IRM | institut du risk management BY Institut

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Sylvie FERON

Titre : **Utilisation des SHAP values pour une tarification du produit RC Industries et Commerces à la maille NAF**

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

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

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

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise :

Nom : AXA France IARD

AXA France I.A.R.D.
Société Anonyme au Capital de 214 799 030 €
Entreprise régie par le Code des Assurances
Siège social : 313, Terrasses de l'Arche
92727 NANTERRE CEDEX
722 057 460 RCS Nanterre

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : Gérald LUCAS

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

RESUME

L'objectif de ce mémoire est de revoir la tarification du produit Responsabilité Civile Entreprises sur le périmètre Industries et Commerces en optimisant le tarif par une segmentation plus fine des activités et une meilleure connaissance du risque.

La Responsabilité Civile Entreprises est une branche d'intensité à déroulement long. Le faible volume de sinistres, la charge importante de chaque sinistre pour laquelle la modélisation d'un coût moyen est moins pertinente que sur une branche de fréquence, ainsi que la définition de l'historique représentent une première difficulté.

Le nombre de modalités de la variable code NAF (Nomenclature d'Activités Française qui est la nomenclature des activités économiques productives, élaborée pour faciliter l'organisation de l'information économique et sociale) qui ne permet pas son utilisation dans un modèle linéaire est une seconde difficulté dont nous devons nous affranchir.

Le but de nos travaux est donc de répondre simplement à une demande complexe de la direction technique et des souscripteurs : obtenir un obtenir un taux à appliquer au chiffre d'affaires des entreprises différencié selon les 300 codes NAF, segmenté selon les variables explicatives.

Les travaux porteront sur la prédiction de la charge sinistres en exploitant les SHAP values (SHapley Additive exPlanations), dérivées des valeurs de Shapley utilisées dans la théorie des jeux, en complément d'un modèle linéaire généralisé afin de tester la faisabilité de cette méthode pour affiner et différencier nos résultats.

Mots-clefs : Responsabilité civile, IARD Entreprises, Tarification, Modèle Linéaire Généralisé, Résidus, Performance de prédictions, Valeurs de Shapley, SHAP values

ABSTRACT

The objective of this study is to review the pricing of the Corporate Liability product dedicated to industries, wholesale and retail businesses. This work is done by optimizing the tariff by better segmentation of activities and knowledge of the risk.

Corporate Liability is a long-term branch of intensity. The low volume of claims, the high amount of each claim for which the use of an average cost is less relevant than on a frequency branch, as well as the definition of the history represent a first difficulty.

The number of modalities of the NAF code variable (Nomenclature d'Activités Française which is the nomenclature of productive economic activities, developed to facilitate the organization of economic and social information) is a second difficulty because it cannot be used in a linear model.

The purpose of our work is therefore to help the technical management and subscribers to obtain a rate that can be applied to the turnover of companies. This rate should be determined for more than 300 NAF codes. Some explanatory variables of the risk will be added in addition of this rate to calculate the tariff.

The work will focus on predicting claims cost by exploiting SHAP values (SHapley Additive exPlanations). SHAP Values are derived from Shapley values used in game theory. Additionally, the feasibility of this method was tested with a linear model to refine and differentiate our results.

Keywords: Corporate Liability, Corporate P&C, Pricing, Generalized Linear Model, Residues, Prediction Performance, Shapley Values, SHAP Values

REMERCIEMENTS

Par ces quelques lignes, je tiens tout d'abord à remercier mes managers de la Direction des Systèmes d'Informations puis de la Direction Actuariat et Pilotage Entreprises ainsi que les responsables des ressources humaines de ces deux directions pour leur implication et la confiance qu'ils m'ont accordée dans la réussite de ce projet personnel et professionnel initié il y a quatre ans.

Je tiens à remercier tout particulièrement Gérald Lucas, mon directeur de mémoire pour son implication, sa disponibilité, son expertise, ses encouragements, ses nombreux conseils ainsi que le challenge des résultats, tout au long de ces travaux de mémoire.

Je tiens également à remercier tout particulièrement Véronique Marpillat, responsable de l'équipe Actuariat Produits et Data Science de m'avoir accueillie au sein de son équipe pendant ma formation d'actuaire, de m'avoir fait confiance dans le cadre de mon projet de reconversion professionnelle et d'avoir apporté un second regard sur les travaux de ce mémoire.

Je remercie Alexis Bussenaud et Franck Percot, mes collègues qui travaillent sur la Responsabilité Civile Entreprises au sein de l'équipe pour la prise en charge des sujets courants pendant mes travaux de mémoire, pour leur précieux soutien et leur aide sur les données et logiciels utilisés.

Je remercie également tous mes collègues de l'équipe Actuariat Produits et Data Science pour leur bonne humeur et leurs conseils.

Je remercie enfin mon mari et mes filles pour leurs encouragements et leur aide au cours de la formation, notamment en mathématiques et probabilités/statistiques mais aussi pendant la réalisation des travaux de mémoire.

SOMMAIRE

INTRODUCTION

PREAMBULE

En 2020, l'appareil productif français rassemblait 4,2 millions d'entreprises dans les secteurs marchands non agricoles et non financiers (source INSEE : Institut National de la Statistique et des Etudes Economiques). Ces entreprises, du fait de leurs activités, sont susceptibles de causer des dommages aux personnes, des dommages aux biens ou des dommages immatériels. L'assurance Responsabilité Civile (RC) permet aux entreprises de couvrir les dommages potentiels causés à des tiers dans le cadre de leurs activités.

L'assurance RC est réputée pour être difficile à tarifer car les risques couverts sont multiples (faute de l'employeur, accident de travail des employés, travaux réalisés par l'entreprise, biens confiés …) hétérogènes, et difficiles à cerner (atteinte accidentelle à l'environnement, risque accru de dommages immatériels consécutifs et pertes financières, rappel ou retrait de produits, cyber …).

Au vu de l'évolution des risques, il est important pour les assureurs de pouvoir appréhender au mieux les risques de leur portefeuille, afin de personnaliser leurs produits et segmenter les tarifs pour proposer à leurs clients l'offre la plus adaptée à leur risque.

L'assurance RC Entreprises a déjà fait l'objet de plusieurs études de refontes tarifaires qui aboutissent à une segmentation du tarif au niveau de grands regroupements d'activités avec moins de 10 niveaux (Industries, Commerces, Collectivités, Prestataires de services, Professions Réglementées, RC médicale par exemple).

PROBLEMATIQUE

L'objectif de ce mémoire est de revoir la tarification du produit Responsabilité Civile Entreprises sur le périmètre Industries et Commerces en optimisant le tarif par une segmentation plus fine des activités. Nous souhaitons obtenir des tarifs différenciés dans la mesure du possible pour les 300 codes activités regroupés dans les deux niveaux Industries et Commerces.

La Responsabilité Civile Entreprises est une branche d'intensité à déroulement long. Le faible volume de sinistres, la charge importante de chaque sinistre pour laquelle la modélisation d'un coût moyen est moins pertinente que sur une branche de fréquence, ainsi que la définition de l'historique représentent une première difficulté.

Le nombre de modalités de la variable code NAF (Nomenclature d'Activités Française qui est la nomenclature des activités économiques productives, élaborée pour faciliter l'organisation de l'information économique et sociale) qui ne permet pas son utilisation dans un modèle linéaire est une seconde difficulté dont nous devons nous affranchir.

Le but de nos travaux est donc de répondre simplement à une demande complexe de la direction technique et des souscripteurs : obtenir un obtenir un taux à appliquer au chiffre d'affaires des entreprises différencié selon les 300 codes NAF inclus dans le produit mais aussi segmenté selon les variables explicatives. Nous devons également pouvoir expliquer facilement à nos directions métier la méthode utilisée et les résultats.

INTERET DE L'ETUDE

Les travaux porteront sur la prédiction de la charge de sinistres pour chacun des 300 codes NAF constitutifs du produit RC Industries et Commerces en modélisant cette charge au niveau macro à l'aide d'un Modèle Linéaire Généralisé (GLM) puis en exploitant les résidus de ce GLM grâce à la méthode des SHAP (SHapley Additive exPlanations) values.

Les SHAP values sont dérivées des valeurs de Shapley utilisées dans la théorie des jeux. En théorie des jeux, les valeurs de Shapley indiquent comment répartir équitablement le gain entre plusieurs joueurs travaillant en équipe, en garantissant à chaque joueur de gagner autant ou plus que s'il agissait indépendamment.

En 2017, Scott Lundberg a appliqué les valeurs de Shapley aux modèles de Machine Learning de type « boîte noire » au travers de l'algorithme SHAP. Les SHAP values sont depuis utilisées pour l'interprétation de modèles de Machine Learning complexes.
Appliquées aux modèles de Machine Learning, elles sont conçues pour attribuer la différence entre la prédiction d'un modèle et une base de référence moyenne aux différentes caractéristiques utilisées en entrée du modèle.
L'idée est d'extraire des informations du modèle afin de comprendre sur quoi se base le modèle pour effectuer les prédictions, soit globalement (contribution des variables et des modalités au modèle), soit localement pour expliquer une prédiction particulière.

Dans notre cas, l'objectif n'est pas d'utiliser la méthode des SHAP values pour expliquer les prédictions d'un modèle de Machine Learning comme c'est habituellement le cas au travers de l'algorithme SHAP, mais en complément d'un modèle linéaire généralisé afin de tester la faisabilité de cette méthode pour affiner, différencier nos résultats et obtenir un coefficient spécifique à appliquer à chaque activité du niveau le plus fin du référentiel INSEE (code NAF). Le nombre de modalités de la variable code NAF ne permet en effet pas d'utiliser cette variable au niveau du GLM qui ne converge plus lors de l'introduction de cette variable.

Pour atteindre cet objectif, la première partie du mémoire sera consacrée à la définition de la responsabilité civile des entreprises et des garanties incluses dans le produit proposé aux entreprises industrielles et aux commerces. Nous verrons les particularités de cette branche d'intensité à déroulement long qui ne facilitent pas la maîtrise du risque.

La seconde partie présentera notre périmètre d'étude et se concentrera sur la gestion des données et l'analyse de la sinistralité.

La troisième partie décrira la manière dont nous avons modélisé le risque en présentant le modèle utilisé, la sélection des variables discriminantes et la comparaison des modèles. Nous exposerons également les résultats de la modélisation de notre charge de sinistres.

La dernière partie sera consacrée aux SHAP values et à l'application opérationnelle de cette méthode pour affiner notre charge de sinistralité en complément du GLM. Cette dernière partie présentera les SHAP values dans la théorie des jeux puis son utilisation dans notre modèle de prédiction pour répondre à notre problématique. Nous effectuerons ensuite des prédictions sur une base de test afin d'estimer l'apport de cette méthode pour segmenter notre tarif par activité.

NB : Pour des raisons de confidentialité, les résultats chiffrés présentés dans ce mémoire ont été modifiés tout en conservant les ordres de grandeur pour maintenir le sens des conclusions.

Chapitre 1

CADRE DE L'ETUDE ET PRESENTATION DU PRODUIT

1. LA RESPONSABILITE CIVILE ENTREPRISES

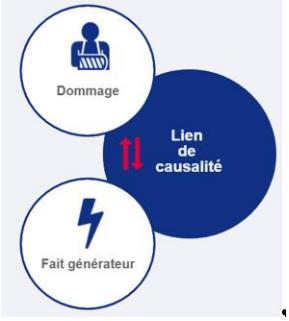
1.1. RAPPEL DES GRANDS PRINCIPES DE LA RC

Le principe général de la Responsabilité Civile (noté RC par la suite) est décrit dans le Code civil. La RC est une obligation légale qui impose à toute personne physique ou morale de réparer les dommages causés à une victime de son fait, de celui des personnes dont elle doit répondre ou des choses dont elle a la charge**.**

Pour couvrir les dommages potentiels, la RC peut faire l'objet d'une assurance. L'assurance RC peut couvrir des particuliers (RC vie privée, scolaire, assurance chasse), des professionnels (commerçants, artisans, professions libérales …) ou des entreprises (industries, prestataires de services, collectivités publiques, barreaux d'avocats, établissements de soins …).

La RC peut être engagée lorsque trois conditions sont réunies :

- la survenance d'un **dommage**
- l'existence d'un **fait générateur**
- et un **lien de causalité** entre ce fait générateur et ce dommage.



Quel que soit le régime de responsabilité, la relation de cause à effet entre le fait dommageable et le dommage doit être établie. Une simple coïncidence ne suffit pas.
C'est à la victime de prouver le dommage.

Pour être réparable, le dommage doit être :

- direct : c'est-à-dire résulter directement du fait reproché au responsable.
- certain : Est certain un préjudice déjà subi (manque à gagner, invalidité …) ou la perte d'une chance si elle était raisonnable. Sont également certaines les conséquences futures si elles sont inévitables (versement d'une rente pour invalidité suite à un accident).
- déterminé : Le dommage peut être matériel, corporel, moral ou économique.
- et consister en la lésion d'un intérêt licite : Il y a exonération en cas de force majeure, fait ou faute de la victime ou fait d'un tiers.

Le fait générateur est le plus souvent un fait personnel, fautif ou parfois non fautif, de l'auteur du dommage. La faute peut être intentionnelle (mauvaise foi, intention de nuire) ou non intentionnelle (négligence ou imprudence).

La RC résulte :

- de la non-exécution d'un contrat ou d'une obligation (**RC contractuelle**)
- d'un dommage causé en dehors de tout contrat (**RC quasi-délictuelle ou délictuelle**) – voir annexe 1
- d'un texte légal mettant à la charge de l'intervenant une responsabilité spécifique (décennale, agent immobilier, hôtelier, sécurité des produits …).

La RC est dite contractuelle lorsque le dommage dont se plaint la victime résulte de la mauvaise exécution ou de l'inexécution d'un contrat ou d'une convention. Ce contrat n'est pas nécessairement écrit, il peut être oral ou tacite. Par définition, la victime et le responsable « se connaissent » s'ils sont co-contractants.

Dans le cas de la RC contractuelle, la cause génératrice du dommage peut être :

- la défaillance du débiteur,
- l'exécution défectueuse,
- la perte de l'objet de l'obligation.

Dans le cas de la RC extra-contractuelle, la faute peut résulter d'une erreur, d'une négligence ou d'une omission.

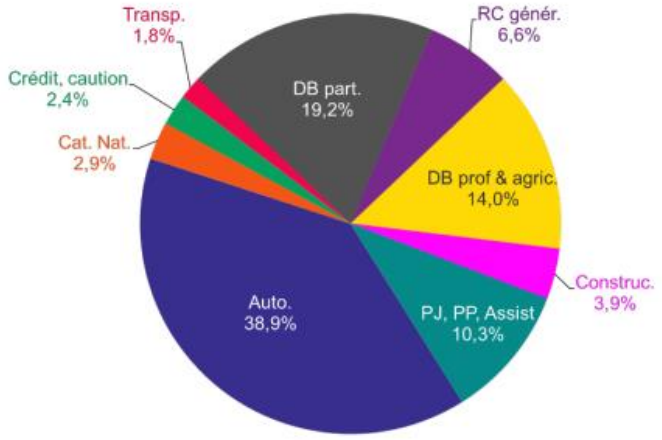
Il existe deux formes de réparation possibles :

- **La réparation en nature**
  - restitution, réparation, remplacement du bien
  - prestation de services pour améliorer la qualité de vie de la victime ayant subi un préjudice corporel
- **Les dommages et intérêts**
  - compensation financière des préjudices subis par la victime
  - sous forme de capital ou de rente.

L'action en responsabilité civile est soumise à la prescription extinctive de droit commun (article 2224 du code civil) : le délai est de 5 ans à partir du moment où le demandeur a eu connaissance des faits permettant d'agir. Par dérogation, ce délai est de dix ans après consolidation du dommage pour les dommages corporels (article 2226 du code civil).

Bien qu'incluse dans de nombreux produits, notamment en assurance des particuliers, l'assurance RC est facultative et elle représente moins de 10% de l'assurance de biens et responsabilité.

Assurance RC générale en chiffres :
En 2019, le montant des cotisations perçues au titre des contrats spécifiques de responsabilité civile générale s'élève à 3,8 milliards d'euros, représentant 6,6 % de l'ensemble des encaissements des assurances de biens et de responsabilité.



Place de l'assurance de responsabilité civile générale dans l'ensemble des cotisations dommages aux biens et responsabilité

Source : L'assurance Responsabilité Civile Générale en 2019 publié en février 2021 par la FFA

1.2. LA BRANCHE RC ENTREPRISES

Dans un contexte de judiciarisation de la société (obligation de sécurité des produits, obligation de résultat, de conseils) et procédurier (les mises en cause sont de plus en plus fréquentes et élevées en montant d'indemnisation), la RC des professionnels est un risque en forte croissance (+4.7% entre 2018 et 2019).

Chiffres clés

RCG des professionnels	2018 (r)	2019 (p)	Evolution 2019/2018
Données comptables			
Montant des cotisations	3 671 M€	3 842 M€	+4,7 %

Source : L'assurance Responsabilité Civile Générale en 2019 publié en février 2021 par la FFA

Par ailleurs, le risque de responsabilité civile est un risque sujet à de nombreuses mutations qu'elles soient économiques, technologiques et sociales ou en lien direct avec l'évolution même du droit. L'apparition de nouveaux risques inhérents à de nouveaux métiers et à de nouvelles technologies (cyber, biotechnologies, insectes dans l'alimentation …) et la croissance de certains risques (amiante, champs électromagnétiques, OGM, hydrogène …) nécessitent un suivi particulier.

Les chefs d'entreprises sont donc de plus en plus préoccupés par leur RC Entreprises. S'assurer en RC est essentiel pour une entreprise car cela permet de se protéger contre les mises en cause par des fournisseurs, clients, sous-traitants ou autres tiers lorsque ceux-ci subissent un dommage. Celui-ci peut avoir lieu dans de nombreuses circonstances, qu'il soit du fait de l'entreprise, de son responsable ou de ses salariés.

Une entreprise du fait de son activité est susceptible de causer :

- **des dommages corporels** (dommages **aux personnes**, c'est-à-dire toute atteinte à l'intégrité physique, mentale, ou psychologique d'une personne physique). Exemple : En faisant ses courses, un client se casse le bras en glissant sur une feuille de salade dans un supermarché assuré.
- **des dommages matériels** (dommages **aux biens** appartenant à autrui et qui portent atteinte au patrimoine de la victime : perte, destruction ou disparition d'un bien mobilier ou immobilier)
- ou **des dommages immatériels** (tout dommage subi autre que corporel ou matériel - exemple : privation de jouissance d'un droit, perte financière, interruption d'un service rendu par une personne ou un bien). Au sens du contrat d'assurance RC, les dommages

immatériels peuvent être consécutifs ou non consécutifs à un dommage matériel. Le dommage immatériel est consécutif (DIC) lorsqu'il est la conséquence d'un dommage corporel ou matériel garanti. Par exemple, un retard dans la livraison par une agence d'un catalogue engendre chez le client une perte de clientèle. Le dommage immatériel est non consécutif (DINC) lorsqu'il n'est pas la conséquence d'un dommage corporel ou matériel ou qu'il est la conséquence d'un dommage corporel ou matériel non garanti. Par exemple, à la suite d'une menace d'explosion dans l'entreprise que nous assurons, un commerce voisin doit être évacué.

Toute entreprise a des responsabilités envers :

- **ses préposés**
- **ses co-contractants**
- **des tiers** en général.

Préposés :
L'entreprise est responsable des dommages envers ses employés lorsqu'ils se blessent sur leur lieu de travail. La responsabilité de l'employeur est présumée en cas d'accident du travail ou de maladie professionnelle atteignant l'un de ses préposés. L'employeur est tenu à une obligation de résultats vis-à-vis de ses employés concernant la sécurité.

Tout manquement à cette obligation revêt le caractère de faute inexcusable dès lors que l'employeur avait ou aurait pu avoir conscience du danger, et qu'il n'a pas pris les mesures nécessaires. Exemple de faute inexcusable : un préposé pénètre dans un tunnel d'un convoyeur à bande pour relever le numéro d'identification du moteur électrique du broyeur. Le moteur, qui n'est pas équipé d'un carter de protection, se met brutalement en marche et lui déchire le bras.
Pour autant, la preuve du manquement incombe toujours au salarié.

Co-contractants :
Il existe différents types de co-contractants :

- les donneurs d'ordre pour les marchés de travaux, de fabrication, de sous-traitance …
- les sous-traitants (entreprises, bureaux d'études ….)
- les fournisseurs pour les achats de machines, outillages, matières premières …
- les clients lors de l'exécution des travaux, la commercialisation de produits …

En RC, la garantie peut être mise en œuvre de deux façons :

- en base **fait dommageable** survenu pendant la période de validité du contrat, base obligatoire pour l'assurance RC des non-professionnels
- en base **réclamation** faite pendant la période de validité du contrat.

Pour les risques professionnels, l'assureur a le choix entre ces deux bases. AXA, comme la plupart des assureurs du marché a choisi la base réclamation. La base réclamation a pour conséquences :

- la reprise du passé inconnu : réclamation imputable à des faits antérieurs à la souscription à condition que ces faits soient ignorés lors de la souscription.
- la prise en charge de toute réclamation formulée pendant 5 ans à compter de l'expiration de la garantie.

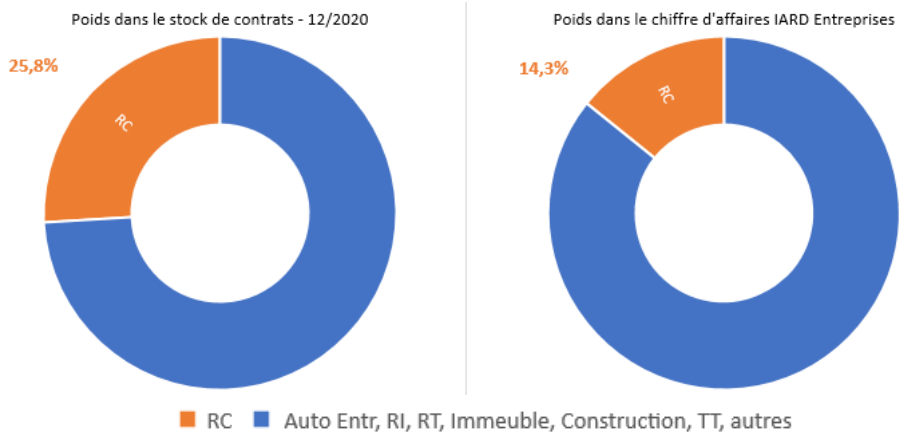
Frontières de la garantie RC :
La RC est à la frontière d'autres garanties. Ainsi, la RC Entreprises AXA ne couvre pas personnellement le dirigeant contre des poursuites pénales. Une assurance responsabilité des Dirigeants (RDD) spécifique doit être souscrite pour prendre en charge les frais de défense et les conséquences pécuniaires de l'assuré mis en cause au titre de ses fonctions (exemple : le dirigeant prend un engagement dépassant les pouvoirs conférés par les statuts). De même, la RC Entreprises AXA ne couvre pas la responsabilité de l'employeur (RDE). Une assurance RDE spécifique doit être souscrite pour prendre en charge les frais de défense et les conséquences pécuniaires de l'assuré mis en cause pour des actes de discrimination ou de harcèlement (par exemple lorsqu'un candidat envoie une demande amiable pour non-respect d'une promesse d'embauche, lorsqu'un salarié porte plainte pour une évaluation professionnelle erronée …).

La RC doit être distinguée de la responsabilité pénale. La responsabilité pénale est l'obligation de répondre de ses actes délictueux en subissant une sanction pénale dans les conditions et selon les formes prescrites par la loi. En responsabilité pénale, le dommage est lié à une infraction au regard des dispositions du code pénal. Les infractions sont classées selon la gravité de la faute : contravention, délit, crime et la sanction est une amende ou un emprisonnement.
Parfois le juge pénal est saisi d'une plainte pour infraction et d'une demande de réparation civile (dommages et intérêts). Seule cette dernière demande pourra être prise en charge au titre du contrat d'assurance RC. La responsabilité pénale et les sanctions pénales ne sont pas assurables.

Pilotée au sein de la direction IARD Entreprises, la branche RC Entreprises a un poids significatif dans le portefeuille AXA France IARD Entreprises. Elle représente :

- 25,8% du poids dans le stock des contrats fin 2020
- 14,3% du chiffre d'affaires.

Répartition de la branche RC au sein du portefeuille AXA France IARD Entreprises

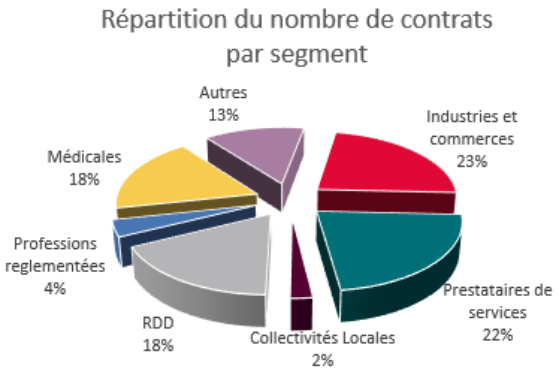


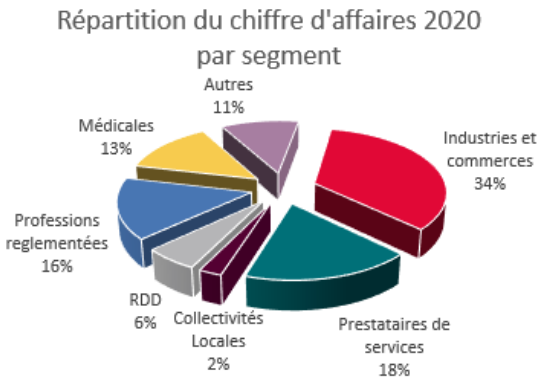
1.3. LA SEGMENTATION DE LA BRANCHE RC ENTREPRISES

Les risques couverts par la branche RC Entreprises sont naturellement très hétérogènes et les entreprises souscrivent à des contrats d'assurance RC selon leur domaine d'activité.
Afin de prendre en compte l'hétérogénéité des risques assurés et en fonction de la réglementation susceptible d'encadrer l'activité (RC Médicale, Professions réglementées …), la branche RC Entreprises AXA, qui représente plus de 400 millions de primes, est organisée autour de plusieurs segments d'activité (Industries et commerces, Prestataires de services, Collectivités locales, Professions réglementées, RC médicale …).

La segmentation des entreprises est réalisée à partir de la nomenclature d'activité professionnelle de l'entreprise à l'aide du code NAF (ou APE) délivré par l'INSEE.

Notre étude porte sur le segment Industries et Commerces (en rouge sur les graphiques) qui représente 23% du nombre de contrats et 34% du chiffre d'affaires de la branche RC Entreprises en 2020.





Exemples d'activités incluses dans le périmètre des entreprises industrielles et commerciales :

- Activités de fabrication (textile, agroalimentaire, chimique …)
- Industries manufacturières
- Activités de sidérurgie, de fonderie et de métallurgie
- Commerces de gros et de détail.

2. LE PRODUIT RC INDUSTRIES ET COMMERCES

AXA a mis en place une solution dédiée aux entreprises industrielles et commerciales, qui couvre l'ensemble des risques de responsabilité civile, découlant de l'exercice de l'activité. Le contrat RC Entreprises garantit les conséquences pécuniaires de la responsabilité civile incombant à l'assuré en raison des dommages corporels, matériels et immatériels causés au tiers :

- **avant livraison** d'un produit où l'achèvement d'une prestation ou de travaux
- **après livraison** d'un produit où l'achèvement d'une prestation ou de travaux.

RC avant livraison des produits ou réception des travaux (ou RC exploitation) :
Le contrat s'exerce du fait :

- Des biens que l'assuré exploite
- Des moyens humains et matériels qu'il met en œuvre
- D'une prestation ou de travaux.

Illustration : Un salarié renverse un pot de peinture sur un client. Celui-ci, mécontent, réclame le remboursement de son costume.

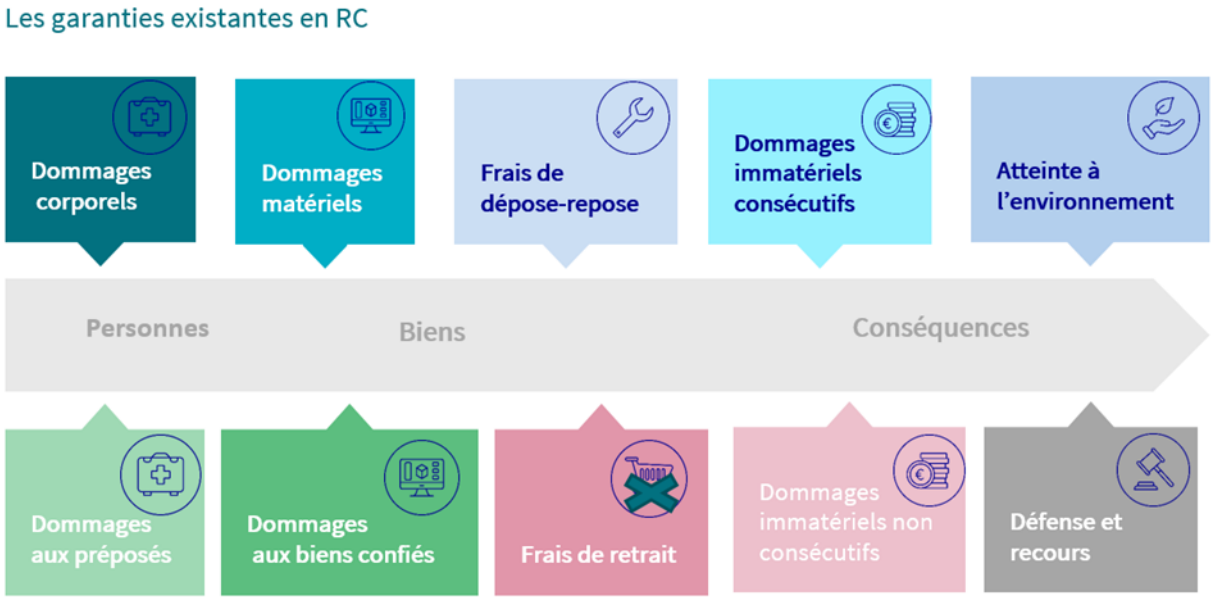
RC après livraison des produits ou réception des travaux :
Le contrat s'exerce en raison des dommages ayant pour origine :

- Une erreur de conception

- Un vice caché de fabrication, de montage, de matière
- Un défaut de sécurité
- Une erreur dans l'exécution de prestations
- Une erreur dans la rédaction des instructions et préconisations d'emploi, des documents techniques et d'entretien de ces produits, matériaux ou travaux
- Un conditionnement défectueux
- Une malfaçon des travaux effectués
- Un défaut de conseil lors de la vente.

Illustration : Le fabricant assuré vend à un client une trieuse de courrier, qui après livraison déchiquette le papier. En réparation du préjudice subi, le client réclame la réfection des courriers et les frais de main-d'œuvre supplémentaire pour trier manuellement le courrier.

Garanties Existantes en RC Entreprises :
Le contrat RC Entreprises AXA propose l'ensemble des garanties suivantes.



Certaines garanties sont incluses de base dans le contrat RC Entreprises, d'autres sont optionnelles.

2.1. LES GARANTIES DE BASE

Les garanties de base incluses dans le produit RC Entreprises sont les suivantes :

- La RC avant livraison (ou RC exploitation) et la RC après livraison
- Les dommages subis par les préposés et les clients
- Les dommages immatériels consécutifs et non consécutifs
- Les dommages subis par des biens confiés à l'entreprise
- Les frais de dépose/repose et frais de retrait exposés par les tiers
- Les risques environnementaux.

Dommages subis par les préposés et les clients

Illustrations :	
Dommage subi par un préposé	Un salarié est blessé à la tête lors du montage d'un équipement sur son lieu de travail. Son responsable avait omis de lui préciser l'obligation de mettre un casque.
Dommage subi par un client	Lors de la visite d'une entreprise, un client glisse sur le sol mal nettoyé et se blesse. Il réclame le remboursement de son préjudice corporel.

Dommages immatériels consécutifs et non consécutifs

On appelle dommage immatériel tout dommage autre que corporel ou matériel et notamment tout préjudice qui résulte :

- De la privation de jouissance d'un droit
- De la perte d'un bénéfice
- De l'interruption d'un service rendu par une personne ou un bien.

Ils visent le préjudice financier consécutif au dommage matériel, dans certains cas à l'immatériel (exemple : perte de chiffre d'affaires suite à un accident de l'homme-clé) mais également le préjudice financier en dehors de tout dommage.

