Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/11/2022

Par : **Islem Garaouch**

Titre : **Création de nouvelles variables explicatives de la sinistralité grave en assurance Auto**

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

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

Membres présents du jury de la filière

Nom : BARADEL Nicolas

Membres présents du jury de l'Institut des Actuaires

ZEC Nicolas, FLICHY Etienne ,
NESSI Jean-Marie

Entreprise : Generali

Signature :

Directeur du mémoire en entreprise :

Nom : SAMBA Gaud Cyrius Eyrich
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Remerciements

Je tiens à exprimer toute ma reconnaissance à Monsieur SAMBA Gaud Cyrius Eyrich, mon tuteur de stage à l'entreprise, pour m'avoir encadrée, orientée, aidée et conseillée tout au long de mon stage.

J'adresse également ma gratitude à tous les membres de l'équipe techniques actuarielles de Generali, pour la qualité de leur travail et leur bonne humeur à toute épreuve.

Mes sincères remerciements vont aussi à Monsieur BARADEL Nicolas, mon tuteur académique pour le suivi de ce mémoire et pour ses conseils précieux.

Je remercie également les professeurs et intervenants professionnels à l'ENSAE. Je les remercie pour l'enseignement qu'ils m'ont apporté.

Résumé

Mots clés : Sinistre grave, modèle linéaire généralisé (GLM), modèle XGBoost

Résumé :
Le but de ce mémoire est la construction de nouvelles variables qui pourraient, en plus des variables tarifaires classiques modéliser la propension d'un assuré à avoir un sinistre grave. On s'intéresse aux sinistres corporels liés au produit « L'Auto Generali ». Le seuil de sinistre graves retenu est 50 000 euros, défini en utilisant la méthode statistique de QQ-plot.
Tout d'abord, on commence par créer de nouvelles variables explicatives décrivant les avenants antérieurs des assurés. La base de modélisation créée gardera ainsi la structure initiale de la base tarifaire à laquelle on ajoute quatre variables principales pour décrire le comportement antérieur de l'assuré. Premièrement, le changement de la zone qui permet de détecter un changement du niveau de risque lié à un changement de la zone d'habitation dans le cas d'un déménagement. Deuxièmement, le changement de la date de naissance du conducteur principal qui permet de détecter la différence d'âge entre l'ancien conducteur principal et le nouveau conducteur principal. Troisièmement, le changement de la date de permis qui permet de détecter la différence entre la date d'obtention du permis de l'ancien conducteur principal et celle du nouveau conducteur principal. Quatrièmement, le changement de groupe qui permet de détecter le changement du groupe de véhicule dans le cas d'un avenant de véhicule.
L'étape suivante consiste à modéliser la probabilité d'avoir un sinistre grave en fonction des différentes variables explicatives construites. Deux approches sont adoptées. On commence par un modèle GLM couramment utilisé dans les processus de tarification. La première étape consiste à modéliser la variable cible en fonction des variables tarifaires. On sélectionne ensuite le modèle le plus pertinent parmi ceux proposés par l'outil utilisé. La deuxième étape de la modélisation consiste à enrichir le modèle avec les nouvelles variables construites pour étudier leurs impacts. Seulement deux parmi les quatre variables précédemment décrites ont été retenues : le changement de zone et le changement de date de naissance du conducteur principal. On s'aperçoit également que ces variables permettent d'améliorer le pouvoir explicatif du modèle.
Afin de tenter d'améliorer les performances du modèle retenu, un autre algorithme très utilisé pour les problématiques de classification et assez reconnu pour ses performances a été implémenté : l'algorithme XGBoost. On s'aperçoit qu'on obtient des résultats meilleurs avec cette approche. Les variables retenues lors de l'étape de sélection des variables sont les mêmes que celles proposées par le modèle GLM. Cependant, l'ordre d'importance des variables diffère entre les deux cas.
La dernière étape de cette étude consiste à mettre en application ces résultats obtenus : On définit une nouvelle règle de mise sous surveillance basée sur les nouvelles variables construites et on étudie son impact. On s'aperçoit qu'on arrive à améliorer les résultats obtenus uniquement avec le processus actuel de surveillance.

Abstract

Key words : Severe claim, generalized linear model (GLM), XGBoost model

Abstract :
The aim of this study is the construction of new variables which could, in addition to the classic tariff variables, describe the propensity of an insured to have a severe claim. We are interested in claims related to the product "L'Auto Generali". The threshold of severe claims used is 50 000 euros, defined using the QQ-plot statistical method.
First, new explanatory variables describing insurers' past amendments were created. The modelling base created will thus keep the initial structure of the tariff base to which four main variables have been added to describe the past behaviour of the insured. Firstly, the change in the zone that allows the detection of a change in the level of the risk due to a change in the area of residence. Secondly, the change in the date of birth of the main driver, that allows the detection of the age difference between the old and the new main driver. Thirdly, the change in the date of the driving licence obtaining that informs about the difference between driving licence obtaining date of the old main driver and the new main driver. Fourthly, the change of group that detects the change of vehicle group in the case of a vehicle amendment.
The next step is to model the probability of having a severe claim as a function of the different explanatory variables constructed. Two approaches were tested. A GLM model, commonly used in underwriting processes, was tested first. The first step is to model our target variable as a function of the tariff variables. Then, the most relevant model among those proposed by the tool used is selected. The second step consists in enriching our model with the new variables constructed to study their impacts. Only two of the four variables previously identified were retained : The change in zone and the change in the date of birth of the main driver. In order to try to improve the performance of the selected model, another algorithm widely used for classification problems and well known for its performance was implemented : the XGBoost algorithm. It can be noticed that better results are obtained with this approach. The variables retained during the variable selection process are the same as those proposed by the GLM model. However, the order of importance of the variables differs between the two models.
The last step of this study consists in applying concretely these results : We define a new supervision rule based on the new variables and study its impact. The results obtained with the current monitoring process alone are then improved.

Sommaire

1 Introduction

L'assurance automobile est un secteur pilier de l'assurance en France. Il ne cesse de se développer incitant ainsi de plus en plus la concurrence entre les différents acteurs. Il y a même de plus en plus de nouveaux entrants alors que le marché est déjà saturé, notamment par la présence des bancassureurs. Ce contexte concurrentiel rend de plus en plus important la bonne maîtrise du portefeuille et notamment des sinistres graves, sources d'importantes pertes financières pour les assureurs.

Les sinistres graves sont considérés comme des évènements rares observés au sein du portefeuille lors d'une augmentation inhabituelle des coûts des sinistres. Contrairement aux sinistres ordinaires, les sinistres graves sont caractérisés par un coût assez élevé et une fréquence très faible. Malgré leur rareté, ils impactent fortement le processus de tarification. Leur prédiction est à fort enjeu en assurance IARD (incendies, accidents et risques divers).

L'objectif de ce mémoire est de construire de nouvelles variables explicatives pertinentes pour prédire l'occurence d'un sinistre grave. Ces variables seront principalement déduites des avenants effectués par l'assuré. Il s'agit donc de modéliser la probabilité pour un assuré donné d'avoir un sinistre grave en fonction d'un panel de variables explicatives comprenant à la fois des variables tarifaires classiquement utilisées pour la modélisation de la fréquence ainsi que les nouvelles variables construites pour cette étude déduites des avenants antérieurs.

Deux approches différentes seront explorées dans ce mémoire : le modèle GLM et le modèle XGBoost. Plus précisément, on s'intéresse dans cette étude à la garantie RCC (Responsabilité civile pour les sinistres corporels) pour les contrats de Véhicules 4 roues « pur » : Hors camping-car, remorque et caravane, et hors usages spéciaux pour le produit : « L'Auto Generali ». Le choix de la garantie RCC en particulier découle du fait qu'on s'intéresse aux sinistres graves qui sont majoritairement des sinistres corporels. L'objectif étant de construire de nouvelles variables pouvant potentiellement prédire l'occurence d'un sinistre grave et d'étudier leurs pertinences.

La première partie du mémoire sera consacrée à une présentation générale du contexte de

l'étude et du produit qui nous intéresse pour cette étude : « L'Auto Generali ». On se focalisera ensuite sur la méthodologie de construction de la base de modélisation. Les bases utilisées ainsi que les différentes variables explicatives construites seront présentées. La troisième partie explorera les différents résultats obtenus par les modèles cités précédemment. La dernière partie aura pour objectif de présenter un cas d'application pratique des différents résultats obtenus pour le processus de surveillance.

2

Contexte de l'étude

Le marché de l'assurance automobile est un marché pilier de l'assurance IARD (Incendie, Accidents et Risques Divers) en France. L'assurance automobile concerne tout véhicule terrestre à moteur assuré en France et destiné à circuler en France ou au sein de l'espace européen. Rendue obligatoire le 27 Février 1958, cette dernière est régie comme toute autre assurance par le code des assurances. Elle permet la protection contre les dégâts qui peuvent être engendrés par les accidents de route suivant le type de garantie choisie.

On présentera, tout d'abord, quelques chiffres témoignant de la place importante du marché de l'assurance automobile en France. On détaillera, par la suite, les différentes caractéristiques du produit « L'Auto Generali » ainsi que les différentes garanties proposées.

