631	16 844
	1 (= Liquidée à 0€)	17 324	38 811

Précision : 70% - Spécificité : 69% - Sensibilité : 69%

Figure 31 - Matrice de confusion avec maximisation de la précision

Scénario 2 : 80% de spécificité minimum		Valeurs de références	
		0 (= Liquidée > à 0€)	1 (= Liquidée à 0€)
Valeurs prédites	0 (= Liquidée > à 0€)	33 366	11 097
	1 (= Liquidée à 0€)	23 589	44 558

Précision : 69% - Spécificité : 80% - Sensibilité : 58%

Figure 32 - Matrice de confusion avec 80% de spécificité minimum

Pour le scénario 2, on retrouve les remarques réalisées ci-dessus : fixer un niveau de spécificité implique forcément une perte de précision, mais celle-ci est moindre dans ce cas. On conserve donc le scénario 2 et on applique cette méthode aux résultats du modèle *XG-Boost*. Voici la matrice de confusion obtenue :

XG Boost		Valeurs de références	
		0 (= Liquidée > à 0€)	1 (= Liquidée à 0€)
Valeurs prédites	0 (= Liquidée > à 0€)	39 092	11 128
	1 (= Liquidée à 0€)	17 863	44 527

Précision : 75% - Spécificité : 80% - Sensibilité : 68%

Figure 33 - Matrice de confusion XG-Boost

Finalement, on observe via ses comparaisons, effectuées sur la base de test, que le modèle *XG-Boost* permet d'augmenter la précision pour un même niveau de spécificité.

Pour conclure, le modèle *XG-Boost* me permet déjà d'améliorer cette première étape de la modélisation de 6 points ! Précisons que dans la suite de l'étude, on conservera la probabilité prédite par le modèle : cela me permettra d'affecter un montant même aux observations présentant une probabilité forte de clôture à 0€ toujours par soucis de prudence.

18. Modélisation du montant de la $PSAP_{ultime}$

Dans cette partie, on ne conserve que les sinistres dont la $PSAP_{ultime}$ a été liquidée à un montant strictement positif.

18.1. Sinistres attritionnels et sinistres graves

18.1.1. Introduction à la Théorie des Valeurs Extrêmes

La théorie des valeurs extrêmes (TVE) est un domaine important de l'Actuariat permettant d'étudier et de déterminer le comportement de valeurs extrêmes présentes au sein d'un échantillon de variables aléatoires.

En effet, voici un graphique représentant le montant de la $PSAP_{ultime}$ de nos données :

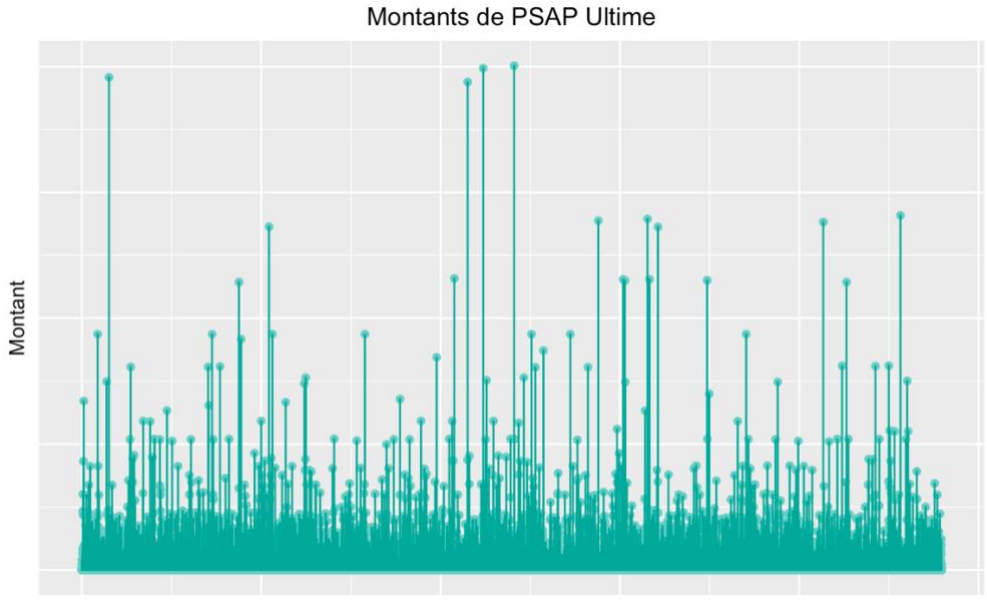


Figure 34 - Montants de $PSAP_{ultime}$

Comme on l'observe les montants sont relativement différents : certains avoisinent les millions d'euros (une faible proportion heureusement), tandis que d'autres sont autour d'une centaine d'euros. La TVE intervient alors afin de pouvoir distinguer les sinistres en deux groupes ayant des caractéristiques différentes à l'aide d'un seuil des sinistres graves. Ainsi, les sinistres dont la $PSAP_{ultime}$ se situe sous le seuil seront identifiés comme des sinistres attritionnels, à l'instar des sinistres dont la $PSAP_{ultime}$ est supérieure au seuil qui seront identifiés comme des sinistres graves.

Toutefois, on observe une dizaine de $PSAP_{ultime}$ ayant des montants qui semblent particulièrement élevés. Sur ces gros dossiers (qui sont en l'occurrence les sinistres les plus importants de ces 20 dernières années sur ce marché) une attention particulière est portée par le service indemnisation, on estime que l'évaluation de la $PSAP_{dernière\ vision}$ est à un niveau correct. En effet, ces sinistres importants, presque exceptionnels, sont suivis de manière régulière et leurs évaluations sont revues au fur et à mesure de l'avancée de l'instruction du dossier. On isole cette dizaine de sinistres en se fixant une limite, appelée seuil des sinistres exceptionnels, au-delà de laquelle on conserve la $PSAP_{dernière\ vision}$.

18.1.2. Détermination du seuil

Pour déterminer le seuil de ses sinistres graves, on se propose d'utiliser une méthode graphique : le mean-excess plot. Ce graphique permet de tracer la fonction mean-excess d'un échantillon de variables aléatoires. La méthode utilisée est décrite en Annexe 3. Observons à nouveau le graphique de la partie 18.1.1 cette fois avec le niveau des seuils :

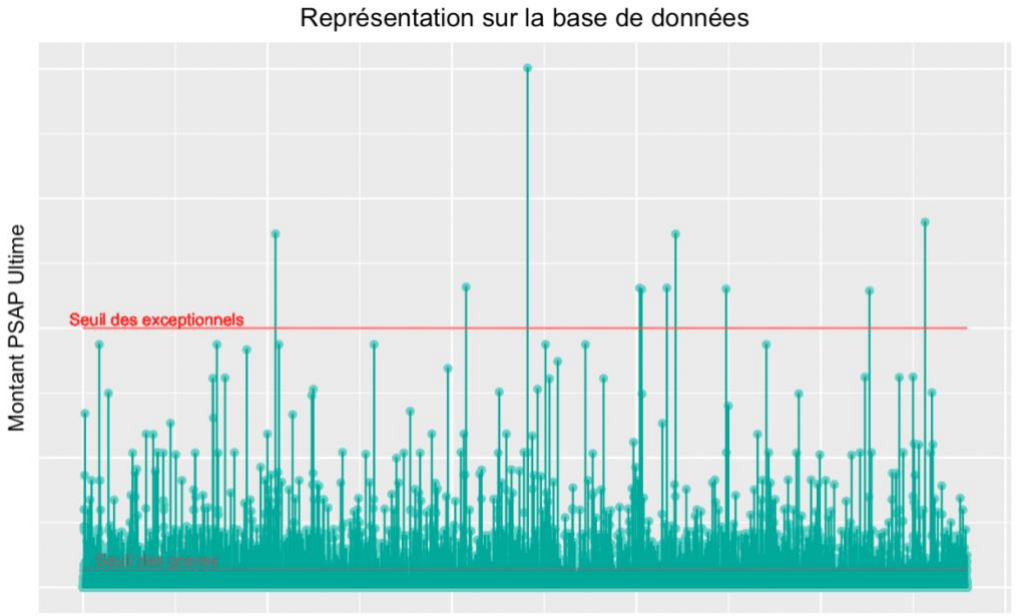


Figure 35 - Montants de $PSAP_{ultime}$ avec le seuil de sinistres graves et de sinistres exceptionnels

Les écarts importants entre les valeurs de $PSAP_{ultime}$ nous encouragent à aborder la modélisation de trois manières différentes :

- Un modèle pour les sinistres attritionnels.
- Un modèle pour les sinistres graves.
- Pour les sinistres exceptionnels : pas de modèle, on reste sur l'observé.

18.2. Modélisation du montant de $PSAP_{ultime}$ des sinistres attritionnels

18.2.1. La Régression Log-Normale

Dans la méthode de tarification existante, c'est via une régression log-normale que les montants de $PSAP_{ultime}$ étaient prédits. On se propose de vérifier la bonne adéquation de cette loi sur nos données afin de voir s'il est nécessaire d'approcher la modélisation via une autre loi. Voici l'histogramme des $\log(PSAP_{ultime})$ sur lesquels est ajustée une loi log-normale :

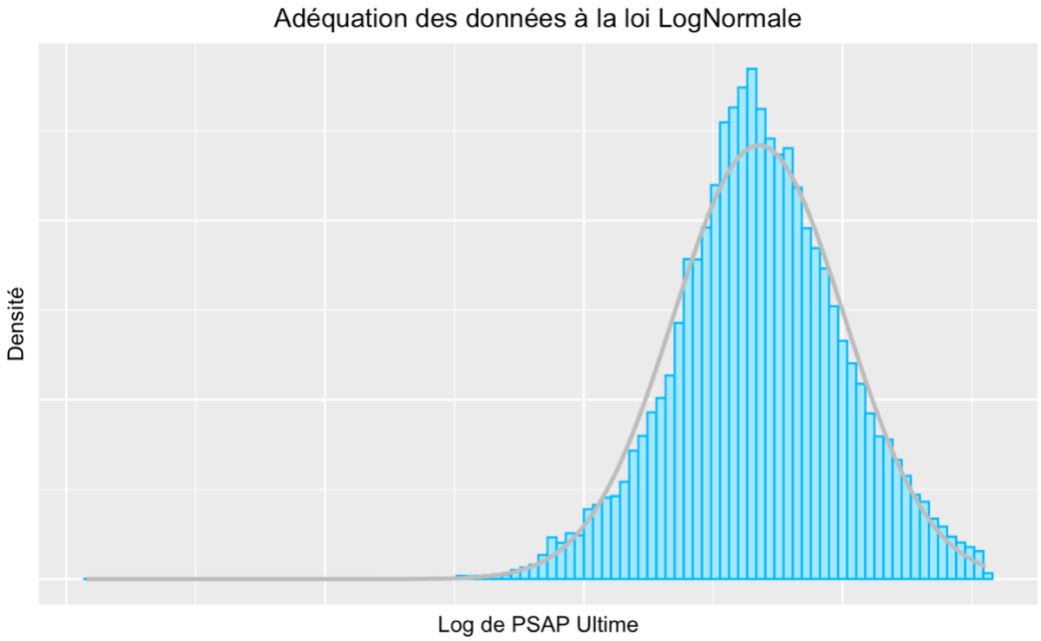


Figure 36 - Adéquation de la loi log normale au $log(PSAP_{ultime})$

L'adéquation à la loi log-normale est plutôt correcte, modulo quelques valeurs, mais dans l'ensemble l'utilisation de cette loi semble justifiée. On prédira donc les $\log(PSAP_{ultime})$ via une loi log-normale. Pour confirmer la loi choisie et valoriser statistiquement l'adéquation on utilise le test de Kolmogorov-Smirnov en comparant les $log(PSAP_{ultime})$ avec un vecteur de même dimension de réalisations de la loi log-normale (avec en moyenne et en écart-type les paramètres estimés sur les $log(PSAP_{ultime})$) : la p-value obtenue est de 0,2705. La p-value du test étant supérieure à 0,05 on accepte l'hypothèse nulle qui stipule que les deux variables sont distribuées selon la même loi.

La modélisation du montant de la $PSAP_{ultime}$ et de sa probabilité de clôture à 0€ présentent les mêmes étapes :

- Préparation des données : ajout de variables, traitement des valeurs manquantes, …
- Travaux préalables : corrélations, *dummification*
- Échantillonnage : en base d'apprentissage et en base de test
- On challenge deux modélisations : avec les variables catégorielles et les variables continues
- Une fois le modèle choisi, on applique une sélection automatique *backward* via la fonction ***stepAIC***

Avant de comparer l'erreur des modèles, on se propose de regarder la prédiction du modèle par rapport à l'observé. Voici les résultats sur la base de test :

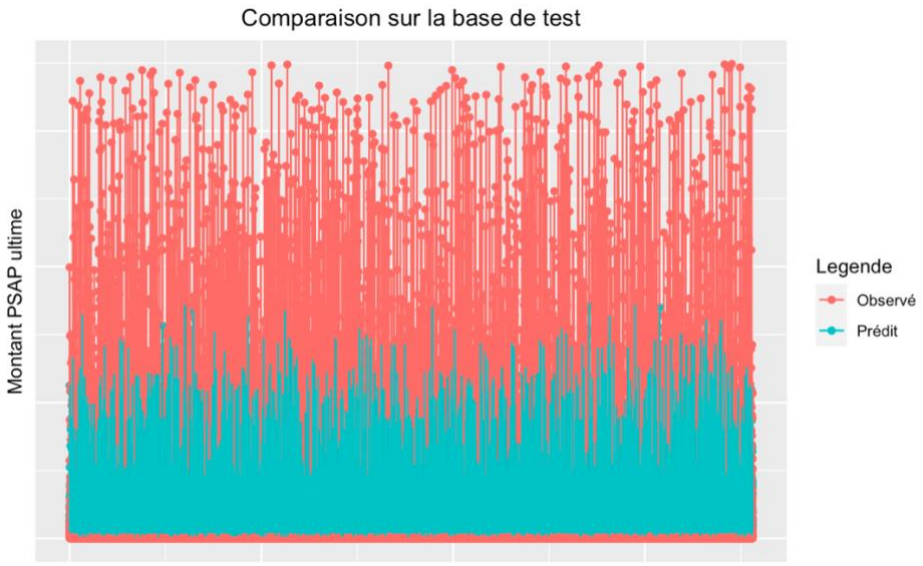


Figure 37 - Comparaison sur la base de test des montants de $PSAP_{ultime}$ (Régression)

On remarque que les résultats prédits sous-estiment les montants importants de $PSAP_{ultime}$. En effet, ce constat avait déjà été fait lors de l'application de la méthode auparavant, cela impliquait une sous-estimation de la charge sinistre ultime et *a fortiori* de la prime pure. Pour comprendre d'où vient cette sous-estimation du modèle, on regarde les résultats obtenus sur la base d'apprentissage (attention : la comparaison entre les résultats observés et les résultats prédits n'a de sens que sur la base de test en termes de performance du modèle ! Ici, on souhaite juste voir si le modèle sous-estime également sur la base d'apprentissage.) :

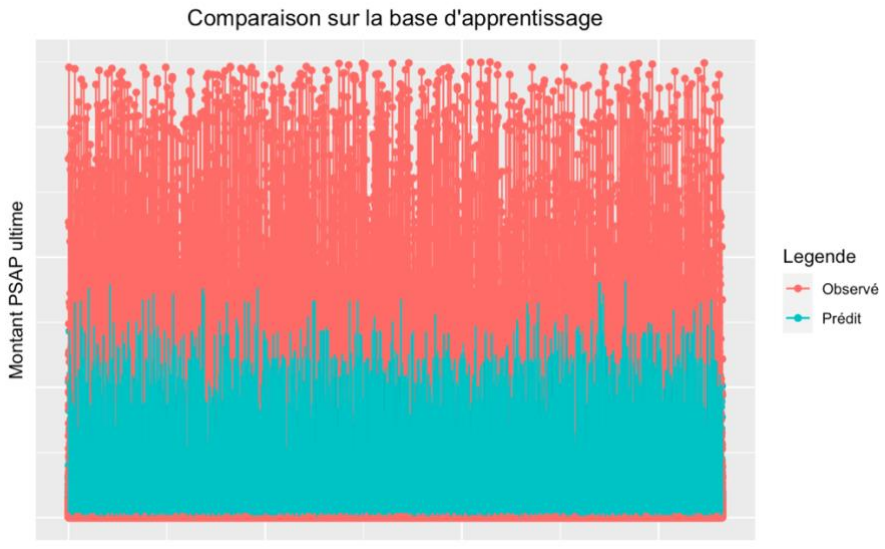


Figure 38 - Comparaison sur la base d'apprentissage des montants de $PSAP_{ultime}$ (Régression)

Finalement, on remarque que même sur la base d'apprentissage le modèle ne parvient pas à capter l'intégralité de l'information. Le modèle n'apprend pas correctement sur les montants les plus importants.

