IRM | institut du risk management BY Institut A

**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 : Ubezzi Robin

Titre : Création d'un score d'appétence client à l'orientation vers un garage agréé et impact sur le tarif technique Automobile

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 :* Generali France

*Signature et Cachet :*

**GENERALI IARD**
Entreprise régie par le Code des Assurances
552 062 663 R.C.S. PARIS
Siège Social : 2, rue Pillet-Will
75009 Paris

*Directeur de mémoire en entreprise :*

*Nom :* Jean-Sébastien Vieu

*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)

# Création d'un score d'appétence client à l'orientation vers un garage agréé et impact sur le tarif technique Automobile

---

**Mémoire d'actuariat**

**Par Robin Ubezzi**

**Directeur de mémoire en Entreprise : Jean-Sébastien Vieu**

---

# Résumé

Dans un contexte marché *Automobile* très concurrentiel, avec un ratio décaissement sur encaissement structurellement supérieur à 100% et des hausses de coûts de pièces détachées supérieures à 6% par an de 2018 à 2020, il est nécessaire d'utiliser toutes les ressources à disposition afin de disposer d'un tarif technique le plus performant possible. Le but de ce mémoire est de mieux segmenter les risques, en enrichissant le modèle de coût moyen Automobiles d'une nouvelle variable discriminante : la probabilité qu'un client répare son véhicule dans un garage agréé du réseau partenaire.

Un maximum d'informations sur le client, son véhicule, les garages agréés ainsi que des données externes, ont été utilisés afin de modéliser cette probabilité. Elle a été calculée en appliquant un modèle de régression logistique, optimisée avec la méthode forward. Les variables les plus discriminantes sont la zone d'habitation du client, les choix antérieurs fait vis-à-vis de l'orientation et le mode de gestion de son sinistre, trois variables qui ne sont pas prises en compte aujourd'hui dans le tarif technique. Une discrétisation a été effectuée sur cette variable, à l'aide de la méthode des Kmeans, afin de créer un score. Cela permet de différencier les clients, et de créer des groupes homogènes d'appétences à l'orientation.

Ce score a été ajouté au modèle de coût de la garantie Responsabilité civile et permet d'en améliorer la performance et donc d'obtenir une segmentation client plus fine.

**Mots clés : Sinistre automobile, IARD, Garantie dommage, Garantie Responsabilité civile, Réseau de garages agréés, Taux d'orientation, Tarif technique, K plus proches voisins, Discrétisation, Kmeans, Modélisation, Régression logistique, Méthode forward, Modèle de coût.**

# Abstract

In a very competitive Automotive market context, with a combined ratio structurally higher than 100% and spare parts cost increases of over 6% per year from 2018 to 2020, it is necessary to use all available resources in order to have the most efficient technical price possible. The purpose of this thesis is to better segment risks, by enriching the Automotive average cost model with a new discriminating variable: the probability that a customer will have his vehicle repaired in an approved garage of the partner network.

A maximum of information on the customer, his vehicle, the approved garages as well as external data, were used to model this probability. It was calculated by applying a logistic regression model, optimized with the forward method. The most discriminating variables are the customer's area of residence, the previous choices made with regard to the orientation and the mode of management of his claim, three variables which are not taken into account today in the technical price. A discretization was carried out on this variable, using the Kmeans method, in order to create a score. This makes it possible to differentiate customers, and to create homogeneous groups of appetites for orientation.

This score has been added to the Civil Liability cover cost model and helps to improve its performance and therefore to obtain a finer-grained customer segmentation.



**Keywords: Motorcar loss, casualty, Damage guarantee, Civil liability guarantee, Network of approved garages, Orientation rate, Technical tariff, K nearest neighbors, Discretization, Kmeans, Modeling, Logistic regression, Forward method, Cost model.**

# Remerciements

Ça y est nous y sommes. Le mémoire est fini.

Un travail qui vient boucler une aventure démarrée il y a plus de 4 ans de cela.

Je pensais avoir un moment de joie intense, de soulagement immense d'arriver au bout du chemin, de clôturer cette partie importante de ma vie professionnelle. Mais non, ce moment n'arrive pas, du moins pas encore et je pense savoir pourquoi : la ligne d'arrivée n'est pas passée. Parce que lorsqu'on accomplit quelque chose d'important ce n'est jamais seul et qu'à cette arrivée toutes les personnes qui ont été là pour vous accompagner dans ce projet sont là, à vous attendre pour partager ce moment avec vous. Ce sont donc ces personnes que je vois tenant une coupe de champagne au loin que je veux remercier.

Tout d'abord, sachant que l'introduction est une partie primordiale, je voudrais remercier,

Mon maitre de stage Monsieur Vieu Jean-Sébastien, non seulement pour son aide et son temps précieux lors de ce projet de longue haleine mais aussi pour être un exemple de professionnalisme et d'empathie que je tente de suivre chaque jour et enfin pour m'avoir permis de rentrer dans le monde de l'assurance par la plus belle des portes, celle de son service.

Un service dans lequel se trouve une certaine Madame Kobana Karélia, pour son soutien, ses nombreuses relectures et remarques toujours à propos, mais évidemment au moins autant pour la bonne humeur qu'elle amène tous les jours au bureau avec elle et qui font qu'un jour chez Generali est toujours une belle journée.

Un service dans lequel se trouvait Monsieur Cepa Vincent, pour son aide précieuse sur ce mémoire de la construction de la base de données jusqu'aux résultats des modèles partagés ensemble mais assurément aussi pour son naturel de tous les instants qui a rendu ce mémoire bien plus vivant.

Et maintenant tout au bout du couloir, je trouve une autre porte toujours ouverte, celle de notre directeur Monsieur Guizouarn Jean-Charles. Pour sa disponibilité lors de ce mémoire, ses remarques toujours pleines de bon sens et surtout pour ce côté humain qui donne à la TA IARD cet air de grande famille.

Dans cette grande famille, j'y trouve Messieurs Michel Remi et Samba Eyrich. Pour m'avoir fait découvrir ce monde intrigant qu'est celui de la tarification Automobile et tout particulièrement celui d'Emblem, et pour l'avoir fait avec calme patience et disponibilité. Et bien entendu Messieurs Geoffroy Romain et Wolf Mathieu, pour m'avoir dirigé dans ce fameux monde qui est devenu magique après quelques explications et beaucoup de rire.

Mais évidemment le soir venu la porte du service se referme et il est temps de rentrer, mais un mémoire ne s'arrête pas à la porte du service il rentre dans celui de la maison et à ce moment-là ma famille a toujours été d'un soutien sans faille.

J'aimerais donc remercier ma mère, évidemment pour ses nombreuses relectures qui ont commencé au balbutiement de ce mémoire et qui finiront bien après son rendu, mais surtout pour être un soutien constant, pour m'avoir toujours donné confiance en moi et poussé à me dépasser, si j'en suis arrivé jusqu'ici elle y est pour énormément.

On dit toujours que le meilleur est pour la fin, et la plupart du temps c'est vrai la conclusion est encore plus importante que l'introduction, mais ce n'est pas toujours le cas, moi le meilleur je l'ai rencontré à mes 19 ans. Je tiens à remercier ma femme qui en plus d'avoir tant relu ce mémoire, qu'elle commence peut-être à le connaitre mieux que moi, me pousse toujours à donner le meilleur de moi-même, et m'a épaulé sur ce projet comme elle m'épaule dans la vie de tous les jours avec force, constance, tendresse, et disons-le, juste de manière parfaite.

Je me sens déjà plus proche de la ligne, hâte de tous vous remercier de vive voix et de fêter cela comme il se doit.

# Table des matières

# Introduction

L'assurance est une merveilleuse invention de l'antiquité, à travers le « secours mutuel » ou encore la « recherche de protection ». Elle permet de bâtir, de se développer, de tenter et tout simplement de prendre des risques en sachant qu'une protection est présente, comme dirait Henry Ford : « New York est la création des assureurs ». De manière plus terre à terre, elle pèse un poids important dans les dépenses de la vie courante d'un français. D'après les chiffres de l'INSEE : (INSEE, 2020), sur la consommation des ménages en 2018, en moyenne 3,2% des dépenses sont dues à des dépenses assurantielles, c'est plus que la part de l'achat de carburant qui est de 2,5%, en tout cas en 2018 ...

Les cotisations sur le marché de l'assurance *Dommages aux biens et de la responsabilité civile* s'élevaient à 59,2 milliards d'euros en 2020, d'après les chiffres de France Assureurs : (FFA, 2021). Ces cotisations sont à mettre au regard des prestations, c'est-à-dire du coût de prise en charge des sinistres, qui étaient de 42,9 milliards d'euros. La branche qui représente le plus, tant au niveau des cotisations que des prestations, est celle de l'automobile. Le coût des prestations pour le seul marché Automobile atteint 17,4 milliards d'euros, soit plus de 40% des prestations globales. Il s'agit d'une branche extrêmement concurrentielle, il est même compliqué d'être rentable dessus comme le montre le niveau du ratio combiné du Marché, structurellement supérieur à 100%. En effet, l'effet de concurrence entraine une stratégie de développement déficitaire : il est fréquent de faire payer aux nouveaux assurés automobile une prime en deçà de leur coût réel, puis d'essayer de la revaloriser année après année afin d'atteindre l'équilibre technique.

Tous les jours, les assureurs français gèrent plus de 32 000 sinistres IARD, et parmi ceux-là plus de 19 000 sont des sinistres Automobiles, (FFA, 2021). Plus particulièrement sur les sinistres matériels, l'augmentation des pièces détachées automobiles est en moyenne supérieure à 5% ces dernières années et tire les coûts de réparation vers le haut, d'après les statistiques SRA. Le projet de loi « Orientation des mobilités » aurait pu commencer à résoudre ce problème en libéralisant le prix des pièces lors de la réparation. Malheureusement, cette loi va libéraliser de « manière progressive » il faudra donc attendre avant d'entrevoir ses premiers effets.

Dans ce contexte difficile pour les assureurs, deux points sont donc primordiaux :

- Maitriser ses coûts à travers les différents gains techniques possibles.
- Disposer de la meilleure segmentation possible afin d'avoir une tarification technique au plus proche des profils de risque.

Ce mémoire porte sur l'analyse de la maitrise des coûts, à travers l'optimisation des coûts des sinistres et plus spécifiquement des sinistres *Dommages*. Le principal levier sur ces sinistres est la négociation des coûts de réparation chez le garagiste. La mise en place d'un partenariat avec des garages permet notamment d'activer ce levier. C'est ce que fait Generali à travers Assercar, son réseau de garages agréés.

Mais l'orientation d'un client vers un de ces garages n'est pas quelque chose d'aisé. Le client est libre du choix de son garage, son assureur ne pouvant lui imposer le garage où faire réparer son véhicule, et ce choix peut dépendre de nombreux facteurs. Il serait très intéressant de savoir à l'avance quel serait le choix de l'assuré, cela pourrait permettre de lui faire payer moins cher sa prime d'assurance puisque lors de la survenance d'un sinistre, celui-ci coûtera moins cher à la compagnie.

C'est l'objectif de ce mémoire : **prédire l'appétence d'un client à aller vers un garage agréé du réseau partenaire et mesurer l'impact de cette nouvelle information sur le modèle tarifaire.**

Dans une époque où la donnée est présente absolument partout et est particulièrement recherchée, des mémoires ont été faits sur la prédiction du comportement de l'assuré, celui de Mme. Callet (Callet, 2015) qui traite de la segmentation du tarification Automobile en fonction des choix internet du client est particulièrement intéressant. Cependant peu de mémoires traitent de la prédiction d'un comportement client lors d'un processus indemnisation. La plupart segmente le tarif en modélisant son comportement extrinsèque vis-à-vis de l'assureur, non pas de son comportement dans la chaîne d'indemnisation.

Le but ici sera d'utiliser des méthodes d'apprentissage supervisé afin de pouvoir segmenter différemment le tarif en s'appuyant sur le levier économique le plus important en indemnisation Automobile à savoir l'orientation d'un client vers un garage partenaire.

Pour atteindre ce but, la première partie contextualisera le sujet, en définissant les principes importants de ce mémoire que sont le déroulement d'un sinistre, le fonctionnement des réseaux de garages agréés dans l'assurance et leur importance, en finissant par décrire le besoin toujours plus important de correctement segmenter sa population de client.

Dans la seconde partie la construction de la base de données sera décrite. La composition de celle-ci sera détaillée tout d'abord au global puis par type de données avant de parler des retraitements nécessaires à sa construction.

Dans la troisième et dernière partie un score d'appétence à l'orientation sera créé et testé sur le tarif automobile. Tout d'abord l'appétence à l'orientation sera modélisée à l'aide d'un modèle de régression logistique, puis cette variable sera discrétisée en sept catégories. Ce scoring sera ensuite appliqué au modèle de coût du tarif automobile et ses impacts sur ce modèle seront analysés.

Tout au long de ce mémoire, une part importante de la chaîne technique de l'assurance Automobile pourra être appréhendée :

- La partie sinistralité, à travers l'analyse de l'orientation du véhicule d'un assuré vers un garage agréé,
- Le lien entre cette sinistralité et les données clients, à travers la modélisation de l'appétence d'un client à l'orientation vers un garage agréé : les données clients en seront ainsi enrichies,
- Cette nouvelle information sur le client permettra d'améliorer la performance du tarif technique, afin notamment de de minimiser le risque d'antisélection et d'améliorer les résultats techniques.

# I. Contextualisation du sujet

Dans cette première partie le but sera d'introduire le sujet, son contexte et ses enjeux. Afin de comprendre le fonctionnement de l'orientation de nos assurés vers un réparateur agréé. La première section détaillera la vie d'un sinistre dommage en assurance automobile. La deuxième permettra de définir ce qu'est un réseau de garages agréés et quels sont les avantages pour chacune des parties de faire partie de ce partenariat. Nous présenterons dans un troisième temps les différents enjeux de l'orientation en insistant sur le gain économique et le suivi opérationnel mis en place pour atteindre les objectifs. Pour finir, le but étant de savoir si l'orientation peut impacter la prime d'un client, nous présenterons le marché concurrentiel qu'est le marché de l'automobile en assurance IARD puis le besoin de segmentation des risques qui en découle.

## I.1. Déroulement d'un sinistre

Dans cette première section nous allons retracer le déroulement d'un sinistre automobile de sa survenance jusqu'à sa réparation et son remboursement par l'assureur. Un sinistre est la réalisation d'un événement prévu au contrat et entraînant la prise en charge financière par l'assureur.

Un « sinistre dommage » sera définie par un sinistre collision avec ou sans tiers, pour lequel l'assuré sera indemnisé. Toutes les autres garanties tels que le bris de glace, le vol ou encore l'incendie seront exclues.

Nous nous intéresserons donc aux deux garanties que sont la garantie **Dommages tous accidents** et la **Responsabilité Civile**. La première est mise en cause lorsque l'assuré est responsable de l'accident ou lorsqu'aucun responsable ne peut être retrouvé (*Exemple : la voiture est rayée pendant la nuit quand elle est stationnée*). La seconde agit lorsque l'assuré n'est pas responsable d'un sinistre et qu'il y a un tiers adverse vers qui se retourner. Voici le schéma du déroulement de la vie de ces sinistres :

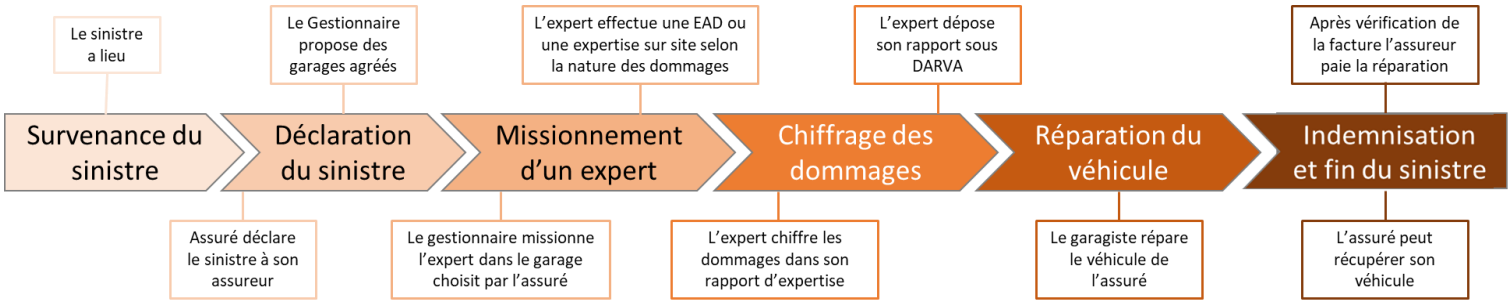



*Figure 1 – Frise de déroulement d'un sinistre*

Lors de la survenance d'un sinistre dommage, l'assuré appelle pour déclarer son sinistre. Le gestionnaire, après avoir vérifié que la garantie mise en jeu était bien acquise par l'assuré lui propose des garages agréés proches de chez lui. L'assuré a ensuite le choix d'emmener son véhicule dans un de ces garages ou dans un autre établissement. Le choix du garage relève, depuis la loi Hamon du 17 mars 2014, du seul choix de l'assuré : « Tout contrat d'assurance souscrit au titre de l'article L. 211-1 mentionne la faculté pour l'assuré, en cas de dommage garanti par le contrat, de choisir le réparateur professionnel auquel il souhaite recourir. Cette information est également délivrée, dans des conditions définies par arrêté, lors de la déclaration du sinistre. », Article L211-5-1 - Code des assurances (Hamon, 2014).

L'assureur gestionnaire du dossier a deux possibilités selon les circonstances du dossier : soit le sinistre concerne deux véhicules assurés chez des assureurs ayant signé la convention **IRSA** (Indemnisation et

Recours entre Sociétés d'Assurance) et l'assureur est obligé de **missionner un expert** ; soit un des deux sinistrés est chez un assureur non signataire de l'**IRSA** (cas assez rare puisque la quasi-totalité des assureurs ont ratifié cette convention) ou alors il n'y a pas de tiers à mettre en jeu et l'assureur peut décider de missionner ou non un expert. Là où le **gré à gré** est répandu en Dommages aux biens, il l'est très peu en Automobile, dans la quasi-totalité des cas un expert est missionné.

Une fois l'expert missionné dans le garage choisi par le client, son rôle est de vérifier la cohérence de la déclaration ainsi que l'acquisition de la garantie par l'assuré puis de procéder à l'évaluation des dommages. Il va selon la nature des dommages, soit se déplacer pour une expertise sur site dans le garage où le véhicule endommagé a été amené, soit effectuer une expertise à distance à partir des photos prises par l'assuré ou par le garagiste. Il doit ensuite, en accord avec le garagiste et avec un outil de chiffrage des pièces de rechange, évaluer le temps de main d'œuvre nécessaire ainsi que le coût des pièces, de la peinture et si nécessaire d'autres éléments sous garantie contractuelle (par exemple une remorque).

Véhicule réparable

| | | | |
|---|---|---|---|
| Montant de l'expertise | | 17851.17 TTC | (14875.98HT) |
| Répartition des chocs : | Choc Arrière | 4694.15 TTC | (3911.80HT) |
| | Choc Avant | 13157.02 TTC | (10964.18HT) |

| Libellé | Vétusté non déduite | | | Vétusté déduite | | | |
|---|---|---|---|---|---|---|---|
| | HT | TVA | TTC | HT | Remise | TVA | TTC |
| Main d'oeuvre | | | | 2337.50 | | 467.50 | 2805.00 |
| Pièces | 11593.48 | 2318.69 | 13912.17 | 11593.48 | | 2318.69 | 13912.17 |
| Ingrédients peintures | 945.00 | 189.00 | 1134.00 | | | | |
| Hors élément SGC | | | | 14875.98 | | 2975.19 | 17851.17 |
| Total général | | | | 14875.98 | | 2975.19 | 17851.17 |

*Tableau 1 - Exemple d'un rapport d'expertise*

Le tableau ci-dessus présente un exemple de chiffrage d'un expert pour un véhicule ayant subi un choc avant et un choc arrière. Le coût toute taxe comprise de ce sinistre est de 17 851,17 € et il se décompose en 2 805€ de main d'œuvre, 13 912,17€ de coût de pièces et 1 134€ d'ingrédient de peinture. La totalité de ce rapport d'expertise sur lequel peut être retrouvé le nombre d'heures de main d'œuvre ou encore le coût de réparation pièce par pièce est présent en *Annexe I : Exemple de rapport d'expertise*.

Une fois le rapport d'expertise rédigé, l'expert le dépose sous *DARVA* afin que l'assureur puisse en prendre connaissance. *DARVA* désigne la plateforme d'échange de données entre les assureurs et leurs partenaires. Le garagiste peut alors effectuer la réparation du véhicule. Après vérification de la cohérence entre la facture du garagiste et l'évaluation du rapport d'expertise Generali procèdera au règlement de l'assuré ou du garage.

## I.2. Réseau de garages agréés

Le choix de garage de l'assuré est libre, mais il existe pour autant, un réel intérêt à l'existence d'un réseau de garages agréés pour les assureurs, les clients et les garages. Un réseau de garages agréés correspond à un ensemble de garages qui a un partenariat avec une ou plusieurs sociétés d'assurance. Le but de cet engagement est de faire bénéficier les trois parties de plusieurs avantages. Le réseau qui nous intéressera par la suite sera le réseau partenaire de Generali : *Assercar*.

### I.2.1. Présentation du réseau Assercar

Afin de mutualiser leurs efforts dans la gestion des prestataires intervenant sur les sinistres IARD, *Aviva* et *Generali* ont créé en 2008 le Groupe d'Intérêt Economique (GIE) KAREO Services. L'objectif de ce GIE est de

mutualiser les ressources, matérielles ou humaines afin de proposer aux clients des services adaptés. Ce groupe est ensuite rejoint par *Pacifica, Sogessur* et *Thélem Assurances* en 2011. Ces trois partenaires disposaient déjà d'un réseau de garages agréés commun de 900 réparateurs, géré par ASSERCAR dont ils étaient actionnaires. Generali et Aviva ont donc logiquement rejoint ce réseau de garages agréés et en sont devenus eux aussi actionnaires. Ensemble, ces assureurs représentaient 4 millions de véhicules assurés soit environ 15% du marché automobile de l'époque.

Le nombre de garages agréés est passé de 900 en 2011 à près de 1400 début 2012 et a ensuite crû progressivement d'une centaine de garages par an.

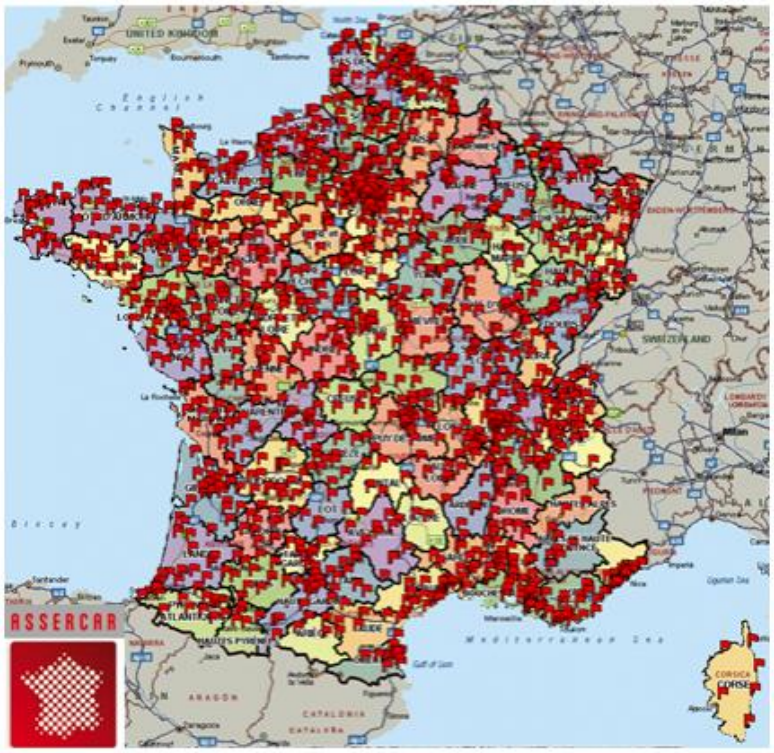



*Figure 2 - Répartition des garages agréés Assercar*

La carte ci-dessus représente la répartition des 1 900 garages du réseau Assercar répartis dans tous les départements français, au 31 décembre 2020.

### I.2.2. Les avantages du réseau

Le réseau Assercar fait partie du GIE *Kareo*, qui a pour but de proposer des solutions innovantes bénéficiant à l'assureur et l'assuré. Les bénéfices pour l'assuré, l'assureur et les garages vont maintenant être décritf.

Tout d'abord, ce partenariat avec le réseau de garage permet de raccourcir les formalités administratives et le délai de traitement du dossier. Le lien se fait plus simplement entre l'expert, le garage et l'assureur ce qui permet une expertise plus rapide et donc une réparation et une récupération de son véhicule plus rapide pour l'assuré. De plus tous les garages du réseau ont des équipements de qualité ainsi que des services supplémentaires pas toujours disponibles dans les autres garages, tel qu'un service de voiturier ou encore de prêt de véhicule. De plus l'avantage du réseau est d'être bien réparti géographiquement et de ne jamais avoir à faire une trop grande distance pour trouver un établissement, en moyenne moins de cinq kilomètres.

Avant la loi du 3 décembre 2020 n°2020-1508 : « Art. L. 211-5-2.-Sont nulles les clauses par lesquelles l'assureur interdit à l'assuré, en cas de dommage garanti par un contrat d'assurance souscrit au titre de l'article L. 211-1, la cession à des tiers des créances d'indemnité d'assurance qu'il détient sur lui. » (LégiFrance, 2020), le client n'avait pas à avancer les frais de réparation s'il allait dans un garage du réseau. Il devait uniquement verser le montant de sa franchise s'il en avait une et le reste du paiement était versé par l'assureur au garage. Depuis cette loi, le client n'a que sa franchise à payer quel que soit le garage choisit.

Le gain pour les garages agréés est tout d'abord un apport de clientèle significatif. Les assureurs s'engageant à envoyer un certain volume dans leur garage. Ils bénéficient également de l'automatisation des flux en étant payé plus rapidement. Et enfin le but du réseau étant toujours d'avoir des garages performants, ils recevront des aides pour monter en gamme. Le réseau pourra proposer des formations aux garagistes ou encore aider à l'achat de nouveaux appareils.

Les assureurs possèdent trois avantages financiers liés à l'envoi des véhicules sinistrés dans des garages du réseau ASSERCAR. Tout d'abord ils ont une réduction sur la main d'œuvre et sur les ingrédients peinture d'environ 30%, négociée tous les ans entre ASSERCAR et ses garages. Le dernier gain est une remise sur le prix global de la facture. Cette remise évolue selon le nombre de véhicules que les assureurs du réseau Assercar envoie dans un même garage chaque mois. Elle est de 4% pour le premier véhicule et évolue jusqu'à atteindre 7% pour le dixième véhicule et pour tous les suivants. Le compte est ensuite remis à zéro chaque début de mois. En cumulant tous ces effets l'économie globale sur une réparation est d'environ 20% du prix de la facture. Nous pouvons comparer les coûts finaux de main d'œuvre et d'ingrédient peinture après ces différentes remises avec le marché de l'assurance grâce aux statistiques marché fourni par la SRA (Sécurité et Réparation Automobiles) (SRA, 2021).

| SRA | SRA 2021 | Generali 2021 Non agree | Generali 2021 Agree |
|---|---|---|---|
| Cout total HT | 1 660 € | 2 052 € | 1 630 € |
| Pièces | 842 € | 1 002 € | 947 € |
| Main d'œuvre | 645 € | 817 € | 525 € |
| MO : Coût horaire | 59,80 € | 72,37 € | 49,28 € |
| MO : Temps | 10,68 | 11,28 | 10,66 |
| Ingrédients Peinture | 173 € | 234 € | 158 € |

*Tableau 2 - Répartition des coûts de réparation marché (SRA) et Generali par poste*

La première colonne du tableau fait partie des statistiques que la SRA met à disposition des assureurs tous les trimestres. Il reprend pour tous les assureurs du marché le coût moyen de réparation hors taxe des sinistres « collision et stationnement », c'est-à-dire les sinistres dommages hors catastrophe naturelle, vol, incendie et bris de glaces. Il divise ensuite cette charge entre les différents postes de coûts de réparation. En moyenne sur le marché le coût de réparation hors taxe pour un sinistre est de 1 660€, environ 50% de cette charge est portée par le coût des pièces, 40% par la main d'œuvre et les 10% restant par le coût des ingrédients peintures.

Sur les deux colonnes de droites nous retrouvons les mêmes informations mais pour la sinistralité Generali. La répartition des postes est quasiment la même pour le SRA et pour les réparations chez les garagistes non

agréés, en revanche cette répartition change chez les agréés. 58 % de la réparation est portée par le coût des pièces, plus que 32% par la main d'œuvre et toujours 10% par les ingrédients peintures. Cet écart s'explique par les négociations sur la main d'œuvre, puisqu'en moyenne le coût horaire de la main d'œuvre est 32% moins cher chez un réparateur non agréé que chez un agréé et même par rapport au marché le coût horaire est 18% plus bas.

Au global le coût moyen de réparation chez Generali est plus haut que sur le marché, ce qui est normal puisque les clients ciblés par Generali sont des chefs d'entreprise ou encore des professionnels.

Le gain sur un sinistre de l'ordre de 20% fait de l'orientation un enjeu majeur, nous allons voir dans la partie suivante quels sont les gains techniques possibles de l'orientation et comment cet indicateur est suivi.

## I.3. Importance de l'orientation

Dans cette partie, nous verrons sur quelle part de la charge sinistre joue l'orientation et quel est le gain potentiel, en millions d'euros, lié à l'orientation des véhicules. Nous décrirons ensuite ce qui est mis en place opérationnellement par Generali pour suivre et améliorer cette orientation du client vers un garage agréé.

### I.3.1. Enjeu technique : Impact sur la charge sinistre

La charge sinistre automobile est le coût total payé par Generali pour l'ensemble de ses sinistres couverts par des contrats Automobile. Elle est divisée en deux grandes familles :

- Les sinistres corporels : Ce sont les sinistres donnant lieu à des victimes. Ils sont couverts par la garantie obligatoire responsabilité civile. Cette garantie protège toutes les personnes à l'intérieur du véhicule de l'assuré ainsi que les personnes que l'assuré pourrait blesser avec son véhicule. Elle prend en charge tous les coûts de santé et de besoin en aide médicale causés par l'accident.
- Les sinistres matériels : Ce sont les sinistres n'impliquant aucune victime. Nous découperons ensuite le coût de ces sinistres en quatre groupes.

| | Corporel | Matériel | | | | | TOTAL |
|---|---|---|---|---|---|---|---|
| | | Réparations | Irréparables | Bris de glace | Autres | Total | |
| Charge 2021 | 125,6 M€ | 151,3 M€ | 55,3 M€ | 32,0 M€ | 8,6 M€ | 247,2 M€ | 372,7 M€ |
| Poids | 33,7% | 40,6% | 14,8% | 8,6% | 2,3% | 66,3% | 100% |

*Tableau 3 - Répartition de la charge Automobile 2021*

Le tableau ci-dessus représente la charge des sinistres Automobile pour l'année de survenance 2021. Nous observons qu'un peu plus d'un tiers de cette charge est portée par les sinistres corporels. La charge matérielle se divise en quatre groupes :

- Le coût des réparations qui pèse pour 40,6% de la charge globale. Il s'agit du coût associé à nos sinistres dommages sur lesquels une réparation est possible. C'est donc sur cette part de la charge que l'orientation d'un véhicule aura un impact.
- Le coût des véhicules irréparables qui pèse pour un peu moins de 15% de la charge globale. Il fait référence au coût associé aux véhicules qui ne pourront pas être réparés ; soit parce que c'est techniquement impossible à cause de pièces qui n'existent plus ou de la trop grande détérioration

d'un véhicule (*Exemple : lors d'un incendie*), soit parce qu'économiquement, réparer le véhicule revient plus cher que la valeur du véhicule avant le sinistre. Dans tous les cas, l'assuré sera remboursé de la valeur de remplacement à dire d'expert de son véhicule avant l'accident.

- Le coût des bris de glaces qui pèse pour 8,6% de la charge globale.
- Tous les autres coûts qui pèsent pour 2,3% de la charge globale. Ils sont composés en grande majorité des frais d'expertises, du paiement à l'assuré adverse du recours conventionné IRSA lorsque notre assuré est responsable et de la récupération de ce recours dans le cas inverse.

L'orientation aura donc un rôle à jouer dans la partie la plus importante en termes de coût qu'est la réparation.

Afin de calculer le gain possible de l'orientation nous allons introduire son principal indicateur de suivi qui est le taux d'orientation. C'est le nombre de véhicules qui vont être réparés dans un garage agréé rapporté au nombre de véhicules qui font l'objet d'une réparation et se définit mathématiquement ainsi :

$$Taux\ d'orientation = \frac{Nombre\ de\ véhicules\ réparés\ dans\ un\ garage\ agréé}{Nombre\ de\ véhicules\ réparés}.$$

Maintenant que cet indicateur est défini, nous allons voir quel impact il a sur la charge sinistre.

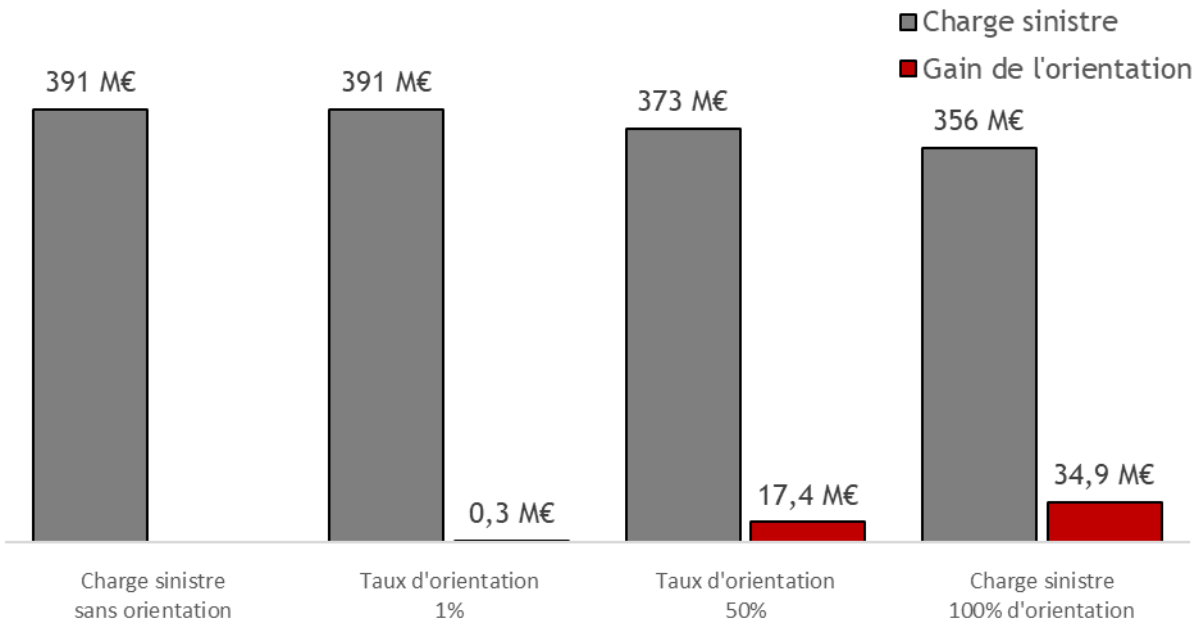



*Figure 3 - Évolution de la charge de sinistre en fonction du taux d'orientation*

L'évolution de la charge sinistre est présentée en fonction du taux d'orientation sur le graphique ci-dessus. Un point d'orientation correspond à un gain d'environ 350K€, ce qui correspond à un gain de 0,1% de sa charge sinistre total. Avec un taux d'orientation de 50% Generali peut économiser plus de 17M€ par an, avec son taux d'orientation en 2021 Generali a économisé plus de 18M€. L'économie totale pouvant être réalisée est de près de 35M€ soit environ 9% de la charge sinistre sans orientation. L'orientation est un levier de gain important et c'est pourquoi Generali compte dessus et le suit de près comme nous allons le voir dans la prochaine partie.

### I.3.2. Enjeu opérationnel

Le fait de conseiller l'assuré sur la marche à suivre après la survenance d'un sinistre, et notamment quel garage choisir pour faire réparer son sinistre, est l'un des principaux rôles de l'assureur. Ce rôle peut être pris par différentes personnes et il dépend tout d'abord de l'intermédiaire choisi lors de la souscription du contrat.

Lorsqu'un contrat est souscrit chez Generali deux types d'intermédiaires différents peuvent être impliqués :

- Le courtier : Il n'est pas affilié à un seul assureur et son rôle est de négocier puis sélectionner les offres des différentes compagnies d'assurances les plus attractives pour son client.
- L'agent Generali : Il est mandataire de Generali et distribue les produits de la compagnie. Son rôle à lui aussi est de négocier les meilleures offres en fonction du profil du client.

Ces deux intermédiaires peuvent jouer à la fois un rôle de commerçant, de conseiller, de négociateur mais aussi de gestionnaire de sinistre. Chez ces deux types d'intermédiaires la gestion du sinistre n'est pas automatique, il existe en effet deux cas de figure :

- La gestion déléguée : Dans ce cas la gestion du sinistre est laissée à la main de l'intermédiaire. C'est donc le courtier ou l'agent qui conseillera l'assuré sur la marche à suivre lors de son sinistre.
- La gestion plateforme Generali : Dans ce cas, c'est le gestionnaire sinistre de Generali travaillant à Tours ou à la plateforme de Reims qui gèrera les sinistres. Les gestionnaires de Tours et de Reims s'occupent respectivement des dossiers des agences et des courtiers, qui ont confié leur gestion. La plateforme de Reims est un GIE créée entre Generali et ses agences qui permet de laisser la gestion des sinistres à des spécialistes de la gestion sinistre et ainsi de libérer du temps aux agents pour de la conquête commerciale.