2.1 L'assurance automobile en France

L'assurance IARD est l'abréviation de l'acronyme « incendies, accidents et risques divers ». Elle permet de couvrir les biens des assurés dans le cas d'incendie, catastrophes naturelles, accidents, etc. Ce secteur occupe une place importante dans le marché de l'assurance en France. Sa part a atteint, d'après le graphique ci-dessous (2.1), 29,8% en 2020. Ce chiffre ne cesse d'évoluer. La plus grande part en assurance IARD est détenue par l'assurance Automobile. Ce produit occupait 11.6% du marché de l'assurance IARD en 2020 (2.1).

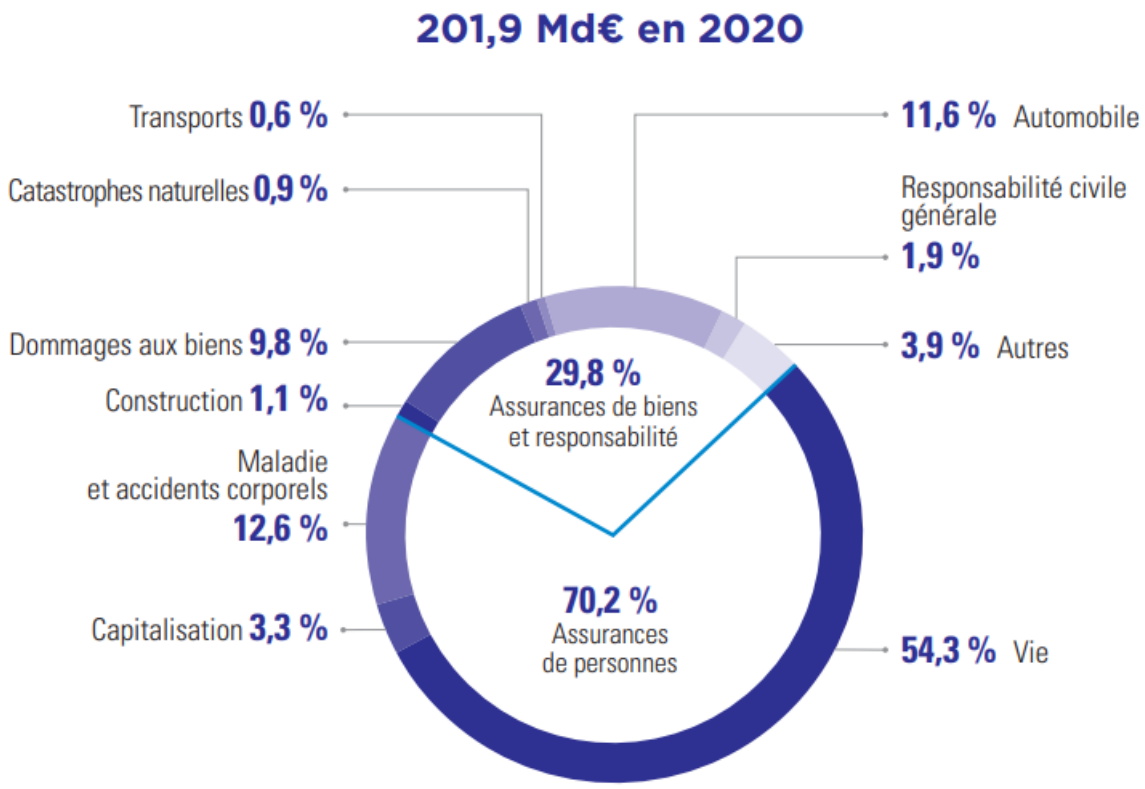


FIGURE 2.1 – Répartition des cotisations pour le marché de l'assurance vie et non vie en France en 2020

	2020 (En Mds €)	2021(p) (En Mds €)	Variation 2021/2020
Assurance automobile	23,5	24,1	+ 2,5 %
Dont particuliers (*)	21,1	21,6	+ 2,2 %
Dont professionnels	2,4	2,5	+ 4,7 %
Autres marchés	36,7	39,1	+ 6,4 %
Total dommages aux biens et responsabilité	**60,2**	**63,2**	**+ 4,9 %**

(p) provisoire *(*) Tous types de véhicules hors flottes d'entreprises*

FIGURE 2.2 – Poids de l'assurance automobile dans les assurances de biens et de responsabilité

Rendue obligatoire en 1958, quelques années après la création des fonds de garantie automobile, l'assurance automobile représente aujourd'hui une part importante du marché de l'assurance IARD. Sa part des cotisations a atteint 24,1 Million d'euros en 2021 avec une

évolution de l'ordre de 2,5% par rapport à l'année 2020 (2.2) donnant ainsi lieu à une intense concurrence entre les différentes compagnies. La figure ci-dessous (2.3) montre le classement suivant la part de marché de l'assurance automobile des dix principaux groupes d'assurance en 2021.

Groupes	Rang 2021	Cotisations (En M€) 2021 (p)	Répartition	Répartition en % cumulé
Covea	1	4 440	18,4 %	18,4 %
Axa	2	3 191	13,3 %	31,7 %
Aema	3	2 710	11,3 %	42,9 %
Groupama-Gan	4	2 358	9,8 %	52,7 %
Allianz	5	1 928	8,0 %	60,7 %
Maif	6	1 578	6,6 %	67,3 %
Crédit Agricole	7	1 499	6,2 %	73,5 %
Generali	8	1 353	5,6 %	79,1 %
Crédit Mutuel	9	1 039	4,3 %	83,4 %
Matmut	10	1 002	4,1 %	87,6 %
Total	///	**21 097**	**87,6 %**	///

(p) provisoire

FIGURE 2.3 – Part de marché des dix principaux groupes d'assurance en 2021

Les dix premiers groupes comptabilisent 87,6 % du marché national de l'assurance automobile. Generali en fait partie et détient 4,3% de l'ensemble des cotisations pour son produit « l'Auto Generali » que l'on présentera dans la partie suivante.

2.2 Présentation du produit « l'Auto Generali »

On s'intéresse dans cette étude au produit d'assurance « L'Auto Generali » destiné aux particuliers pour assurer les véhicules 4 roues « pur » : Hors camping-car, remorque et caravane, et hors usages spéciaux. Generali propose différentes formules qui diffèrent suivant les garanties comprises dans la formule.

Le tableau ci-dessous résume les trois formules proposées par Generali :

	L1	L2	L3
Responsabilité civile	✓	✓	✓
Défense Pénale et Recours Suite à Accident	✓	✓	✓
Événements majeurs (catastrophes naturelles, événements climatiques, attentats, actes de terrorisme et de sabotage, émeutes et mouvements populaires, catastrophes technologiques)	-	✓	✓
Bris de glaces	-	✓	✓
Vol et Tentative de Vol	-	✓	✓
Incendie	-	✓	✓
Dommages Tous Accidents/ vandalisme	-	-	✓

✓ : *Inclus d'office* - : *Option*

Deux notions sont importantes pour bien comprendre le tableau ci-dessus. D'une part, la mention *inclus d'office* signifie que la formule couvre obligatoirement la garantie en question. D'autre part, la mention *Option* indique que la garantie est optionnelle dans cette formule. L'assuré peut donc choisir de l'inclure ou pas.

La formule L1 s'avère ainsi la moins inclusive en termes de garanties. Elle permet uniquement de satisfaire l'obligation légale d'assurance des véhicules terrestres à moteur. Elle assure la prise en charge des préjudices causés. Au contraire, la formule L3 est la plus complète permettant de couvrir un champ plus large de dommages.

2.2.1 Description des garanties

Afin de mieux comprendre la différence entre ces différentes formules, on explicitera ci-dessous les différentes garanties précédemment évoquées.

Responsabilité Civile Automobile

Cette garantie est la garantie minimale qu'un contrat d'assurance *Auto* doit obligatoirement inclure. Elle permet la prise en charge des préjudices causés aux Tiers. Ces préjudices peuvent être des dégâts matériels, blessures, maladie, etc.

Défense Pénale et Recours Suite à Accident DPRSA

Elle permet d'assurer la défense juridique de l'assuré s'il est victime de dommages ou bien s'il est mis en cause dans le cadre de poursuites pénales. Elle permet la défense juridique de l'assuré dans le cas de litige afin de garantir ses droits. Elle est incluse dans toutes les formules proposées par Generali.

Événements majeurs

Cette garantie assure contre les évènements majeurs imprévisibles qui ne peuvent pas être contrôlés par l'assuré. Parmi les évènements majeurs, on peut citer : les catastrophes naturelles, les événements climatiques, les attentats, les actes de terrorisme et de sabotage, les émeutes et mouvements populaires ainsi que les catastrophes technologiques

Vol

Cette garantie permet de couvrir l'assuré contre tout vol ou tentative de vol du véhicule en entier ainsi que tout vol ou tentative de vol des éléments intérieurs ou extérieurs du véhicule (par exemple : serrures endommagées).