Cependant, ce constat n'est pas surprenant : le modèle de régression moyennise le résultat final. Même si en moyenne les résultats ne sont pas si décadrés, lorsque l'on observe la prédiction d'une profession ou d'un grand compte en particulier, on peut se retrouver avec une prime pure finale qui n'est pas à la hauteur du risque.

Une sous-estimation de la prime pure peut se révéler particulièrement fâcheuse un cadre de *LTA*, car le tarif ne pourra pas être revu avant plusieurs années (cf. 4.2.1). On souhaite donc corriger cette sous-estimation en challengeant la méthode de régression via la méthode *XG-Boost*.

18.2.2. Modèle *XG-Boost*

Comme pour la régression, la méthode appliquée pour modéliser le montant de la $PSAP_{ultime}$ est la même que pour modéliser la probabilité de clôture à 0€ modulo les paramètres du *XG-Boost* qui deviennent :

- *objective* = « *reg*:*squarederror* » : car on est sur un problème de régression
- *eval_metric* = « *RMSE* »

Pour le reste : préparation de la base de données, optimisation du nombre d'itérations par validation croisée, la méthode est similaire. On teste deux méthodes de modélisation :

- Méthode 1 : application du modèle *XG-Boost* pour prédire les montants de $PSAP_{ultime}$
- Méthode 2 : application du modèle *XG-Boost* pour prédire les log $(PSAP_{ultime})$

➢ Méthode 1 : montants de $PSAP_{ultime}$

Comme pour la régression, on se propose d'observer le graphique représentant les montants prédits et observés sur la base de test :

Figure 39 - Comparaison sur la base de test des montants de $PSAP_{ultime}$ (XG-Boost)

On fait deux constats majeurs sur ce graphique :

- Le premier constat est que le modèle *XG-Boost* semble mieux capter les montants élevés que le modèle de régression. Cela me rassure dans le choix d'une modélisation via une approche *Machine Learning*.
- Second constat, le modèle prédit des montants négatifs bien qu'il n'y en a pas dans la base d'apprentissage (En effet, seuls les sinistres dont la $PSAP_{ultime}$ est supérieure à 0€ ont été conservés.) Voyons si les résultats sur la base d'apprentissage sont similaires :

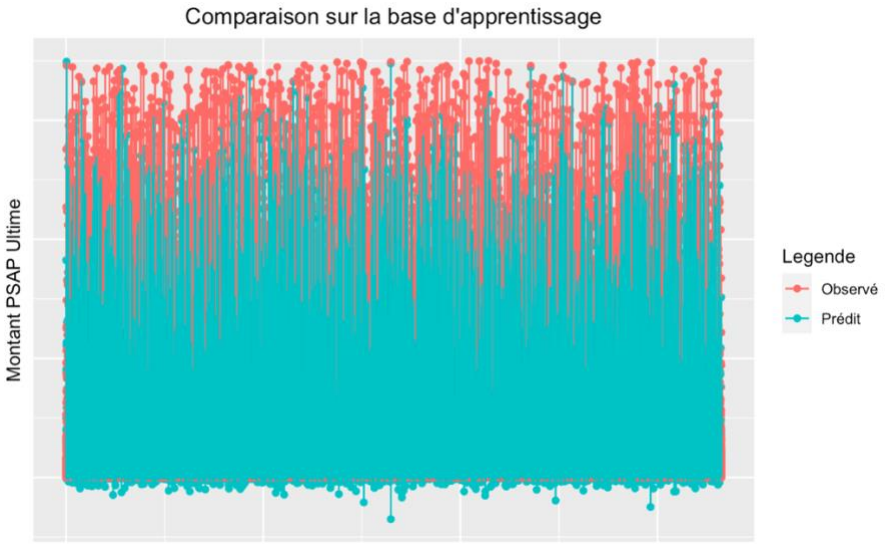


Figure 40 - Comparaison sur la base d'apprentissage des montants de $PSAP_{ultime}$ (XG-Boost)

On remarque que, malgré l'absence d'observations négatives, le modèle prédit des montants négatifs sur la base d'apprentissage. Le modèle *XG-Boost*, peut-il prédire des valeurs en dehors des valeurs minimales et maximales observée ? La réponse est oui. En effet, le *Boosting* fonctionne sur la minimisation des erreurs effectuées par le modèle lors de l'itération précédente (cf. 12.1.1.2). L'algorithme se concentre sur la minimisation de la fonction de perte sans pour autant prendre en compte les limites supérieures, ou inférieures, de l'échantillon de données. Prenons un exemple[39] simple, un modèle générant 2 arbres à l'aide de 2 variables explicatives. On veut prédire la $PSAP_{ultime}$ à l'aide de :

- L'ancienneté du sinistre en années
- La profession sinistrée : A, B ou C.

On suppose que le montant minimal de $PSAP_{ultime}$ constituée pour un sinistre disposant d'un an d'ancienneté est 1000€. Admettons que pour *nrounds*=1, i.e. le premier arbre, seule la variable « Ancienneté » est choisie dans la construction de l'arbre tel que : une année d'ancienneté en plus diminue $PSAP_{ultime}$ de 10€. Cette première modélisation a généré une erreur qu'on note $\mathcal{E}_1$.

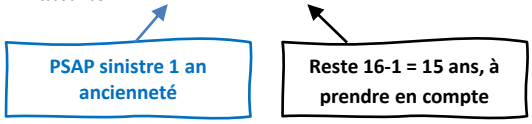
La méthode de *Gradient Boosting* est une méthode d'*Ensemble Learning* qui va partir de l'erreur réalisée par le premier arbre $\mathcal{E}_1$ pour réaliser le second arbre.

Pour ***nrounds***=2, le second arbre va tenter d'expliquer l'erreur réalisée par le premier arbre. La modélisation conclut que les sinistres de la profession

[39] Exemple prédictions XG-Boost hors limites, StackExchange.

A nécessitent la constitution de $PSAP_{ultime}$ moins importante que pour les autres professions : le montant de la $PSAP_{ultime}$ doit être corrigé de $-900€$. Dans notre base d'apprentissage le sinistre le plus ancien à 8 ans d'ancienneté. Dans la base de test, échantillon jamais encore rencontré par le modèle, on doit prédire le montant de $PSAP_{ultime}$ d'un sinistre de la profession A clos 16 ans après son ouverture :

- Le premier arbre va conclure que le montant de la $PSAP_{ultime}$ doit être réduit de : $PSAP_{ultime} = 1000€ - 15 * 10€ = 850€$



- Le second arbre sait que le premier arbre se trompe, d'environ $900€$ pour un sinistre de la profession A, sur le montant de la $PSAP_{ultime}$: $PSAP_{ultime} = 850€ - 900€ = -50€$

Ainsi, on prédit un montant de $PSAP_{ultime}$ négatif bien qu'aucune observation de la base de données initiale ne soit négative.

Pour corriger ce problème plusieurs solutions : la mise en place d'une contrainte dans le modèle ou la transformation via la fonction log par exemple. En effet, en transformant la variable cible via la fonction logarithme, on appliquera la fonction exponentielle aux résultats de la prédiction : de cette manière, on écarte la possibilité de prédire des montants négatifs.

➢ Méthode 2 : $log(PSAP_{ultime})$

On applique la transformation log sur les montants de $PSAP_{ultime}$ et on entraîne les modèles en optimisant le nombre d'itérations par validation croisée. Voici les résultats obtenus :

Figure 41 - Comparaison sur la base de test des montants de$PSAP_{ultime}$ (XG-Boost avec transformation log)

On constate que l'on ne prédit plus de montants négatifs, mais une observation a été prédite à un niveau relativement important (au-dessus du seuil de sinistres graves alors que l'on travaille sur les sinistres attritionnels). Le même constat est fait sur la base d'apprentissage : bien que les sinistres présentent une $PSAP_{ultime}$ inférieure au seuil des graves le modèle prédit sur cette base d'apprentissage une dizaine de montants supérieurs à cette limite.

Finalement, on doit garder un esprit critique : cette modification me permet seulement de déplacer le problème. Cela étant dit, cette méthode n°2 est préférée à la méthode n°1 pour deux raisons :

- La première réside dans le fait que seulement une dizaine d'observations supérieures à la valeur maximale ont été observées tandis que c'est une centaine d'observations prédites inférieures à 0€.
- On préférera toujours prédire un montant de charge sinistre ultime plus élevé plutôt que trop faible.

18.3. Modélisation du montant de $PSAP_{ultime}$ des sinistres graves

La modélisation du montant de $PSAP_{ultime}$ des sinistres graves est la même que pour les sinistres attritionnels, on se propose donc de n'observer que les résultats finaux sur la base de test avec le modèle de régression log-normale :

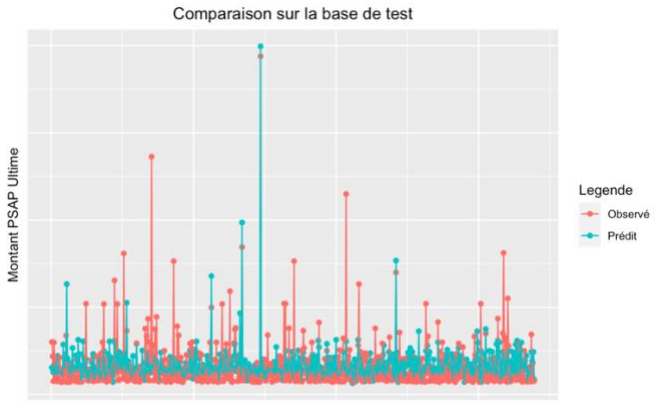


Figure 42 - Comparaison sur la base de test des montants de $PSAP_{ultime}$ graves (Régression)

Et avec le modèle *XG-Boost* :

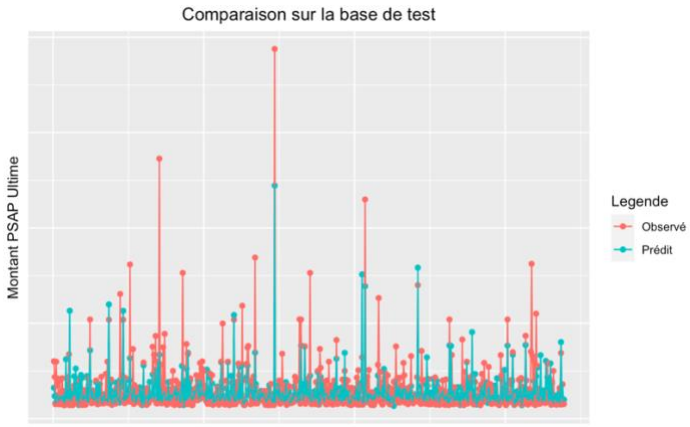


Figure 43 - Comparaison sur la base de test des montants de $PSAP_{ultime}$ graves (XG-Boost)

Le nombre d'observations réduit implique des écarts sur les montants prédits. Cependant on remarque que le modèle de régression a su prédire l'observation la plus importante là où le *XG-Boost* ne l'a prédit qu'à moitié. Rappelons que nous sommes sur des valeurs atteignant des millions d'euros : cette observation va donc avoir un impact considérable lors de la comparaison du modèle *XG-Boost*.

18.4. Comparaison

18.4.1. MAE et $RMSE$

Afin de comparer les modèles entre eux, on se propose d'utiliser deux critères :

- La moyenne arithmétique des valeurs absolues des écarts : la MAE (Mean Absolut Error). Cette quantité est définie par la formule suivante :

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

- La racine de l'erreur quadratique moyenne : la RSME (Root Squared Mean Error). Cette quantité est définie comme suit :

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

Avec n le nombre d'observations, y_i les valeurs observées de notre variable cible (montants de $PSAP_{ultime}$ observés) et $\hat{y_i}$ les valeurs prédites par les modèles de notre variable cible (montants de $PSAP_{ultime}$ prédits).

Pourquoi observer ces deux quantités afin de comparer les modèles ? Premièrement, les deux quantités présentent des similitudes[40] :

- Leur valeur est à apprécier en fonction de la nature de la variable cible : avoir une $RMSE$ ou une MAE importante si notre variable cible prend des valeurs importantes est tout à fait possible, cela ne veut pas forcément dire que le modèle n'est pas performant pour autant.
- Plus leur valeur est faible plus le modèle est performant : si on a $RMSE_{modèle1} < RMSE_{modèle2}$ alors le modèle 1 est plus proche du résultat observé que le modèle 2. Même constat pour la MAE.

Regarder les deux quantités est intéressant car elles présentent des avantages différents[26] :

- La $RMSE$ va accorder plus de poids aux fortes erreurs dû à la présence du carré dans sa formule. Elle va également être plus sensible aux prédictions anormales (comme celle observée dans la partie 18.2.2) cependant elle reste la mesure de choix en termes de performance du modèle.
- La MAE est beaucoup plus facile à interpréter et à communiquer. Dans l'intérêt de notre étude qui est de présenter une nouvelle méthode de modélisation à un public qui n'est pas forcément familier avec ce domaine, on préférera utiliser la MAE.

Rappelons que ces quantités sont à regarder sur la base de test. Voici les résultats obtenus sur la prédiction finale, attritionnelle et graves, sur la base de test :

[40] *RMSE* VS *MAE*, Stephen Allwright, Juillet 2022.