Réussir à conseiller un client dans le choix de son garage n'est pas toujours facile et comme nous allons l'analyser ci-dessous tous les gestionnaires n'ont pas la même réussite dans cet exercice.

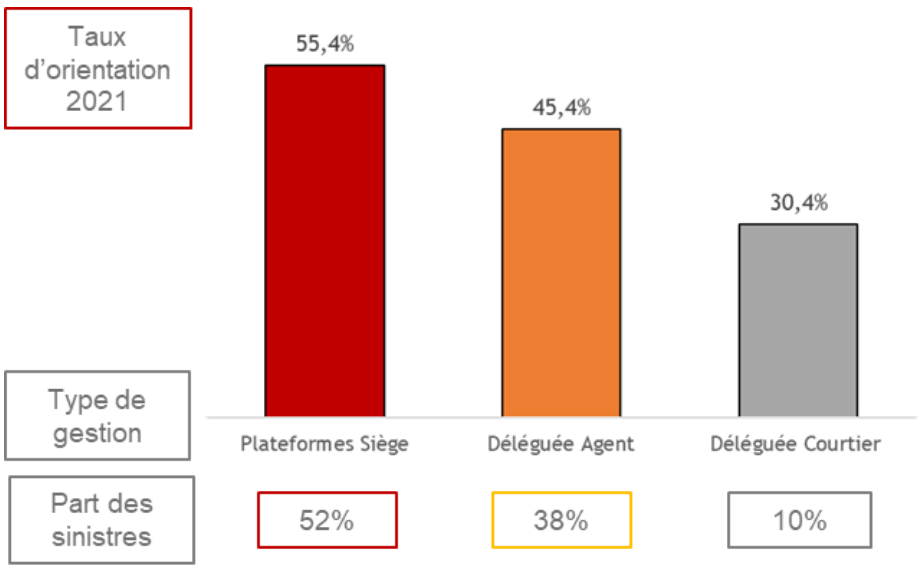



*Figure 4 - Taux d'orientation 2021 par type de gestion de sinistre*

Le graphique ci-dessus décompose le coût d'orientation par type de gestionnaire de sinistre. A gauche, les plateformes Siège sont composées des gestionnaires de la plateforme de Reims et du siège. Ils ont un taux

d'orientation de 55,4%, soit dix points de plus que les sinistres gérés par les agents et vingt-cinq de plus que ceux gérés par les courtiers. Cette différence s'explique non seulement par l'expertise des gestionnaires des plateformes Siège mais aussi par les mesures mises en place par le Siège. Les gestionnaires des plateformes Siège sont objectivés sur leurs performances d'orientation : ce taux est suivi mensuellement par gestionnaire et envoyé à leur manager pour permettre un meilleur pilotage. Quant aux agents, ils ont un taux plus bas mais eux aussi ont des avantages à orienter leurs sinistres, leur capacité à proposer des rabais à leurs clients dépend en partie du niveau de leur taux d'orientation. Pour ce qui est des courtiers, ils sont sensibilisés à cet aspect mais n'étant pas des mandataires Generali, il est plus compliqué d'améliorer leur performance. Heureusement, la part de sinistres gérés chez les courtiers est faible, de l'ordre de 10%. Le principal levier d'amélioration du taux d'orientation est le transfert des sinistres gérés par les agents vers la plateforme de Reims, en plus de l'amélioration de la performance de ces plateformes qui réussissent à atteindre des taux de 60% en début 2022. Jusque-là cette stratégie fonctionne bien puisque la part des sinistres délégués par les agents est passée de 57% en 2018 lors des débuts de cette plateforme à environ 35% en début 2022, et elle permet une amélioration du taux d'orientation.

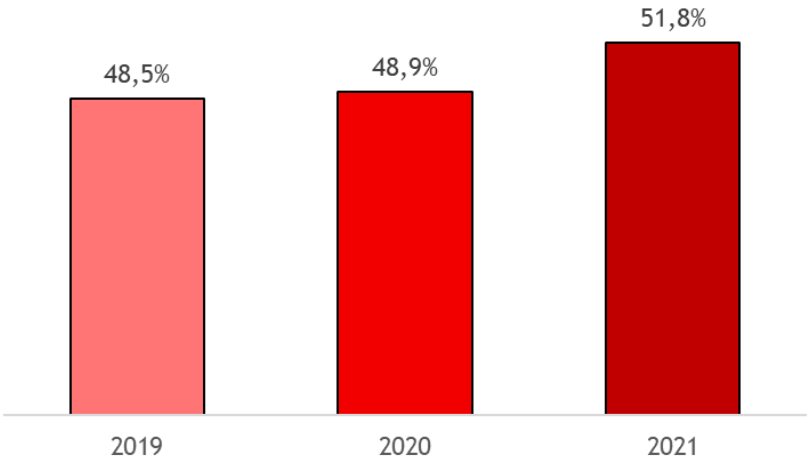



*Figure 5 - Taux d'orientation par année de réparation*

Le graphique ci-dessus met assez bien en avant l'augmentation du taux d'orientation d'année en année. Bien qu'il y ait eu qu'une légère augmentation entre 2019 et 2020, dû à la fermeture des garages lors de la crise du Covid-19, le taux d'orientation s'est amélioré de 3,3 points en deux ans. Il a dépassé pour la première fois les 50% et a atteint 51,8% à la fin de l'année 2021. De plus, nous verrons dans le prochain chapitre que le taux d'orientation s'améliore tous les ans depuis 2015.

## I.4. Le besoin de segmentation en assurance

Dans la partie précédente l'intérêt de l'orientation dans l'optimisation du coût des sinistres a été démontré. Nous allons à présent expliquer en quoi cette information pourrait impacter la tarification. Afin d'introduire le sujet, nous commencerons par décrire le contexte du marché de l'Automobile sur ces dernières années puis nous enchainerons sur la manière dont est calculé le tarif aujourd'hui et comment l'orientation pourrait le faire évoluer.

### I.4.1. Contexte marché

Le marché de l'assurance Automobile est un marché extrêmement concurrentiel et tout particulièrement ces dernières années. Les coûts de réparations sont en constante augmentation, ce qui est principalement dû à l'augmentation des prix des pièces chez les constructeurs automobiles. Par ailleurs, la loi de libéralisation des pièces pour les véhicules de plus de dix ans qui entrera en vigueur le 1er janvier 2023, ne limitera que partiellement cette hausse.

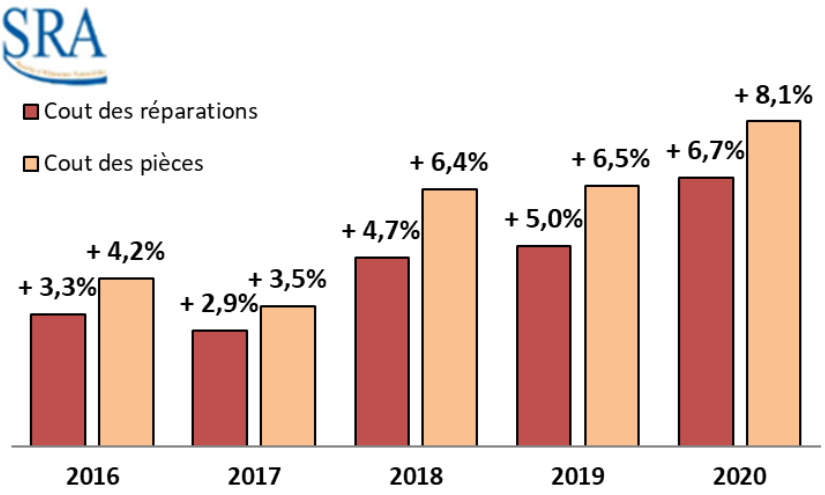



*Figure 6 - Évolution des coûts de réparation et des pièces de 2016 à 2020*

Le graphique ci-dessus décrit l'évolution des coûts de réparation et des coûts des pièces détachées lors de ces réparations. Nous observons que la hausse des coûts est passée d'environ +3% sur les années 2016 et 2017 à près de +5% en 2018 et 2019, pour passer à +6,7% en 2020. Cette constante évolution du coût de réparation est directement portée par la hausse des coûts des pièces qui est passée de +3,5% en 2017 à +8,1% en 2020. Cette hausse est particulièrement présente chez certaines marques telles que Peugeot et Citroën qui ont augmenté le coût de leurs pièces d'environ 9% sur l'année 2020. Vous trouverez ces statistiques et celles des autres constructeurs dans l'*Annexes II : Communication SRA janvier 2021*.

Ces multiples hausses obligent les assureurs à revoir leurs tarifs mais même ainsi il est compliqué pour eux d'être rentable sur la branche Automobile.

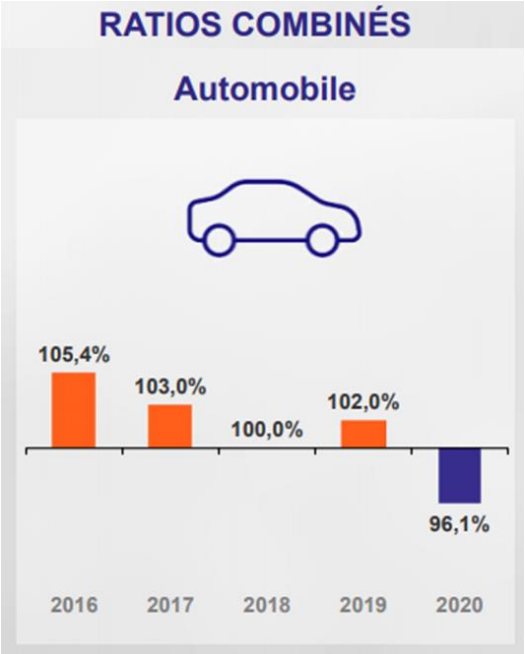



*Figure 7 - Évolution du ratio combiné en Automobile de 2016 à 2020*

Le graphique ci-dessus est extrait d'un communiqué de presse de la Fédération Française de l'Assurance (FFA, 2021). Il présente l'évolution du ratio combiné sur le marché français de 2016 à 2020. Le ratio combiné désigne le rapport entre la totalité des paiements (frais de gestion, commissions versées, remboursement des sinistres …) et les encaissements de primes. Un tel ratio supérieur à 100% indique que sur l'année, l'assureur a plus dépensé qu'il n'a encaissé de primes, et en excluant les gains possibles lors de placements financiers il est donc déficitaire. Nous observons donc qu'entre 2016 et 2019 les assureurs n'ont pas fait de profit sur l'assurance Automobile. Il n'y a qu'en 2020 malgré une forte hausse des coûts de réparation que les assureurs ont eu un ratio combiné inférieur à 100% grâce à une forte baisse de fréquence due aux différents confinements subis lors de la crise de la Covid-19.

### I.4.2.L'orientation une réponse à un contexte tendu

Comme nous avons pu le voir, le marché de l'automobile est particulièrement concurrentiel, c'est pourquoi en plus de savoir maîtriser ses coûts de sinistres, il faut réussir à apprécier ses risques. Nous traiterons ici uniquement du tarif technique, c'est à dire le tarif du modèle permettant de calculer la prime pure d'un assuré. Cette prime pure est le montant payé par l'assuré permettant de couvrir la survenance de ses sinistres, tout effet de rabais, de geste commercial, de coût de gestion du sinistre, de rémunération des intermédiaires ou de comparaison à la concurrence ne seront pas traités. Pour calculer cette prime il faudra trouver les variables les plus discriminantes pour le risque d'un assuré, ce qui permettra d'avoir une bonne segmentation. Il sera ainsi possible de réduire l'effet d'antisélection qui consiste à garder les clients qui vont payer trop peu pour un risque élevé et à l'inverse perdre les bons clients qui eux paient trop cher. La tarification d'une prime pure se distingue en deux modèles : un modèle de fréquence et un modèle de coût. Le but d'un modèle de fréquence est de prédire la probabilité qu'un assuré ait un sinistre. Ce modèle dépendra par exemple du bonus-malus, ou encore du nombre de sinistre antérieur de l'assuré. Dans ce mémoire seul le modèle de coût sera abordé. Nous supposons que l'appétence d'un client Generali à aller dans un garage agréé impact le coût de ses sinistres, mais pas le nombre de ses sinistres.

A ce jour, Generali utilise principalement six variables tarifaires dans ses modèles de tarification des garanties *Dommages* et *Responsabilité Civile* pour les affaires nouvelles. Deux de ces variables concernent l'assuré, les quatre autres concernent le véhicule :

- L'Age de l'assuré
- La classe de prix SRA du véhicule
- L'ancienneté du véhicule
- Le type de carrosserie du véhicule
- La classe de coût à la réparation SRA du véhicule
- Un zonier de coût selon l'adresse de l'assuré, ce zonier n'a été rajouté que récemment dans le modèle de tarification c'est pourquoi nous ne l'avons pas eu à disposition lors de la création de la base pour ce mémoire.

Ce tarif peut ensuite connaitre de légers ajustements selon la version ou encore l'usage du véhicule assuré. Toutes ces variables permettent d'avoir une bonne vision du coût d'un sinistre mais nous allons chercher à savoir si le choix de faire réparer son véhicule chez un garage agréé dépend seulement de ces variables. Dans le cas où de nouvelles variables entreraient en jeu il faudrait les prendre en compte dans le tarif, car comme nous l'avons vu précédemment, le fait d'orienter permet d'économiser 20% du coût du sinistre. Il semble en effet normal qu'une appétence plus grande à l'orientation entraine une prime moins élevée et à l'inverse une aversion à l'orientation devrait entrainer une hausse. Le taux d'orientation a d'ailleurs commencé à être pris en compte dans le cadre du renouvellement des tarifs de 2022. Une partie de la majoration des contrats dépendait en partie du taux d'orientation de l'intermédiaire chez qui était le client.

# II. Création et analyses de la base de données

Dans cette partie la manière dont a été construite la base de données sur laquelle porte l'étude sera détaillée. Cette base est constituée de données internes à Generali mais aussi de données externes en *open data*. Nous définirons le périmètre qui a été retenu pour la création de cette base. Puis nous décrirons chacune des sept bases de données en expliquant quelle est sa provenance et quelles sont les variables retenues. Une fois cette base de données créée nous présenterons les différents traitements de qu'il a fallu mettre en place afin d'avoir des variables complètes et les plus discriminantes possibles.

## II.1. Description de la base

### II.1.1. Base d'étude

Notre base d'étude est une consolidation des bases suivantes :

- Les sinistres survenus de 2015 à 2019
- Les rapports d'expertise de ces véhicules expertisés
- Les données contrats de ces sinistres
- Les données des clients ayant souscrit ces contrats
- Les données des véhicules de ces clients
- Des données externes de densité de population en Open data
- Des données spécifiques à notre réseau de garages agréés.

Les années de survenance de 2015 à 2019 ont été retenues car les informations sur les variables clients sur ces années étaient les mêmes et avaient un bon taux de remplissage. De plus la méthode pour distinguer si un véhicule était orienté était la même sur ces cinq années ce qui a conforté notre choix. Avant cette année des données étaient manquantes sur la liste de nos garages partenaires. L'exercice 2020 a été exclu en raison de la crise sanitaire, les tendances d'orientation étaient biaisées par la fermeture des garages lors des mois de confinement.

Nous n'avons gardé pour notre étude, que les trois typologies de sinistres suivantes : les dommages sur les véhicules stationnés, sur les véhicules roulants et sur les dommages lors d'accidents entre deux véhicules ou plus. Ces trois distinctions ne seront pas conservées dans nos variables explicatives et plus généralement aucune information connue après la survenance du sinistre car le but est d'évaluer l'appétence d'un client à orienter son véhicule vers un garage agréé quel que soit la circonstance du sinistre. C'est pour cette même raison que les sinistres irréparables seront inclus dans notre base d'étude même s'ils ne permettent pas d'effectuer de gains financiers lors de l'orientation.

La base avant retraitement des données manquantes est constituée de 215 099 observations pour 32 variables explicatives et une variable à expliquer : le *Top Agréé*. Ce top est composé de deux modalités : *oui,*

si le véhicule est réparé dans un garage agréé ou *non*, dans le cas inverse. Le taux d'orientation moyen sur cette base est de 45,4%.

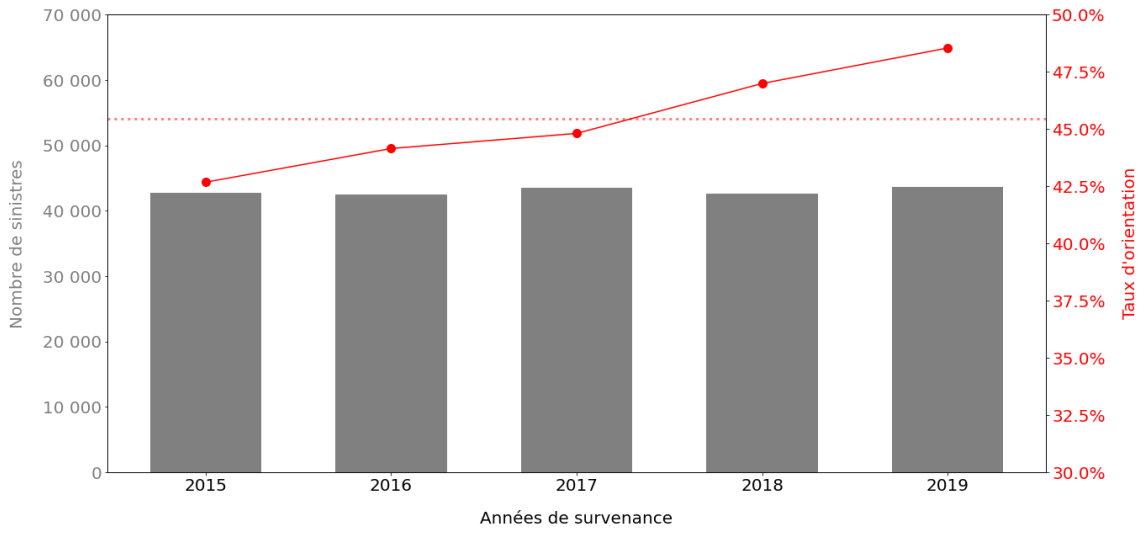



*Figure 8 - Taux d'orientation et nombre de sinistres par année de survenance*

Sur le graphe ci-dessus nous avons un premier aperçu du volume de sinistres par an et du taux d'orientation associé. Les bâtons en gris représentent le nombre de véhicules sinistrés (échelle de gauche) et la courbe rouge représente le taux d'orientation (échelle de droite). Le taux d'orientation moyen a été ajouté en pointillé rouge pour pouvoir comparer chaque valeur avec la moyenne. Il est important de noter que le portefeuille sinistré Auto-mono est stable dans le temps, il y a un peu plus de 40 000 sinistres expertisés par an. De plus, nous pouvons observer à travers ce graphique l'amélioration chaque année du taux d'orientation avec un gain de plus d'un point par an en moyenne, puisqu'il évolue de 42,7% d'orientation en 2015 à 48,5% en 2019. Le taux d'orientation moyen est lui dépassé en 2018.

Dans la suite de ce chapitre nous allons décrire les variables utilisées en les répartissant selon différentes catégories :

- Jointure : les variables servant de jointures, d'une base à une autre.
- Cible : la variable à expliquer.
- Explicative : les variables permettant de discriminer le taux d'orientation.
- Base : les variables ne servant qu'à la création de la base ou à la création de nouvelles variables et qui disparaitront au moment de la modélisation.
- Nouvelle variable : les variables explicatives créées à partir d'autres variables.

Il est à noter, que ce graphique sera le seul dans la partie dédiée à l'analyse qui a été réalisé à partir de la base globale. Nous allons dès à présent séparer la base de données en deux : un jeu d'entrainement et un jeu de test. Le but de cette séparation est de savoir comment les futurs modèles vont se comporter sur de nouvelles observations. Ainsi, les modèles seront entrainés sur le jeu d'entrainement puis les résultats testés sur le jeu de test. Nous séparons notre base dès à présent afin d'éviter le biais « d'espionnage de données ». C'est le fait d'analyser les données sur la base de test, de trouver une structure intéressante sur ces données et de choisir ensuite un modèle adapté à ces données. En général la séparation de la base de données est de 80% pour l'entrainement et 20% pour le test, c'est ce que nous ferons. (Géron, 2017)

## II.2. Les bases de données

### *II.2.1. Données sinistres*

La première base présentée est celle des sinistres Automobiles. Presque aucune variable de cette base ne sera conservée mais elle est essentielle dans la création de la base finale.

| Nom des variables | Description | Type de variable |
|---|---|---|
| Numéro de sinistre | Code alphanumérique permettant d'identifier un sinistre de façon unique | Jointure |
| Numéro de contrat | Code alphanumérique permettant d'identifier un contrat de façon unique | Jointure |
| Année d'enregistrement | Année d'enregistrement du sinistre | Explicative |
| Année de survenance | Année de survenance du sinistre | Explicative |
| Nature du sinistre | Information permettant d'identifier la nature du dommage | Base |
| Département du sinistre | Département de survenance du sinistre | Base |
| Produit | Codification permettant d'identifier à quel produit est rattaché le contrat | Base |

*Tableau 4 - Liste et description des variables de la base de données sinistres*

Nous récupèrerons tous les sinistres de nature dommages sur les produits Automobile du particulier et du professionnel, survenus entre 2015 et 2019. Nous pourrons ensuite croiser avec les bases présentées ci-dessous.

### *II.2.2. Données rapports d'expertise*

La deuxième base qui permet de créer le socle de notre base d'étude est celle des rapports d'expertise. Une fois croisée avec la base des sinistres, nous ne garderons que les numéros de sinistre qui se retrouvent dans les deux bases. Le taux de perte est inférieur à 1% de la base lors de ce croisement. La base de rapports d'expertise est remplie manuellement par les experts et malheureusement dans certains cas les données dans le champ *numéro de sinistre* sont erronées.

| Nom des variables | Description | Type de variable |
|---|---|---|
| Numéro de sinistre | Code alphanumérique permettant d'identifier un sinistre de façon unique | Jointure |
| Top agréé | Top oui/non indiquant si le rapport d'expertise a eu lieu dans un garage agréé Assercar | Cible |
| Taux d'orientation antérieur | Taux d'orientation du client sur la période précédant le sinistre, entre 2015 et la date de survenance du sinistre | Nouvelle variable |

*Tableau 5 - Liste et description des variables de la base de données rapports d'expertise*

Cette base est celle qui a le moins de colonnes mais c'est l'une des plus importantes car elle contient la variable cible ainsi que le taux d'orientation sur les sinistres antérieur qui est très explicative. Comme nous pouvons le voir sur la figure ci-dessous cette variable n'est pas souvent remplie, cependant elle est discriminante pour le comportement du sinistre.

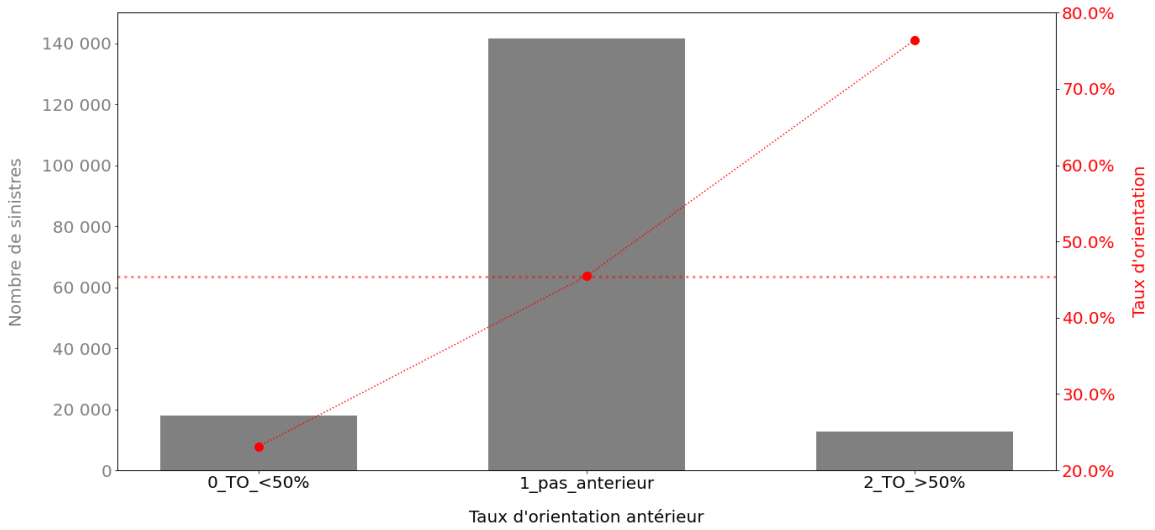



*Figure 9 - Taux d'orientation et nombre de sinistres en fonction du taux d'orientation antérieur*

La variable taux d'orientation sur les sinistres antérieurs a été regroupée en trois valeurs :

- Les clients n'ayant pas eu de sinistre *Dommage* antérieurs. Ce groupe de clients possède un taux d'orientation très proche de la moyenne de la base d'entrainement et donc apprendra peu de chose sur l'appétence à aller dans un garage agréé, à la différence des deux autres. Environ 80% des données ont cette valeur.
- Les clients ayant eu au moins un sinistre *Dommage* antérieur et étant allé dans un garage agréé moins d'une fois sur deux. Ces clients n'iront qu'une fois sur quatre en moyenne dans un garage agréé lors d'une prochaine réparation.
- Les clients ayant eu au moins un sinistre *Dommage* antérieur et étant allé dans un garage agréé une fois sur deux, ou plus. Ces clients iront quant à eux plus de trois fois sur quatre en moyenne dans un garage agréé lors d'une prochaine réparation.

### *II.2.3. Données clients*

Les données clients seront intéressantes à analyser. Elles donneront un premier aperçu du comportement d'un client Generali par rapport à l'orientation.

| Nom des variables | Description | Type de variable |
|---|---|---|
| Numéro de contrat | Code alphanumérique permettant d'identifier un contrat de façon unique | Jointure |
| Code postal | Code numérique permettant d'identifier le code postal dans lequel habite l'assuré | Explicative |
| Code INSEE | Code alphanumérique permettant d'identifier la commune dans laquelle habite l'assuré | Explicative |
| Longitude X assuré | Point x indiquant la longitude du domicile de l'assuré | Base |
| Latitude Y assuré | Point y indiquant la latitude du domicile de l'assuré | Base |
| Ancienneté permis | Temps entre l'obtention du permis et la date de l'accident | Explicative |
| Sexe | Sexe de l'assuré | Explicative |
| Profession | Profession de l'assuré | Explicative |
| Age | Age de l'assuré à la survenance du sinistre | Explicative |
| Sinistre dans son département | Top oui/non permettant de savoir si le sinistre a eu lieu dans le département de résidence de l'assuré | Nouvelle variable |

*Tableau 6 - Liste et description des variables de la base de données clients*

Certaines variables comme l'âge de l'assuré ou encore le code INSEE donne des premières indications très intéressantes mais elles ont besoin de retraitement pour donner leur pleine mesure. Elles seront traitées dans le prochain chapitre. Pour l'instant, analysons les résultats de la variable sexe.

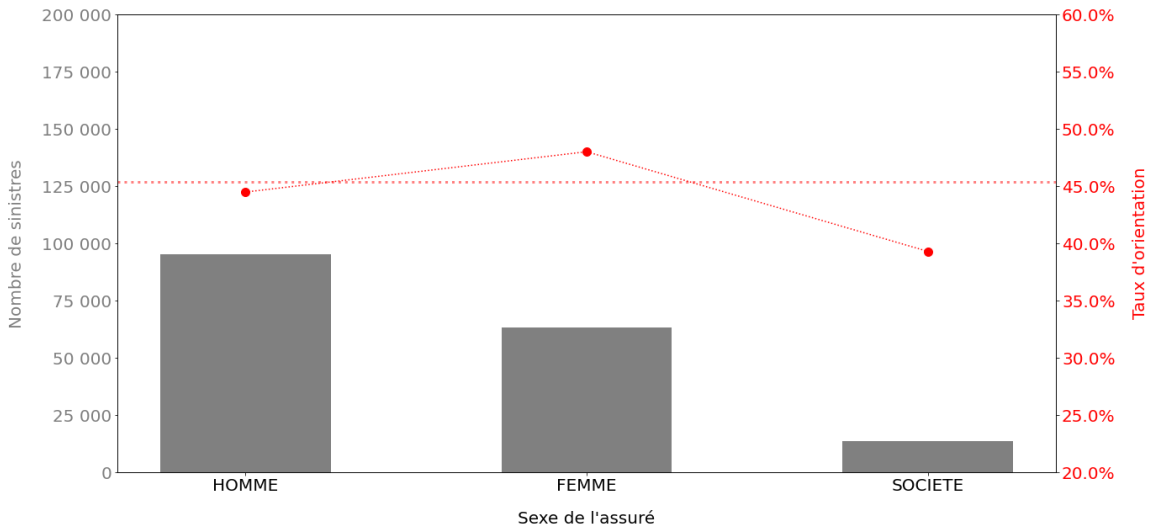



*Figure 10 - Taux d'orientation et nombre de sinistres en fonction du sexe de l'assuré*

Nous remarquons que notre portefeuille sinistré est composé de plus d'hommes que de femmes et qu'un peu moins de 10 % de notre base correspondent à des contrats professionnels sur lesquels nous n'avons pas d'information sur le sexe. En moyenne les femmes paraissent avoir plus tendance à aller dans des garages agréés, là où les hommes et surtout les professionnels, semblent y être plus réticents.

### *II.2.4. Données contrats*

En complément des données clients, les données contrats apportent de la profondeur sur l'appétence au risque du client ou encore sur sa fidélité. Tout comme les données du client, les informations sur le contrat sont vues à la date de survenance du sinistre.

| Nom des variables | Description | Type de variable |
|---|---|---|
| Numéro de contrat | Code alphanumérique permettant d'identifier un contrat de façon unique | Jointure |
| Code APSAD | Code alphanumérique permettant d'identifier un modèle de véhicule de façon unique | Jointure |
| Ancienneté du contrat | Temps depuis la signature du contrat chez Generali | Explicative |
| Formule garantie | Type de formule choisie lors de la souscription : RC seule, ajout des garanties Vol et incendie ou ajout du dommage tout accident | Explicative |
| Leasing | Top oui/non indiquant si le véhicule est en leasing | Explicative |
| Offre 8000 km | Top oui/non indiquant si le contrat ne couvre l'assuré que sur 8000km lors de l'année de souscription | Explicative |
| Nombre de sinistre antérieur responsable | Codification permettant de savoir le nombre de sinistre dont l'assuré est responsable lors des trois dernières années | Explicative |

| | | |
|---|---|---|
| Nombre de sinistre antérieur non responsable | Codification permettant de savoir le nombre de sinistre lors desquels l'assuré est non responsable ces trois dernières années | Explicative |
| Bonus-Malus | Codification permettant de savoir quel est le coefficient Bonus-Malus | Explicative |
| Type d'intermédiaire | Type d'intermédiaire par lequel le client a souscrit son contrat (Agent ou Courtier) | Base |
| Mode de gestion | Information sur la gestion du sinistre (gestion intermédiaire ou gestion siège) | Base |
| Gestion | Croisement du mode de gestion et du type d'intermédiaire en trois modalités : Gestion plateformes, déléguée Agents, déléguée Courtiers | Nouvelle variable |

*Tableau 7 - Liste et description des variables de la base de données contrats*

Cette partie traitera de la variable type de gestion appliquée au sinistre. Comme nous l'avons vu au chapitre précédent, elle semble être une variable particulièrement discriminante. Les trois gestions possibles pour cette variable sont les suivantes :

- Déléguée Agents : Le sinistre est géré par un agent.
- Déléguée Courtiers : Le sinistre est géré par un courtier.
- Gestion plateformes : Le sinistre est géré par les équipes de gestionnaires de Generali.

Ces trois cas diffèrent énormément comme nous avons pu le voir dans le chapitre précédent. Sur la seule année 2021 il y avait un écart de vingt-cinq points entre la gestion plateformes et la gestion déléguée chez les courtiers.

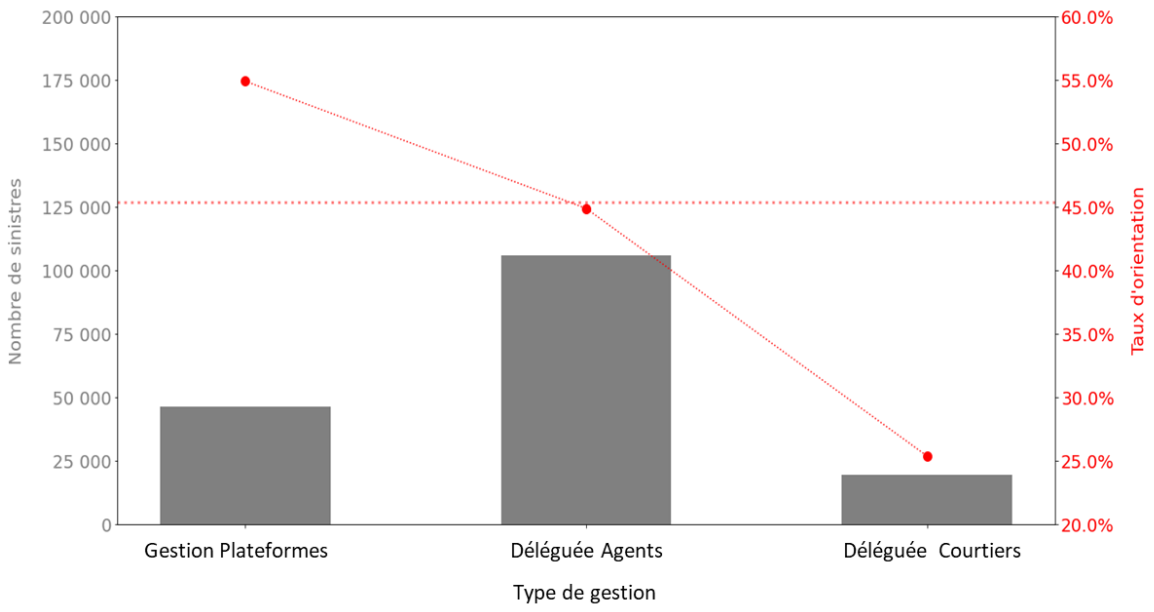



*Figure 11 - Taux d'orientation et nombre de sinistres en fonction du type de Gestion*

Nous observons que les statistiques sur notre base de 2015 à 2019 exacerbent encore plus les écarts qui existent entre les différentes gestions en 2021. Les sinistres en gestion plateformes ont toujours un taux d'orientation avoisinant les 55%, ceux délégués chez les agents ont un taux d'orientation autour de la moyenne de notre base d'entrainement à 45% alors que les courtiers en local sont proches de 25%, soit une

différence de trente points entre les deux extrêmes. Au niveau de la répartition de nos sinistres la différence avec 2021 est d'autant plus notable, puisqu'environ 15% des sinistres gérés par les courtiers, 30% en gestion interne et plus de la moitié des sinistres qui sont géré par les agents en locaux.

### *II.2.5. Données véhicules*

Afin d'analyser plus en détail le comportement des clients, les caractéristiques de leurs véhicules sont particulièrement intéressantes. Ces données étant pour la plupart déjà prises dans le modèle de tarification, notre but sera de voir si nous pouvons ressortir un effet autre que celui qui joue sur le tarif. Ces données ont été récupérées pour certaines dans les bases contrats et pour d'autres dans la base du Système d'Immatriculation des Véhicules (SIV).

| Nom de variables | Description | Type de variable |
|---|---|---|
| Code APSAD | Code alphanumérique permettant d'identifier un modèle de véhicule de façon unique | Jointure |
| Ancienneté du véhicule | Temps entre la fabrication du véhicule et la date du sinistre | Explicative |
| Ancienneté d'acquisition du véhicule | Temps entre l'acquisition du véhicule et la date du sinistre | Explicative |
| Genre du véhicule | Libellé distinguant les véhicules en trois catégories, les véhicules particuliers, les camionnettes et les autres (poids lourds par exemple) | Explicative |
| Série limitée | Top oui/non indiquant si le véhicule fait partie d'une série limitée | Explicative |
| Dernier tarif | Coût en euro du tarif le plus récent pour chaque véhicule neuf | Explicative |
| Puissance administrative | Unité administrative calculée, en partie, à partir de la puissance réelle du moteur | Explicative |
| Classe de prix | Unité permettant de séparer les véhicules en fonction de leur prix à neuf | Explicative |
| Classe de réparation | Unité permettant de séparer les véhicules en fonction du coût estimé de leur réparation | Explicative |
| Groupe SRA | Unité permettant de séparer les véhicules en fonction de leur dangerosité, nous la détaillerons plus ci-dessous | Explicative |

*Tableau 8 - Liste et description des variables de la base de données véhicules*

La variable qui sera détaillée dans cette sous partie est le groupe SRA. Il permet de donner un indicateur liant pour un véhicule : sa puissance réelle, sa masse, sa vitesse maximale et un indicateur de sa sécurité globale. Le groupe SRA ayant remplacé le groupe APSAD, il commence à partir de la valeur 20 afin de ne pas les confondre. Sa formule est la suivante :

$$Groupe\ SRA = 20 + \left(27.88 * \frac{Puissance\ réelle}{Masse\ vide + 200kg} + 0.00283 * (Vitesse\ \max - 130Km/h) + \frac{1}{13} * PTAC\right) * (1 + 0.02 * Note\ de\ conception)$$

La note de conception est une note allant de -2 pour un véhicule excellent et de +2 pour un mauvais véhicule. Elle prend en compte les innovations améliorant la sécurité active et passive du véhicule.

La variable devrait être décroissante avec des véhicules plus sécuritaires et moins puissant plus simple à orienter.