Illustrations :	
Dommage immatériel consécutif à un dommage matériel ou corporel	Lors de l'installation d'une machine chez un client, le préposé de la société assurée provoque un court-circuit dans l'installation électrique, ce qui entraîne pour le client une perte d'exploitation.
Dommage immatériel non consécutif	1er cas : Dommage immatériel consécutif à un dommage corporel ou matériel non garanti au titre du contrat de RC : A la suite d'une explosion dans l'entreprise assurée (dommage non couvert par le contrat RC mais par le contrat risque industriel de l'entreprise), les commerces voisins doivent être évacués. Cette explosion a aussi pour conséquence un retard dans la livraison que l'assuré devait effectuer chez un client. 2ème cas : Dommage immatériel constaté en l'absence de tout dommage matériel et/ou corporel (communément appelé « Dommage immatériel pur ») : A la suite d'une menace d'explosion dans l'entreprise assurée, un commerce voisin doit être évacué.

Dommages subis par des biens confiés à l'entreprise

On entend par bien confié tout bien meuble appartenant à un tiers y compris aux clients de l'assuré dont ce dernier a le dépôt, la garde ou la détention à un titre quelconque.

Il peut s'agir de :

- Biens appartenant à un tiers et confiés à l'assuré en vue d'un travail. Par exemple : des textiles sont confiés pour la teinture, des lames de robots ménagers sont confiées à une entreprise d'affûtage et de polissage, des clichés de valeur sont confiés à un imprimeur pour la confection d'un catalogue

- Biens appartenant à un tiers et utilisés par l'assuré dans le cadre de ses activités, à condition que ces biens lui soient prêtés à titre gratuit. Par exemple : une société fabrique des flacons en utilisant des moules qui lui sont confiés, appartenant au propriétaire de la marque de parfums.

Illustration : Un client confie à un réparateur un appareil pour lequel un boulon doit être changé. Lors de l'intervention, le réparateur endommage le circuit électrique de l'appareil.

Les risques environnementaux
Les risques environnementaux incluent :

- La responsabilité environnementale
- La RC atteinte à l'environnement accidentelle
- Le préjudice écologique. Depuis la promulgation de la loi sur la biodiversité d'août 2016, il est maintenant admis que « toute personne responsable d'un préjudice écologique est tenue de le réparer » - Art. 1246.

Les dommages aux biens de l'assuré lui-même ne sont pas couverts par le contrat RC Entreprises (mais par un contrat complémentaire GREEN). Lorsqu'une entreprise est classée SEVESO, elle doit souscrire une assurance spécifique pour ses risques d'atteintes à l'environnement.

Illustrations :	
Responsabilité environnementale	Un stock d'huiles usagées mal conditionné situé sur le terrain d'un garagiste se déverse dans une rivière. La garantie responsabilité environnementale, mise en œuvre après injonction du préfet, permet de couvrir les 1ers frais destinés à la réparation des dommages environnementaux.
RC atteinte à l'environnement accidentelle	Lors de l'intervention sur un site client, une entreprise de maintenance renverse un produit toxique utilisé dans le cadre de son activité et pollue l'eau d'un réservoir. Le sol de l'entreprise doit être décapé, le réservoir vidé et nettoyé. Le client réclame le coût du nettoyage.
Préjudice écologique	En mettant en place un drainage de captage d'eau pour son exploitation, le client assèche accidentellement une zone humide de reproduction d'une espèce protégée. Des associations du département et le conseil général portent réclamation et exigent la restauration de zones propices au développement de cette espèce.

2.2. LES GARANTIES OPTIONNELLES

Les garanties optionnelles du produit RC Entreprises sont les suivantes :

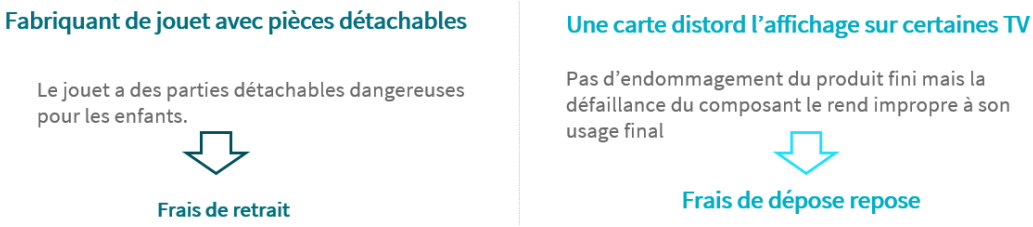
- Les frais de dépose/repose engagés par l'assuré
- Les frais de retrait engagés par l'assuré
- Les exportations directes aux USA/Canada
- La protection juridique

- Les frais de prévention. Cette garantie couvre des frais engagés par l'assuré pour éviter ou limiter un sinistre qui serait couvert par le contrat, avec l'accord préalable de l'assureur.

Obligation de sécurité
Deux directives européennes intégrées dans le droit français, imposent aux producteurs et à leurs intermédiaires dans la chaîne de fabrication d'un produit :

- Une obligation de sécurité : prendre toutes les mesures et dispositions nécessaires pour empêcher l'apparition d'un risque pouvant affecter la vie ou la santé des tiers
- Une obligation de suivi :
  - Observer les produits mis sur le marché pendant toute la durée prévisible de leur utilisation
  - Prendre dans les plus brefs délais toutes les mesures qui s'imposent en cas de danger. Cela peut aller jusqu'au rappel ou au retrait des produits.

En fonction de sa nature, un produit vicié peut soit faire l'objet d'une réparation, soit d'un retrait pur et simple.



En présence d'un produit qui peut faire l'objet d'une réparation (par exemple une machine-outil), sa réparation peut nécessiter un rapatriement dans l'atelier du fabricant assuré. On parlera de frais de dépose/repose. Les frais de dépose/repose exposés par les tiers sont pris en charge par la garantie de base du contrat au titre des dommages immatériels.

Dans certains cas, l'assuré peut procéder à ses frais à la dépose/repose du produit livré défectueux : il s'agit d'une garantie optionnelle.

Illustration : Les clients de notre assuré sont victimes d'une intoxication alimentaire. Son entreprise est reconnue responsable et l'article parait dans les médias. Une autorité administrative exige de retirer les produits susceptibles de présenter un risque pour la santé des consommateurs. Les frais engagés pour organiser le retrait rapide des produits font l'objet d'une garantie optionnelle. Les frais liés à l'indemnisation des victimes ainsi que les frais de mise en place du plan de communication font parties des garanties de base du contrat RC Entreprises.

Option frais de dépose/repose exposés par l'assuré :
Le contrat RC Entreprises prévoit une extension de garantie avec la prise en charge des frais de dépose/repose.
Sont garantis les frais de dépose repose engagés par l'assuré pour les produits livrés par ses soins pour autant que sa responsabilité soit recherchée du fait :

- d'un vice caché où défaut non apparent des produits fournis
- d'un défaut de sécurité des produits fournis
- d'une erreur commise dans les instructions d'emploi de ces produits
- d'une erreur commise dans l'exécution des prestations

dans la mesure où ce vice caché, ce défaut où cette erreur s'est révélé après livraison.

Les frais de dépose/repose comprennent les dépenses de main d'œuvre et de transport ainsi que les dépenses en matériel et moyens nécessitées par les opérations de remplacement du produit.

Illustration : Notre assuré fabrique des canalisations et livre ses produits à des installateurs professionnels. Peu après des travaux de pose dans une entreprise de jouets, cette dernière constate que l'eau s'évacue mal et provoque le début d'une inondation. L'installateur professionnel présente une réclamation auprès de notre assuré. Ce dernier teste les coudes en stock dans son atelier et constate leur défectuosité. Les coudes doivent être remplacés. Par souci d'efficacité, notre assuré prend l'initiative de déposer les coudes installés dans l'entreprise de jouets et de les remplacer. Les frais de dépose repose sont pris en charge par l'assureur (hors coût du produit).

Option frais de retrait exposés par l'assuré :
En présence d'un produit qui ne peut pas faire l'objet d'une réparation (par exemple : jouets, produits alimentaires), la RC Entreprises garantit les frais de retrait du marché. Ce sont les frais engagés pour procéder :

- à une mise en garde du public ou des détenteurs de biens
- au retrait du marché des produits mis en circulation par l'assuré, en vue de les repérer, de les isoler de les rappeler et éventuellement de les détruire. Ces opérations sont entreprises en cas de menace ou de survenance de dommages corporels ou dommages matériels garantis.

La garantie frais de retrait s'applique lorsque les opérations de retrait sont entreprises :

- pour répondre à une injonction d'une autorité publique compétente
- ou en raison d'un vice ou d'un défaut de sécurité du produit livré ou d'une faute commise par l'assuré ou une personne dont il est responsable.

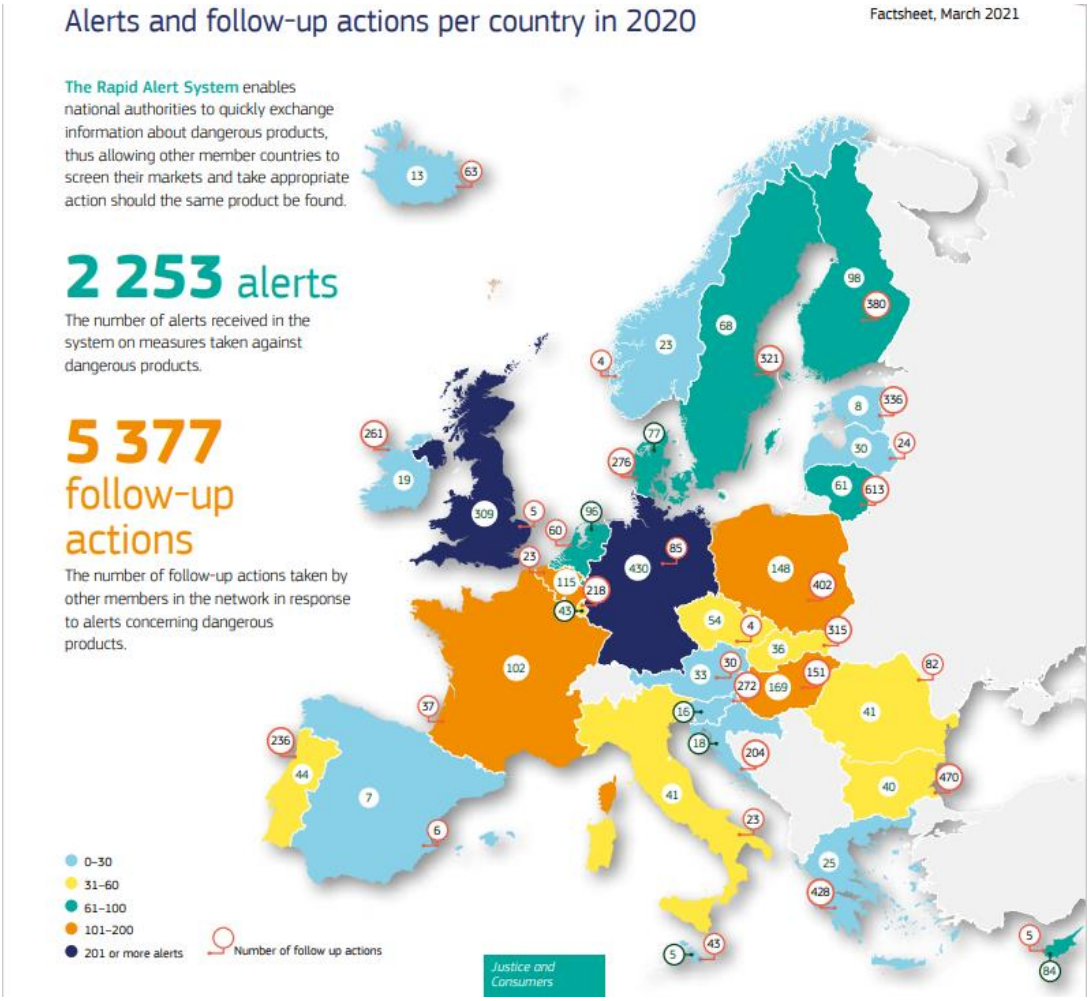
Illustration : Notre assuré importe des ours en peluche. Peu après leur mise sur le marché, les réclamations affluent. Les yeux des ours ont été mal cousus et risquent de provoquer l'étouffement des enfants destinataires de ces jouets. L'entreprise doit intervenir pour retirer tous les ours du marché. L'assureur prend en charge les frais engagés par l'assuré pour le retrait du produit (hors coût du produit).

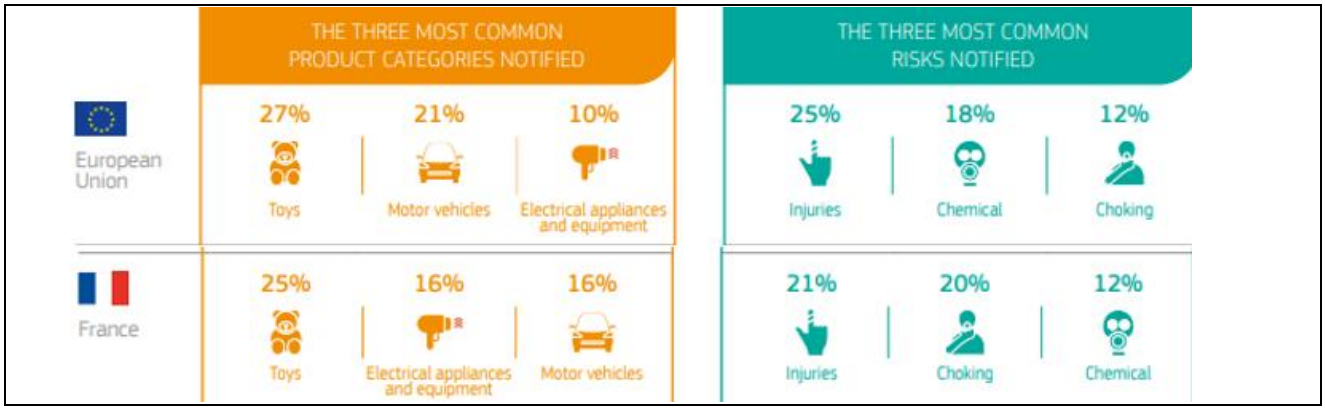
Les frais de retrait engagés par les tiers sont pris en charge par la garantie de base du contrat au titre des dommages immatériels.

Les garanties frais de dépose/repose et frais de retrait sont optionnelles et non automatiques. La garantie frais de retrait n'est accordée que si l'entreprise met en place des mesures de qualité et de prévention. Enfin, pour certaines activités, la garantie frais de retrait peut être renforcée et prévoir la couverture des produits contaminés (Garantie contamination des produits en agroalimentaire).
L'analyse du risque sur les frais de retrait porte sur l'existence ou non d'un système d'assurance qualité, y compris de contrôle des produits, la durée de vie du produit, la traçabilité du produit tout au long de la chaîne, le mode de distribution du produit, la destination du produit et sa sensibilité au risque de dommage corporel, la diffusion géographique, la quantité d'unités mises sur le marché par séries ou par lots ainsi que la nature des produits.

Les rappels concernent tous les produits : jouets, cosmétiques, produits électroménagers, automobiles, produits agro-alimentaires … et tous les acteurs dans la chaîne de commercialisation d'un produit : importateur, fabricant, distributeur, fournisseur, mandataire …

Les rappels de produits sont enregistrés dans la base européenne RAPEX. Au cours de l'année 2020, il y a eu 102 alertes RAPEX enregistrées en France. La Commission européenne publie chaque semaine sur Internet un bilan des produits dangereux et des mesures correctives prises.





Source : European Commission - safety_gate_statisticsAndAnualReports_2020_RAPEX_2020

Exemple : début septembre 2020, la France a été concernée par une mesure de rappel de produits agroalimentaires contenant des graines de sésame importées d'Inde et présentant des résidus d'un produit chimique, l'oxyde d'éthylène, à une teneur supérieure à la limite maximum réglementaire (250 tonnes de graines de Sésame contaminées, plus de 4300 produits rappelés en France et 700 marques concernées).

Option exportations USA/Canada :
Lorsqu'un assuré procède à des exportations aux Etats-Unis et au Canada, il est en général soumis à des contrôles drastiques sur ces produits. En contrepartie, les procédures sont nombreuses et les règles de droit s'appliquent par Etat. Les montants des indemnisations sont très élevés et défendus par des avocats dont la rémunération est liée au seuil des indemnisations. La garantie de l'assurance RC doit tenir compte de ces spécificités au travers d'une garantie optionnelle.

Option frais de prévention :
En cas de sinistre, l'assuré a l'obligation de prendre les mesures nécessaires pour éviter, diminuer ou supprimer tout préjudice susceptible d'entraîner la mise en jeu des garanties du contrat. La garantie frais de prévention prend en charge ces frais de prévention engagés par l'assuré, sous réserve d'un accord préalable.

2.3. LE TABLEAU SYNOPTIQUE DU PRODUIT RC ENTREPRISES

Le tableau ci-dessous résume les garanties de base et options du contrat RC Entreprises AXA.

En conclusion, l'assurance RC est réputée pour être difficile à tarifer car les risques couverts sont :

- multiples (faute de l'employeur, accident de travail des employés, travaux que l'entreprise réalise à l'extérieur de son établissement, chez un client, biens confiés …)
- hétérogènes du fait de la variété des activités couvertes et de la réglementation susceptible d'encadrer l'activité
- et difficiles à cerner (atteinte accidentelle à l'environnement, risque accru de dommages immatériels consécutifs et pertes financières, rappel ou retrait de produits après livraison, cyber …).

Le risque peut être aggravé si ces travaux sont réalisés en milieu sensible : industries du bois, de la chimie, des matières plastiques, des hydrocarbures. De plus, une entreprise peut envisager sa responsabilité avant la livraison ou après livraison des produits ou réception des travaux sur une période qui peut être longue (rappel de produits pour des problèmes de sécurité par exemple).

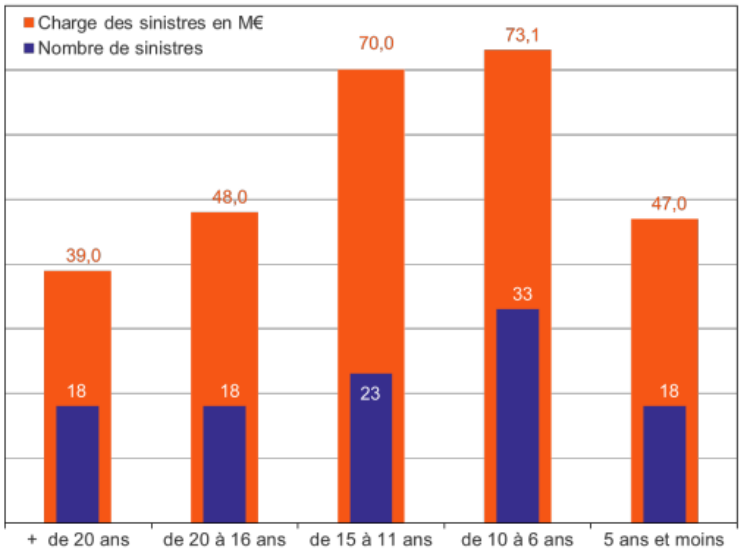
3. LA RC ENTREPRISES : UNE BRANCHE D'INTENSITE

La RC Entreprises est un risque difficile à évaluer car il faut prendre en compte le facteur temps.

3.1 UNE BRANCHE A DEVELOPPEMENT LONG

La RC fait partie de la famille des sinistres à développement long, même si en pratique une partie des sinistres sont à déroulement court. Plusieurs dizaines d'années peuvent parfois s'écouler entre la date du fait à l'origine du dommage, la date de sa manifestation, celle de sa constatation, jusqu'à l'action en responsabilité et la mise en jeu éventuelle du contrat d'assurance du responsable.
De plus, pour les sinistres les plus graves, il faut parfois attendre plusieurs années avant la liquidation (actions en justice, sinistres corporels).

La ventilation des sinistres de la branche RC Entreprises supérieurs à 1 M€ clos en 2019 déclarés à la Fédération Française de l'Assurance montre que sur les 110 sinistres clos, plus de la moitié (59 sinistres), sont survenus il y a plus de 10 ans. Ils totalisent 56% de la charge sinistres.



Source : L'assurance Responsabilité Civile Générale en 2019 publié en février 2021 par la FFA

Sur le produit RC Industries et Commerces AXA que nous étudions qui représente 35% de la RC Entreprises, la répartition des sinistres (hors sinistres sans suite) sur les 10 dernières années (2011-2020) est la suivante.

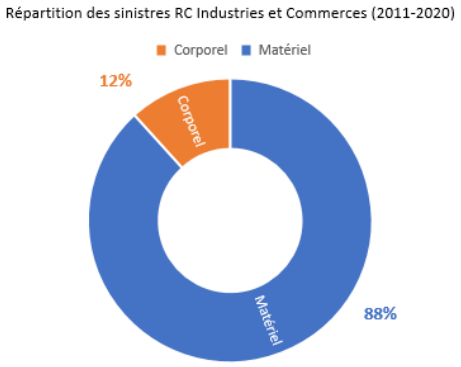
SINISTRES (nbre)	Année clôture										
Année survenance	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**Non Clos**
2011	515	691	280	119	86	72	53	41	27	23	57
2012		597	748	311	104	87	54	45	39	36	104
2013			559	591	297	130	87	64	38	38	131
2014				460	625	317	120	71	66	42	179
2015					455	574	280	151	73	46	217
2016						484	562	263	139	101	320
2017							486	577	293	129	396
2018								480	639	287	550
2019									540	596	885
2020										470	1749
Total général	**515**	**1288**	**1587**	**1481**	**1567**	**1664**	**1642**	**1692**	**1854**	**1768**	**4588**

En % cumulés	Année clôture										
Année survenance	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Non Clos
2011	26%	61%	76%	82%	86%	90%	92%	95%	96%	97%	100%
2012		28%	63%	78%	83%	87%	89%	92%	93%	95%	100%
2013			29%	59%	75%	82%	86%	89%	91%	93%	100%
2014				24%	58%	75%	81%	85%	88%	90%	100%
2015					25%	57%	73%	81%	85%	88%	100%
2016						26%	56%	70%	77%	83%	100%
2017							26%	57%	72%	79%	100%
2018								25%	57%	72%	100%
2019									27%	56%	100%
2020										21%	100%

Nous constatons donc qu'environ (en pourcentages cumulés) :

- 25% des sinistres sont à développement court (clôturés dans l'année)
- 58% des sinistres sont clôturés la deuxième année
- 75% des sinistres sont clôturés la troisième année
- 80% des sinistres sont clôturés la quatrième année
- 90% des sinistres sont clôturés la septième année
- 97% des sinistres sont clôturés la dixième année.

Notre base comprend 12% de sinistres corporels et 88% de sinistres matériels.



Le déroulement de sinistres est plus long sur les sinistres corporels qui représentent 12% du total de nos sinistres (en pourcentages cumulés) :

- 1% des sinistres corporels sont clôturés dans l'année
- 15% la deuxième année
- 55% la cinquième année
- 70% la septième année
- 90% des sinistres corporels sont clôturés la neuvième année.

3.2 UNE BRANCHE D'INTENSITE

La branche RC Entreprises est considérée comme une branche d'intensité : ce n'est pas la fréquence de sinistralité qui la caractérise, mais plutôt l'intensité des sinistres.

Les chiffres ci-dessous présentent la répartition des sinistres AXA clôturés en 2020 de l'ensemble de la branche RC Entreprises. Ces chiffres reflètent l'intensité des sinistres de la branche : 2% des sinistres représentent 50% du montant des sinistres et 16 sinistres de plus d'un million totalisent 18% des montants de sinistres alors qu'ils ne représentent que 0,2% du nombre de sinistres.

Tranche de coût du sinistre (K€)	Nombre de sinistres (1)	Répartition nombre de sinistres (%)	Répartition montant des sinistres (%)
]0 ; 15]	8 340	83,9%	12%
]15 ; 150]	1 412	14,2%	38%
]150 ; 1000]	176	1,8%	32%
> 1000	16	0,2%	18%
Ensemble	**9 944**		

Hors sinistres sans suite
Source : Données AXA – sinistres RC Générale des entreprises clos en 2020

La répartition du nombre de sinistres RC Entreprises AXA est similaire à celle de la FFA (Sinistres clos en 2019). La répartition du montant des sinistres diffère sensiblement de celle de la FFA puisque nos sinistres de plus d'un million représentent 18% de la charge alors que ces sinistres représentent 29% de la charge dans les données FFA tableau ci-dessous.

Tranches de coût	Nombre de sinistres	Montant des sinistres en M€	Part en nombre de sinistre	Poids en montant de sinistre	Coût moyen de la tranche
0 à 15 K€	46 247	122,7	86,2 %	12,7 %	2,7 K€
15 à 150 K€	6 437	295,2	12,0 %	30,5 %	45,9 K€
150 à 1 000 K€	848	268,8	1,6 %	27,8 %	317,0 K€
> 1 000 K€	111	280,5	0,2 %	29,0 %	2 526,8 K€
Ensemble	**53 643**	**967,2**	**100,0 %**	**100,0 %**	**18,0 K€**

Source : L'assurance Responsabilité Civile Générale en 2019 publié en février 2021 par la FFA

Notre base de sinistres correspondant au produit Industries et Commerces contient 48 sinistres de plus d'un million d'euros (entre un million et quatre millions de charge) dont 15 sont clos sur l'historique 2011-2020.
Parmi les sinistres clôturés en 2019 ou 2020 de plus d'un million sur le périmètre Industries et Commerces (4 sinistres 2019 et 3 sinistres 2020), 3 sinistres concernaient des défauts dans des produits (machine de chantier, composants d'un train, balises informatiques), 2 sinistres concernaient des incendies survenus pendant des travaux et 2 sinistres concernaient des produits alimentaires viciés.

En conclusion, l'assurance RC Entreprises fait partie de la famille des sinistres à développement long en raison des procédures judiciaires et des délais entre la date des faits à l'origine du dommage et sa constatation ou réclamation.

La branche RC Entreprises est une branche d'intensité caractérisée par une sinistralité peut fréquente mais des sinistres graves. Notre périmètre d'étude (RC Industries et Commerces) concentre 20 à 25% des sinistres graves de la branche RC Entreprises.

Chapitre 2

PRESENTATION DES DONNEES

Le premier chapitre nous a permis de définir les bases de la responsabilité civile entreprises. Ce deuxième chapitre est consacré au contexte dans lequel s'inscrivent nos travaux et à la description de la base de données qui nous servira d'étude. En effet, l'assurance est un ensemble de marchés très concurrentiels où les données constituent le socle d'une modélisation fiable et performante. La maîtrise de la qualité des données s'impose donc comme un levier de compétitivité et est une préoccupation centrale dans notre démarche de tarification. Une part importante des travaux est consacrée à la collecte, l'analyse, la préparation et le retraitement des données.
Si cette étape est chronophage, elle est essentielle avant d'établir un modèle tarifaire. Nous devons ainsi construire une base de données adaptée qui permettra d'avoir un premier avis sur les variables discriminantes susceptibles d'expliquer la sinistralité avant d'entamer la phase de modélisation.

L'objet de ce chapitre est donc de présenter la méthode de construction de la base de données, identifier les variables disponibles que nous pouvons introduire par la suite dans nos modèles, analyser la sinistralité de notre portefeuille et apprécier la qualité de nos données.

1. LE CONTEXTE DE L'ETUDE

La refonte du tarif concerne le produit de la branche Responsabilité Civile Entreprises disponible sur le périmètre des Industries et des Commerces et quelques activités du transport. Actuellement, la tarification de ce produit peut se faire à l'aide de deux outils de souscription :

- Un ancien outil pour lequel le tarif est calculé à partir du code activité de l'entreprise et de son chiffre d'affaires. Cet outil fournit aux souscripteurs un taux (en pour mille) à appliquer au chiffre d'affaires de l'entreprise pour obtenir la prime d'assurance.
- Un nouvel outil déployé en 2019 et pour lequel le tarif est calculé à partir de différents paramètres renseignés lors de la souscription (code activité, chiffre d'affaires, note Crédit Safe, nombre de sinistres antérieurs, fractionnement de la prime, région : métropole, DOM ou Monaco, distributeur, activité de sous-traitance ou non, activité destination, posture commerciale) et en fonction des options souscrites (export USA/Canada, chiffre d'affaires USA/Canada, protection juridique, frais de retrait, frais de dépose/repose, prévention). Sur cet outil, dont le tarif est segmenté, la souscription n'est possible que pour les entreprises ayant un chiffre d'affaires inférieur à deux millions d'euros.

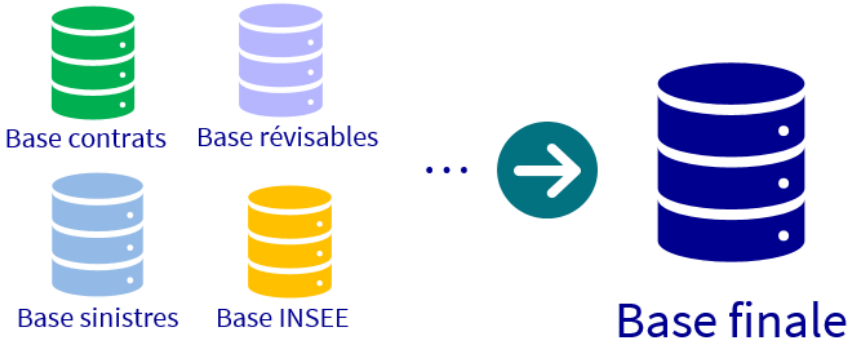
Le nouveau tarif devra prendre en compte les entreprises dont le chiffres d'affaires est compris entre zéro et vingt millions en segmentant le plus justement possible les risques.

Pour refléter l'hétérogénéité et l'exhaustivité des risques, notre historique doit être suffisant pour vérifier la stabilité des données à travers le temps, et pour apprécier correctement le risque. La base de données sera donc constituée initialement à partir d'un historique 2011-2020 (base 10 ans). Nous travaillons sur une branche à déroulement long et il est difficile d'arbitrer entre un historique suffisant pour la robustesse du modèle et un historique suffisamment récent pour mieux refléter le marché.

Pour l'étude du produit Industries et Commerces, nous disposons d'un portefeuille de 44 800 contrats et 40 000 sinistres (dont 21 000 sinistres sans suite) sur notre base 10 ans (2011-2020).

2. LES DONNES DISPONIBLES

Dans les bases de données AXA, nous disposons de nombreuses informations sur les entreprises. Nous disposons également de bases externes pour collecter des informations supplémentaires. Le but est de rassembler pour chaque numéro de contrat toutes les informations utiles pour expliquer la charge de sinistres (bases contrats et sinistres, bases révisables, base des risques, base clients du marketing et bases de l'INSEE qui fournissent des informations détaillées sur les entreprises).
INSEE : Institut National de la Statistique et des Etudes Economiques



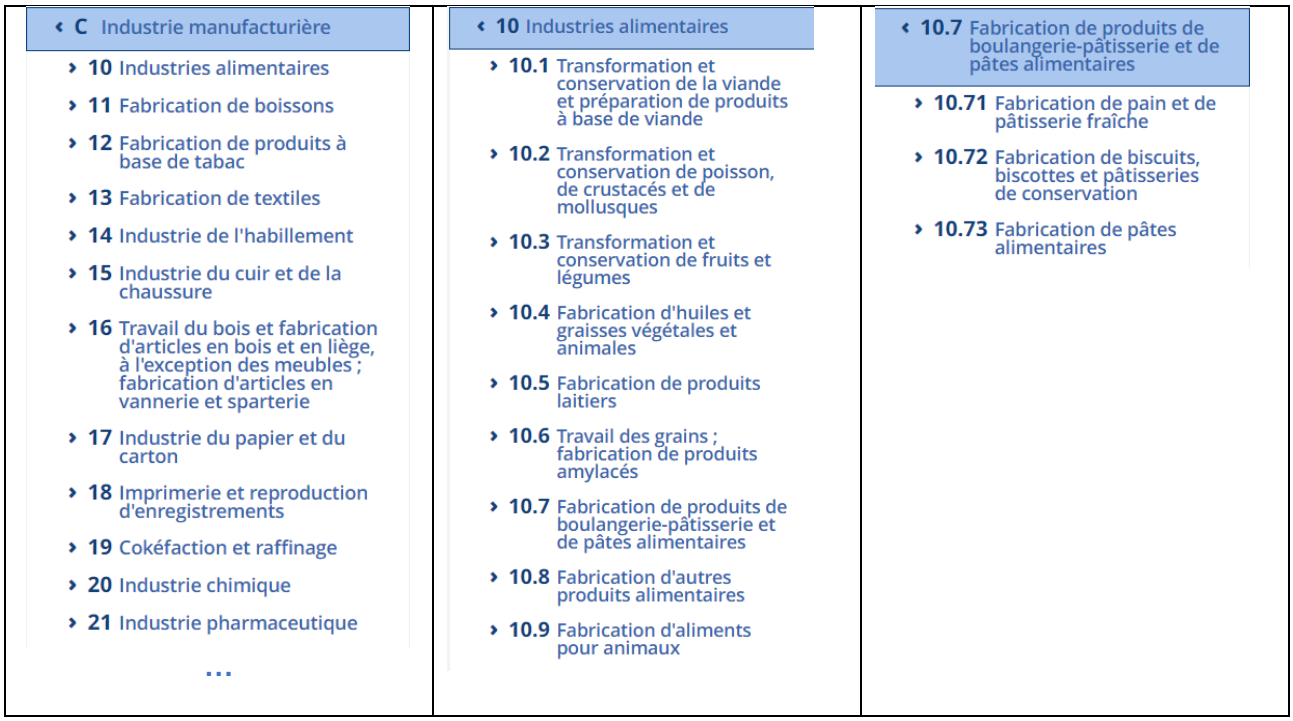
2.1. LA BASE CONTRATS

La première étape de la constitution de la base de données pour la tarification est d'isoler tous les contrats qui concernent le produit étudié, sur un historique de 10 ans : contrats actuellement en cours mais également contrats résiliés à ce jour mais qui étaient en cours sur la période d'étude.
Les contrats RC AXA sont disponibles dans la base Contrats (chaque contrat a un numéro de contrat unique).