Incendie

Cette garantie couvre l'assureur suite à la survenance des évènements cités ci-dessous

- Incendie
- Explosion
- Combustion spontanée
- Chute de la foudre

Bris de glaces

Cette garantie permet, selon une limite contractuellement définie, le remboursement du coût de remplacement (pièces et main d'œuvre) ou de la réparation suite au bris de l'un des éléments suivants :

- Pare brise
- Glaces latérales
- Lunette arrière
- Toits vitrés, toits ouvrants
- Optiques de phares avant, clignotants avant et antibrouillards avant

Dommages Tous Accidents

Cette garantie assure la réparation de tous les dommages matériels subis par l'assuré sans tenir compte de sa responsabilité.

3

Éléments théoriques liés à la modélisation

Ce chapitre sera consacré à la présentation des éléments théoriques de la modélisation. L'objectif est de modéliser la probabilité qu'un assuré sinistré ait un sinistre grave et d'expliquer les facteurs déterminants de cet évènement. En d'autres termes, le but de la modélisation est de chercher les variables les plus pertinentes qui pourraient expliquer une sinistralité grave.
On s'intéresse à la modélisation des sinistres graves pour les contrats du produit « L'Auto Generali » destiné aux particuliers pour assurer les véhicules 4 roues « pur » : Hors camping-car, remorque et caravane, et hors usages spéciaux. La première partie abordera la méthode sélectionnée pour choisir le seuil à partir duquel un sinistre sera classé comme un sinistre grave.
On détaillera ensuite la théorie des modèles linéaires simples et généralisés utilisés pour la modélisation de la fréquence, l'outil utilisé ainsi que les éléments théoriques liés à l'algorithme XGBoost.

3.1 Éléments théoriques liés à la méthode de choix du seuil

3.1.1 Rappel sur la théorie des valeurs extrêmes

La théorie des valeurs extrêmes permet d'étudier la queue de la loi de probabilité d'une distribution donnée. Plus précisément :
On considère :

- $(X_1, ..., X_n)$ n variables aléatoires
- F(x)=P($X_i \leq x$) la fonction de répartition correspondante à la variable aléatoire X_i

- M_n =max$(X_1, ..., X_n)$

On s'intéresse à l'étude du comportement de la variable aléatoire M_n.

Avant de détailler les différents cas de figures que l'on peut avoir, rappelons la définition d'une *GEV (Generalized Extreme Value)* :

$X \sim GEV(\mu, \sigma, \kappa)$ avec $\sigma > 0$ si sa fonction de répartition s'écrit suivant la valeur de κ

- $\kappa \neq 0 : F(x, \mu, \sigma, \kappa) = exp(-[1 + \kappa\frac{(x-\mu)}{\sigma}]^{\frac{-1}{\kappa}})$ définie pour x tel que $[1 + \kappa\frac{(x-\mu)}{\sigma}] > 0$
- $\kappa = 0 : F(x, \mu, \sigma, \kappa) = exp(-exp([-\frac{x-\mu}{\sigma}]))$ définie pour x dans $\mathbb{R}$

On définit ainsi différentes lois suivant la valeur de κ :

- $\kappa = 0$ correspond à la loi Gumbel
- $\kappa < 0$ correspond à la loi Weibull
- $\kappa > 0$ correspond à la loi Fréchet

Théorème de Fisher-Tippet

Pour la suite de variables aléatoires précédemment définies, s'ils existent deux suites a_n et b_n telles que $P(\frac{(M_n-b_n)}{a_n} \leq z) \to G(z)$ et G non dégénérée alors on dit que G est la fonction de répartition d'une loi GEV.

Le paramètre κ permet de décrire l'épaisseur de la queue de la distribution étudiée et donc déterminer le poids des valeurs extrêmes.
en effet, on peut distinguer trois cas différents :

- Loi Weibull : queue fine
- Loi Gumbel : queue fine ou moyenne
- Loi Fréchet : queue épaisse

Aprés avoir introduit la notion de queue fine, épaisse et moyenne, on s'intéresse dans la suite à la méthode de QQ-plot qui sera utilisée pour déterminer le seuil des sinistres graves.

3.1.2 Méthode de QQ-plot

Un QQ-plot est un graphique qui permet de comparer deux distributions : une distribution empirique et une distribution théorique. Plus précisément, on trace les quantiles de la distribution empirique en fonction de ceux de la distribution théorique. Ceci permet de déduire si les valeurs de l'échantillon étudié suivent une loi théorique. Dans ce cas, le graphique sera une droite linéaire.
Lorsqu'on cherche à déterminer des valeurs extrêmes, le graphique qui est utilisé est celui de la distribution de la loi exponentielle. Cette comparaison permet de déterminer si oui ou non l'échantillon étudié comprend des valeurs extrêmes. En effet, trois graphes sont possibles :

- Une droite linéaire : Ce cas correspond à une queue trés légère. La distribution empirique est confondue avec celle de la loi exponentielle.
- Une courbe concave : Ce cas indique une forte présence de valeurs extrêmes dans l'échantillon étudié. On parle ainsi d'une queue épaisse.
- Une courbe convexe : Ce cas indique une faible présence de valeurs extrêmes dans l'échantillon étudié. On parle ainsi d'une queue légère.

Dans le cas de notre étude des sinistres graves, on s'attend ainsi à ce que le graphique de QQ-plot soit concave.

3.2 Modèle linéaire généralisé

3.2.1 Rappel sur les modèles de régression simples

L'objectif du modèle linéaire simple est d'expliquer la variable d'intérêt Y à partir d'un panel de variables explicatives $X = (X^1, ..., X^k)$

l'équation du modèle linéaire simple s'écrit comme suit :

$$Y = X\beta + \varepsilon = \beta_0 + \sum_{j=1}^{k} \beta_j X^j + \varepsilon$$

Avec

- Y la varible aléatoire endogène qu'on cherche à expliquer
- $(\beta_0, \beta_1, ..., \beta_k)$ les coefficients du modèle
- $(X^1, ..., X^k)$ les variables exogènes explicatives
- ε le terme d'erreur du modèle vérifiant : $Var(\varepsilon) = \sigma^2 > 0$ et $E(\varepsilon) = 0$ où σ^2 est un paramètre inconnu à estimer

Les hypothèses du modèle : Normalité et homoscédasticité

On suppose ε suit une loi normale centrée de variance σ^2 (homoscédasticité).
Par conséquent, la variable à expliquer Y suit également la loi normale :

$$Y \sim N(X\beta, \sigma 2)$$

Cette hypothèse de normalité est très restrictive et rend l'utilisation de ce modèle assez limitée, d'où l'intérêt des modèles linéaires généralisés.

3.2.2 Présentation des modèles linéaires généralisés

A l'instar des modèles linéaires classiques, les modèles linéaires généralisés cherchent à prédire la relation entre la variable à expliquer Y et l'ensemble des variables explicatives.

Ces modèles sont souvent utilisés dans le processus de tarification en assurance IARD. Ils permettent d'alléger les hypothèses retenues par les modèles de régressions classiques : normalité, homoscédasticité, effets additifs.

Par ailleurs, on ne suppose plus que la relation entre la variable à expliquer et l'ensemble des variables explicatives est affine. Elle est plutôt représentée par une fonction de lien monotone et différentiable. On parle ici d'une relation fonctionnelle entre la combinaison linéaire des variables explicatives et l'espérance conditionnelle de la variable conditionnelle sachant les valeurs prises par ces premiers prédicteurs.

Avant de présenter l'expression mathématiques du modèle ainsi que ses hypothèses et la méthode d'estimation des paramètres, on commence par introduire une notion importante pour la suite : la famille exponentielle.

La famille exponentielle

Soit la fonction f définie par :

$$f(y, \theta, \varphi) = \exp(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)) \ (1)$$

Avec :
- θ le paramètre naturel, inconnu
- ϕ le paramètre de dispersion supposé connu
- a(.) et c(.) sont des fonctions dérivables
- b(.) est de classe C^3 et de dérivée première inversible

La famille des exponentielles est l'ensemble des lois dont la densité f peut s'écrire sous le format présenté ci-dessus. Soit Y une variable aléatoire appartenant à la famille des exponentielles. Elle vérifie :

- $E(Y) = b'(\theta) = \mu$
- $Var(Y) = b''(\theta)$ x $a(\phi)$

Avec $b'(\theta)$ et $b''(\theta)$ les dérivées premières et secondes de b par rapport à θ.

Comme b' est supposée inversible, on peut écrire : $\theta = (b')^{-1}(\mu)$

Un modèle exponentiel admet donc ces deux propriétés principales :

- La loi de distribution est caractérisée par la moyenne et la variance.
- La variance est une fonction de la moyenne.

Après avoir explicité les propriétés de la famille exponentielle, on donnera dans la suite quelques exemples de lois usuelles qui appartiennent à cette famille de loi.