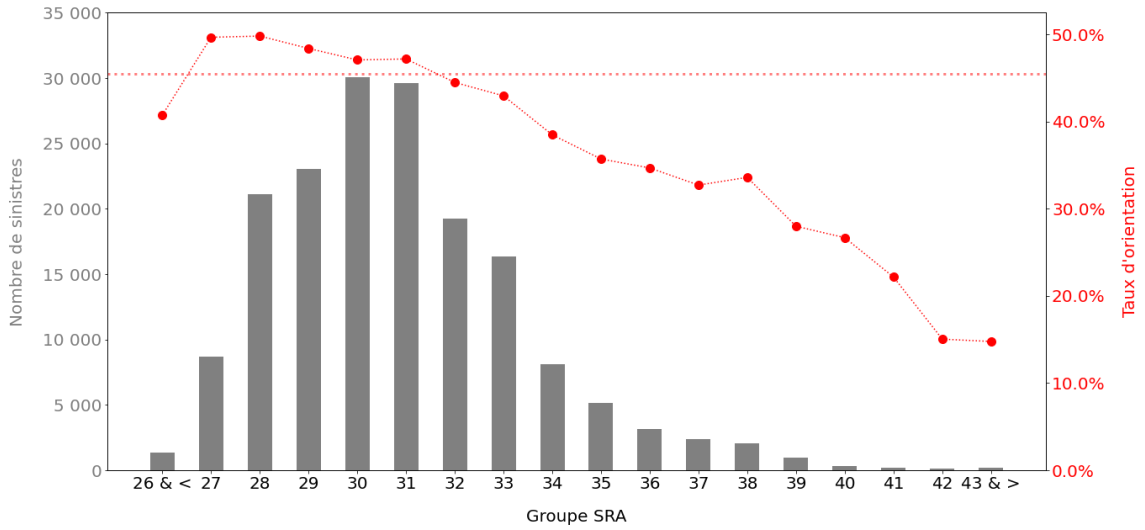



*Figure 12 - Taux d'orientation et nombre de sinistres en fonction du Groupe SRA*

L'analyse de ce graphique confirme presque notre intuition. Il n'y a que la première valeur, celle qui englobe les groupes SRA inférieurs à 26 qui échappe à la tendance décroissante. Il est possible que ces véhicules qui sont peu puissants soient des véhicules anciens ou de collection et donc plus compliqués à orienter. Pour le reste il y a une décroissance qui démarre du groupe 27 dans lequel les véhicules sont orientés dans un cas sur deux et nous finissons pour les véhicules avec une valeur supérieure ou égale à 43 avec un taux d'orientation autour de 15%. Le taux d'orientation est très différent entre les premiers et les derniers groupes SRA. Cependant, la grande majorité des sinistres se trouve entre les groupes SRA 27 et 34 qui ont un taux d'orientation qui varie peu, entre 50% et 40%. Il n'est donc pas certain que cette variable soit importante par la suite.

### *II.2.6. Données externes*

Comme nous le verrons par la suite, les données géographiques jouent un rôle essentiel dans l'explication du comportement du client dans le choix de l'orientation. C'est pourquoi nous sommes allés chercher dans les bases INSEE des données sur la densité de population ainsi que sur le type de zone géographique de l'assuré. La question était savoir si le fait d'avoir plus de garage non agréé dans son département ou dans son code postal impactait le choix du client pour l'orientation du véhicule. Le nombre de garage a été comptabilisé par code postal afin d'être comparé avec notre nombre de garages agréés.

| Nom de variables | Description | Type de variable |
|---|---|---|
| Code postal | Code numérique permettant d'identifier le code postal | Jointure |
| Superficie code postal | Superficie en mètre carré du code postal | Base |
| Garages CP | Nombre de garages par code postal | Base |
| Densité garage CP | Densité de garage par code postal, est égale à la division du nombre de garage par la superficie du code postal | Nouvelle variable |
| Densité garage DPT | Densité de garage par département, est égale à la division du nombre de garage par la superficie du département | Nouvelle variable |
| Densité population | Densité de population par code postal | Explicative |

| | | |
|---|---|---|
| Statut commune | Code indiquant si la commune est une ville-centre, une banlieue, une ville isolée ou hors unité urbaine | Explicative |
| Tranche unité urbaine | Code indiquant la tranche de taille de l'unité urbaine à laquelle appartient la commune selon le recensement de la population 2017 | Explicative |

*Tableau 9 - Liste et description des variables de la base de données externes*

La tranche d'unité urbaine permet de connaitre le type de zone dans laquelle se trouve la commune, il y en a huit :

- La commune est hors unité urbaine.
- La commune appartient à une unité urbaine entre 2 000 et 4 999 habitants.
- La commune appartient à une unité urbaine entre 5 000 et 9 999 habitants.
- La commune appartient à une unité urbaine entre 10 000 et 19 999 habitants.
- La commune appartient à une unité urbaine entre 20 000 et 49 999 habitants.
- La commune appartient à une unité urbaine entre 50 000 et 99 999 habitants.
- La commune appartient à une unité urbaine entre 100 000 et 199 999 habitants.
- La commune appartient à une unité urbaine entre 200 000 et 1 999 999 habitants.
- La commune appartient à l'unité urbaine de Paris.

La tranche de 200 000 à 1 999 999 habitants contenant trop de données nous l'avons transformé en rajoutant les zones urbaines de grandes villes ayant plus de 5 000 sinistres dans notre base. Nous pouvons observer ci-dessous les résultats pour cette variable :

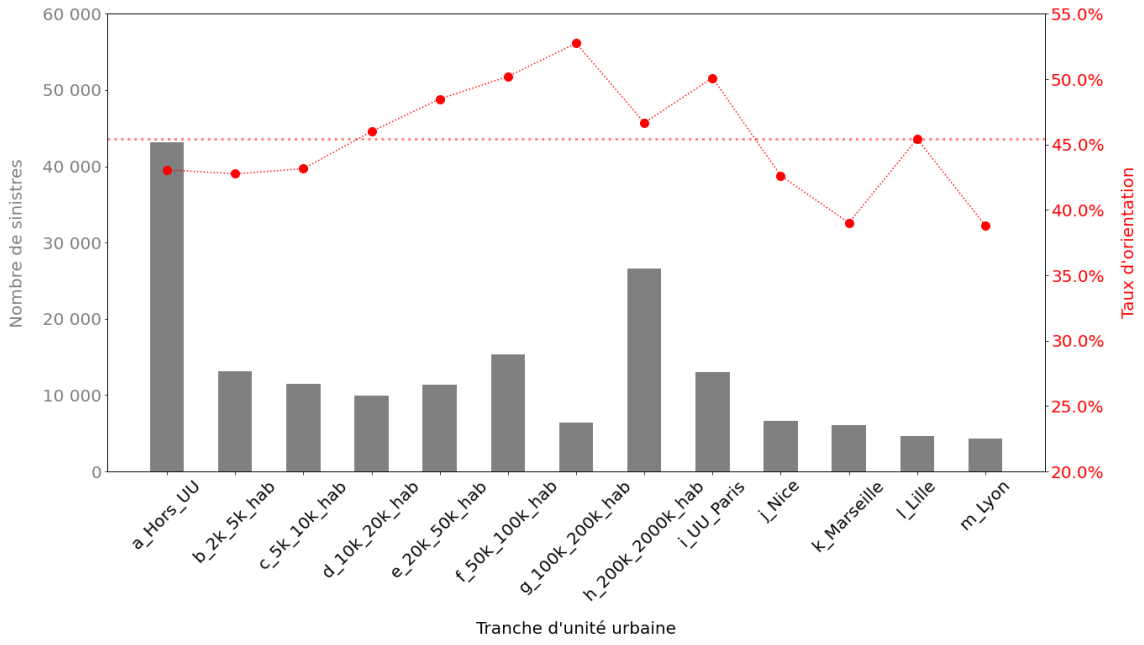



*Figure 13 - Taux d'orientation et nombre de sinistres en fonction de la tranche d'unité urbaine*

Le taux d'orientation est dans un premier temps croissant par rapport à la taille de la zone urbaine de la commune de résidence de l'assuré. Les habitants des zones urbaines inférieures à 10 000 habitants ont un taux d'orientation autour de 42.5% alors que ceux entre 50 000 et 100 000 habitants sont plus proches de 53%. Puis à partir de la zone urbaine de plus de 200 000 habitants nous observons des comportements assez erratiques. Les assurés habitant Paris vont une fois sur deux dans un garage agréé alors qu'à Lyon ou Marseille ils y vont moins de 40% du temps. Cette variable est intéressante car elle met en avant la différence de comportement des assurés par type de zones urbaines mais aussi à l'intérieur des grandes villes françaises.

### *II.2.7. Données réseau de garages*

La dernière source d'information qui va être explorée est celle traitant des données concernant nos garages agréés. Nous nous sommes donc servis des informations contenues dans nos bases, en les rapprochant de nos données contrats ainsi que de données externes afin de ressortir les indicateurs qui paraissaient les plus pertinents.

| Nom de variables | Description | Type de variable |
|---|---|---|
| Identifiant garage | Code alphanumérique permettant d'identifier un garage du réseau Assercar de façon unique | Base |
| Code postal garage | Code numérique permettant d'identifier le code postal dans lequel habite l'assuré | Base |
| Longitude X garage | Point x indiquant la longitude du garage agréé | Base |
| Latitude Y garage | Point y indiquant la latitude du garage agréé | Base |
| Densité CP garages agréés | Densité de garages agréés par code postal, est égale à la division du nombre de garages agréés par la superficie du code postal | Nouvelle variable |
| Densité Département garages agréés | Densité de garages agréés par département, est égale à la division du nombre de garages agréés par la superficie du département | Nouvelle variable |
| Taux de garages agréés CP | Part de garages agréés par rapport au nombre de garages sur ce code postal | Nouvelle variable |
| Taux de garages agréés Département | Part de garages agréés par rapport au nombre de garages sur ce département. | Nouvelle variable |
| Distance assuré garage agréé le plus proche | Distance entre l'adresse de l'assuré et le garage agréé le plus proche | Nouvelle variable |

*Tableau 10 - Liste et description des variables de la base de données réseau de garages*

Une des premières intuitions que nous pouvons avoir concernant le choix d'un client à aller vers un garage agréé est le fait qu'il en soit géographiquement proche. C'est pourquoi la distance entre l'adresse de l'assuré et le garage agréé le plus proche a été mesurée en géocodant notre portefeuille de contrats sinistrés et notre réseau de garages agréés. La distance entre l'adresse de coordonnées ($x_{assuré}$, $y_{assuré}$) et chacune des adresses des garages agréés au moment du sinistre ($x_{garage}$, $y_{garage}$) a ensuite été calculée, puis la distance minimum a été retenue.

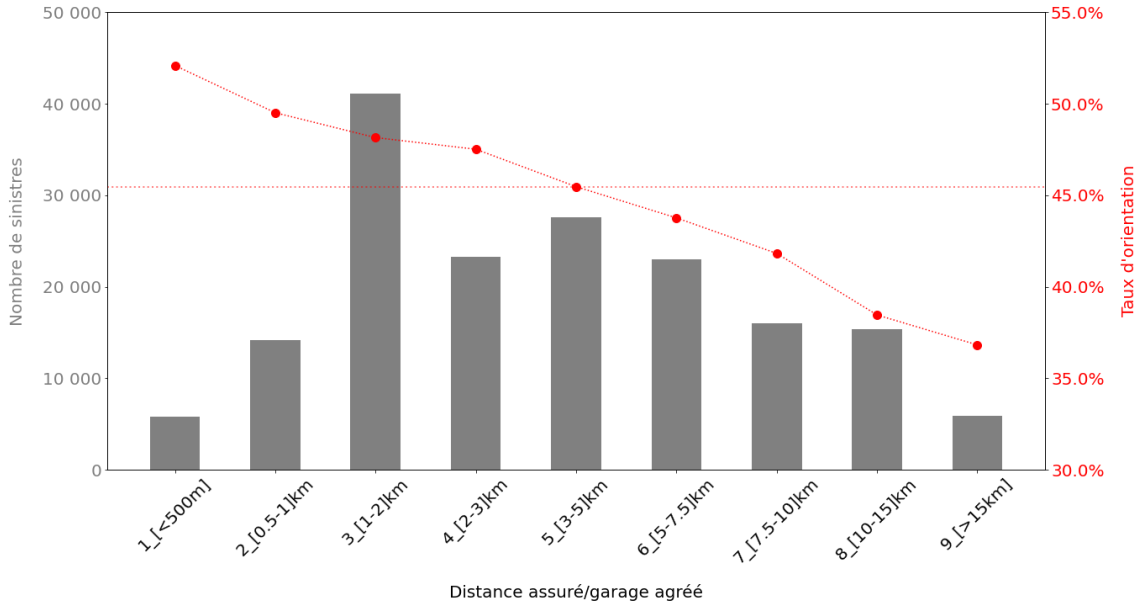



*Figure 14 - Taux d'orientation et nombre de sinistres en fonction de la distance entre l'assuré et le garage agréé le plus proche*

Une décroissance nette de l'orientation d'un client vis-à-vis de la distance du garage agréé le plus proche est constatée. Un assuré qui habite à moins d'un kilomètre d'un garage partenaire aura plus d'une chance sur deux d'y aller que lorsqu'il habite à plus de 15 kilomètres, il n'aura alors plus qu'environ 37% de chance seulement d'y aller. Cette décroissance est assez marquée, nous allons vérifier qu'elle ne soit pas trop contenue dans d'autres variables que nous avons déjà dans notre base. Intuitivement, la variable qui semble la plus discriminante sur le fait d'habiter proche d'un garage agréé est celle de vivre dans une grande ville ou une plus petite.  Pour vérifier nous allons utiliser une boite à moustache avec en abscisses, les tranches de zones urbaines et en ordonnées, la distance entre l'adresse de l'assuré et le garage agréé le plus proche.

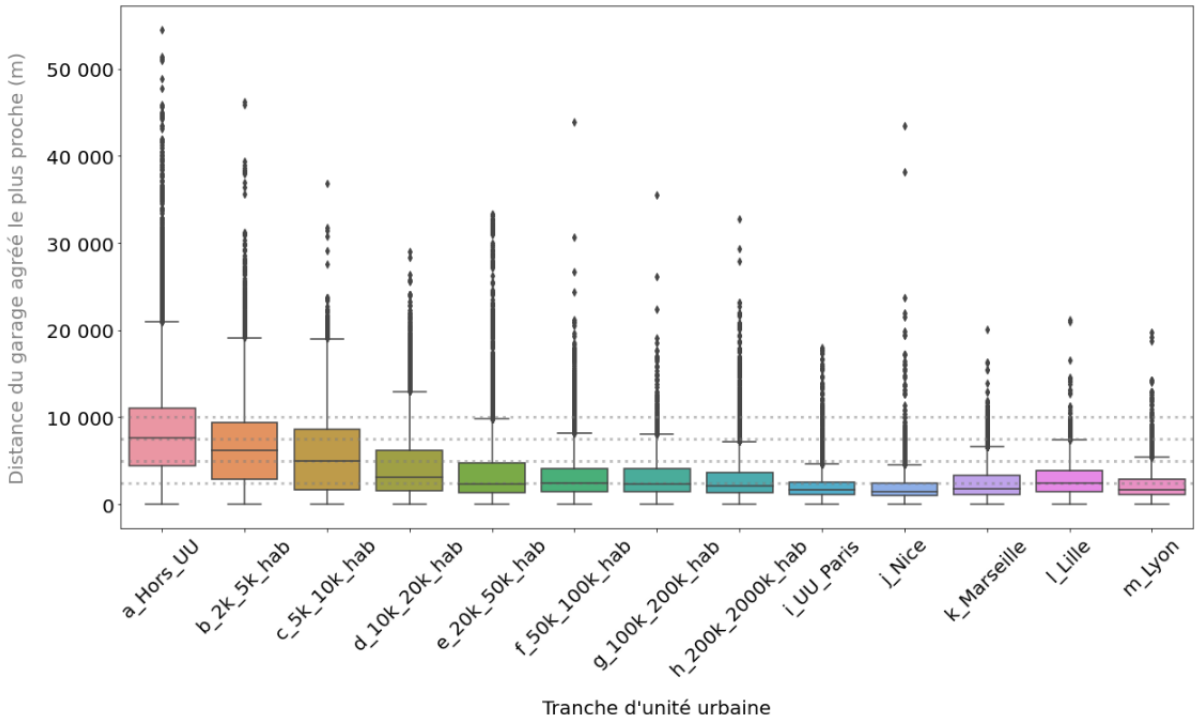



*Figure 15 -Boxplot de la distance entre l'assuré et le garage agréé en fonction de la tranche de zone urbaine*

Dans le graphique ci-dessus, plus la zone urbaine est peuplée plus la chance d'être loin d'un garage agréé est faible. Par exemple, hors zone urbaine plus d'un quart de la population a un garage à plus de 10 kilomètres de chez lui alors qu'à Paris ou à Nice 75% de la population habite à moins de 2,5 kilomètres d'un garage agréé. Maintenant que nous avons observé cette différence de comportement, il est intéressant de voir comment se comporte l'orientation et la répartition des sinistres en fonction de ces deux variables c'est ce que cherchera à révéler le graphe ci-dessous. Il sépare notre population entre les zones urbaines de plus de 20 000 habitants et celles de moins.

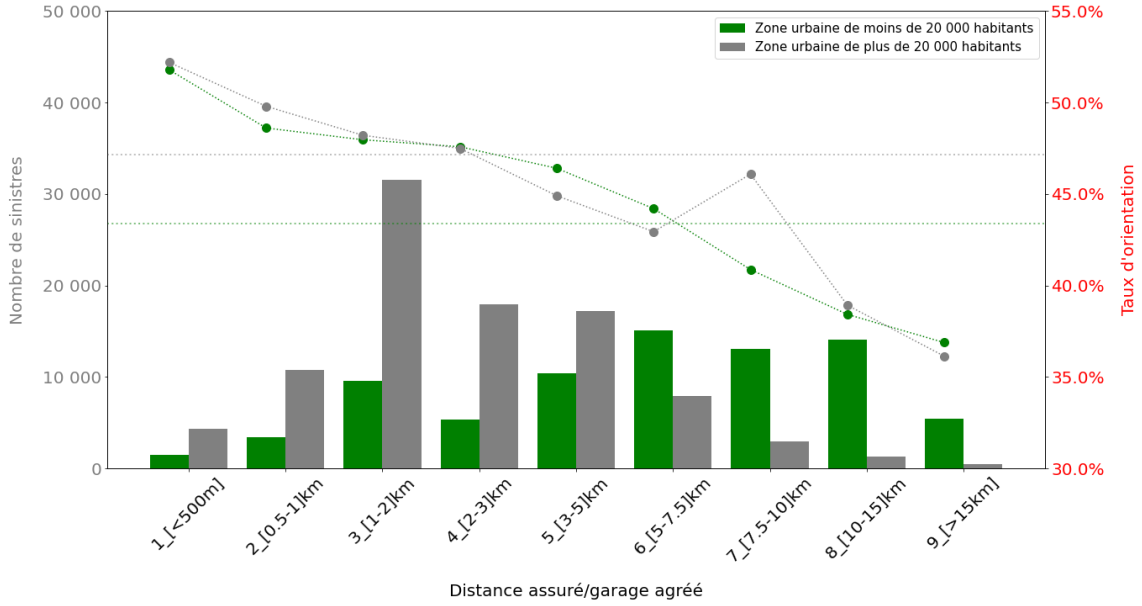



*Figure 16 -Analyse de la distance selon l'importance de la zone urbaine*

La distinction des deux populations est très nette sur ce graphique : les sinistres en zone urbaine plus peuplée ont un taux d'orientation autour de 47,5% alors que ce taux n'est que de 44% pour le reste. Pour les zones de plus de 20 000 habitants la plupart de nos assurés sinistrés ont un garage agréé entre 1 et 5 kilomètres de chez eux alors que pour la deuxième population la plupart se trouve entre 3 et 15 kilomètres. Au-delà de 15 kilomètres 5 000 assurés sont recensés pour la zone de moins de 20 000 habitants contre moins de 500 pour l'autre zone. L'autre point intéressant de ce graphique est la proximité des courbes de taux d'orientation. Exceptée la tranche de 7,5 à 10 kilomètres sur laquelle nous avons peu de données pour la zone grise, pour toutes les autres tranches le taux d'orientation des deux zones est très proche. Cela signifie que quelle que soit la zone, c'est principalement la distance au garage qui importe, l'écart de 3,5 points d'orientation entre ces deux populations s'explique donc par l'éloignement plus marqué à un garage agréé dans les zones peu peuplées plutôt qu'à un comportement différent des habitants en fonction de leur zone urbaine.

Nous avons vu dans cette première sous partie comment avait été créée notre base de données et quelles variables principales la constituaient. Nous allons maintenant aller un peu plus loin dans l'analyse des bases de données en présentant le regroupement des variables numériques ainsi que de la corrélation des variables.

## II.3. Retraitement des données

Dans la partie précédente, nous avons parlé de la base de données brutes mais avant de pouvoir commencer à exploiter ces données avec des modèles, il s'agit de les rendre intelligibles. C'est pourquoi dans cette partie nous présenterons les différents retraitements effectués sur les variables.

### II.3.1. Données manquantes

Dans toutes les bases de données, le premier problème qui se pose est le traitement des données manquantes. Dans nos travaux, nous avons utilisé différentes méthodes en fonction de la variable et c'est ce que nous allons présenter dans cette première sous partie.

#### *II.3.1.1. Problème des données manquantes et premières solutions*

Dans certains cas le fait de ne pas avoir la donnée est une information en soit : par exemple, l'âge du conducteur n'était jamais rempli pour les véhicules de société. Nous possédions déjà cette information avec la variable sexe mais dans le cas contraire, il aurait fallu garder l'information. La difficulté c'est que, à part dans ces cas-là, le fait de garder des valeurs manquantes regroupe des populations qui ne sont pas forcément les mêmes. Par exemple si nous avons des données manquantes sur la marque du véhicule pour deux assurés ça les regrouperait au sein de la même catégorie alors que l'un peut avoir une Ferrari et l'autre une Clio. En somme, garder des valeurs manquantes renvoie donc une fausse information et peut biaiser nos modèles.

Une première méthode est le remplacement des valeurs manquantes par la médiane. C'est une méthode qui peut s'avérer utile lorsqu'il y a peu de valeurs manquantes et que la variable est numérique. La médiane étant la valeur la plus neutre ce traitement revient à dire que pour ce sinistre cette variable ne donnera pas d'information.

La méthode la plus basique pour gérer les données manquantes est la suppression de l'observation. Cette méthode s'applique lorsque :

- Le nombre de données manquantes est très faible et ne peut pas être rempli par une autre méthode. Par exemple nous avions quatre valeurs manquantes pour le Code Insee et pour le code Postal, ces variables n'étant pas numérique la méthode de la médiane ne fonctionnera pas.
- L'observation a trop peu d'informations remplies. Par exemple nous avions des rapports d'expertises dans nos bases avec un numéro de sinistre faux. Le croisement avec les autres bases n'étant pas possible ce sont des observations que nous n'avons pas gardées dans notre base.

Enfin quand trop d'informations sont manquantes il est possible d'utiliser des modèles mathématiques pour « deviner » quelle valeur devrait être là. Dans notre cas la méthode choisie est celle des K plus proches voisins.

Le tableau ci-dessous présente les variables avec des valeurs manquantes, le taux de vide associé et la méthode utilisée pour les combler.

| | **Code INSEE** | **Code Postal** | **Anciennete du contrat** | **Age du conducteur** | **Anciennete de permis** | **Classe de réparation** |
|---|---|---|---|---|---|---|
| Taux de vide | 0,002% | 0,002% | 3,609% | 7,979% | 7,979% | 0,357% |
| Méthode utilisée | Suppression des données | | K plus proches voisins | | | Médiane |

*Tableau 11 - Tableau des variables avec des valeurs manquantes et solutions utilisées*

Nous avons 4 données que nous supprimons car elles sont manquantes à travers le Code Insee et le code Postal, la base finale a donc 215 095 observations.

#### *II.3.1.2. Une méthode plus mathématique : les K plus proches voisins*

##### II.3.1.2.1. Définition de la méthode

L'algorithme des K plus proches voisins fait partie des méthodes d'apprentissage supervisé. Le principe de ces méthodes est de classer des nouvelles données à partir de données étiquetées, à la différence de l'apprentissage non supervisé qui regroupe des données entre elles sans aucun à priori. Cette méthode est qualifiée de « *Lazy Learning* » car le modèle ne retient rien lors de la phase d'apprentissage, il cherche juste les k plus proches voisins et sélectionne ensuite la valeur moyenne de ces k voisins. Nous détaillerons ci-dessous, l'algorithme qui est mis en place pour tout d'abord tester notre modèle des k plus proches voisins puis l'appliquer (Chavent, 2021-2022).

1. Séparer notre jeu de données sans valeur manquantes en un jeu d'entrainement et un jeu de test sur lequel nous vérifierons les résultats.
2. Identifier les variables corrélées avec la variable à approcher et les retirer afin de ne pas faire tourner notre modèle sur toute la base. Il est ainsi possible de gagner beaucoup de temps de calcul pour peu de perte de résultat.
3. Standardiser ces données afin qu'elles aient toutes le même poids.
4. Choisir un nombre de voisins qu'il faudra ensuite optimiser, nous commencerons par trois.
5. Détecter les k voisins les plus proches de chaque nouvelle donnée.

Pour ce faire, il est nécessaire de définir la notion de distance, afin de calculer quel voisin est le plus proche. La distance utilisée dans ce but se nomme la distance euclidienne, noté $d$. Pour un nombre de variables *N,* nous cherchons à minimiser la distance entre chaque point de la base *p* de coordonnées ($p_1$, …, $p_N$) et notre nouveau point *n* ($n_1$, …, $n_N$) la distance :

$$d(p,n) = \sqrt{\sum_{i=1}^{N}(p_i - n_i)^2}.$$

D'où l'importance de standardiser les valeurs afin que chaque variable ait le même poids dans le calcul de la distance.

6. Allouer la moyenne des k voisins les plus proches à chaque nouvelle valeur.

A ce stade de notre algorithme, les résultats de notre méthode des k plus proches voisins sur notre base test sont disponibles. Puisque nous connaissons les valeurs prises par la variable estimée, nous allons pouvoir mesurer le pouvoir prédictif de notre algorithme.

7. Calculer l'écart entre les valeurs prédites et les valeurs réelles, nous le ferons en calculant *le Root Mean Square Error*. C'est la racine de la moyenne des erreurs au carré. Soit *n* le nombre d'individus et $\varepsilon_i$ l'erreur entre la valeur prédite et celle estimée :

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} {\varepsilon_i}^2}.$$

8. Optimiser ensuite notre nombre de voisins en choisissant le couple RMSE et nombre de voisins le plus faible. Pour ce faire nous utiliserons la méthode du coude, c'est-à-dire regarder le point qui

correspond à la plus grosse cassure sur la courbe qui a pour ordonnée le RMSE et en abscisse le nombre de voisins utilisé.

9. Valider notre modèle en comparant son RMSE avec celui d'autres modèles tels que le remplissage par la moyenne ou la médiane.
10. Appliquer le modèle obtenu aux données manquantes et analyser ses effets sur la variable.

#### II.3.1.2.2. Exemple : L'âge du conducteur

Comme nous l'avons vu dans le tableau précédent, cette méthode de remplissage a été appliquée sur plusieurs variables numériques. Nous allons illustrer sa mise en œuvre sur l'âge du conducteur qui est une variable discriminante pour le taux d'orientation.

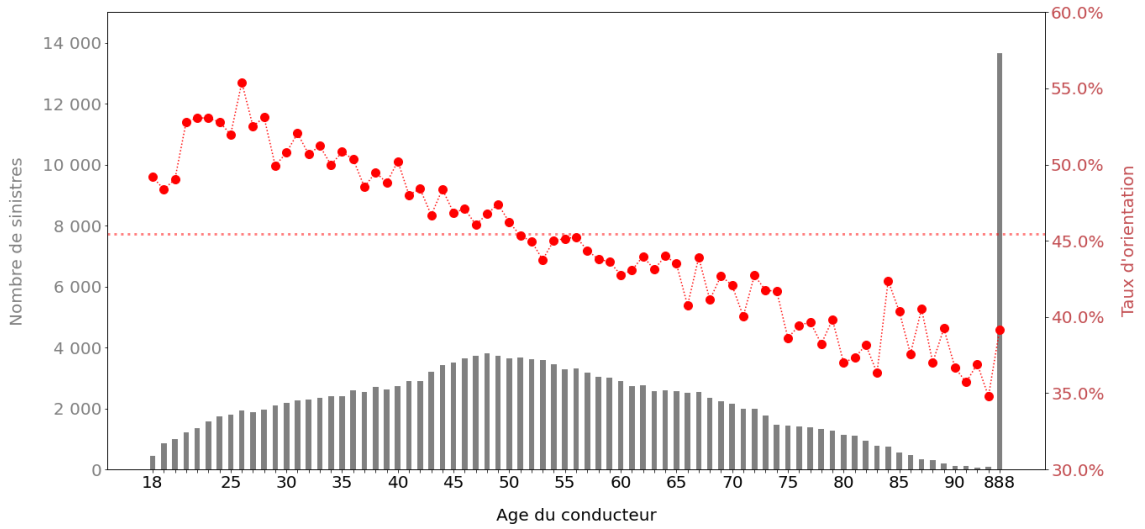



*Figure 17 - Taux d'orientation et nombre de sinistres en fonction de l'âge du conducteur*

Le nombre de sinistres par âge suit une cloche : il est plus faible sur les âges extrêmes que sont 18 et 93 ans avec moins de 500, alors qu'à son maximum, qui est de 48 ans, il est au-dessus de 4 000. Le taux d'orientation quant à lui, a une tendance décroissante, atteignant presque 55% pour les conducteurs de 26 ans et 35% pour ceux de 93 ans. Toutes les personnes âgées de plus de 50 ans ont un taux d'orientation plus faible que la moyenne de notre base d'entrainement. Enfin, il y a environ 14 000 valeurs manquantes représentées par 888, soit plus de trois fois plus que notre maximum. Ces données manquantes ont un taux d'orientation autour de 40%, qui s'avère être bien plus faible que la moyenne de la base d'entrainement. Ceci paraît assez logique car les données manquantes au niveau de l'âge s'expliquent par le fait que ce sont des professionnels et non des particuliers, et selon la figure 4, ils ont un taux d'orientation plus faible que la moyenne.

La méthode des k plus proches voisins sera utilisée pour remplacer les valeurs manquantes car leur proportion est importante : elles représentent près de 8% des données de la variable. Il n'est pas envisageable de supprimer ces lignes, et d'appliquer une méthode plus simpliste telle que le remplissage par la médiane car cela biaiserait fortement la proportion d'une seule donnée.

Nous avons donc créé une nouvelle base composée de nos données sans les valeurs manquantes afin de paramétrer et tester notre modèle. Nous séparons cette base en une base d'entrainement composée de 80% des données choisies aléatoirement et les 20% restants seront utilisés pour valider notre modèle sur des données jamais utilisées.

Nous avons sélectionné les variables les plus corrélées avec la variable âge du conducteur et avons créé la matrice de corrélation ci-dessous.

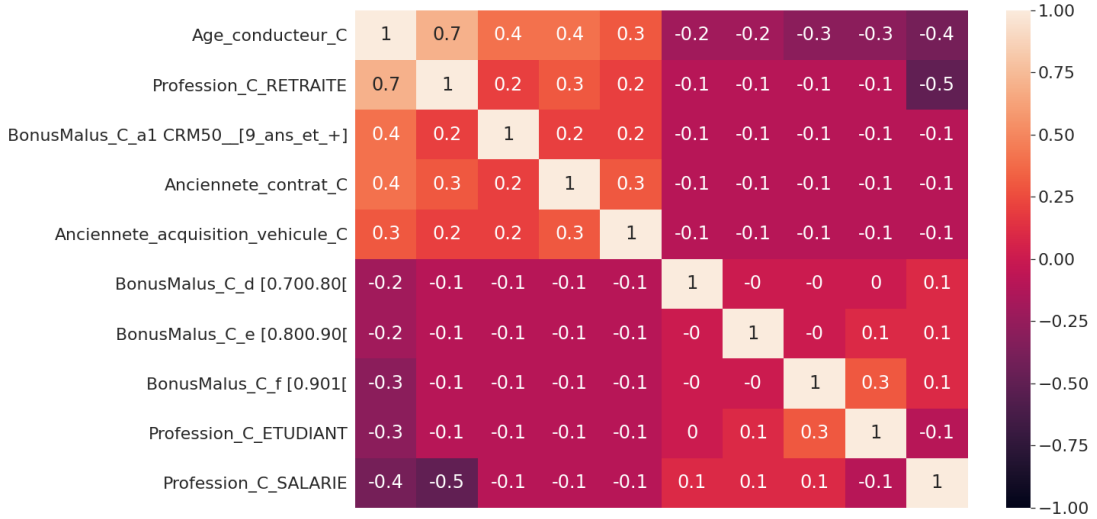



*Figure 18 - Matrice de corrélation des variables les plus corrélées à l'âge du conducteur*

Les variables corrélées à l'âge du conducteur sont pour la plupart assez intuitives. Celle qui implique que l'assuré soit âgé est le fait que l'assuré soit à la retraite tandis qu'être salarié ou étudiant implique qu'il est plus jeune. Les données sur le coefficient bonus-malus ressortent aussi : plus le bonus est proche de 0,5 plus l'assuré a eu le temps d'accumuler du temps sans sinistre alors qu'un jeune conducteur aura un coefficient plus proche de 1. La variable qui aurait pu être la plus explicative est l'ancienneté du permis, malheureusement lorsque la donnée était manquante pour l'âge du conducteur, elle l'était aussi pour l'ancienneté du permis.

Les variables qui vont servir pour notre méthode étant sélectionnées, nous pouvons maintenant l'appliquer pour plusieurs nombres K de plus proches voisins et choisir le nombre optimal.

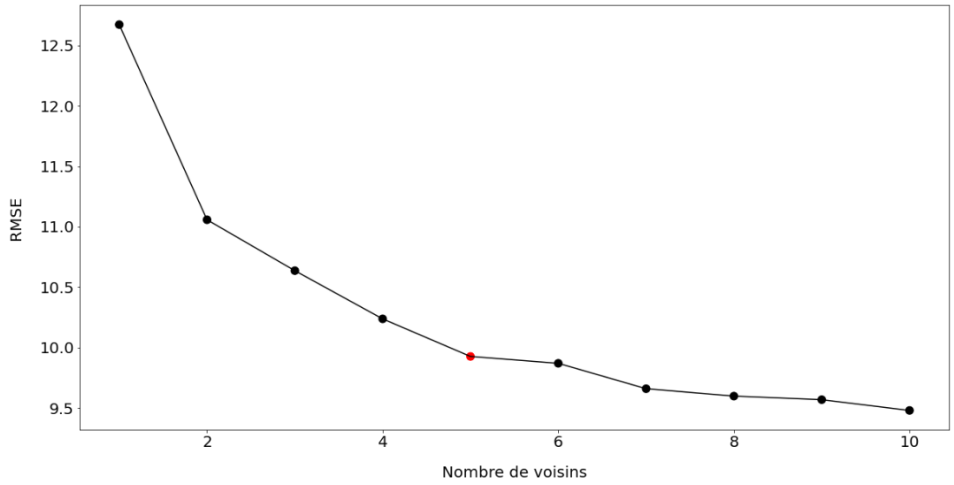



*Figure 19 - RMSE des modèles des k plus proches voisins par rapport au nombre de voisins sélectionné*

Afin de sélectionner le nombre optimal de voisins, la méthode du coude a été employée (Sangnier, 2019), c'est-à-dire analyser la courbe et voir à quel moment une cassure est observée. Dans cet exemple la cassure

n'est pas évidente mais nous observons une décroissance plus faible entre les points cinq et six. Nous choisirons donc cinq comme nombre de voisins optimal.

Une fois notre modèle choisi il s'agit de le paramétrer sur la base d'entrainement puis le tester sur la base de test et comparer les RMSE avec d'autres modèles plus basiques.

| Modèles | Âge aléatoire entre 18 et 93 ans | Médiane de la base d'entrainement | Moyenne de la base d'entrainement | K plus proches voisins |
|---|---|---|---|---|
| RMSE | 28,2 | 16,4 | 16,4 | 10,2 |

*Tableau 12 - RMSE en fonction de la méthode de complétion utilisée*

Nous avons donc dans l'ordre du plus mauvais remplacement au meilleur : le fait de remplacer le nombre manquant par un nombre au hasard entre 18 et 93, puis les méthodes de remplacement par la médiane et la moyenne qui sont très proches l'une de l'autre, et enfin le modèle des k plus proches voisins. Ce modèle est 2,8 fois meilleur que le hasard et 1,6 fois meilleur que les méthodes de la médiane et de la moyenne.

Nous appliquons alors le modèle paramétré sur les données d'entrainement pour remplir les valeurs manquantes de notre base. En dernière étape, nous pouvons comparer sur le graphique ci-dessous les différences entre le remplacement par la méthode de la médiane ou des k plus proches voisins.