Nous sélectionnons les contrats présents au moins une journée entre 2011 et 2020.

Nous sélectionnons les contrats correspondants à notre produit sur la base d'une liste de codes NAF. Au moment de son immatriculation, une entreprise reçoit un code NAF (Nomenclature d'Activités Française qui est la nomenclature des activités économiques productives, élaborée pour faciliter l'organisation de l'information économique et sociale). Le code NAF (ou code APE) permet de déterminer l'activité principale de l'entreprise. Il est composé de 4 chiffres et d'une lettre. Les deux premiers chiffres indiquent le secteur d'activité de l'entreprise ou la division. Le troisième et le quatrième chiffre indiquent plus précisément l'activité exercée (le troisième chiffre indique le groupe et le quatrième indique la classe). Quant à la lettre, elle représente la spécificité de l'économie française.

Exemple : code NAF 10.71D (Pâtisserie)
La pâtisserie fait partie des industries manufacturières alimentaires (10) et plus précisément des industries de fabrication de produits de boulangerie-pâtisserie et de pâtes alimentaires (10.7) et plus précisément encore de la fabrication de pain et de pâtisserie fraîche (10.71).



Pour effectuer notre tarification, les codes NAF de l'INSEE sont transformés en code NAF AXA qui sont soit des regroupements de codes NAF INSEE, soit une division plus précise du code INSEE.

Par exemple le code10.71D (Pâtisserie) est divisé en 4 codes NAF AXA :

NAF AXA	Libellé NAF AXA
158A100	Fabrication boulangerie et/ou pâtisserie industrielle
158A110	Fabrication Pains et Viennoiseries
158A120	Fabrication de pâtisserie fraîche
158C100	Fabrication artisanale et commerce de détail Boulangerie - pâtisserie

Le code NAF AXA 158A100 Fabrication boulangerie et/ou pâtisserie industrielle regroupe les codes NAF INSEE suivants :

NAF INSEE	Libellé NAF INSEE
10.71A	Fabrication industrielle de pain et de pâtisserie fraîche
10.71B	Cuisson de produits de boulangerie
10.71C	Boulangerie et boulangerie-pâtisserie
10.71D	Pâtisserie
10.72Z	Fabrication de biscuits, biscottes et pâtisseries de conservation
10.85Z	Fabrication de plats préparés

Comme dans la base INSEE, nous avons différents niveaux de regroupement des codes NAF (Niveaux 2 à 6) et un niveau de regroupement par secteurs (Agroalimentaire, Métallurgie, Textile/habillement, …).

Dans notre base Contrats, nous sélectionnons les contrats révisables, c'est-à-dire les contrats dont la prime est indexée sur le chiffre d'affaires ou sur une autre métrique (par exemple nombre de tête de bétail). Sur le périmètre Industries et Commerces, le chiffre d'affaires constitue notre métrique. Une partie de nos contrats à la marge ont une prime forfaitaire et ces contrats sont donc supprimés de notre base.

Notre base initiale comprend 44 800 contrats et 215 000 lignes. Nous avons en effet pour chaque contrat une ligne par année de vision du contrat dans nos bases (entre 2011 et 2020).

La base Contrats contient les numéros de SIREN et SIRET des entreprises. Le numéro SIREN est l'identifiant national de l'entreprise. Il est composé d'une suite de 9 chiffres. Il est attribué à vie. Le numéro SIRET permet d'identifier géographiquement chaque entreprise ou chaque établissement de la même entreprise. Il identifie le lieu où est produit l'activité du ou des établissements que détient l'entreprise. Si l'activité est produite dans plusieurs locaux, chacun d'eux recevra un numéro SIRET différent. Il est composé de 14 chiffres : le numéro SIREN (9 chiffres) + le numéro NIC (5 chiffres). Le NIC (numéro interne au classement) définit l'emplacement géographique.

Le numéro SIRET est utilisé pour rechercher des données complémentaires dans les bases INSEE mais ce numéro n'est pas toujours bien renseigné dans notre base contrats. Les numéros de SIRET sont vérifiés et complétés à l'aide de nos base Marketing Clients et de la base INSEE.

La base Contrats contient toutes les informations relatives au contrat mais également à la souscription (garanties, montants de garantie, fractionnement de la prime, …) ainsi que les données géographiques de l'entreprise.

Pour la tarification, nous devons ajouter à notre base contrats des informations sur le chiffre d'affaires (issu de la base Révisables) et des informations sur l'entreprise issues de la base INSEE.

2.2. LA BASE DES REVISABLES

La base Révisables contient les contrats dont la prime est révisée régulièrement selon certains critères (chiffre d'affaires, nombre de salariés, …). Elle contient les informations que l'assuré nous déclare afin que sa prime soit révisée et également la façon dont la prime a été recalculée (en fonction des déclarations de l'assuré). Chaque année, nous demandons au client d'actualiser les informations fournies lors de la souscription. Nous récupérons dans cette base le chiffre d'affaires de l'entreprise, la prime du produit Industries et Commerces étant révisée suivant le chiffre d'affaires.

2.3. LA BASE INSEE

Disponible sur le site de l'INSEE, la base Sirene fournit des données d'identité des entreprises et des établissements. Elle nous permet à l'aide du code SIRET d'enrichir nos données contrats en créant les variables suivantes qui seront étudiées pour la tarification :

- Monoactivité (indique si l'entreprise exerce une seule activité à 100% ou le % de son activité principale exemple 80% si elle exerce plusieurs activités)
- Nombre d'établissements actifs du SIREN
- Saisonnalité de l'activité
- Age du dirigeant
- Tranche de chiffre d'affaires de l'entreprise
- Taille de la commune de l'établissement
- Note Credit Safe. Cette note, calculée par Credit Safe à partir des données financières de l'entreprise, nous donne des informations sur la santé financière de nos clients. C'est une note qui va de 1 à 20, plus elle est élevée et plus l'entreprise est solvable.
- Ancienneté de l'entreprise

- Ancienneté de l'établissement
- Lieu d'activité qui pourra être utilisé pour un zonier …

2.4. LA BASE SINISTRES

La tarification d'une garantie se base sur la modélisation de la charge de sinistre observée sur cette garantie. Ainsi, après avoir récupéré toutes les données relatives au contrat, il faut ajouter les données de sinistralité par contrat et par année d'observation sur la période de l'étude. Nous devons donc sélectionner tous les sinistres survenus entre le 01/01/2011 et 31/12/2020 sur les contrats de notre périmètre (avec une vision à la fin de chaque année N mais également une vision à fin décembre 2021).

Nous devons également extraire 5 années supplémentaires pour obtenir les sinistres antérieurs sur 5 ans, cette information étant demandée sur notre nouvel outil de souscription.

La création de la sinistralité associée à la modélisation de l'étude s'opère en plusieurs étapes. Nous devons obtenir les variables suivantes :

- Etat du sinistre (en cours, clos, sans suite ou annulé),
- Date de survenance, date d'ouverture
- Garanties mises en jeu
- Nature du sinistre (corporel, matériel)
- Charges : différentes visions (N, N+1 …) et différents montants : règlements, recours, frais, évaluations de règlements et de recours
- Contrat auquel le sinistre est attaché
- Activité sinistrée
- Lieu du sinistre
- Date de clôture
- Date d'observation …

A noter qu'en RC Entreprises, une part importante de sinistres sont classés sans suite. Chez AXA, le taux de sans suite en RC Entreprises se situe entre 50 % et 60 %. Sur les garanties RC, nous sommes en défense et la partie adverse doit faire la preuve de la responsabilité de nos assurés. Beaucoup d'expertises aboutissent à la conclusion de « non responsabilité ». Les dommages peuvent également être inférieurs à la franchise ou non justifiés. Beaucoup de dossiers sont donc clôturés « sans suite » avec les seuls frais de l'expertise.

Notre base Sinistres contient :

- 18 500 sinistres en cours ou clos de 2011 à 2020
- 21 100 sinistres sans suite ou annulés sur cette même période.

Dans la constitution de notre base de données sinistres, nous séparons la charge de sinistres attritionnels, la charge de sinistres graves et la charge de sinistres atypiques. Les seuils de sinistres graves et atypiques sont fixés et ne sont pas l'objet de cette étude.

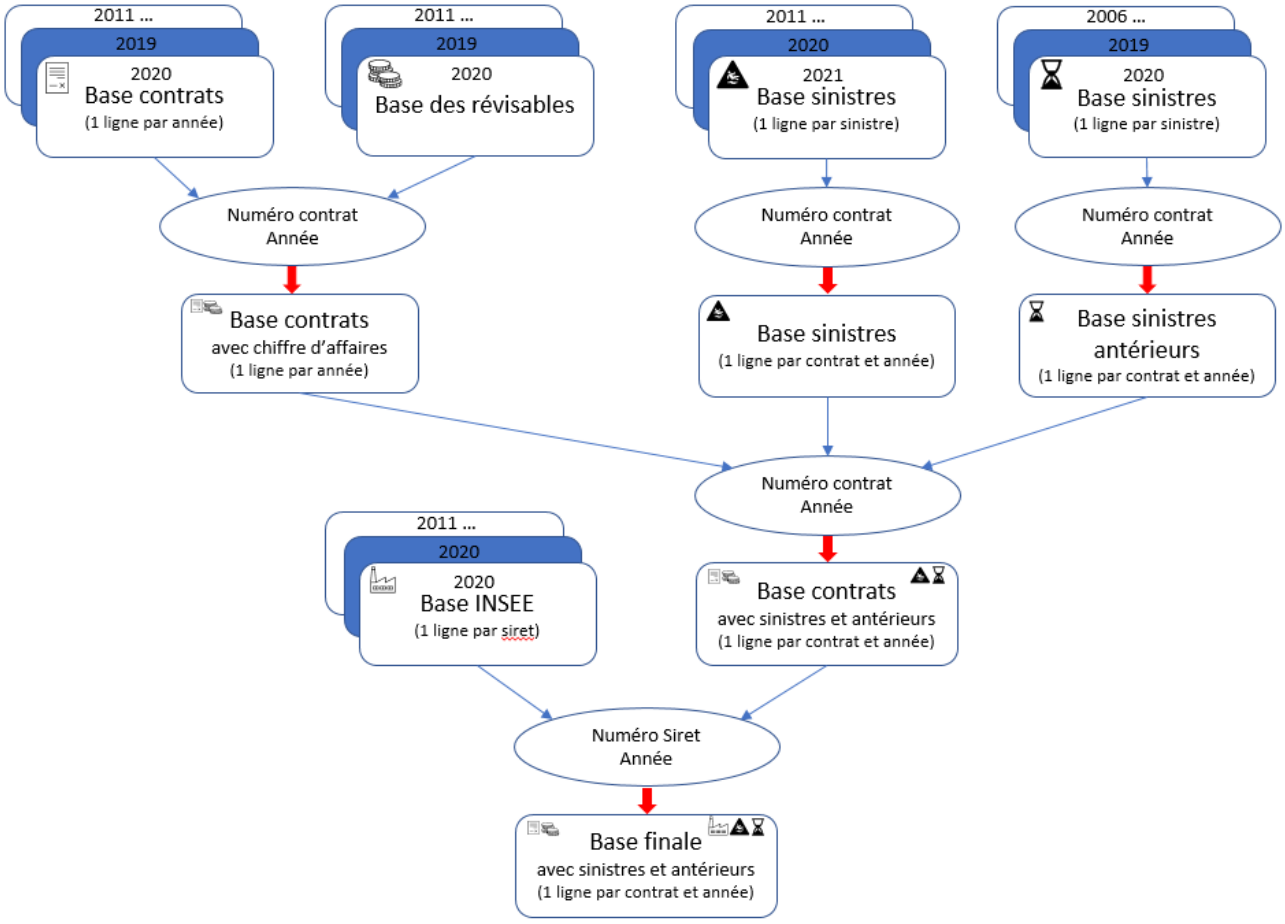
Un sinistre est considéré comme attritionnel lorsque son montant n'excède pas 150 000 euros. Un sinistre est grave lorsque sa charge est comprise 150 000 euros et un million d'euros. Enfin, un sinistre est atypique lorsque son montant dépasse un million d'euros. Les valeurs de ces seuils ont

été déterminées lors de précédentes études et sont communes à plusieurs produits et branches de l'assurance Entreprises.

Notre base d'étude initiale (2011-2020) agrégée contient :

- 44 800 contrats qui représentent 215 000 lignes (une ligne par année de vision du contrat).
- 18 500 sinistres en cours ou clos
- 21 100 sinistres sans suite ou annulés.

Le schéma ci-dessous décrit de manière synthétique la base de données utilisée.



3. LA SINISTRALITE

Nous avons sélectionné dans notre base à fin décembre 2021, tous les sinistres survenus entre 2011 et 2020 ainsi que tous les sinistres antérieurs depuis 2006. Il est important de prendre conscience de la nature de la sinistralité du portefeuille. Dans cette partie, nous allons donc étudier la distribution du nombre de sinistres ainsi que les charges sur notre période d'étude (2011-2020).

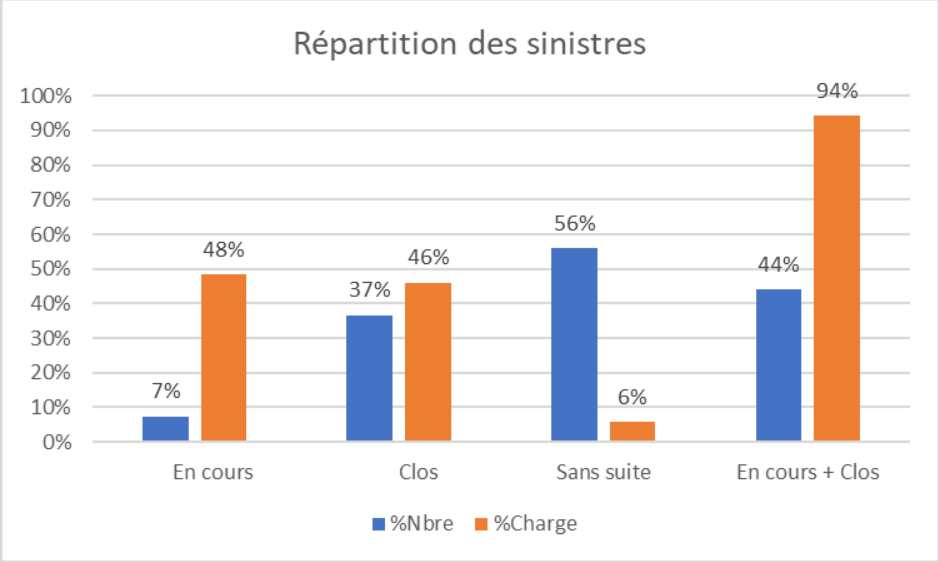
3.1. L'ANALYSE DE LA SINISTRALITE

Nous avons en moyenne 4500 sinistres par an sur notre produit Industries et Commerces.

Survenance	Nbre sinistres vus fin N
2011	4761
2012	5025
2013	4591
2014	4429
2015	4393
2016	4366
2017	4270
2018	4353
2019	4353
2020	4138
Moyenne	4468

Ces sinistres sont répartis en trois groupes en fonction de leur état :

- En cours
- Clos
- Sans suite ou annulé.



Nos sinistres en cours et clos représentent plus de 94% de notre charge pour 44% des sinistres en volumes. Nos sinistres sans suite représentent 6% de notre charge pour un volume de sinistres de 56%.

Comme indiqué précédemment ce taux important de sinistres sans suite est lié au fait que nous sommes en défense, la partie adverse devant apporter la preuve de la responsabilité de nos assurés. Beaucoup de dossiers sont donc clôturés « sans suite » avec les seuls frais de l'expertise (environ 1200€).

Nous allons donc isoler les sinistres sans suite pour en calculer la charge inflatée et vieillie, avant de supprimer ces sinistres sans suite pour ne pas biaiser notre étude (volumes importants pour une charge faible). Leur charge vieillie inflatée sera réintroduite lors du passage de la prime pure à notre prime commerciale.

En première analyse, nous vérifions les cadences d'ouverture et de clôture de nos sinistres en nombre et en pourcentage en supprimant les sinistres sans suite qui sont ouverts et clôturés rapidement.

Sinistres par année de survenance selon l'année d'ouverture

En nombre de sinistres (hors sinistres sans suite)

Ouverture	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total général
Survenance												
2011	1583	354	14	6	3	1	1	2				1964
2012		1750	344	23	5	1	1		1			2125
2013			1606	301	22		4	2				1935
2014				1601	249	21	5	2	1	1		1880
2015					1539	237	11		2	2	5	1796
2016						1601	247	11	7	2	1	1869
2017							1655	202	18	4	2	1881
2018								1740	197	10	9	1956
2019									1789	194	38	2021
2020										1897	322	2219
Total général	1583	2104	1964	1931	1818	1861	1924	1959	2015	2110	377	19646

En répartition cumulée

N	N+1	N+2	N+3

Ouverture	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total général
Survenance												
2011	80,60%	98,63%	99,34%	99,64%	99,80%	99,85%	99,90%	100,00%	100,00%	100,00%	100,00%	
2012		82,35%	98,54%	99,62%	99,86%	99,91%	99,95%	99,95%	100,00%	100,00%	100,00%	
2013			83,00%	98,55%	99,69%	99,69%	99,90%	100,00%	100,00%	100,00%	100,00%	
2014				85,16%	98,40%	99,52%	99,79%	99,89%	99,95%	100,00%	100,00%	
2015					85,69%	98,89%	99,50%	99,50%	99,61%	99,72%	100,00%	
2016						85,66%	98,88%	99,46%	99,84%	99,95%	100,00%	
2017							87,99%	98,72%	99,68%	99,89%	100,00%	
2018								88,96%	99,03%	99,54%	100,00%	
2019									88,52%	98,12%	100,00%	
2020										85,49%	100,00%	
Total général	1583	2104	1964	1931	1818	1861	1924	1959	2015	2110	377	19646

Nous constatons que nos sinistres sont ouverts à plus de 80% dans l'année où ils surviennent (cellules en vert). A partir de N+3, nous n'ouvrons quasiment plus de sinistres sur la survenance N. Sur les survenances les plus anciennes, moins de 10 sinistres supplémentaires seront ouverts en cumulé sur les années de survenance au-delà de N+3. Nous pouvons donc considérer qu'en N+3, nous avons la quasi-totalité de nos sinistres (moins de 0,3% manquants) sur lesquels nos gestionnaires ont positionné des provisions après étude du sinistre.

Nous réitérons cette analyse en remplaçant la date d'ouverture par la date de clôture afin de vérifier la cadence de clôture de nos sinistres.

Sinistres par année de survenance selon l'année de clôture

En nombre de sinistres (hors sinistres sans suite)

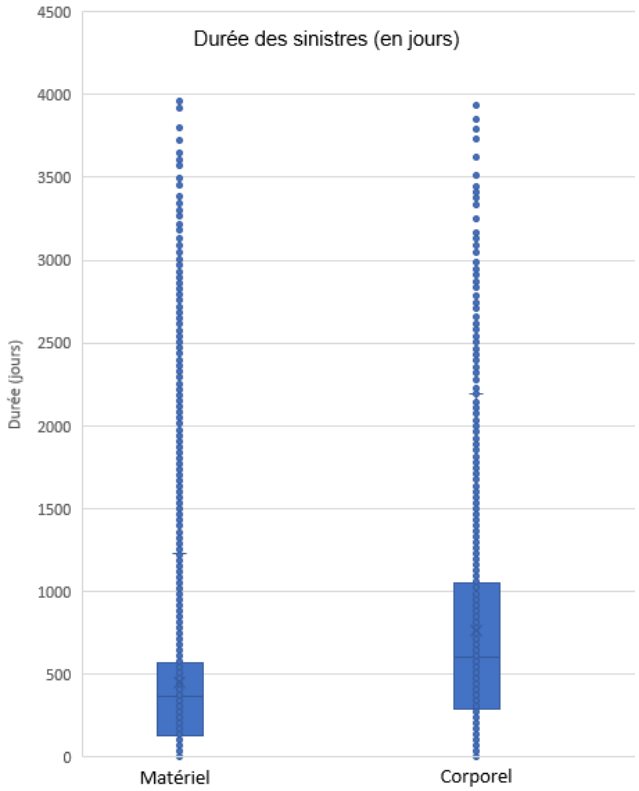
Clôture	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Non clos	Total général
Survenance													
2011	515	691	280	119	86	72	53	41	27	23	13	44	1964
2012		597	748	311	104	87	54	45	39	36	24	80	2125
2013			559	591	297	130	87	64	38	38	34	97	1935
2014				460	625	317	120	71	66	42	42	137	1880
2015					455	574	280	151	73	46	31	186	1796
2016						484	562	263	139	101	70	250	1869
2017							486	577	293	129	78	318	1881
2018								480	639	287	136	414	1956
2019									540	596	262	623	2021
2020										470	617	1132	2219
Total général	515	1288	1587	1481	1567	1664	1642	1692	1854	1768	1307	3281	19646

En répartition cumulée

	N	N+1	N+2	N+3

Clôture	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Non clos	Total général
Survenance													
2011	26,22%	61,41%	75,66%	81,72%	86,10%	89,77%	92,46%	94,55%	95,93%	97,10%	97,76%	2,24%	
2012		28,09%	63,29%	77,93%	82,82%	86,92%	89,46%	91,58%	93,41%	95,11%	96,24%	3,76%	
2013			28,89%	59,43%	74,78%	81,50%	85,99%	89,30%	91,27%	93,23%	94,99%	5,01%	
2014				24,47%	57,71%	74,57%	80,96%	84,73%	88,24%	90,48%	92,71%	7,29%	
2015					25,33%	57,29%	72,88%	81,29%	85,36%	87,92%	89,64%	10,36%	
2016						25,90%	55,97%	70,04%	77,47%	82,88%	86,62%	13,38%	
2017							25,84%	56,51%	72,09%	78,95%	83,09%	16,91%	
2018								24,54%	57,21%	71,88%	78,83%	21,17%	
2019									26,72%	56,21%	69,17%	30,83%	
2020										21,18%	48,99%	51,01%	
Total général	515	1288	1587	1481	1567	1664	1642	1692	1854	1768	1307	3281	19646

Nous constatons que nos sinistres sont clôturés à plus de 80% en N+3 sur les survenances les plus anciennes et à un peu moins de 80% sur les survenances les plus récentes (cellules en orange). En N+3, nous avons donc une part très importante de sinistres clôturés.



L'analyse de la durée des sinistres en jours confirme ce point. La durée médiane des sinistres matériels est de 376 jours et celle des sinistres corporels de 707 jours.

82% des sinistres matériels ont une durée inférieure à 2 ans et 91% une durée inférieure à 3 ans.

52% des sinistres corporels ont une durée inférieure à 2 ans et 72% une durée inférieure à 3 ans.

Les sinistres matériels représentent en volumes 89% de notre base.

La vision de nos sinistres en année N+3 semble pertinente puisque, pour chaque année de survenance : plus 99,5% des sinistres du produit RC Industries et Commerces sont ouverts en N+3 :

- plus de 99,5% ont été déclarés et donc provisionnés par les gestionnaires sinistres
- environ 80% de nos sinistres sont clos. Nous avons donc une charge nette de sinistre sécurisée sur 80% du périmètre.

Nous allons maintenant étudier l'évolution de notre charge nette de sinistres sur notre historique. La charge nette disponible dans notre base Sinistres est le cumul des règlements en indemnités, en frais accessoires et les évaluations de règlements desquels sont déduits les recours encaissés et les évaluations de recours à percevoir.

3.2. LA CHARGE SINISTRE

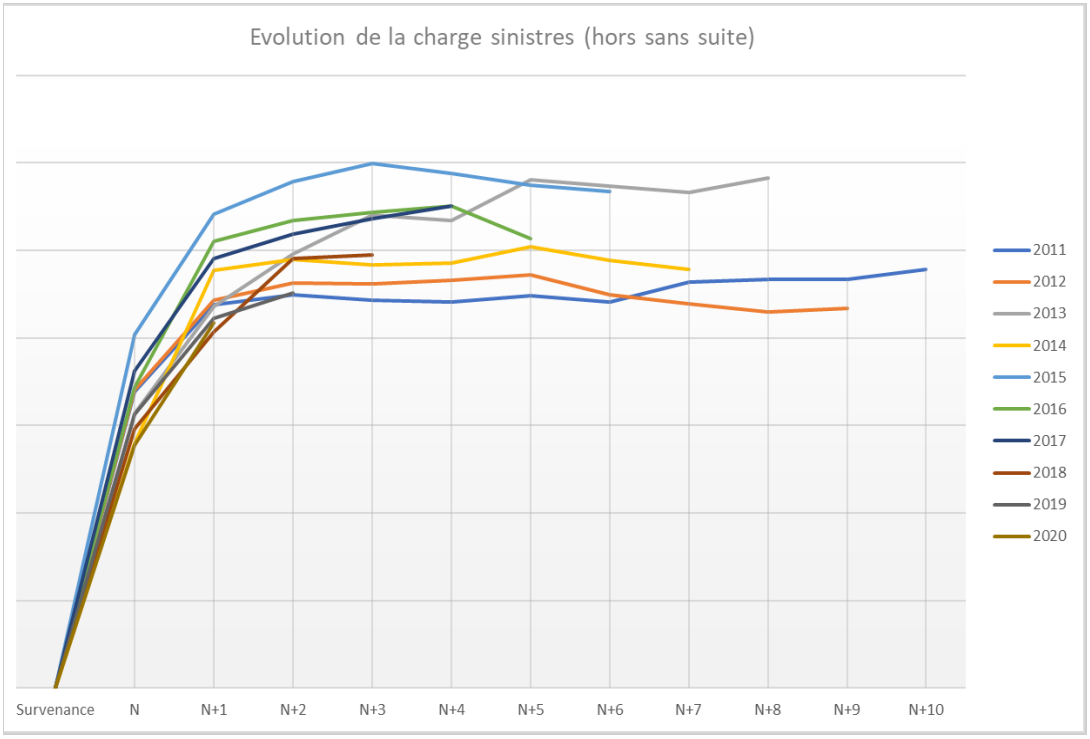
Nous travaillons sur une branche à déroulement long et il est difficile d'arbitrer entre un historique suffisant pour la robustesse du modèle et un historique suffisamment récent pour mieux refléter le marché.

3.2.1. Vieillissement des charges de sinistres

Ne disposant pas des outils ni données nécessaires à la projection des provisions pour sinistres non encore manifestés, ni des charges ultimes à la maille contrat ou case tarifaire (ces chiffres sont calculés au global sur le produit), nous décidons de modéliser une charge Dossier / Dossier (D/D)

plutôt qu'une Charge Finale Prévisible que nous ne pouvons pas calculer par contrat. C'est la charge modélisée qui sera ensuite projetée à l'ultime, à travers un coefficient de passage.
Nous définissons le coefficient de proportionnalité = Charge année N / Charge 2021. Notre objectif est de trouver un âge de survenance qui maximise ce ratio tout en ayant un périmètre satisfaisant.

Le graphique ci-dessous présente l'évolution de notre charge D/D par année de survenance (pour les sinistres clos et en cours uniquement, hors sinistres sans suite). Comme pour le nombre de sinistres qui se stabilise en N+3, la courbe d'évolution des charges de chaque année étudiée commence à se stabiliser en N+3, ce qui semble le meilleur compromis pour conserver un historique suffisant.



En effet, si nous choisissions par exemple 5 années de vieillissement (N+5) pour notre charge, nous ne pourrions conserver que l'historique 2011-2016 pour lequel nous aurions moins d'observations et une sinistralité plus ancienne reflétant moins bien la sinistralité actuelle et future.

Nous souhaitons conserver au minimum 80% de la charge 2021, cela implique donc de modéliser la charge N+3. Le périmètre se définit donc directement comme allant de l'année 2011 à l'année 2018 pour laquelle nous avons la charge à fin 2021. Il est important que chaque année de survenance ait la même longueur d'historique pour ne pas déformer la sinistralité par contrat.
En effet, si nous conservions un historique plus long pour les années de survenance les plus anciennes, en agrégeant nos sinistres par secteur d'activité par exemple dans les modèles, nous apporterions une déformation entre les secteurs qui ont des contrats anciens et ceux dont les contrats sont plus récents, les activités éligibles au produit Industries et Commerces pouvant évoluer en fonction des politiques de souscription.

Sur la base d'étude charge N+3 (sans les sinistres 2019 et 2020), les sinistres sans suite représentent 6,21% de la charge et 57,43% des volumes. La charge N+3 totale (y compris les sinistres sans suite en vision N+3) représente 98,86% de la charge finale 2021. Cette même charge N+3 hors sinistres sans suite représente 92,59% de la charge finale 2021.

Nous modélisons donc directement la charge D/D vue en N+3 sur l'historique 2011-2018. Nous appliquerons ensuite le coefficient de proportionnalité pour obtenir :

Prime pure à l'ultime = Prime pure modélisée D/D vue en N+3 * coefficient de proportionnalité

3.2.2. Inflation de notre charge de sinistres

Compte-tenu de la profondeur de notre historique de 10 ans (première année prise en compte : 2011), nous devons revaloriser nos coûts de sinistres.

L'inflation nous permet de ramener nos charges sinistres à des euros comparables (As if). Il s'agit ici de rapporter toutes les données passées à une vision fin 2021. Cette inflation est indispensable pour comparer les charges issues de survenances différentes. En effet, ce qui coûtait 1€ en 2011 coûte plus cher en 2021 du fait de l'inflation. Nous séparons les sinistres matériels et corporels qui sont inflatés de manière différente :

- L'indice pris en compte pour inflater nos sinistres matériels est celui du coût de la construction produit par la Fédération Française du Bâtiment (FFB). L'indice FFB est utilisé pour réviser le montant des primes d'assurance de la plupart des contrats d'assurance notamment habitation. Nous l'utilisons également sur la branche RC. Ce mécanisme a pour finalité de refléter l'augmentation des coûts dans le temps.
- Nous intégrons une inflation de 5% chaque année pour les sinistres corporels.

Cette inflation est appliquée sur notre charge N+3 par rapport à la survenance. Ainsi pour un sinistre survenu en 2011, nous conservons sa charge à fin 2014 que nous inflatons à l'aide du tableau suivant pour la ramener à des euros fin 2021 :

Année	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Sinistre matériel	17,9%	14,9%	12,6%	11,4%	11,6%	10,1%	6,4%	5,0%	4,3%	3,7%	0,0%
Sinistre corporel	62,9%	55,1%	47,7%	40,7%	34,0%	27,6%	21,6%	21,6%	15,8%	5,0%	0,0%
Survenance				2011	2012	2013	2014	2015	2016	2017	2018

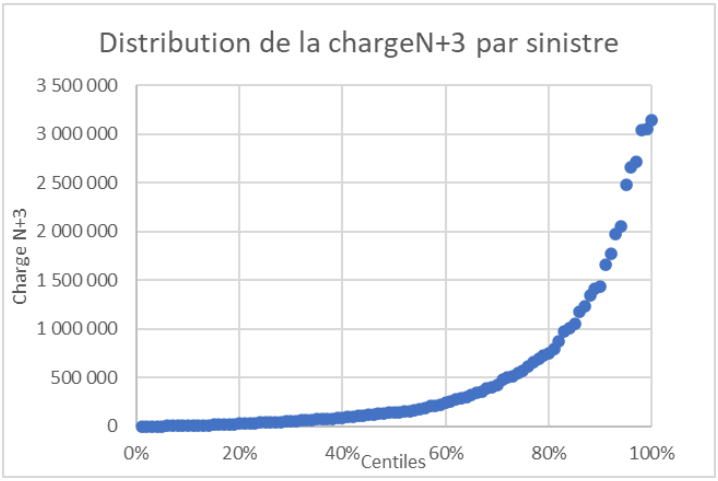
Sur un recul de dix ans, l'impact sur la charge N+3 au global reste modéré (+8,3%), ce qui n'est en revanche pas le cas sur les exercices les plus anciens et sur les sinistres corporels.