La loi Poisson

Soit Y une variable aléatoire qui suit une loi de poisson de paramètres $\mu = E(Y)$

La fonction densité de Y s'écrit pout tout y=0,1,2,... :

$$f(y) = \frac{\mu^y}{y!} e^{-\mu}$$

Cette fonction peut se réécrire sous le même format que celui de l'équation (1) avec :

$$\theta = log(\mu)$$
$$a(\phi) = 1$$
$$b(\theta) = e^{\theta}$$
$$c(y, \phi) = -ln(y!)$$

Loi Normale

Soit Y une variable aléatoire qui suit une loi normale d'espérance μ et de variabe σ^2. La fonction densité de Y s'écrit :

$$f(y) = \frac{1}{\sqrt{2\Pi\sigma^2}} exp(\frac{-1}{2} \frac{(y-\mu)^2}{\sigma^2})$$

$f(y)$ se réécrit sous le format suivant :

$$f(y) = exp(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} log(2\pi\sigma^2))$$

On peut donc constater que la loi normale appartient à la famille exponentielle avec :

$$\theta = \mu$$
$$\phi = \sigma^2 \; a(\phi) = \phi$$
$$b(\theta) = \frac{1}{2}\theta^2$$
$$c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2} log(2\pi\sigma^2)$$

Loi binomiale

Soit Y une variable aléatoire qui suit une loi binomiale de paramètres n et p.

Sa fonction de répartition s'écrit :

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$f(y) = exp(log(f(y))) = ylog(\tfrac{p}{1-p}) + log(\binom{n}{y})$$

Cette fonction peut se réécrire sous le même format que celui de l'équation (1) avec :

$$\theta = log(\tfrac{p}{1-p})$$

$$a(\phi) = 1$$

$$b(\theta) = nlog(1 + e^{\theta})$$

Loi Gamma

Soit Y une variable aléatoire qui suit une loi gamma. Sa fonction de densité s'écrit sous la forme :

$$f(t, \lambda, y) = \tfrac{1}{\Gamma(t)} \lambda^t y^{t-1} e^{-\lambda\mu} \text{ avec } t > 0,\ y > 0 \text{ et } \lambda > 0$$

De même que le cas précédent :

$$f(y) = exp(tlog(\lambda) - +(t-1)log(y) - log(\Gamma(t)))$$

Cette expression correspond à la forme générale des fonctions exponentielles avec :

$$\theta = -\lambda$$

$$\phi = 1$$

$$a(\phi) = 1$$

$$b(\theta) = -tlog(-\theta)$$

$$c(y, \phi) = (t-1)log(y) - log(\Gamma(t))$$

Expression mathématique

Après avoir introduit la notion de famille exponentielle, on présentera dans cette partie l'expression mathématique des modèles GLM. Introduits par Nelder Wedderburn en 1972, les modèles GLM supposent l'existence d'une fonction de lien g inversible qui relient entre la combinaison linéaire des variables explicatives $X = (X^1, ..., X^k)$ et l'espérance de la variable Y.

Mathématiquement, cette équation s'écrit sous la forme suivante :

$$g(E(Y \mid X)) = g(\mu(X)) = X\beta$$

Avec,

- $Y \mid X$ est la variable aléatoire à expliquer dont la densité appartient à la famille exponentielle
- g est inversible
- Le prédicteur μ est construit à partir de la combinaison linéaire des variables explicatives.

$$\mu = X\beta$$

Il s'avère donc essentiel pour construire un modèle GLM de choisir au préalable la loi de la variable à expliquer ainsi que la fonction de lien adéquate. C'est la différence majeure que l'on peut noter par rapport aux modèles linéaires simplistes. Ces derniers représentent un cas particulier des modèles GLM en choisissant la fonction d'identité pour la fonction de lien et la loi normale centrée pour la variable Y. le choix de la fonction de lien dépend de la loi de la variable aléatoire Y. Le tableau ci-dessous récapitule quelques exemples de lois couramment utilisées ainsi que leurs fonctions de lien correspondantes.

Loi	**Fonction de lien**
Bernoulli/Binomiale	$g(\mu) = logit(\mu) = ln(\frac{\mu}{1-\mu})$
Normale	$g(\mu) = \mu$
Poisson	$g(\mu) = ln(\mu)$
Gamma	$\frac{-1}{\mu}$

Estimation des paramètres du modèle

Si on dispose de n observations de la variable Y, l'expression mathématique présentée ci-dessus peut se réécrire sous le format suivant :

$$g(E[Y_i \mid X_i]) = \beta_0 + \sum_{j=1}^{k} \beta_j X_i^j$$

Une fois la loi et la fonction de lien fixée, on cherche à trouver la meilleure approche pour estimer la valeur des coefficients du modèle. On explicitera dans cette partie la méthode d'estimation par maximum de vraisemblance.

Estimation par maximum de vraisemblance

Cette méthode consiste à trouver les coefficients qui maximisent la fonction de vraisemblance du modèle.
On définit la fonction de vraisemblance de la manière suivante :

$$\mathfrak{L}(\beta, y) = \prod_{i=1}^{N} f(y_i, \beta)$$

où f est la fonction de densité de la loi de Y. On pose : $l(\beta, y) = ln(\mathfrak{L}(\beta, y)) = \sum_{i=1}^{n} f(y_i, \beta)$

Ainsi, Les coefficients $\hat{\beta}_j$ pour $j = 0...k$, estimateurs des coefficients du modèle, s'obtiennent en résolvant pour chaque j :

$$\frac{\partial l(\beta_0,...,\beta_k,y))}{\partial \beta_j} = 0$$

Par ailleurs, ces coefficients doivent vérifier la condition du second ordre : $\frac{\partial^2 l(\beta_0,...,\beta_k,y))}{\partial \beta_j^2} < 0$

Cas particulier : Régression logistique

Après avoir introduit la théorie des modèles GLM, on s'intéresse dans cette partie à un cas particulier des modèles GLM qui sera utilisé pour la modélisation des sinistres graves : Le modèle de régression logistique (modèle logit). Il est essentiellement utilisé quand la variable à expliquer est binaire. C'est le cas de la variable cible de cette étude target_SNBSIN2GRAV : Elle vaut 1 si l'assuré a eu un sinistre grave et 0 sinon.

Expression mathématique du modèle logit

Soit Y la variable à expliquer et $X = (X_1, ..., X_k)$ le vecteur des variables explicatives. Dans le cas de cette étude, la variable Y traduit l'occurence d'un sinistre grave pour un assuré donné.

Soit $\tau = P(Y = 1 \mid X = x)$ et $1 - \tau = P(Y = 0 \mid X = x)$ avec $\tau \in \left[0 \quad , \quad 1\right]$. Ainsi, Y suit donc une loi de Bernoulli de paramètre τ.
Le modèle s'écrit ainsi avec la fonction de lien logit :

$$g(\tau) = ln(\tfrac{\tau}{1-\tau}) = X'\beta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \ (2)$$

où :

- g est la fonction de lien qui est supposée inversible
- le vecteur $\beta = (\beta_0, ..., \beta_k)$ est le vecteur des coefficients du modèle

En appliquant la fonction exponentielle à l'équation 2, on peut écrire :

$$\tau(x) = \tfrac{exp(x'\beta)}{1+exp(x'\beta)} = \tfrac{1}{1+exp(-x'\beta)} \ (3)$$

Estimation des paramètres du modèle

Une fois l'équation du modèle est posée, il reste à estimer le vecteur des coefficients β. On a choisi la méthode d'estimation par vraisemblance.

On suppose que l'on dispose d'un échantillon de n observations i.i.d (y_i, x_i). La fonction de vraisemblance s'écrit :

$$\mathfrak{L} = \prod_{i=1}^{n} \left(\frac{exp(\beta_0+\beta' x_i)}{1+exp(\beta_0+\beta' x_i)}\right)^{y_i} \left(\frac{exp(\beta_0+\beta' x_i)}{1+exp(\beta_0+\beta' x_i)}\right)^{1-y_i}$$

$$\mathfrak{L} = \prod_{i=1}^{n} \tau(x_i)^{y} i (1-\tau(x_i))^{y} i$$

Ainsi, on peut écrire la fonction log-vraisemblance :

$$l = ln(\mathfrak{L}) = \sum_{i=1}^{n} y_i ln(\tau(x_i))(1-y_i) ln(1-\tau(x_i))$$

On applique ensuite les conditions de premier ordre. Les coefficients du modèle sont donc solution du système suivant :

$$\begin{cases} \frac{\partial l(\beta_0,...,\beta_k)}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \tau(x_i)) = 0 \\ \frac{\partial l(\beta_0,...,\beta_k)}{\partial \beta_j} = \sum_{i=1}^{n} x_i^j (y_i - \tau(x_i)) = 0 \text{ pour } j = 1, ..., k \end{cases}$$

La solution analytique de cette équation n'est pas facile à calculer. La résolution nécessite l'utilisation des méthode itératives de résolution telles que l'algorithme de Newton-Raphson ou la méthode des scores de Fisher.

Après avoir présenté les modèles linéaires généralisés et plus précisément la régression logistique qui sera utilisée pour modéliser les sinistres graves. On introduira dans la suite l'outil utilisé pour la modélisation, les critères de choix des variables ainsi que les critères de choix du modèle.

3.3 L'outil utilisé et les critères de sélection des variables

Le secteur d'assurance se caractérise par l'inversion du cycle de production. En d'autres termes, l'assureur doit fixer le prix de son produit sans avoir de visibilité sur les coûts qu'il aura à régler, d'où l'importance de l'étape de tarification. Cette dernière permet aux assureurs de satisfaire leurs engagements vis-à-vis de leurs assurés. Plusieurs solutions sont aujourd'hui à disposition des assureurs pour leur permettre d'implémenter les modèles GLM et d'effectuer la modélisation des primes pures. Parmi ces solutions, on peut citer la plateforme Akur8 utilisée à Generali pour implémenter les modèles de risque. En effet, Akur8 est une assurtech qui propose aux assureurs un logiciel de modélisation du risque s'appuyant sur des algorithmes de machine Learning. Cette solution Saas (Software as a service) permet d'automatiser le processus de tarification tout en gardant une transparence et un contrôle complet de la part des assureurs sur leurs modèles.