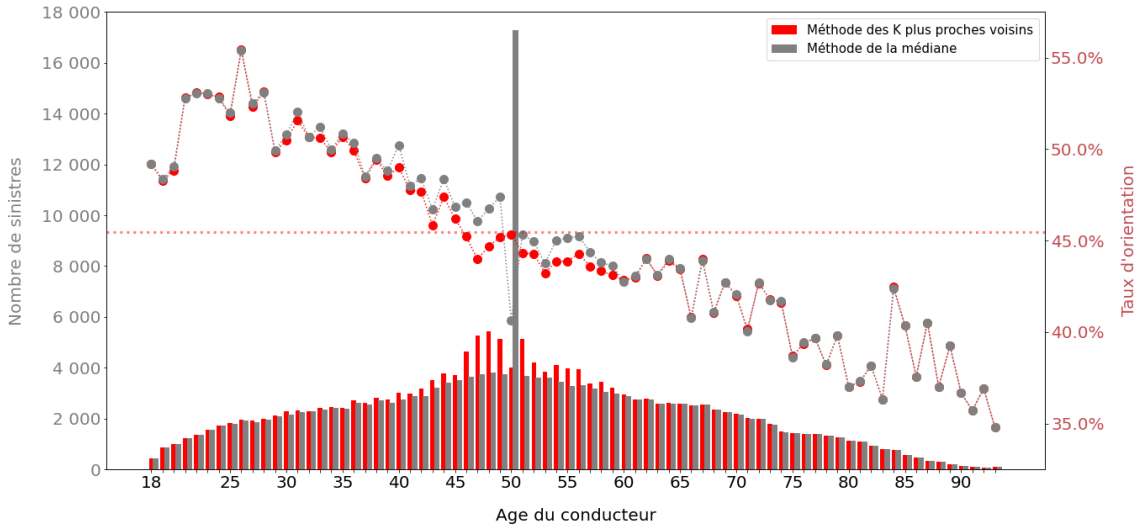



*Figure 20 - Taux d'orientation et nombre de sinistres en fonction de l'âge du conducteur et de la méthode de complétion utilisée*

Nous voyons que la méthode de la médiane biaise énormément la valeur 50, non seulement en la rendant très importante au niveau du nombre mais aussi en baissant fortement son taux d'orientation. La méthode des k plus proches voisins permet de lisser cet effet en rajoutant du volume principalement entre 40 et 60 ans et en baissant légèrement le taux d'orientation sur ces âges. Grace à l'application de cette méthode nous gardons une allure de courbe décroissante.

Cependant cette variable n'est pas parfaitement décroissante. Nous allons voir dans la partie suivante comment retravailler les variables afin de les rendre plus compréhensibles pour nos modèles et d'éviter le surapprentissage.

### II.3.2. Discrétisation des variables

Lorsque nous travaillons sur des données, il y a souvent des variables qui ne sont pas ou peu compréhensibles si elles ne sont pas retravaillées. Certaines variables, comme le code postal ou le code Insee de l'assuré, ne sont pas explicatives en premier lieu à cause du trop grand nombre de modalités qu'elles contiennent. D'autres variables, comme les variables numériques, auront besoin d'être découpées en plusieurs tranches afin de pouvoir les inclure dans un modèle de régression logistique et qu'elles ne soient pas interprétées comme ayant un rapport linéaire avec notre variable à expliquer. Enfin, nous retrouvons aussi un autre type de variables qui auront trop peu de données sur certaines modalités pour être explicatives et qui gagneront à être regroupées.

#### *II.3.2.1. Premiers cas pratiques*

Il y a différents types de variables à retravailler et pour cela nous avons utilisé plusieurs méthodes. La première méthode est celle de regroupement des modalités trop peu représentées et à même comportement. Par exemple dans la *Figure 6*, la dernière valeur du groupe SRA est « 43 & > ».

| Groupe | 43 | 44 | 45 | 46 | 47 | 48 | 49 | Inconnu |
|---|---|---|---|---|---|---|---|---|
| Nombre de sinistres | 59 | 34 | 23 | 17 | 10 | 3 | 1 | 43 |
| Part dans la base | 0.03% | 0.02% | 0.01% | 0.01% | 0.01% | 0.001% | 0.001% | 0.03% |
| Taux d'orientation | 22% | 15% | 8% | 12% | 10% | 33% | 100% | 7% |

*Tableau 13 - Répartition des valeurs supérieurs à 42 du groupe SRA*

Nous voyons sur le tableau ci-dessus qu'elle contient les sept groupes supérieurs à 42 et le groupe Inconnu qui à eux tous ne comptent que pour 0.1% de la base. De plus ils ont des taux d'orientation semblables à part pour les groupes 48 et 49 qui ont trop peu de valeurs pour que le taux soit robuste.

Un autre retraitement à faire pour pouvoir utiliser des variables catégorielles, tels que la profession ou encore la tranche urbaine, est l'encodage « One-Hot ». Cette transformation revient à transformer une variable catégorielle ayant *N* valeurs distinctes en N variables binaire, composées chacune soit de 0 soit de 1. Cette transformation est très utile dans les algorithmes d'apprentissages tels que la régression logistique pour éviter de donner une impression de lien linéaire. Par exemple pour la tranche d'unité urbaine, sur la figure 7, l'évolution du taux d'orientation n'est pas linéaire. Il sera donc intéressant de séparer cette variable de 13 valeurs distinctes en 13 variables binaires.

#### *II.3.2.2. La méthode des quantiles*

Une méthode de discrétisation que nous allons utiliser est celle des quantiles, aussi appelée méthode des effectifs égaux. (Hunault, 2021)

Le but de cette méthode est d'obtenir le même nombre de données dans chaque groupe. Cette méthode permet de transformer une variable quantitative en variable factorielle. Nous allons détailler son principe :

1. Tout d'abord choisir le nombre de groupes à créer *n*.
2. Trier les valeurs de la variable de manière croissante.
3. Afin d'obtenir le nombre d'individus par groupe, diviser le nombre de valeurs totales de la variable *N*, par le nombre de groupes :

$$i = \frac{N}{n}.$$

4. Mettre les *i* premiers individus, ($N_1$, … , $N_i$) dans le premier groupe.

a. Si les individus $N_{i+1}$, …, $N_{i+k}$ ont la même valeur que l'individu $N_i$ les mettre eux aussi dans le premier groupe.

5. Procéder ainsi de suite jusqu'à avoir les n groupes constituées.

C'est la méthode que nous avons utilisée pour le regroupement de l'âge du conducteur ou encore pour la variable « Dernier tarif », qui donne pour chaque véhicule sinistré le dernier tarif à date en euro. Pour cette dernière variable nous avons décidé de tester cette méthode avec dix groupes.

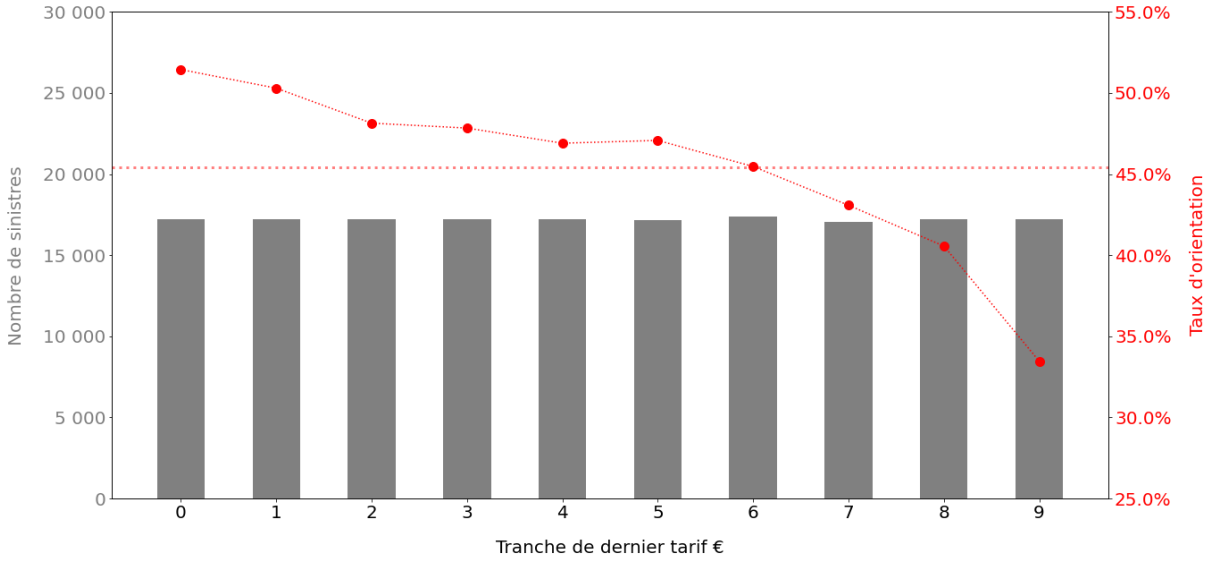



*Figure 21 - Taux d'orientation et nombre de sinistres en fonction du regroupement de la variable « dernier tarif »*

Le graphique ci-dessus représente le taux d'orientation et le nombre de sinistres sur les dix groupes que nous avons créés. Les environ 17 000 véhicules avec le coût le plus bas se trouvent dans le groupe 0, puis les 17 000 suivant dans le groupe 1 et ainsi de suite jusqu'au 17 000 derniers avec les plus hauts coûts du groupe 9. Nous observons bien environ le même nombre de sinistres pour chacun des groupes même si le groupe 7 semble être un peu moins rempli, ce léger écart est dû aux valeurs égales remplissant d'avantage certains groupes comme nous l'avons vu dans le point 4.a de la méthode. Le taux d'orientation sur cette variable est clairement décroissant, mais il parait se stabiliser sur certains groupes, comme sur le 2 et le 3 par exemple, et au contraire il parait fortement décroissant pour d'autres, comme sur le passage des groupes 8 à 9.

En retravaillant légèrement cette variable nous pourrions avoir une décroissance linéaire qui serait plus appropriée pour un modèle de régression logistique. Un modèle de régression logistique est un modèle de régression linéaire généralisé il est donc plus performant sur des variables qui ont une tendance linéaire.

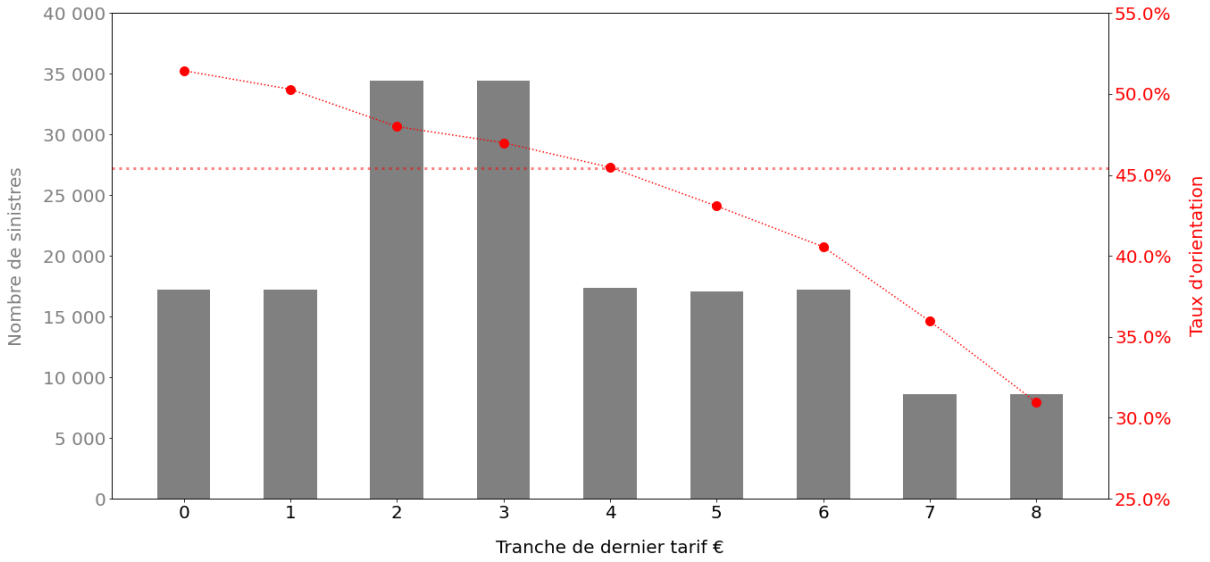



*Figure 22 - Taux d'orientation et nombre de sinistres en fonction du regroupement revu de la variable « dernier tarif »*

Sur ce graphique nous avons regroupé les groupes 2 et 3 et les groupes 4 et 5 du regroupement précédent pour les mettre respectivement dans les groupes 2 et 3, nous avons aussi divisé le groupe 9 en deux pour créer les groupes 7 et 8. Nous obtenons ainsi un taux d'orientation décroissant qui passe de 52% pour le groupe 0 qui représente les véhicules les moins chers à un taux d'orientation de 32% pour le groupe des véhicules les plus chers. Cette décroissance est quasiment linéaire et sera donc mieux prise en compte dans un modèle de régression.

Ainsi, bien que la méthode des quantiles permette de faire un premier regroupement sans a priori sur la variable, nous avons vu qu'il avait besoin d'être retravaillé par la suite, pour être optimal. Nous allons donc présenter dans la prochaine partie, la méthode des Kmeans qui permet de passer outre cette phase de retraitement tout en optimisant le nombre de groupes.

#### *II.3.2.3. La méthode des Kmeans*

##### II.3.2.3.1. Définition mathématique

Le but de cette méthode est de regrouper des variables factorielles dans des groupes de taux d'orientation homogène. Pour cela, nous allons utiliser un algorithme d'apprentissage non supervisé, c'est-à-dire nous allons partir sans a priori sur le nombre de classes et les regroupements. La méthode que nous avons choisie est celle des Kmeans. (Rakotamalala, 2016)

Le principe est de créer les groupes avec la distance la plus faible, au sens d'une métrique, entre chaque point de ce groupe et le centre de ce groupe. Dans notre cas, nous utiliserons la distance euclidienne que nous avons définie dans la partie **Erreur ! Source du renvoi introuvable.**. Dans notre cas, la formule est plus simple c ar nous ne ferons la distance qu'en 2 dimensions. Elle se résume donc pour chaque couple $o_1 = (x_1, y_1)$ et $o_2 = (x_2, y_2)$ à faire $\mathrm{d}(o_1, o_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

Nous allons voir ci-dessous l'algorithme des Kmeans détaillé pas à pas :

- Tout d'abord nous choisissons un nombre *n* de groupes à créer. Nous verrons dans la deuxième partie comment optimiser ce nombre de groupes.
- Après ce choix, *n* points sont placés de manière aléatoire : ce seront nos centres de gravité de groupes provisoires. Nous créons *n* groupes autour de ces points en minimisant pour chacun la distance entre le point et ces centres.
- Une fois tous les points répartis en *n* groupes, nous calculons le centre de gravité de ce groupe au sens encore une fois de la distance euclidienne.
- Puis nous répétons le même algorithme avec cette fois les *n* réels centres de gravité jusqu'à ce que tous les points retombent à chaque fois dans le groupe dans lequel ils étaient au tour précédent.

#### II.3.2.3.2. Nombre de groupes optimal

La méthode des Kmeans est assez intuitive. Cependant, à la différence des méthodes hiérarchiques qui aident au choix du nombre de groupes, il faut ici tester plusieurs nombres de groupes *n* afin de voir lequel est le plus judicieux. Nous utiliserons pour cela l'inertie intra-groupes. Elle représente l'homogénéité à l'intérieur d'un groupe : moins ce groupe sera homogène plus son inertie sera importante.

Mathématiquement, l'inertie intra-groupes est la somme des carrés des distances entre chaque point et son centre de groupe, divisée par le nombre de points au total. Soit une classification en *n* groupes d'effectifs $e_1, ..., e_n$, les *k* individus de l'échantillon étant des points d'un espace euclidien. Notons les groupes $G_1, ..., G_n$ et $g_1, ..., g_n$ leurs centres de gravité :

$$I_{intra} = \sum_{i=1}^{n} \sum_{o \in G_i} d^2(o, g_i).$$

Plus nous faisons de groupes, plus cette inertie sera faible. Mais le but est de garder le plus d'information possible apporté par la variable tout en minimisant son nombre de groupes afin d'éviter du sur-apprentissage. Il faut donc réussir à minimiser à la fois le nombre de groupes et l'inertie intra-groupes, pour cela nous utiliserons la méthode du coude.

#### II.3.2.3.3. Exemple : Le code INSEE

En analysant les variables de la base de données, nous avons découvert que les données géographiques faisaient parties des plus intéressantes, car les plus discriminantes pour le taux d'orientation.

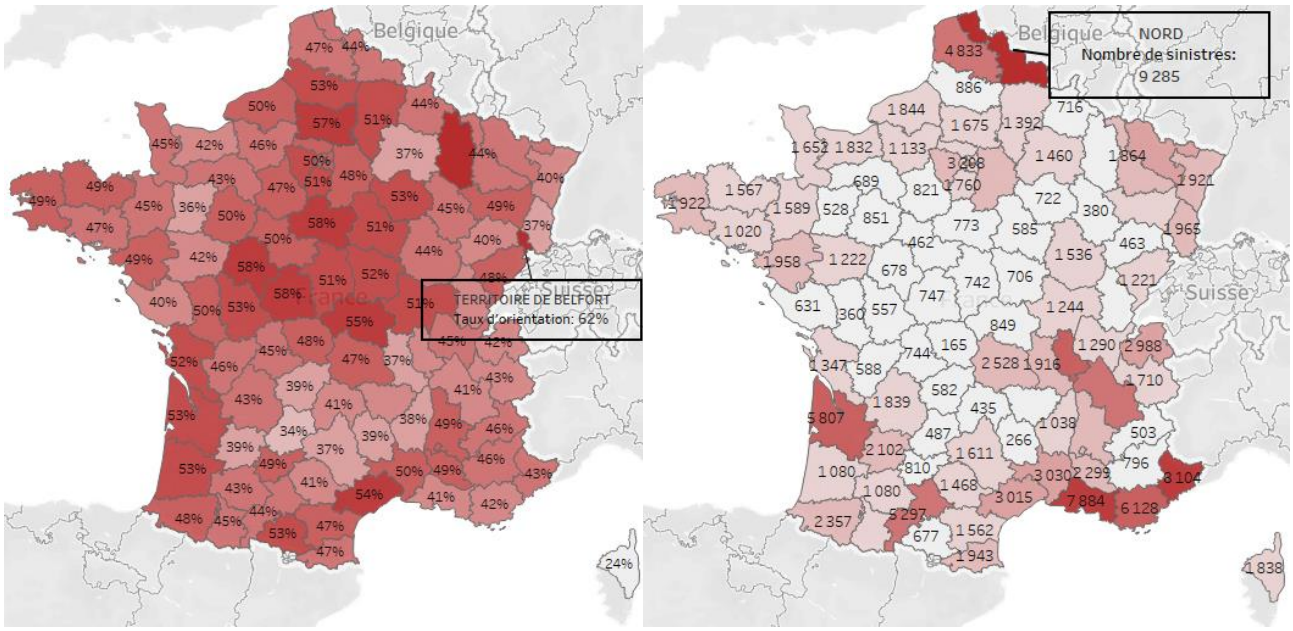



*Figure 23 - Carte de France du taux d'orientation et du nombre de sinistres par département*

Les cartes ci-dessus montrent à gauche le taux d'orientation par département et à droite le nombre de sinistres par département dans notre base. Le taux d'orientation minimal est de 24% en Corse et *heureusement* son nombre de sinistre n'est que de 1 838 soit un peu moins de 1% de la base. Le taux d'orientation maximal est de 62% sur le territoire de Belfort, *malheureusement* pour seulement 484 sinistres. Les trois départements les plus sinistrés pour Generali sont le Nord, les Alpes-Maritimes et les Bouches-du-Rhône et ils ont tous les trois un taux d'orientation légèrement en dessous de la moyenne.

Cette discrimination géographique du taux d'orientation s'accentue encore plus lorsque nous regardons plus précisément à la maille code postal ou même plus finement sur le code INSEE.

La variable code INSEE a 21 357 modalités ce qui la rend impossible à interpréter, nous avons donc besoin de regrouper les modalités qui se ressemblent et apportent la même information sur le taux d'orientation. Pour ça nous allons mettre en application l'algorithme des Kmeans décrit précédemment.

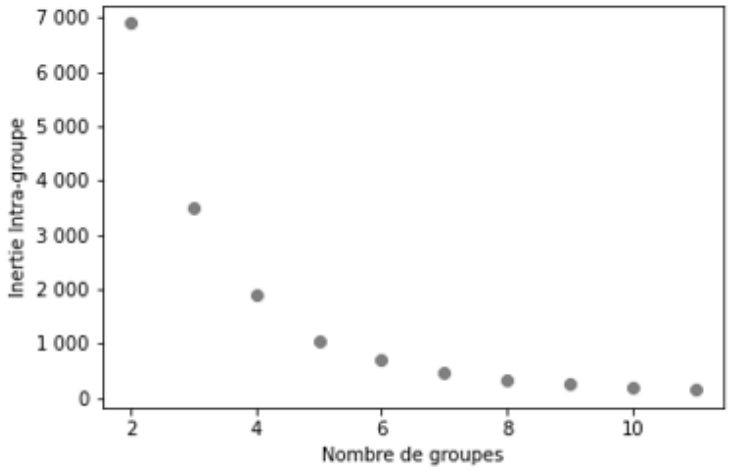
Nous allons créer une base qui pour chaque sinistre associe le code INSEE et la moyenne de taux d'orientation des sinistres dans ce code INSEE.

| Code Insee | 31555 | 06088 | 33063 | 06029 | 83137 |
|---|---|---|---|---|---|
| Taux d'orientation (Point X et Y de la base) | 45.5% | 39.2% | 49.9% | 46.8% | 40.6% |
| Nombre de sinistres | 1 191 | 1 070 | 807 | 712 | 705 |

*Tableau 14 - Nombre de sinistres et taux d'orientation des plus grands code INSEE*

Le tableau ci-dessus présente les cinq codes INSEE les plus sinistrés avec leurs taux d'orientation et le nombre de sinistres associés. Le code INSEE, le plus sinistré est celui de Toulouse. Il cumule un peu moins de 1 200 sinistres et un taux d'orientation de 45.5%. Même en étant le Code INSEE le plus représenté, il ne compte que pour 0.7% de notre base d'entrainement.

Une fois cette base créée, nous pouvons appliquer la méthode des Kmeans pour plusieurs nombres de groupes et analyser pour chaque méthode l'inertie intra-groupes.




| Nombre de groupes | Inertie Intra-groupe | Différence Groupe N/N-1 |
|---|---|---|
| 2 | 6 898 | |
| 3 | 3 508 | 97% |
| 4 | 1 897 | 85% |
| 5 | 1 064 | 78% |
| 6 | 709 | 50% |
| 7 | 484 | 46% |
| 8 | 348 | 39% |
| 9 | 263 | 32% |
| 10 | 198 | 33% |
| 11 | 158 | 26% |

*Figure 24 - Inertie Intra-groupes en fonction du nombre de groupes*

Le graphique ci-dessus représente l'évolution de l'inertie intra-groupes en fonction du nombre de groupes utilisés pour la méthode des Kmeans. Dans le tableau de droite nous avons noté la valeur exacte de chaque inertie et nous avons calculé l'évolution de l'inertie en fonction du groupe précédent. Pour cela nous avons appliqué la formule suivante :

$$Diff\ Groupe\ N/N\text{-}1\ = \frac{Inertie\ Groupe\ N\text{-}1 - Inertie\ Groupe\ N}{Inertie\ Groupe\ N\text{-}1}.$$

Cette différence d'inertie intra-groupe diminue jusqu'à une séparation en 6 groupes ou elle est de 50% et celle de 7 groupes est de 46% : il s'agit du premier ralentissement net dans la décroissance de cette différence. En complément l'allure de la courbe fait un coude entre les groupes 4 et 7, nous prendrons donc 6 comme nombre de groupes pour cette classification.

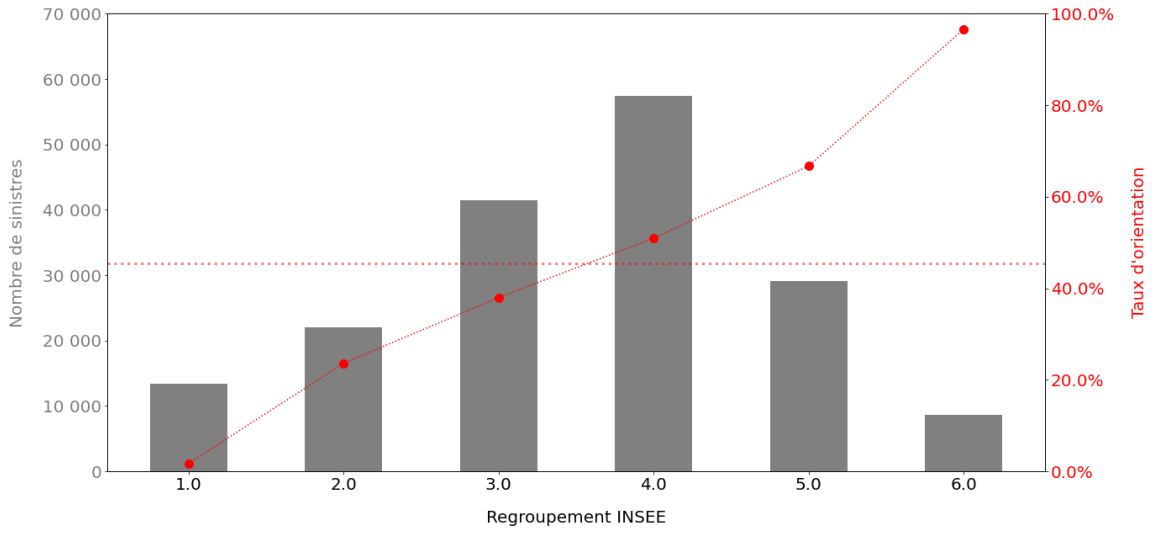



*Figure 25 - Taux d'orientation et nombre de sinistres en fonction du regroupement par code INSEE*

La figure ci-dessus présente le nombre de sinistres et le taux d'orientation par classe de *Code Insee* créés lors du regroupement avec la méthode des Kmeans. Nous observons que ce regroupement discrimine énormément le taux d'orientation puisqu'il va à près de 0% pour le groupe 1 et à près de 100% pour le groupe 6. De plus le nombre de sinistres dans ces deux groupes est assez faible. Le problème est que la maille code INSEE est très fine la plupart des code INSEE présents dans le groupe 1 sont des codes INSEE avec peu de sinistres. Apprendre sur de telles données peut être compliqué et entrainer du sur-apprentissage.

Cette méthode des Kmeans sera utilisé sur une dizaine de variable telles que le modèle et la marque du véhicule, ou encore le code postal de l'assuré. Plus généralement dans la suite de l'étude toutes les variables que nous rencontrerons et qui seront notées « Regroupement » ou « Reg » dans les sorties logiciels auront été discrétiser par une des deux méthodes vues précédemment.

### II.3.3. Corrélation des variables

La base de données est, à ce stade, nettoyée. En effet, les variables sont complètes, sans données manquantes et sont regroupées de manière intelligibles pour les premières analyses ainsi que pour les futurs modèles. Dans cette partie nous allons chercher à supprimer les variables qui apportent la même information. Il faudra donc supprimer les variables trop corrélées entre elles et pour ce faire nous commencerons par définir mathématiquement ce qu'est la corrélation. (Rakotomalala, 2017)

#### *II.3.3.1. Corrélation de Pearson*

Le coefficient de corrélation de Pearson met en avant une relation linéaire entre deux variables continues. Dans notre étude, l'indice de corrélation linéaire de Pearson sera suffisant, nous n'aurons pas besoin de nous pencher sur des corrélations plus complexes. Cet indice nous permettra de supprimer les variables qui sont trop corrélées. Nous allons utiliser deux types de modèles :

- La régression logistique, qui fonctionne avec des variables indépendantes,
- Les arbres de décision, qui ne sont pas impactés par les variables corrélées mais qui tournent bien plus vite avec moins de variables.

La formule de cet indice, appliqué à deux variables *X* et *Y* de taille *n*, est :

$$\mathrm{Corr}_{\mathrm{pearson}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})*(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}*\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.$$

Avec $\bar{X}$ *et* $\bar{Y}$ respectivement la moyenne de l'échantillon pour les variables *X* et *Y*.

Cet indice fluctue dans l'intervalle [-1 ;1] et s'il est égal à :

- 1 la relation est parfaitement linéaire, dans le sens positif, c'est-à-dire quand *X* croit *Y* croit,
- -1 la relation est parfaitement linéaire, mais négative,
- 0 la relation n'est pas du tout linéaire.

Cet indice permettra d'identifier les variables qui apportent la même information et donc les variables qui pourront être supprimées ou remplacées dans nos modèles.

#### *II.3.3.2. Matrice de corrélation*

Afin d'obtenir cet indice pour toutes les variables de notre base, nous allons créer la matrice de corrélation qui donne pour chaque croisement de variable le coefficient de corrélation.

Il y a dans notre base plus de 90 variables explicatives, après la création de nouvelles variables et de l'encodage « One-Hot » sur les variables catégorielles. Nous ne pouvons donc pas représenter toutes les corrélations de nos variables, nous allons donc mettre en avant les groupes de variables les plus corrélées entre elles.

Pour commencer, nous allons nous intéresser aux 10 variables qui expliquent le mieux linéairement le taux d'orientation sur la base d'entrainement, donc les 10 variables ayant les corrélations les plus fortes en valeurs absolues avec le « TOP_AGREE ».

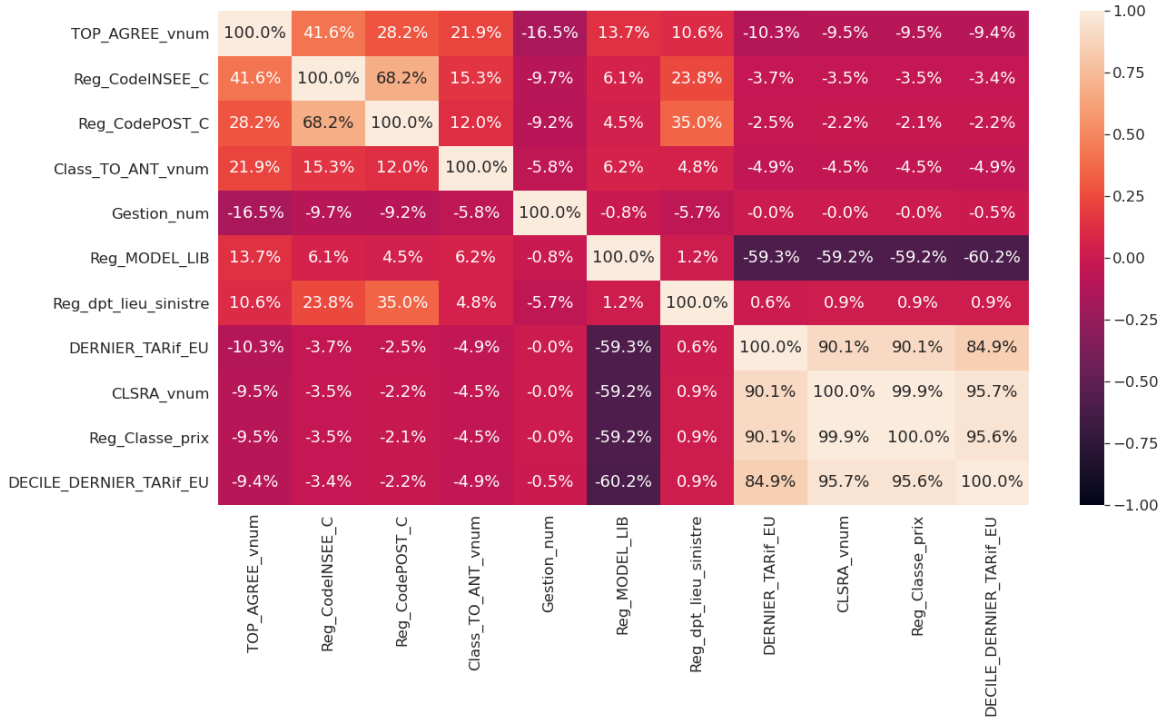



*Figure 26 - Matrice de corrélation des variables les plus corrélées au taux d'orientation sur la base d'entrainement*

Le type de variable ressortant comme le plus important à première vue est l'aspect géographique. Le regroupement de codes INSEE, de codes postaux ou encore de départements dans lequel a eu lieu le sinistre sont 3 des 6 premières variables. Nous avons ensuite la variable du taux d'orientation antérieur qui ressort en troisième position. La gestion ressort quant à elle en quatrième place. Ces variables ne sont pour l'instant pas prises en compte dans le calcul du tarif chez Generali. Le fait que ces variables ressortent en haut de la liste conforte l'intérêt de notre démarche pour le reste de notre étude.

Les 5 autres variables parmi les 10 premières qui ressortent sont des variables véhicules. La première à ressortir est le regroupement de modèle et les 4 autres sont des variables en rapport avec le prix du véhicule. Ces 4 variables sont très corrélées les unes avec les autres, avec une corrélation minimale de 84.9%. Elles font partie des variables sur lesquelles il faudra faire des choix, il ne sera en effet pas possible de garder toutes ces variables lors du lancement de nos modèles.

Lors de l'analyse des corrélations entre ces variables, nous pouvons voir qu'il y a une relation, mais assez faible, entre les regroupements de codes INSEE et de codes postaux avec la gestion et le taux d'orientation sur antérieur de l'ordre de 10%. De la même manière, entre ces deux dernières variables la corrélation est de -5.8%, donc existante mais faible. Le même constat peut se faire sur presque toutes les variables, excepté sur la gestion qui a une corrélation de moins de 1% avec toutes les variables véhicules. Enfin les deux variables

de regroupement du codes INSEE et du codes postaux apportent beaucoup d'information au taux d'orientation mais ont une forte corrélation entre elles, de 68.2%. Il va donc falloir choisir une seule de ces deux variables dans la perspective de nos modèles de régression logistique. Nous avons donc fait les travaux en ne gardant à chaque fois qu'une de ces deux variables mais nous ne présenterons dans la suite de nos travaux que la version avec le *regroupement de code Postale* car comme nous le verrons au début de la partie 0 la variable de *regroupement de code INSEE* entrainé trop de sur-apprentissage.

Plus d'une trentaine de variables ont été supprimés grâce à ces études de corrélations.

La construction de la base de données est à présent terminée. Après avoir consolidé la base d'études à partir de sept bases de données, il a fallu la nettoyer. Ce travail a été effectué en complétant les données manquantes à l'aide de modèles mathématiques, puis en discrétisant les variables numériques ou avec trop de modalités. Pour finir nous nous sommes intéressés à aux corrélations des données ce qui a permis la suppression de nombreuses variables similaires. L'étape suivante, qui est la *construction* et le choix des modèles mathématiques afin de scorer l'appétence d'orientation des clients sinistrés, peut à présent commencer.

# III. Création d'un scoring d'appétence client à l'orientation

La phase de constitution de la base de données étant terminée, il est maintenant temps de créer notre scoring d'appétence client à l'orientation. Le but est d'appliquer des modèles d'apprentissage supervisé afin de deviner si un sinistre est allé dans l'un des garages du réseau Assercar. Mathématiquement, c'est un problème de classification binaire à résoudre. Nous aurons d'un côté, pour chaque sinistre orienté des 1 et pour ceux non orientés des 0, et de l'autre, le résultat de nos modèles qui donnera une probabilité entre 0 et 1. Nous mesurerons ensuite la performance de nos modèles en transformant cette probabilité en 0 et 1. Une fois le meilleur modèle choisi nous l'utiliserons pour créer notre scoring en regroupant les individus par taux d'orientation homogène.

## III.1. Modélisation de l'appétence à l'orientation

Nous allons dans cette partie calculer la probabilité qu'un client Generali aille dans un garage agréé en utilisant une régression logistique. C'est l'un des modèles privilégiés pour la résolution de problèmes de classification. Il est relativement simple à mettre en œuvre et surtout, à la différence de certains modèles « boite noire », il permet de disposer de résultats interprétables, en identifiant les variables significatives et en explicitant leur importance. Mais avant de créer ce modèle il va falloir commencer par présenter son fonctionnement.

### III.1.1. Définition de la régression logistique

#### *III.1.1.1. Préquel : régression linéaire multiple*

La régression logistique est un modèle de régression linéaire auquel nous appliquons une fonction de coût particulière. Nous allons donc commencer par définir le fonctionnement de la régression linéaire.

Un modèle linéaire effectue une prédiction en calculant une somme pondérée des variables d'entrée à laquelle il rajoute une constante. L'équation obtenue est la suivante :

$$\hat{y} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n.$$

- $\hat{y}$ est la valeur prédite pour une observation,
- $x_i$ est la valeur de la $i^{ème}$ variable pour cette observation,
- $\theta_j$ est le $j^{ème}$ paramètre du modèle pour cette observation et $\theta_0$ est la constante.

Matriciellement cette formule est :

$$\hat{y} = {}^t\theta . X.$$

- Où $\theta$ est le vecteur $\begin{pmatrix} \theta_0 \\ \theta_1 \\ \cdots \\ \theta_{n-1} \\ \theta_n \end{pmatrix}$ et $X$ est le vecteur $\begin{pmatrix} 1 \\ x_1 \\ \cdots \\ x_{n-1} \\ x_n \end{pmatrix}$

Pour donner un exemple concret, prédire la donnée *âge du conducteur,* grâce aux variables $Y_1$ : *ancienneté du contrat* et $Y_2$ : *Top Retraite Oui/Non* qui aurait 1 pour oui et -1 pour non et avec un conducteur d'une ancienneté de contrat de 5 ans et toujours en activité, revient à cette équation :

$$\hat{y}_{age\_conducteur} = (47{,}2 \quad 0{,}6 \quad 12{,}4).\begin{pmatrix} 1 \\ 5 \\ -1 \end{pmatrix} = 37{,}8\ ans$$

#### *III.1.1.2. Estimation des probabilités*

Nous venons de voir *la base* de la régression linéaire multiple, la régression logistique fonctionne de la même manière en calculant une somme pondérée des variables d'entrées puis en rajoutant une constante, mais au lieu d'envoyer le résultat, il renvoie une probabilité du résultat. Le modèle de régression logistique ayant pour but de prédire des 0 et des 1, la distribution associée à ce modèle sera une binomiale, il renvoie une probabilité contenue dans ces deux bornes. De manière *basique*, quand cette probabilité est supérieure à 0,5 le résultat est censé être 1 et quand il est inférieur à 0.