	RMSE	MAE
Régression LogNormale	65 400	19 200
XG-Boost	59 060	16 670

Figure 44 - Comparaison MAE et RMSE

En regardant les résultats sur la $RMSE$ on remarque que celle-ci est plus faible pour le modèle *XG-Boost*, celui-ci est donc plus performant sur le montant de la $PSAP_{ultime}$ que le modèle de régression. Même constat sur la MAE, celle-ci est plus faible pour le modèle *XG-Boost* : on est donc plus proche de la variable observée (en valeur absolue) avec le modèle *XG-Boost*. Ces résultats viennent confirmer ce que l'on observe sur les graphiques observé/prédit sur la base de test. Le modèle *XG-Boost* semble mieux capter l'information, notamment sur les montants plus élevés des sinistres attritionnels.

18.4.2. Charge sinistre ultime

On peut également comparer la charge sinistre ultime sur la base test entre la charge sinistre ultime observée, la charge sinistre ultime prédite par la régression log-normale et la charge sinistre ultime prédite par le modèle *XG-Boost* :

- Sur la charge ultime des sinistres attritionnels, on observe environ 1,8 % d'écart entre les résultats de la régression log-normale et les données observées. Pour le modèle *XG-Boost*, on observe un écart de 0,6 % entre les résultats de la modélisation et l'observé. L'écart a donc diminué de 65 %.
- Sur la charge ultime des sinistres graves, le constat est différent : on observe en valeur absolue 7 % d'écart entre la régression log-normale et l'observé contre 10 % avec le modèle *XG-Boost*. Dû au nombre faible d'observations, on pourrait adopter une approche différente pour ces sinistres graves en utilisant plutôt la liquidation moyenne de ces sinistres issues de la loi GPD.

19. Mise en production sur la base des sinistres en cours

Une fois les modèles validés, on les relance sur l'ensemble de la base de liquidation (sans distinction apprentissage et test) pour utiliser l'ensemble de l'information disponible. Ces modèles sont ensuite appliqués à la base des sinistres en cours sur laquelle on a effectué les mêmes transformations de variables :

- Discrétisation des variables continues en variables discrètes
- Traitement des valeurs manquantes
- …

On applique ensuite aux observations les deux approches :

- Modèle *XG-Boost* pour obtenir la probabilité de clôture à 0€ de la $PSAP_{ultime}$.
- Modèle *XG-Boost* pour obtenir le montant $PSAP_{ultime}$ en ajustant :

    - Un modèle « attritionnel » pour les sinistres dont la $PSAP_{dernière\ vision}$ est inférieure au seuil des graves.
    - Un modèle « grave » pour les sinistres dont la $PSAP_{dernière\ vision}$ est inférieure au seuil des exceptionnels mais supérieure au seuil des graves.
    - Pour les sinistres dont la $PSAP_{dernière\ vision}$ est supérieure au seuil des exceptionnels : on conserve la valeur actuelle, ces sinistres sont surveillés de manière régulière pour adapter cette valeur.

La mise en place de cette nouvelle méthode permettant de prouver les apports du *Machine Learning* sur des problématiques d'estimation de la provision ultime a été partagée, puis adoptée, par les différents interlocuteurs à savoir : actuariat, pilotage et souscription. La base des sinistres en cours, dont la provision est poussée à l'ultime par le modèle *XG-Boost*, est ensuite intégrée à l'outil qui servira à effectuer le renouvellement des grands comptes depuis l'année 2022.

L'étude de la sinistralité d'un grand compte peut être encouragée par :

- L'approche de la date de fin de *LTA* (***Long Term Agreement***) et donc la possibilité d'un nouvel appel d'offres qui impliquerait des propositions concurrentes.
- Une demande d'évolution sur le contrat actuel : par exemple en termes de plafond de garantie ou de franchises.
- …

L'Actuariat fournit une note reprenant le contexte de la demande autour du grand compte, les statistiques observées notamment les courbes de liquidation, les résultats de la modélisation *XG-Boost*, ainsi qu'une régression linéaire simple ayant vocation à prendre en compte la partie des sinistres *IBNR*. L'outil permet ainsi de proposer une estimation de la charge sinistre ultime puis de calculer une prime pure à hauteur de la charge sinistre ultime prédite en intégrant les différents chargements : frais généraux, réassurance et commissions.

Ci-dessous, un exemple d'écran fourni par l'outil *RShiny*, et par la suite intégré dans la note finale, en abscisse sont représentés les exercices et en ordonnée les niveaux du rapport $\frac{Sinistres}{Cotisations}$:

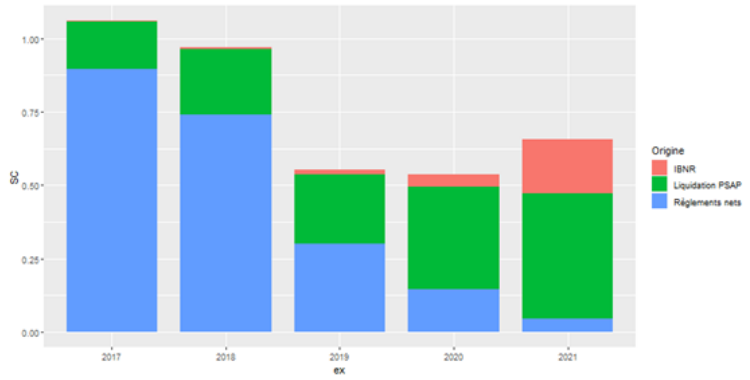


Figure 45 - Mise en production de la méthode sur la base des sinistres en cours

- La partie bleue représente le niveau des règlements nets déjà effectués sur les dossiers clos, mais également sur les dossiers en cours.
- La partie verte représente le montant de provision ultime estimée par le modèle *XG-Boost*
- La partie rouge représente la charge apportée par les sinistres *IBNR*.

20. Limites du modèle *XG-Boost*

Cette introduction du modèle *XG-Boost*, et plus généralement du *Machine Learning* pour une problématique de provisionnement, a montré un intérêt au point que son application a été introduite dès l'année 2022. Toutefois, le modèle présente des limites :

1. Premièrement, le fait d'omettre toute l'information disponible sur les sinistres en cours. En effet, ces sinistres, censurés à droite, ne peuvent pas être intégrés de la même manière que les sinistres clos, car leur niveau de liquidation est susceptible d'évoluer, mais pour autant ne pas les considérer dans la modélisation nous amène à ne pas profiter de l'information qu'ils apportent. L'histogramme de la partie 9.3 nous montre que la base des sinistres en cours contient des dossiers anciens, en proportion plus importante que sur la base des sinistres clos. Ces sinistres, une fois leurs procédures judiciaires ou la résolution amiable terminée, peuvent venir dégrader des exercices anciens. De plus, nous savons que la forte judiciarisation associée aux engagements importants sur cette garantie implique une liquidation moyenne de plusieurs années. D'ailleurs, on observe un écart de 30% entre le délai de traitement moyen sur les sinistres en cours et sur les sinistres clos.
   Ainsi, sur les exercices récents, seuls les dossiers à la résolution facile sont clos, les dossiers complexes présentant une liquidation plus longue notamment à cause de la judiciarisation. En effet, pour les sinistres ouverts en 2022 seul 11 % des dossiers sont clos. Ces dossiers clos présentent un taux de judiciarisation de 11 % contre un taux de judiciarisation de 32 % pour les sinistres ouverts en 2022 toujours en cours. Pour confirmer ces intuitions observons les taux de judiciarisation sur les sinistres clos et sur les sinistres en cours :

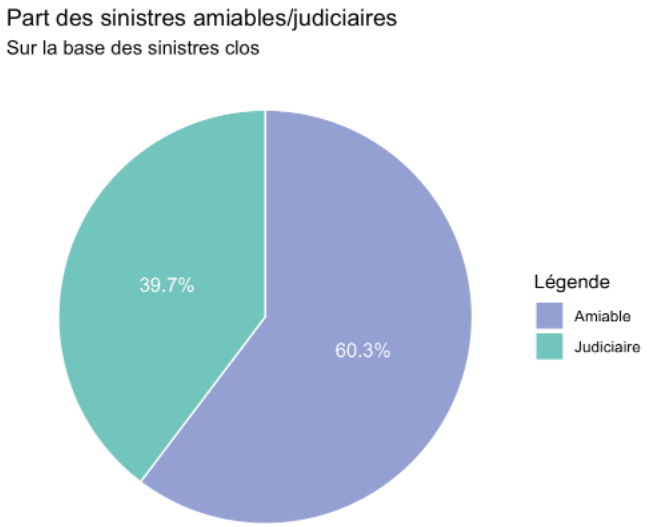


Figure 46 - Taux de judiciarisation sur la base des sinistres clos

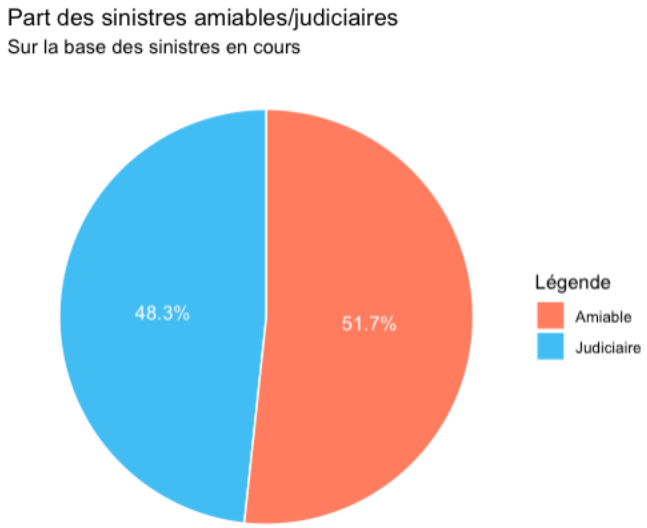


Figure 47 - Taux de judiciarisation sur la base des sinistres en cours

Les deux graphiques démontrent une proportion plus importante de sinistres judiciaires sur la base des sinistres en cours. Outre la judiciarisation, on a tendance également à admettre que la charge est plus importante sur les sinistres dont le traitement n'est pas terminé. Considérons les distributions de charges suivantes (afin de pouvoir observer correctement les résultats, on observe d'une part les sinistres attritionnels et d'autre part les sinistres graves) :

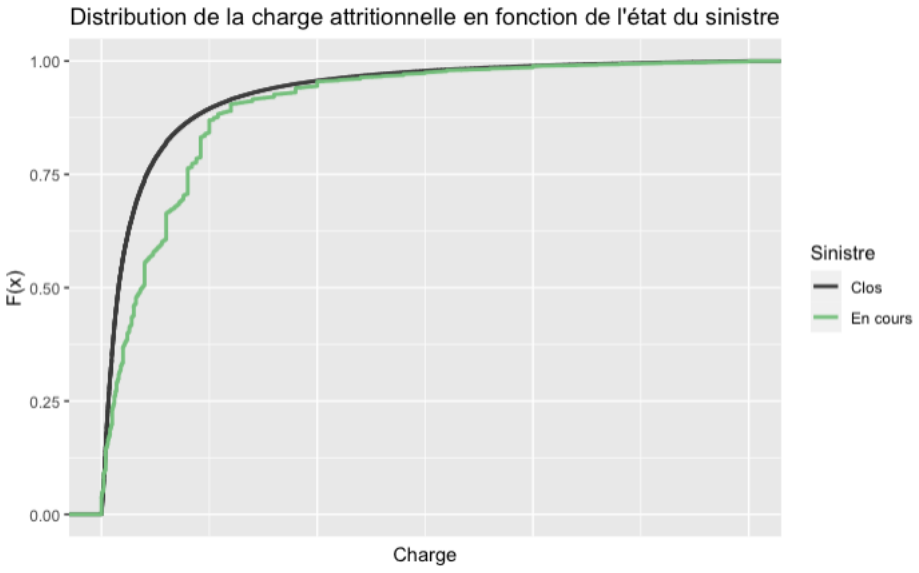


Figure 48 - Distribution de la charge des sinistres attritionnels selon leur état

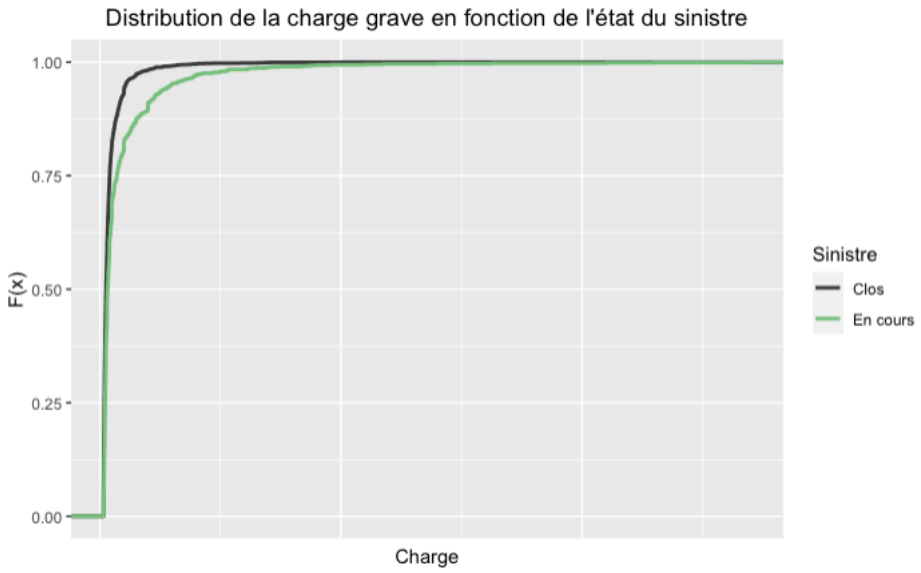


Figure 49 - Distribution de la charge des sinistres graves selon leur état

La distribution verte étant légèrement inférieure, on en conclut que la charge sur la base des sinistres en cours est plus importante sur les sinistres attritionnels, mais également sur les sinistres graves. Pour confirmer cette observation, un test statistique est réalisé : le test de Wilcoxon-Mann-Whitney[41]. Le test de Wilcoxon-Mann-Whitney permet de tester l'hypothèse selon laquelle des observations issues de deux groupes différents suivent une même distribution. Ce test est un test non-paramétrique et donc une bonne alternative au test de Student. De plus, c'est un test performant même sur des échantillons petits.

Contrairement aux autres tests, la statistique de test repose sur les rangs des observations. Ces rangs correspondent à la position de chaque observation dans la série, une fois les données des deux groupes mélangées et triées par ordre croissant. En considérant les rangs au lieu des valeurs des observations cela permet au test d'être moins influencé par la présence de valeurs extrêmes. Notons ces groupes A et B, la statistique de test, notée U, est définie comme :

$$U = min\left(n_A n_B + \frac{n_A(n_A+1)}{2} - R_A, n_A n_B + \frac{n_B(n_B+1)}{2} - R_B\right)$$

Avec R_A la somme des rangs du groupe A et R_B la somme des rangs du groupe B.