On présentera ci-dessous les méthodes utilisées pour sélectionner les variables explicatives les plus pertinentes à inclure dans le modèle. Par un souci de confidentialité, le processus de sélection des variables explicatives utilisé par Akur8 ne pourra pas être détaillé dans ce mémoire.

3.3.1 Critères de Sélection des variables

Méthode classique de sélection des variables

L'étape de sélection des variables est primordiale pour avoir un bon modèle GLM permettant de décrire au mieux la variable à expliquer. Les variables retenues dans le modèle final doivent satisfaire deux conditions essentielles. D'une part, on cherche à trouver le panel de variables explicatives qui permet d'expliquer de la manière la plus réaliste et la plus complète la sinistralité grave. D'autre part, on souhaite avoir un nombre juste nécessaire de variables explicatives : Toutes variables non significatives ajoutées au modèle diminuent la précision des coefficients estimés.

Cette étape sera primordiale dans la modélisation de la probabilité d'avoir un sinistre grave pour l'ensemble des assurés sinistrés. Comme on l'a explicité dans les parties précédentes, l'objectif est d'enrichir le modèle tarifaire déjà utilisé à Generali et trouver de nouvelles variables qui permettent de mieux expliquer la sinistralité grave. Cette étape de sélection de variables nous permettra ainsi de s'assurer de la pertinence des nouvelles variables ajoutées à l'ensemble des variables tarifaires pour décrire la probabilité d'un sinistre grave.

Pour sélectionner les variables qui expliquent au mieux le risque, l'assureur se base sur un certain nombre d'exigences : légales, techniques et opérationnelles. Les variables retenues par un modèle tarifaire (respectivement de fréquence) doivent satisfaire les exigences légales imposées par le régulateur. La variable sexe, par exemple, ne peut plus être utilisée comme variable discriminante du risque depuis décembre 2012 (La cour de justice européenne). Par ailleurs, Il est préférable que ces variables ne soient pas très coûteuses opérationnellement. Finalement, ces variables doivent être stables et pertinentes pour décrire la sinistralité. On détaillera dans la suite deux procédures utilisées pour une sélection optimale des variables explicatives à savoir la sélection pas à pas et la sélection exhaustive.

Sélection pas à pas

La méthode de sélection pas à pas est souvent utilisée et permet d'avoir une première idée sur les variables les plus significatives (une autre mesure de la significativité des variables est le spread qui sera explicité dans la partie suivante). Il existe trois approches pour cette procédure.

Le principe de la sélection pas à pas est le suivant : On sélectionne un certain nombre de variables puis on effectue un test de significativité à chaque itération permettant de réduire ou augmenter l'échantillon initial de variables explicatives par suppression ou ajout de nouvelles variables.

- La sélection ascendante (forward)

On part d'un modèle vide (avec aucune variable sélectionnée au préalable). La sélection s'effectue au fur et à mesure en ajoutant la variable qui augmente les performances du modèle. En d'autres termes, à chaque étape de la sélection, on choisit la variable qui permet d'améliorer les mesures de performance du modèle. On peut à ce propos utiliser le test de Wald ou le test de Student qui sont asymptotiquement équivalents ou bien le R^2.

- La sélection descendante (backward)

Dans cette approche, on fait le chemin inverse de l'approche précédente. On part d'un modèle avec toutes les variables. On élimine au fur et à mesure la variable la moins significative suivant les tests de significativité effectués. Le processus sera arrêté dès qu'on n'arrive plus à détecter une variable pertinente permettant d'augmenter les performances du modèle.

L'inconvénient de ces deux approches est le fait que la variable ne sera testée qu'une seule fois. Dans le premier cas de sélection ascendante, une fois la variable retenue, elle ne pourra plus être éliminée. Dans le second cas de sélection descendante, une fois la variable n'a pas été choisie, on ne pourra plus la sélectionner dans le modèle. Il s'avère ainsi que l'ordre de test des variables est déterminant pour les deux approches.

- La sélection mixte (stepwise)

La sélection mixte permet, en partie, de remédier à ce problème. Cette méthode est une combinaison entre les deux précédentes. On modifie l'approche de sélection ascendante : A chaque itération, on revérifie la significativité de toutes les variables déjà retenues. Elles seront enlevées si leur significativité a baissé par rapport à un certain seuil de tolérance.

Sélection exhaustive

Cette approche est différente de celle décrite ci-dessus. Elle est utilisée notamment si on dispose d'un certain nombre de variables qui n'est pas trop élevé.

On note k le nombre de variables explicatives à disposition. Ainsi, le nombre total de modèles faisant intervenir exactement r variables que l'on pourra construire est : $C_k^r = \frac{k!}{r!(k-r)!}$
Si on considère toutes les valeurs possibles de r, on peut construire au total $\sum C_k^r = 2^k$ modèles. Ainsi, le modèle sélectionné est celui qui a de meilleurs performances parmi tous ceux qui sont à disposition.

Cette méthode présente l'avantage de permettre de couvrir tout le champ des modèles possibles avec les variables explicatives à disposition. Cependant, elle s'avère très coûteuse si ce nombre est assez élevé.

3.3.2 Critères de choix du modèle

On a détaillé dans la partie précédente les différents critères utilisés pour sélectionner les variables explicatives les plus pertinentes à inclure dans le modèle. On s'intéresse dans la suite aux critères de choix du modèle. En d'autres termes, on va présenter les indicateurs qui permettent de choisir le modèle le plus performant.

Le coefficient de Gini

Le coefficient de Gini est une mesure d'inégalité. Souvent utilisé en économie pour étudier l'inégalité des rémunérations dans un pays donné. Ce coefficient est également une mesure de performance des modèles. Plus précisément, le coefficient de Gini mesure la qualité de segmentation du modèle. Il peut être visualisé à l'aide de la courbe de Lorenz ci-dessous 3.1. Dans notre cas, l'axe des abscisses représente la part cumulée de l'exposition et l'axe des ordonnées représente la part cumulée de la fréquence.

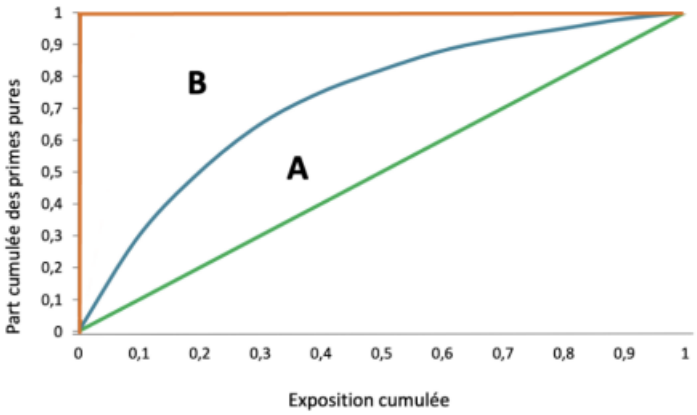


FIGURE 3.1 – Un exemple de la courbe de Lorenz

On peut mesurer ce coefficient à l'aide de la courbe de Lorenz à l'aide de la formule suivante :

Indice de GINI $= \frac{AireA}{AireA+AireB}$

Avec,
- A l'aire délimitée par la courbe de Lorenz (en bleu) et la droite d'égalité parfaite (en vert)
- B l'aire au-dessus de la courbe de Lorenz (en bleu)

La courbe Lift

La courbe lift est une mesure du pouvoir prédictif du modèle. Celle-ci est construite en triant les prédictions par ordre croissant puis en les répartissant en vingt groupes qui représentent chacun 5% des prédictions. On calcule ensuite les prédictions moyennes pour chaque valeur observée. L'outil utilisé génère, d'une part, une courbe avec les valeurs observées et, d'autre part, une courbe avec les valeurs prédites.
Si la courbe des risques prédits et observés se suivent (comme le montre le graphique ci-dessous 3.2), le modèle s'adapte bien aux données. Au contraire, une courbe plus chaotique signifie que les prédictions ne sont pas très fiables.

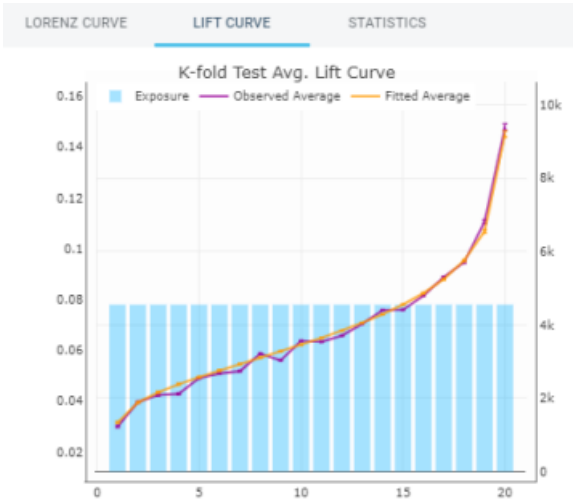


FIGURE 3.2 – Un exemple de représentation de la courbe Lift

La courbe ROC

La courbe de ROC (Receiver Operating Characteristic) est une représentation graphique (La courbe en rouge dans le graphique ci-dessous 3.3) du taux des vrais positifs en fonction du taux des faux positifs. Elle est couramment utilisée pour mesurer l'efficacité en terme de prédiction d'un modèle de classification binaire.