Cette probabilité estimée par le modèle, est une transformation de la régression linéaire par la fonction logistique qui est en forme de S et qui est définie par : sig(t) $= \frac{1}{1+e^{-t}}$

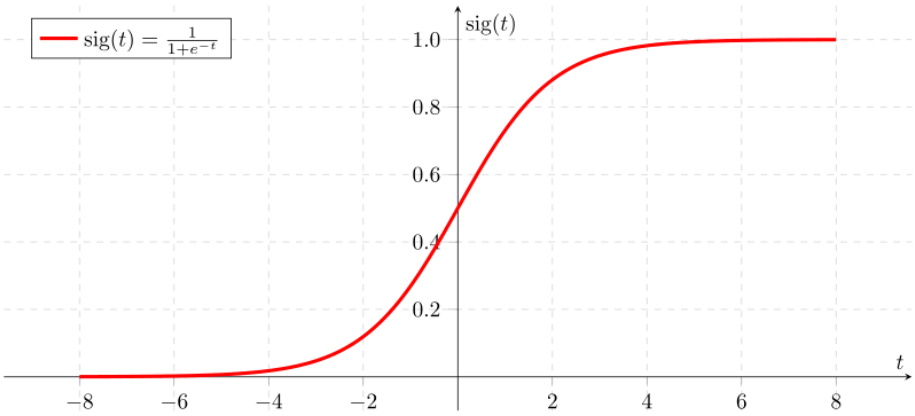



*Figure 27 - Graphique de la fonction logistique*

La probabilité du modèle se définit par :

$$\hat{p} = sig({}^{t}\theta.X).$$

Puis la prédiction de ce modèle se définit par :

$$\hat{y} = \begin{cases} 0 & si\ \hat{p} < 0{,}5 \\ 1 & si\ \hat{p} \geq 0{,}5 \end{cases}$$

Sur la figure ci-dessus, *sig(t)* est plus grand que 0,5 quand *t* est plus grand que 0 et cette fonction est plus petite que 0,5 quand *t* est négatif. Le modèle de régression logistique prédira donc 1 si $\theta^T.X$ est positif et 0 si $\theta^T.X$ est négatif.

#### *III.1.1.3. Estimation des paramètres*

Nous venons de voir comment se compose un modèle de régression logistique, mais il nous reste à voir comment déterminer les paramètres ($\theta_0, \quad \theta_1, \quad \dots \quad \theta_n$) de ce modèle.

Le but est de trouver une fonction qui dépend de $\theta$, et qui pénalise fortement les mauvais paramètres, c'est à dire qui estime des probabilités élevées pour les observations positives et des probabilités basses pour les observations négatives. Cette fonction s'appelle la fonction de coût. Il faudra ensuite minimiser son erreur

afin de minimiser l'erreur $\hat{y}$ et $y$. Pour une observation de la régression logistique, la fonction de coût pour une seule observation se définit par :

$$c(\theta) = \begin{cases} -\log(\hat{p}) & si\ y = 1 \\ -\log(1-\hat{p}) & si\ y = 0 \end{cases}$$

Cette fonction de coût fonctionne bien parce que si $y = 1$ :

- Pour que $c(\theta)$ soit proche de 0 il faut que $\hat{p}$ soit proche de 1,
- A l'inverse si $\hat{p}$ décroit et se rapproche de 0 cette fonction de coût croit très vite vers $+\infty$ grâce à la fonction log.

La même logique peut être appliquée pour le cas inverse ou $y = 0$.

Sur l'ensemble des $n$ observations du modèle la fonction de coût, aussi appelée Log Vraisemblance ou Log Likelihood, et qui sera notée LL, est donc :

$$LL(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log\left(\hat{p}^{(i)}\right) + (1-y^{(i)})\log\left(1-\hat{p}^{(i)}\right)\right].$$

Il n'existe pas de $\theta$ unique qui minimise cette équation, mais cette équation étant convexe il est certain de pouvoir trouver un minimum global grâce, par exemple, à une descente de gradient ordinaire. L'idée globale de la descente de gradient est de tester comment évolue le résultat en faisant évoluer, petit pas par petit pas, le paramètre $\theta$. Pour ce faire, les dérivées partielles pour chaque élément de $\theta$ sont utilisées. La dérivée partielle de la fonction de coût du *j*ème paramètre $\theta_j$ se calcule ainsi :

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \frac{1}{n}\sum_{i=1}^{n}\left[(sig(\theta^T.X^{(i)}) - y^{(i)})X_j^{(i)}\right].$$

Le principe est de cette équation est de tester, pour une variable $\theta_j$ de la base et pour chaque élément $X_j^{(i)}$ quelle est l'erreur entre la probabilité calculée et la véritable valeur $y^{(i)}$ puis de multiplier cette erreur par la valeur de l'élément $X_j^{(i)}$. Puis finalement normaliser la valeur en divisant par le nombre d'individus de la base. D'autres techniques, comme la descente de gradient stochastiques ou par mini-lots, existent, mais nous avons présenté ici une manière de résoudre le problème de minimisation de la fonction de coût $C(\theta)$.

### III.1.2. Application du modèle

#### *III.1.2.1. Premier modèle*

Nous avons vu précédemment quel était le principe de la régression logistique, nous pouvons maintenant lancer notre modèle sur la base d'entrainement. Pour ce faire un modèle de régression logistique sera retenu et implémenté sous Python et plus précisément avec la bibliothèque statsmodels.api. Cette bibliothèque permet de créer des modèles de régression logistiques et donne accès à de nombreuses informations sur le modèle telles que la p-value des différentes variables, qui sont indisponibles sur la bibliothèque *sklearn*. Pour commencer nous allons travailler sur la base avec toutes les variables afin de voir quel résultat nous obtenons puis nous tenterons d'optimiser ce modèle. Nous noterons dans la suite ce modèle le modèle « tot ».

Logit Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | TOP_AGREE_vnum | **No. Observations:** | 172080 |
| **Model:** | Logit | **Df Residuals:** | 172021 |
| **Method:** | MLE | **Df Model:** | 58 |
| **Date:** | Wed, 30 Mar 2022 | **Pseudo R-squ.:** | 0.1204 |
| **Time:** | 11:54:36 | **Log-Likelihood:** | -1.0429e+05 |
| **converged:** | True | **LL-Null:** | -1.1857e+05 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 0.000 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.2197 | 0.005 | -41.519 | 0.000 | -0.230 | -0.209 |
| **ZU_TUU_j_Nice** | 0.0254 | 0.008 | 3.120 | 0.002 | 0.009 | 0.041 |
| **ZU_Statut_commune_C** | 0.0018 | 0.012 | 0.151 | 0.880 | -0.021 | 0.025 |
| **Reg_dpt_lieu_sinistre** | 0.0151 | 0.006 | 2.447 | 0.014 | 0.003 | 0.027 |
| **Reg_MARQUE_LIB** | 0.0167 | 0.007 | 2.311 | 0.021 | 0.003 | 0.031 |
| **Reg_MODEL_LIB** | 0.1992 | 0.008 | 24.166 | 0.000 | 0.183 | 0.215 |

*Tableau 15 - Sortie Python régression logistique base totale, image arrêtée à la sixième variable*

Nous avons affiché les informations de bases apportées par la sortie Python sur le modèle de régression logistique. Nous avons coupé la sortie à la sixième variable pour plus de lisibilité. Le cadre du haut réuni des informations utiles sur le modèle telles que :

- La variable à prédire, ici le *TOP AGREE*
- Le modèle utilisé, ici logit pour régression logistique
- La méthode utilisée pour trouver les paramètres du modèle : ici MLE pour *Maximum Likelihood*. Par défaut c'est le maximum de la log vraisemblance qui est utilisée pour trouver les paramètres du modèle.
- Est-ce que le modèle converge ou pas, ici oui il converge.
- Le nombre d'observations dans notre base d'entrainement : 172 080
- Combien de variables sont utilisés dans le modèle : 58
- La log vraisemblance du modèle : -104 290, notée *LL-Model*.
- La log vraisemblance du modèle nulle, c'est-à-dire sans aucune autre variable que la constante : -118 570, notée *LL-Null*.
- Le $Pseudo\ R\text{-}Square$ est une mesure de l'amélioration du modèle grâce aux variables implémentées par rapport au modèle avec la constante seule, sa formule est :

  $$Pseudo\ R\text{-}Square = 1 - \frac{\text{LL-Null}}{\text{LL-Model}}.$$

  Ici il est égal à 0,1204
- Le *LLR p-value* est un test de significativité du modèle nul par rapport au modèle avec des variables. L'hypothèse de base est que le modèle nul apporte autant d'information que le modèle complet, donc plus la valeur est proche de 0 plus nous pouvons rejeter cette hypothèse et dire que le modèle

avec des variables apporte plus d'information. Ici elle est très proche de 0 vu que le résultat indique 0,000 nous avons donc un modèle performant.

Le cadre en dessous donne les statistiques par variable ci-dessous :

- La colonne *Coef* présente la liste des coefficients à appliquer à nos données afin d'avoir le résultat de la prédiction, c'est le paramètre $\theta$ définit dans la partie précédente.
- La colonne *std err* contient pour chaque coefficient l'écart type.
- *z* est la statistique du test de Wald, ce test permet de vérifier l'importance d'une variable dans le modèle. Il se calcule ainsi pour chaque variable :
$$z_{Wald} = \frac{Coef}{std\ error}.$$
- P>|z| est la p-value du test de Wald. L'hypothèse de base du test de Wald est que la variable n'a pas d'importance, plus la p-value est petite plus cette hypothèse est fausse.
Ce test de Wald part du principe que notre variable $z_{Wald}$ suit une loi du Chi-2 à 1 degré de liberté vu que la significativité d'une seule variable est testée.
- Les deux dernières colonnes représentent les bornes haute et basse de notre coefficient, toujours en nous servant de ce test de Wald.

Dans ce tableau, nous n'avons pas trié les variables par ordre d'importance afin de montrer l'écart qui pouvait exister parmi des variables choisies au hasard. La première variable qui ressort est la constante, c'est le $\theta_0$ de notre équation de régression défini dans la partie *Préquel : régression linéaire multiple*. C'est une valeur qui prend la valeur du coefficient, ici -0,2197 à toutes les données. Nous avons dans ce tableau les trois types de variables, au niveau de l'importance, que nous pouvons retrouver dans le modèle :

- Les variables peu importantes : C'est le cas du *Statut de la zone urbaine*, du *Regroupement de département du sinistre* et du *Regroupement de marque du véhicul*e. Ce sont des variables qui ont une p-value supérieur à 5%, qui ont donc moins de 95% de probabilité d'être significatif au sens du test de Wald. Et du coup elles ont des coefficients très faibles et ne feront que très peu varier la prédiction.
- Les variables importantes mais pas essentielles : C'est le cas de *la Zone Urbaine de Nice*. Elles ont une p-value inférieur à 5% elles sont donc significatives au sens du test de Wald, mais elles ont un coefficient assez faible, inférieur à 5%.
- Les variables cruciales : C'est le cas du *Regroupement de modèle de véhicule*. Elles ont des p-value inférieur à 0,1% et des coefficients forts, supérieurs à 5%.

Quelque soit le type de variables, il faudra tester s'il est pertinent de les garder mais ça nous donne un premier à priori sur ces variables.

Le modèle étant significatif, les variables qui sont importantes d'après le test de Wald sont identifiées et seulement celles-ci seront gardées ultérieurement. Le modèle peut donc être appliqué sur les données et les résultats peuvent être analysés.

| | Score orientation | Prédiction orientation | Réel orientation |
|---|---|---|---|
| **0** | 0.403794 | 0 | 0 |
| **1** | 0.482684 | 0 | 1 |
| **2** | 0.853677 | 1 | 1 |
| **3** | 0.393227 | 0 | 0 |
| **4** | 0.148114 | 0 | 0 |

*Tableau 16 - Résultats de la régression logistique sur les 5 premiers individus*

Ce tableau représente le résultat du modèle sur les cinq premières lignes de la base d'entrainement. Le score d'orientation est la probabilité que le client emmène son véhicule dans un garage agréé d'après notre modèle. La colonne *Prédiction orientation* vaut 1 lorsque la probabilité est supérieure à 50% et 0 dans le cas inverse. La colonne *Réel orientation* vaut 1 lorsque l'assuré est allé chez un garage agréé et 0 sinon. Sur cette sous échantillon, quatre des cinq observations prédites sont bonnes. L'individu 1 qui est la seule mauvaise prédiction a une probabilité très proche de 50% ce qui la rend compliqué à prédire.

Maintenant qu'un premier modèle a été créé, nous allons avoir besoin d'indicateurs qui nous permettent de mesurer la performance de notre modèle et de le comparer avec d'autres modèles.

### *III.1.2.2. Mesures de performance*

#### III.1.2.2.1. La p-value

Une première mesure de performance interne au modèle de régression logistique est la p-value du test de Wald pour chaque variable. Plus elle est proche de 0, plus la variable est censée apporter de l'information au modèle. La plupart du temps, les coefficients les plus grands sont ceux avec les p-values les plus fortes et inversement pour les variables à faible p-value.

| Variable | Regroupement code postal | Taux d'orientation sur antérieur | Gestion | ... | Offre 8000 km | Classe SRA | ... | Sexe : FEMME | Formule_garanties : RC+DR | GENRE_VEH : VASP |
|---|---|---|---|---|---|---|---|---|---|---|
| Ordre d'importance selon la p value | 1 | 2 | 3 | ... | 22 | 23 | ... | 56 | 57 | 58 |
| Coefficient | 0,5513 | 0,4322 | -0,297 | | -0,0076 | -0,1742 | | 0,0031 | 0,0031 | 0,001 |
| P-value | 0 | 0 | 0 | | 0,159 | 0,2 | | 1 | 1 | 1 |

*Tableau 17 - Classement de l'importance des variables selon la p-value*

Le tableau ci-dessus présente les variables classées de la p-value la plus faible à la p-value la plus forte, avec pour chaque variable le coefficient du modèle associé. Les trois variables les plus importantes sont le regroupement de code postal, le taux d'orientation sur antérieur et la gestion du sinistre, les trois moins importantes sont le fait que le genre de véhicule soit une VASP, que la formule de garantie soit Responsabilité civile et défense recours exclusivement et que le conducteur soit une femme. Pour ces exemples l'importance en valeur absolu du coefficient suit parfaitement le classement de la p-value. Cependant la p-value de l'offre 8 000 km est plus faible que celle de la classe SRA et pour autant le coefficient de la classe SRA est bien plus fort que celui de l'offre 8 000 Km. Cette incongruité peut s'expliquer par la corrélation de la classe SRA avec d'autres variables qui ressortent avant au niveau de la p-value comme le groupe de prix, la puissance administrative ou le dernier tarif en euro (cf. figure 20 chapitre II-B-3-b). Cet effet sera mieux expliqué une

fois le un modèle forward appliqué à nos variables selon la p-value dans la sous partie III.1.2.3 Optimisation du modèle : Méthode Forward.

#### III.1.2.2.2. Matrice de confusion et indicateurs associés

Dans la partie précédente, une mesure de performance interne au modèle a été vue. Une mesure de performance du résultat du modèle sera traitée, c'est-à-dire à quel point le modèle réussit à correctement prédire le comportement des assurés. Pour ce faire, la matrice de confusion sera utilisée et les différentes mesures d'erreurs qui lui sont associées seront détaillées.

La matrice de confusion permet de découper les résultats d'un modèle de régression logistique binaire en quatre composantes :

- Les vrais positifs : Ce sont les prédictions qui valent 1 quand les données réelles valent 1.
- Les faux positifs : Ce sont les prédictions qui valent 1 quand les données réelles valent 0.
- Les vrais négatifs : Ce sont les prédictions qui valent 0 quand les données réelles valent 0.
- Les faux négatifs : Ce sont les prédictions qui valent 0 quand les données réelles valent 1.

Concrètement dans notre cas, les vrais positifs seront les assurés qui sont allés dans un garage agréé et que notre modèle a bien prédit alors que les faux négatifs seront les assurés qui sont allés dans un garage agréé mais que le modèle a prédit à 0.

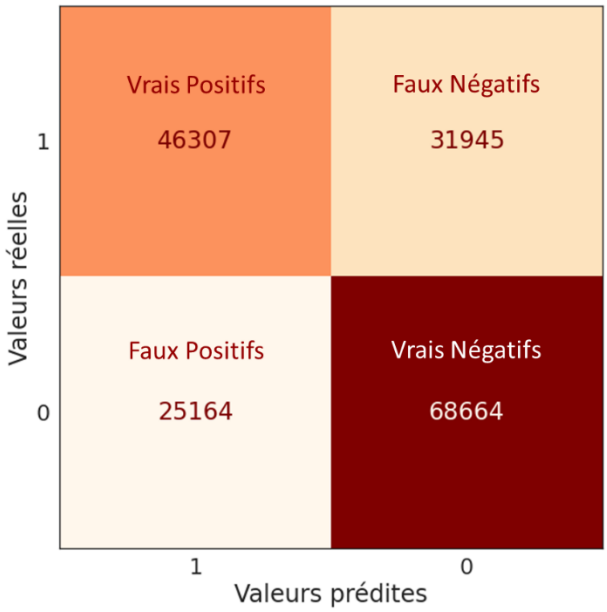



*Figure 28 - Matrice de confusion avec un seuil de 50% sur la base d'entrainement*

La matrice de confusion ci-dessus est celle de notre modèle de régression logistique avec toutes les variables de la base sur les données d'entrainement, pour lesquels les prédictions sont séparés en deux à partir d'une probabilité de 50%. Il est à noter qu'en faisant varier ce seuil ou *threshold* de 50%, la matrice varie. Les premières observations que nous pouvons faire à ce sujet sont que sur chaque colonne de valeurs prédites, les *vrais* sont plus grands que les *faux*, ce qui prouve que le modèle est utile et que les *négatifs* semblent être mieux prédits que les *positifs*.

Mais ces quelques observations ne sont pas suffisantes pour savoir si notre modèle est bon ou pour le comparer à d'autres modèles, pour ça nous aurons besoin de nouveaux indicateurs :

- On appelle *précision*, le taux de vrais positifs sur la totalité des positifs prédits par le modèle. Cet indicateur permet de mesurer à quel point notre modèle prédit correctement les positifs. Sa formule mathématique est :

$$Precision = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Positifs}.$$

  Pour cette matrice de confusion la précision est de 64.8%.

- Le rappel désigne le taux de vrais positifs sur la totalité des réels positifs dans la base de données. Cet indicateur indique la part de positif que notre modèle réussit à capter, tout en minimisant la part de négatifs. Sa formule mathématique est :

$$Recall = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Négatifs}.$$

  Pour cette matrice de confusion le rappel est de 59.2%.

- La justesse ou accuracy est la part de bonnes prédictions sur la totalité de la base. C'est un des indicateurs le plus utilisé pour jauger un modèle.

$$Accuracy = \frac{Vrais\ Positifs + Vrais\ Négatifs}{Vrais\ Positifs + Faux\ Positifs + Faux\ Négatifs + Vrais\ Négatifs}.$$

  Pour cette matrice de confusion l'accuracy est de 66,8%.

- Un autre indice que nous prendrons en compte pour juger notre modèle est si le taux d'orientation prédit est proche du taux d'orientation réel. La formule avec les informations de la matrice de confusion est :

$$Taux\ d'orientation = \frac{Vrais\ Positifs + Faux\ Positifs}{Vrais\ Positifs + Faux\ Positifs + Faux\ Négatifs + Vrais\ Négatifs}.$$

  Pour cette matrice de confusion il est de 41,5% alors que pour rappel, sur la base d'entrainement, il est de 45,5%.

Ce taux d'orientation proche de 50% nous place dans une situation de données quasiment équilibrées. L'accuracy sera donc une bonne métrique de mesure de qualité du modèle. Au contraire lorsqu'une base de données est composée de données à prédire très déséquilibrées, il est simple d'avoir une bonne accuracy avec un mauvais modèle. Par exemple si nous avions que 5% de 1 à prédire il suffirait de tout prédire à 0 et nous aurions une accuracy de 95%. Ici ce n'est pas le cas, si nous prédisons trop de 1 ça portera préjudice au résultat sur les 0 et inversement. L'accuracy est donc une bonne métrique de mesure de qualité de notre modèle ou encore de comparaison de modèle.

Pour avoir des accuracy comparables entre modèle il faut cependant trouver un endroit fixe pour la prendre et si possible à l'endroit où nous considérons que notre modèle est le meilleur. Nous utiliserons pour cela le rappel et la précision. Ce sont deux indicateurs qui comme l'accuracy d'une matrice varie selon le seuil utilisé pour la prédiction de 1 et de 0. Ce sont deux métriques opposées, c'est-à-dire que plus l'une grandit plus

l'autre décroit. Le croisement de ces deux indicateurs est donc un des points ou notre modèle est le plus performant.

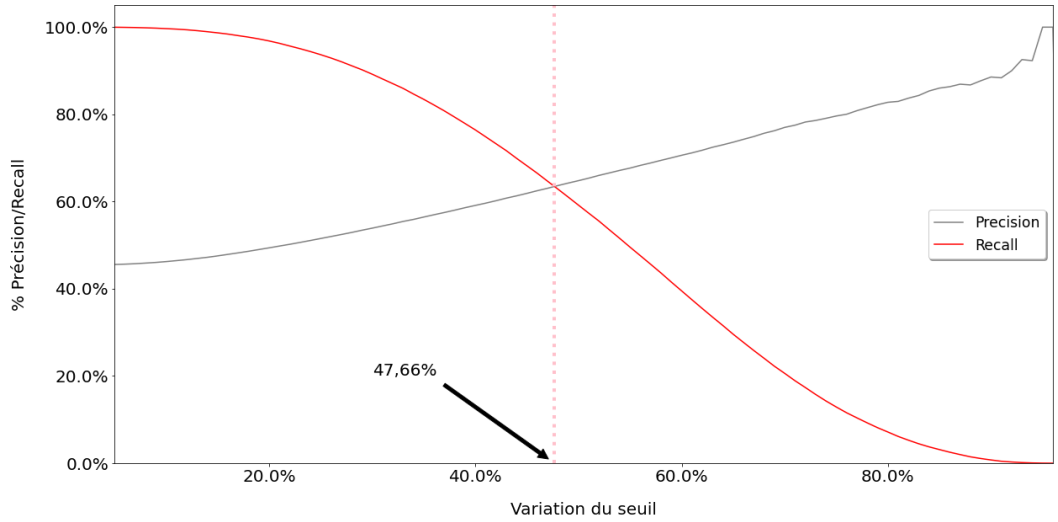
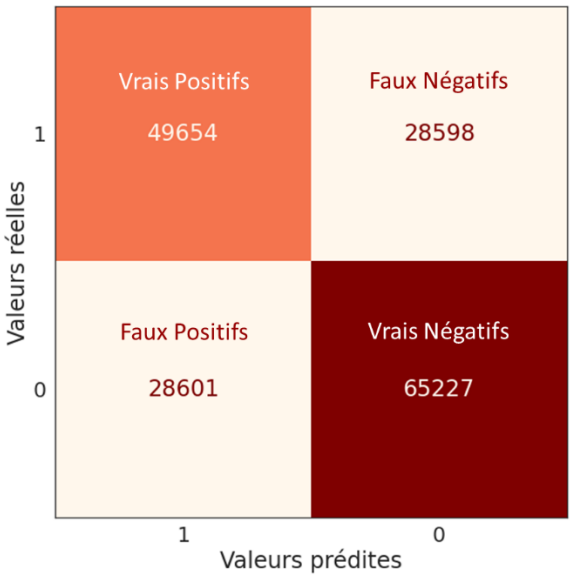



*Figure 29 - Précision et rappel sur le modèle total en fonction du seuil*

Le rappel est proche de 100% lorsque le seuil est inférieur à 5%, ce qui est assez logique vu qu'à ce seuil toutes les valeurs ayant selon le modèle plus de 5% de chance d'aller dans un garage agréé sont prédits dans un garage agréé. Il n'y a donc que les plus récalcitrant qui sont prédits comme 0 et pour ceux-là l'erreur est très faible, il n'y a donc plus de faux négatifs. A l'inverse lorsque le seuil approche de 100% la précision atteint 100% car nous ne prédisons comme allant dans un garage agréé ceux qui ont énormément de chance d'après le modèle d'y aller il n'y a donc plus de faux positifs.

En poussant plus loin l'analyse, nous observons que le rappel reste au-dessus de 80% jusqu'à un seuil d'environ 35%, ce qui signifie que nous minimisons bien la part de faux négatifs par rapport aux vrais positifs. Cependant la précision quant à elle pour arriver à 80% a besoin d'arriver à un seuil de 75% et nous avons même besoin d'arriver à un seuil de 95% pour atteindre les 100% de précision. Ça signifie que nous avons plus de mal à capter l'information sur les positifs. Le modèle permet tout de même d'être vraiment bon sur ces deux zones. Ce qui laisse une zone grise entre 40 et 75% ce qui est assez normal car les clients qui sont dans la moyenne sont plus compliqués à prédire. En résumé même si notre modèle réussit à bien prédire les personnes qui auront envie d'aller dans un garage agréé il est meilleur sur ceux qui n'auront pas envie d'y aller.

Ces deux courbes se croisent pour le seuil de 47,66% nous allons donc sortir la matrice de confusion et les différentes métriques vu précédemment pour ce seuil.




| | Taux |
|---|---|
| Seuil | 47,7% |
| Precision | 63,5% |
| Recall | 63,5% |
| Accuracy | 66,8% |
| Orientation | 45,5% |

*Figure 30 - Matrice de confusion et métriques associées avec un seuil optimisé sur la base d'entrainement*

Avec ce nouveau seuil non seulement l'accuracy est resté stable mais nous avons optimisé la précision et le rappel et le taux d'orientation est de 45,5% comme sur la base d'entrainement. C'est donc le point parfait pour comparer nos futurs modèles. Ce premier modèle est encourageant, nous réussissons dans plus de 2 cas sur 3 à correctement prédire quel assuré va aller ou non dans un garage agréé. De plus comme nous l'avons vu sur la figure 25 notre modèle prédit bien les clients qui n'auront pas envie d'aller dans un garage agréé et plutôt bien ceux qui auront envie d'y aller mais il existe une zone grise entre les deux plus compliqué à prédire.

#### *III.1.2.3. Optimisation du modèle : Méthode Forward*

Nous avons lancé un premier modèle avec la totalité des variables qui est encourageant, mais il a besoin de beaucoup de variables afin de tourner. Il demande donc de la puissance et des prétraitements et peut amener à sur-apprendre. Nous verrons dans cette partie comment supprimer certaines variables tout en gardant la qualité de notre modèle sur la base d'entrainement puis nous regarderons dans la prochaine partie comment tous ces modèles se comportent sur la base de test. Pour ça nous allons utiliser la méthode forward sur trois mesures d'erreurs vues précédemment. Mais avant nous définirons cette méthode.

##### III.1.2.3.1. Définition de la méthode

Le principe de cette méthode est de tester une à une les variables de notre modèle et de voir comment varie l'erreur choisit à chaque ajout de variables. Nous détaillerons ci-dessous l'algorithme de cette méthode.

1. Tout d'abord choisir la mesure d'erreur à utiliser, dans notre cas nous utiliserons en premier lieu le $Pseudo\ R\text{-}Square$ du modèle que nous prendrons donc comme exemple pour cet algorithme.
2. Initialiser :
    a. Une liste de variable nul dans laquelle nous mettrons à chaque pas la meilleure variable, noté : *List_var*.
    b. Un meilleur $Pseudo\ R\text{-}Square$ à 0 pour commencer, noté : *Best_prs*.

c. Une condition d'amélioration du $Pseudo\ R\text{-}Square$, noté : *Cdt_prs*. Pour se jauger celui de base avec toutes les variables est de 0,1204. Nous allons donc prendre une amélioration de 0,001.

3. Lancer notre modèle de régression logistique avec la première variable de la liste :
   a. Si le modèle avec la première variable améliore le $Pseudo\ R\text{-}Square$ de 0,001 alors :
      i. *List_var* est composée de cette première variable.
      ii. *Best_prs* est remplacé par le $Pseudo\ R\text{-}Square$ de ce modèle
   b. Sinon rien ne se passe et nous testons la seconde variable.

4. Lancer notre modèle de régression logistique avec la deuxième variable de la liste :
   a. Si le modèle avec la deuxième variable a un meilleur $Pseudo\ R\text{-}Square$ que celui avec la première :
      i. *List_var* est composée de cette deuxième variable qui remplace la première.
      ii. *Best_prs* est remplacé par le $Pseudo\ R\text{-}Square$ de ce modèle.
   b. Sinon rien ne se passe et nous testons la troisième variable.

5. Faire ainsi de suite avec toutes les variables du modèle, à la fin de ce premier tour nous avons donc *List_var* avec une seule variable, celle qui améliore le plus le $Pseudo\ R\text{-}Square$.

6. Relancer le même processus avec toutes les variables restantes afin d'obtenir la deuxième variable la plus importante et ainsi de suite jusqu'à avoir les variables les plus discriminantes pour le $Pseudo\ R\text{-}Square$. La condition d'arrêt est le fait d'avoir tester toutes les variables restantes et qu'aucune d'entre elles n'apporte au moins 0,001 de plus au $Pseudo\ R\text{-}Square$.

#### III.1.2.3.2. Méthode forward appliquée au $Pseudo\ R\text{-}Square$

Cet algorithme n'est malheureusement pas implémenté dans Python, nous l'avons donc créé en nous aidant de la page internet : (Schumacher, 2015), nous trouverons le code dans l'Annexes III : Code de la méthode forward sous Python. Une fois cette étape accomplie, nous avons pu le lancer sur notre base de données et récupérer les résultats ci-dessous :

| Variables | Null | Regroupement code postal | Taux d'orientation sur antérieur | Gestion | Regroupement de modèle | Regroupement age conducteur | Année de survenance | Toutes les variables | Cdt_prs |
|---|---|---|---|---|---|---|---|---|---|
| Pseudo R-Square cumulé | 0 | 0,06 | 0,089 | 0,1028 | 0,1138 | 0,1163 | 0,1185 | 0,1204 | 0,001 |
| Gain sur le Pseudo R-Square total (%) | | + 49,8% | + 24,1% | + 11,5% | + 9,1% | + 2,1% | + 1,8% | + 1,6% | + 0,8% |

*Tableau 18 - $Pseudo\ R\text{-}Square$* en fonction des variables ajoutées

Dans ce tableau nous trouvons de gauche à droite les variables qui impactent le plus le Pseudo R-square et donc qui ont été conservées par le modèle forward. Le principe de la méthode étant de choisir quelle variable rajouter au modèle à chaque pas lorsqu'on est au niveau de la variable « Taux d'orientation sur antérieur », le score du pseudo R square est celui du modèle comprenant les deux variables « Regroupement code postal » et « Taux d'orientation sur antérieur », et ainsi de suite après l'ajout de chaque variable. Le gain calculé en dessous est en pourcentage l'apport de chaque variable au Pseudo R-Square total, il est calculé ainsi :

$$Gain\ Pseudo\ R\text{-}Square = \frac{Ps\ R\text{-}Sq\ nouveau\ modèle - Ps\ R\text{-}Sq\ ancien\ modèle}{Ps\ R\text{-}Sq\ modèle\ total - Ps\ R\text{-}Sq\ modèle\ Null}.$$

En utilisant cet indicateur nous remarquons que près de la moitié de l'information est amenée par le « Regroupement code postal » et presque un quart par le « Taux d'orientation sur antérieur ». Nous avons ensuite la « Gestion » du sinistre qui apporte un peu plus de 10% de l'information alors que le « Regroupement de modèle » de véhicule en amène un peu moins de 10%. Les deux variables restantes sont le « Regroupement âge du conducteur » et l'« Année de survenance » qui pèsent environ 2% du $Pseudo\ R\text{-}Square$ total. Une fois ces six variables rentrées dans notre modèle de régression logistique nous atteignons un $Pseudo\ R\text{-}Square$ de 0,1185, il ne reste que 1,6% d'information manquante portée par les 52 variables restantes. De plus la condition que nous avions mis pour que le modèle garde la variable suivante est qu'elle apporte 0,001 au $Pseudo\ R\text{-}Square$ ce qui correspond à 0,8% soit la moitié de l'effet restant. Ce n'est donc pas étonnant que le modèle s'arrête une fois ces six variables ajoutées car elle capte à elle seule 98,4% de l'information au titre du $Pseudo\ R\text{-}Square$.

#### III.1.2.3.3. Méthode forward appliquée à la p-value

Nous avons appliqué ce même algorithme mais cette fois à la p-value de chaque variable en gardant celles qui avaient une p-value inférieur à 1%.

| Variables | Regroupement code postal | Classe SRA | Regroupement de modèle | Gestion | Taux d'orientation sur antérieur | Regroupement age conducteur | ... | Profession Autre | ... | Densité de garage agréé par code postale | Cdt_pvalue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ordre d'entrée de la variable | 1 | 2 | 3 | 4 | 5 | 6 | ... | 9 | … | 23 | |
| P-value lors de l'arrivée de la | 0 | 0 | 0 | 0 | 0 | 0 | ... | 4,30E-31 | … | 0,0037 | 0,01 |
| P-value lors du modèle final | 0 | **0,042** | 0 | 0 | 0 | 0 | ... | **0,214** | ... | 0,0037 | |

*Tableau 19 - $P\text{-}value$* en fonction des variables rentrées dans le modèle

Le tableau ci-dessus, nous retrouvons de gauche à droite les variables de notre modèle qui sont ressorties comme ayant la p-value la plus faible. Le premier enseignement de ce tableau est que cinq des six variables du modèle précédent ressortent dans les p-value les plus faibles pas exactement dans le même ordre mais pas loin.

La seule variable nouvelle qui se loge parmi ce top six est la classe SRA. Ce qui est intéressant c'est que cette variable est ressortie en deuxième position, ça veut donc dire que lorsqu'on a un modèle avec seulement le regroupement code postal et une seule autre variable c'est celle avec la p-value la plus faible. Cependant sur la figure 24 dans notre chapitre III-2-b-(1), la p-value de notre variable lorsque le modèle utilisé est celui avec toutes les variables est de 0,2 et même ici sur la troisième ligne du tableau la p-value de cette variable passe à 0,042 lorsque les 23 variables sont rentrées dans le modèle. Ça signifie que la Classe SRA apporte beaucoup d'information mais que cette information est déjà contenue dans d'autres variables qui sont ajoutées plus tard dans le modèle. Nous ne garderons donc ni la classe SRA ni la variable Profession Autre qui sont les deux variables avec une p-value supérieur à 0,01 à la fin de notre algorithme.

#### III.1.2.3.4. Méthode forward appliquée à l'accuracy

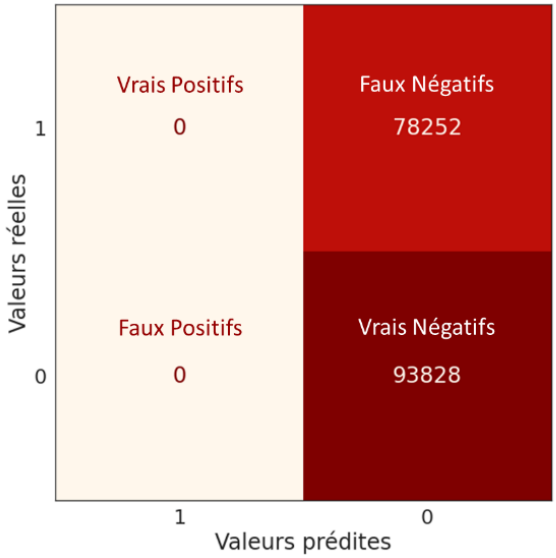
La dernière mesure que nous utiliserons avec cette méthode est l'accuracy. Nous prendrons cette métrique à un seuil de 50%, nous optimiserons les résultats des trois modèles créés en faisant jouer la précision et le rappel dans un second temps.

| Variables | Null | Regroupement code postal | Taux d'orientation sur antérieur | Gestion | Regroupement de modèle | Regroupement age conducteur | Groupe SRA | Toutes les variables | Cdt_acc |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy cumulé | 54,5% | 61,7% | 64,2% | 65,3% | 66,3% | 66,5% | 66,6% | 66,8% | 0,1% |
| Gain sur l'accuracy total (%) | | + 58,5% | + 20,3% | + 8,9% | + 8,1% | + 1,6% | + 0,8% | + 1,6% | + 0,8% |

*Tableau 20* - Accuracy en fonction des variables rentrée dans le modèle

Le premier enseignement de ce tableau est que l'accuracy de notre modèle sans aucune variable est de 54,5%. C'est un résultat étonnant car de première vue nous nous attendions plutôt à un modèle qui prédit aussi bien que le hasard, mais pas mieux, avec un score autour de 50%. Cet effet peut s'expliquer par les biais dont nous avions parlé plus haut sur les différentes mesures d'erreur de la matrice de confusion et ci-dessous nous allons détailler ces effets à l'aide de la matrice de confusion de ce modèle sans variable :

| | Score orientation | Prédiction orientation | Réel orientation |
|---|---|---|---|
| 0 | 0.454742 | 0 | 0 |
| 1 | 0.454742 | 0 | 1 |
| 2 | 0.454742 | 0 | 1 |
| 3 | 0.454742 | 0 | 0 |
| 4 | 0.454742 | 0 | 0 |




| | Taux |
|---|---|
| Seuil | 50,0% |
| Precision | #DIV/0! |
| Recall | 0,0% |
| Accuracy | 54,5% |
| Orientation | 0,0% |

*Figure 31 - Indicateur de performance du modèle sans variable*

Le modèle « Null » n'a que la constante, il prédit donc que tous les individus ont la même chance d'être orienté et cette probabilité est le taux d'orientation moyen de la base d'entrainement : 45,47%. Le seuil utilisé de base permettant de séparer les prédictions de manière binaire étant de 50%, tous les individus sont prédits comme non orientés. Et justement le taux d'orientation étant de 45,5% en classant tous les individus comme non orientés nous obtenons le complémentaire qui est 54,5% ce qui nous donne l'accuracy du modèle « Null ». Ce modèle est un parfait exemple de l'importance d'avoir plusieurs mesures d'erreurs. Il a une accuracy au-dessus de la moyenne mais avec un Rappel à 0% et une précision non calculable car avec des 0 sur la colonne de prédiction 1, un modèle qui classe aux hasards les assurés serait préférable à ce modèle même si son accuracy serait autour de 50%.