Sur la base de notre étude du déroulement des sinistres et de la charge sinistres, nous conservons une charge N+3, qui nous permet d'avoir une base suffisamment robuste.

- Notre charge N+3 (y compris sinistres sans suite) représente 98,86% de notre charge
- Les sinistres sans suite représentent 6,21% de la charge pour 57,43% des volumes
- Nous ramenons la charge N+3 à des euros 2021 pour chaque sinistre retenu dans notre base cible.

3.2.3. Analyse de la charge de sinistres

Nous allons nous intéresser maintenant à la distribution de la charge N+3 des sinistres. Sur le graphique ci-dessous, nous constatons que nous avons un seuil à un million au-delà duquel notre courbe n'est plus continue, le nombre de sinistres diminuant fortement. Ce seuil d'un million d'euros correspond à notre seuil fixé pour un sinistre atypique sur le périmètre de la Responsabilité Civile.



Le tableau suivant présente la répartition du nombre de sinistres et de la charge en fonction de différents seuils fixés (y compris sinistres sans suite). La médiane des charges cumulées se situe à un seuil d'environ 145 150€ qui correspond au seuil fixé de 150 000€ pour un sinistre grave sur le périmètre de la Responsabilité Civile.

Seuil fixé	Nbre de dossiers	Part des dossiers	Charge cumulée	Charge moyenne	Part des charges
0	36 101	100%	431 118 000	11 942	100%
1 000	18 162	50,31%	426 738 159	23 496	98,98%
5 000	7 486	20,74%	401 807 608	53 675	93,20%
10 000	4 847	13,43%	383 138 430	79 047	88,87%
50 000	1 602	4,44%	307 662 296	192 049	71,36%
100 000	816	2,26%	252 484 797	309 418	58,57%
140 000	533	1,48%	219 051 523	410 978	50,81%
150 000	473	1,31%	210 303 883	444 617	48,78%
200 000	341	0,94%	188 078 457	551 550	43,63%
500 000	119	0,33%	122 024 378	1 025 415	28,30%
1 000 000	43	0,12%	71 637 681	1 665 993	16,62%
1 500 000	19	0,05%	42 733 652	2 249 140	9,91%

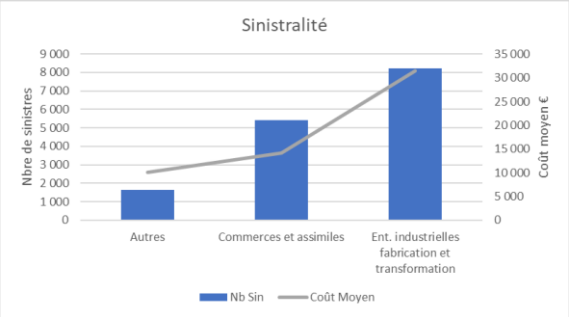
Sur notre périmètre produit, le nombre de sinistres atypiques est très faible (moins de 0,2% des sinistres sur un historique de 10 ans). Nous ne procédons donc pas à leur modélisation. Une charge supplémentaire sera ajoutée à notre prime pure modélisée afin de tenir compte du poids de ces sinistres en fonction de leur espérance de survenance.

Les sinistres graves (environ 500 sinistres et 32% de la charge totale) seront modélisés avec les sinistres attritionnels de manière à ce qu'ils impactent directement les regroupements d'activités.

Les sinistres graves impactent en effet plus fréquemment le secteur de l'industrie que celui du commerce.

RC Industrie et Commerces - Années 2011-2020
Sinistres vus à N+3 inflatés

Regroupement de secteurs d'activités (3 niveaux)		Nombre de sinistres					Charge Sinistres			
	Nb contrats	Total Attri+Graves	Attritionnels	(%)	Graves	(%)	Total Attri+Graves	Attritionnels	(%)	Graves	(%)
Autres	13 100	1 658	1 643	99%	15	1%	16 713 416	12 569 811	75%	4 143 606	25%
Commerces	58 828	5 415	5 330	98%	85	2%	76 747 243	51 477 378	67%	25 269 865	33%
Ent. Industrielles	110 060	8 239	7 854	95%	385	5%	259 309 632	136 203 110	53%	123 106 521	47%
Total	**181 988**	**15 312**	**14 827**	97%	**485**	3%	**352 770 291**	**200 250 299**	57%	**152 519 992**	43%



Sinistralité attritionnelle et grave selon le regroupement en 3 secteurs d'activités

4. L'EXPLORATION DES DONNEES

L'analyse de la sinistralité (volumes et charges de sinistres) a permis de déterminer les variables sinistres à prendre en compte pour réaliser notre modélisation. Il convient dans cette partie d'analyser les données de notre base (tris à plat), d'étudier la complétion des données et de retraiter/regrouper les modalités de certaines variables afin d'obtenir des groupes d'étude de taille suffisante.

4.1. LES VARIABLES

Pour chaque contrat et chaque année, nous disposons d'informations relatives :

- A l'activité de l'entreprise (chiffre d'affaires, code NAF, sensibilité du NAF, segmentation, secteur d'activité, différents niveaux de regroupements d'activités)
- Aux données de souscription (exposition, année de souscription, fractionnement de la prime, réseau de distribution, tranche de montant garantie ainsi que les différentes garanties souscrites : export USA, export hors USA, protection juridique, faute inexcusable de l'employeur, pollution et frais de retrait)
- Aux données INSEE (catégorie juridique d'entreprise, tranche d'effectifs, note Credit Safe, âge du dirigeant, âge de l'établissement, nombre d'établissement actifs, tranche INSEE de CA)
- Aux sinistres (exercice de survenance, charge de sinistres antérieurs, nombre de sinistres antérieurs, nombre de sinistres sans suite, attritionnels, graves et atypiques et charges N+3 inflatées de sinistres sans suite, attritionnels, graves et atypiques).
- Aux données géographiques de l'entreprise (région, département, commune et différents niveaux de regroupement géographiques).

Le tableau en annexe 2 décrit l'ensemble des variables disponibles pour notre étude.

4.2. LA QUALITE DES DONNEES

Pour avoir une meilleure connaissance du portefeuille et de la qualité de notre base, il est important d'étudier chaque variable individuellement. Pour chaque variable, nous réalisons des tableaux et graphiques permettant de visualiser pour chaque modalité de chaque variable :

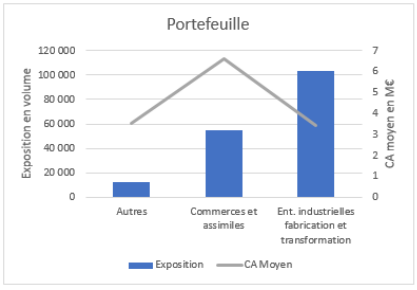
- L'exposition (1 si le contrat est présent en portefeuille toute l'année),
- Le chiffre d'affaires (CA) des entreprises,
- Le CA moyen,
- Le nombre de contrats,
- Le nombre de sinistres (hors sinistres sans suite),
- La charge sinistres (attritionnelle, grave et atypique),
- Le coût moyen d'un sinistre,
- Et un indicateur de charge sinistres divisée par le CA de l'entreprise.

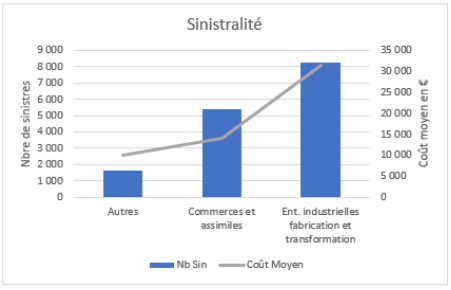
L'exemple ci-dessous présente la variable «Nivlib2_group » qui segmente les entreprises en 3 modalités :

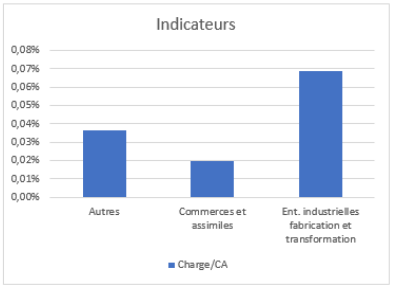
- Commerces et assimilés
- Entreprises industrielles de fabrication et transformation
- Autres.

RC Industrie et Commerces
Années 2011-2020
Sinistres vus à N+3 inflatés

ENT_nlibniv2_group	Portefeuille			Nb contrats	Nombre de sinistres				Charge Sinistres				Indicateurs	
	Exposition	CA	CA moy (M€)		Total Attri+Graves	Attritionnels	Graves	Atypiques	Total Attri+Graves	Attritionnels	Graves	Atypiques	Coût Moyen	Charge / CA
Autres	12 258 7%	45 877 233 141 6%	3,50	13 100 7%	1 658	1 643	15	0	16 713 416	12 569 811	4 143 606	0	10 080,5	0,0364%
Commerces et assimiles	55 213 32%	389 496 914 660 48%	6,62	58 828 32%	5 415	5 330	85	5	76 747 243	51 477 378	25 269 865	7 041 244	14 173,1	0,0197%
Ent. industrielles fabrication et transformation	103 617 61%	375 835 173 050 46%	3,41	110 060 60%	8 239	7 854	385	40	259 309 632	136 203 110	123 106 521	72 380 379	31 473,4	0,0690%
Total	**171 089** 100%	**811 209 320 851** 100%	**4,46**	**181 988** 100%	**15 312**	**14 827**	**485**	**45**	**352 770 291**	**200 250 299**	**152 519 992**	**79 421 624**	**23 038,8**	**0,0435%**







Sur cet exemple, nous avons majoritairement en portefeuille des entreprises industrielles. Les commerces ont un chiffre d'affaires en moyenne plus élevé que celui des entreprises industrielles. En ce qui concerne la sinistralité, la fréquence de sinistres est équivalente (histogramme du nombre de sinistres identique à l'histogramme d'exposition), cependant, le coût moyen d'un sinistre est plus élevé pour les entreprises industrielles. En rapportant la charge de sinistres au chiffre d'affaires de l'entreprise, les entreprises industrielles et autres sont plus sinistrées que les commerces dont le chiffre d'affaires est plus élevé. Cette variable sera donc intéressante pour notre modèle.

A travers ces tris à plat, nous cherchons à identifier les variables avec des données manquantes, notamment pour les données INSEE (modalité 00-NR), ainsi les variables pour lesquelles l'information est concentrée sur une modalité.

L'analyse univariée de chaque variable permet également d'apprécier la qualité des données à notre disposition.

Notre base de données initiale comporte un nombre élevé de variables explicatives, un nombre de modalités important pour certaines variables ainsi que des valeurs manquantes.

Une étape préalable consiste donc à vérifier la qualité des données et effectuer des prétraitements sur les variables avant de les utiliser dans notre modèle :

- Traitement des valeurs manquantes et aberrantes
- Discrétisation des variables avec recodage en tranches lorsque le nombre de modalités est trop important
- Etude et sélection des variables corrélées à la cible
- Etude des corrélations entre les variables explicatives sélectionnées.

4.2.1. Gestion des données manquantes

Les valeurs manquantes déséquilibrent les analyses et sont génératrices de biais dans nos futures conclusions. Il est donc important de prendre en compte et corriger si possible ces données avant toute modélisation.

En responsabilité civile, les variables qui intuitivement semblent influer sur le tarif sont le chiffre d'affaires de l'entreprise et le secteur d'activité de l'entreprise. Une attention particulière doit donc être portée à ces variables.

Notre variable chiffre d'affaires de l'entreprise comporte 9% de valeurs manquantes ou incohérentes (chiffre d'affaires à 10€ par exemple). Après analyse, nous constatons que les contrats sont en général présents plusieurs années dans nos bases et que toutes les valeurs de chiffre d'affaires ne sont pas manquantes (dans la plupart des cas, le chiffre d'affaires est manquant sur l'une des années uniquement). Pour corriger cette donnée, nous imputons un chiffre d'affaires moyen par calcul sur les années du contrat où la donnée est renseignée. Nous recherchons également cette donnée dans d'autres bases internes (base client et base marketing) lorsque nécessaire. Après correction de la variable chiffre d'affaires, nous avons encore 2,4% de nos lignes avec un CA vide ou erroné. Nous supprimons ces lignes, la variable chiffre d'affaires étant corrélée à notre charge de sinistres et notre tarif actuel étant établi et revu en fonction du chiffre d'affaires. Il s'agit en général de contrats anciens pour lesquels l'entreprise n'a été présente qu'une seule année en portefeuille.

Concernant les activités, tous nos contrats ont un code NAF qui est associé à 6 niveaux de regroupements internes et un groupe d'activités. La qualité des données activités est excellente puisque seuls une dizaine de contrats ont des niveaux de regroupement qui ne sont pas en phase avec leur code NAF.

Dans nos bases AXA, nous n'avons que deux types de variables avec des valeurs manquantes en dehors du CA : le montant de garantie (16% pour les contrats les plus anciens) et les variables géographiques comme la région (10%) et ou la commune (77%). En revanche, nous avons 10 à 20% de valeurs manquantes pour les variables de la base INSEE.
Ces valeurs manquantes ne peuvent pas être traitées en les associant de manière logique à une autre modalité ou par affectation d'une moyenne ou d'une médiane. Elles peuvent être traitées de la manière suivante :

- suppression de l'observation
- suppression de la variable

- regroupement avec la modalité ayant la fréquence la plus proche
- regroupement avec la modalité majoritaire, c'est-à-dire celle ayant l'effectif le plus élevé
- ou création d'une modalité « valeur manquante » à part entière lorsque l'effectif est suffisant en considérant que cette valeur manquante peut être le reflet de caractéristiques spécifiques du contrat (anciens contrats pour lesquels les données INSEE ne sont pas disponibles : note Crédit safe par exemple).

C'est cette dernière option qui est utilisée pour traiter les valeurs manquantes : création d'une modalité « 00_NR », plutôt que de supprimer l'observation ne sachant pas si les variables avec des valeurs manquantes feront partie des variables sélectionnées par le modèle. Ce choix pourra bien évidemment être remis en cause en utilisant des techniques plus sophistiquées comme par exemple une régression à partir d'autres variables correctement renseignées ou en utilisant la méthode des k plus proches voisins.
La variable Commune est supprimée, 77% des informations étant manquantes pour cette variable.

Pour réaliser notre modèle, il est primordial d'éliminer les valeurs erronées ou aberrantes concernant la sinistralité. Ainsi nous supprimons 35 sinistres dont la charge est négative et un contrat dont la charge sinistre pondérée par le chiffre d'affaires est aberrante.
Nous supprimons également toutes les lignes pour lesquelles le chiffre d'affaires est inférieur à 10000€ (soit 1,2% de nos contrats représentant 0,13% de notre charge sinistres), ces entreprises étant du ressort du périmètre Particuliers et Professionnels et non du périmètre Entreprises.

Les modifications effectuées sur la base initiale (suppression des lignes avec un chiffre d'affaires manquant ou inférieur à 10000€, suppression de 36 lignes avec une charge sinistre négative ou aberrante) sont mineures.

4.2.2. Discrétisation avec recodage en tranches

Plusieurs variables ont un nombre de modalités importantes (NAF, nombre de sinistres antérieurs par exemple) ou sont continues (chiffre d'affaires de l'entreprise). Elles doivent donc faire l'objet d'un regroupement. Les modalités sont soit regroupées en fonction de leur sinistralité pondérée par le chiffre d'affaires afin d'avoir des classes homogènes soit regroupées lorsqu'elles sont contigües (nombre de sinistres antérieurs), en vérifiant dans les deux cas que chaque classe de regroupement est composée d'une volumétrie minimale. Pour vérifier la validité des regroupements, nous effectuons une analyse de la corrélation de la variable avant et après regroupement avec la variable explicative.

Pour discrétiser la variable continue chiffre d'affaires, nous étudions les quantiles (nombre de contrats, pourcentage de CA cumulé et pourcentage de charge sinistres cumulée). Le CA est découpé en tranches en veillant à obtenir un nombre de contrats, un chiffre d'affaires cumulé et un total de charges sinistres cumulé suffisant dans chaque classe et des classes relativement équilibrées entre le nombre de contrats, le CA cumulé et la charge cumulée pour ne pas biaiser le modèle.

Lors de l'étape de discrétisation, nous supprimons certaines variables qui ont une modalité regroupant plus de 95% de l'effectif (après vérification de leur corrélation à la variable cible) car le poids est concentré sur une seule modalité (la variable géographique métropole concentre 90% de l'exposition sur les modalités vides et métropole, alors que les modalités DOM et Monaco ne représentent que 0,002% des contrats).

4.2.3. Tests de cohérence

Lors de l'étude de la qualité de notre base, nous effectuons également des tests de cohérence et des comparaisons de volumes avec les reporting existants.

Après étude des données disponibles, nous avons construit notre base de données (années 2011 à 2018 avec 5 années d'antériorité de sinistres) à partir de variables contrats, INSEE, géographiques et sinistres en conservant une charge de sinistralité en vision N+3 inflatée (As If 2021).

L'analyse de chaque variable et la vérification de la qualité de nos données ont été coûteuses en temps, mais c'est un point essentiel avant toute modélisation pour mettre en exergue les incohérences et proposer le cas échéant des corrections et retraitements.

Chapitre 3

MODELISATION DU RISQUE

Le deuxième chapitre s'est attaché à la description de la base de données pour la modélisation. Notre tarif sera basé sur la charge de sinistralité en vision N+3 inflatée (As If 2021). Nous avons également calculé notre coefficient de proportionnalité qui nous permettra de passer de la prime pure calculée sur la charge en vision N+3 à la prime pure finale.
Dans cette partie nous nous intéressons au modèle utilisé, aux étapes préliminaires à la modélisation (à savoir choix des variables explicatives et liens les reliant), ainsi qu'aux indicateurs qui nous permettront de comparer les différents modèles entre eux.

1. LE CHOIX DU MODELE

Pour la tarification, il est possible d'utiliser une approche prime pure directement, c'est-à-dire en modélisant la valeur de la charge sinistres observée sur le portefeuille ou une approche Fréquence-Coût Moyen. La RC Entreprises étant une branche d'intensité, c'est la première méthode (prime pure) qui sera retenue. La prime pure sera donc modélisée directement en ne considérant que le coût des sinistres. De plus, pour répondre au besoin des souscripteurs sur l'un de nos deux outils qui permet de tarifer les entreprises qui ont les chiffres d'affaires les plus élevés, notre tarif devra se présenter sous la forme d'un taux de prime pure à appliquer au chiffre d'affaires. Cette spécificité nous conforte donc dans la modélisation de la charge.

1.1. LE MODELE LINEAIRE GENERALISE

Le Modèle Linéaire Généralisé (GLM) est la principale méthode de tarification adoptée en assurance non-vie. A travers une fonction de lien, le GLM permet de modéliser la relation entre la variable réponse Y (charge de sinistres vision N+3 inflatée) et les variables explicatives X_i (chiffre d'affaires, secteur d'activité, nombre de sinistres antérieurs, …).

Le GLM se distingue des modèles linéaires classiques à travers la prise en compte partielle des effets non linéaires grâce au choix d'une fonction de lien.
Dans une régression linéaire classique, on suppose que l'espérance d'une variable aléatoire Y à prédire est liée par une relation linéaire aux variables explicatives X_i :

$$\mathbb{E}[Y|X] = \beta_0 + \sum_i \beta_i X_i$$

Avec pour une observation donnée y,

$$y = \beta_0 + \sum_i \beta_i x_i + \varepsilon \quad \text{où} \quad \varepsilon \sim N(0, \sigma).$$

Cependant, l'utilisation de la régression linéaire classique est très limitée en pratique car elle suppose que :

- L'espérance de la variable Y à prédire est une simple fonction linéaire des variables
- ε le terme d'erreur suit une loi normale de variance fixe.

Le GLM permet de dépasser ces limites de la régression linéaire classique en supportant de modéliser une relation non-linéaire entre la variable à expliquer Y (plus précisément son espérance) et les variables explicatives X_i.Le caractère normal de la variable à expliquer n'est plus imposé ni la normalité des distributions des résidus. Seule l'appartenance à une famille exponentielle est nécessaire.

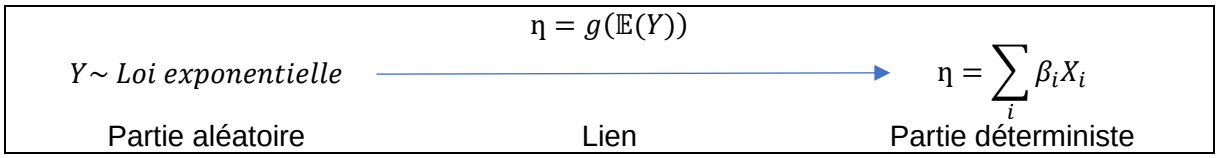
Nous avons un échantillon d'observations postulées indépendantes, (Y, X), où Y est notre variable à expliquer et X_1, X_2, …, X_p nos p variables expliquatives qu'on suppose non-stochastiques. Nous cherchons à construire un modèle pour l'espérance de la variable Y à l'aide d'une combinaison des variables X_1, X_2, …, X_p.

Un GLM est composé de trois éléments :

- une composante aléatoire : une loi de probabilité pour la variable Y à expliquer qui appartienne à la famille de lois exponentielles de dispersion
- une composante déterministe : le prédicteur linéaire $\eta = X\beta$, où X est une matrice $n * p$ avec les colonnes X_1,X_2,…,X_p qui représentent chacune des variables explicatives pour les n observations et β un vecteur de p paramètres ou coefficients, $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$
- une fonction de lien g différentiable et monotone, qui lie le prédicteur linéaire η à l'espérance de Y :

$$g(\mathbb{E}(Y)) = X\beta = \eta$$
Cela revient à écrire $g(\mathbb{E}(Y)) = \sum_i \beta_i X_i$

La fonction de lien permet de relier la composante aléatoire à la composante déterministe comme le montre le schéma ci-dessous.



La fonction de lien est la fonction qui permet de lier les variables explicatives X à la prédiction de l'espérance de Y. Elle permet donc d'expliquer l'évolution de l'espérance de la variable à prédire en fonction des variables explicatives.

Il existe de nombreuses fonctions de lien, dont les plus utilisées pour les GLM sont :

- la fonction de lien identité $g(X) = X$ qui est utilisée dans le modèle linéaire classique pour obtenir un modèle additif
- et la fonction de lien logarithmique $g(X) = \ln(X)$ qui est utilisée pour obtenir un modèle multiplicatif. En effet,

$$g(\mathbb{E}(Y)) = ln(\mathbb{E}(Y)) = \sum_i \beta_i X_i$$
Soit $e^{ln(\mathbb{E}(Y))} = \mathbb{E}(Y) = e^{\sum_i \beta_i X_i} = \prod_i e^{\beta_i X_i}$

Nous souhaitons obtenir une structure tarifaire multiplicative qui est la forme la plus adaptée aux données en assurance non-vie, nous utiliserons donc une fonction logarithmique.

Le GLM a pour avantage sa facilité d'interprétation, ainsi que sa capacité à mettre en évidence l'impact des variables explicatives sur la variable à expliquer. L'équation tarifaire issue d'un GLM est facile à implémenter. Le GLM est un modèle robuste, moins susceptible de faire du surapprentissage

que d'autres techniques de modélisation. Il a pour inconvénient la tendance à sous-apprendre, c'est un modèle paramétrique et il est nécessaire de spécifier les interactions entre les variables avant la modélisation.

1.2. LA FAMILLE EXPONENTIELLE ET LA LOI DE TWEEDIE

1.2.1. La famille de lois exponentielles

Pour la composante aléatoire du GLM, nous devons utiliser une loi de probabilité pour la variable Y à expliquer qui appartienne à la famille de lois exponentielles. La densité d'une loi de probabilité de la famille des lois exponentielles s'exprime de la façon suivante :

$$f_Y(y,\theta,\boldsymbol{\phi}) = exp\left\{\frac{y\theta - v(\theta)}{u(\boldsymbol{\phi})} + w(y,\boldsymbol{\phi})\right\}$$

- θ est appelé le paramètre canonique ou de la moyenne et $\boldsymbol{\phi}$ le paramètre de dispersion. $\boldsymbol{\phi}$ sera souvent considéré comme un paramètre de nuisance.
- $u(\boldsymbol{\phi})$ est une fonction définie sur $\mathbb{R}$ non nulle.
- $v(\theta)$ est une fonction définie sur $\mathbb{R}$ deux fois dérivable.
- $w(y,\boldsymbol{\phi})$ est une fonction définie sur $\mathbb{R}^2$ et qui ne dépend pas de θ.

L'espérance et la variance s'écrivent de la manière suivante :

- $\mathbb{E}(Y) = v'^{(\theta)}$
- $V(Y) = v''^{(\theta)} * u(\boldsymbol{\phi})$

La famille exponentielle contient la plupart des lois usuelles telles que la loi de Bernouilli ou la loi de Poisson pour les variables discrètes et la loi Normale, la loi Exponentielle, la loi Gamma, la loi Log-normale ou la loi de Tweedie pour les variables continues. La distribution de Tweedie peut être vue comme une distribution Poisson composée.

Le tableau ci-dessous résume quelques exemples de lois de la famille exponentielle avec la valeur des paramètres permettant d'écrire la densité sous la forme :

$$f_Y(y,\theta,\boldsymbol{\phi}) = exp\left\{\frac{y\theta - v(\theta)}{u(\boldsymbol{\phi})} + w(y,\boldsymbol{\phi})\right\}$$

Loi	θ paramètre canonique	$\boldsymbol{\phi}$ paramètre de dispersion	Fonction $u(\boldsymbol{\phi})$	Fonction $v(\theta)$	Fonction $w(y,\boldsymbol{\phi})$
Bernouilli $\mathcal{B}(p)$	$ln\left(\frac{p}{1-p}\right)$	1	1	$ln(1+e^{\theta})$	0
Normale $\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	σ^2	$\frac{\mu^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)\right)$
Poisson $\mathcal{P}(\lambda)$	$ln(\lambda)$	1	1	λ	$-ln(\mathrm{y!})$
Gamma $\mathcal{G}(p,\lambda)$	$-\frac{1}{p}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$	$-ln\left(\frac{1}{p}\right)$	$(\lambda-1)ln(\mathrm{y}) - ln(\boldsymbol{\Gamma}(\lambda))$

Le tableau ci-dessous résume quelques exemples de fonctions de liens canoniques des lois de la famille exponentielle.

Loi	fonction de lien
Bernouilli $\mathcal{B}(p)$	Lien logit : $g(\mu) = logit(\mu)$
Normale $\mathcal{N}(\mu, \sigma^2)$	Lien identité : $g(\mu) = \mu$
Poisson $\mathcal{P}(\lambda)$	Lien log : $g(\mu) = ln(\mu)$
Gamma $\mathcal{G}(p, \lambda)$	Lien inverse : $g(\mu) = \frac{1}{\mu}$

1.2.2. La loi de Tweedie

La distribution de Tweedie (ou distribution Poisson composée) permet d'estimer directement la prime pure d'un contrat d'assurance sans utiliser la décomposition classique fréquence * coût moyen. En effet, cette distribution présente la particularité d'avoir une masse de distribution en 0, autrement dit, une valeur positive en 0 et une densité continue sur]0;1[. Cette masse de distribution en 0 permet de prendre en compte l'ensemble des assurés qu'ils aient un sinistre ou non. En RC Entreprises, un grand nombre de contrats sont en effet non sinistrés et ont par conséquent, une charge sinistre nulle (95% de notre base après suppression des sinistres sans suite). Les contrats sinistrés ont alors une charge continue et positive, ce qui justifie le choix de cette loi.

La distribution de Tweedie appartient à la famille de lois exponentielles. Sa densité s'exprime sous la forme suivante :

$$f_Y(y, \theta, \boldsymbol{\phi}) = exp\left\{\frac{y\theta - v(\theta)}{u(\boldsymbol{\phi})} + w(y, \boldsymbol{\phi})\right\}$$

avec :

- $\theta = \begin{cases} \frac{\mathbb{E}(Y)^{1-p}}{1-p} \; si \; p \neq 1 \\ ln(\mathbb{E}(Y)) \; si \; p = 1 \end{cases}$
- $u(\boldsymbol{\phi}) = \boldsymbol{\phi}$
- $v(\theta) = \begin{cases} \frac{\mathbb{E}(Y)^{2-p}}{2-p} \; si \; p \neq 2 \\ ln(\mathbb{E}(Y)) \; si \; p = 2 \end{cases}$

La loi de Tweedie se caractérise également par la relation particulière entre sa variance et son espérance :

$V(Y) = \boldsymbol{\phi}[\mathbb{E}(Y)]^p$

$\boldsymbol{\phi}$ est le paramètre de dispersion et p réel est le paramètre de puissance (ou paramètre de forme de la distribution).

En fonction de la valeur de ce paramètre de puissance p, les lois usuelles suivantes peuvent être retrouvées car ce sont des cas particuliers de la loi de Tweedie :

- La loi normale lorsque p=0
  En effet, si $Y \sim \mathcal{N}(\mu, \sigma^2)$, alors
  $$\boldsymbol{\phi}[\mathbb{E}(Y)]^0 = \boldsymbol{\phi} = \sigma^2 = V(Y)$$
- La loi de Poisson lorsque p=1
  En effet, si $Y \sim \mathcal{P}(\lambda)$, alors
  $$\boldsymbol{\phi}[\mathbb{E}(Y)]^1 = \mathbb{E}(Y) = \lambda = V(Y) \text{ avec } \boldsymbol{\phi} = \mathbf{1}$$
- La loi Gamma lorsque p=2
  En effet, si $Y \sim \mathcal{G}(p, \lambda)$, alors
  $$\boldsymbol{\phi}[\mathbb{E}(Y)]^2 = \frac{1}{p} * \frac{p^2}{\lambda^2} = \frac{p}{\lambda^2} = V(Y)$$

Si p=3, on retrouve une loi Gaussienne Inverse. Pour les valeurs de p supérieures à 3, les distributions sont définies mais plus difficiles à estimer car elles ne peuvent pas être écrites sous une forme finie.

Le paramètre p qui nous intéresse pour modéliser notre charge de sinistre est le cas où p est strictement compris entre 1 et 2. Comme vu précédemment, si p=1, nous avons une loi de Poisson et si p=2, nous avons une loi Gamma. Lorsque 1<p<2, la loi de Tweedie est une loi de Poisson composée avec des sauts suivants une loi Gamma. Cette loi est positive, très asymétrique à droite avec une masse de distribution en 0, contrairement à la loi Gamma, ce qui permet de gérer notre nombre important de contrats avec une charge de sinistres nulle. La loi de Tweedie permet également de modéliser en même temps la fréquence et la sévérité.

La fonction de lien utilisée ici sera la fonction log, qui est une fonction de lien adéquate pour la loi de Tweedie. Nous utiliserons une valeur de p de 1,5.

Notre modélisation de la charge de sinistres s'effectuera :

- A l'aide d'un Modèle Linéaire Généralisé
- Avec une loi de Tweedie de la famille exponentielle, adaptée avec une masse de distribution en 0 pour gérer le nombre important de contrats avec une charge de sinistres nulle
- Et une fonction de lien log.