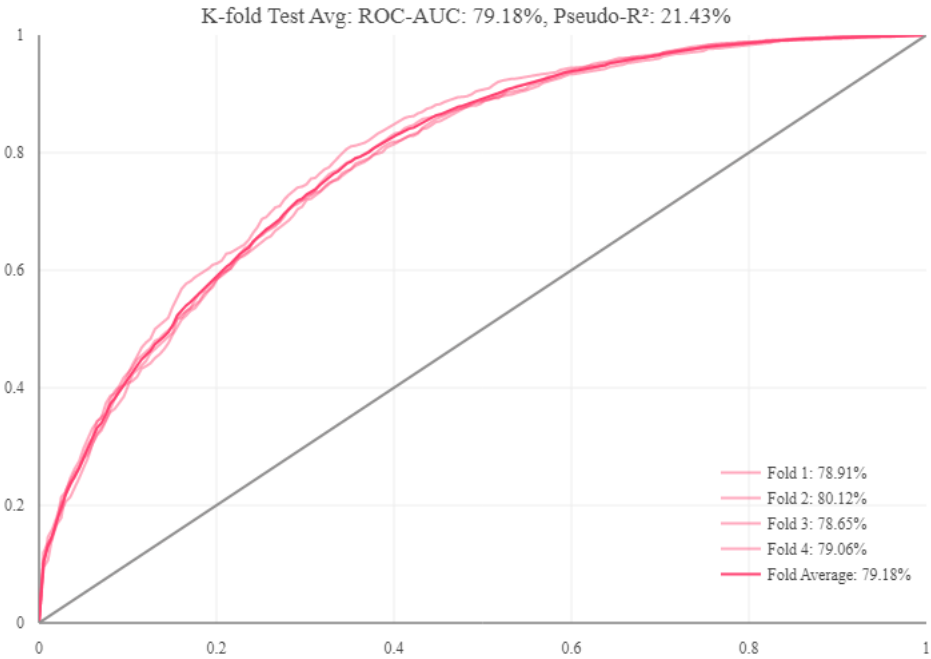


FIGURE 3.3 – Un exemple de représentation de la courbe ROC

La courbe ROC est souvent exploitée pour calculer l'AUC correspondant à l'aire sous cette courbe. Ce coefficient permet de mesurer la capacité prédictive du modèle. Il indique la probabilité de prédire un vrai positif. Cette valeur est comprise entre 0 et 1. Si le coefficient AUC vaut 0.5, les prédictions du modèle sont considérées aléatoires et le modèle n'est pas fiable. On peut considérer que le modèle est fiable à partir d'un AUC égal à 0.7.

Les résidus

En plus des outils graphiques présentés ci-dessous pour analyser la performance du modèle, une analyse soigneuse des résidus s'avère essentielle. Les résidus représentent le bruit du modèle ou l'erreur observée. Cette mesure de distance entre les prédictions et les valeurs observées permet de s'assurer de la robustesse du modèle. On explicitera ci-dessous les mesures de résidus couramment utilisées pour les modèles GLM.
La formule générique de calcul des résidus est la suivante :

$$r_{Pi} = y_i - \hat{y}_i$$

On présentera dans la suite trois types de résidus permettant d'évaluer la qualité des modèles GLM.

Les résidus de pearson

Les résidus de Pearson se calculent comme l'écart entre les valeurs prédites et les valeurs observées rapportées par l'écart type estimé des valeurs prédites.

$$r_{Pi} = \frac{y_i - \hat{y}_i}{s_i}$$

Avec,
- y_i les valeurs observées
- $\hat{y}_i$ les valeurs prédites
- s_i l'écart type de $\hat{y}_i$

Cependant, tels qu'ils sont présentés ci-dessous, ces résidus ne sont pas homogènes à une variance. Il est donc difficile de les interpréter. On calcule souvent les résidus standardisés comme suit :

$$r_{Psi} = \frac{y_i - \hat{y}_i}{s_i}$$

Les résidus de déviance

Les résidus de déviance sont calculés en fonction de la déviance du modèle qui dépend de la distribution choisie selon la formule suivant :

$$res_{Deviance} = signe(y - \hat{y})\sqrt{Deviance(y, \hat{y}, w)}$$

La déviance peut être calculée en fonction de la distribution du modèle de la manière suivante :

$$PoissonDeviance = 2.(yln(\tfrac{y}{\hat{y}}) - (y - \hat{y}))$$

$$GammaDeviance = 2.(yln(\tfrac{y}{\hat{y}}) - \tfrac{y-\hat{y}}{\hat{y}}))$$

$$GaussianDeviance = (y - \hat{y})^2$$

À l'exception des distributions gaussiennes, il n'existe aucune garantie théorique quant à la forme que devrait avoir les résidus de déviance d'un modèle bien ajusté.

Les résidus quantiles

les résidus quantiles sont une alternative aux résidus de déviance. Ils sont calculés en transformant la distribution du modèle en une distribution normale comme suit :

$$Res_{quantile} = \Phi^{-1}(\mathcal{F}(y, \hat{y}))$$

Avec :

- ϕ est la fonction de distribution d'une loi normale
- $\mathcal{F}$ est la fonction de répartition de la distribution choisie pour le modèle

L'outil utilisé propose deux types de résidus : les résidus de déviance et les résidus quantiles. Il génère pour chaque modèle une représentation graphique des résidus en fonction des prédictions obtenues. Cependant, ces graphiques sont difficilement interprétables. Il n'y a aucune garantie théorique de la forme de ces courbes. Cependant, il est primordial de vérifier que la courbe des résidus obtenue est symétrique centrée autour de 0 avant la validation du modèle.

La RMSE : ROOT OF MEAN SQUARE ERROR

L'erreur quadratique moyenne est la racine carrée de la somme des carrés de la différence entre l'observé et le prédit. Elle est équivalente à la log-vraisemblance d'un modèle gaussien. Cette métrique est impactée par les valeurs ou erreurs extrêmes. Elle représente l'une des métriques les plus importantes à prendre en considération lors de la validation d'un modèle.

La RMSE se calcule à l'aide de la formule suivante :

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

Cette expression ne tient pas compte du poids de chaque observation (l'exposition). En pratique, on utilise le plus souvent une mesure pondérée de la RMSE calculée à l'aide de la formule suivante :

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

Le spread

Le spread est un indicateur calculé pour chaque variable. Il permet de mesurer le degré d'importance des variables à travers la dispersion des coefficients calibrés. On distingue deux mesures différentes du spread : le spread 100/0 et le spread 95/5. La première mesure est calculée en prenant en compte tout le portefeuille. La deuxième, quant à elle, se calcule en retirant 5% de l'exposition la plus risquée (coefficients les plus élevés) et 5% de l'exposition la moins risquée (coefficients les plus faibles).

Le spread 100/0

Le spread 100/0 est calculé, pour une variable donnée, à l'aide de la formule suivante en fonctiton du coefficient maximal et du coefficient minimal :

$$\frac{coefficient\_max+1}{coefficient\_min+1} - 1$$

Le graphique ci-dessous 3.4 montre la distribution des coefficients de la variable *agecon-ducteur* ainsi que l'exposition correspondante. On peut ainsi déduire le spread 100/0 qui est égal à : $\frac{2.17}{0.97} - 1 \sim 1.23$

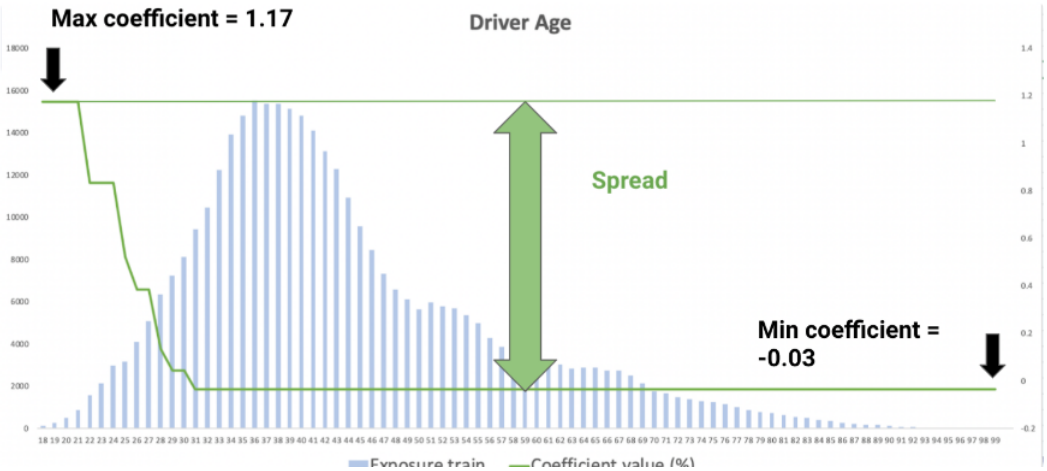


FIGURE 3.4 – Schéma explicatif de la méthode de calcul du Spread 100/0

Le spread 95/5

Le spread 95/5 (3.5) se calcule de la même manière que le spread 100/0 en supprimant les 5 % de l'exposition la plus risquée et les 5 % de l'exposition la moins risquée.