Ce test permet ensuite d'accepter ou de rejeter l'hypothèse nulle H_0 selon laquelle les observations des deux groupes suivent une même distribution. Dans notre cas, on testera donc si l'état du sinistre, « clos » ou « en cours », à un effet sur la distribution du montant de charge nette des sinistres graves et des sinistres attritionnels. Les p-values obtenues sont inférieures à 0,05, on peut donc rejeter l'hypothèse nulle selon laquelle les charges suivent une même distribution suivant l'état de traitement du dossier. En particulier, la charge

[41] Mann Whitney U Test (Wilcoxon Rank Sum Test), Université de Boston

moyenne observée sur la base des sinistres en cours est 3 fois plus importante à la charge observée sur les sinistres clos. Cependant, sur les sinistres dont le traitement n'est pas encore terminé, on s'attend à observer des mouvements au niveau de la charge et donc de la provision, mais ceux-ci ne doivent pas être pour autant écartés de la modélisation, car ils sont porteurs d'information.

2. La seconde limite, qui est la plus structurante, concerne l'indépendance entre les observations lors de la construction de la base servant à entraîner le modèle de liquidation. Afin d'appliquer le modèle en ligne à ligne, on transforme la base initiale contenant les différentes visions organisées en colonne. En réorganisant la base de liquidation des sinistres clos en ligne sans pour autant prendre en compte le lien temporel existant entre les visions d'un même sinistre, on considère que les différentes visions d'un même sinistre sont des observations indépendantes.

Ces limites me poussent à considérer une autre méthode permettant de prendre en compte l'intégralité de l'information disponible : en intégrant les dossiers en cours ainsi que les différentes visions des variables comme la nature juridique, la présence d'un règlement accessoire ou principal.

21. Résultats du modèle « *weighted CART* »

La partie théorique du modèle « *weighted CART* » est disponible dans la partie 13.4.

21.1. Construction de la base de données

Première étape nécessaire pour réaliser la modélisation à l'aide des arbres *CART* pondérés, ou *weighted CART*, est la constitution de la base de données.

En effet, pour appliquer les deux premières méthodes nous avions scindé les données sur la base du critère de clôture du dossier sinistre, construisant ainsi une base contenant les sinistres clos et une base regroupant les sinistres en cours. L'application de cette méthode étant motivée par la prise en compte des sinistres en cours dans la modélisation, il est nécessaire de construire une nouvelle base de données plus adaptée.

La méthode présentée dans Lopez et al.[32] (2016), Lopez et al.[33] (2019) et Lopez et al.[34] (2020) est un cadre général se basant sur les dernières visions des variables explicatives. Toutefois, on peut très facilement adapter cette méthode pour l'alimenter des différentes visions des variables explicatives. Ainsi, pour chaque ancienneté atteinte, on pourra fournir à l'algorithme la vision de la variable explicative à cet instant. Les informations disponibles dans les bases de données sinistres me permettront ainsi de récupérer pour les variables listées ci-dessous les différentes visions à chaque fin d'année :

- La nature juridique depuis 2011 (pour les sinistres ouverts avant cette date, nous n'avons pas l'information.)
- La présence d'un règlement principal.
- La présence d'un règlement accessoire.
- Le montant de la provision.

L'objectif étant de permettre au modèle d'identifier l'effet sur la variable cible des changements de nature juridique ou de règlement sur le sinistre. Pour toutes ces raisons, une nouvelle base doit donc être constituée :

- Contenant l'ensemble des sinistres : clos ou en cours
- Contenant les visions de chaque variable explicative afin de récupérer, pour toutes les observations, la première vision, puis la seconde, et ainsi de suite jusqu'à la dernière vision.

En d'autres termes, à partir de la base sinistre constituée ainsi :

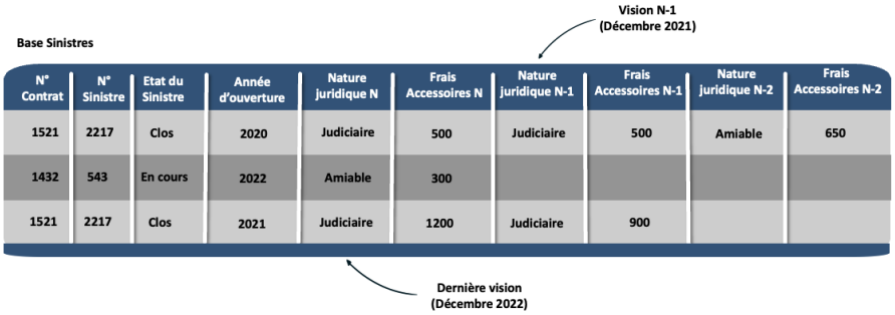


Figure 50 - Extraction de données initiale

On crée la base de modélisation qui nous permettra de prédire, pour chaque vision le montant final du sinistre :

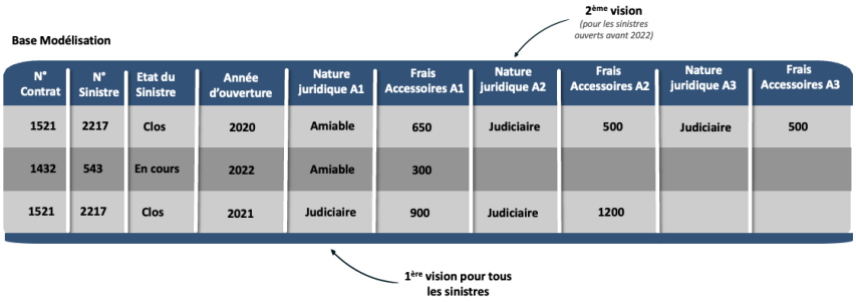


Figure 51 - Construction de la base de données pour la modélisation wCART

Comme nous l'avons décrit dans la partie 13.2 la présence d'observations censurées nous amène à ajouter des variables. La première variable à ajouter est la variable δ qui prendra la valeur 1 si la période d'observation du sinistre est complète (*i.e.* les sinistres clos) et 0 si le sinistre est en cours. Cette variable sera utile dans la suite pour calculer les poids de Kaplan-Meier, mais également la variable N. Comme nous n'observons pas le montant final de la provision pour les sinistres en cours, on crée la variable N qui, *a fortiori*, sera égale à 0 si le sinistre est en cours et au montant final de provision dans le cadre d'un sinistre clos. Enfin, pour les mêmes raisons, la durée de liquidation du sinistre sera représentée par la variable Y. Pour les sinistres clos la variable Y sera égale à la durée de la gestion du dossier, et pour les sinistres en cours Y représentera le temps écoulé entre l'ouverture du dossier et la date de l'extraction.

21.2. Modèle de survie

Après avoir échantillonné la base de modélisation contenant environ 230 000 lignes au départ en une base d'apprentissage contenant 160 000 lignes et une base de test contenant 70 000 lignes, on commence par le modèle de survie qui nous permettra par la suite de calculer nos poids de Kaplan-Meier. Pour cette modélisation, j'utilise le package *survival* sur R et particulièrement la fonction *survfit* qui permet de calculer une courbe de survie pour des données censurées à l'aide de la méthode de Kaplan-Meier détaillée dans la partie 13.3.1. Les courbes obtenues de l'estimateur de Kaplan-Meier sont un outil graphique performant nous permettant de représenter la probabilité d'un événement par rapport à un intervalle de temps. Dans notre étude, l'événement est représenté par la clôture du dossier.

Le modèle tente d'expliquer l'ancienneté du sinistre en années, représentée par la variable Y, suivant son état censuré ou non $(\delta = 0 \; ou \; \delta = 1)$ en fonction des variables explicatives fournies. Voici un exemple de courbe de survie obtenue lorsque la variable explicative choisie est la nature juridique :

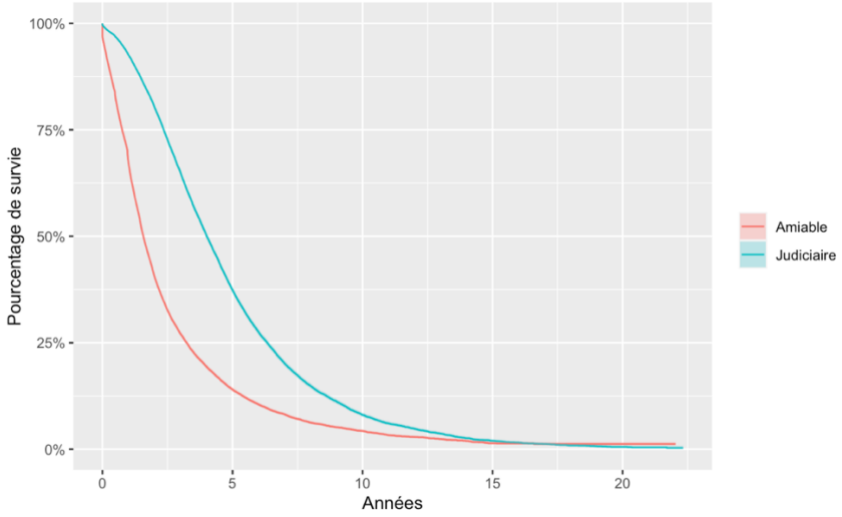


Figure 52 - Courbe de survie en fonction de la nature juridique du sinistre

Le modèle de survie semble bien confirmer les constats que nous avions déjà fait auparavant : les sinistres judiciaires ont une probabilité de survie supérieure aux sinistres amiables. Si la probabilité que le traitement d'un dossier amiable dure 5 ans est d'environ 12 %, elle est d'environ 37 % pour un sinistre judiciaire. Le modèle nous permet également de produire un tableau reprenant quelques indicateurs statistiques. Pour chaque modalité de la variable explicative, on y retrouve la quantité *median* qui nous donne le temps de survie médian, c'est-à-dire le temps pour lequel la probabilité de survie $S(t)$ vaut 0,5. Voici les résultats obtenus pour ce modèle simplifié :

Nature Juridique	Median*
Judiciaire	6
Amiable	2,25

**Les chiffres ont été modifiés*

Ce résultat confirme l'interprétation graphique faite depuis les courbes de survies : les sinistres judiciaires présentent une durée de vie plus longue, comparativement à un sinistre amiable.

Cette observation doit cependant être confirmée par un test statistique afin de voir si la différence entre les deux types de sinistres est statistiquement significative. Le test de *« Log-Rank »* est le test le plus utilisé pour comparer des courbes de survie entre elles ; l'hypothèse nulle suggère que les deux courbes de survie que l'on notera $S_{Amiable,t}$ et $S_{Judiciaire,t}$ sont identiques pour tout temps t. Autrement dit, on a :

$$\begin{cases} H_0 : S_{Amiable,t} = S_{Judiciaire,t} \\ H_1 : S_{Amiable,t} \neq S_{Judiciaire,t} \end{cases}$$

La statistique de test calculée par la fonction *survdiff* de R est une généralisation du test log-rank proposée par Harrington et al.[42] (1982) :

$$\frac{\sum_{i=1}^{D} \omega_i (d_{Amiable,i} - E_i)}{\sqrt{\sum_{i=1}^{D} \omega_i^2 V_i}} \xrightarrow{\mathcal{L}} \chi^2(1)$$

Avec :

- Avec $t_i, i = 1,2, \dots D$ les temps tels que $t_1 < t_2 < \cdots < t_D$
- $d_{Amiable,i}$ le nombre observé de sinistres amiables clos
- E_i l'espérance du nombre de sinistres clos
- V_i la variance du nombre de sinistres clos
- $\omega_i = \hat{S}_i^{\rho}$ famille G-rho d'Harrington et Fleming

La p-value obtenue pour le test Log-rank est inférieure à 0,05 : on rejette donc l'hypothèse nulle selon laquelle il n'y a pas de différence significative entre les deux courbes de survie.

On complète finalement le modèle avec les autres variables explicatives dont nous disposons :

- La profession croisée de la circonstance sinistrée,
- La présence d'un règlement accessoire ou principal,
- L'ouverture du sinistre au forfait ou non.

Ce modèle étant entraîné sur l'intégralité de la base d'apprentissage en amont de la procédure de *weighted CART*, on conserve uniquement la dernière vision de chaque variable pour chacune des observations.

21.3. Poids de Kaplan-Meier

A partir du modèle de l'estimateur de Kaplan-Meier construit lors de l'étape précédente, nous déduisons les probabilités de survie. Ces probabilités de survie nous permettent ainsi d'obtenir les poids qui seront intégrés dans les arbres *CART* grâce à la formule suivante pour toute observation i tel que $i \in \{1, n\}$:

[42] Harrington et al. (1982)

$$\omega_i = \frac{\delta_i}{n\left(1 - \hat{G}(Y_i -)\right)}$$

Avec :

- $\delta_i = 1$ pour un sinistre clos et $\delta_i = 0$ pour un sinistre en cours.
- n le nombre d'observations.
- Y_i la durée de vie.
- $\hat{G}(Y_i -)$ l'estimateur de Kaplan-Meier de la fonction de répartition de la variable C définit par $G(t) = \mathbb{P}(C \leq t)$ obtenu à partir de notre modèle.

Ces poids de Kaplan-Meier ont pour objectif d'accorder un poids plus important aux observations complètes ayant la durée de traitement la plus longue. En effet, les sinistres que l'on a totalement observés pendant une période importante sont rares, mais sont bien présents dans la base des sinistres en cours.

En observant les poids de Kaplan-Meier obtenus avec le modèle sur les sinistres clos (ceux-ci étant nuls sur les sinistres en cours), on obtient le graphique suivant :

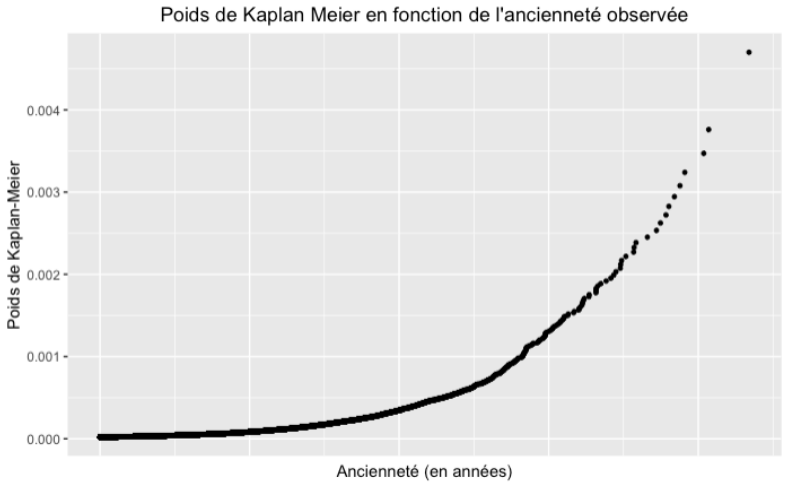


Figure 53 - Poids de Kaplan-Meier

Plus le sinistre est ancien plus le poids qui lui est accordé est important.