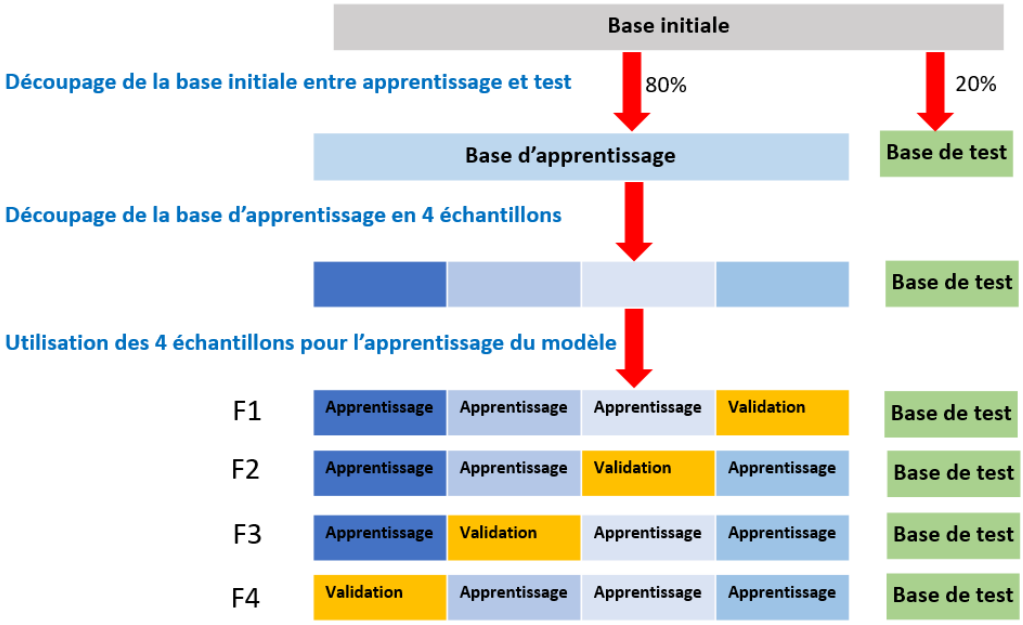
1.3. LA VALIDATION CROISEE

Pour entraîner nos modèles, nous fractionnons notre base de données en deux bases avec des données tirées aléatoirement sans remise :

- Une base d'apprentissage, sur laquelle le modèle sera construit, équivalente à 80% de la base initiale
- Une base de test avec les 20% restants, indépendante de la base d'apprentissage, qui servira à vérifier que le modèle s'ajuste correctement aux données d'une base qui n'a pas servi à déterminer les variables explicatives ni les coefficients.

Pour vérifier la qualité des modélisations sur la base d'apprentissage avant de vérifier le modèle final sur la base de test, nous utilisons la validation croisée. Pour cela, nous partageons la base d'apprentissage en plusieurs jeux de données équilibrés afin de vérifier les différentes modélisations. Cette technique permet d'éviter l'overfitting, c'est-à-dire d'avoir un modèle qui représenterait mal les données à partir duquel il a été entraîné. La validation croisée nous assure qu'il n'existe pas de cases tarifaires trop petites qui auraient tendance à biaiser les résultats en forçant le modèle à s'ajuster.

Le principe de la validation croisée est de découper la base de données en *k* échantillons (k-folds) ayant la même taille. L'un de ces sous-ensembles est considéré comme la base de validation, les *k* − 1 sous-ensembles restants sont utilisés comme base d'apprentissage. Cette opération est répétée en prenant chacun des *k* échantillons comme base de validation. Ainsi, chacun des k-modèles a un score de performance. Le score de performance du modèle final est calculé en prenant la moyenne de la performance de k-modèles.



Principe de la validation croisée

Nous découpons notre base d'apprentissage en 4 échantillons (folds). Ainsi, nous obtiendrons les métriques de mesure de notre modèle (GINI, RMSE) sur les 4 échantillons, le score de performance final sur la base d'apprentissage et sur la base de test étant calculé en prenant la moyenne de la performance des 4 échantillons (k-folds).

2. LA SELECTION DES VARIABLES

2.1. LA VARIABLE REPONSE

Comme indiqué précédemment, la contrainte imposée par l'un de nos outils étant une modélisation en taux de prime pure et la RC étant une branche d'intensité, notre variable Y à expliquer sera la charge de sinistres attritionnels et graves hors sans suite en vision N+3 par rapport à la survenance, charge inflatée (As If 2021).

Nous choisissons de spécifier une variable temps qui sera l'année de présence du contrat (2011 à 2018) et servira à effectuer des contrôles de cohérence temporelle.

Nous choisissons une variable d'exposition pour pondérer les observations. Notre variable d'exposition est égale au produit de la durée de présence du contrat dans l'année (durée comprise entre 0 strictement et 1) et du chiffre d'affaires de l'entreprise. En effet, la charge de sinistre annuelle d'une entreprise ayant un chiffre d'affaires de 20 000 euros est difficilement comparable à celle d'une entreprise dont le chiffre d'affaires est de 15 millions d'euros. Intuitivement, l'entreprise avec un chiffre d'affaires de 15 millions d'euros a un nombre plus important d'employés et de clients que celle ayant un chiffre d'affaires de 20 000 euros.
Notre variable réponse sera normalisée en fonction de cette exposition.

2.2. LES VARIABLES EXPLICATIVES

La loi et la fonction de lien étant choisies, la variable réponse et l'exposition étant définies, il convient avant d'entamer la modélisation d'analyser si les variables explicatives sont significatives, c'est-à-

dire si elles expliquent notre charge de sinistre, ainsi que d'étudier les liens qui les relient entre elles au travers des corrélations qui pourraient nuire à la qualité du modèle.
En effet, la quantité d'informations récoltée est riche, mais pour autant notre objectif est de retenir uniquement les données les plus adaptées à la problématique et d'extraire les variables significativement discriminantes.

2.2.1. Etude des variables corrélées à la cible

Afin de réduire le nombre de variables à injecter dans nos modèles GLM, nous étudions les corrélations entre les variables explicatives et la variable à expliquer puis entre les variables explicatives retenues.
La majorité de nos variables sont des variables catégorielles, les variables quantitatives comme par exemple le chiffre d'affaires ou le nombre de sinistres antérieurs ayant été discrétisés en tranches.
La corrélation est mesurée à l'aide du V de Cramer. Les variables quantitatives sont assimilées à des variables qualitatives contenant un grand nombre de modalités.

Traditionnellement, pour établir s'il existe un lien entre deux variables qualitatives croisées dans un tableau de contingence, on utilise le test du Khi2 (χ^2).

Cependant, le test du Khi2 permet de savoir s'il existe une dépendance entre deux variables, mais il ne donne pas d'indication sur l'intensité de cette relation. De plus le Khi2 dépend de la taille de l'échantillon et du degré de liberté (nombre de lignes -1)*(nombre de colonnes -1).
Le V de Cramer représente une forme centrée réduite du Khi2, donc plus pertinente puisqu'il permet de comparer l'intensité du lien entre les deux variables étudiées.
Le V de Cramer se base sur le Khi2 maximum que le tableau de contingence pourrait théoriquement produire.

$$V_{Cramer} = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{\boldsymbol{n}*(\boldsymbol{min}(\boldsymbol{l},\boldsymbol{c})-\boldsymbol{1})}}$$

avec n le nombre d'individus de l'échantillon, et l et c le nombre de lignes et de colonnes du tableau de contingence.

Plus le V de Cramer est proche de 0, moins les variables étudiées sont dépendantes. Au contraire, plus il est proche de 1, plus la liaison entre les deux variables étudiées est forte.

Nous n'avons aucune variable explicative avec un V de Cramer inférieur à 0,1 (absence de corrélation ou corrélation très faible). La variable la moins corrélée à notre variable d'intérêt (V de Cramer de 0,19) est la variable métropole qui comporte les modalités (métropole, DOM, Monaco et Non Renseigné). Les variables les plus corrélées à notre charge de sinistres sont la tranche de nombre de sinistres antérieurs normalisé par le chiffre d'affaires, les différents niveaux de regroupement de NAF, la tranche d'effectif de l'entreprise, la tranche de montant de garantie assuré, ainsi que les garanties Export USA, faute inexcusable, pollution, Région et forme juridique (INSEE).

2.2.2. Etude des corrélations entre les variables explicatives sélectionnées

L'étude des corrélations entre les variables révèle des corrélations importantes entre certaines variables. Afin de ne pas biaiser les modèles :

- si les variables sont très corrélées et discriminantes, on crée une variable croisée
- si les variables sont moyennement ou peu corrélées avec la variable cible, on conserve la variable la plus discriminante et on supprime les autres (ex : la variable tranche

d'effectif de l'établissement est corrélée avec le chiffre d'affaires, les garanties faute inexcusable et pollution sont extrêmement corrélées). A noter que tous nos niveaux de regroupement de NAF sont corrélés mais aucun niveau de regroupement ne se détache des autres, nous conservons donc tous les niveaux et les testerons dans le GLM.

2.2.3. Sélection des variables corrélées à la cible

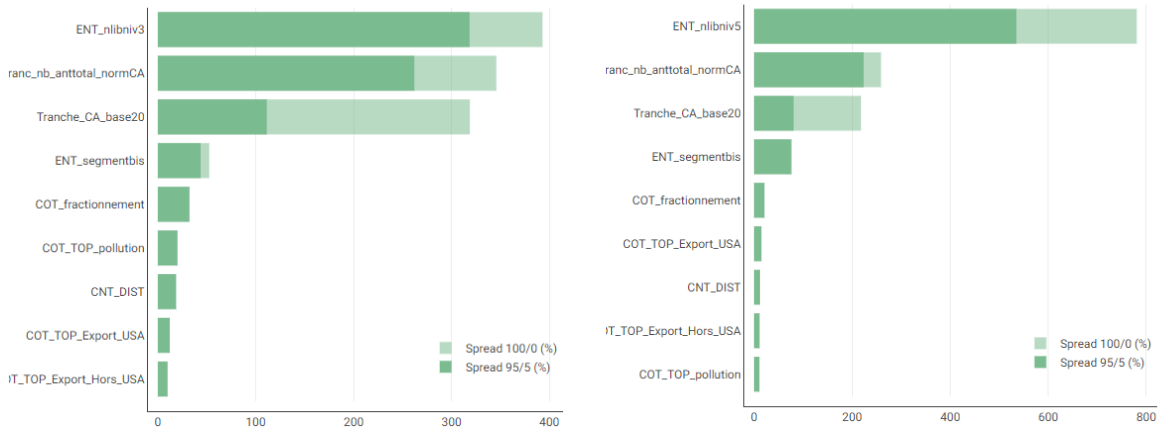
Pour construire notre modèle, nous devons sélectionner les variables qui ont de l'effet sur notre variable Y charge de sinistres, c'est-à-dire les variables X_i dont l'indice β_i est différent de zéro. Pour cela, nous utilisons une méthode de sélection pas à pas.

La méthode descendante (ou Backward) consiste à étudier un premier modèle qui prend en compte toutes les variables explicatives. Si toutes les variables sont significatives (au sens du test de Student), on garde le modèle. Sinon, on cherche la variable la moins significative (p-value la plus grande) et on la supprime du modèle. La procédure s'arrête lorsque la suppression d'une nouvelle variable détériore le modèle.

La méthode ascendante (ou Forward) consiste à étudier tous les modèles à une variable et garder celui qui est le plus significatif (p-value la plus faible du test de Student). Si aucun modèle à une variable n'est significatif, la sélection s'arrête. Sinon, on introduit une nouvelle variable explicative. La procédure s'arrête lorsque l'ajout d'une variable supplémentaire n'améliore plus significativement le modèle.

La méthode progressive (ou Stepwise) est une combinaison des méthodes Forward et Backward, qui teste à chaque étape les variables à inclure ou exclure. A chaque étape, on réexamine toutes les variables introduites précédemment dans le modèle. En effet, une variable considérée comme la plus significative à une étape peut à une étape suivante devenir moins significative en raison de sa corrélation avec d'autres variables introduites ultérieurement dans le modèle.

Les sélections de variables suivantes ressortent :



Les 3 premières variables sélectionnées et ayant un impact important sur notre charge de sinistres sont :

- l'activité de l'entreprise (le niveau 3 et le niveau 5 sont un regroupement de codes NAF et comportent respectivement 17 et 122 modalités)
- le nombre de sinistres antérieurs (variable qui a été discrétisée en tranches et normalisée en la divisant par le chiffre d'affaires car plus corrélée à notre charge si elle est normalisée)

- la tranche de chiffre d'affaires de l'entreprise.

A noter que la variable NAF, qui est utilisée lors de la souscription, n'a pas pu être sélectionnée car son introduction ne permet pas au GLM de converger. Les deux niveaux de regroupement de NAF (niveau 3 et niveau 5) seront donc utilisés et nous devrons voir comment affiner les coefficients du GLM obtenus pour ces deux variables, afin d'appliquer des coefficients différents aux codes NAF regroupés dans un même niveau.

On trouve ensuite un groupe de variables sélectionnées ayant un impact moins important sur notre charge de sinistres :

- la sensibilité du NAF, indicateur valorisé par la direction technique (standard, sensible, réservé, exclu, construction)
- le fractionnement de la prime (mensuel, trimestriel, semestriel ou annuel)
- certaines garanties souscrites (pollution, export vers les USA, export hors USA)
- et le réseau de distribution.

Les variables susceptibles d'expliquer notre charge de sinistres pour les entreprises industrielles et les commerces sont les suivantes :

- Activité de l'entreprise (via les regroupements de codes NAF Niveau 3 et 5 et le segment qui indique la sensibilité du code NAF évaluée par la Direction Technique)
- Le chiffre d'affaires de l'entreprise
- Le nombre de sinistres antérieurs normalisé (ie divisé par le chiffre d'affaires)
- Des caractéristiques de souscription (fractionnement de la prime et réseau de distribution)
- Ainsi que les garanties souscrites (pollution, export USA et export hors USA).

3. LA COMPARAISON DES MODELES ET LES RESULTATS

Le modèle GLM et la sélection des variables ayant été présentés, il convient de s'intéresser aux résultats de modélisation et à la qualité de l'ajustement du modèle.

Nous avons modélisé notre charge de sinistres attritionnelle et grave en utilisant plusieurs GLM issus des sélections de variables précédentes (plusieurs sélections de variables étant possibles avec des coefficients très proches). A noter que nous avons également testé ces GLM sur la charge attritionnelle uniquement en lissant ensuite la charge de graves sur l'ensemble des cases tarifaires obtenues. Cependant, ces modèles étaient moins performants pour prédire la charge.

La première étape après modélisation est de comparer les qualités d'ajustement des différents GLM. La comparaison des modèles se fera :

- Sur la base d'indicateurs statistiques : métriques d'indicateurs d'erreurs, notamment le coefficient de Gini et la racine de l'erreur quadratique moyenne (RMSE),
- Via l'analyse des résidus.
- Sur la capacité à prédire le bon montant de charge sinistres.

Un modèle sera plus pertinent qu'un autre si son taux d'erreur est minimal, c'est-à-dire si les métriques sont les plus faibles sur la base de test et si la charge prédite sur la base de test est la plus proche de celle observée.

3.1. LA MESURE DE PERTINENCE DES MODELES

L'erreur d'estimation d'un modèle correspond à l'écart entre la valeur observée et la valeur prédite. Nous calculons les métriques de mesure de l'erreur de prédiction suivantes : RMSE et coefficient de Gini.

RMSE (Root Mean Squared Error)

L'erreur quadratique moyenne (MSE : Mean Squared Error) est une mesure d'erreur classique correspondant à la somme des carrés des écarts entre les valeurs prédites $\hat{y}_i$ et les valeurs à prédire (valeurs observées) y_i rapportée au nombre n des observations.
Pour chaque individu i, notons y_i la valeur observée et $\hat{y}_i$ la valeur prédite par le modèle. L'erreur quadratique moyenne est alors calculée de la manière suivante :

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

La MSE présente l'avantage de pénaliser plus fortement les fortes erreurs à travers le carré que d'autres mesures de performance.

On définit également usuellement la RMSE comme la racine carrée de la MSE :

$$RMSE = \sqrt{MSE}$$

La RMSE a l'avantage de s'exprimer dans la même unité que la variable à expliquer.

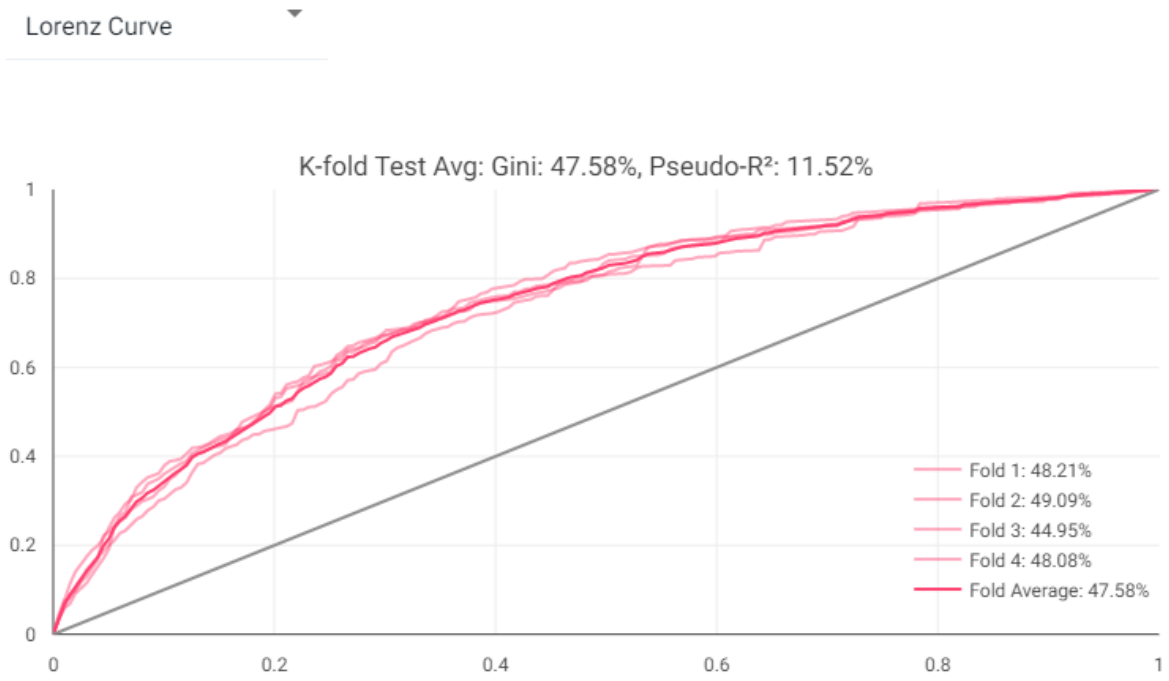
Coefficient de Gini et courbe de Lorenz

Le coefficient de Gini est un indicateur de dispersion d'une distribution dans une population. Il ne permet pas de mesurer la qualité d'ajustement du modèle aux données, mais il permet d'évaluer les qualités de segmentation d'un modèle c'est-à-dire la qualité du classement des individus.

Dans un modèle de fréquence, le coefficient de Gini permet de vérifier que les assurés ayant le moins de sinistre sont modélisés par un risque plus faible que les assurés ayant plus de sinistres.
Dans un modèle de coût, il permet de vérifier que les sinistres les plus coûteux sont modélisés par une charge plus élevée que les sinistres moins onéreux.

Le calcul du coefficient de Gini se fait à partir de la courbe de Lorenz. L'objectif est de vérifier que notre modélisation segmentée est meilleure que la modélisation aléatoire ou uniforme, représentée sur le graphique de la courbe de Lorenz par la bissectrice. Dans le cadre de notre étude de la charge de sinistres, il est possible de représenter la courbe de Lorentz en plaçant sur l'axe des abscisses la part cumulée de la population triée par prédiction de risque et en ordonnée, la part cumulée de la grandeur détenue (ici notre charge sinistres). Les observations sont triées de la prévision la plus élevée (risque élevé) à la plus basse (risque faible). Une prédiction aléatoire du risque suivrait la diagonale en noir (bissectrice où x% de la population représentent x% de la charge de sinistres).

La courbe de Lorenz (en rouge sur le graphique) représente la part cumulée des contrats qui créent les risques les plus importants (en charge de sinistres). Ainsi 20% de la population des entreprises assurées créent 50% de notre charge de sinistres.



Exemple de courbe de Lorenz obtenue pour le modèle ayant les regroupements d'activités de niveau 3

Le coefficient de Gini est lié à la courbe de Lorentz. L'AUC (Area Under Curve) mesure l'aire entre la courbe de Lorentz moyenne (en rouge sur le graphique) et l'axe des abscisses. Le coefficient de Gini est calculé grâce à l'AUC via la formule suivante :
$Coefficient\ de\ Gini\ = 2 * AUC - 1.$

Le coefficient varie de 0 (ou 0%) à 1 (ou 100%). Par exemple, la segmentation aléatoire (bissectrice en noir sur le graphique) correspond à une AUC de 1 et donc un coefficient de Gini de 0.
Plus le coefficient de Gini est élevé, plus la puissance de classement du modèle est élevée, donc le modèle segmente plus la population. Ainsi plus le coefficient de Gini est élevé, plus le modèle permet de différencier les assurés en fonction de leur risque. Il permet en complément de la RMSE d'éviter de choisir un modèle qui effectuerait un mauvais classement des assurés en fonction de leur risque et évite ainsi l'antisélection.

Lors de l'étape de choix des variables, nous avons obtenu 2 listes de variables susceptibles d'expliquer notre charge de sinistres avec une variable différente correspondant au niveau de regroupement des activités (NAF) :

- La variable ENT_Nivlib3 qui regroupe les codes NAF en 17 classes
- Et la variable ENT_Nivlib5 qui regroupe les codes NAF en 122 classes

Nous allons donc tester deux modèles GLM avec une loi de distribution de Tweedie.
La variable à expliquer sera la charge de sinistres attritionnels et graves hors sans suite en vision N+3 par rapport à la survenance inflatée (As If 2021).

Les 9 variables explicatives seront :

- L'activité de l'entreprise (regroupement Niveau 3 ou Niveau 5)
- Le nombre de sinistres antérieurs divisé par le chiffre d'affaires de l'entreprise
- La tranche de chiffre d'affaires de l'entreprise
- La sensibilité du NAF (standard, sensible, réservé, exclu, construction)

- Le fractionnement de la prime (mensuel, trimestriel, semestriel ou annuel)
- Le réseau de distribution
- Les garanties :
    - Pollution
    - Export vers les USA
    - Export hors USA.

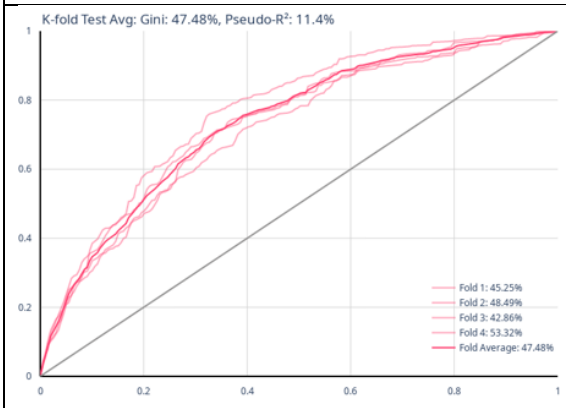
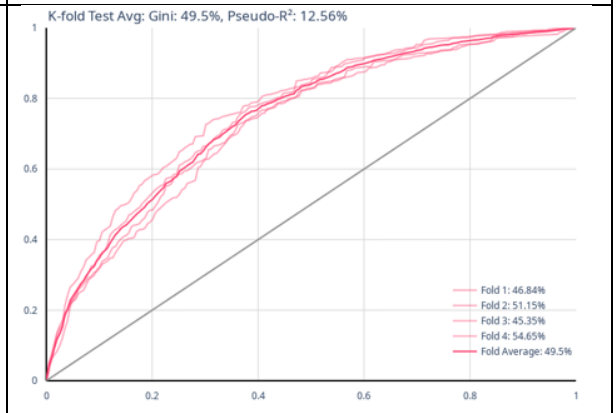
Nous avons également testé un modèle pour modéliser uniquement la charge attritionnelle. La RMSE était meilleure, le GINI légèrement moins bon. La charge de sinistres graves a ensuite été ajoutée de manière uniforme aux prédictions (% de charge grave) mais ce modèle a été écarté car les prédictions finales étaient plus éloignées de l'observé que les prédictions obtenues en modélisant directement la charge attritionnelle et grave.

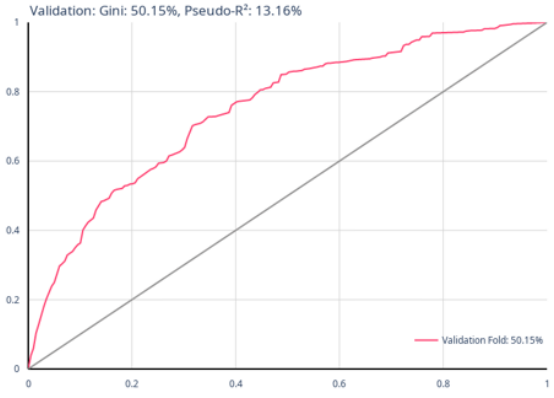
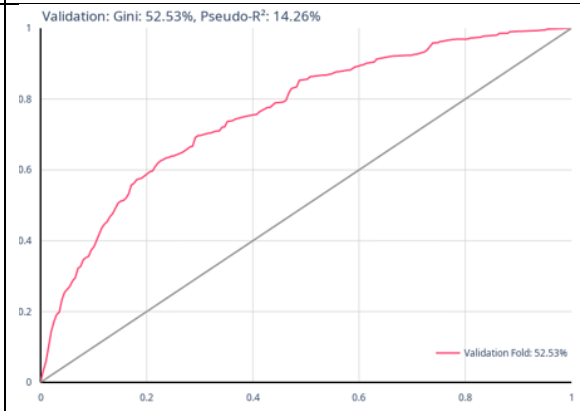
Les sinistres graves sont en effet plus concentrés sur certaines activités. Ainsi, l'industrie lourde et la fabrication d'équipements industriels concentrent 46% de la charge de sinistres graves et représentent 20% du chiffre d'affaires des entreprises. A contrario, les activités hôtels et restaurants n'ont enregistré aucun sinistre grave sur la période observée de 10 ans.

Les résultats de nos deux modèles GLM avec les neuf variables sélectionnées sont présentés dans le tableau ci-dessous :

Modèle	Modèle avec regroupement Niveau3		Modèle avec regroupement Niveau5	
Indicateurs	Base d'apprentissage	Base de test	Base d'apprentissage	Base de test
Coefficient de Gini	47.48%	50.15%	49.50%	52.53%
RMSE	0.01	0.009	0.01	0.009

Le modèle avec le regroupement des activités au niveau 5 a un coefficient de Gini légèrement meilleur sur la base de test. Les RMSE des deux modèles sont comparables.

Modèle avec regroupement Niveau3	Modèle avec regroupement Niveau5
Courbe de Lorentz sur base d'apprentissage	Courbe de Lorentz sur base d'apprentissage
	

Modèle avec regroupement Niveau3	Modèle avec regroupement Niveau5
Courbe de Lorentz sur base de test	Courbe de Lorentz sur base de test
	

3.2. L'ANALYSE DES RESIDUS

L'étude de la RMSE et du coefficient de Gini ne suffisent pas pour valider un modèle. Il est en effet nécessaire d'analyser les écarts entre les valeurs observées y_i et les valeurs prédites par le modèle $\hat{y}_i$ afin de vérifier si le modèle peut être amélioré.

L'analyse des résidus permet de vérifier si l'erreur est aléatoire et de détecter des valeurs aberrantes ou trop influentes qui pénaliseraient le modèle ou encore de détecter des effets non linéaires.

Les résidus r_i sont définis pour chaque individu i, comme l'écart entre la valeur observée y_i et la valeur prédite par le modèle $\hat{y}_i$:

$$r_i = y_i - \hat{y}_i$$

Les résidus peuvent être pondérés par l'écart type (résidus de Pearson). Les résidus de Pearson sont calculés en effectuant le rapport entre les résidus r_i et l'écart type estimé de $\hat{y}_i$:

$$rp_i = \frac{y_i - \hat{y}_i}{\sqrt{V(\hat{y}_i)}}$$

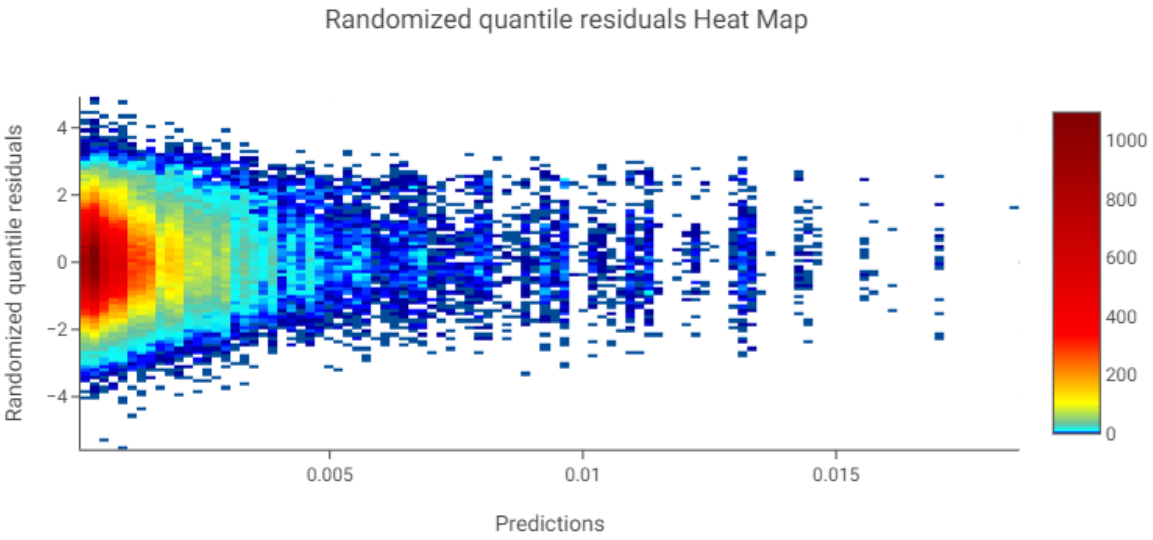
Graphiquement, plus les résidus seront centrés en 0, plus l'erreur de modélisation sera faible.

Nous analysons les résidus quantiles. Sur le graphique, ils doivent être normalement distribués en supposant que la distribution correcte est utilisée. Les résidus quantiles sont calculés en transformant la distribution en une distribution normale : un modèle bien ajusté avec une hypothèse de distribution correcte devrait présenter des résidus quantiles qui sont à peu près normalement distribués. En pratique, pour une distribution continue, ils sont calculés comme suit :

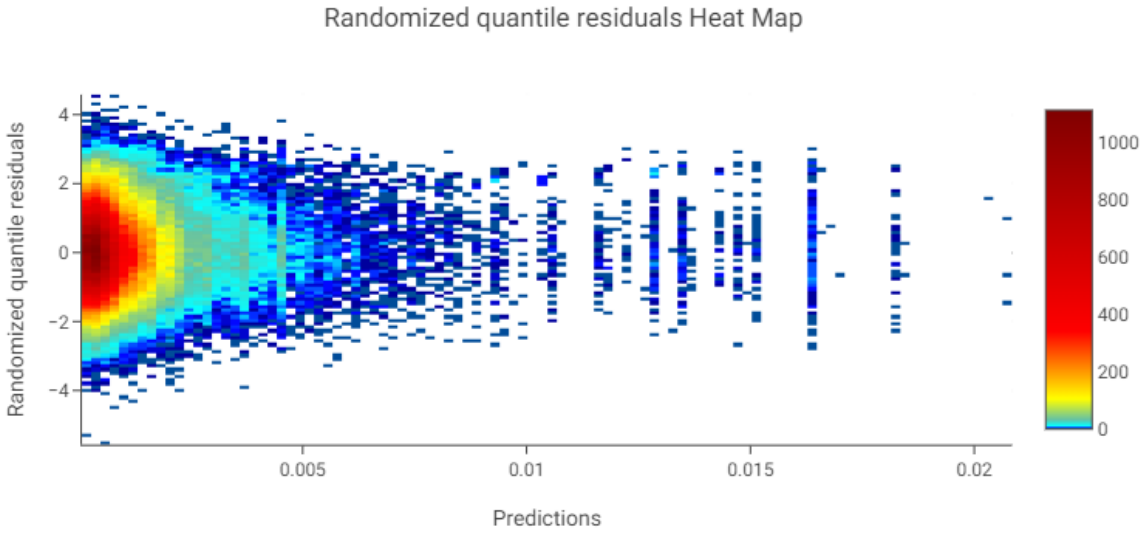
$$rquantile = \Phi^{-1}(\mathcal{F}(y\,;\hat{\mu}))$$

où $\mathcal{F}$ est la fonction de distribution cumulative de la distribution choisie (distribution de Tweedie dans notre cas) et $\boldsymbol{\Phi}$ la fonction de distribution cumulative de la loi normale centrée réduite. Pour les distributions discrètes, une randomisation est appliquée.

Les graphiques ci-dessous présentent les résidus quantiles de nos deux modèles.