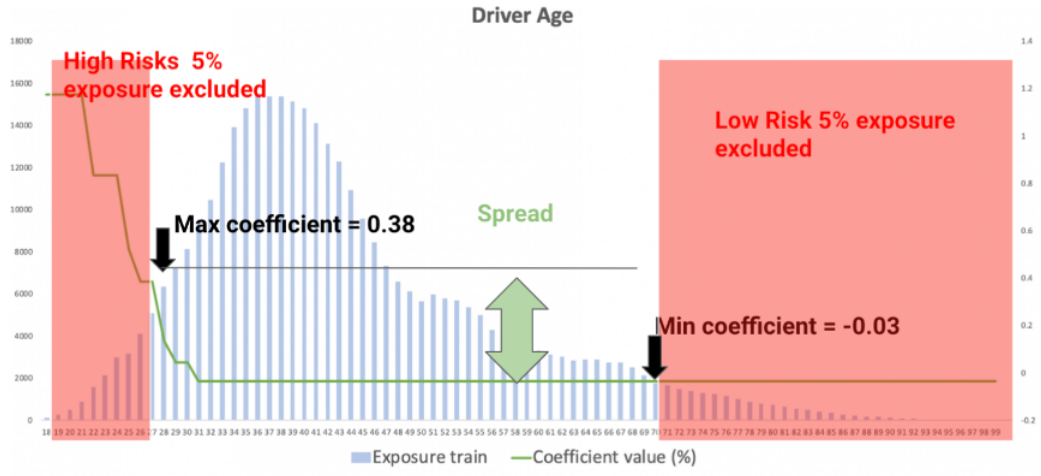


FIGURE 3.5 – Schéma explicatif de la méthode de calcul du Spread 95/5

La validation croisée

L'étape de validation croisée est primordiale pour s'assurer de la robustesse du modèle obtenue. Elle consiste à évaluer le modèle obtenu sur une base test distincte de la base d'apprentissage. Cette évaluation est indisponsable pour s'assurer de la possibilité de généralisation du modèle à d'autres bases de données.

La validation croisée se fait en utilisant le mécanisme K-folds cross validation suivant les étapes suivantes :

- Étape 1 : la première étape consiste à définir la base d'apprentissage et la base de test. On prend généralement 80% de la base de modélisation pour l'apprentissage et 20% pour le test.

- Étape 2 : la deuxième étape consiste à découper la base d'apprentissage et la base test en k échantillons (souvent dits *k-folds)*).

- Étape 3 : le troisième étape consiste à répéter les procédures d'entrainement et de test k-fois. A chaque itération, on prend k-1 des échantillons comme base d'apprentissage et l'échantillon restant servira comme base test.

Le graphique ci-dessous (3.6) montre un exemple de la procédure de validation croisée effectuée sur 4 échantillons.

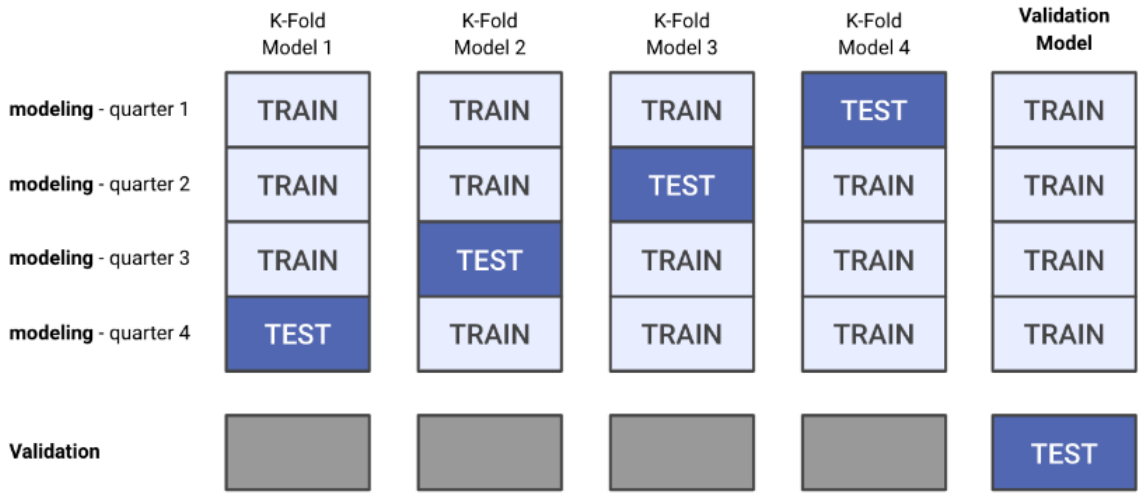


FIGURE 3.6 – Schéma explicatif de la procédure de validation croisée effectuée sur 4 échantillons

Outre les modèles GLM, on a utilisé l'algorithme XGBoost pour la modélisation de la probabilité de la sinistralité grave. On présentera dans la partie suivante le principe de ce modèle.

3.4 Méthode d'apprentissage automatique : XGBoost

Cette partie présentera les éléments théoriques liés à la compréhension de l'algorithme XGBoost.
Afin de bien comprendre le fonctionnement de cet algorithme d'apprentissage automatique, il est nécessaire d'introduire la notion de boosting.

3.4.1 Le boosting

Le boosting est une technique ensembliste qui consiste à agréger différents classifieurs faibles (ayant une grande valeur d'erreur) pour en former un meilleur classifieur. Cette amélioration du pouvoir prédictif du classifieur faible se fait à travers la construction de modèles en séquence. Ce processus séquentiel s'effectue à travers les différents poids attribués aux différentes observations. A chaque nouvelle itération d'un algorithme de boosting, les valeurs qui ont été le plus mal ajustées aux précédent classifieur se verront attribuées les poids les moins importants.

3.4.2 Présentation de l'algorithme XGboost

Le terme XGBoost signifie Extreme Gradient Boosting. Il est considéré comme une autre désignation de l'approche Gradient Boosting Machine introduite par Friedman en 1999. Cet algorithme est également considéré comme la version la plus améliorée parmi les modèles d'agrégations qui utilisent l'algorithme CART. Il utilise la méthode de boosting précédemment décrite ainsi que la méthode du gradient. Cette dernière fait également partie des méthodes de résolution des problèmes d'optimisation différentiable. L'idée est de suivre le sens de descente du gradient pour définir le minimum de la fonction en question.

Afin de mieux comprendre les différentes étapes de cet algorithme, on présentera dans la suite son formalisme mathématique.

Notations

- M le nombre total des étapes de l'algorithme, on note $m \in \{1, ..., M\}$
- $X = (x_i)_{1<i<n}$ où x_i est le vecteur des caractéristiques pour chaque individu i
- $Y = (y_i)_{1<i<n}$ où y_i est la variabele à prédire pour chaque individu i
- $\mathfrak{M}$ est l'ensemble des modèles qu'on cherchera à optimiser
- f une fonction de coût choisie au prélable pour évaluer les modèles (par exemple la MSE : MEAN SQUARE ERROR)
- $\hat{\phi}^{(m)}$ est la fonction qui renvoit le vecteur de prédiction à la sortie d'un modèle M_m à l'étape b de l'algorithme
- $\hat{y}^{(m)}$ l'estimateur correspondant à l'étape m
- $\hat{r}^{(m)}$ est l'estimateur du vecteur des résidus au début de l'étape m
- Ω est une fonction qui pour un modèle donné renvoie une mesure de sa complexité

On peut déduire la relation suivante entre les différents éléments détaillés ci-dessus :

$$\hat{y}_i^{(m)} = \sum_{k=0}^{(m)} \hat{r}_i^{(k)} = \hat{y}_i^{(m-1)} \text{ et } \hat{y}_i^0 = \hat{\phi}(x_i) \text{ avec } 1 \leq i \leq n \text{ et } 1 \leq m \leq M$$

Etapes de l'algorithme

- Etape 1 :
  On définit un premier modèle constant renvoyant les mêmes prédictions pour tous les vecteurs des variables explicatives $\phi^{(0)}(x_i) = Constante$
- Etape 2 :
  Pour m=1 à M, la fonction $\hat{\phi}^{(m)}$ associée au modèle M_m renvoie une estimation des résidus du modèle précédent : $\hat{\phi}^{(m)}(x_i)$ cherche à estimer $y_i - \hat{y}_i^{(m-1)}$. Pour ce faire, il faut optimiser la fonction de coût du modèle et chercher $\hat{\phi}^{(m)}$ solution du problème d'optimisation suivant :

$$min(\sum_{i=1}^{n} f(y_i, \hat{y}_i^{(m)}) + \sum_{k=1}^{M} \Omega(\phi_k))$$

La solution de ce problème d'optimisation dépendra du choix de la fonction de coût f.