Si nous regardons maintenant le modèle avec toutes les variables, il a une accuracy de 66,8%. C'est pourquoi nous prendrons comme condition d'amélioration de cette mesure 0,1%, car ça revient à approcher le modèle total de près de 1% avec la même formule de comparaison que celle pour le $Pseudo\ R\text{-}Square$ :

$$Gain\ Accuracy = \frac{Accuracy\ nouveau\ modèle - Accuracy\ ancien\ modèle}{Accuracy\ modèle\ total - Accuracy\ modèle\ Null}.$$

Une fois encore cinq des six variables de notre modèle sont les mêmes que dans les deux modèles précédents. Les deux variables les plus importantes sont : le « Regroupement de code postal » apporte près

de 60% de l'information, et le « Taux d'orientation sur antérieur » qui apporte un cinquième de l'information. Puis la « Gestion » et le « Regroupement de modèle » apportent un peu plus de 8% de l'information et enfin le « Regroupement âge conducteur » et le « Groupe SRA » apportent à elles deux environ 2,5% d'information. Les cinquante-deux variables restantes n'offrent que 1,6% d'information supplémentaires. Nous avons donc un modèle robuste et qui est très proche du modèle avec cinquante-huit variables avec seulement six variables, ce qui aidera pour la rapidité de notre modèle mais aussi pour éviter le sur-apprentissage comme nous le verrons par la suite.

Nous avons à ce stade trois modèles de régression logistique à comparer avec notre modèle global, nous allons voir dans la prochaine partie lequel des modèles est le plus performant.

### III.1.3. Analyse des résultats

Dans cette partie nous allons comparer les trois modèles que nous venons de créer à l'aide des mesures d'erreurs de la matrice de confusion. Nous allons comparer les résultats non seulement sur la base d'entrainement mais aussi sur la base de test que nous n'avons pas encore utilisée jusqu'ici. Nous pourrons ainsi comparer nos modèles sur des données nouvelles mais aussi sur des données sur lesquelles le modèle n'a pas encore appris.

#### *Sur la base d'entrainement*

Tout comme nous l'avons fait pour le modèle global, la première étape sera de trouver, pour chacun des trois modèles, le point qui maximise à la fois le rappel et la précision.

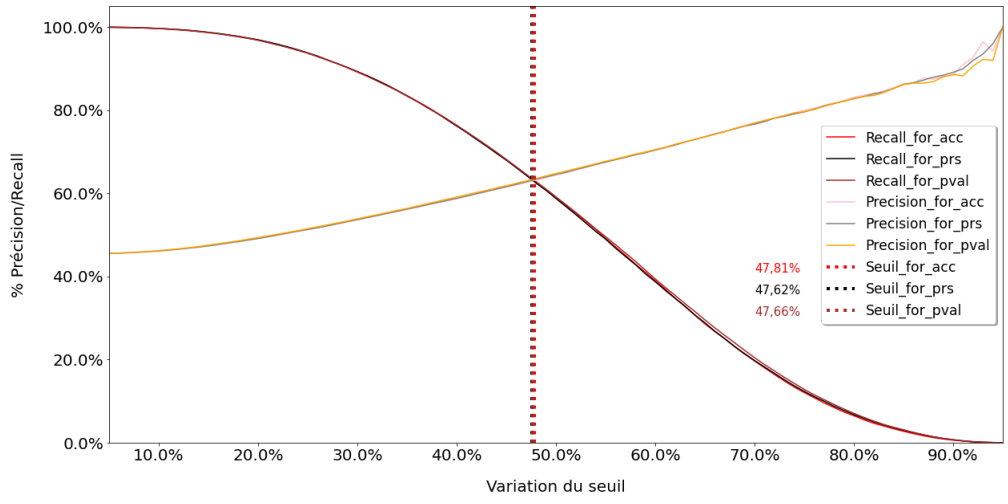



*Figure 32 - Evolution de la précision et du rappel en fonction du seuil et des trois modèles sur la base d'entrainement*

Sur le graphique ci-dessus, la précision et le rappel, en fonction du seuil a été tracé pour chacun des trois modèles créés avec la méthode *forward* dans la partie précédente. Ils sont distingués sur le graphique par leur suffixe : *« Acc »* pour l'accuracy, « *Prs* » pour le Pseudo R-Square et « *Pval* » pour la p-value, notations que nous réutiliserons par la suite. Quel que soit le modèle, les courbes de précision et les courbes de rappel sont très proches les unes des autres. Les seules différences notables se trouvent sur les courbes de précision au-delà du seuil de 80%. A ce niveau, la précision est moins bonne pour le modèle obtenu avec les 21 variables du modèle *Pval* qu'avec les deux autres modèles. Cela signifie que ce modèle prédit moins bien les clients qui ont plus de chance d'aller dans les garages agréés. Sur cette même section le modèle *Prs* semble plus régulier que le modèle *Acc*. Sur la figure ci-dessus, les trois seuils où se croisent le rappel et la précision, sont

également renseignés : ces derniers sont très proches les uns des autres et sont compris entre 47,62% et 47,81%. À présent l'accuracy de ces trois modèles va pouvoir être comparé à ces points d'intersection.

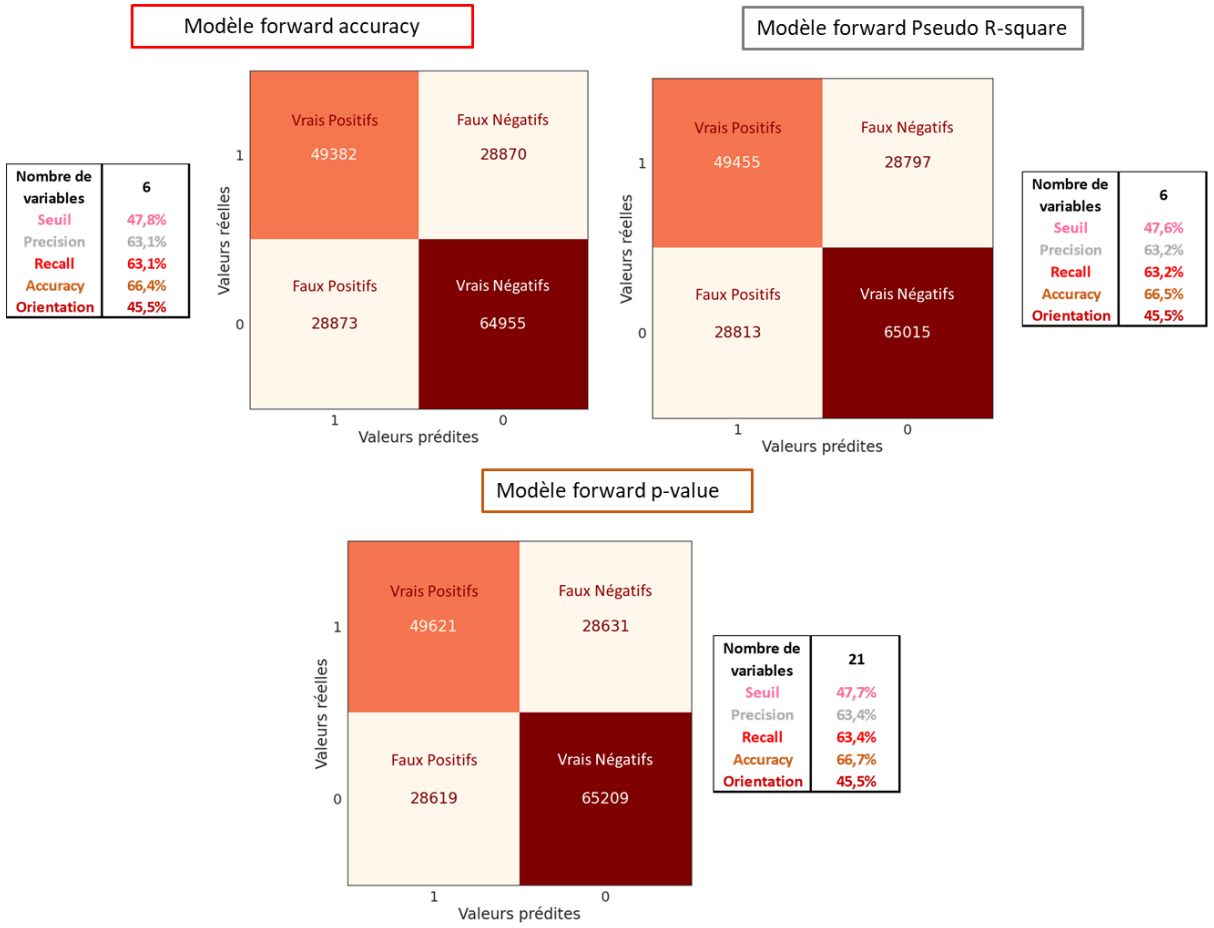



*Figure 33 - Matrice de confusion et métriques associées des trois modèles retenus sur la base d'entrainement*

La figure ci-dessus présente les matrices de confusion et les mesures d'erreurs associées, pour chacun des trois modèles créés dans la partie précédente. Les indicateurs importants, que sont le seuil auquel nous regardons ces mesures et le nombre de variables utilisées pour chaque modèle, sont également retrouvés. Nous observons que le modèle *Pval* a des résultats très proches de ceux du modèle avec la totalité des cinquante-huit variables dont nous pouvons retrouver la matrice de confusion en figure 26. Son seuil est le même et tous les autres indicateurs ne perdent que 0,1 points : par exemple, l'accuracy est de 66,7% lorsque celui du modèle global est de 66,8%. Nous ne perdons quasiment pas d'informations sur la base d'entrainement en retirant les trente-sept variables les moins discriminantes. Les deux matrices au-dessus nous apprennent qu'en ne gardant que six variables nous pouvons avoir des résultats presque aussi bons puisqu'on ne perd que 0,2 points sur tous nos indicateurs avec les variables du modèle *Prs* et 0,3 points sur le modèle *Acc*. Ce qui est intéressant en comparant ces deux modèles, c'est que le modèle *Prs* a un meilleur accuracy que le modèle *Acc* qui a pourtant été créé en optimisant l'accuracy. L'explication à cet effet est le seuil que nous avons utilisé pour la méthode forward qui était de 50% alors qu'ici, nous comparons nos modèles à un seuil optimal selon la précision et le rappel.

Après cette partie nous avons une première idée sur la qualité de nos modèles. Le modèle *Prs* semble le meilleur compromis entre l'optimisation du nombre de variables et de nos indicateurs d'erreurs. Nous allons vérifier cette théorie avec le jeu de données sur lequel le modèle n'a pas appris le jeu de test.

### *Sur la base de test*

Avoir mis de côté la base de test permet de vérifier que les résultats obtenus par notre modèle ne sont pas que le résultat d'un apprentissage sur la base mais qu'ils peuvent se généraliser au reste de nos données. Un exemple de variable qui est trop fine et entraine du sur-apprentissage est le regroupement de code INSEE.

| | Modèle total | | | | | |
|---|---|---|---|---|---|---|
| | Avec Regroupement code INSEE | | Avec Regroupement code Postal | | Sans ces deux regroupements | |
| | Entrainement | Test | Entrainement | Test | Entrainement | Test |
| Accuracy Seuil 50% | **70,1%** | 62,7% | **66,8%** | 64,1% | 64,1% | 63,6% |

*Tableau 21 - Accuracy du modèle total avec ou sans les variables de regroupement géographique*

Le tableau ci-dessus présente l'accuracy calculé sur le modèle avec toutes les variables :

- Avec le regroupement de code INSEE et sans le regroupement de code Postal, que nous noterons *Tot_INSEE*,
- Avec le regroupement de code Postal et sans le regroupement de code INSEE, qui est notre modèle total habituel noté *Tot*,
- Sans aucun de ces deux regroupements, que nous noterons *Tot_ss_GEO*,

Dans le modèle *Tot_INSEE* l'accuracy sur la base d'entrainement est 7,4 points plus haut que celui sur la base d'entrainement. Cet écart entre le jeu d'entrainement et le jeu de test s'appelle l'erreur de généralisation. C'est un écart très important qui démontre un problème de sur-apprentissage dans ce modèle. Par comparaison l'erreur de généralisation sur le modèle *Tot* n'est que de 2,6 points. Nous en déduisons donc que la variable de regroupement de code INSEE provoque presque 5 points de sur-apprentissage de plus que celle du regroupement code postal. De plus, si nous comparons le résultat de ces deux modèles sur la base test, nous observons que le modèle *Tot* est un meilleur prédicteur. C'est pour cette raison que nous avons écarté la variable de regroupement de code INSEE des bases de données. Nous pourrions nous poser la même question du sur-apprentissage sur la variable de regroupement du code postal, car dans le modèle *Tot_ss_GEO* l'erreur de généralisation n'est que de 0,5 points ; mais d'une part le modèle *Tot* prédit mieux sur la base de test et de l'autre pour réussir à approcher l'information qu'elle apporte sur la base de test il faut la remplacer par cinq variables. Nous accepterons donc le biais de sur-apprentissage apporté par cette variable en échange de son pouvoir prédictif fort.

Comme dans la partie précédente, la précision et le rappel seront comparés pour chacun des trois modèles créés et pour le modèle *Tot* qui n'a pas encore été essayé sur la base de test.

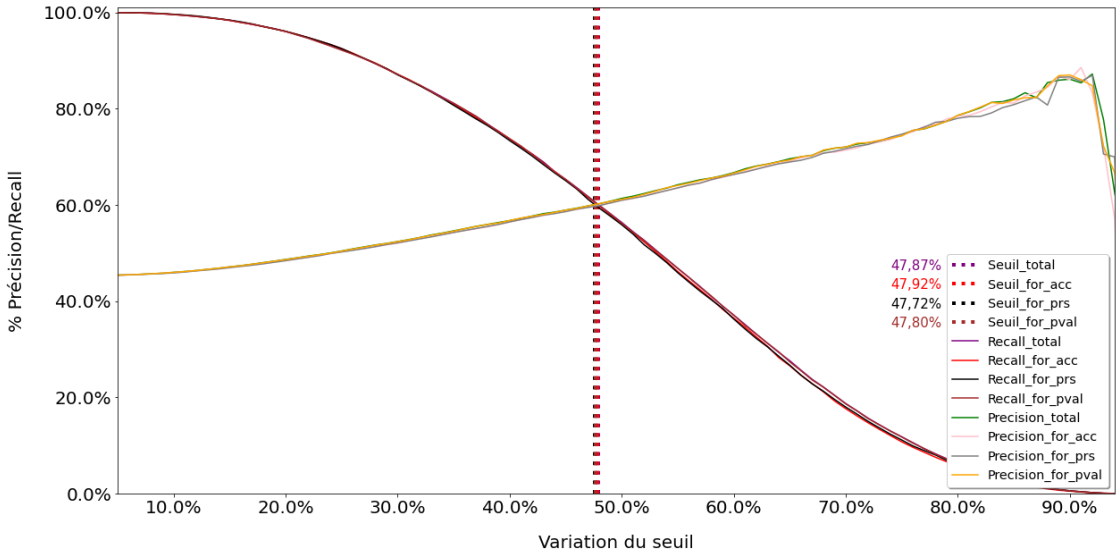



*Figure 34 - Évolution de la précision et du rappel en fonction du seuil et des quatre modèles sur la base de test*

Nous observons sur le graphique ci-dessus les mêmes tendances générales que sur la base d'entrainement. Quelques soient le modèle toutes les courbes de précision et de rappel sont proches les unes des autres et les principales différences observables se trouvent sur les courbes de précision lorsque le seuil est élevé. Sur la base de test, à l'inverse de la base d'entrainement, c'est le modèle *Prs* qui parait moins performant et plus erratique sur le haut de la courbe. L'autre point important est la précision de notre modèle qui chute au seuil de 90%. Cette baisse signifie que nous ne réussissons pas à prédire les quelques assurés qui ont l'appétence la plus forte à l'orientation, nous verrons dans la prochaine partie que nous prédisons plus facilement l'inverse.

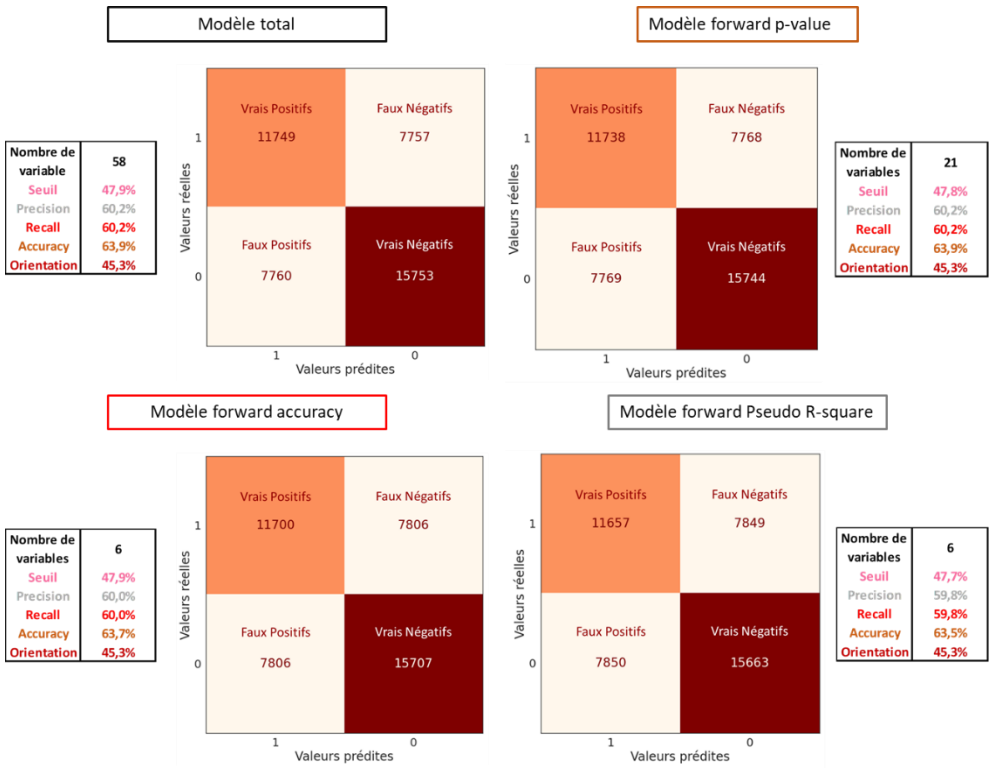



*Figure 35 - Matrice de confusion et métriques associées des quatre modèles sur la base de test*

La figure ci-dessus présente les matrices de confusion pour les trois modèles créés et pour le modèle *Tot* sur la base de test. Tout d'abord, le taux d'orientation de la base de test qui est de 45,3% est bien retrouvé pour chaque modèle, ce qui confirme le seuil de comparaison. En comparant les deux premiers modèles, il apparait que le résultat sur la base de test est la même avec les vingt et une variables ayant les plus petites p-value et toutes les variables de notre base. Le modèle *Pval* est donc plus intéressant car il est aussi performant que le modèle *Tot*, et ce, avec moins d'informations. La différence avec les modèles en-dessous est respectivement de 0,2 points et 0,4 points pour toutes les mesures de performances des modèles *Acc* et *Prs*. À l'inverse des résultats sur la base d'entrainement, le modèle *Acc* est donc plus performant que le modèle *Prs*.

Le modèle *Acc* sera gardé pour le scoring car il donne des résultats très proches des meilleurs modèles, avec seulement six variables. Afin d'optimiser notre modèle, le test de retirer la variable « Groupe SRA », a été effectué. C'était la variable qui apportait le moins d'information : il y a une perte de 0,2 points sur chaque mesure d'erreurs, ce qui correspond au même écart qu'en retirant les quinze variables de différence entre les modèles *Pval* et *Acc*. Cette variable sera donc gardée et le modèle *Acc* sera utilisé dans la suite du mémoire.

### *Description du modèle final*

Dans cette partie, la composition du modèle choisi pour créer le scoring, sera analysée.

| | |
|---|---|
| No. Observations: | 172080 |
| Df Residuals: | 172073 |
| Df Model: | 6 |
| Pseudo R-squ.: | 0.1169 |
| Log-Likelihood: | -1.0471e+05 |
| LL-Null: | -1.1857e+05 |
| LLR p-value: | 0.000 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] | Poids du coeff |
|---|---|---|---|---|---|---|---|
| const | -0.2196 | 0.005 | -41.634 | 0.000 | -0.230 | -0.209 | (hors const) |
| Reg_CodePOST_C | 0.5649 | 0.006 | 101.542 | 0.000 | 0.554 | 0.576 | 33% |
| Class_TO_ANT_vnum | 0.4316 | 0.006 | 72.785 | 0.000 | 0.420 | 0.443 | 25% |
| Gestion_num | -0.3070 | 0.005 | -56.963 | 0.000 | -0.318 | -0.296 | 18% |
| Reg_MODEL_LIB | 0.2310 | 0.006 | 38.278 | 0.000 | 0.219 | 0.243 | 13% |
| Reg_Age_conducteur | -0.1319 | 0.005 | -25.080 | 0.000 | -0.142 | -0.122 | 8% |
| Groupe_vnum | -0.0705 | 0.006 | -11.627 | 0.000 | -0.082 | -0.059 | 4% |

*Tableau 22 – Résultats de la régression logistique avec le modèle Acc*

La sortie Python ci-dessus représente les principaux indicateurs du modèle de régression logistique. Tout d'abord, le $Pseudo\ R\text{-}Square$ est de 11,69 %, celui du modèle *Tot* étant de 12,04% nous avons plus de 97% de l'information récupérée par nos six variables. Le LLR p-value est 0 ce qui signifie que le test de significativité visant à savoir si notre modèle apporte plus que le modèle sans variable, est particulièrement concluant.
Le tableau de droite présente tout d'abord les coefficients du modèle, qui sont des informations qui vont permettre de réappliquer ce modèle à toutes nouvelles données. Ces coefficients sont triés, en valeur absolue, du plus grand au plus petit. Il est intéressant de noter que ce tri suit le même que l'ordre d'entrée de nos variables dans le modèle *forward*. Le poids des coefficients n'est cependant pas le même que celui de nos variables dans l'apport à l'accuracy. Nous avons près d'un tiers de l'information qui est porté par le « regroupement code postal », un quart par le « Taux d'orientation sur antérieur », 18% par la gestion et un quart par les trois autres variables. L'importance des variables est plus répartie que dans l'importance de l'accuracy qui était porté par près de 60% par la première variable (voir Tableau 20).
De plus, nous pouvons vérifier que les p-values sont toutes à 0 il n'y a donc aucune variable présente qui est sans intérêt pour le modèle. La *standard error* pour chacun des coefficients est inférieure à 1% nous avons donc des coefficients robustes. Cet aspect se retrouve aussi sur les deux dernières colonnes. Nous sommes

sûrs à 95% que nos quatre premières variables sont dans une fourchette de plus ou moins 5% de leur valeur. Par exemple pour le « regroupement code postal » la fourchette est même de 1,9 % :

| Coeff Regroupement CP | | |
|---|---|---|
| 0,025 | 0,5 | 0,975 |
| 0,554 | 0,5649 | 0,576 |

| Besoin pour etre sur à 95% | |
|---|---|
| +/- | part du coeff |
| 0,011 | 1,9% |

*Tableau 23 - Calcul de l'écart au coefficient pour un ecart type de 95%*

Maintenant que l'analyse des paramètres du modèle final a été effectuée, l'analyse des corrélations des variables qui le composent, peut débuter.

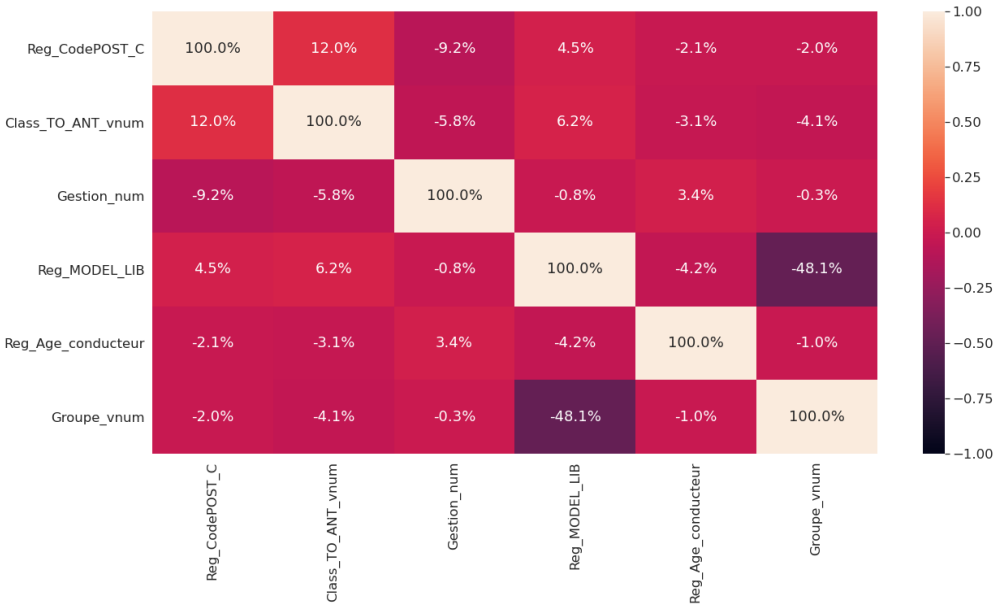



*Figure 36 - Matrice de corrélation des variables de la base final*

Nous observons que dans l'ensemble les variables sont peu corrélées les unes aux autres. Les deux seules variables qui sont corrélées de manière significatives sont le *Regroupement de modèle* et le *Groupe SRA*. Elles sont corrélées à un peu moins de 50% ce qui veut quand même dire que plus de la moitié de l'information contenue dans chacune de ces variables apporte quelque chose au modèle. De plus, ce sont deux des trois dernières variables en termes d'importance ce qui n'impactera que dans une moindre mesure notre modèle. Pour finir, nous avons testé le modèle sans le *Groupe SRA* et il était moins performant. Nous n'exclurons donc pas cette variable. Sur le reste de la base, les corrélations entre les variables sont faibles (en dessous de 12,0%), nous avons donc des variables avec une faible dépendance les unes par rapport aux autres.

La dernière analyse que nous ferons sur notre modèle est la comparaison de la précision sur nos véhicules agréés et non agréés, respectivement les 1 et les 0 de notre matrice de confusion. Pour ça nous utiliserons un nouvel indicateur :

- La précision négative ou Negative Predictive Value est le taux de vrais négatifs sur la totalité des négatifs prédits par le modèle. Cet indicateur permet de mesurer à quel point notre modèle prédit correctement les négatifs. Sa formule mathématique est :

$$Negative\ predictive\ value = \frac{Vrais\ Négatifs}{Vrais\ négatifs + Faux\ Négatifs}.$$

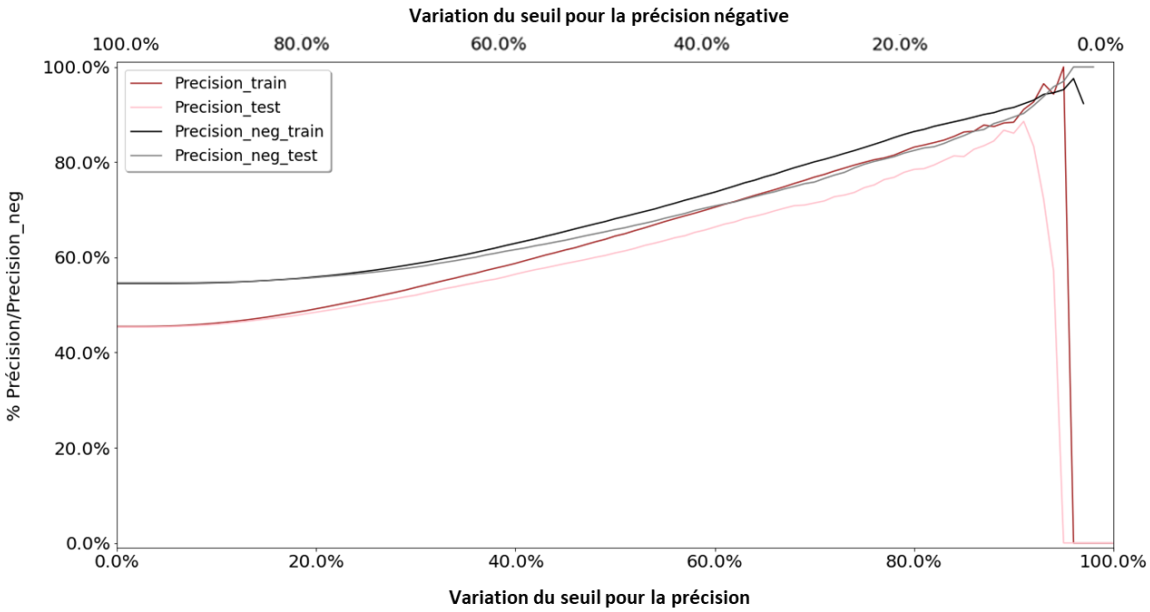



*Figure 37 - Précision et précision négative en fonction du seuil*

Le graphique ci-dessus présente la variation de la précision en fonction du seuil qui évolue de manière croissante et la précision négative avec cette fois, un seuil décroissant. Le seuil évolue dans un sens opposé pour ces deux métriques afin de pouvoir comparer leur évolution. Nous observons au début des courbes un écart de 10 points entre les deux mesures d'erreurs, avec une précision autour de 45% et une précision autour de 55%. Cet écart est normal, pour la précision le seuil étant très bas la matrice au début de la courbe ne prédit que des 1. Nous trouvons donc tous les vrais positifs et le reste des données étant des faux positifs, c'est normal que nous retrouvions un taux proche du taux d'orientation. Le même effet se produit pour la précision négative, mais comme il y a 55% de 0 c'est ce taux que nous retrouvons. En examinant les courbes sur la base d'entrainement, l'écart se réduit progressivement mais elles ne se rejoignent qu'à la fin. Si notre modèle prédisait aussi bien les positifs que les négatifs nos deux courbes se rejoindrait dès 50%. Cet effet est encore plus visible sur notre base de test, la courbe de précision négative reste au-dessus tout du long avec un écart minimal de 2 points au seuil de 90%. La courbe de la précision chute ensuite car nous n'arrivons pas à prédire les clients avec la plus grande chance d'être orientés, elle chute même jusqu'à 0 ce qui signifie que le client avec la plus forte prédiction n'est pas allé dans un garage agréé. A l'inverse, pour la précision négative, la courbe de test atteint les 100%, ce qui signifie que les clients avec la plus faible probabilité d'aller dans un garage agréé, ne s'y sont pas rendus. L'espace vide à la fin de cette courbe en deçà du seuil de précision négative de 2% traduit le fait qu'aucun client n'a une probabilité plus faible que 2% d'aller dans un garage agréé selon notre modèle.

Nous avons prédit pour chacun de nos clients sinistrés, quelle était sa probabilité d'aller dans un garage agréé. Dans la prochaine partie nous créerons, à partir de cette probabilité, des groupes de clients ayant un score d'appétence proche.

## III.2. Création du scoring

Maintenant que nous connaissons l'appétence de chaque client à aller vers un réparateur du réseau de garages agréés Assercar, le but sera de regrouper notre population en groupes de clients homogènes au regard de ce nouvel indicateur. Créer ces groupes nous permettra de mettre une note à chacun de nos clients.

Il s'agit du même principe que la note google qui va d'une à cinq étoiles. Ce regroupement permettra de faire ressortir les groupes de « bons » et « mauvais » clients au regard de l'orientation. Comme nous le verrons plus tard, cet aspect sera intéressant d'un point de vue opérationnel. D'autre part, comme nous l'avons expliqué dans la partie précédente sur la discrétisation, constituer des regroupements permettra de rendre la variable plus simple à interpréter par les modèles et d'éviter le sur-apprentissage.

### III.2.1. Application de trois méthodes de discrétisation

Dans cette partie nous allons créer un scoring à partir de trois méthodes différentes : La méthode des amplitudes, la méthode des quantiles et la méthode des Kmeans. Nous comparerons ensuite ces trois méthodes afin de choisir celle qui regroupe le mieux nos données. Mais tout d'abord nous allons commencer par décrire la création du scoring sur la seule des trois méthodes qui n'a pas encore été définie, la méthode des amplitudes.

#### *III.2.1.1. La méthode des amplitudes*

Le but de cette méthode est d'avoir la même amplitude d'orientation dans chacune des classes créées. Nous définissons l'amplitude comme étant la différence entre la plus petite et la plus grande valeur. Nous allons détailler les différentes étapes de cette méthode :

1. Tout d'abord choisir le nombre *n* de groupes à créer.
2. Trier les valeurs du taux d'orientation de manière croissante.
3. Diviser la différence entre la plus petite valeur *v* et la plus grande valeur *V* de taux d'orientation par le nombre de groupes afin d'obtenir la même amplitude *a* pour chaque groupe :
$$a = \frac{V-v}{n}.$$
4. Répartir les individus par taux d'orientation dans les n groupes créés :

$$[v\,;\ v+a],]\mathrm{v}+a\,;\ v+2a],\ldots,]v+(n-1)\times a\,;\ v+n\times a] \text{ avec } V = v+n\times a.$$

Nous allons initialiser cette méthode en choisissant notre nombre de groupes. Pour ce faire nous allons lancer la méthode présentée ci-dessus à notre base de données avec un nombre de groupe allant de deux à trente. Ensuite nous minimiserons l'inertie intra-groupe et le nombre de groupes en utilisant la méthode du coude.

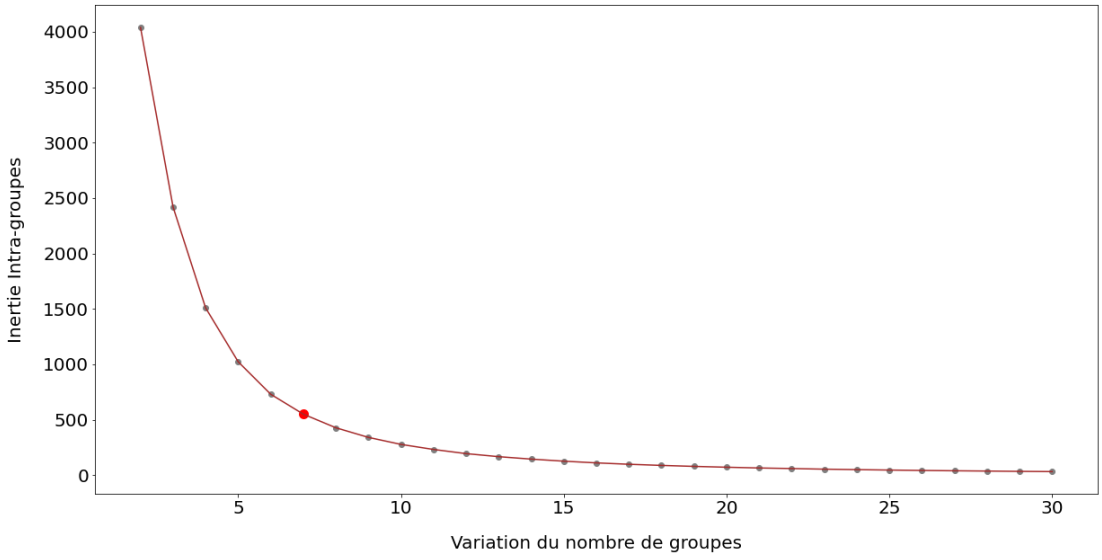



*Figure 38 - Inertie Intra-groupes en fonction du nombre de groupes avec la méthode des amplitudes*

Le graphique ci-dessus représente l'évolution de l'inertie intra-groupes en fonction du nombre de groupes utilisés lors de la méthode des amplitudes. L'inertie passe d'environ 4 900 lorsque nos données ne sont séparées qu'en deux à environ 35 lorsque nous séparons nos données en trente groupes. La cassure sur cette courbe se fait entre les nombres de groupes 5 et 9 c'est pourquoi nous avons choisi le point entre les deux, 7 comme le meilleur nombre de classes.