Graphique des résidus quantiles pour le modèle avec regroupement de niveau 3



Graphique des résidus quantiles pour le modèle avec regroupement de niveau 5

Les résidus quantiles de nos deux modèles sont comparables. Dans les deux cas, l'histogramme des résidus quantiles montre que nos résidus sont normalement distribués et que la variabilité des résidus diminue avec la valeur de la prédiction $\hat{y}_i$. Nos modèles sont donc bien ajustés aux données.

3.3. LA PERFORMANCE DE PREDICTION DE LA CHARGE

3.3.1. Lift Curve

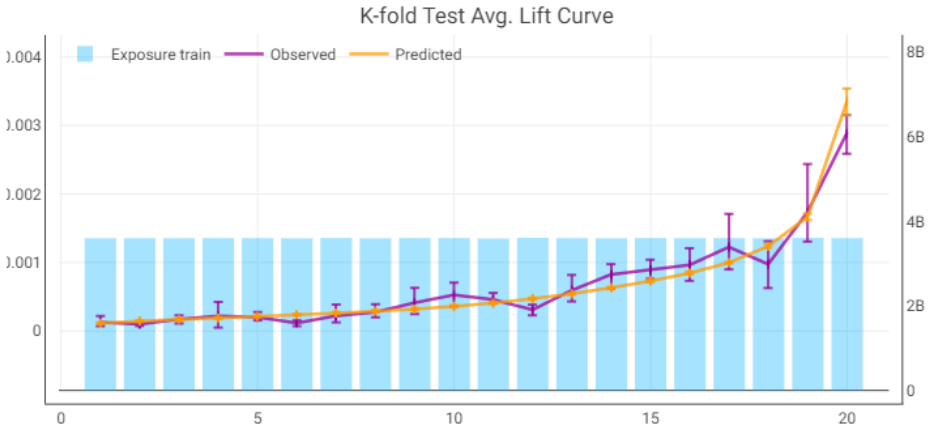
En complément du coefficient de Gini, la Lift Curve permet de visualiser l'adéquation du modèle aux données risques.

L'idée est de trier la base de données en ordonnant les prédictions, de la prédiction la plus basse à la plus élevée, puis en les répartissant en 20 classes de risque. Chaque classe contient donc 5% des prédictions. Pour chaque classe, nous calculons l'observation moyenne du risque modélisé (charge de sinistres) et nous pouvons la comparer à la prédiction. Les valeurs moyennes prédites et observées sont représentées sur la Lift Curve.

Ensuite, pour chacune des classes, un écart d'adéquation des données sur les premiers et derniers quantiles permet de savoir que nous avons sous ou sur modélisé les entreprises ayant un risque de sinistralité faible ou important. Autour des moyennes, des barres d'erreur permettent donc de visualiser la variabilité du risque afin d'évaluer la qualité des prédictions du modèle.
L'objectif de la Lift Curve est de refléter le pouvoir discriminant du modèle (différence entre les risques extrêmes prédits et observés).

L'analyse d'adéquation des données est réalisée sur l'ensemble des bases : apprentissage (train et validation) et test pour s'assurer que nous n'avons pas d'overfitting lié à une spécificité de la base d'apprentissage.

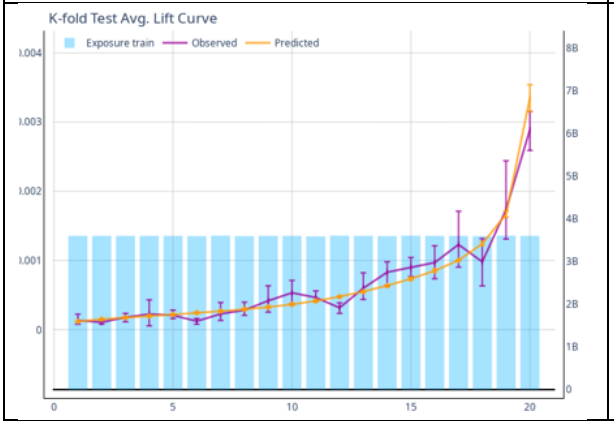
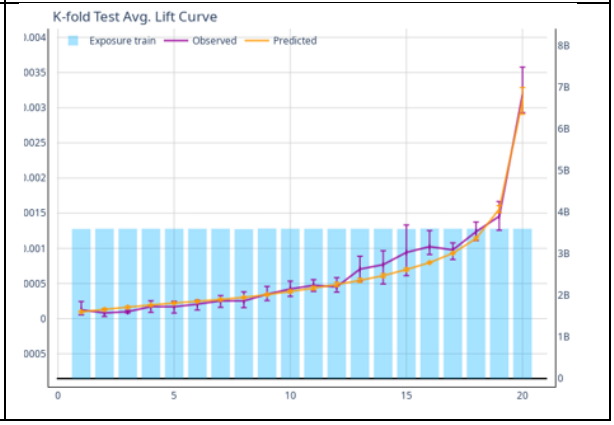
Sur le graphique représentant la Lift Curve, si les courbes moyenne observée et moyenne prédite se superposent, cela signifie que le modèle prédit bien les valeurs observées, tandis qu'une courbe plus chaotique signifie que les prédictions sont moins bien ajustées aux données.

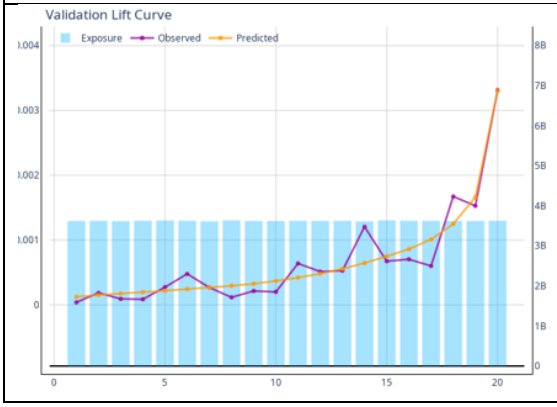
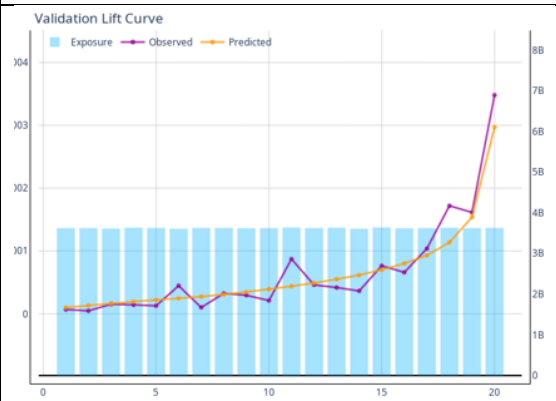


Exemple de Lift Curve sur la base de validation de notre modèle avec regroupement niveau 3

Sur le graphique, la courbe en violet représente la moyenne observée avec affichage de la variabilité du risque et la courbe en jaune représente la modélisation moyenne pour chacune des 20 classes. Dans l'exemple ci-dessus (base de validation avec le modèle incluant le niveau 3 des activités), nous constatons que les prédictions sont bien ajustées aux données puisque les moyennes prédites sont dans l'intervalle des moyennes observées à l'exception de la sixième et la dernière classe. Cependant la moyenne prédite reste très proche de la moyenne observée pour ces deux classes.

Les Lift Curve de nos deux modèles GLM avec les neuf variables sélectionnées sont présentées dans le tableau ci-dessous.

Modèle avec regroupement Niveau3	Modèle avec regroupement Niveau5
Lift Curve sur base d'apprentissage	Lift Curve sur base d'apprentissage
	

Modèle avec regroupement Niveau3	Modèle avec regroupement Niveau5
Lift Curve sur base de test	Lift Curve sur base de test
	

La lift curve de gauche (modèle niveau 3) sur la base de validation de l'échantillon d'apprentissage indique une possible surestimation pour les entreprises dans les classes 6, 12 et 20, la courbe jaune qui représente les prédictions étant au-dessus de la courbe violette qui représente les observations.
La lift curve de gauche (modèle niveau 5) indique une possible sous-estimation pour les entreprises des classes 13 et 16, la courbe jaune des prédictions étant sous celle des observations et une légère surestimation pour les classes 2 et 3.

Cependant, pour les deux modèles, malgré quelques petits écarts, la modélisation s'ajuste correctement aux données.
Pour différentier l'un de nos deux modèles, nous allons donc étudier plus finement la charge totale prédite par rapport à la charge observée.

3.3.2. Charge estimée

Les résultats d'estimation des charges de sinistres de nos deux modèles GLM avec les neuf variables sélectionnées sont présentés dans le tableau ci-dessous.

Modèle	Modèle avec regroupement Niveau 3		Modèle avec regroupement Niveau 5	
Indicateurs	Base d'apprentissage	Base de test	Base d'apprentissage	Base de test
Moyenne observée	6677E-4	6659E-4	6677E-4	6659E-4
Moyenne prédite	6644E-4	6651E-4	6326E-4	6279E-4

Le modèle avec le regroupement des activités au niveau 3 prédit une charge de sinistres en moyenne plus proche et même très proche de la charge observée sur la base de test.

Les charges observées et estimées par les deux modèles sur les bases d'apprentissage et de test sont exposées dans le tableau suivant.

Modèle	**Modèle Niveau 3**		**Modèle Niveau 5**	
Base	Base d'entraînement	Base de test	Base d'entraînement	Base de test
Charge observée	192 442 089	48 273 440	192 442 089	48 273 440
Charge prédite	191 389 858	48 210 951	182 207 397	45 515 039
Ecart (obs-pred)	1 052 231	62 489	10 234 692	2 758 402
Ecart % (obs/pred)	99,5%	99,9%	94,7%	94,3%

Une analyse plus détaillée des charges prédites selon le regroupement niveau 3 en 17 classes des codes NAF confirme que ce modèle prédit mieux pour chaque classe la charge réelle observée. Seule une classe comportant très peu d'observations (moins de 50) n'est pas correctement prédite par ce modèle. Dans le modèle avec le regroupement au niveau 5 (122 classes de codes NAF), il y a des écarts plus importants entre la charge prédite et la charge observée pour la moitié des classes environ (sous ou surestimation).

Le modèle avec le regroupement niveau 3 est donc meilleur pour prédire notre charge de sinistralité.

Nous avons testé deux modèles comportant neuf variables pour prédire notre charge de sinistres attritionnelle et grave, le modèle utilisé pour prédire uniquement la charge attritionnelle étant moins performant pour évaluer la charge attritionnelle et grave.

Les indicateurs statistiques (coefficient de Gini, courbe de Lorenz et RMSE) sont très proches pour nos deux modèles et confirment que les deux modélisations s'ajustent à nos données.

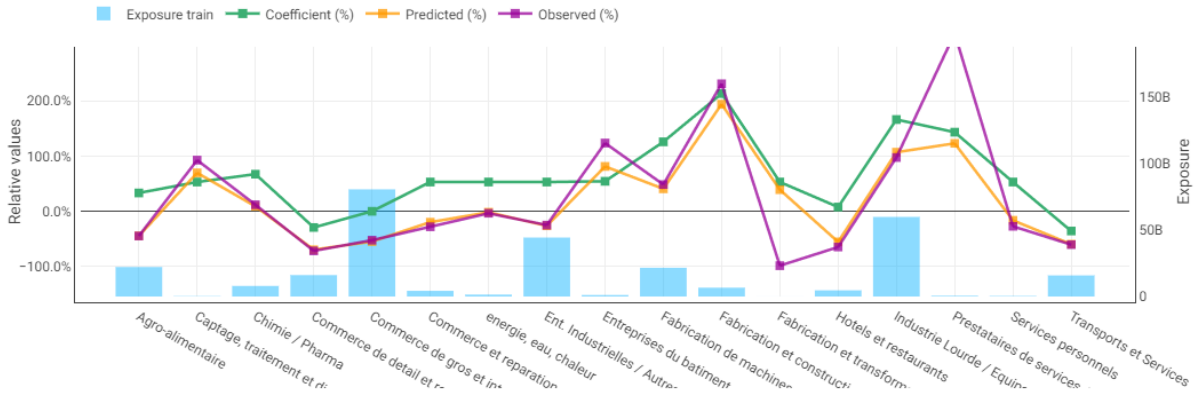
L'analyse des résidus quantiles et des Lift Curve des deux modèles confirment également que les deux modèles sont adaptés pour prédire notre charge de sinistres.

Nous choisissons donc le modèle avec la variable niveau 3 qui regroupe les codes NAF de nos entreprises en 17 classes, ce modèle étant plus précis lorsque l'on étudie plus en détail les charges prédites en fonction des charges réellement observées.

3.4. LES COEFFICIENTS DES VARIABLES

Le modèle ayant été sélectionné, nous pouvons étudier les variables utilisées pour prédire la charge de sinistre. En effet, il est important de comprendre l'apport de chacune des variables dans le modèle mais également de s'assurer qu'il n'y a pas de sur-dispersion des coefficients dans le GLM. Afin de comprendre l'influence de chaque modalité, nous présentons les valeurs relatives des coefficients de chaque variable sélectionnée dans le modèle (donnée en vert sur le graphique), la charge de sinistres observée (Observed Average en violet sur le graphique) ainsi que la charge moyenne prédite (Predicted Average en jaune sur le graphique).

<u>Activité de l'entreprise</u>
La variable « ENT_nlibniv3 » permet de segmenter le risque selon l'activité des entreprises : les codes NAF sont regroupés en 17 classes.



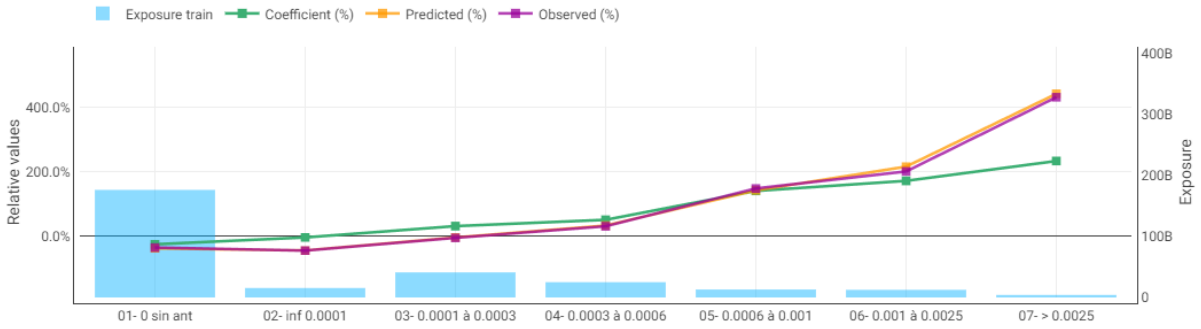
Coefficient de la variable Nivlib3 décrivant l'activité de l'entreprise (17 niveaux)

Sans surprise, les entreprises exerçant une activité commerciale (notamment les commerces de détail et les commerces de gros), les hôtels et restaurants ainsi que les entreprises de transport ont les coefficients les plus bas.
Les entreprises industrielles de fabrication ont les coefficients les plus élevés, notamment :

- les entreprises de fabrication et construction de matériels de transports (construction navale, construction de matériel ferroviaire, aéronautique, véhicules automobiles, motos)
- les entreprises de fabrication de machines
- l'industrie lourde et les équipements industriels (forge, métallurgie, transformation de l'acier, production de métaux, …).

Nombre de sinistres antérieurs

La variable « Tranc_nb_anttotal_normCA » permet de segmenter le risque selon les antécédents de sinistres sur les 5 dernières années. Afin de pouvoir comparer le nombre de sinistres d'une entreprise ayant un chiffre d'affaires de 10 000 euros avec celui d'une entreprise ayant un chiffre d'affaires de 20 millions d'euros (donc ayant en toute logique plus de sinistres), le nombre de sinistres antérieurs est rapporté au chiffre d'affaires.

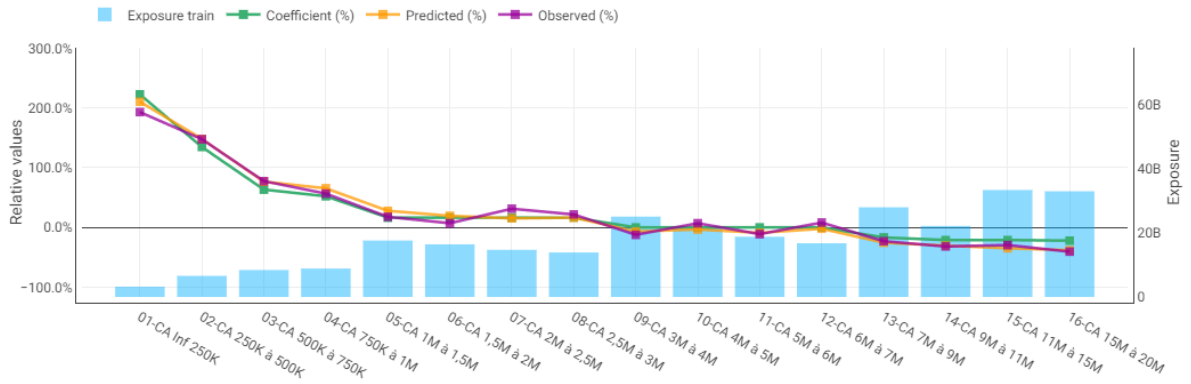


Coefficient de la variable nombre de sinistres antérieurs divisé par le chiffre d'affaires

Sans surprise, comme dans la majorité des risques IARD, la sinistralité passée est un bon indicateur de la sinistralité future. Nous observons que la courbe des coefficients est croissante à mesure que le nombre de sinistres augmente. Une entreprise n'ayant pas eu de sinistre dans les 5 dernières années a 25% de risque de moins que la moyenne d'avoir un sinistre. Les entreprises ayant la plus forte antériorité de sinistres ont 200% de risque en plus que la moyenne d'avoir un nouveau sinistre.

Tranche de chiffre d'affaires

La variable « Tranche_CA_base20 » permet de segmenter le risque selon la tranche de chiffre d'affaires de l'entreprise. Le chiffre d'affaires a été segmenté en 16 tranches.



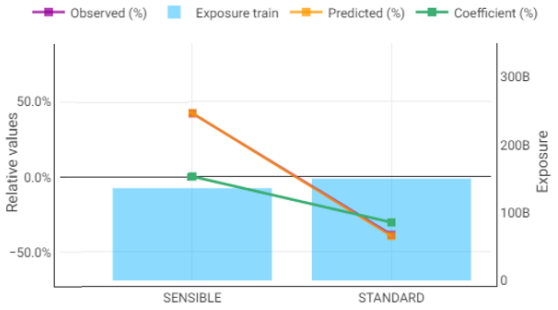
Coefficient de la variable tranche de chiffre d'affaires

Dans notre modèle, nous avons utilisé en variable d'exposition le chiffre d'affaires de l'entreprise multiplié par l'exposition c'est-à-dire le nombre de jours où le contrat a été présent dans l'année (exposition comprise entre 0 et 1, avec une valeur de 1 si le contrat a été présent toute l'année). La charge prédite par le modèle est donc un taux, comme demandé par notre direction technique pour l'un de nos outils, que nous devons appliquer au chiffre d'affaires de l'entreprise. Nous allons donc multiplier les coefficients obtenus dans le GLM par le chiffre d'affaires et l'exposition.

Nous constatons que la charge de sinistres n'augmente pas proportionnellement au chiffre d'affaires. Ainsi pour la tranche 01 (entreprises avec un CA inférieur à 250K€), la charge totale de sinistres est de 7 millions d'euros pour un CA cumulé de 4 milliards. Pour la tranche 08 (2,5 à 3 millions de CA), la charge est de 13 millions d'euros pour un CA cumulé de 18 milliards. Pour la dernière tranche (15 à 20 millions de CA), la charge de sinistres est de 17 millions d'euros pour un CA cumulé de 43 milliards. La charge de sinistres cumulée croît donc moins vite que le CA cumulé. Proportionnellement au CA, les entreprises ayant un CA plus faible ont un risque plus important, ce qui explique la décroissance des coefficients. Les coefficients sont plus élevés pour les entreprises ayant un CA inférieur à 1 million d'euros et plus bas pour celles dont le CA est supérieur à 7 millions.

<u>Sensibilité du code NAF</u>

La variable « ENT_segmentbis » est une variable renseignée par la Direction Technique et qui indique la sensibilité du NAF.

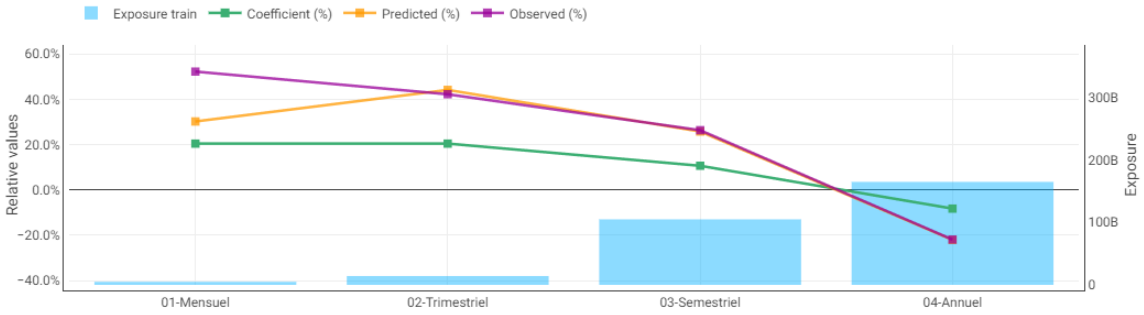


Coefficient de la variable sensibilité du NAF

On constate que les modalités construction, exclu et réservé sont sous-représentées et par conséquent, sans charge sinistre suffisante pour pouvoir modéliser ces modalités. La modalité «standard» qui représente des activités ayant un risque non aggravé a un coefficient plus faible (-30%) que les autres modalités.

<u>Fractionnement de la prime</u>

La variable « COT_fractionnement » est une variable qui indique la modalité de paiement de la prime choisie par l'entreprise au moment de la souscription : mensuelle, trimestrielle, semestrielle ou annuelle.

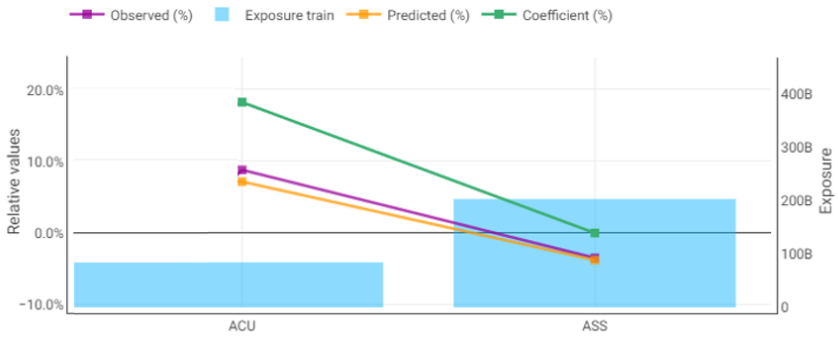


Coefficient de la variable fractionnement

Les contrats dont la prime est annualisée ont un coefficient de -8% par rapport à la moyenne, ce qui laisse à penser que les entreprises qui règlent leur prime en une fois peuvent être considérées comme des « bons » assurés. Par ailleurs, il est justifié de faire bénéficier l'entreprise qui paie sa prime en une fois d'une réduction tarifaire au titre d'actes de gestion diminués. Les entreprises qui paient une cotisation semestrielle ont un coefficient de +10% par rapport à la moyenne et les entreprises qui règlent leur cotisation trimestriellement ou mensuellement ont un coefficient de +20% par rapport à la moyenne.

Distributeur

La variable « CNT_DIST » est une variable qui indique le réseau via lequel le contrat est souscrit (Agent, Courtier ou Salarié commercial).

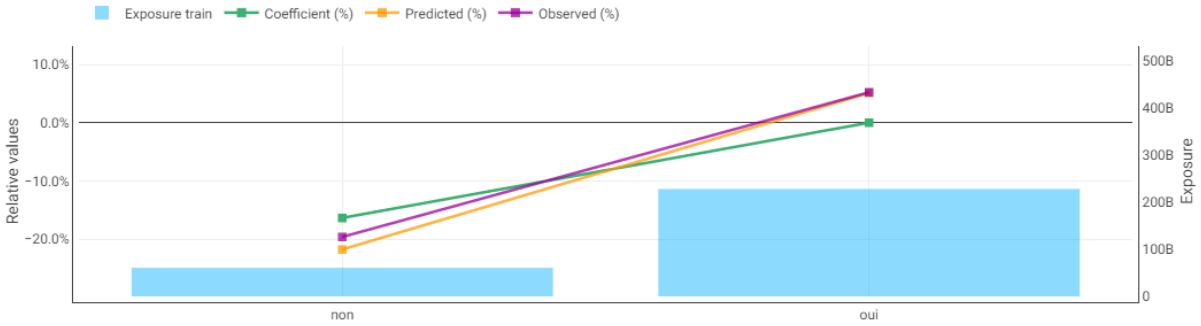


Coefficient de la variable distributeur

La modalité Salarié commercial est sous-représentée et par conséquent, sans charge sinistre suffisante. Les entreprises en portefeuille Agents sont généralement moins risquées que les entreprises apportées par les courtiers dont le coefficient est majoré de 18%.

Garanties pollution, export USA et export Hors USA

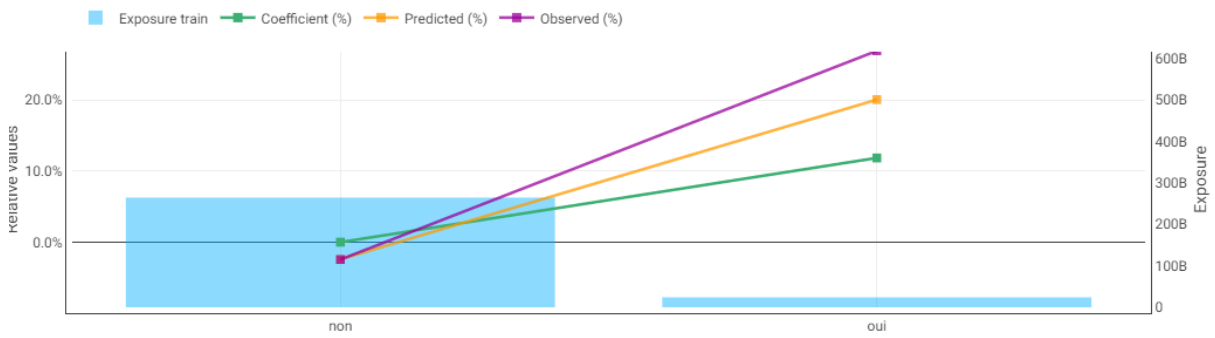
Enfin les variables « COT_TOP_pollution », « COT_TOP_EXPORT_USA » et « COT_TOP_EXPORT_Hors_USA » définissent si les garanties pollution, export vers les Etats Unis et Export vers d'autres pays que les Etats Unis sont souscrites.



Coefficient de la variable Top_pollution

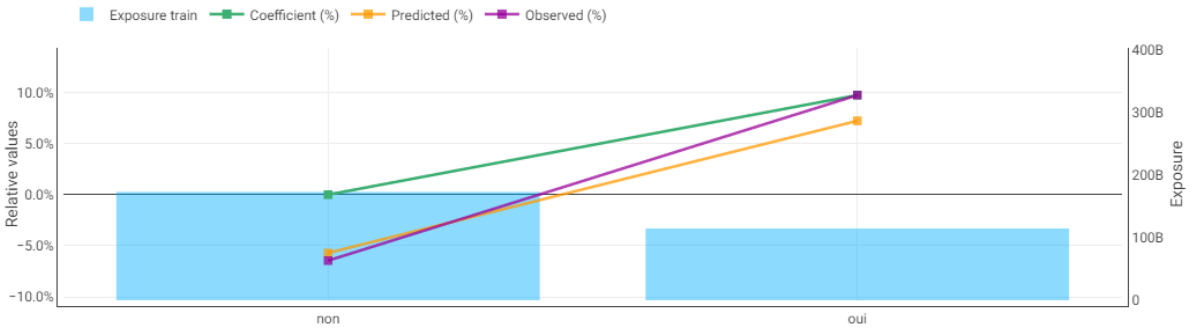
Sans surprise, l'absence de la garantie pollution au contrat fait baisser le coefficient de 16%.
Il en est de même pour les garanties exportations hors USA et export USA.

Les entreprises qui exportent vers les Etats-Unis ont un coefficient de +12%. Les procédures judiciaires sont nombreuses aux Etats-Unis, les montants des indemnisations sont très élevés et les assurés sont défendus par des avocats dont la rémunération est liée au seuil des indemnisations.



Coefficient de la variable Export_USA

Les entreprises qui effectuent des exportations vers d'autres pays que les Etats-Unis ont un coefficient de +10%.



Coefficient de la variable Export_hors_USA

A noter que la garantie frais de retrait qui permet la mise en garde du public ou des détenteurs de biens ainsi que le retrait de produits mis en circulation par l'assuré en cas de menace ou de survenance de dommages corporels ou dommages matériels garantis apparaissait en dixième position dans la liste des variables. L'intégration de cette variable frais de retrait faisait baisser les performances du modèle et son coefficient était inférieur à +1%, les autres variables du modèle captant déjà l'information exprimée par cette variable.

La variable Top_faute_inexcusable, qui indique si la garantie faute inexcusable de l'employeur a été souscrite, n'avait pas été sélectionnée car trop fortement corrélée à la variable Top_pollution (99% de corrélation).

Nous réussissons donc à expliquer la charge de sinistre attritionnelle et grave de nos contrats entreprises industries et commerces via un modèle linéaire généralisé composé de neuf variables.

Les indicateurs statistiques sont équivalents sur les bases d'apprentissage et de test, ce qui montre la stabilité du modèle et son ajustement aux données ainsi que son caractère prédictif. L'analyse des résidus quantiles confirme que notre modèle est adapté pour prédire notre charge de sinistre.

Chapitre 4

UTILISATION DES SHAP VALUES POUR SEGMENTER LES CODES ACTIVITES

L'utilisation d'un GLM pour modéliser notre charge de sinistres attritionnelle et grave nous permet de déterminer une prime pure segmentée selon les neuf variables explicatives du risque. Nous avons obtenu une structure tarifaire multiplicative, adaptée aux données en assurance non-vie. Pour une nouvelle observation, nous obtenons le montant de la prime pure pour la partie attritionnelle et grave en multipliant les coefficients de chaque variable explicative à l'intercept et à notre offset (exposition multipliée par le chiffre d'affaires de l'entreprise).

Cependant, nous avons injecté dans notre GLM une variable explicative pour l'activité de l'entreprise qui est un regroupement d'activités en 17 classes. Or les souscripteurs ont besoin d'un coefficient distinct pour des activités différentes à la maille du NAF dans chacune des classes. Le nombre de modalités de la variable NAF (plus de 300 NAF) n'a pas permis d'utiliser cette variable dans notre GLM qui ne convergeait plus lorsque nous l'introduisions.

Cette dernière partie s'attache à la reprise des résultats du GLM sur les activités pour descendre à un niveau plus fin que celui utilisé dans notre modèle. L'idée est de voir dans quelle mesure la méthode des SHAP (SHapley Additive exPlanations) values peut nous aider à descendre au niveau code NAF en mesurant la contribution de chaque NAF dans les regroupements qui ont servi d'input au GLM et ainsi différencier les coefficients pour chaque activité.

1. LES SHAP VALUES : PRESENTATION ET DEFINITION

1.1 LES SHAP VALUES DANS LA THEORIE DES JEUX

Les valeurs de Shapley sont une méthode initialement utilisée dans la théorie des jeux coopératifs. On parle de jeu coopératif lorsque deux ou plusieurs joueurs sont impliqués dans une stratégie visant à obtenir un gain ou un résultat.
En théorie des jeux, n joueurs collaborent ensemble pour obtenir un gain. La question est de savoir comment répartir de manière équitable le gain entre ces n joueurs, c'est à dire en prenant en compte la contribution juste de chaque joueur au gain. Cela signifie que l'on ne rémunère pas un joueur uniquement en fonction du gain qu'il peut apporter seul mais en fonction de sa contribution au groupe lorsqu'il interagit avec les autres joueurs. La valeur de Shapley s'applique principalement dans les situations où les contributions des joueurs sont inégales.
Cette valeur incite les joueurs à collaborer en garantissant à chaque joueur de gagner autant ou plus que s'il jouait indépendamment des autres.

On suppose que l'on a un jeu coopératif avec n joueurs. Une coalition S est un sous-ensemble de joueurs. On appelle $N = \{1, .., n\}$ la grande coalition.
Le nombre réel v(S) s'interprète comme le gain dont les membres de S bénéficient dans le cas où la coalition S est jouée.