21.4. Algorithme de prédiction de la PSAP ultime

21.4.1. Description de l'algorithme

Nous avons décrit dans la partie 13.4 les éléments théoriques nécessaires à la mise en place de l'algorithme *weighted CART*, simplifié en *wCART.* Voyons désormais quelles sont les étapes de l'algorithme dont l'objectif est de prédire la provision pour sinistres à payer pour chaque dossier. L'algorithme initial proposé dans Lopez et al.[33] (2019) est appliqué au montant final du sinistre, dans le cadre de notre étude nous adaptons la variable cible en remplaçant le montant final du sinistre par la provision pour sinistres à payer finale. En effet, notons M le montant final du sinistre. Cette quantité se décompose comme étant :

$$M = Réglements_{observés} + PSAP - Recours_{observés}$$

Les quantités $Réglements_{observés}$ et $Recours_{observés}$ sont déterministes, nous disposons donc de la valeur réelle de ces quantités (même pour les sinistres en cours). Il est préférable de lancer l'algorithme uniquement sur la partie aléatoire de ce calcul soit la quantité $PSAP$.

21.4.1.1. Algorithme simplifié

Commençons par introduire simplement l'algorithme à l'aide d'un cas simplifié : pour chaque jour où le dossier sinistre reste ouvert, la PSAP augmente de 1€. Autrement dit, dans ce cas particulier, la PSAP finale correspondra à la durée de vie résiduelle, c'est-à-dire $P = T$. Si l'observation est censurée (i.e. le sinistre est toujours en cours) alors il existe un lien direct entre la durée de vie du sinistre et la PSAP ultime. En effet, dans le cadre d'un sinistre en cours dont la durée de vie est $Y = k$, on a $\delta = 0$ et la PSAP finale que l'on souhaite prédire, P^*, est définie par la quantité suivante :

$$P^* = \mathbb{E}[T|\delta = 0, Y = k, X = x] = \mathbb{E}[T|T \geq k, X = x]$$

avec X les variables explicatives.

Soit k_i la durée de vie de la $i^{ème}$ observation.

- **1ère étape** : calcul des poids de Kaplan-Meier sur l'ensemble de la base de données. Cette étape est réalisée une unique fois au début de la procédure. Calculer les poids sur l'ensemble des observations permet d'assurer la consistance de l'estimateur.
- **2ème étape** : pour chaque durée de vie atteinte, on filtre les données en ne conservant que les observations dont la durée de vie est supérieure à k_i. Précisions ici que nous conservons également les sinistres en cours soit censurés à droite.
- **3ème étape :** prédiction de la durée de vie résiduelle $T - k_i|X = x, T > k_i$ à l'aide d'une procédure *CART* adaptée afin de pouvoir y introduire les poids de Kaplan-Meier.
- **4ème étape :** optimisation de l'arbre obtenu à l'aide d'une procédure de *pruning*.
- **5ème étape :** Prédiction de la durée de vie résiduelle représentée par la quantité $\mathbb{E}[T - k_i|T > k_i, X = x]$
- **6ème étape :** passage à la durée de vie atteinte suivante puis retour à l'étape 2.

21.4.1.2. Algorithme complet

Nous avons vu comment fonctionnait l'algorithme simplifié dans la sous partie précédente, voyons désormais quelles sont les étapes à suivre lorsque la provision n'augmente pas de manière constante pour chaque durée de vie atteinte. Dans le cadre simplifié $P = T$, mais dans un cadre plus général, on a :

$$P^* = \mathbb{E}[P|\delta = 0, Y = k, X = x] = \mathbb{E}[P|T \geq k, X = x]$$

A l'aide de l'approche bayésienne on obtient :

$$P^* = \mathbb{E}[P|T \geq k, X = x] = \frac{\mathbb{E}[P1_{T \geq k}|X = x]}{\mathbb{E}[1_{T \geq k}|X = x]}$$

Finalement pour prédire la quantité P^*, deux arbres *wCART* sont entraînés : un pour le numérateur et un pour le dénominateur.

21.4.1.3. Commentaires sur l'algorithme

Au regard de la procédure présentée dans les sous-parties 21.4.1.1 et 21.4.1.2 on identifie rapidement que, pour les sinistres aux durées de vie les plus longues, l'algorithme risque de ne pas être performant. En effet, au fur et à mesure des itérations la base sélectionnée au départ est de plus en plus diminuée, dû à la faible représentation des sinistres à la gestion longue. Ainsi les sinistres ayant une durée de vie supérieure à 10 ans ne représentent que 1,5 % des sinistres clos.

Notons également que d'autres méthodes peuvent être substituées à l'approche bayésienne pour estimer la quantité P^*comme l'approche de plug-in décrite dans la suite du papier. Cette méthode admet qu'une quantité donnée peut être approchée par la même quantité calculée à partir d'un échantillon d'observations tirées de la distribution de départ. Dans le cadre de la procédure *wCART* cela implique d'entraîner un premier modèle noté $\hat{\pi}$ pour estimer $\mathbb{E}[P|T = t, X = x]$ puis entraîner un deuxième modèle pour prédire la quantité $T|T \geq k, X = x$, prédiction que l'on notera $\hat{T}(k, x)$. Enfin, afin d'obtenir la prédiction finale P^*on applique le principe de plug-in :

$$\widehat{P^*} = \hat{\pi}\left(\hat{T}(k, x), x\right)$$

Plusieurs approches peuvent ensuite être utilisées pour prédire $\hat{T}$: par exemple, on pourrait appliquer la méthode de prédiction utilisée pour prédire le montant de la PSAP en remplaçant la quantité P par la durée de vie observée T. Dans le cadre de cette étude, seul l'approche bayésienne est considérée pour prédire le montant ultime de provisions.

21.4.2. Procédure *wCART*

21.4.2.1. Variables explicatives

Une fois les poids de Kaplan-Meier obtenus, et la méthode d'estimation de la provision ultime définie, les premiers arbres peuvent être entraînés. Les arbres *CART* présentent un avantage important comparativement aux modèles de régressions : la gestion des valeurs manquantes et des corrélations est faite de manière automatique. Ainsi, ces étapes de pré-traitement de la base de données ne sont pas nécessaires. Nous pouvons ainsi fournir au modèle l'ensemble des variables explicatives dont nous disposons à savoir :

- Des variables numériques discrètes comme : le montant d'évaluation à l'origine.
- Des variables binaires : sinistres corporels (oui/non), ouverture au forfait (oui/non), la profession croisée de la circonstance sinistre (pour des raisons de confidentialité cette variable est anonymisée)

Les variables nature juridique du sinistre, présence d'un règlement principal, présence d'un règlement accessoire sont des variables qui peuvent être amenées à changer en fonction de la vision du dossier sinistre. Un historique d'une dizaine d'année est disponible pour ces variables permettant de les reconstruire pour chaque vision du dossier sinistre, et cela, pour tous les sinistres ouverts après 2010. L'algorithme étant appliqué pour chacune des durées de vie atteinte par le dossier, il y a un réel intérêt de fournir aux deux arbres *wCART* les valeurs de ces variables pour chaque vision.

En effet, on permet au modèle d'identifier d'éventuels changements dans le montant de la provision liés par exemple :

- La judiciarisation d'un dossier.
- La présence d'un paiement sur le dossier hors frais accessoires.
- La présence d'un paiement lié à des frais accessoires sur le dossier sinistre.

21.4.2.2. Arbres optimaux et *pruning*

Après avoir défini l'algorithme utilisé ainsi que les variables explicatives du modèle, on peut lancer la procédure. Voici les arbres *wCART* prédisant la durée de vie résiduelle obtenue pour la première année de vie du sinistre sans optimisation des paramètres :

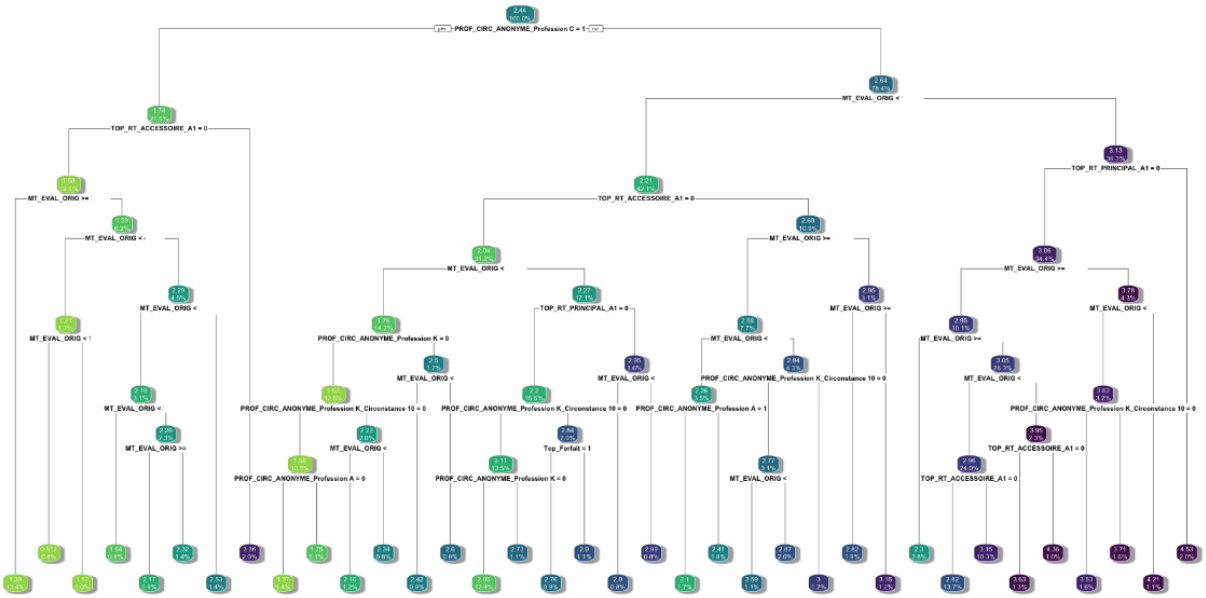


Figure 54 - Arbre CART de la durée de vie résiduelle avant pruning

L'arbre obtenu avant l'étape de *pruning*, ou élagage, présente un nombre important de nœuds et de feuilles. Un arbre aussi segmentant n'est pas optimal : avoir aussi peu d'observations par feuille peut être à l'origine de sur-apprentissage. L'étape de *pruning* permet de retirer de l'arbre obtenu les « branches » de l'arbre *CART* n'améliorant pas la performance du modèle. Deux stratégies sont possibles afin d'élaguer l'arbre (celles-ci sont détaillées dans la partie « L'optimisation des arbres ».) :

- soit en arrêtant la création de l'arbre dès lors qu'il n'y a plus d'amélioration de la performance : on parle alors de *pre-pruning* ou *early-stopping*.
- soit après la création de l'arbre : on parle alors de *post-pruning*.

La méthode retenue pour optimiser l'arbre est la seconde méthode. Plus précisément, le critère retenu pour cette étape de *pruning* est la minimisation de l'erreur de cross-validation. Le paramètre à optimiser est noté CP ou *Complexity Parameter*. Il est défini comme étant la quantité minimale d'amélioration que le modèle doit obtenir à chaque nœud.

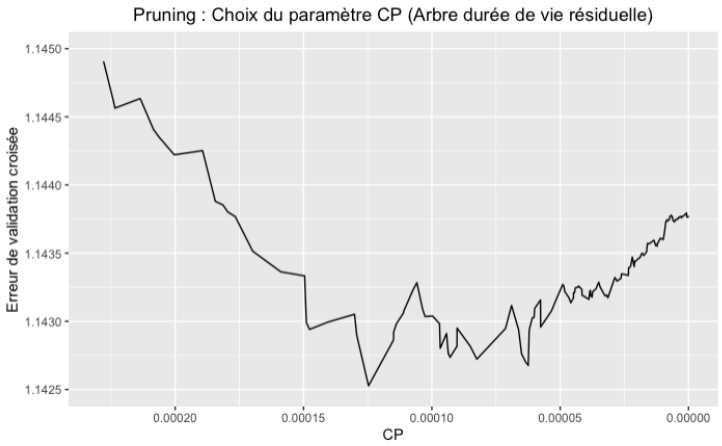


Figure 55 - Optimisation du paramètre de complexité

Le graphique ci-dessus représente l'erreur de validation croisée obtenue en fonction du *Complexity Parameter.* Le CP qui minimise l'erreur de validation croisée est 0.000125. On relance ainsi le modèle avec ce paramètre optimisé :

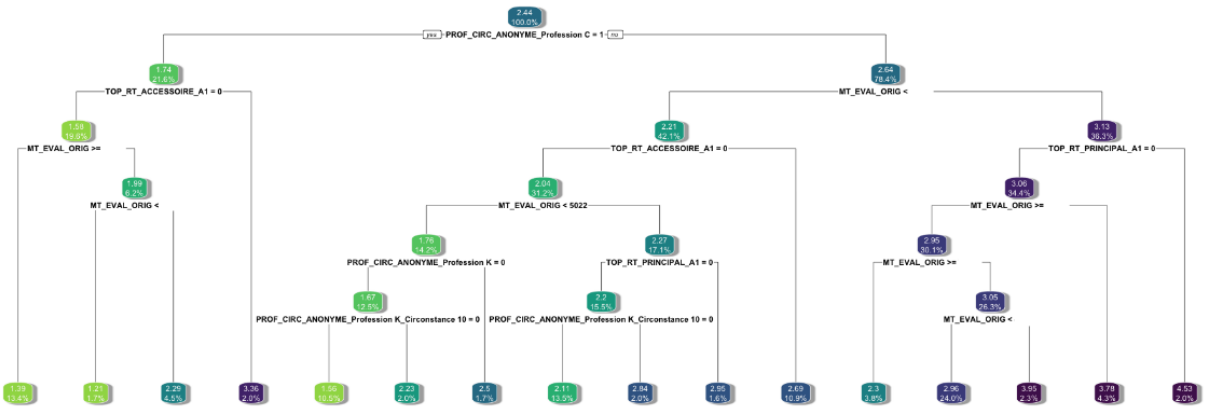


Figure 56 - Arbre de la durée de vie résiduelle après pruning

On obtient un arbre relativement moins important : la profondeur de l'arbre est réduite ainsi que le nombre de feuilles obtenues. L'étape de *pruning* a permis de retirer du premier arbre les « branches » ne permettant pas d'améliorer la prédiction du modèle d'au moins 0,000125.

21.4.2.3. Importance des variables

Les algorithmes de *Machine Learning* apparaissent souvent comme de véritables « boîtes noires » rendant leur exploitation difficile. Les arbres *CART* présentent un avantage : leur interprétabilité est rendue possible à la suite de nombreux travaux autour de l'importance des variables par exemple. La structure de l'arbre nous renseigne déjà sur l'importance de chaque variable : en effet plus la variable étant à l'origine d'un *split* est proche de la racine plus elle est déterminante dans le modèle. Breiman et al.[20] propose, dans l'ouvrage introduisant les arbres *CART* au début des années 80, un indice plus précis reposant sur la notion de découpes de substitution. Pour cela, on observe comment se comporte l'erreur de prédiction du modèle lorsque le lien existant entre la variable explicative et la variable cible est rompue.