Nous pouvons maintenant appliquer la méthode des amplitudes et séparer en 7 groupes les prédictions d'orientation sur la base totale.

| Scoring_OR | Nombre de client | Prédiction min | Prédiction max | Prédiction moyenne | Taux d'orientation réel | Écart Préd/Tx Or |
|---|---|---|---|---|---|---|
| 1 | 11 469 | 1.6% | 15.0% | 10.9% | 11.7% | -0.8pts |
| 2 | 35 346 | 15.0% | 28.4% | 22.3% | 22.5% | -0.2pts |
| 3 | 47 618 | 28.4% | 41.8% | 35.2% | 35.7% | -0.5pts |
| 4 | 50 086 | 41.8% | 55.1% | 48.4% | 48.6% | -0.2pts |
| 5 | 42 142 | 55.1% | 68.5% | 61.5% | 61.4% | 0.1pts |
| 6 | 22 995 | 68.5% | 81.9% | 74.2% | 72.8% | 1.4pts |
| 7 | 5 443 | 81.9% | 95.3% | 86.0% | 83.2% | 2.8pts |

*Tableau 24 - Discrétisation de l'appétence client prédite en sept classes avec la méthode des amplitudes*

Le tableau ci-dessous décrit la répartition de nos prédictions, l'appétence du client à aller vers un garage agréé, dans les sept groupes que nous avons créés avec la méthode des amplitudes. Nous allons détailler les sept colonnes qui le composent :

- Le scoring est la discrétisation de notre variable prédiction créée à l'aide de la méthode des amplitudes. C'est donc le score du client, il va de 1 pour un assuré qui a très peu de chance d'aller dans un garage agréé à 7 pour ceux qui, à l'inverse, ont plus de propension à l'orientation.
- Le nombre de client est le nombre d'individus se trouvant dans chaque classe.
- La prédiction *min* est la plus petite probabilité qu'un client aille dans un garage agréé pour chaque classe.
- La prédiction *max* est la plus forte probabilité qu'un client aille dans un garage agréé pour chaque classe.
- Le taux d'orientation réel est le taux d'orientation moyen sur les clients, calculé sur chaque classe.
- La prédiction moyenne est la moyenne de notre variable de prédiction de probabilité qu'un client aille dans un garage agréé, construite dans la partie précédente, calculée pour chacune des classes.
- La dernière colonne est la différence entre la prédiction moyenne et le taux d'orientation réel, il permet de voir l'erreur de notre modèle à l'intérieur de chaque classe du scoring.

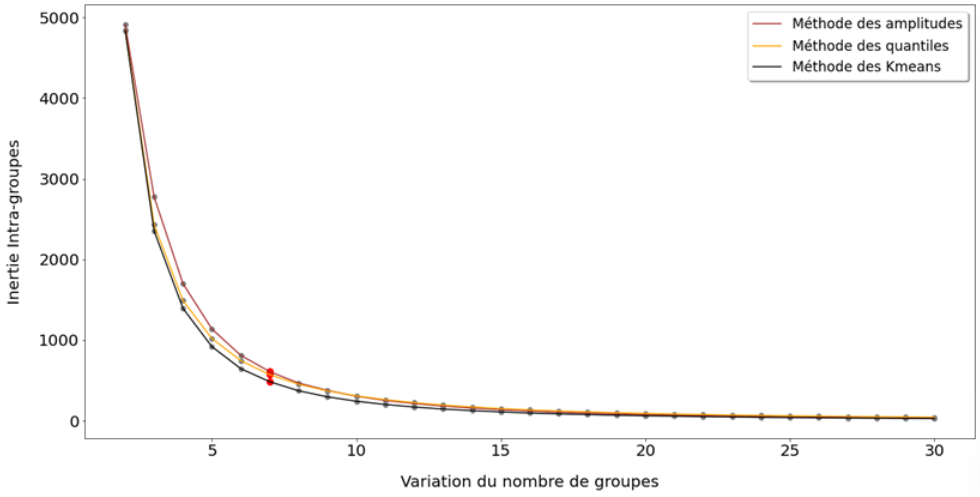
En comparant nos colonnes de taux d'orientation réel et de prédiction moyenne, nous nous apercevons qu'ils sont dans la majorité des cas assez proches. Il n'y a que sur le groupe 7 que nous prédisons un taux d'orientation de 86,0%, supérieur de 2,8 points à l'orientation réel pour ce groupe. Cet écart sur ce groupe peut s'expliquer par le faible nombre de clients le composant, c'est le groupe avec le nombre de clients le plus faible. Nous n'avons que 5 443 clients dans ce groupe, soit 2,5% de la base ; là où le groupe 4 qui est le plus grand, contient 50 086 soit 23,3% de la base. Ce ne sont pas que deux cas isolés : le nombre de clients

n'est pas réparti équitablement à travers les 7 classes. Nous avons beaucoup de clients sur les classes du milieu et peu sur les extrêmes. Une trop faible représentativité dans une classe peut induire un biais, comme pour le groupe 7. Les colonnes *Prédiction min* et *Prédiction max* quant à elles décrivent les bornes d'orientation de chacune de nos classes. L'amplitude de chacune de nos classes est d'environ 13,4 points et nos prédictions vont de 1,6 % de chance d'aller dans un garage agréé pour la plus faible à 95,3% pour la plus forte.

Dans cette partie nous avons décrit une première méthode de discrétisation de nos prédictions, nous allons maintenant comparer les résultats obtenus avec deux autres méthodes.

#### *III.2.1.2. Comparaison de nos trois méthodes*

Dans la partie précédente, sur la discrétisation des variables, nous avions vu deux autres méthodes mathématiques pour regrouper des variables : la méthode des quantiles et la méthode des Kmeans. Dans cette partie nous allons donc tester les trois méthodes de discrétisation vues dans ce mémoire, et comparer les résultats.




| Nb groupes | Kmeans | Quantile | Amplitude |
|---|---|---|---|
| 2 | 4839.04 | 4843.31 | 4912.59 |
| 3 | 2355.19 | 2430.56 | 2775.90 |
| 4 | 1393.71 | 1489.72 | 1700.08 |
| 5 | 919.88 | 1021.15 | 1138.92 |
| 6 | 646.83 | 746.34 | 811.42 |
| 7 | 484.86 | 572.02 | 610.18 |
| 8 | 372.65 | 453.66 | 468.29 |
| 9 | 297.24 | 371.51 | 377.48 |
| 10 | 243.16 | 309.21 | 304.38 |
| 29 | 30.17 | 47.89 | 37.11 |
| 30 | 28.27 | 45.09 | 34.94 |

*Figure 39 - Évolution de l'inertie intra-groupes en fonction du nombre de groupes choisis et de la méthode utilisé*

La figure ci-dessus présente l'inertie intra-groupes en fonction du nombre de groupes choisis et des trois méthodes utilisées. Nous pouvons noter que ces trois courbes ont la même allure. La cassure se fera donc pour les trois au niveau de 7 groupes. Nous remarquons que la courbe de la méthode des amplitudes est au-dessus des deux autres courbes jusqu'à 9 groupes, c'est donc la moins adaptée pour un faible nombre de groupes. L'inertie intra-groupes de cette méthode est ensuite plus basse que celle de la méthode des quantiles, et elle reste en dessous de celle-ci jusqu'à la fin de notre courbe ; la méthode des amplitudes est donc plus adaptée que celle des quantiles pour un nombre de groupes plus important. La méthode des Kmeans a une inertie inférieure aux deux autres méthodes quelques soit le nombre de groupes choisi. C'est donc cette méthode que nous privilégierons, au premier abord.

Nous allons maintenant appliquer ces trois méthodes avec 7 groupes à notre base de données et regarder comment sont répartis le nombre et la probabilité de notre prédiction dans chacun de nos scoring. Cette analyse permettra de confirmer ou de réfuter l'utilisation de la méthode des Kmeans.

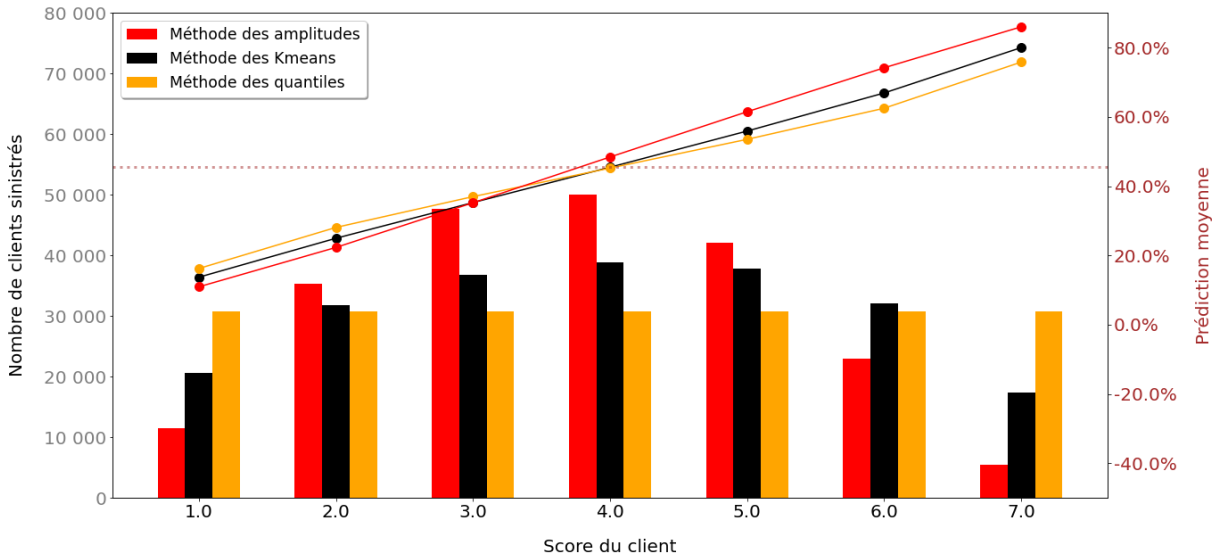



*Figure 40 - Nombre de clients sinistrés et prédiction moyenne de leur appétence à l'orientation en fonction du score client*

Le graphique ci-dessus présente, pour nos trois méthodes, le nombre de clients sinistrés et la prédiction de l'appétence à l'orientation moyenne en fonction du score du client.

En analysant tout d'abord le diagramme en bâton, nous observons que le nombre de clients est très différent d'une méthode à l'autre.

Avec la méthode des amplitudes ce nombre est faible sur les extrémités, avec un minimum à un peu plus de 5 000 pour le dernier groupe ; cependant c'est celui qui a le plus de clients sur les scores 2 à 5, avec notamment un pic à plus de 50 000 pour le groupe 4, soit un écart de plus de 10 000 avec les deux autres groupes. Cette trop grande volatilité sur le nombre de clients est un problème car il donne trop de poids à notre score sur les scores entre 3 et 5 et peut entrainer un sur-apprentissage sur les scores extrêmes.

Avec la méthode des quantiles, du fait de sa construction, le nombre de clients par groupe est parfaitement stable. Cet aspect est très positif pour l'apprentissage d'un futur modèle sur ces données.

La méthode des Kmeans, quant à elle, donne un nombre de sinistres qui se trouve entre les deux autres méthodes pour tous les groupes, à part pour le 6 où il est légèrement supérieur à la méthode des quantiles. Elle a, au minimum, environ 17 300 clients dans le groupe 7, soit 12 700 de moins que la méthode des quantiles, mais plus de trois fois plus que le nombre de clients de la méthode des amplitudes. Elle n'est donc, au niveau du nombre, pas aussi stable que la méthode des quantiles mais elle est bien meilleure que celle des amplitudes.

En examinant la courbe de prédiction moyenne obtenu avec la méthode des amplitudes, nous observons qu'elle est parfaitement linéaire, c'est partiellement dû à la méthode utilisée pour construire le score. Il s'agit d'un avantage dans la prise en compte future de cette variable dans un modèle de régression. Cette courbe a la prédiction moyenne la plus faible pour le plus petit score, de 11%, et la plus forte pour le plus grand score, de 86%. Ceci est assez logique car c'est là où elle a le moins de volume et donc elle n'a que les plus faibles prédictions en 1 et les plus fortes en 7. Elle a donc l'amplitude la plus forte des trois méthodes, d'environ 75 points, ce qui veut dire que c'est celle qui discrimine le mieux les prédictions d'appétence à

l'orientation. Cependant, comme nous l'avons vu précédemment, le faible nombre sur les extrêmes, induit du sur-apprentissage.

La méthode des quantiles a logiquement le comportement inverse, c'est-à-dire qu'elle a l'amplitude de prédiction moyenne la plus faible, d'environ 60 points. De plus c'est la seule des trois courbes à ne pas suivre une tendance très légèrement sinusoïdale et pas parfaitement linéaire. Un avantage de cette courbe est sa prédiction moyenne du groupe 4 qui est égale au taux d'orientation moyen de notre base, cela lui permettrait d'être une bonne modalité de référence, pour notre futur modèle de coût moyen.

La méthode des Kmeans est, comme pour le nombre de client, toujours entre les deux autres courbes. Elle a une amplitude de 66 points, ce qui la rend plus discriminante que la méthode des quantiles. De plus elle a une tendance linéaire et a une prédiction moyenne de 45,5% sur le groupe 4 et qui est seulement 0,1 point de plus que le taux d'orientation moyen sur la base globale. C'est donc la meilleure méthode en termes de prédiction moyenne, car elle possède les avantages des deux méthodes en évitant leurs inconvénients.

La méthode des Kmeans est donc la meilleure méthode selon l'inertie intra-groupe, la deuxième meilleure méthode sur la répartition du nombre de nos clients assez proche derrière la méthode des quantiles et la meilleure méthode sur la répartition de la prédiction moyenne. C'est donc la méthode que nous utiliserons afin de regrouper nos prédictions.

### III.2.2. Analyse du scoring final

Maintenant que nous avons choisi la méthode de scoring la plus adaptée, nous allons l'appliquer sur notre base de données. Le but de cette partie sera d'analyser les résultats de notre scoring. Pour ce faire nous allons commencer l'analyse sur le scoring global avant d'étudier plus en détail les erreurs.

| Scoring_OR | Nombre de client | Prédiction min | Prédiction max | Prédiction moyenne | Taux d'orientation réel | Écart Préd/Tx Or |
|---|---|---|---|---|---|---|
| **1.0** | 20 537 | 1.6% | 19.3% | 13.7% | 14.0% | -0.3pts |
| **2.0** | 31 775 | 19.3% | 30.0% | 24.9% | 25.2% | -0.2pts |
| **3.0** | 36 766 | 30.0% | 40.3% | 35.3% | 35.9% | -0.6pts |
| **4.0** | 38 820 | 40.3% | 50.7% | 45.5% | 45.9% | -0.4pts |
| **5.0** | 37 855 | 50.7% | 61.4% | 55.9% | 55.6% | 0.3pts |
| **6.0** | 32 020 | 61.4% | 73.5% | 66.9% | 66.6% | 0.3pts |
| **7.0** | 17 326 | 73.5% | 95.3% | 80.0% | 77.9% | 2.1pts |

*Tableau 25 - Résultats du scoring avec la méthode des Kmeans sur la base complète*

Le tableau ci-dessus présente les résultats du scoring final choisi pour regrouper nos prédictions. Le scoring 4 est celui avec la prédiction moyenne très proche du taux d'orientation moyen, de plus c'est celui qui contient le plus d'individus ce sera donc une bonne modalité de référence pour notre futur modèle de coût. Les scores aux extrêmes sont les deux groupes avec le nombre d'individus le plus faible, et pourtant leur amplitude de prédiction est plus forte que les cinq autres groupes. Pour le groupe 1 l'amplitude est d'environ 17 points, puis nous passons à 10 ou 11 points selon les groupes du groupe 2 à 5, puis nous avons 12,1 points pour le groupe 6 et 12,2 points pour le groupe 7. Cette dernière semble un peu faible au regard du peu de clients sinistrés qu'elle contient. D'ailleurs, en analysant la différence entre le taux d'orientation réel et la prédiction moyenne pour chaque score, nous observons le plus gros écart est de 2,1 points et se trouve sur

le groupe 7, sur tous les autres classes l'écart, en valeur absolue, est toujours inférieur à 0,6 points. Dans le *Tableau 24* sur la méthode des amplitudes l'écart est de 2,8 points pour ce groupe, nous avons bien un effet de sur-apprentissage moins important avec la méthode des Kmeans choisie. Cet écart entre notre prédiction et le réel est un bon résultat, mais afin de vérifier la robustesse de notre modèle, nous allons analyser nos résultats sur la seule base de test.

| Scoring_OR | Nombre de client | Prédiction min | Prédiction max | Prédiction moyenne | Taux d'orientation réel | Écart Préd/Tx Or |
|---|---|---|---|---|---|---|
| **1.0** | 4 029 | 1.6% | 19.3% | 13.8% | 17.2% | -3.4pts |
| **2.0** | 6 330 | 19.3% | 30.0% | 25.0% | 28.7% | -3.7pts |
| **3.0** | 7 417 | 30.1% | 40.3% | 35.2% | 36.5% | -1.3pts |
| **4.0** | 7 708 | 40.3% | 50.7% | 45.5% | 46.3% | -0.9pts |
| **5.0** | 7 697 | 50.7% | 61.4% | 55.9% | 53.5% | 2.4pts |
| **6.0** | 6 449 | 61.4% | 73.5% | 66.9% | 63.9% | 3.0pts |
| **7.0** | 3 389 | 73.5% | 95.0% | 79.8% | 73.3% | 6.6pts |

*Tableau 26 - Résultats du scoring avec la méthode des Kmeans sur la base de test*

Le tableau ci-dessus expose les résultats de notre scoring sur la seule base de test. L'amplitude de chacune de nos classes restent la même, nous remarquons que la plus faible prédiction se trouve dans la base de test alors que la plus forte n'est pas dedans, puisque la prédiction maximum du score 7 est de 95,0% et non de 95,3% comme sur la base totale. La répartition de nos clients reste elle aussi la même que sur notre base totale, puisque nous avons toujours le moins de client sur les classes 1 et 7 avec respectivement 9% et 8%, et le plus grand nombre sur le groupe 4 avec 18% de notre base. L'écart entre notre taux d'orientation réel et nos prédiction moyenne est lui plus important que sur la base totale. Il est, encore une fois, le plus fort sur le groupe 7 avec 6,5 points, mais il est important sur les autres groupes aussi. Nous observons que sur les extrêmes cet écart est le plus important puisque nous avons en valeur absolue au moins 3 points d'écart sur les groupes 1, 2, 6 et 7, là où les autres sont en dessous de ce seuil, le groupe 4 étant même en dessous de 1 point d'écart. Le modèle de régression utilisé amplifie la propension du client à aller vers un garage agréé. Lorsqu'un assuré a une faible probabilité d'y aller, le score prédit sera encore plus faible que cette dernière. À l'inverse lorsqu'un assuré aura une forte probabilité d'aller dans un garage agréé le score prédit sera plus haut que cette probabilité.

Afin de comprendre pourquoi nous avons cet effet de sur-apprentissage, nous allons observer sur le scoring 7, qui a le plus grand écart, les résultats sur les trois variables les plus discriminantes dans notre modèle.

| Reg_CodePOST_C | Class_TO_ANT | Gestion | Nombre de clients | Prédiction moyenne | Taux d'orientation réel | Écart Préd/Tx Or |
|---|---|---|---|---|---|---|
| 4 | 2_TO_>50% | Gestion Plateformes | 216 | 78.9% | 77.8% | 1.1pts |
| 5 | 1_pas_anterieur | Gestion Plateformes | 207 | 74.9% | 68.1% | 6.8pts |
| | 2_TO_>50% | Déléguée Agents | 367 | 78.4% | 73.3% | 5.1pts |
| | | Gestion Plateformes | 285 | 83.6% | 79.6% | 3.9pts |
| 6 | 1_pas_anterieur | Gestion Plateformes | 729 | 77.6% | 67.5% | 10.1pts |
| | 2_TO_>50% | Déléguée Agents | 385 | 82.0% | 79.5% | 2.5pts |
| | | Gestion Plateformes | 195 | 87.6% | 84.6% | 3.0pts |
| 7 | 1_pas_anterieur | Déléguée Agents | 273 | 76.4% | 64.8% | 11.6pts |
| | | Gestion Plateformes | 352 | 80.9% | 67.3% | 13.6pts |
| | 2_TO_>50% | Déléguée Agents | 119 | 86.2% | 84.9% | 1.3pts |

*Tableau 27 - Zoom sur les résultats de la classe 7*

Le tableau ci-dessus représente les résultats de la classe 7 sur la base de test. Nous avons fait un focus sur les trois variables les plus discriminantes du modèle que sont : *Le regroupement code postal*, *Le taux d'orientation sur antérieur* et *la gestion*. Nous avons ensuite sélectionné les groupes avec plus de 100 clients sinistrés afin d'avoir suffisemment d'individus dans chaque groupe. Par ailleurs, ce filtre n'implique qu'une perte minime d'information puis qu'il nous permet d'analyser nos résultats sur 92% des clients de la classe 7. Nous observons que les écarts entre le taux d'orientation réél et la prédiction moyenne ne sont pas toujours importants. Le modèle réussit, même sur la base de test et sur un zoom assez précis puisque nous sommes sur moins de 500 sinistres en moyenne, à prédire l'orientation des clients de manière satisfaisante. La moitié des groupes à un écart de moins de 4 points entre le réel taux d'orientation et la prédiction moyenne. Et sur les groupes avec un écart de plus de 6,5 points, nous remarquons qu'ils ont tous la même valeur sur la variable *Taux d'orientation sur antérieur* c'est la valeur « *Pas d'antérieur* ». Cette valeur signifie que l'assuré n'a pas eu de sinistres avant et que donc il n'a pas encore choisi d'aller ou non dans un garage agréé. Là où les autres variables de notre modèle dépendent soit des choix de vie du client (Code postal, Modèle et groupe du véhicule), soit de choix de la compagnie (Type de gestion de son sinistre), cette dernière variable pointe du doigt le précédent choix fait par un assuré en matière d'orientation. C'est donc une variable extrêmement discriminante et comme nous l'avons vu lors de l'analyse de la variable sur la *Figure 9*, elle est composée à 80% de la valeur « *Pas d'antérieur* ». Un axe d'amélioration fort pour notre modèle pourrait être de réussir à compléter cette donnée. Pour ce faire Generali pourrait mettre en place un questionnaire à la souscription pour savoir son attrait à aller dans un garage agréé. Il faudrait évidemment vérifier la cohérence de ces réponses avec la réalité lorsque les clients ont des sinistres.

Nous avons, à présent, créé et analysé les performances de notre scoring d'appétence d'un client Generali à aller dans un garage agréé, dans la suite nous l'appellerons *Scoring d'orientation*. Les résultats étant concluants, nous pouvons appliquer cette nouvelle variable créée au modèle de coût moyen qui permet de calculer la prime pure d'un contrat Automobile chez Generali.

## III.3. Application du scoring au modèle de coût automobile

Nous allons, dans cette partie, ajouter le *Scoring d'orientation* que nous venons de créer dans un modèle de coût moyen Automobile. Nous commencerons par décider sur quel modèle nous testerons cette variable et pour cela nous ferons une première analyse de son effet sur le coût moyen des sinistres.

### III.3.1. Impact du scoring sur le coût des garanties Dommages et RC

Dans la modélisation du scoring nous avons utilisé des sinistres rattachés aux deux garanties dommages et responsabilité civile. En effet, le but était de déceler quel serait le comportement de l'assuré, indépendamment des circonstances de l'accident.

Lors d'un modèle de coût nous ne pouvons pas agréger ces deux garanties pour deux raisons. Premièrement procéder ainsi biaiserait le modèle car ces deux garanties n'ont pas les mêmes coûts en moyenne, deuxièmement la garantie responsabilité civile est obligatoire mais pas celle dommage. Nous sommes donc obligés de distinguer les deux, c'est pourquoi nous allons analyser l'évolution des coûts moyens de ces deux garanties par rapport au scoring d'appétence client à l'orientation afin de choisir le meilleur modèle à décrire. Le nombre de sinistres sur chacune de ces deux garanties est assez proche puisque nous avons 110 057 sinistres *Dommages* et 105 038 sinistres *Responsabilités civiles.*

Comme pour la création du scoring, nous rentrons dans une partie de modélisation dans laquelle nous séparons notre base en deux : 80% de nos données dans la base d'entrainement et les 20 % restants dans la base de test. Nous n'utiliserons, une fois encore, la base de test qu'à la toute fin, dans le but de vérifier nos résultats et d'éviter le biais « d'espionnage de données ».

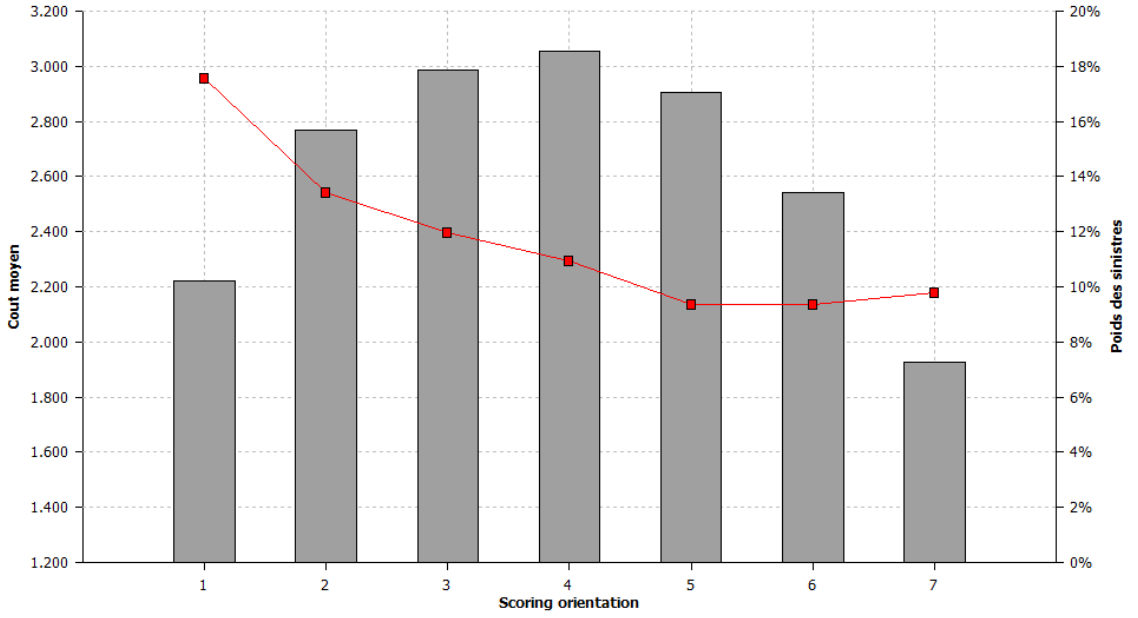



*Figure 41 - Coût moyen d'un sinistre Dommage en fonction du Scoring d'orientation sur la base d'entrainement*

La figure ci-dessus représente le coût moyen des sinistres *Dommages* en fonction du *Scoring d'orientation*. La courbe rouge désigne le coût moyen des sinistres (échelle à gauche), et les bâtons gris correspondent au poids des sinistres (échelle à droite). La répartition en nombre est la même que sur le scoring global. Elle a une forme de cloche, avec nettement moins de clients sur le score 1 et surtout sur le 7, et le score 4 est le plus représenté. Au niveau du coût moyen des sinistres, il suit une tendance nettement décroissante du score

1 à 5. Il passe de près de 3 000€ pour les clients qui ont le moins d'appétence à l'orientation, à environ 2 150€ pour ceux du groupe 5. Ce coût moyen remonte ensuite légèrement entre le groupe 5 et le groupe 7 pour se rapprocher de 2 200€. Compte tenu que ces trois derniers groupes ont un coût moyen très proche, si nous modélisons sur cette garantie il faudra sûrement les regrouper. Cet effet de *stagnation* peut être dû à une qualité de prédiction un peu moins bonne sur nos « bons » clients comme cela avait été mis en avant lors de l'analyse des résultats du modèle de prédiction. Finalement, il y a un écart de près de 800€ entre le groupe 1 et le groupe 7, soit une différence de 27%, en rapportant cet écart au coût du groupe 1. Le *Scoring d'orientation* a un effet très net sur le coût moyen d'un sinistre *Dommages*. L'analyse doit maintenant être faite et comparé les sinistres *Responsabilité Civile.*

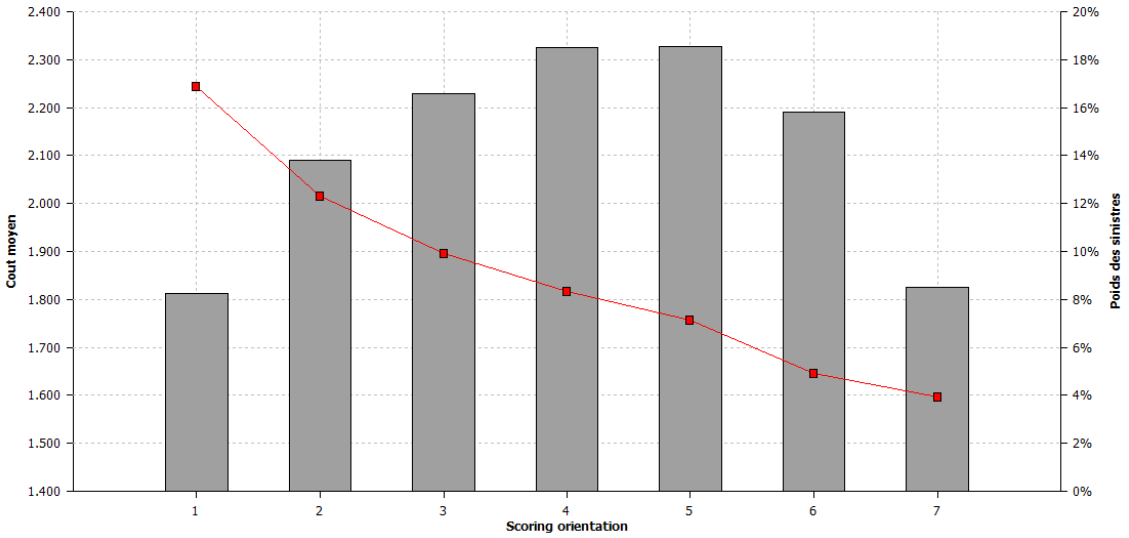



*Figure 42 - Coût moyen d'un sinistre RC en fonction du Scoring d'orientation sur la base d'entrainement*

La figure ci-dessus représente le coût moyen des sinistres *Responsabilité Civile* en fonction du *Scoring d'orientation*. Le nombre de sinistres ne suit pas exactement la même répartition que sur le scoring global. L'effet de cloche est retrouvé mais cette fois, la répartition des scores 1 et 7 est exactement la même avec environ 8% de la base, et le score 4 n'est plus le plus important car il est très légèrement plus bas que le score 5. Au niveau du coût moyen, la décroissance est parfaite du score 1 jusqu'au score 7 avec un coût qui passe de 2 250 € pour les clients les plus averses à l'orientation à 1 600 € pour les plus appétents. Il y a donc une amplitude de 650 €, qui, rapportée au coût du score 1 représente 29% du sinistre. L'effet de l'orientation est encore plus marqué sur cette garantie, c'est pourquoi la suite de l'étude sera effectuée dessus.

### III.3.2. Impact sur le modèle de coût

Maintenant que le choix de la garantie a été effectuée, il faudra appliquer la variable de score au modèle de coût. Ce modèle est basé, lui aussi, sur la régression linéaire mais utilise dans la plupart des cas une fonction de coût log et une distribution suivant une loi Gamma où log-normale (Planchet & Miseray, 2017). Le but de ce mémoire n'étant pas de créer un tarif mais de tester une nouvelle variable dans celui existant, le modèle de coût de la garantie *Responsabilité civile* pour la revalorisation des contrats 2020 sera utilisé. Ce modèle se base sur les six variables vues dans la partie *I.4.2 : L'orientation une réponse à un contexte tendu* et utilise la loi Gamma pour sa distribution, il sera implémenté sous le logiciel *Emblem*.

Afin de tester l'importance de la variable *Scoring d'orientation*, deux indicateurs de performance seront nécessaires :

- Le test de Wald : cet indicateur permet de vérifier l'importance d'une variable dans le modèle, à l'aide du test du Chi-2,
- L'AIC : Cet indicateur pénalise la log-vraisemblance du modèle, *LL*, avec le nombre de paramètres *k* du modèle :

$$AIC = 2 \times k - 2\ln(LL).$$

Dans le but de tester le gain d'AIC de cette variable et de le comparer à l'apport des autres variables, une méthode *Backward* a été appliquée au modèle. Le principe est, à l'inverse de la méthode *Forward*, de partir du modèle avec toutes les variables choisies puis de les retirer une à une en observant la différence entre le modèle avec ou sans cette variable.

| | Variables | Test de Wald | AIC | Gain d'AIC par variable | Poids de l'AIC par variable |
|---|---|---|---|---|---|
| | Modèle avec les 7 variables | - | 1 320 590,57 | - | - |
| Modèle sans : | Ancienneté_vehicule_C | 0,00% | 1 321 614,57 | 1 024 | 41,9% |
| | Classe_prix_C | 0,00% | 1 321 083,23 | 493 | 20,2% |
| | Zone_RC_C | 0,00% | 1 320 918,32 | 328 | 13,4% |
| | Scoring_Orientation | 0,00% | 1 320 846,20 | 256 | 10,5% |
| | Classe_reparation_C | 0,00% | 1 320 723,86 | 133 | 5,5% |
| | Carrosserie_C | 0,00% | 1 320 699,64 | 109 | 4,5% |
| | Age_conducteur_C | 0,00% | 1 320 689,22 | 99 | 4,0% |

*Tableau 28 - Importance des variables du modèle de coût au sens de l'AIC et du test de Wald*

Le tableau ci-dessus présente dans les colonnes de gauche à droite :

- Les variables présentes dans le modèle,
- Les résultats du test de Wald sur ces variables,
- L'AIC sur le modèle avec les sept variables, noté *AIC tot*, puis l'AIC sur le modèle sans la variable indiquée, noté *AIC var*,
- Le Gain d'AIC calculé en faisant pour chaque variable :

$$Gain\ AIC\ var = AIC\ tot - AIC\ var,$$

- Le Poids de chaque variable par rapport à l'AIC qui se calcule en faisant le rapport entre le gain d'AIC de la variable et la somme du gain d'AIC de toutes les variables :

$$Poids\ AIC\ var = \frac{Gain\ AIC\ var}{\sum_{var} Gain\ AIC\ var}.$$

Les p-value du test de Wald sont inférieures à 0,00% pour toutes les variables, elles sont donc toutes très significatives. Cet effet est normal pour les six variables utilisées habituellement dans le tarif mais c'est un premier indicateur important pour la variable *Scoring d'orientation*. Les variables sont ensuite triées de haut en bas en fonction de leur apport au Gain d'AIC. Le *Scoring d'orientation* est la quatrième variable apportant le plus d'information, avec plus de 10% d'information supplémentaire. C'est donc un premier résultat très intéressant.

Maintenant que l'importance de prendre en compte cette variable de score a été démontrée, l'impact de cette variable sur le modèle va pouvoir être analysé. La première de ces analyses se fera sur la base d'entrainement.

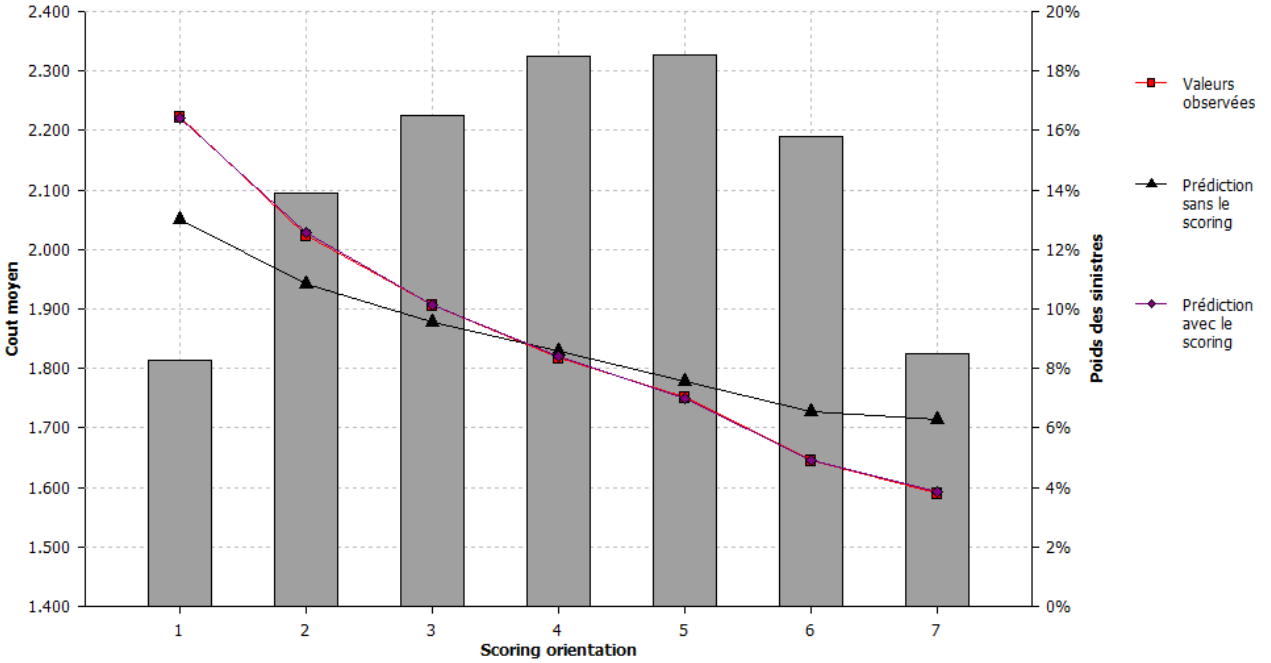



*Figure 43 -Evolution du coût moyen par Scoring d'orientation prédit et réel en fonction de la prise en compte de cette variable, sur la base d'entrainement*

Le graphique ci-dessus présente le coût moyen en fonction du *Scoring d'orientation* sur trois courbes :

- En rouge : Les valeurs observées qui sont donc les réels coûts moyens à l'intérieur de ces groupes,
- En noir : Les prédictions de coût moyen avec le modèle de coût moyen historique, donc sans la variable *Scoring orientation*,
- En violet : Les prédictions de coût moyen avec le modèle de coût moyen incluant la variable *Scoring orientation*.

Une première observation est que la courbe de prédiction avec le *Scoring d'orientation* est parfaitement alignée avec les valeurs réelles. C'est quelque chose de normal, étant donné que le modèle apprend sur cette base de données, il est donc construit justement pour donner ces résultats.