Lloyd Shapley (mathématicien et économiste américain) a énoncé qu'il est possible de déterminer une répartition unique du gain généré par la grande coalition N de manière équitable entre les joueurs dans un jeu coopératif en introduisant des axiomes intuitifs.

Il existe donc une fonction ϕ unique :

$$\phi \begin{cases} \mathrm{N} & \rightarrow \mathbb{R} \\ i & \rightarrow \phi_i(v) \end{cases}$$

qui satisfait les quatre propriétés suivantes afin de répartir le gain d'un jeu coopératif :

- L'efficience : la somme des rétributions de chaque joueur doit être égale au gain total.

$$\sum_{i \in N} \phi_i(v) = v(\mathrm{N})$$

- La symétrie : les numéros des joueurs ne jouent aucun rôle dans la détermination de leur rétribution. Si deux joueurs contribuent de la même façon dans toutes les coalitions dans lesquelles ils jouent, leurs rétributions doivent être égales.

$$\forall S, v(S \cup \{i\}) = v(S \cup \{j\}) \Rightarrow \phi_i(v) = \phi_j(v)$$

- La nullité : si toutes les coalitions dans lesquelles un joueur est présent ont le même gain avec et sans ce joueur, alors la rétribution de ce joueur est nulle.

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

- L'additivité : si v et w sont deux jeux, alors la rétribution de chaque joueur pour le jeu somme v+w est égale à la somme de leurs rétributions pour les jeux v et w.

$$\phi_i(v + w) = \phi_i(v) + \phi_i(w)$$

Donnée sous sa forme probabiliste, la fonction ϕ vérifie :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!\,(n - s - 1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right)$$

avec s le cardinal de la colation S, c'est-à-dire le nombre de joueurs dans ce sous-ensemble S.

ϕ_i est la contribution du joueur i au gain. Pour obtenir ϕ_i, il faut donc calculer pour toutes les coalitions possibles dans lesquelles le joueur i n'apparaît pas, la différence de gain $v(S \cup \{i\}) - v(S)$. Cela permet de calculer le gain obtenu par la coalition avec et sans le joueur i afin de mesurer l'impact du joueur i lorsqu'il collabore avec les joueurs de cette coalition.

Si la différence $v(S \cup \{i\}) - v(S)$ est positive, cela signifie que le joueur i contribue positivement au gain de la colaition S. A contrario, si cette différence est négative, cela signifie que le joueur i pénalise les autres joueurs de la coalition. Enfin si $v(S \cup \{i\}) - v(S) = 0$, alors la contribution du joueur i au groupe est nulle.

On calcule ensuite la moyenne de ces différences pour toutes les coalitions S dans lesquelles le joueur i n'est pas présent.

Pour cela on classe les coalitions S par cardinal s et on fait la moyenne des différences $v(S \cup \{i\}) - v(S)$ pour toute les coalitions S ayant le même cardinal, sachant que pour une taille de coalition k, il y a $\binom{n-1}{k}$ coalitions possibles. On fait ensuite la moyenne de ces moyennes, sachant qu'il y a n tailles de coalitions différentes. A noter que la coalition de taille 0 est l'ensemble vide.

L'équation

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!\,(n-s-1)!}{n!}\,(v(S \cup \{i\}) - v(S))$$

peut donc être réécrite sous la forme de moyennes de différences :

$$\phi_i(v) = \frac{1}{n}\sum_{k=0}^{n-1} \binom{n-1}{k}^{-1} \sum_{\substack{S \subseteq N \setminus \{i\} \\ s=k}} (v(S \cup \{i\}) - v(S))$$

Lloyd Shapley a énoncé qu'il est possible de déterminer une répartition unique du gain généré dans un jeu coopératif de manière équitable entre les joueurs à l'aide d'axiomes intuitifs qui permettent de calculer les valeurs de Shapley.

1.2. LES SHAP VALUES ET LES MODELES DE PREDICTION

Le calcul des SHAP values dans la théorie des jeux peut être utilisé dans les modèles de machine learning. Les joueurs vont alors devenir les variables explicatives et comme pour la théorie des jeux nous allons calculer leur contribution au modèle.
L'idée est d'extraire des informations du modèle afin de comprendre sur quoi se base le modèle pour effectuer les prédictions, soit globalement (contribution des variables et des modalités au modèle), soit localement pour expliquer une prédiction particulière.

Voici un exemple illustré par un schéma pour comprendre comment les SHAP values peuvent permettre de comprendre un modèle de machine learning localement au niveau d'une prédiction. Un hôpital utilise un modèle prédictif de type boîte noire pour calculer la probabilité de décès d'un patient en fonction de différentes caractéristiques physiques comme l'âge, la pression artérielle, …

Dans le schéma ci-dessous, nous voyons le cas d'une patiente dont la probabilité de décès est supérieure à la moyenne mais nous ne connaissons pas les raisons pour lesquelles cette probabilité est supérieure. Les SHAP values vont permettre d'attribuer un coefficient à chacune des caractéristiques physiques décrivant cette personne et d'expliquer les raisons qui ont fait augmenter sa probabilité de décès par rapport à la moyenne (boîte à droite du graphique).

Source : GitHub - slundberg/shap: A game theoretic approach to explain the output of any machine learning model.

Nous pouvons ainsi voir sur cet exemple que l'âge de la patiente, sa pression artérielle et son IMC ont fait augmenter la probabilité de décès par rapport à la valeur moyenne de référence (0,1) et dans quelle proportion. Nous pouvons aussi voir qu'à contrario, le sexe a fait baisser le risque de décès de manière importante (-0,3).

Pour notre problème d'affectation d'un coefficient à chaque activité, l'idée est de calculer comme en théorie des jeux, une contribution "juste" de chaque variable dans la prédiction d'une instance.
Dans notre cas, les joueurs vont donc devenir les modalités de nos variables explicatives et le gain à répartir devient la différence entre la prévision et la moyenne des prévisions.

L'utilisation des SHAP values pour interpréter les modèles de machine learning a été introduite par Scott Lundberg en 2017.
Pour chaque prédiction d'un modèle de machine learning, l'algorithme SHAP calcule les contributions de chaque variable explicative à cette prédiction. L'explication est exprimée comme une fonction linéaire des n variables. Au voisinage d'une observation x, la prédiction $f(x)$ peut être approximée par la somme des effets des variables comme suit :

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \mathbb{1}_i$$

avec ϕ_0 la valeur de base du modèle et $\mathbb{1}_i$ la fonction indicatrice qui vaut 1 si la variable est présente et 0 sinon.

L'algorithme SHAP a été conçu de manière à vérifier les propriétés définies par Lloyd Shapley. Ainsi, la somme des valeurs de contribution des variables est égale à la sortie du modèle. Les variables manquantes n'ont aucun impact. De plus si le modèle change de telle sorte que la contribution d'une variable X_i augmente ou reste la même, alors la valeur de contribution de cette variable ne doit jamais diminuer.

Dans le cas du machine learning, nous avons donc un modèle XGBoost qui retourne des prédictions à partir de variables explicatives. Ce modèle XGBoost est difficile à interpréter. Cependant, nous connaissons la moyenne des prédictions. A partir de cette moyenne, l'objectif est d'expliquer l'écart entre la prédiction et la moyenne des prédictions grâce aux variables explicatives observées. L'objectif de la méthode SHAP proposée par Scott Lundberg est d'attribuer pour chaque prédiction une importance à chaque variable explicative et de savoir dans quelle proportion les variables ont contribué à l'augmentation ou la diminution de la valeur de la prédiction par rapport à une valeur de base (qui est la moyenne des prédictions).

Par analogie au jeu :

- Les joueurs sont remplacés par les variables explicatives X
- La différence entre la prédiction et la prédiction moyenne est distribuée de manière équitable entre les différentes variables utilisées par le modèle.

Pour obtenir la valeur ϕ_i qui donne l'importance de la contribution de la i^{ème} variable au calcul de la prédiction, comme dans la théorie des jeux, on observe l'impact de l'absence de cette i^{ème} variable sur la prédiction. Pour une observation donnée, on calcule la différence de prédiction entre le modèle prenant en compte cette i^{ème} variable et le modèle ne la prenant pas en compte.

Importance de la variable X_i = prédiction du modèle avec X_i – prédiction du modèle sans X_i

L'impact du retrait de la i^{ème} variable sur la prédiction dépend des autres variables. En effet cette i^{ème} variable interagit avec les autres variables, notamment au travers des corrélations. Si deux variables X_1 et X_2 interagissent, l'absence de X_1 va entraîner une augmentation de la contribution de X_2 et

inversement. Il faut donc comme dans le calcul des valeurs de Shapley dans la théorie des jeux, considérer toutes les coalitions S possibles (c'est-à-dire les sous-ensembles de variables explicatives parmi les n variables explicatives) et entraîner le modèle f sur chaque sous-ensemble S et sur chaque sous-ensemble SU$\{i\}$.
On note N l'ensemble de toutes les variables explicatives et s le nombre de variables dans la coalition S.

L'équation de la théorie des jeux qui permet de calculer la valeur de Shapley devient pour une observation x, pour le modèle :

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!\,(n-s-1)!}{n!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]$$

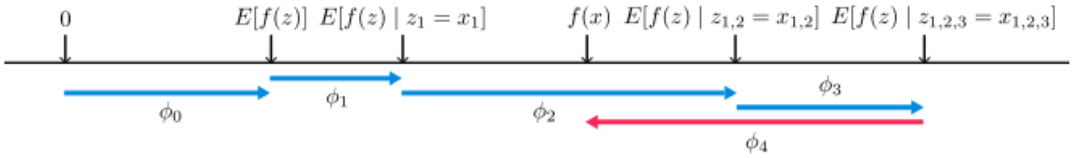
Avec f_S le modèle complexe entraîné sur les variables de la coalition S et $f_{S\cup\{i\}}$ le modèle complexe entraîné sur les variables de la coalition S et la variable i.

Cette équation peut encore s'écrire sous la forme :

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!\,(n-s-1)!}{n!} \left[\mathbb{E}[f(X) \,|\, X_{S\cup\{i\}} = x_{S\cup\{i\}}] - \mathbb{E}[f(X) \,|\, X_S = x_S]\right]$$

Le graphique ci-dessous explique cette formule d'obtention des SHAP values pour un ordre unique d'incrémentation de quatre variables explicatives dans un modèle f.
Chaque SHAP value ϕ_i est obtenue par la variation de valeur de l'espérance conditionnelle de $f(X)$ lorsque nous ajoutons la condition par rapport à la variable X_i.

En partant de l'espérance des prédictions, c'est-à-dire $\mathbb{E}[f(X)]$, nous obtenons la valeur de base ϕ_0. Ensuite, nous ajoutons la variable X_1 au modèle f pour voir quelle information elle apporte à la prédiction $f(X)$. Nous obtenons la valeur de ϕ_1 en calculant la différence $\mathbb{E}[f(X) \,|\, X_1] - \mathbb{E}[f(X)]$. Nous ajoutons ainsi une à une les quatre variables du modèle.



Source : A Unified Approach to Interpreting Model Predictions. Scott Lundberg

Toutefois, l'ordre dans lequel nous ajoutons les variables n'est pas indépendant car il y a des interactions et corrélations entre les variables. C'est pourquoi dans la formule de calcul des SHAP values il faut considérer tous les ordres d'ajouts possibles. Deux variables prises isolément peuvent n'avoir aucun pouvoir de prédiction alors qu'ensemble elles peuvent être très informatives de la prédiction.
Les SHAP values résultent de la moyenne des valeurs ϕ_i sur tous les ordres possibles.

Les SHAP values attribuent ainsi à chaque variable la modification de la prédiction attendue du modèle lors du conditionnement sur cette variable (voir le graphique). La valeur de base $\phi_0 = \mathbb{E}[f(X)]$ est la valeur qui serait prédite si nous ne connaissions aucune caractéristique de la sortie.

A noter que c'est la quantité $f(x) - \mathbb{E}[f(X)]$ qui est répartie entre les variables (c'est-à-dire les joueurs) et non la prévision elle-même. Le coefficient ϕ_i explique donc comment les valeurs x_i contribuent à décaler la prévision $f(x)$ de la moyenne des prévisions $\mathbb{E}[f(X)]$.

La somme de ϕ_0 et des ϕ_i approxime la prédiction $f(x)$.

Si nous reprenons le graphique, pour une observation x, nous pouvons vérifier que la somme des SHAP values est égale à la différence entre la prédiction $f(x)$ et la moyenne des prédictions $\mathbb{E}[f(X)]$. Dans l'exemple du graphique, la somme $\phi_1 + \phi_2 + \phi_3 + \phi_4$ explique la position de la prédiction $f(x)$ au milieu du graphique par rapport à $\phi_0 = \mathbb{E}[f(X)]$ à gauche du graphique. Les SHAP values ϕ_1, ϕ_2 et ϕ_3 ont contribué à l'augmentation de la valeur de la prédiction par rapport à la moyenne des prédictions alors que la SHAP value ϕ_4 a contribué négativement à la prédiction $f(x)$.

Si la définition mathématique des SHAP Values est relativement simple, leur estimation numérique est plus compliquée. Il faut en effet estimer les espérances conditionnelles et gérer le nombre important de combinaisons du nombre de coalitions à parcourir lorsque le nombre de variables augmente. Le nombre de colations est en effet exponentiel.

Le but de l'algorithme SHAP introduit par Scott Lundberg est de calculer les SHAP values en expliquant localement (c'est-à-dire au voisinage d'une observation x) par un modèle simple g notre modèle complexe f (XGBoost). Nous cherchons donc g tel que :

$$\forall z \; tel \; que \; z \approx x, alors \; g(z) \approx f(x)$$

Dans l'algorithme SHAP, le modèle g est une fonction linéaire de variables binaires qui assigne une contribution ϕ_i à chacune des n variables observées.

$$g(z) = \phi_0 + \sum_{i=1}^{n} \phi_i \mathbb{1}_i \approx f(x)$$

En 2017, Scott Lundberg a appliqué les valeurs de Shapley aux modèles de Machine Learning de type « boîte noire » au travers de l'algorithme SHAP pour déterminer la contribution de chaque variable explicative dans un modèle.

L'algorithme SHAP permet d'extraire des informations d'un modèle complexe pour expliquer globalement sur quoi se base le modèle pour prédire ou pour expliquer localement chaque prédiction. Les SHAP values calculées via cet algorithme définissent les contributions des variables et des modalités dans un modèle.

2. L'UTILISATION DES SHAP VALUES POUR SEGMENTER LES ACTIVITES DES ENTREPRISES

2.1. LA PREPARATION DES DONNEES

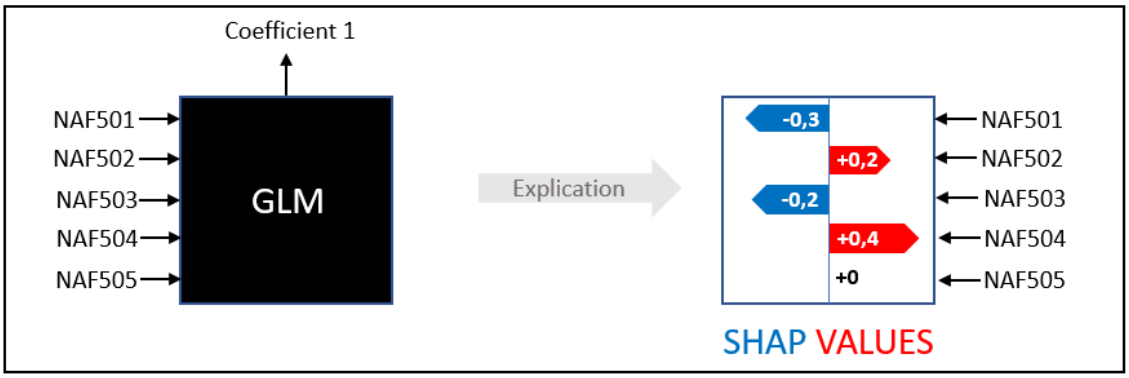
Notre modélisation GLM de la charge de sinistres a donné des coefficients pour chaque modalité des neuf variables explicatives, notamment pour la variable ENT_nlibniv3 qui regroupe les activités des entreprises en 17 classes.

Regroupement d'activités	Coefficient GLM
Agroalimentaire	0,87
Captage, traitement et distribution d'eau	1
Chimie / Pharmacie	1,09
Commerce de détail et réparation d'articles domestiques	0,46
Commerce de gros et intermédiaires du commerce	0,65
Commerce et réparation automobile	1
Energie, eau, chaleur	1
Entreprises Industrielles / Autres	1
Entreprises du bâtiment et des travaux publics	1
Fabrication de machines, équipements et instruments	1,48
Fabrication et construction de matériels de transport	2,05
Fabrication et transformation de matériaux	1
Hôtels et restaurants	0,70
Industrie Lourde / Equipements industriels	1,74
Prestataires de services / Autres	1,59
Services personnels	1
Transports et Services	0,42

Prenons l'exemple de la modalité niveau 3 « Commerce et réparation automobile » de notre GLM qui a obtenu un coefficient de 1. Ce regroupement d'activités comprend cinq activités (NAF) pour lesquelles nous constatons une sinistralité différente dans les tris à plats réalisés sur notre base de données :

- NAF 501Z000 : Vente de véhicules automobiles, remorques, semi-remorques et caravanes
- NAF 502Z000 : Entretien et réparation de véhicules automobiles
- NAF 5030000 : Vente en gros et au détail d'équipements automobiles
- NAF 5040000 : Vente en gros et au détail de motocycles
- NAF 505Z000 : Vente au détail de carburants (station-service).

Ce regroupement est l'équivalent d'une boîte noire à laquelle les activités 501, 502, 503, 504 et 505 ont participé pour obtenir un coefficient 1 mais dont nous ne connaissons pas la contribution au calcul du coefficient obtenu dans le GLM, d'où l'analogie avec un modèle de machine learning. L'idée est d'utiliser les SHAP values pour déterminer la contribution de chacune des ces activités.



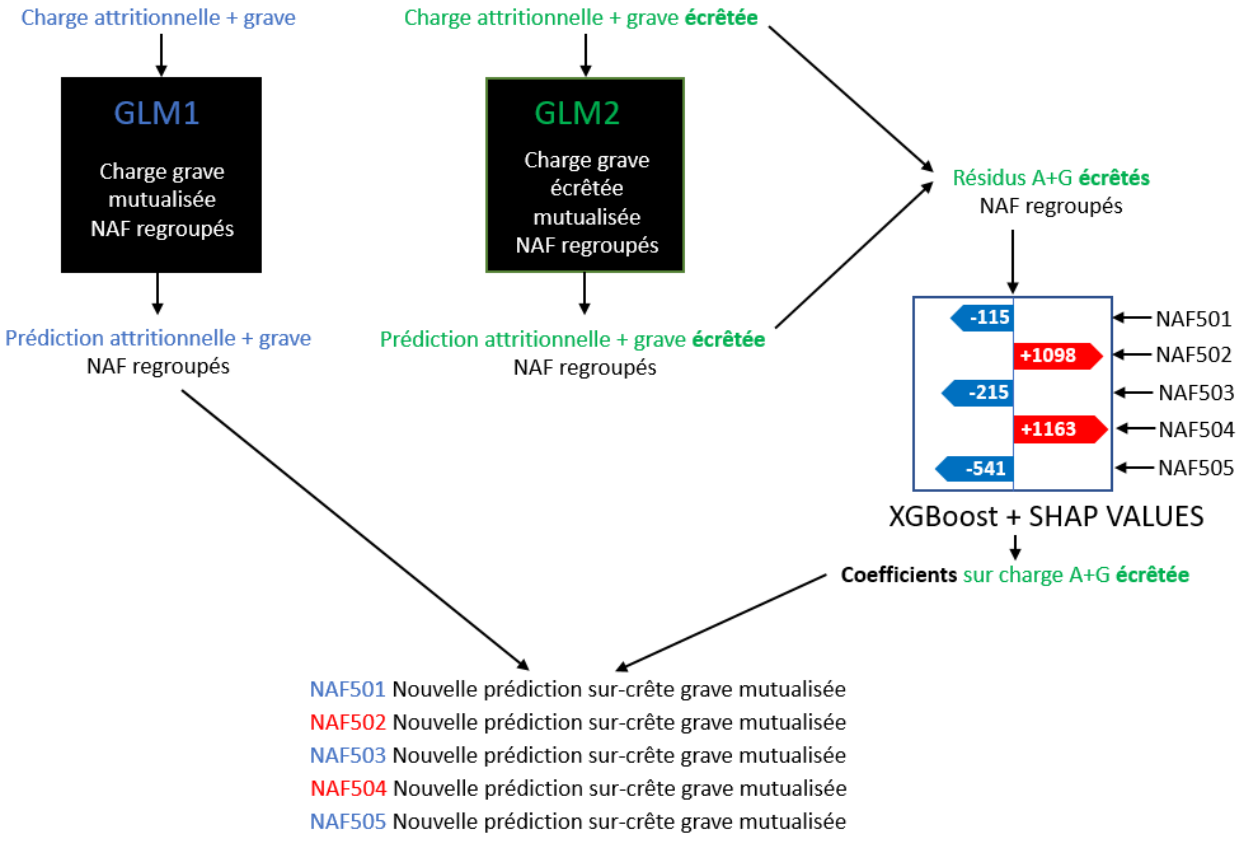
Utilisation des SHAP values pour déterminer les coefficients des activités

Pour déterminer l'impact de chacun des NAF associé à un niveau 3 d'activités, nous utilisons les résidus du modèle GLM. En effet, les résidus ont la particularité de représenter la partie non expliquée par le modèle. Pour rappel, les résidus r_i sont définis pour chaque individu i, comme l'écart entre la valeur observée y_i (charge réelle de sinistres) et la valeur prédite par le modèle $\hat{y}_i$:

$$r_i = y_i - \hat{y}_i$$

Nous avons réalisé notre GLM en modélisant une charge de sinistres attritionnels et graves et avons obtenu une prédiction de charge attritionnelle et grave. Cependant, pour la suite de notre étude, nous souhaitons conserver une partie de la charge grave au niveau du regroupement de niveau 3, qui correspond à la sur-crête, afin de mutualiser cette sur-crête de charge de sinistres graves. En effet, si nous attribuons la charge de sinistres graves à un niveau plus fin c'est-à-dire au niveau des codes NAF, nous ne lissons plus la charge de sinistres graves. Nous obtenons une charge trop importante pour les codes NAF sur lesquels un sinistre grave a été enregistré sur l'historique 2011-2018 par rapport à notre référentiel tarifaire actuel ainsi qu'une prime pure trop faible sur d'autres activités. De plus, certaines activités ont un volume de sinistres insuffisant pour absorber ces sinistres graves alors qu'ils sont correctement mutualisés lorsque nous modélisons le niveau 3 de regroupement (17 classes).

Pour calculer la contribution de chaque code NAF à un regroupement d'activités niveau 3 via les SHAP values tout en mutualisant la sur-crête des sinistres graves, nous reprenons notre GLM sur la même base de données avec les mêmes variables explicatives en ajustant la variable réponse. Nous modélisons la charge de sinistre attritionnelle et grave écrêtée à 150 000 euros qui est notre seuil de sinistre grave. Cette nouvelle modélisation sert uniquement pour obtenir les résidus nécessaires au calcul de la contribution de chaque NAF.
Nos résidus sont donc égaux à la charge de sinistres attritionnels et graves écrêtés à 150 000 euros moins la prédiction du nouveau GLM multipliée par l'offset. Ils représentent toujours la partie non expliquée par le GLM. Ils correspondent à un montant en euros.



Méthode de modélisation

Les résidus du nouveau GLM sur charge écrêtée vont servir pour un modèle de Gradient Boosting, étape préalable à l'obtention des SHAP values. Les résidus constituent la variable à expliquer du XGBoost et les variables explicatives correspondant aux codes NAF du niveau 3 à analyser, variables explicatives que nous transformons en indicatrices.

Si nous reprenons l'exemple précédent du regroupement de niveau 3 « Commerce et réparation automobile » qui comprend cinq activités, nous aurons cinq variables : NAF501 à NAF505 discrétisées avec la valeur 1 si l'activité de l'observation correspond au code NAF de la variable et la valeur 0 dans le cas contraire.
Nous obtenons ainsi une base de données des résidus de la forme suivante à utiliser dans le modèle XGBoost.

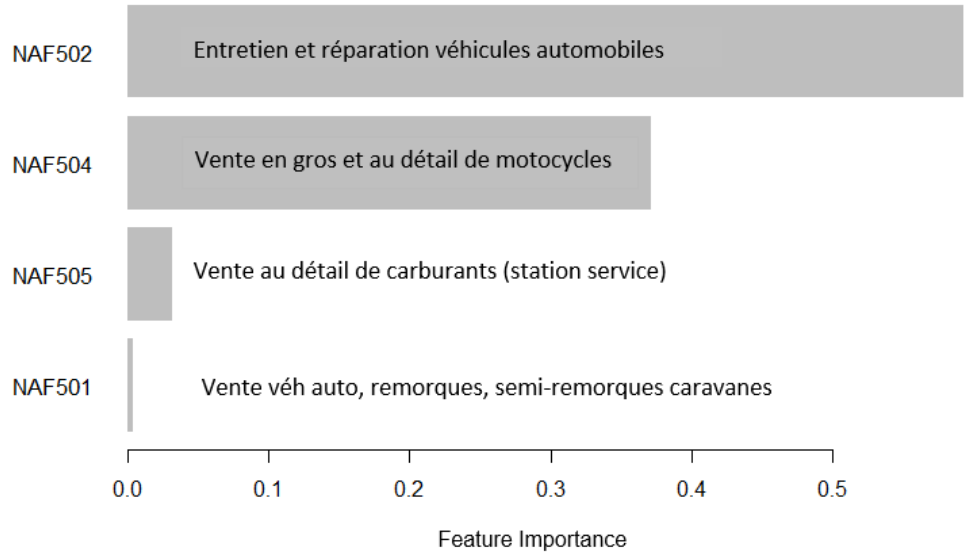
Variable	NAF501	NAF502	NAF503	NAF504	NAF505	Résidu
Observation 1	1	0	0	0	0	-10.32316
Observation 2	1	0	0	0	0	-11.55806

Cette base de données est scindée en deux bases :

- Une base d'entraînement qui contient 70% des observations
- Une base de test avec 30% des observations restantes qui nous servira à valider si les SHAP values apportent ou non de la précision au modèle GLM. Nous utilisons 30% des observations pour obtenir une base de test de taille suffisante.

Une des difficultés rencontrées lors de la constitution des bases d'entraînement et de test pour permettre le calcul des SHAP values est le faible volume de sinistres ainsi que la volumétrie de données lorsque nous descendons à un niveau de segmentation plus fin. La responsabilité civile entreprises étant une branche d'intensité, seules 5% de nos lignes environ ont une charge de sinistres. Il faut donc veiller à ce que les bases d'entraînement et de test représentent environ respectivement 70% et 30% de la charge sinistres sur chacune des activités.

Le graphique ci-dessous restitue l'ordre d'importance de nos cinq variables dans la construction du modèle XGBoost des résidus sur la base d'entraînement.



Ordre d'importance des variables dans le modèle XGBoost (base d'entraînement)

Variable	Importance de la variable dans le modèle
NAF502 Entretien et réparation de véhicules automobiles	0,593
NAF504 Vente en gros et au détail de motocycles	0,371
NAF505 Vente au détail de carburants (station-service)	0,0318
NAF501 Vente de véhicules automobiles, remorques, semi-remorques et caravanes	0,004

Les activités liées à l'entretien et la réparation de véhicules automobiles ainsi que la vente en gros et au détail de motocycles arrivent en premier par ordre d'importance pour expliquer les résidus et donc la charge de sinistres. Les activités de vente au détail de carburants et de vente de véhicules automobiles, remorques, semi-remorques et caravanes ont un ordre moindre d'importance dans le modèle. Enfin, la variable NAF503 (Vente en gros et au détail d'équipements automobiles) n'apparaît pas (importance proche de 0).

La sinistralité sur les activités d'entretien et de réparation de véhicules automobiles s'explique notamment par des dommages sur les véhicules lors des réparations ou de l'entretien ou des dysfonctionnements après réparation. Les garanties mises en jeu sont la responsabilité civile exploitation ou la responsabilité civile après livraison. L'activité d'entretien et de réparation enregistre également le plus grand nombre de sinistres corporels : salariés blessés lors de l'utilisation de machines ou renversés par des véhicules dans les locaux ou sur les parkings (garantie faute inexcusable de l'employeur) ou clients blessés lors de l'utilisation de machines. A noter que les stations de lavage sont particulièrement touchées par des sinistres matériels : rouleaux ou portiques qui endommagent les véhicules ou corporels : salariés ou clients gravement blessés par ces mêmes portiques ou rouleaux en voulant par exemple sortir du véhicule alors que le programme de lavage s'est déclenché ou en restant trop proches des portiques lors du lavage.

Pour l'activité de vente en gros et au détail de motocycles qui est le deuxième code activité par ordre d'importance dans le modèle XGBoost, la plupart des sinistres mettent en jeu la responsabilité civile après livraison (dysfonctionnements ou incendie du moteur après la vente). Plus de la moitié des sinistres de ce secteur sont imputables à la vente de quads qui prennent feu lors de l'utilisation.

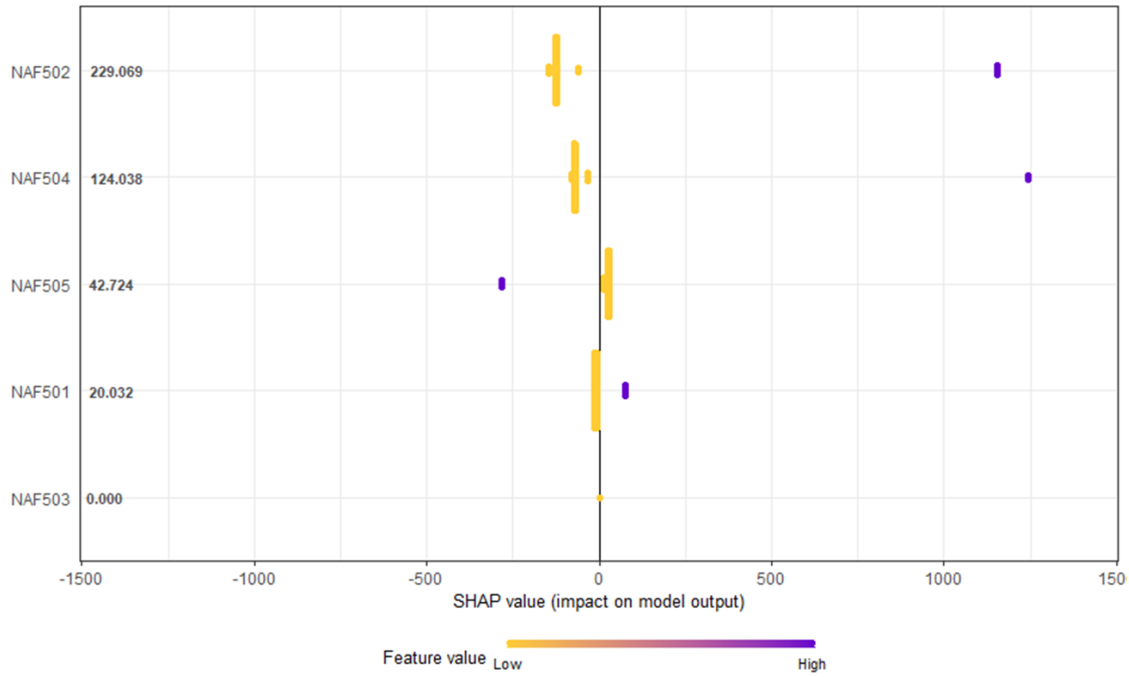
Pour les entreprises effectuant de la vente au détail de carburants, les sinistres sont souvent liés à des problèmes de pistolets défectueux aux pompes qui dans la plupart des cas tâchent les vêtements ou chaussures des clients. Il n'y a pas de dommages matériels importants car pas de véhicules endommagés ni de sinistres corporels constatés sur cette activité sur notre historique. La charge de sinistres est donc très faible sur cette activité.

À noter que la charge de sinistres totale sur l'activité NAF503 (vente d'équipements automobiles) qui n'apparaît pas dans le graphique est élevée au global mais cette activité concentre 65% des entreprises du regroupement. Rapportée au volume de contrats, la charge est donc moindre.

2.2. L'ANALYSE DES SHAP VALUES

Les fonctions du package SHAP de R servent à convertir les résultats de sortie du XGBoost en un tableau de coefficients pour chacune de nos variables NAF501 à NAF505 correspondant aux SHAP values et obtenir des représentations visuelles.

Le graphique ci-dessous restitue le nuage des SHAP values calculé pour chaque variable explicative sur la base d'entraînement.