Il propose notamment d'effectuer des permutations sur la variable explicative afin de briser ce lien. Formalisons cette notion en reprenant les notations de Ghattas[43] (2006) : la construction d'un arbre de régression repose sur un algorithme itératif recherchant des règles de division, ou *split*, binaire. Notons $d = d(x_m, s)$ cette règle, avec x_m une réalisation de la $m^{ième}$ composante du vecteur des covariables $X = (X_1, X_2, \dots, X_m, \dots, X_n)$. La règle de division d partage notre échantillon de données au nœud t en deux sous-ensembles t_g et t_d (selon que $x_m \leq s$ ou que $x_m > s$ avec $s \in \mathbb{R}$). Il existe alors de multiples découpages possibles, la procédure va donc favoriser le *split* en maximisant l'indice $\Delta\hat{R}(d,t)$. Cet indice est défini comme suit : $\Delta\hat{R}(d,t) = \hat{R}(t) - \hat{R}(t_d) - \hat{R}(t_g)$ avec $\hat{R}(t)$ l'erreur quadratique moyenne définit comme : $\hat{R}(t) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_t})^2$ avec n le nombre d'observations contenues dans l'échantillon initial. La division de substitution $\tilde{d}_m$ est parmi l'ensemble des divisions possibles sur $X_m$celle maximisant la probabilité de prédire la division optimale notée d^*.

Finalement, l'importance de la variable X_m dans l'arbre de régression, C est donnée par :

$$I(X_m) = \sum_{t \in C} \hat{R}\left(\tilde{d}_m(t), t\right)$$

Avec $\tilde{d}_m(t)$ la division de substitution au nœud t reposant sur la m[ième] variable. Autrement dit, l'importance de chaque variable est calculée comme étant la somme des diminutions de l'erreur quadratique moyenne si on brisait le lien entre la variable à expliquer et la variable explicative en remplaçant à la règle d par la règle de substitution $\tilde{d}$.

Les figures ci-dessous représentent l'importance des variables obtenue pour la première vision de liquidation :

- Pour l'arbre prédisant la durée de vie résiduelle :

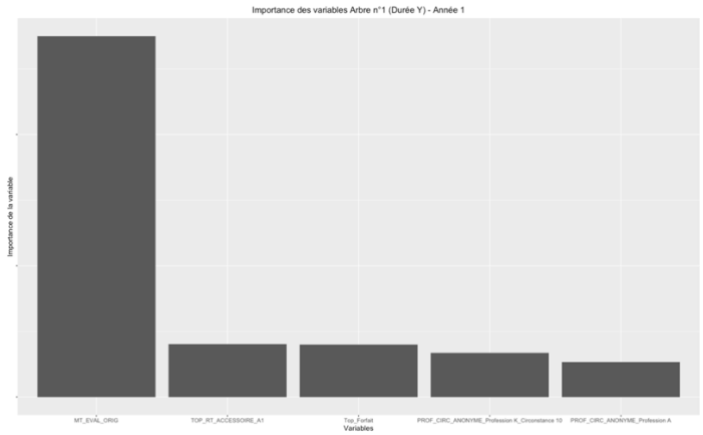


Figure 57 - Importance des variables : Arbre de durée résiduelle (année 1)

[43] Ghattas (1999)

La variable ayant le plus d'importance dans le modèle, pour cette première année de liquidation, est le montant d'évaluation d'origine. Ce montant, positionné à l'ouverture du dossier, correspond soit :

- à un niveau défini au préalable en cas de manque d'informations. On parle alors de forfait d'ouverture.
- pour les dossiers disposant d'informations de l'évaluation, à hauteur d'une première expertise.

On sait que les sinistres importants ont une gestion plus longue. On peut confirmer cette idée en affichant le montant moyen du niveau d'évaluation d'origine des sinistres par durée de liquidation :

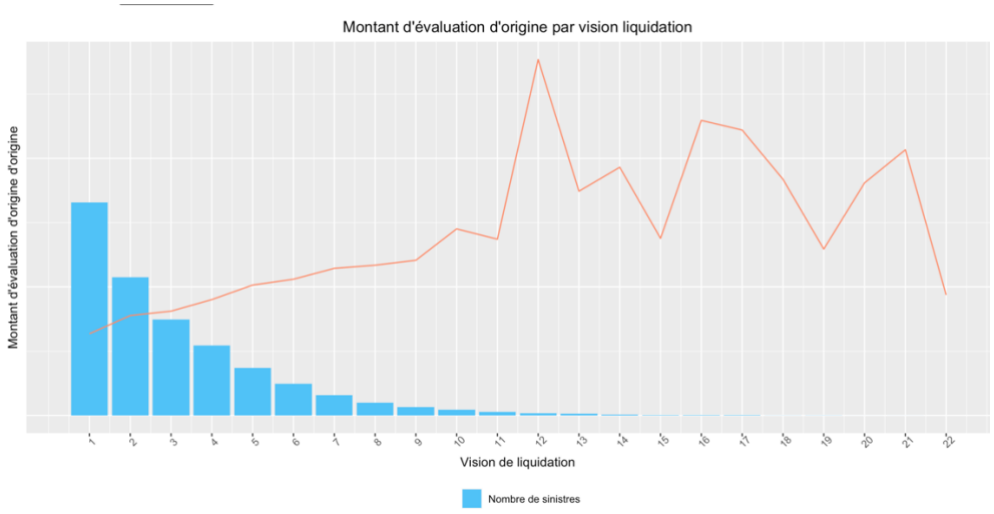


Figure 58 - Montant d'évaluation d'origine par vision liquidation

Ce graphique confirme cette intuition : plus la vision de liquidation est importante, plus la moyenne des montants d'évaluation d'origine des sinistres est importante.

- Pour l'arbre prédisant le montant final de provision :

Figure 59 - Importance des variables : Arbre du montant de provision (année 1)

Encore une fois le montant d'évaluation d'origine est la variable la plus importante cependant, contrairement à l'arbre prédisant la durée résiduelle, les tendances changent par la suite. Ce résultat semble plutôt logique : le montant d'évaluation a un impact certain dans la prédiction du montant final particulièrement la première année.

Les mêmes graphiques sur les années de liquidation suivantes sont proposés en Annexe.

21.4.3. Résultats du modèle *wCART*

21.4.3.1. Prédictions en fonction de l'exercice

Suite à l'étape de *pruning*, notre algorithme est fin prêt à être entraîné sur une base d'apprentissage composée de 70 % des observations de la base initiale choisies de manière aléatoire. Nous disposons désormais de deux arbres nous permettant de prédire, pour chaque année de vie des sinistres, le montant final de la PSAP. Afin d'apprécier la performance des modèles *wCART* on se propose d'appliquer les modèles sur la base de test contenant les 20 % d'observations restantes. Précisons que dans la base de validation, nous disposons de sinistres clos et de sinistres censurés. Sur les sinistres clos, il est possible de comparer la valeur prédite par le modèle avec la valeur observée, mais pour les sinistres censurés nous ne disposons pas de la provision ultime observée. Le graphique ci-dessous représente donc :

- Sur le diagramme en bleu le montant de la provision des sinistres clos observée,
- Sur le diagramme en orange le montant de la provision des sinistres clos prédite,
- Sur une courbe orange plus foncée le montant de provision finale prédite apportée par les sinistres censurés.

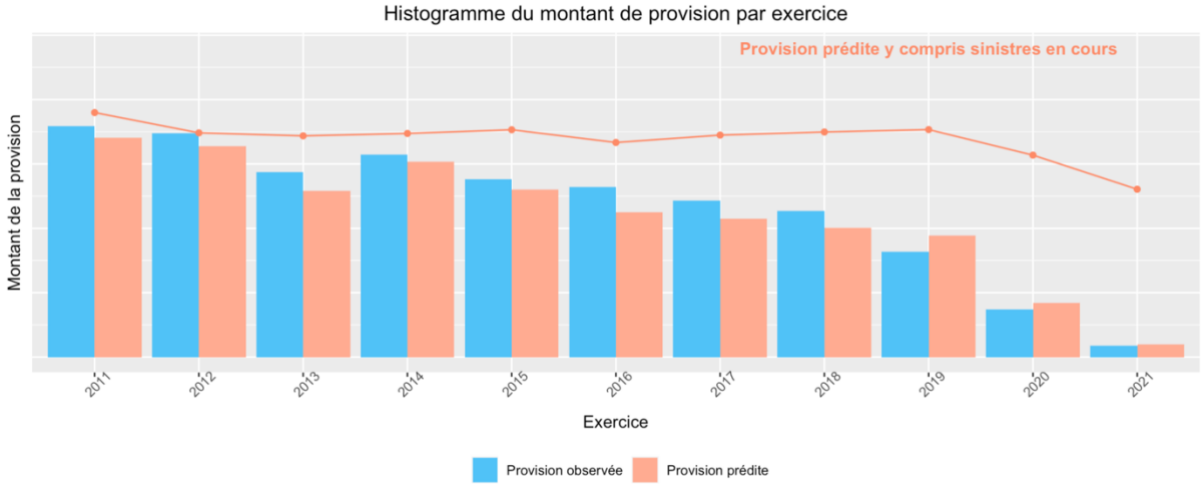


Figure 60 - Histogramme du montant de provision observé et prédit par exercice

La provision prédite sur les sinistres en cours semble décroître pour les exercices les plus récents, mais pour rappel, dans le cadre de ce mémoire on ne prédit que la provision des sinistres *RBNS* et aucun traitement n'est réalisé sur les sinistres *IBNR*. La part des sinistres *IBNR* est plus importante sur les exercices récents et vient compenser le nombre plus faible d'ouverture constatée de dossiers sinistres.

En observant les prédictions sur des sinistres comparables (i.e. clos), représentées par les histogrammes, on observe que l'algorithme semble plutôt correctement prédire la provision. Certains écarts sont observés sur quelques exercices, dus à la présence de provisions avec des montants plus importants.

Précisions que les observations extrêmes (provisions exceptionnelles à plusieurs millions d'euros définies dans 18.1.1) ont été retirées de notre base initiale : celles-ci nécessitent un traitement spécifique et ne peuvent être intégrées dans la modélisation au même titre que les autres observations.

21.4.3.2. Prédictions en fonction de la vision de liquidation

Cette partie vient compléter le point précédent afin de répondre à une question : est-ce que le modèle est plus performant sur certaines visions de liquidation et moins performant sur d'autres ? Pour cela, nous représentons les mêmes indicateurs que dans la partie précédente cette fois par vision de liquidation :

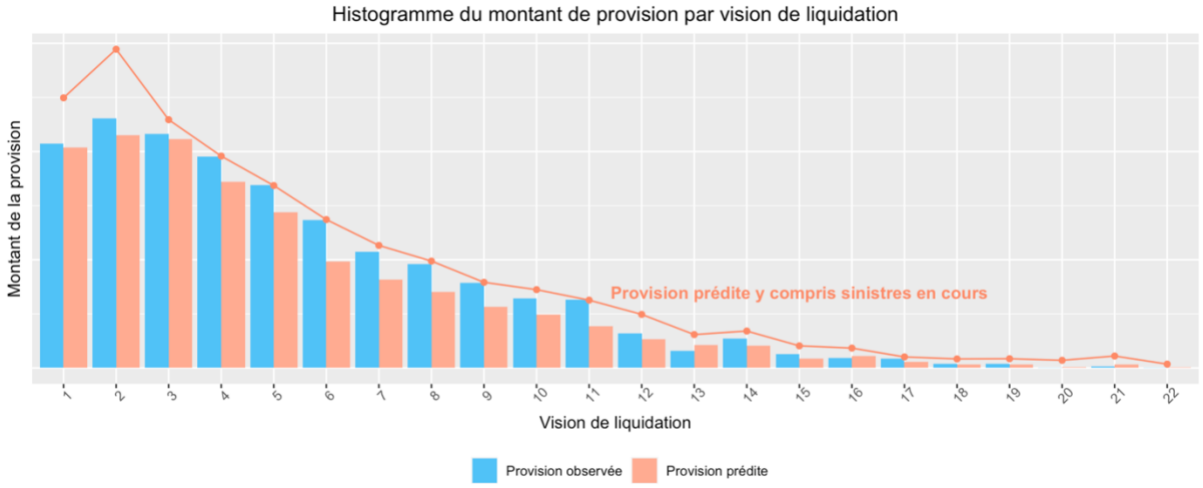


Figure 61 - Histogramme du montant de provision observé et prédit par vision de liquidation

Les montants prédits et observés sur les données non censurées semblent correspondre. On retrouve ce même phénomène d'augmentation en deuxième année de la provision que l'on observait sur les données constatées (cf. 9.2). Plus les observations ont une durée de liquidation importante, plus les modèles deviennent volatiles : en effet le manque d'observations impact fortement les résultats (seuls 2 % des observations censurées ont une durée de liquidation supérieure ou égale à dix ans.)

21.5. Comparaison des modèles

21.5.1. *RMSE* et *MAE*

Les mesures *RMSE* et *MAE* ont été définies dans la partie 18.4.1: nous avions exposé les avantages de considérer ces deux quantités. Suite à la mise en place de la méthode *wCART* nous pouvons désormais calculer, toujours sur la base de test, les deux indicateurs afin de les comparer avec les deux premières méthodes, voici les résultats obtenus :

	RMSE	MAE
Régression LogNormale	65 400	19 200
XG-Boost	59 060	16 670
wCART	43 747	13 478

Figure 62 - Comparaison des MAE et RMSE pour les 3 modèles

La prise en compte des observations censurées via le modèle *wCART* nous permet non seulement d'améliorer :

- Les résultats lorsque l'on accorde le même poids aux écarts : la *MAE* est réduite.
- Les erreurs de prédictions importantes : la *RMSE* est réduite de 25 % avec l'introduction de ce modèle.