Ce qui va être intéressant à observer c'est l'écart entre les valeurs observées et la prédiction sans le *Scoring d'orientation*, car il représente l'apport de cette nouvelle variable à la segmentation du modèle. Cet écart est très faible, de l'ordre de 20€, sur les groupes centraux que sont les groupes 3,4 et 5. C'est normal, ces groupes discriminent très peu les clients, car ils contiennent les clients qui ont une chance moyenne, entre 30% et 60%, d'aller dans un garage agréé. Mais les effets de la variable sont plus significatifs au niveau des extrêmes, lorsque les clients seront appétents à l'orientation le coût moyen prédit sera surestimé alors qu'à l'inverse lorsque les clients seront peu disposés à aller dans un garage leur coût moyen sera sous-estimé. Sur le groupe 2 et 6, cet écart est d'environ 80 €, sur le groupe 7 il est de 120€ et il monte même jusqu'à 150€ pour le groupe 1.

Finalement les premiers résultats sur cette variable permettent une nouvelle différenciation tarifaire, il va falloir vérifier que cet effet persiste sur la base de test et n'est pas juste un effet de sur-apprentissage sur la base d'entrainement.

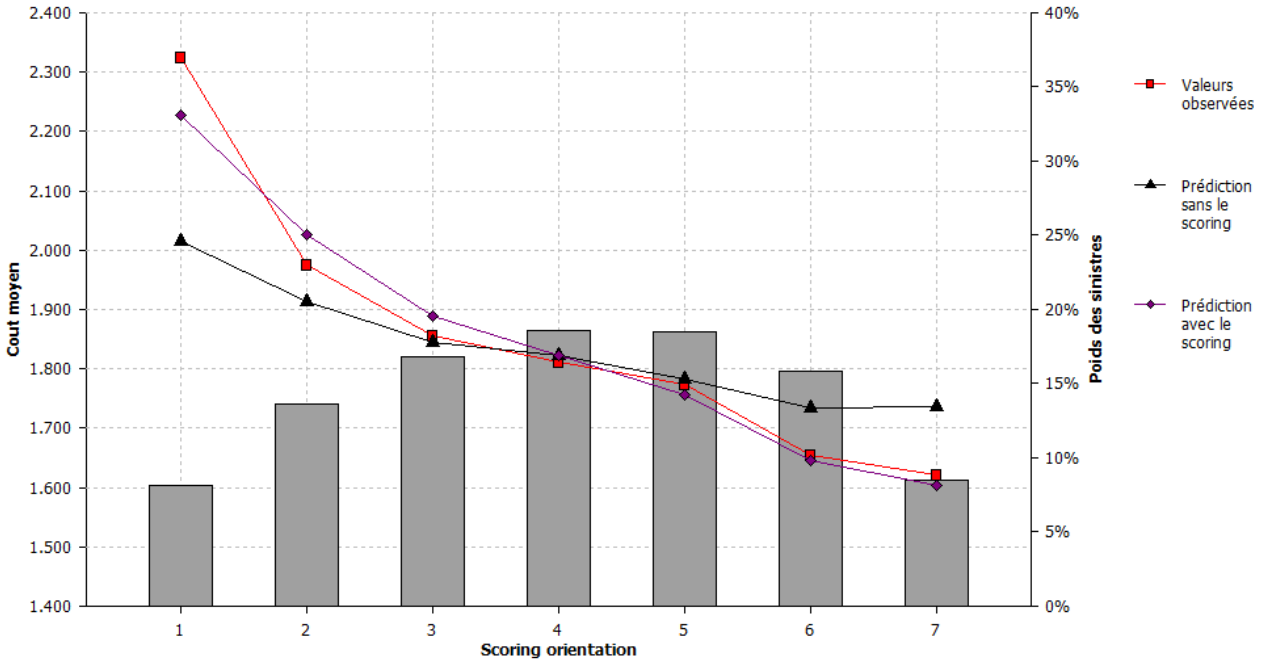



*Figure 44 - Evolution du coût moyen par Scoring d'orientation prédit et réel en fonction de la prise en compte de cette variable, sur la base de test*

Le graphique ci-dessus représente les mêmes informations que la figure précédente mais cette fois sur la base de test. Tout d'abord, il est à noter que la répartition des clients par *Scoring d'orientation* reste la même sur la base de test, cela conforte la qualité de la séparation de la base.

Sur le coût moyen, les impacts observés sur la base d'entrainement sont conservés sur la base de test. Il y a quatre cas différents qui ressortent selon les groupes :

- Un impact quasiment nul : C'est le cas des groupes 4 et 5 l'impact de la variable est très faible, les deux modèles prédisent bien le coût moyen de ces groupes,
- Un léger biais entrainé par le *Scoring d'orientation* : C'est le cas du groupe 3 sur lequel le coût moyen réel est très proche de la prédiction sans le scoring mais avec il est surestimé de 35 €.
- Un coût moyen entre les deux modèles : C'est le cas pour le groupe 2. Le coût moyen est surestimé de 52€ par le modèle comprenant le *Scoring d'orientation* mais il est sous-estimé de 61€ sans cette variable.
- Un impact fort et juste du *Scoring d'orientation* : C'est le cas des groupes 1, 6 et 7.

En détaillant ce dernier point, la prédiction avec le *Scoring d'orientation* des points 6 et 7 sous-estime le coût moyen respectivement de 9€ et 16€, soit un écart de l'ordre de 1% sur le coût moyen d'un sinistre sur ces tranches ; la prédiction sans cette variable a respectivement un écart de 79€ et 116€ soit en moyenne pour les deux plus de 5% d'écart. Sur le groupe 7, le modèle avant le scoring sous-estime de 300€, soit un écart de plus de 15%, alors qu'après la sous-estimation persiste mais est de moins de 100€.

Les résultats sur la base de test confirment donc bien que cette variable a un impact important et qu'elle sera déterminante dans la différenciation des clients.

Lors de cette dernière partie, la modélisation de l'appétence client à se diriger vers un garage agréé a été réalisée. Les trois variables les plus importantes de ce modèle étant : Le *Regroupement code postal*, le *Taux d'orientation sur antérieur*, et le mode de *Gestion* du contrat du client. Cette donnée a ensuite pu être discrétisée et transformée en un score très différenciant allant de 1 pour les clients averses à l'orientation jusqu'à 7 pour les clients particulièrement appétents. Enfin la conclusion de cette partie qui était de tester son impact sur le modèle de coût moyen Responsabilité civile a été un succès, puisque son pouvoir de différenciation sur les scores élevés a été mis en avant. L'ajout de cette variable dans le modèle de coût permettra donc, grâce à une meilleure connaissance client, d'avoir une meilleure segmentation du tarif technique lors du renouvellement des primes Automobile en 2022.

# Conclusion

Le contexte marché de l'automobile étant particulièrement concurrentiel, un des leviers mis en place par Generali afin d'optimiser ses coûts de sinistres est l'orientation de ses clients vers son réseau de garage agréé. Ce levier permet d'économiser 20% du coût d'un sinistre, ce qui est considérable, d'autant plus dans une période de forte inflation. Le but de ce mémoire était donc de prédire l'appétence d'un client à aller faire réparer son véhicule dans un garage agréé suite à un sinistre afin de segmenter plus finement le tarif technique et ainsi améliorer ses résultats sur le marché de l'automobile.

La constitution de la base de données a été une étape importante dans la perspective de disposer du plus d'informations possible sur le client. Parmi les sept bases utilisées, cinq ont permis un apport important à l'étude :

- La base sinistre pour la définition du périmètre,
- La base des rapports d'expertises pour la variable cible, le *Top_Agree,* et le *taux d'orientation sur antérieur,*
- La base des données contrats, avec le mode de *Gestion* du sinistre,
- La base de données véhicule pour le *Regroupement de modèles de véhicule* et le *Groupe SRA,*
- La base de données clients pour le *Regroupement de code postaux*.

Aucune variable issue des bases de réseau de garages et de données externes n'a été retenue dans le modèle final mais leur analyse a été intéressante et la variable de distance entre l'assuré et le garage agréé le plus proche sera utilisée par la direction Indemnisation afin de challenger le maillage géographique du réseau d'*Assercar*.

Le nettoyage des données, et notamment la discrétisation des variables à l'aide de la méthode des Kmeans, a été cruciale. En effet, deux des six variables constituant le modèle sont discriminantes grâce à ce retraitement : le *Regroupement de modèles*, qui regroupe en six groupes de véhicules différencient plus de 3 000 modalités différentes, et surtout le *Regroupement de code postaux*. Cette dernière est la plus discriminante pour le modèle, bien qu'elle induise un léger sur-apprentissage. Des travaux plus approfondis sur cette variable et sur le *Code INSEE*, en regroupant les codes ayant peu de sinistres avec ceux environnant, permettrait surement une amélioration du pouvoir prédictif de cette variable et donc in fine du modèle.

L'application du modèle d'apprentissage qu'est la régression logistique à la base de données ainsi constituée a permis d'obtenir des résultats significatifs. Tout d'abord, le seuil à utiliser pour la détermination des 1 et des 0 dans la régression a été trouvé en optimisant la précision et le rappel. Puis l'accuracy a été calculé sur ce seuil pour quatre modèles différents, et a permis de choisir le modèle qui maximisait cette métrique tout en minimisant le nombre de variables. L'accuracy pour ce modèle sur la base de test était de 63,9%, ce qui signifie que dans près de deux cas sur trois la prédiction était juste. D'autres méthodes, telles que le support vecteur machine, ou encore les forêts aléatoires, ont été testées et ne donnent pas de résultats significativement meilleurs au premier abord. Cependant, l'optimisation des paramètres, voire l'application de *boosting* sur ces méthodes, n'a pas été testée, et pourrait améliorer ces résultats.

La prédiction a ensuite été regroupée en sept classes de clients homogènes au regard de ce nouvel indicateur, grâce à la méthode des Kmeans. Ce regroupement a permis de créer un score d'appétence client à l'orientation. La probabilité prédite sur le score était en moyenne proche des réels taux d'orientation, il existait néanmoins des écarts sur les groupes extrêmes, le premier mais surtout le dernier. Celui-ci, qui

correspond aux clients les plus appétents à l'orientation, est surestimé par le modèle, c'est-à-dire qu'il prédit une probabilité plus forte pour ces clients. Cet effet pourrait être en partie résolu en affinant le travail effectué sur le code Postal et le code INSEE, mais la variable qui semble la plus significative pour une bonne prédiction sur ce groupe est le taux d'orientation sur antérieur. Or beaucoup de clients n'ont pas de sinistres antérieurs, il manque donc des informations sur leur comportement passé. Une solution pour l'approcher pourrait être de proposer, lors de la souscription du contrat, un questionnaire permettant d'évaluer l'appétence à l'orientation aux clients puis de tester ces résultats lors de réels sinistres.

Pour finir, nous avons analysé le score d'appétence sur les deux garanties *Dommage et Responsabilité civile*. Cependant, afin de pouvoir tester son apport sur le modèle de coût moyen existant, il était nécessaire de choisir l'une de ces deux garanties. Le choix a été fait de se concentrer sur la garantie *Responsabilité civile* car c'est sur celle-ci que notre nouvelle variable semble être la plus discriminante et apporter le plus d'information.

Une fois ce score ajouté aux variables du modèle de coût de la garantie *Responsabilité civile*, sa significativité a pu être démontrée. C'est la quatrième variable, parmi les sept retenues pour le modèle, apportant le plus d'information. Ce nouveau modèle a ensuite été appliqué sur la base de test et les résultats sont concluants : sur les groupes 1, 6 et 7, il réussit à fortement discriminer le coût moyen d'un assuré et même sur le groupe 2 sa prise en compte serait intéressant. Pour le groupe 1, il permet même d'éviter une sous-estimation du coût moyen de 15%. L'objectif de réussir à segmenter plus finement le tarif est donc atteint. Ce score d'appétence à l'orientation devrait permettre une meilleure différenciation des majorations lors du renouvellement du tarif des garantie *Responsabilité civile* en 2022.

Sa prise en compte dans le modèle de tarification va aussi pouvoir être étudié mais il manquera une information importante qu'est le taux d'orientation sur antérieur. Il faudra vérifier que même sans cette variable le modèle reste discriminant. L'application de ce score au modèle de *Dommages,* est aussi une prochaine étape dans l'utilisation de ces données.

Cette approche, consistant à prédire le comportement client, ici, dans le cadre d'un processus d'indemnisation, pourrait être transposée à d'autres problématiques. De nombreuses applications concrètes pourraient en effet en découler, la même méthode serait applicable :

- En Automobile : sur l'orientation d'un client vers les réparateurs partenaires bris de glaces,
- En Dommages aux biens : en prédisant l'envie d'un client de passer outre l'expertise et accepter un autre mode d'indemnisation tel que le gré à gré et ainsi d'économiser le coût de l'expertise,
- Ou encore en Assurance vie : avec l'appétence à la téléconsultation

Prédire le comportement d'un assuré à travers des modèles de Machine learning ouvre de nouveaux horizons aux assureurs afin de continuer d'adapter au mieux leur stratégie aux aspirations de leurs clients de manière rentable.

# Table des illustrations

## Tableaux



## Figures

# Références


Callet, B. (2015, Juillet). *Influence du comportement des internautes sur la sinistralité de contrats d'assurance Automobile*. Récupéré sur Ressources actuarielles: http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/8cd70228996458a5c1257f93004db08b/$FILE/CALLET.002.pdf/CALLET.pdf

Chavent, M. (2021-2022). *Apprentissage supervisé - présentation générale*. Récupéré sur Master MAS - Bordeaux: http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/cours1.pdf

FFA. (2021, Mars 24). *Fédération Française de l'Assurance*. Récupéré sur FFA: https://www.franceassureurs.fr/espace-presse/communiques-de-presse/conference-presse-resultats-2020/

Géron, A. (2017). *Machine Learning avec Scikit-learn.* Paris: Dunod.

Gorrand, R. (2020, Avril). Assurance Dommage. *Tarification à priori*. Paris, France.

Hamon, B. (2014, Mars 19). *Article L211-5-1*. Récupéré sur Code des assurances: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000028742662/

Hunault, G. (2021). *Découpage en classes et discrétisation*. Récupéré sur Université d'Angers: https://gilles-hunault.leria-info.univ-angers.fr/wstat/discr.php#:~:text=1.-,D%C3%A9finition,r%C3%A9aliser%20un%20d%C3%A9coupage%20en%20classes%22.

INSEE. (2020, 02 27). *Consommation des ménages*. Récupéré sur Insee.fr: https://www.insee.fr/fr/statistiques/4277709?sommaire=4318291

LégiFrance. (2020, Décembre 3). *LOI n° 2020-1508*. Récupéré sur Code des assurances: https://www.legifrance.gouv.fr/jorf/id/JORFSCTA000042607096

Planchet, F., & Miseray, A. (2017, Mars). *Tarification IARD : Introduction aux techniques avancées*. Récupéré sur Ressources actuarielles: http://www.ressources-actuarielles.net/C1256F13006585B2/0/457A36A8ECC541AEC1257D740067EEC4/$FILE/GLM_FP.pdf?OpenElement

Rakotamalala, R. (2016). *Méthode des centres mobiles*. Récupéré sur Université de Lyon 2: https://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf

Rakotomalala, R. (2017, Décembre 27). *Analyse de corrélation*. Récupéré sur Université de Lyon 2: https://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf

Sangnier, M. (2019, Mars 21). Unsupervised learning. *Data sciences CEA 2019*. Paris.

Schumacher, A. (2015, Avril 23). *Forward Selection with statsmodels*. Récupéré sur https://planspace.org/: https://planspace.org/20150423-forward_selection_with_statsmodels/

SRA. (2021, Décembre 31). *SRA Statistiques membres*. Récupéré sur Sécurité et Réparation Automobiles: https://www.sra.asso.fr/statistiques/Communication

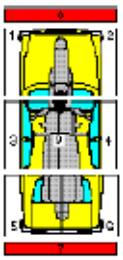
Tardy, J. (2018). *Amélioration de la qualité des données en assurance*. Récupéré sur Institut des actuaires: https://www.institutdesactuaires.com/docs/mem/ab894a7f2edac8f35a2661dd6efed84c.pdf

# Annexes

## Annexe I : Exemple de rapport d'expertise

[redacted]

Tél. :
Fax :
Email : [redacted]



Date du rapport : 07/02/2022
Numéro référence :
Nom :
Société :
Numéro de Contrat :
Référence Société : [redacted]
Référence Emetteur :
Date évènement : 14/01/2022
Date Mission : 21/01/2022

### RAPPORT D'EXPERTISE EURO

Je vous adresse le rapport que j'ai établi au titre de la mission en référence.
Restant à votre disposition, je vous prie d'agréer l'expression de mes salutations distinguées

**Destinataire :** [redacted]

**VEHICULE EXPERTISE :**

| | | | |
|---|---|---|---|
| **Immatriculation :** | [redacted] | | |
| **Marque** | : AUDI | **Genre** | : VP |
| **Modèle** | : RS3 SPORTBACK QUATTRO 2. | **Carrosserie** | : CI 5 |
| | | **Energie** | : ES |
| | | **Puissance** | : 29 CV |
| **Finition** | : . | **Couleur** | : Gris Standard |
| **Type** | : M10AUDVP026K771 | **Nombre places** | : 5 |
| **Numéro série** | : WUAZZZ8V9JA909242 | **Poids à vide** | : |
| **Mise en circul.** | : 29/06/2018 | **PTAC** | : 2010 |
| **Kilométrage** | : 89351 km | **Etat général** | : Normal |
| **Usure pneus** | : AV G :50 % AR G :10 % AV D :50 % AR D :10 % | | |

**MANDANT :**
GENERALI ASSURANCES
2 RUE PILLET WILL
PARIS

**REPARATEUR :** [redacted]

**CIRCONSTANCE DE L'EXPERTISE**
-Vu Avant travaux le 24/01/2022 personnes présentes
[redacted] Expert, [redacted] Réparateur

**PIECES COMMUNIQUEES :**
**VEHICULE ECONOMIQUEMENT REPARABLE :** Oui
**VEHICULE TECHNIQUEMENT REPARABLE :** Oui
**IMMOBILISATION THEORIQUE :** 4.0 jour(s)
**DOMMAGES IMPUTABLES :** Arrière, Moyenne, 180°
Avant, Forte, 0° (ou 360°)
**ACCORD PRIS AVEC LE REPARATEUR :**Oui **ACCORD DE REGLEMENT DIRECT DEFINITIF :**Non

### Véhicule réparable

| | | | |
|---|---|---|---|
| Montant de l'expertise | | **17851.17 TTC** | (14875.98HT) |
| Répartition des chocs : | Choc Arrière | **4694.15 TTC** | (3911.80HT) |
| | Choc Avant | **13157.02 TTC** | (10964.18HT) |
| Montant facturé par le réparateur (facture n° 339) | | **17851.18** | |

TVA récupérable : Non

Dans le cadre de l'expertise de votre véhicule, nous sommes amenés à traiter vos données personnelles (noms, prénoms, coordonnées, etc...). Ces données sont destinées au cabinet d'expertise et à ses sous-traitants (éditeurs de logiciels notamment), au propriétaire du véhicule, au réparateur, et le cas échéant, à l'assureur et au Ministère de l'Intérieur. Elles sont conservées pendant la durée strictement nécessaire à la réalisation de notre mission, puis archivées conformément aux règles de prescription légale. Vous bénéficiez d'un droit d'accès, de rectification, de limitation, de portabilité et d'effacement de vos données, et d'un droit d'opposition pour des motifs légitimes auprès de votre assureur, et lorsque la mission nous a été confiée par vous-même à l'adresse suivante : **contact30@idea-expertises.com.**
Enfin, vous avez le droit d'introduire une réclamation auprès de la CNIL (Commission nationale de l'informatique et des libertés), autorité de contrôle en charge du respect des obligations en matière de protection des données à caractère personnel.



### ANNEXE au RAPPORT D'EXPERTISE
### Numéro [redacted]

**Total Général**

| Libellé | Vétusté non déduite | | | Vétusté déduite | | | |
|---|---|---|---|---|---|---|---|
| | HT | TVA | TTC | HT | Remise | TVA | TTC |
| **Main d'oeuvre** | | | | 2337.50 | | 467.50 | 2805.00 |
| **Pièces** | 11593.48 | 2318.69 | 13912.17 | 11593.48 | | 2318.69 | 13912.17 |

# ANNEXE au RAPPORT D'EXPERTISE
## Numéro [redacted]
## Chiffrage du Choc numéro 1

### Descriptif

| | | | | | |
|---|---|---|---|---|---|
| **Point d'impact** | Arrière | **Intensité** | Moyenne | **Immobilisation théorique** | 1.5 jour(s) |
| **Angle** | 180° | **Zone de déformation** | Arrière | | |

### Sous total par choc

| Libellé | Vétusté non déduite | | | Vétusté déduite | | | |
|---|---|---|---|---|---|---|---|
| | HT | TVA | TTC | HT | Remise | TVA | TTC |
| **Main d'oeuvre** | | | | 935.00 | | 187.00 | 1122.00 |
| **Pièces** | 2626.80 | 525.35 | 3152.15 | 2626.80 | | 525.35 | 3152.15 |
| **Ingrédients peintures** | 350.00 | 70.00 | 420.00 | | | | |
| **Hors élément SGC** | | | | 3911.80 | | 782.35 | 4694.15 |
| **SOUS TOTAL GENERAL** | | | | **3911.80** | | **782.35** | **4694.15** |

### Main d'oeuvre par choc et par qualification

| Libellé | Nbr heures | P.U. | HT brut | Montant TVA | Remise |
|---|---|---|---|---|---|
| Tôlerie T1 | 4.50 | 85.00 | 382.50 | 76.50 | |
| Tôlerie T2 | 1.50 | 85.00 | 127.50 | 25.50 | |
| Peinture PEINT1 | 5.00 | 85.00 | 425.00 | 85.00 | |

(DSP : Débosselage Sans Peinture, MES : Mécanique Electricité Sellerie, PEINT : Peinture, T : Tôlerie)

### Pièces par choc

**Remplacement**

| Libellé | Abatt. Usure | Quantité | HT brut | Taux TVA | Remise | HT net | Peinture |
|---|---|---|---|---|---|---|---|
| PIECES A REMPLACER: | | 1.0 | 0.00 | 0.00 % | | | Non |
| PC AR | 0.00 % | 1.0 | 679.10 | 20.00 % | | 679.10 | Oui |
| SPOILER PC AR | 0.00 % | 1.0 | 790.16 | 20.00 % | | 790.16 | Non |
| FEU ARG VOLET AR | 0.00 % | 1.0 | 237.40 | 20.00 % | | 237.40 | Non |
| MONOGRAMME MODELE AR | 0.00 % | 1.0 | 38.67 | 20.00 % | | 38.67 | Non |
| FEU ARG | 0.00 % | 1.0 | 237.40 | 20.00 % | | 237.40 | Non |
| VOLET AR | 0.00 % | 1.0 | 550.20 | 20.00 % | | 550.20 | Oui |
| EMBLEME MARQUE AR | 0.00 % | 1.0 | 35.87 | 20.00 % | | 35.87 | Non |
| COLLE 2K LUNETTE | 0.00 % | 1.0 | 58.00 | 20.00 % | | 58.00 | Non |

**Opération**

| | |
|---|---|
| PREPARATION PEINTURE | Opération de peinture seule |

### Ingrédients peintures par choc

| Libellé | Quantité | P.U. | HT brut | Taux TVA | Remise |
|---|---|---|---|---|---|
| Métal vernis Temps H | 5.00 | 70.00 | 350.00 | 20.00 % | |

# ANNEXE au RAPPORT D'EXPERTISE
# Numéro [redacted]
## Chiffrage du Choc numéro 2

## Descriptif

| | | | | | |
|---|---|---|---|---|---|
| **Point d'impact** | Avant | **Intensité** | Forte | **Immobilisation théorique** | 2.5 jour(s) |
| **Angle** | 0° (ou 360°) | **Zone de déformation** | Avant | | |

## Sous total par choc

| Libellé | Vétusté non déduite | | | Vétusté déduite | | | |
|---|---|---|---|---|---|---|---|
| | HT | TVA | TTC | HT | Remise | TVA | TTC |
| **Main d'oeuvre** | | | | 1402.50 | | 280.50 | 1683.00 |
| **Pièces** | 8966.68 | 1793.34 | 10760.02 | 8966.68 | | 1793.34 | 10760.02 |
| **Ingrédients peintures** | 595.00 | 119.00 | 714.00 | | | | |
| **Hors élément SGC** | | | | 10964.18 | | 2192.84 | 13157.02 |
| **SOUS TOTAL GENERAL** | | | | **10964.18** | | **2192.84** | **13157.02** |

## Main d'oeuvre par choc et par qualification

| Libellé | Nbr heures | P.U. | HT brut | Montant TVA | Remise |
|---|---|---|---|---|---|
| Tôlerie T1 | 6.50 | 85.00 | 552.50 | 110.50 | |
| Tôlerie T2 | 1.00 | 85.00 | 85.00 | 17.00 | |
| Mécanique électricité sellerie MES1 | 0.50 | 85.00 | 42.50 | 8.50 | |
| Peinture PEINT1 | 8.50 | 85.00 | 722.50 | 144.50 | |

(DSP : Débosselage Sans Peinture, MES : Mécanique Electricité Sellerie, PEINT : Peinture, T : Tôlerie)

## Pièces par choc

**Remplacement**

| Libellé | Abatt. Usure | Quantité | HT brut | Taux TVA | Remise | HT net | Peinture |
|---|---|---|---|---|---|---|---|
| PLAQUE POLICE AV | 0.00 % | 1.0 | 11.50 | 20.00 % | | 11.50 | Non |
| GRILLE CALANDRE | 0.00 % | 1.0 | 1331.44 | 20.00 % | | 1331.44 | Non |
| SUPPORT PLAQUE POLIC | 0.00 % | 1.0 | 163.88 | 20.00 % | | 163.88 | Non |
| PC AV | 0.00 % | 1.0 | 1517.30 | 20.00 % | | 1517.30 | Oui |
| ENJOLIVEUR G PC AV | 0.00 % | 1.0 | 168.64 | 20.00 % | | 168.64 | Oui |
| PROJECTEUR D | 0.00 % | 1.0 | 1626.80 | 20.00 % | | 1626.80 | Non |
| PROJECTEUR G | 0.00 % | 1.0 | 1626.80 | 20.00 % | | 1626.80 | Non |
| AILE AVD | 0.00 % | 1.0 | 598.60 | 20.00 % | | 598.60 | Oui |
| CAPOT-MOTEUR | 0.00 % | 1.0 | 790.80 | 20.00 % | | 790.80 | Oui |
| ECHANGEUR AIR-AIR | 0.00 % | 1.0 | 1084.00 | 20.00 % | | 1084.00 | Non |
| GUIDE G PC AV | 0.00 % | 1.0 | 23.46 | 20.00 % | | 23.46 | Non |
| GUIDE D PC AV | 0.00 % | 1.0 | 23.46 | 20.00 % | | 23.46 | Non |

**Opération**

| | |
|---|---|
| REGLAGE: | Opération de contrôle (sans banc) |
| PROJECTEURS G, D | Opération de contrôle (sans banc) |
| REPARATION / PEINTUR | Opération de redressage (sans marbre) |
| AILE AVG | Opération de redressage (sans marbre) |

## Ingrédients peintures par choc

| Libellé | Quantité | P.U. | HT brut | Taux TVA | Remise |
|---|---|---|---|---|---|
| Métal vernis Temps H | 8.50 | 70.00 | 595.00 | 20.00 % | |

## Annexes II : Communication SRA janvier 2021

SRA

# COMMUNICATION STATISTIQUE
janvier 2021

**Évolutions des principaux éléments constituant le coût de la réparation des VP / VUL au 4ème trimestre 2020.**

### ANALYSE DE L'ENSEMBLE DES EXPERTISES DU MARCHÉ

**Source** : Base des expertises de réparation-collision (hors catastrophe naturelle, vol, incendie, BDG)

| 12 mois 2020 par rapport aux 12 mois 2019 | au cours des 12 derniers mois |
|---|---|
| *Coût total* : augmentation du coût total de la réparation | |
| → + 6,7 % | → + 6,7 % |
| *Pièces* : augmentation du coût moyen des pièces de rechange consommées | |
| → + 8,1 % | → + 8,1 % |
| *Main-d'œuvre carrosserie* : augmentation du coût horaire moyen de la main-d'œuvre totale | |
| → + 3,2 % | → + 3,2 % |
| *Peinture* : augmentation de l'équivalent horaire moyen du coût des ingrédients peinture de réparation-collision | |
| → + 4,8 % | → + 4,8 % |
| *Répartition du coût total* : | |
| ⇨ 50,9% pièces<br>⇨ 38,6% MO totale<br>⇨ 10,3% Ing. peinture | ⇨ 50,9% pièces<br>⇨ 38,6% MO totale<br>⇨ 10,3% Ing. peinture |

### OBSERVATION DES PRIX ET DES TARIFS AFFICHÉS

au cours des 12 derniers mois

**Source** : Paniers de pièces SRA constitués à partir des prix catalogues constructeurs

*Pièces* : augmentation du coût des paniers de pièces de rechange (1)

→ + 5,9 %

**Source** : Panel SRA de réparateurs

*Main-d'œuvre carrosserie* : augmentation du taux horaire moyen à partir des tarifs affichés dans les garages

→ + 3,3 %

*Peinture* : augmentation de l'équivalent horaire du prix des ingrédients peinture à partir des tarifs affichés dans les garages

→ + 4,4 %

(1) : augmentation du coût des pièces par les constructeurs en prenant en compte les variations tarifaires lors des changements de génération de véhicules et l'augmentation des coûts de pièces consécutive à la pénétration des nouveaux modèles sur le marché automobile

Pour plus d'informations, vous pouvez consulter notre site sur www.sra.asso.fr, rubrique **Statistiques** :

- Immatriculations 2019 par marque, carrosserie et pays
- Variation du prix et du coût des pare-brise en 2020

# PIÈCES DE RECHANGE

*Méthodologie :*

**SRA suit chaque trimestre, un panier de pièces de rechange établi à partir :**

➔ d'un échantillon de véhicules représentatifs du parc français (231 modèles-générations en 2020). A chaque véhicule est associée une pondération, déterminée en fonction de l'importance du parc accidenté et mise à jour chaque début d'année.

➔ d'un panier de pièces d'origine constructeur, provenant du tarif complet. La liste des pièces comprend 50 à 60 pièces de carrosserie et le radiateur. A chaque pièce est attribuée une pondération déterminée en fonction de sa fréquence de remplacement.

**A partir de ces paniers, SRA calcule :**

➔ l'évolution du prix des pièces de rechange à modèle constant

➔ l'évolution du coût des pièces de rechange en prenant en compte :

+ la pénétration des nouveaux modèles sur le marché
+ les variations de prix et du nombre de pièces lors des changements de génération

## Indice et évolution du prix et du coût des pièces de rechange hors TVA (base 100 : année 2015)

| Année | 1er trimestre (janv. à mars) | 2ème trimestre (avril à juin) | 3ème trimestre (juillet à sept.) | 4ème trimestre (oct. à déc.) | Moyenne des 4 trimestres |
|---|---|---|---|---|---|
| **Indice et évolution du prix des pièces de rechange** | | | | | |
| 2019 | 111,7 | 112,0 | 112,1 | 113,1 | 112,2 |
| 2020 | 115,4 | 117,4 | 117,8 | 118,1 | 117,1 |
| Variation 18/17 | 2,7% | 1,6% | 2,8% | 2,7% | 2,4% |
| Variation 19/18 | 6,1% | 6,3% | 5,1% | 6,8% | 6,0% |
| Variation 20/19 | 3,3% | 4,8% | 5,1% | 4,4% | 4,4% |
| **Indice et évolution du coût des pièces de rechange** | | | | | |
| 2019 | 119,2 | 119,6 | 119,7 | 121,5 | 120,0 |
| 2020 | 124,9 | 127,0 | 127,5 | 126,7 | 127,0 |
| Variation 18/17 | 4,8%® | 3,6% | 4,7% | 4,6% | 4,4% |
| Variation 19/18 | 7,9% | 8,1% | 6,9% | 6,6% | 7,6% |
| Variation 20/19 | 4,8% | 6,2% | 6,5% | 5,9% | 5,8% |

*Observations :*

Le prix des pièces de carrosserie à modèle constant augmente de **4,4 %** en 2020.

En prenant en compte, en plus, les augmentations des paniers lors des changements de modèle de véhicules, on observe une augmentation moyenne de **5,8 %** en 2020.

## Évolution du prix et du coût des pièces de rechange du panier SRA par marque

| Marque | Variation % du prix des pièces | | | | | Variation % du coût des pièces | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4è trim 2020 / 4è trim 2019 (sur 12 mois) | 2020 | | | | 4è trim 2020 / 4è trim 2019 (sur 12 mois) | 2020 | | | |
| | | 1er trim. | 2è trim. | 3è trim. | 4è trim. | | 1er trim. | 2è trim. | 3è trim. | 4è trim. |
| **ALFA ROMEO** | **8,07** | 1,55 | 4,45 | - | - | **8,81** | 2,07 | 4,45 | - | - |
| **AUDI** | **8,87** | 2,82 | - | 3,74 | - | **8,08** | 2,28 | - | 3,74 | - |
| **BMW** | **2,29** | 1,78 | - | 0,50 | - | **2,86** | 2,44 | - | 0,50 | - |
| **CITROEN / DS** | **7,18** | 1,82 | 2,77 | - | 2,61 | **8,77** | 3,15 | 2,77 | - | 2,61 |
| **DACIA** | **2,43** | -0,35 | 2,79 | - | - | **2,46** | -0,33 | 2,79 | - | - |
| **FIAT** | **4,28** | 0,68 | 3,56 | - | - | **6,83** | 3,16 | 3,56 | - | - |
| **FORD** | **2,80** | 0,19 | 2,47 | – | -0,06 | **3,10** | 0,68 | 2,47 | – | -0,06 |
| **HONDA** | **2,37** | – | - | - | 2,37 | **3,33** | 0,94 | - | - | 2,37 |
| **HYUNDAI** | **7,00** | 1,22 | 5,71 | - | - | **8,64** | 2,77 | 5,71 | - | - |
| **KIA** | **4,65** | 2,90 | 1,70 | - | -- | **4,42** | 2,67 | 1,70 | - | -- |
| **MAZDA** | **5,91** | 5,95 | - | -0,04 | - | **6,17** | 6,21 | - | -0,04 | - |
| **MERCEDES** | **1,38** | 1,38 | - | - | - | **2,29** | 2,29 | - | - | - |
| **MINI** | **1,90** | 1,61 | - | 0,29 | - | **2,85** | 2,35 | - | 0,29 | - |
| **NISSAN** | **0,95** | 0,95 | - | - | - | **3,82** | 3,82 | - | - | - |
| **OPEL** | **6,50** | – | 1,84 | - | 4,58 | **4,68** | -1,80 | 1,84 | - | 4,58 |
| **PEUGEOT** | **7,31** | 1,75 | 2,79 | - | 2,60 | **9,29** | 3,83 | 2,79 | - | 2,60 |
| **RENAULT** | **1,58** | -0,13 | 1,72 | - | - | **4,28** | 2,52 | 1,72 | - | - |
| **SEAT** | **8,38** | 4,37 | - | 1,91 | - | **7,89** | 5,67 | - | 1,91 | - |
| **SKODA** | **5,03** | 2,93 | - | 2,04 | - | **5,88** | 3,74 | - | 2,04 | - |
| **SUZUKI** | **1,01** | 1,01 | - | - | - | **0,78** | 0,76 | - | - | - |
| **TOYOTA** | **2,87** | 1,53 | - | 1,32 | - | **4,17** | 2,81 | - | 1,32 | - |
| **VOLKSWAGEN** | **5,71** | 3,05 | - | 2,58 | - | **8,26** | 3,58 | - | 2,58 | - |
| **VOLVO** | **- 1,04** | -1,04 | - | - | - | **0,18** | 0,18 | - | - | - |
| **Toutes marques** | **4,44** | **1,27** | **1,87** | **0,48** | **0,88** | **6,86** | **2,86** | **1,67** | **0,48** | **0,88** |

*Commentaire :*

A noter ce trimestre, une 3ème augmentation cumulée des prix de pièces chez **PSA** entraine une hausse globale de leur coût : **+ 8,77 %** chez **CITROEN / DS** et **+ 9,29 %** chez **PEUGEOT** au cours de l'année 2020.

## Annexes III : Code de la méthode forward sous Python

```
#Chargement des packages
import statsmodels.api as sm
import pandas as pd
from sklearn.metrics import accuracy_score
#Création de la fonction sous Python
def forward_regression_pval(X,y):
#Initialisation de nos paramètres
  initial_list = []
  threshold_in = 0.01
  verbose = True
  included = list(initial_list)
#Lancement de la boucle pour le forward
  while True:
    changed = False
    # Étape forward
    excluded = list(set(X.columns) - set(included))
    new_pval = pd.Series(index = excluded,dtype = np.float64)
    for new_column in excluded:
      #Initialisation du modèle
      model = sm.Logit(y,sm.add_constant(pd.DataFrame(X[included + [new_column]]))).fit()
      new_pval[new_column] = model.pvalues[new_column]
    best_pval = new_pval.min()
    #Test si la nouvelle p - value est plus performante que la meilleure retenue jusqu'ici
    if best_pval < threshold_in:
      best_feature = new_pval.argmin()
      included.append(excluded[best_feature])
      changed = True
      #Affichage de l'accuracy et de la p - value pour le meilleur modèle
      if verbose:
        model = sm.Logit(y,sm.add_constant(pd.DataFrame(X[included]))).fit()
        print('Train accuracy
              = ',accuracy_score(y,list(map(round,model.predict(sm.add_constant(X[included]))))))
        print('Add ' + excluded[best_feature] + ' with p - value ' + str(model.pvalues[len(included) - 1]))
    if not changed:
      break
  return included
#Lancer la base d'entrainement
forward_regression_pval(X_train,Y_train)
```