Influence des variables explicatives (base d'entraînement)

Le nuage en violet correspond aux valeurs élevées de la variable et le nuage en jaune aux valeurs faibles.

En abscisse, nous trouvons les SHAP values. Une SHAP value positive signifie que la variable contribue positivement au résidu et une SHAP value négative signifie que la variable contribue négativement à ce résidu. En ordonnée, les variables apparaissent par ordre d'importance de haut en bas. La valeur sur cet axe est la valeur absolue de la valeur SHAP moyenne.

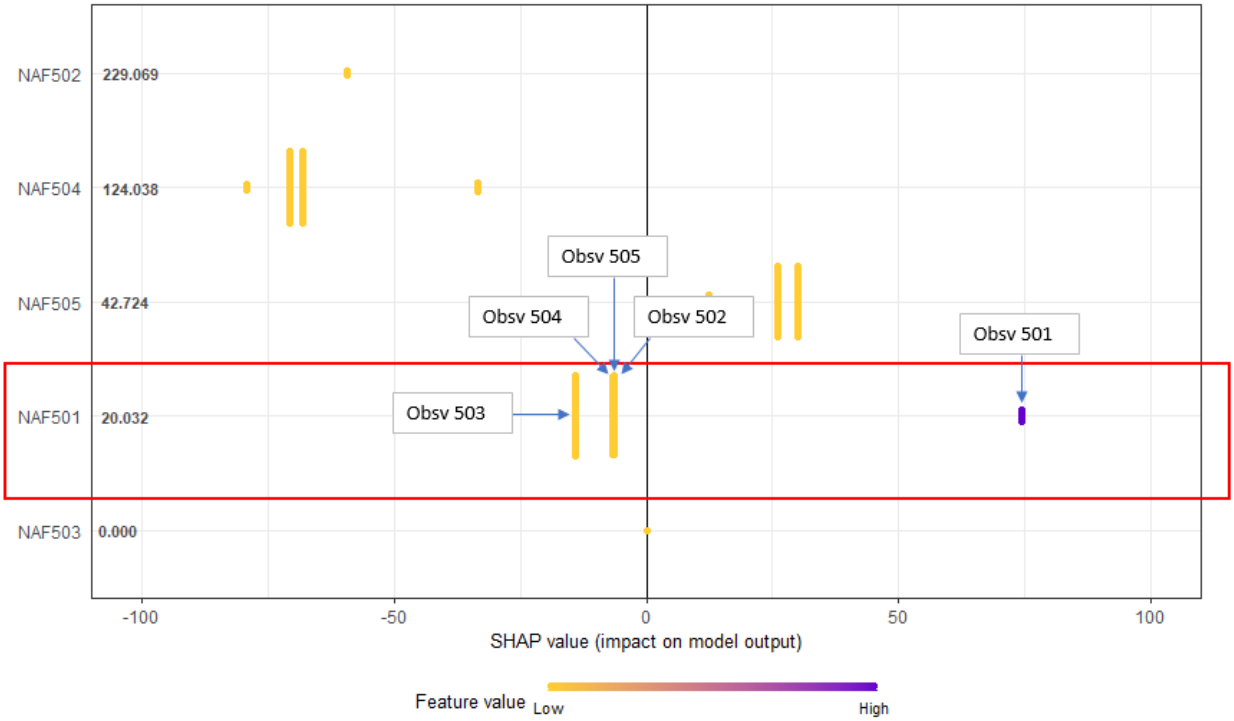
Les SHAP values suivantes sont obtenues sur la base d'entraînement.

Observation Variable /SHAP VALUE	**NAF501** ϕ_1	**NAF502** ϕ_2	**NAF503** ϕ_3	**NAF504** ϕ_4	**NAF505** ϕ_5	**BIAS** ϕ_0
Observation avec un NAF 501Z000	74,4674	-124,0946	0	-68,18912	29,87564	-26,98495
Observation avec un NAF 502Z000	-6,548689	1152,705	0	-33,53965	12,24858	-26,98495
Observation avec un NAF 5030000	-14,18428	-128,8941	0	-70,71577	25,90077	-26,98495
Observation avec un NAF 5040000	-6,66817	-59,38	0	1244,039	12,46714	-26,98495
Observation avec un NAF 505Z000	-6,59227	-145,2848	0	-79,34113	-283,2897	-26,98495

Le BIAS de -26,98495 correspond au coefficient $\phi_0 = \mathbb{E}[f(X)]$, la moyenne des résidus en euros. Le coefficient ϕ_1 est la SHAP value de la variable NAF501, ϕ_2 la SHAP value de la variable NAF502, … pour chacune des observations.

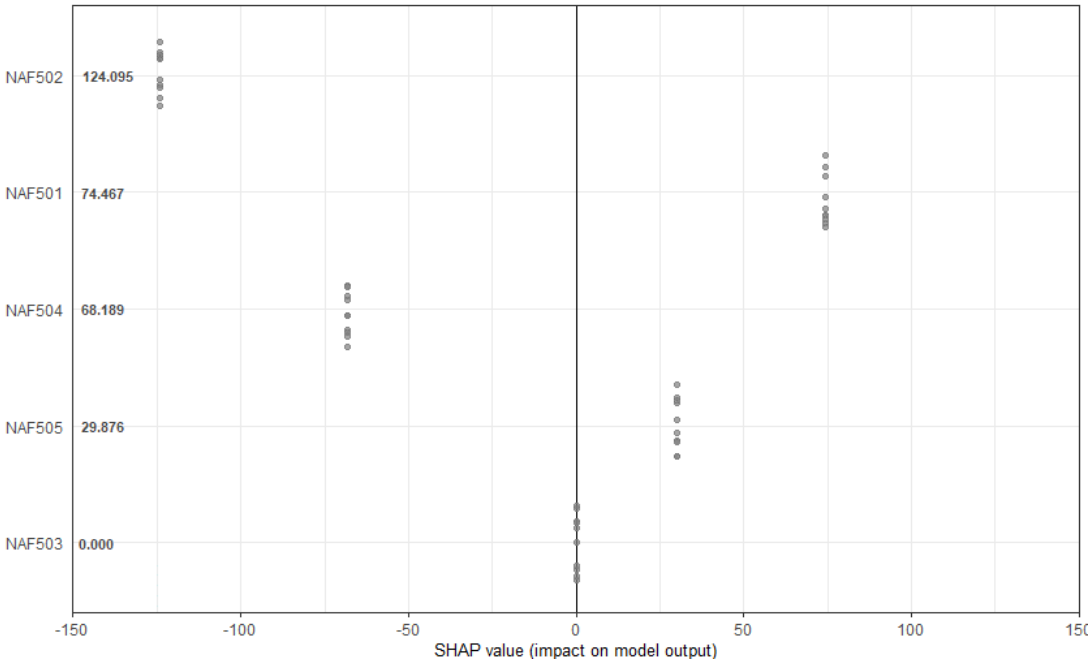
Si nous grossissons le graphique précédent et que nous observons la ligne correspondant à la variable NAF501, nous retrouvons les valeurs ϕ_1 de la variable NAF501 pour les différentes observations. Ainsi les observations qui ont un NAF 501Z000 contribuent positivement à la variable NAF501 et les observations qui ont un code NAF 502Z000, 5030000, 5040000 ou 505Z000 contribuent négativement à la variable NAF501.

Nous retrouvons dans ce graphique les colonnes du tableau des SHAP values. Sur la ligne d'ordonnée NAF501, nous avons les valeurs ϕ_1 de la colonne NAF501 pour les différentes observations.



Influence des variables explicatives : zoom sur la variable NAF501

Si nous sélectionnons 10 observations ayant un code 501Z000, nous obtenons le graphique ci-dessous.



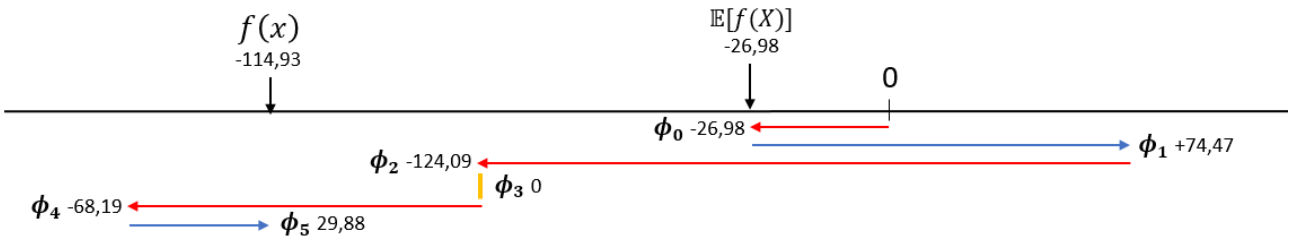
Affichage des SHAP values pour 10 observations ayant un code NAF501Z000

Nous retrouvons sur la première ligne (variable NAF502), le coefficient ϕ_2 correspondant à ces 10 observations qui ont un code NAF 501Z000 (-124,095), sur la deuxième ligne, le coefficient ϕ_1 de ces mêmes observations (74,467), … Nous retrouvons donc les lignes du tableau des SHAP values.

Nous connaissons ainsi la contribution des observations qui ont une activité 501Z000 (Vente de véhicules automobiles, remorques, semi-remorques et caravanes) aux variables NAF501, …, NAF505.

Si nous reprenons le graphique de Scott Lundberg, nous pouvons ainsi positionner les coefficients pour les observations avec un NAF 501Z000.

Observation Variable /SHAP VALUE	**NAF501 ϕ_1**	**NAF502 ϕ_2**	**NAF503 ϕ_3**	**NAF504 ϕ_4**	**NAF505 ϕ_5**	**BIAS ϕ_0**
Observation avec un NAF 501Z000	74,4674	-124,0946	0	-68,18912	29,87564	-26,98495



Impact des variables sur une observation ayant un code NAF501Z000

Pour une observation avec un NAF 501Z000 sur le graphique, la somme $\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5$ explique la position du résidu $f(x)$ à gauche du graphique par rapport à $\phi_0 = \mathbb{E}[f(X)]$. Les SHAP values ϕ_2 et ϕ_4 ont contribué à la diminution de la valeur du résidu par rapport à la moyenne des résidus alors que les SHAP value ϕ_1 et ϕ_5 ont contribué à l'augmentation de la valeur du résidu $f(x)$. La SHAP value ϕ_3 qui est nulle n'a pas eu d'impact sur la valeur du résidu.
La somme $\phi_0 + \phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5$ (-114,93 euros) explique la position du résidu $f(x)$ par rapport à l'origine et nous donne la contribution des observations ayant un NAF 501Z000 donc la contribution de la variable NAF501 au résidu du GLM.

2.3. LA PRESENTATION DES RESULTATS

Les résidus du GLM ont été calculés comme l'écart entre la valeur observée y_i (charge réelle de sinistres) et la valeur prédite par le modèle $\hat{y}_i$. Les observations ayant un code NAF 501Z000 ont contribué négativement au résidu du GLM (-114,93 euros). Or dans le GLM, nous avons affecté le même coefficient 1 aux cinq codes NAF de notre regroupement. Dans le GLM, les prédictions des observations ayant un code NAF 501Z000 ont donc été surestimées de 114,93 euros lorsqu'elles ont contribué au regroupement de niveau 3 par rapport aux observations ayant un code NAF différent dans ce même regroupement.

Grâce aux valeurs des SHAP values, nous pouvons calculer pour chacun de nos codes NAF, leur contribution aux résidus du GLM (8ème colonne du tableau : somme des SHAP values).

Observation SHAP VALUE	**NAF501 ϕ_1**	**NAF502 ϕ_2**	**NAF503 ϕ_3**	**NAF504 ϕ_4**	**NAF505 ϕ_5**	**BIAS ϕ_0**	**Somme des SHAP values**	**Impact GLM**
NAF 501Z000	74,47	-124,09	0	-68,19	29,884	-26,98	-114,94	Surestime
NAF 502Z000	-6,55	1152,71	0	-33,54	12,25	-26,98	1097,88	Sous-estime
NAF 5030000	-14,18	-128,89	0	-70,72	25,90	-26,98	-214,88	Surestime
NAF 5040000	-6,67	-59,38	0	1244,04	12,47	-26,98	1163,47	Sous-estime
NAF 505Z000	-6,59	-145,28	0	-79,34	-283,29	-26,98	-541,49	Surestime

Il est ensuite possible de transformer ce montant en coefficient pour un code NAF (1 + charge totale surestimée rapportée à la somme des prédictions).

Observation	Coefficient
NAF 501Z000	0,81
NAF 502Z000	2,82
NAF 5030000	0,78
NAF 5040000	2,18
NAF 505Z000	0,97

Ces nouveaux coefficients obtenus grâce à l'utilisation de la méthode des SHAP values nous permettent de segmenter le portefeuille de manière plus fine. Nous obtenons ainsi des coefficients différenciés à la place de notre coefficient unique de niveau de regroupement 3 du GLM de modélisation de la charge attritionnelle et grave pour le regroupement « Commerce et réparation automobile ». Le coefficient initial du GLM qui était de 1 pour ce regroupement sera multiplié par un coefficient spécifique (calculé dans le tableau ci-dessus) pour chacune des cinq activités.

Analyse des résultats

L'étape suivante consiste à appliquer les nouveaux coefficients issus de la méthode des SHAP values aux prédictions calculées par le GLM avec neuf variables (dont la variable du niveau de regroupement 3) sur la base d'entraînement et sur la base de test.

Le tableau ci-dessous présente les résultats de prédictions de la charge attritionnelle et grave écrêtée sur la base d'entraînement : comparaison de la charge écrêtée (en bleu) aux anciennes prédictions du GLM niveau 3 (en gris) et aux nouvelles prédictions niveau NAF grâce aux SHAP values (en vert). Les deux colonnes de droite montrent les écarts entre la charge réellement observée et la charge prédite.

Base d'entraînement : charge attritionnelle et grave écrêtée

Code NAF	Charge attri et grave écrêtée	Anciennes prédict (GLM niveau 3)	Nouvelles prédict via SHAP value	Ecart charge obs / anciennes pred Niv3	Ecart charge obs / nouvelles pred NAF
501Z000	95 848	118 393	95 867	-22 544	-19
502Z000	268 709	95 098	268 563	173 611	146
5030000	767 944	989 232	768 122	-221 288	-178
5040000	163 245	74 742	163 166	88 503	79
505Z000	9 861	70 560	9 913	-60 699	-52
Total	1 305 607	1 348 025	1 305 631	-42 418	-25

Nous constatons que pour chaque code NAF l'application du coefficient issu de la méthode des SHAP values (colonne du tableau en vert) permet de mieux différencier chaque activité que le coefficient GLM de la variable niveau de regroupement 3 (colonne du tableau en gris) qui était similaire pour les cinq activités. La nouvelle charge prédite pour chacune des activités via les SHAP values est très proche de la charge observée, contrairement à la charge prédite via le GLM pour laquelle les activités ne sont pas segmentées. Cependant, c'est sur cette base d'entraînement que les coefficients appliqués pour calculer les nouvelles prédictions de charges ont été calculés.

Il faut donc vérifier les résultats sur la base de test.

Le tableau suivant présente les résultats de prédictions de la charge attritionnelle et grave écrêtée sur la base test : comparaison de la charge écrêtée (en bleu) aux anciennes prédictions du GLM niveau 3 (en gris) et aux nouvelles prédictions niveau NAF grâce aux SHAP values (en vert).

Base de test : charge attritionnelle et grave écrêtée

Code NAF	Charge attri et grave écrêtée	Anciennes prédict (GLM niveau 3)	Nouvelles prédict via SHAP value	Ecart charge obs / anciennes pred Niv3	Ecart charge obs / nouvelles pred NAF
501Z000	30 238	78 436	63 512	-48 197	-33 274
502Z000	109 846	39 144	110 546	70 702	-700
5030000	401 623	457 565	355 292	-55 942	46 331
5040000	128 053	26 147	57 079	101 907	70 974
505Z000	3 396	28 335	3 981	-24 939	-585
Total	673 157	629 627	590 411	43 530	82 746

Les résultats sur la base de test confirment que la nouvelle charge prédite pour chacun des codes NAF via les SHAP value est plus proche de la charge observée que l'ancienne charge prédite via le GLM (sans différenciation des codes NAF).

Les résultats sur la base de test indiquent que l'utilisation des coefficients calculés via la méthode des SHAP values permet de segmenter nos activités au niveau NAF en améliorant nos prédictions.

Puisque nous souhaitons conserver une mutualisation de la sur-crête des sinistres graves au niveau d'activité 3, nous devons maintenant appliquer les coefficients que nous avons obtenus à l'aide des SHAP values en utilisant les résidus attritionnels plus graves écrêtés aux prédictions attritionnelles et graves du GLM et les comparer aux charges observées de sinistres attritionnels et graves.

Le tableau suivant présente les résultats de prédictions de la charge attritionnelle et grave (non écrêtée) sur l'ensemble de la base, cette base n'ayant pas servi au calcul des SHAP values.

Base complète : charge attritionnelle et grave

Code NAF	Charge attri et grave observée	Anciennes prédict (GLM niveau 3)	Nouvelles prédict via SHAP value	Ecart charge obs / anciennes pred Niv3	Ecart charge obs / nouvelles pred NAF
501Z000	126 086	280 316	226 983	-154 230	-100 897
502Z000	386 683	182 052	514 127	204 632	-127 444
5030000	1 550 159	1 999 416	1 552 513	-449 257	-2 355
5040000	610 621	142 318	310 689	468 302	299 932
505Z000	13 256	132 068	18 554	-118 811	-5 298
Total	2 686 805	2 736 170	2 622 866	-49 364	63 939

En appliquant les coefficients calculés à l'aide des résidus écrêtés sur les anciennes prédictions de charge attritionnelle et grave du GLM, les résultats des prédictions sont améliorés car plus proches de la charge observée. Nous pouvons donc effectuer des prédictions meilleures sur le niveau plus fin.

Compte-tenu des résultats obtenus sur l'amélioration des prédictions, le coefficient de la variable niveau de regroupement 3 du GLM peut donc être remplacé par le coefficient initial du GLM de niveau 3 multiplié par le coefficient calculé à l'aide de la méthode des SHAP values pour chacun des NAF.

Ainsi nous pouvons segmenter le tarif des activités :

- NAF 501Z000 : Vente de véhicules automobiles, remorques, semi-remorques et caravanes
- NAF 502Z000 : Entretien et réparation de véhicules automobiles
- NAF 5030000 : Vente en gros et au détail d'équipements automobiles
- NAF 5040000 : Vente en gros et au détail de motocycles
- NAF 505Z000 : Vente au détail de carburants (station-service).

et leur attribuer un coefficient tarifaire spécifique tout en mutualisant la sur-crête de sinistres graves au niveau du regroupement d'activités.

Les activités de vente de véhicules (NAF 501Z000), de vente d'équipements (NAF 5030000) et les stations-services (NAF505Z000) auront un montant de prime pure revu à la baisse par rapport au coefficient initialement calculé en mutualisant ces activités et les activités entretien et réparation (NAF502Z000) et vente de motocycles (NAF5040000) auront un montant de prime pure revu à la hausse, ces activités étant plus risquées.

En utilisant la méthode des SHAP values, issue de la théorie des jeux, il a été possible de segmenter nos activités et de déterminer des coefficients, en améliorant nos prédictions, ce que nous n'avions pas pu faire via le GLM en raison du nombre de modalités trop important de la variable « code NAF ».

CONCLUSION

L'objectif de ce mémoire était de revoir la tarification du produit Responsabilité Civile Entreprises sur le périmètre Industries et Commerces. Il s'agissait d'obtenir un taux à appliquer au chiffre d'affaires des entreprises différencié selon les 300 codes NAF, segmenté selon les variables explicatives. De plus, la méthode utilisée et les résultats devaient pouvoir être facilement expliqués pour présentation aux directions métier.

Les travaux ont porté sur la prédiction de la charge de sinistres attritionnelle et grave en utilisant un Modèle Linéaire Généralisé. L'avantage du GLM réside dans la possibilité de décomposer la prédiction comme le produit des effets de chacune des variables dans un modèle multiplicatif. La prédiction est donc facilement interprétable et il est possible d'obtenir la prime pure très simplement à partir d'une calculette tarifaire. Les GLM constituent donc l'outil de référence utilisé par les actuaires pour la tarification des produits d'assurance non-vie.

Nous avons expliqué et prédit la charge de sinistre attritionnelle et grave de nos contrats entreprises Industries et Commerces via un modèle linéaire généralisé composé de neuf variables sans la variable « code NAF » mais avec une variable « Regroupement d'activités » qui est un regroupement de nos codes activités en 17 classes. Les neuf variables explicatives de notre charge sinistre sont l'activité de l'entreprise (via le regroupement de codes NAF de niveau 3), le nombre de sinistres antérieurs normalisé (ie divisé par le chiffre d'affaires), le chiffre d'affaires de l'entreprise, la sensibilité du code NAF, le fractionnement de la prime, la garantie pollution, le réseau de distribution, la garantie Export USA et la garantie Export hors USA.

L'inconvénient du GLM est qu'il est difficile de modéliser des variables avec un nombre important de modalités : plus de 300 codes activités dans notre cas.

Via le GLM nous avons obtenu une prédiction de charge attritionnelle et grave. Cependant, pour la suite de notre étude, nous souhaitions conserver une partie de la charge grave au niveau du regroupement de niveau 3, qui correspond à la sur-crête, afin de mutualiser cette sur-crête de charge de sinistres graves. Pour mutualiser la sur-crête des sinistres graves, nous avons repris notre GLM sur la même base de données avec les mêmes variables explicatives en ajustant la variable réponse. Nous avons modélisé la charge de sinistre attritionnelle et grave écrêtée à 150 000 euros qui est notre seuil de sinistre grave. Cette nouvelle modélisation a servi à obtenir des résidus. Ces résidus sont égaux à la charge de sinistres attritionnels et graves écrêtés à 150 000 euros moins la prédiction du nouveau GLM multipliée par l'offset. Ils représentent la partie non expliquée par le GLM et correspondent à un montant en euros.

Pour déterminer l'impact de chacun des codes NAF associé à un niveau 3 d'activités, nous avons ensuite utilisé l'algorithme SHAP :

- pour injecter ces résidus dans un modèle XGBoost. Les variables explicatives étaient nos différents codes NAF et la variable à expliquer le résidu du GLM égal à la différence entre la charge observée et la charge prédite
- pour calculer les SHAP values et déterminer la contribution de chaque code NAF dans le niveau 3 de regroupement.

La méthode des SHAP values quantifie pour une activité donnée l'impact de cette activité sur la prédiction qui lui est associée dans le modèle. Ainsi, nous savons quel code NAF a le plus contribué à augmenter ou diminuer notre résidu et donc notre charge de sinistres.

Les SHAP values obtenues sur une base d'entraînement équivalente à 70% des données ont été converties en coefficients. Ces coefficients ont ensuite été appliqués sur les 30% de données restantes pour calculer de nouvelles prédictions de la charge de sinistres et valider l'apport des SHAP values.

Les résultats ensuite obtenus sur la base complète des prédictions attritionnelles et graves en appliquant à chaque code NAF le coefficient calculé via la méthode des SHAP values montrent que la nouvelle prédiction de charge de sinistre attritionnelle et grave est plus proche de la charge observée que la charge prédite par le GLM sur le regroupement d'activités.

En utilisant la méthode des SHAP values, issue de la théorie des jeux, il a donc été possible de segmenter nos activités au niveau demandé tout en améliorant nos prédictions.

En conclusion, à partir d'une modélisation GLM qui segmente le risque selon neuf variables tarifaires dont une variable « Regroupement d'activités » comprenant 17 groupes d'activités à laquelle nous ajoutons une segmentation par activité via les SHAP values, nous répondons à la problématique de départ. Nous avons obtenu un taux à appliquer au chiffre d'affaires des entreprises différencié selon les 300 codes NAF, segmenté selon les variables explicatives.

Au cours des prochains mois, ces travaux vont se prolonger par la mise en place opérationnelle des résultats en commençant par les comparer au tarif appliqué actuellement par les souscripteurs puis en analysant les évolutions de la production selon cette nouvelle segmentation.

ANNEXES

Annexe 1 : RC quasi-délictuelle ou délictuelle :

La RC est dite délictuelle ou quasi délictuelle lorsqu'elle a pour origine un fait qui a causé un dommage à autrui en dehors de tout lien contractuel. Le dommage naît du fait juridique, sans se rattacher à l'exécution d'un contrat. C'est une obligation légale prévue aux articles 1240 à 1244 du Code Civil.

Article 1240 du Code Civil : RC délictuelle « Tout fait quelconque de l'homme qui cause à autrui un dommage oblige celui par la faute duquel il est arrivé à le réparer ».

Article 1241 du Code Civil : RC quasi-délictuelle « Chacun est responsable du dommage qu'il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence ».

Annexe 2 : tableau des variables disponibles pour notre étude :

Le tableau ci-dessous décrit l'ensemble des variables disponibles pour notre étude.

Informations relatives au contrat souscrit (enrichies des données INSEE) :

Variable	Description de la variable et informations apportées	Nbre de modalités	Valeurs manquantes
COT_exposition	Exposition annuelle : quotient du nombre de jours où le contrat est présent en portefeuille et du nombre de jours annuel	Continue]0,1]	Non
ENT_annee_ace	Année police	2011 à 2018	Non
COT_fractionnement	Fractionnement (modalité de paiement de la prime : mensuelle, trimestrielle, semestrielle ou annuelle). Le fractionnement choisi peut révéler la santé financière de l'entreprise.	4	Non
CNT_DIST	Réseau de distribution du contrat (Agent, Courtier ou Salarié commercial). Les entreprises en portefeuille Agents ont généralement un chiffre d'affaires plus faible et les affaires apportées par les Agents sont moins risquées	3	Non
ENT_NAF	Code NAF AXA	302	Non
ENT_segment	Variable indiquant la sensibilité du NAF définie par la direction technique (standard, sensible, exclu, …)	6	Non
ENT_segmentation	Segmentation des entreprises pour analyse de la rentabilité	9	Non
ENT_ACTI_GROUPE	Secteur d'activité de l'entreprise (agroalimentaire, ameublement, …, textile, transports)	17	Non

ENT_nlibniv2_group	Regroupement d'activités AXA Niv2 agrégé (Commerces, Entreprises industrielles, Autres)	3	Non
ENT_nlibniv2	Regroupement d'activités AXA Niv2	11	Non
ENT_nlibniv3	Regroupement d'activités AXA Niv3	17	Non
ENT_nlibniv4	Regroupement d'activités AXA Niv4	38	Non
ENT_nlibniv5	Regroupement d'activités AXA Niv5	122	Non
ENT_nlibniv6	Regroupement d'activités AXA Niv6	248	Non
CNT_tranche_montant_garantie	Montant maximum garanti. Variable discrétisée en 8 tranches	8	16%
COT_TOP_Export_Hors_USA	Booléen indiquant si la garantie Export hors USA est souscrite (29% des contrats ont cette garantie)	2	Non
COT_TOP_Export_USA	Booléen indiquant si la garantie Export USA est souscrite (5% des contrats ont cette garantie)	2	Non
COT_top_export	Booléen indiquant si une des 2 garanties précédentes est souscrite (30% des contrats ont une garantie export)	2	Non
COT_TOP_PJ	Booléen indiquant si la garantie Protection Juridique est souscrite (1% des contrats ont cette garantie)	2	Non
COT_TOP_faute_inex	Booléen indiquant si la garantie faute inexcusable est souscrite (16% des contrats ont cette garantie)	2	Non
COT_TOP_pollution	Booléen indiquant si la garantie Pollution est souscrite (17% des contrats ont cette garantie)	2	Non
COT_TOP_retrait	Booléen indiquant si la garantie Frais de retrait est souscrite (20% des contrats ont cette garantie)	3	Non
ENT_LIB_CJ	Catégorie juridique d'entreprise (INSEE)	30	17%
ENT_Effectif_Eta	Tranche d'effectif de l'établissement (INSEE)	10	17%
ENT_Note_CS	Tranche de note Crédit Safe (INSEE)	5	16%
ENT_age_dirigeant	Tranche d'âge du dirigeant de l'entreprise (INSEE)	7	22%
ENT_age_etablissement	Tranche d'âge de l'établissement (INSEE)	9	13%
ENT_nb_eta_actif_siren	Tranche de nombre d'établissements actifs pour le SIREN (INSEE)	7	14%
ENT_tranche_CA_INSEE	Tranche de chiffre d'affaires (INSEE)	11	20%
ENT_Tranche_chargeant_horsatyp	Tranche (quantiles) de charge sinistre antérieurs sur 5 ans hors atypique	7	Non
ENT_Tranche_nb_anttotal	Tranche de nombre de sinistres antérieurs sur 5 ans	7	Non
ENT_CA_Total_retrait	Chiffre d'affaires annuel de l'entreprise	Continue	9%

Données géographiques :

Variable	Description de la variable et informations apportées	Nbre de modalités	Valeurs manquantes
GEO_DEPART	Département dans lequel est implanté l'entreprise	99	11%
GEO_DEPCOMEN	Commune	9747	77%
GEO_Metropole	Regroupement géographique (Métropole, DOM ou Monaco)	3	10%
GEO_Nouvelle_Region	Région	18	10%
GEO_REGION	Regroupement de régions	5	10%
GEO_Zonier_group	Région	34	10%

Informations relatives aux sinistres pour un contrat et une année :

Variable	Description de la variable et informations apportées	Nbre de modalités	Valeurs manquantes
SIN_surv	Année de survenance du (des) sinistre(s). Cette année de survenance est égale à l'année où le contrat est présent dans notre base (une ligne par année et par contrat)	2011 à 2018	Non
SIN_nb3sinSS	Nombre de sinistres sans suite (Vision N+3) pour cette survenance	Continue	Non
SIN_tot_C3SS	Charge de sinistres sans suite inflatée (Vision N+3)	Continue	Non
SIN_nb3sinattriHSS	Nombre de sinistres attritionnels hors sans suite (Vision N+3)	Continue	Non
SIN_tot_C3AttriHSS	Charge de sinistres attritionnels inflatée hors sans suite (Vision N+3)	Continue	Non
SIN_nb3singraveHSS	Nombre de sinistres graves hors sans suite (Vision N+3)	Continue	Non
SIN_tot_C3GraveHSS	Charge de sinistres graves inflatée hors sans suite (Vision N+3)	Continue	Non
SIN_nb3sinatypHSS	Nombre de sinistres atypiques hors sans suite (Vision N+3)	Continue	
SIN_tot_C3AtypHSS	Charge de sinistres atypiques inflatée hors sans suite (Vision N+3)	Continue	

BIBLIOGRAPHIE

INSEE : Institut Nationale de la Statistique et des Etudes économiques (2022), Les entreprises en France, Edition 2022, https://www.insee.fr/fr/statistiques/6667157

INSEE : Institut Nationale de la Statistique et des Etudes économiques (2022), Nomenclature d'activités française, https://www.insee.fr/fr/information/2406147

FFA : Fédération Française de l'Assurance (Février 2021), L'assurance Responsabilité Civile Générale en 2019

European Commission (2020) Safety_gate_statisticsAndAnualReports_2020_RAPEX_2020

A. Charpentier, (2010-2011). Statistique de l'assurance. STT 6705V Statistique de l'assurance II. partie 1 - assurance non-vie tarification & provisionnement. 3rd cycle. Université de Rennes 1 et Université de Montréal

A. Charpentier. Actuariat IARD - ACT2040 Partie 4 - modèles linéaires généralisés. http ://freakonometrics.hypotheses.org/ Université du Québec à Montréal

C. Clovis, M. Sembona (2016). Caractérisations des modèles multivariés de stables-Tweedie multiples. Thèse de doctorat de Mathématiques. Laboratoire de Mathématiques de Besançon

E. W. Frees, G. Meyers, A. D. Cummings (2014). Insurance Ratemaking and a Gini Index. Journal of Risk and Insurance, Vol. 81, Issue 2, pp. 335-366

L. S. Shapley (1953). A value for n-person games. Contributions to the Theory of Games, pages 307–317

F. Lange (2007), Exploration de la valeur de Shapley et des indices d'interaction pour les jeux définis sur des ensembles ordonnés. Thèse de doctorat de Mathématiques. Université Panthéon-Sorbonne - Paris I

S. Lundberg, S. Lee (2017). A unified approach to interpreting model predictions. NeurIPS (*selected for oral presentation)*

S. Lundberg, G. Erion, S. Lee (2018). Consistent Individualized Feature Attribution for Tree Ensembles. arXiv

U. L. Paris (2020). Repousser les limites d'explicabilité – un guide avancé de SHAP