21.5.2. Prédiction sur les sinistres en cours

On peut représenter pour les trois approches de modélisation les provisions ultimes prédites sur les sinistres en cours :

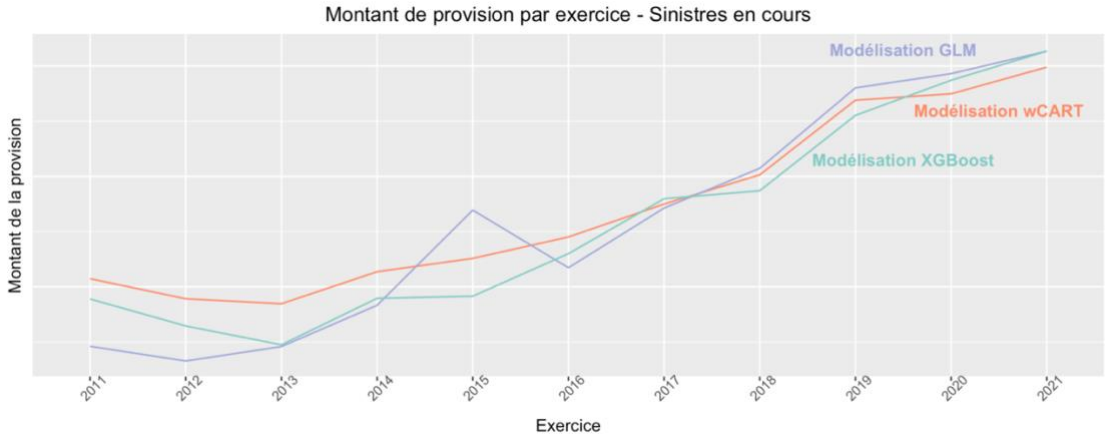


Figure 63 - Prédiction des modèles sur la base des sinistres en cours

Finalement, sur les exercices récents, les prédictions sont relativement proches entre les modèles. Remarquons tout de même que le montant ultime de provisions du modèle *wCART,* pour les sinistres en cours de l'exercice 2021, est légèrement inférieur aux prédictions des autres modèles. Toutefois, sur les exercices plus anciens, on observe que le modèle *wCART* prédit des provisions plus importantes que les autres modèles. En incluant dans notre modélisation les sinistres en cours, sur les exercices anciens cela nous amène à augmenter le montant des provisions à considérer. Ce résultat n'est pas étonnant. En effet, nous avions vérifié sur nos données observées le postulat selon lequel les sinistres importants ont une gestion plus longue. Notons également que pour l'exercice 2015, le modèle faisant appel aux régressions, logistique pour la fréquence et log-normale pour le coût, prédit un montant ultime étonnamment élevé. Ce résultat est dû à la présence d'un dossier important en cours ; le modèle de régression parvient à estimer ce montant à l'aide du modèle des graves, grâce à la propriété de moyennisation des modèles de régression. Cette provision étant trop extrême pour être modélisée, je fais le choix de placer directement le niveau observé de celle-ci. Les enjeux étant importants, ce dossier est régulièrement suivi et mis à jour par la direction Indemnisation.

Conclusion et Perspectives

La première partie de ce mémoire me permet d'introduire la typologie de la sinistralité : une forte judiciarisation impliquant une liquidation particulièrement longue. Déterminer l'issue d'un dossier peut s'avérer très difficile, et pour des besoins de suivi de la rentabilité de la garantie RC Pro nous devons tenir compte des provisions pour sinistres à payer dans notre charge de sinistralité à l'ultime. Pour cela, il est nécessaire de trouver le juste niveau de liquidation des $PSAP_{ultime}$ en prenant en compte le maximum d'informations disponibles sur chaque sinistre : c'est ce que permettent les méthodes de provisionnement en ligne à ligne.

La méthode existante est appliquée sur la base des sinistres clos uniquement. Elle s'articule en deux étapes, dues au nombre important de dossiers dont la $PSAP_{ultime}$ est liquidée à 0€. J'ai challengé la méthode de régression existante par un algorithme de *Machine Learning* : le *XG-Boost*. Après avoir effectué le traitement des données, j'ai amélioré les modèles de régression en appliquant une sélection automatique des variables explicatives. La première variable cible modélisée est la probabilité de liquidation de la $PSAP_{ultime}$ à 0 €. Pour ce problème de classification, la mise en place de la méthode *XG-Boost* m'a permis d'augmenter la précision de la prédiction de 6 points. La seconde variable modélisée, le montant de la $PSAP_{ultime}$, a nécessité de définir le seuil des sinistres graves. Après avoir obtenu ce seuil, deux modèles sont testés : un sur les sinistres attritionnels (i.e. dont le montant de la $PSAP_{ultime}$ est inférieur au seuil des graves) et un sur les sinistres graves. Sur les sinistres attritionnels, l'algorithme *XG-Boost* permet de diminuer l'écart entre les $PSAP_{ultime}$ observées et les $PSAP_{ultime}$ prédites de 65 % par rapport à la régression log-normale. Pour les sinistres graves, c'est la régression log-normale qui présente un écart plus faible que le *XG-Boost*. Les modèles, une fois validés, sont ensuite appliqués sur la base des sinistres en cours.

Cette première étape permet d'apporter une nouvelle méthode réduisant les écarts en l'appliquant de la même manière que la méthode précédente, c'est-à-dire en entraînant les modèles sur la base des sinistres clos. Elle prouve l'apport du *Machine Learning* pour des problématiques de provisionnement et, une fois présentée à l'ensemble des acteurs de la tarification sur le marché des PCD, est mise en production dès l'année 2022. Cependant, en faisant ce choix de modélisation, j'omets une partie de l'information conséquente portée par les sinistres en cours depuis plusieurs années et dont la liquidation à l'ultime de la provision peut avoir un impact considérable. J'oppose donc aux deux méthodes précédentes, une méthode permettant de tenir compte de ces sinistres en cours dans la modélisation. Cette méthode repose sur le calcul de poids de Kaplan-Meier obtenus depuis l'estimateur éponyme, et d'une procédure *CART* aménagée pour la censure. L'application de cette nouvelle méthode prouve l'intérêt de prendre en compte ces sinistres en cours puisque, sur les sinistres clos, la *MAE* diminue de 20% et la *RMSE* de 25%. Cette méthode permet une adaptation facile de l'algorithme à de nouvelles données ainsi qu'à de possibles dérives de sinistralité (liées à une évolution des pratiques de la profession par exemple).

Toutefois, perdurent quelques points d'attention notamment sur le traitement de l'inflation. Le sujet n'a pas été abordé au cours de ce mémoire, car il reste difficile d'apprécier correctement les effets de l'inflation sur le coût moyen des garanties de responsabilité civile professionnelle. En effet, pour les sinistres dommages aux biens, des indices permettent de prendre en compte cette inflation. On peut citer l'indice SRA (Sécurité et Réparation Automobile) pour les sinistres autos, ou encore l'indice BT01 regroupant différents indicateurs comme le coût de matériaux ou encore les frais de transport et permettant d'actualiser les coûts des sinistres de la construction. Sur la RC Pro, la nature des sinistres peut être totalement différente. Les coûts des sinistres ne représentent pas le même préjudice. Pour un agent immobilier par exemple, le coût du sinistre peut être équivalent au montant de la transaction immobilière sur laquelle porte le litige. Pour un sinistre lié à un défaut de conseil, engendrant un redressement fiscal, le client peut demander la prise en charge par le professionnel du montant des intérêts de retard au titre de sa RC Pro. Face à la multitude de causes possibles de sinistre, aucun indice ne permet de tenir compte, efficacement et de manière homogène sur l'ensemble des professions, du contexte inflationniste. Deuxième point d'amélioration, la méthode Lopez et al.[32] comme évoqué précédemment, a été élargie en 2020 afin de tenir compte de la troncature à gauche, autrement dit elle permet également le traitement des sinistres *IBNR*. Pour cela, la méthode de calcul des poids de Kaplan-Meier est revue afin de tenir compte du phénomène de troncature. On pourrait donc facilement adapter la méthode présentée dans ce mémoire pour qu'elle puisse tenir compte de problématique plus large comme la présence de sinistres tardifs.

Table des illustrations

Bibliographie

[1] BPI France, Activités réglementées
[2] Regulated professions database
[3] Les professions réglementées, article internet du Journal du Net
[4] Page Wikipédia sur les professions réglementées
[5] Salaires moyens : source hellowork
[6] Rapport Augier, 2012.
[7] Rapport Attali, 2014.
[8] Rapport Ferrand, 2014.
[9] Loi Macron pour la croissance, l'activité et l'égalité des chances économiques
[10] La Responsabilité Civile du fait personnel
[11] Législation Avocats
[12] Radier Associes, Le sursis à statuer, Septembre 2019.
[13] Article R331-6 du Code des Assurances sur les provisions techniques
[14] J.A NELDER et R.W.M. WEDDERBURN. *Generalized Linear Models*. Journal of the Royal Statistical Society, Series A, 135, 370-384, 1972.
[15] M. TENENHAUS, *GLM*.
[16] Rasoir d'Ockham, principe de parcimonie.
[17] Régression Logistique, IBM.com
[18] La régression logistique, S. NEJI et A.-H. JIGOREL, 2013.
[19] Y. LUO, Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage, 2016.
[20] L. BREIMAN, J. FRIEDMAN, R. OLSHEN et C. STONE. *Classification and Regression Trees*. Chapman and Hall, Wadsworth, New York, 1984.
[21] Periclès Group, Machine Learning: Du *GLM* à l'arbre de CART en passant par le Random Forest.
[22] R. GENUER et J.-M. POGGI*. Arbres CART et Forêts aléatoires, Importance et sélection de variables*, 2017.
[23] C. CHESNEAU (Université de Caen-Normandie). *Introduction aux arbres de décision (de type CART)*, 2020.
[24] Université de Toulouse, « Arbre binaire de décision »
[25] XG-Boost, blent.ia
[26] Y. FREUND et R. E. SCHAPIRE. *A Short Introduction to Boosting*, Journal of Japanese Society of Artificial Intelligence, Vol. 14, No. 5, pp. 771-780, 1999.
[27] Descente de gradient, datascientest.com
[28] T. CHEN et C. GUESTRIN. *XG-Boost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, 2016.

[29] G. KE, Q. MENG, T. FINLEY, T. WANG, W. CHEN, W. MA, Q. YE et T.-Y. LIU. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, 3149-3157, December 2017.
[30] M. WUTHRICH. *Machine Learning in Individual Claims Reserving*. Swiss Finance Institute Research Paper No. 16-67, (November 11, 2016).
[31] K. KUO. *DeepTriangle: A Deep Learning Approach to Loss Reserving*. *Risks* 2019, *7*, 97.
[32] O. LOPEZ, X. MILHAUD et P.-E. THEROND. *Tree-based censored regression with applications in insurance*, Electronic Journal of Statistics, Electron. J. Statist. 10(2), 2685-2716, 2016.
[33] O. LOPEZ, X. MILHAUD. *Individual reserving and nonparametric estimation of claim amounts subject to large reporting delays*. Scandinavian Actuarial Journal, pp.34 – 53, 2021.
[34] O. LOPEZ, X. MILHAUD et P.-E. THEROND. *A tree-based algorithm adapted to microlevel reserving and long development claims*. ASTIN Bulletin. 49. 1-22, 2019.
[35] E.L. KAPLAN et P. MEIER. *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association, 53, 457-481, 1958.
[36] Test de Wald, stringfixer.com
[37] Charpentier, Residuals from a logistic regression, 2013.
[38] J. PARRY. XG-Boost importance des variables, Décembre 2018
[39] Exemple prédictions XG-Boost hors limites, StackExchange.
[40] S. ALLWRIGHT. *RMSE* VS *MAE*, 2022.
[41] Mann Whitney U Test (Wilcoxon Rank Sum Test), Université de Boston
[42] D.P. HARRINGTON, T.R. FLEMING. *A class of rank test procedures for censored survival data*, *Biometrika*, Volume 69, Issue 3, Pages 553–566, December 1982.
[43] B. GHATTAS. *Importance des variables dans les méthodes CART*, *Modulad*, N°24, pp. 29-39, Décembre 1999.
[44] A. CABOS, Sensibilité et spécificité, Août 2022
[45] S. GHOSH et S. RESNICK, *A discussion on mean excess plots, Stochastic Processes and their Applications*, Volume 120, Issue 8, Pages 1492-1517, 2010.
[46] Generalized Pareto Distribution

Annexes

1. Procédure weighted CART

D'après les notations de Lopez et al.[33] (2019), on souhaite estimer $\pi(x) = \mathbb{E}[\varphi(P)|X = x]$ avec X un ensemble de représentations de variables explicatives. Pour pouvoir intégrer les poids de Kaplan-Meier, l'algorithme *CART* classique doit être modifié. Plus généralement, à chaque étape de l'algorithme des règles notées $R_j(z)$ permettent de diviser les observations pour chaque valeur possible des variables explicatives $x = \left(x^{(1)}, \dots, x^{(d)}\right)$. Ainsi on a $R_j(x) = \begin{cases} 1 \\ 0 \end{cases}$ selon que z satisfait la règle $R_j(x)$ ou non (*a fortiori* on a $R_j(x)\, R_{j'}(x) = 0$ pour tout $j \neq j'$ et $\sum_j R_j(x) = 1$.

L'algorithme *CART* pondéré des poids de Kaplan-Meier devient alors :

- **Étape 1 :** $R_1(x) = 1$ pour tout z et $n_1 = 1$

- **Étape $k + 1$:** On note $\left(R_1, \dots R_{n_k}\right)$ les règles obtenues à l'étape k. Pour $j = 1, \dots, n_k$:
  - Si toutes les observations telles que $\delta_j R_j(\mathrm{X}_i) = 1$ prennent la même valeur de x alors on conserve la règle j.
  - Sinon, deux nouvelles règles R'_{j1} et R'_{j2} sont déterminés à l'aide de x_l avec $x_l = \arg\ min_x\, m_l(R_j, x)$

    Et où

$$m_l(R_j, x) = \sum_{i=1}^{n} \omega_i \left(\phi(N_i, T_i, \mathrm{X}_i) - \bar{n}_{l-}(x, R_j)\right)^2 1_{X_i^{(l)} \leq x} R_j(\mathbb{X}) + \sum_{i=1}^{n} \omega_i \left(\phi(N_i, T_i, \mathrm{X}_i) - \bar{n}_{l-}(x, R_j)\right)^2 1_{X_i^{(l)} > x} R_j(\mathbb{X})$$

    On définit $\bar{n}_{l-}(x, R_j)$ et $\bar{n}_{l-}(x, R_j)$ comme :

$$\bar{n}_{l-}(x, R_j) = \frac{\sum_{i=1}^{n} \omega_i \phi(N_i, T_i, \mathrm{X}_i) 1_{X_i^{(l)} \leq x} R_j(\mathbb{X})}{\sum_{k=1}^{n} \omega_k 1_{X_k^{(l)} \leq x} R_j(\mathbb{X})} \text{ et } \bar{n}_{l+}(x, R_j) = \frac{\sum_{i=1}^{n} \omega_i \phi(N_i, T_i, \mathrm{X}_i) 1_{X_i^{(l)} > x} R_j(\mathbb{X})}{\sum_{k=1}^{n} \omega_k 1_{X_k^{(l)} > x} R_j(\mathbb{X})}$$

    On choisit ensuite l'indice $l = \arg\ \widehat{max_l}\, m_l(R_j, x_l)$ et on définit
    $R'_{j1}(\mathbb{X}) = R_j(\mathbb{X}) 1_{x^{(l)} \leq x_l}$ et $R'_{j2}(\mathbb{X}) = R_j(\mathbb{X}) 1_{x^{(l)} > x_l}$

  - Le nombre de règles est désormais porté à n_{k+1}

- **Dernière étape :** l'algorithme s'arrête dès lors que $n_{k+1} = n_k$

2. Optimisation du seuil

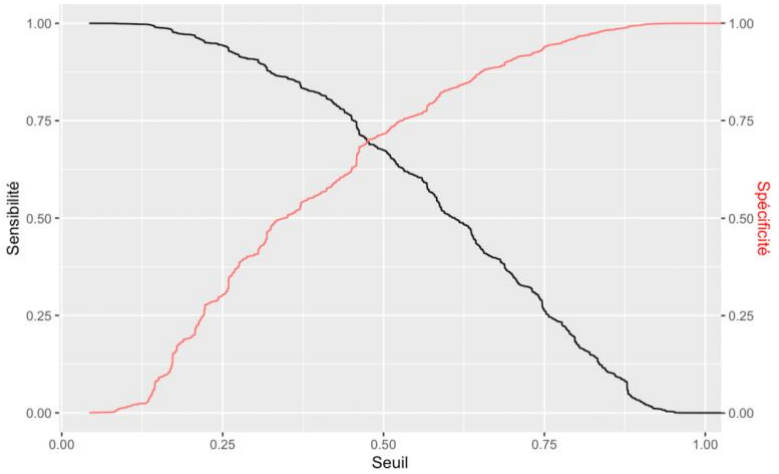
Le seuil généralement envisagé est celui de 0,5, on se propose d'étudier ce seuil et de l'améliorer en le choisissant en fonction de nos données. Il influe directement sur la précision du modèle c'est-à-dire sur le nombre d'observations qui seront effectivement classées correctement. On peut donc partir d'une première idée : trouver le seuil qui maximise cette précision. Revenons d'abord sur des notions importantes en classification de données : la sensibilité et la spécificité[44]. Pour rappel voici comment sont calculées les deux quantités :

$$Sensibilité = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Négatif} \qquad Spécificité = \frac{Vrai\ Négatif}{Vrai\ Négatif + Faux\ Positif}$$

Le contexte sanitaire actuel nous a rendus familier avec les tests, mais pour faire un rapide parallèle avec la santé, on peut résumer les deux quantités ainsi :

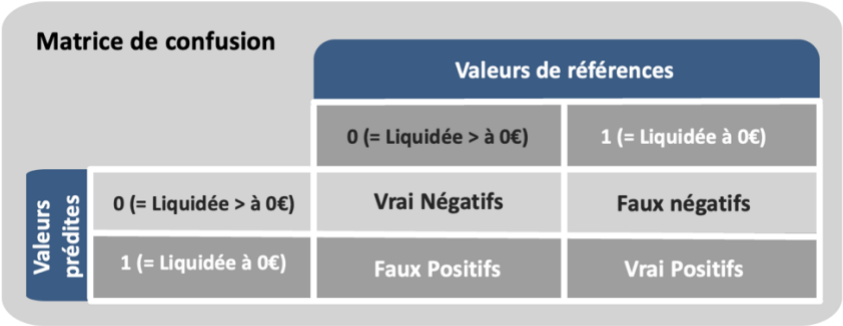
- On maximise la **spécificité** lorsque l'on effectue un test censé confirmer un diagnostic. (autrement dit lorsque l'on s'attend à un résultat positif)
- On maximise la **sensibilité** lorsque l'on effectue un test censé exclure un diagnostic. (autrement dit lorsque l'on s'attend à un résultat négatif)

On trace ainsi la spécificité et la sensibilité du modèle pour des seuils de 0 à 1, voici le résultat obtenu pour la régression logistique sur la base de test :



L'intersection des deux courbes nous indique le seuil maximisant la sensibilité, c'est-à-dire ici 0,47. Cette première méthode permet déjà d'adapter le seuil aux données, mais on peut aller encore plus loin et favoriser une quantité plus que l'autre. En effet, au regard de leur formule, on ne peut pas augmenter la spécificité sans diminuer la sensibilité : il existe une dualité entre les deux quantités. Nonobstant, cela devient intéressant lorsque l'on souhaite en favoriser une plus que l'autre. Prenons la matrice de confusion adaptée à nos données :

[44] Sensibilité et spécificité, Anthony Cabos, Août 2022



Dans notre cas, on préférera un taux plus élevé de Faux Négatifs, c'est-à-dire que l'on prédit des $PSAP_{ultime}$ comme liquidée à un montant supérieur 0€ alors qu'elles seront en réalité liquidées à 0€. C'est une prise de décision plutôt prudente : une fois arrivé à l'ultime, on devrait se retrouver avec des bonis de liquidation. On accepte donc une quantité plus importante de Faux négatifs ce qui a un impact direct sur la sensibilité : elle diminue. À cause de la dualité qui existe entre spécificité et sensibilité, il revient finalement à maximiser la spécificité ou du moins de fixer une limite sous laquelle on ne souhaite pas descendre : dans notre cas, la limite choisie est 80 %. Ce choix implique inexorablement une diminution de la précision, en effet le seuil qui la maximise est celui maximisant les deux quantités en même temps. Ainsi, le seuil obtenu en maximisant cette spécificité est environ de 0,4.

3. Détermination du seuil des sinistres graves

Définition 1. ***Excess distribution***[45]. Soit X une variable aléatoire avec F sa fonction de répartition et s un seuil. L'excess distribution au-dessus du seuil s est définit comme :

$$F_s(x) = P(X - s \leq x | X > s) = \frac{F(x+s) - F(s)}{1 - F(s)}$$

Définition 2. ***Fonction Mean Excess***[45]. Soit X un variable aléatoire et s un seuil. La fonction mean excess de la variable X s'écrit comme :

$$e(s) = E[X - s \mid X > s]$$

Le mean excess plot est un outil utilisé lorsque l'on souhaite confirmer (ou infirmer) l'ajustement d'une loi GPD, Generalised Pareto Distribution, à un échantillon.

Définition 3. ***Generalised Pareto Distribution***[46]. Soit X une variable aléatoire, la fonction de répartition de la loi GPD est donné par :

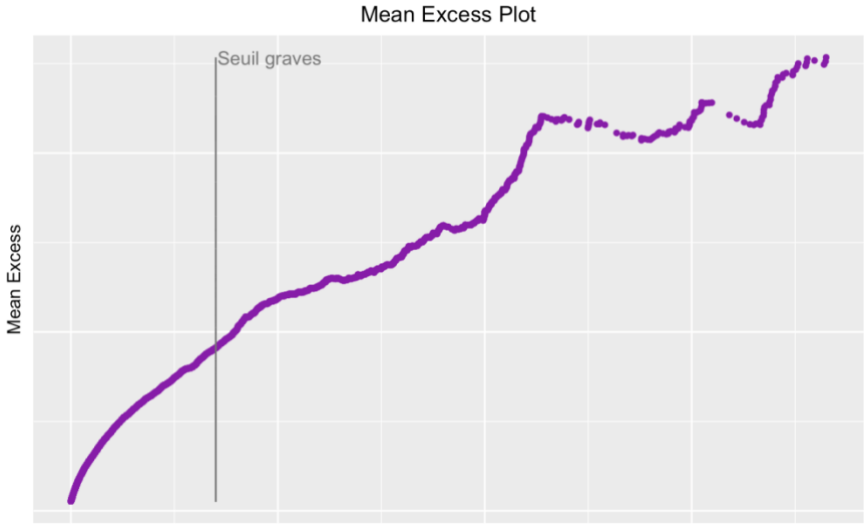
$$GPD(\gamma, \sigma, x) = \begin{cases} 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}} & si\ \gamma \neq 0 \\ 1 - e^{\left(-\frac{x}{\sigma}\right)} & si\ \gamma = 0 \end{cases}$$

D'après cette distribution GPD on déduit que la mean excess fonction s'écrit comme :

$$e(s) = E[X - s \mid X > s] = \frac{\sigma + \gamma s}{1 - \gamma}$$

Autrement dit, lorsque les valeurs extrêmes de l'échantillon peuvent être modélisées par une GPD la *mean excess* fonction est une fonction linéaire du seuil. Ainsi le *mean excess plot* obtenue doit afficher une courbe linéaire dans l'ensemble (dans le cas contraire, on s'intéressera à ajuster une autre loi : une courbe horizontale pourrait être révélateur d'un ajustement exponentiel par exemple.) Ci-dessous le *mean excess plot* obtenu pour notre échantillon de données de montant de la $PSAP_{ultime}$:

[45] GHOSH S., RESNICK S. [2010]
[46] Generalized Pareto Distribution

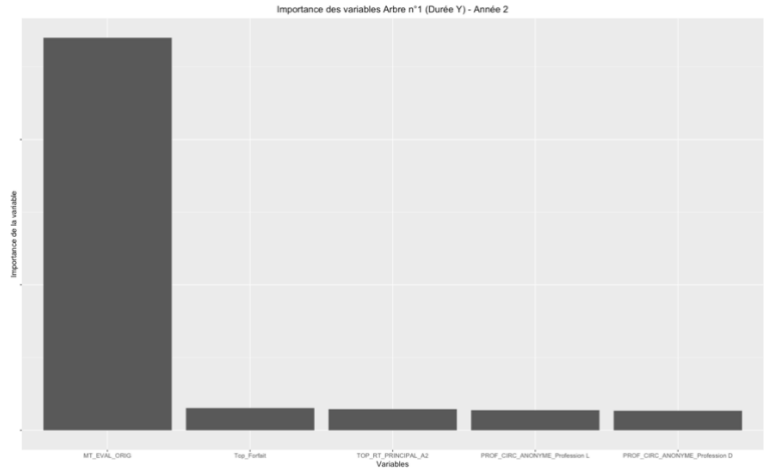


La courbe obtenue est linéaire dans l'ensemble : on pourrait ajuster une loi GPD sur cet échantillon. On détermine le seuil au moment où la courbe n'est plus une droite, en regardant en parallèle la proportion de sinistres concernés sur notre base. En effet, la proportion de sinistres graves sur la base doit rester correcte afin de ne pas considérer une partie trop importante de celle-ci comme appartenant à la sinistralité grave. Pour cela, on compare la valeur du quantile à 99 % de nos données afin de rester sur une proportion de graves autour des 1 % sur l'intégralité de la base.

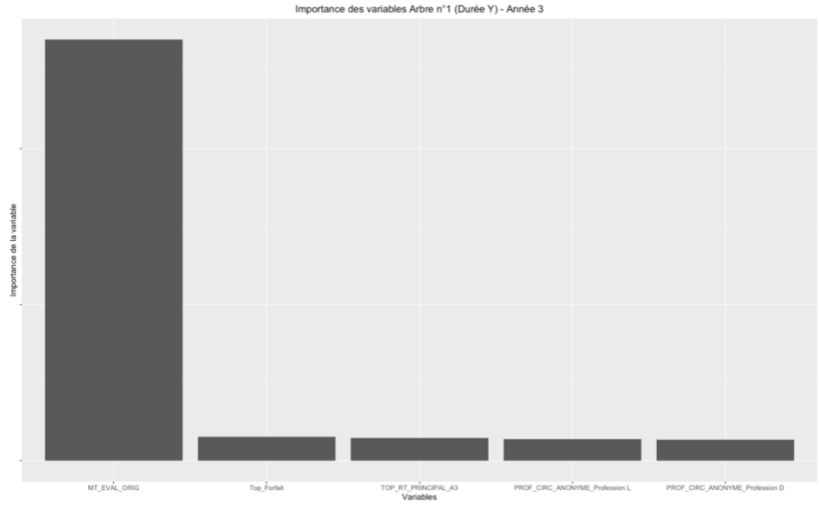
4. Importance des variables

3.1. Arbre durée de vie résiduelle

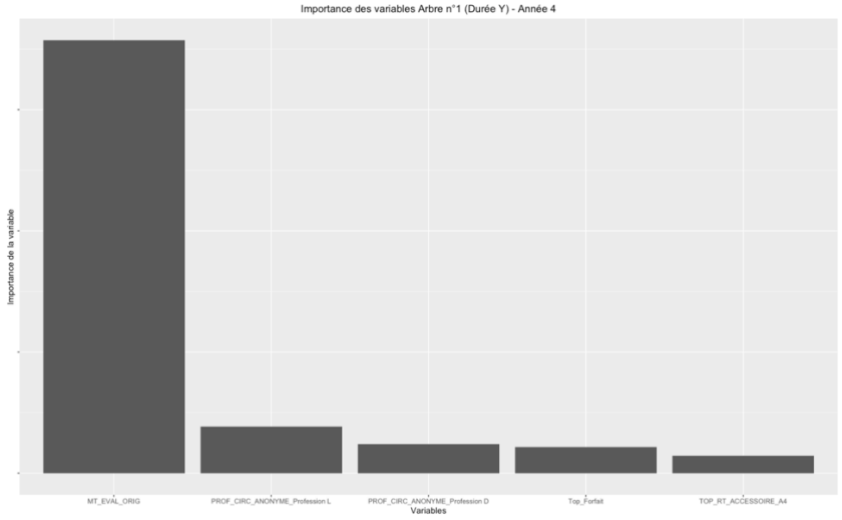
- Année n°2 :



- Année n°3 :

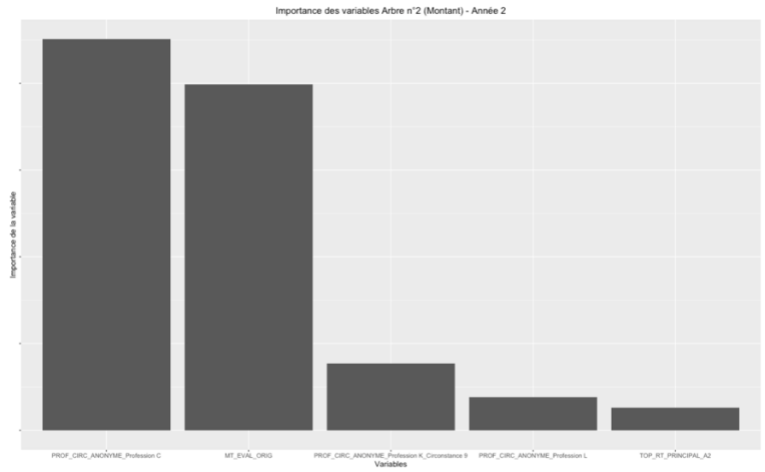


- Année n°4 :

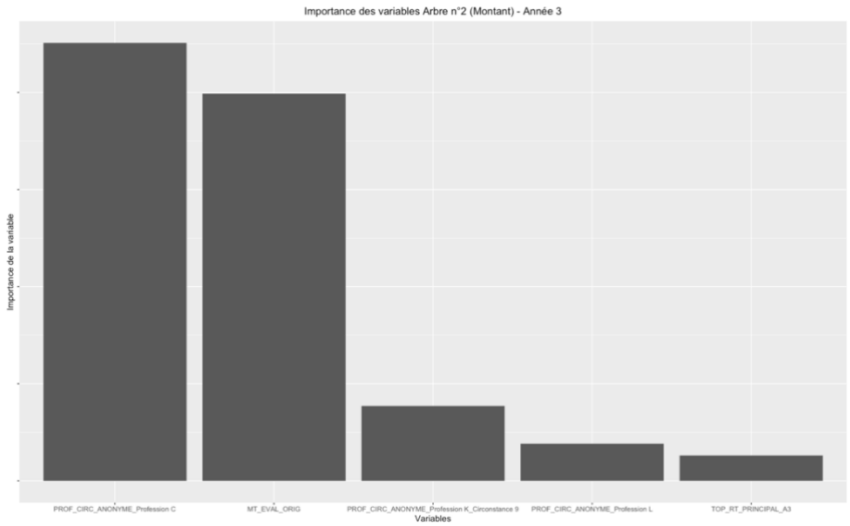


1.2. Importance des variables arbre montant

- Année n°2 :



- Année n°3 :



- Année n°4 :