- Etape 3 :
  Les itérations de l'étape 2 permettent à la fin d'obtenir la fonction associée au modèle M_m. En appliquant cette fonction au vecteur des variables explicatives, on obtient l'output du modèle : les prédictions $\hat{y}_i$ pour chaque y_i.

4

Création de la base de modélisation

Dans la partie précédente, on a détaillé les éléments théoriques utilisés pour la modélisation notamment la théorie des modèles linéaires généralisés et le principe du modèle XGBoost. On s'intéresse dans ce chapitre à la construction de la base de modélisation, étape primordiale pour assurer des résultats robustes.

On commence, tout d'abord, par présenter la méthode de choix du seuil des sinistres graves ainsi que les différentes bases de Generali qui vont servir pour la création de la base de modélisation. L'objectif est d'arriver à trouver de nouvelles variables explicatives, principalement des variables de la base des avenants, qui décrivent bien, en plus des variables tarifaires, la sinistralité grave.
Deux bases principales ont servi pour cette étude : la base tarifaire classique utilisée pour créer les modèles de risque au sein de Generali et la base *« Avenant »* qui sera construite à partir des bases brutes mensuelles. Ces deux bases seront explicitées dans ce chapitre.

L'étape suivante consiste à détailler la méthodologie utilisée pour créer la base de modélisation après avoir préparée et nettoyée la base tarifaire et la base *« Avenant »*. On expliquera principalement les différentes pistes explorées, les résultats obtenus et la méthode retenue.

4.1 Choix du seuil avec la méthode du QQ-plot

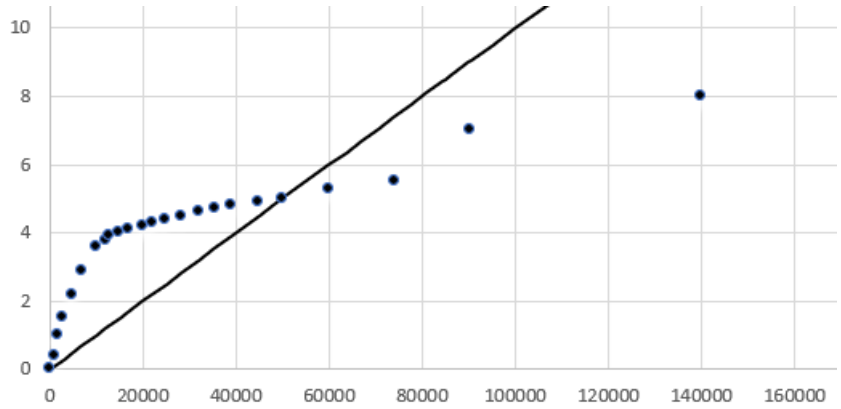


FIGURE 4.1 – Loi empirique des montants des sinistres versus la loi exponentielle

La distribution théorique utilisée afin de déterminer le seuil des sinistres graves est celle de la distribution exponentielle. D'après le graphique représenté ci-dessus, on observe un changement de la concavité de la courbe pour des valeurs comprises entre 40 000 et 60 000. La distribution des données présente une queue plus épaisse que la loi exponentielle. On retient le seuil de 50 000 par cette méthode.

4.2 Les bases utilisées pour l'étude

4.2.1 La base tarifaire

La base tarifaire ainsi construite regroupe tous les contrats présents dans le portefeuille du produit « L'Auto Generali » entre la mensuelle du mois de janvier 2018 et la mensuelle du mois de décembre 2020. Chaque ligne correspond à une situation de risque. Une situation de risque est, par définition, une période durant laquelle les variables liées à un contrat donné sont restées stables. Chaque changement dans les critères tarifaires génère donc une nouvelle période d'exposition.

La base ainsi obtenue contient des variables décrivant le profil de l'assuré et de son véhicule auxquelles on ajoutera les informations concernant la sinistralité antérieure du contrat ainsi que les caractéristiques du conducteur secondaire.

Le tableau ci-dessous montre un exemple fictif représentatif d'un extrait de la base tarifaire.

La colonne IDBASE représente l'identifiant unique pour chaque ligne dans la base tarifaire. Il est construit en concaténant le numéro de contrat, le numéro du véhicule ainsi que les

dates de début et de fin de la situation de risque. Les colonnes target_SNBSIN2RCC et target_SNBSIN2GRAV sont, quant à elles, des variables binaires qui inidiquent respectivement l'occurence ou non d'un sinistre ainsi que l'occurence ou non d'un sinistre grave.On a choisi ici uniquement quelques variables tarifaires (âge, ancienneté du permis,....) pour expliciter la structure de la base tarifaire.

IDBASE	ageconducteur	anciennetepermis	target_SNBSIN2RCC	target_SNBSIN2GRAV
id1	21	3	1	0
id2	32	3	1	1

Présentation des variables tarifaires

On peut distinguer les variables tarifaires suivant le risque qu'elles décrivent :

Le risque véhicule

Les variables qui décrivent le risque lié au véhicule dans la base tarifaire sont les suivantes :

- *L'âge du véhicule*

- *L'usage du véhicule*

- *La classe et le groupe SRA*

La classe et le groupe SRA sont des informations fournies par l'organisme Sécurité et Réparation Automobile (SRA). Ils permettent à l'assureur d'avoir une visibilité sur les caractéristiques techniques du véhicule notamment sa marque, son modèle, sa valeur, sa puissance, la qualité de ses airbags, etc.

Plus précisément, le groupe SRA est indicateur synthétique de la puissance de la vitesse et du poids. Il renseigne l'assureur sur la puissance du véhicule. Cette variable est numérique et prend des valeurs qui varient entre 20 et 50. 20 signifie que le véhicule est de très faible puissance alors que 50 est un indicateur d'une forte puissance.

La classe SRA, quant à elle, indique la tranche de coût du véhicule. Cette variable est alphanumérique. Elle prend des valeurs comprises entre A et ZA. La valeur A correspond à un véhicule peu coûteux tant dis que la valeur ZA correspond à un véhicule très coûteux.

- *Le CRM*

Le CRM est l'abréviation du Coefficient de Réduction- Majoration souvent appelé bonus-malus. Cette variable est numérique et comprise entre 0.5 et 3.5. Ce coefficient est mis à jour à chaque échéance annuelle. A la souscription du contrat, l'individu se voit attribué un coefficient de CRM égal à 1. Si l'assureur passe 13 ans sans sinistres responsables, ce coefficient passe à 0.5. Pour chaque sinistre responsable, il est multiplié par 1.25.

Ce coefficient agit sur le montant de prime payé par l'assuré. Plus le nombre de sinistres responsables augmente, plus le montant des primes augmente.

- *Présence ou non d'un garage*

Le risque conducteur

Les variables tarifaires qui décrivent le risque lié au conducteur sont les suivantes :

- *Le sexe du conducteur*

- *L'ancienneté du permis*

- *La situation maritale*

- *La catégorie socio-professionnelle*

Les variables présentées ci-dessus permettent de dresser un profil pour le conducteur et donc du risque de sinistralité qu'il peut présenter : Un conducteur jeune avec un permis récent a plus de chance d'avoir un sinistre qu'un conducteur âgé.

Le risque géographique

La zone géographique est un facteur explicatif de la sinistralité. En effet, de manière schématisée on peut dire qu'une zone encombrée augmente fortement la probabilité d'occurrence d'un sinistre. Inversement, les lieux d'habitation peu peuplé baisse le risque.

Les carcatéristiques du contrat

En plus des variables qui décrivent le profil du conducteur ainsi que son véhicule, pour chaque contrat, on dispose de ses caractéristiques. Ces dernières explicitent les garanties souscrites, les différentes franchises, etc.

On peut résumer les variables décrivant les caractéristiques du contrat dans la liste ci-dessous :

- *La présence ou non de la garantie BDG*

- *La formule de garantie : L1, L2, l3*

- *Les franchises vol*

- *Les franchises BDG*

- *Les franchises conducteur*

- *L'ancieneté du contrat*

- *Le fractionnement de la prime*

4.2.2 La base avenant

Après avoir récupéré les informations contenues dans la base tarifaires pour tous les contrats dans la période d'étude (de 2018 à 2020), on cherche à enrichir cette base avec des variables *« avenants »*. On s'intéresse au comportement antérieur de l'assuré et à tous les changements effectués qui ont affecté sa sinistralité. Plus précisément, on souhaite récupérer tous les mouvements qui ont eu lieu pour chaque contrat de la base tarifaire entre les années 2015 et 2020, pour obtenir une information d'avenants sur les 3 années antérieures à chaque situation de risque de la base tarifaire.

On commence par récupérer tous les contrats dans le portefeuille du produit de « L'Auto Generali » pour chacune des années de l'étude à partir des bases brutes mensuelles. L'étape suivante consiste à construire les variables *« avenants »*. Un avenant correspond, par définition, à une modification du contrat d'assurance. Cette modification peut être liée au conducteur principal, aux caractéristiques du véhicule, aux différentes garanties du contrat, etc. On va restreindre notre étude à 4 changements principaux qui paraissent les plus pertinents pour expliquer la sinistralité grave.

On s'intéresse principalement aux changements suivants :

Un changement de groupe représenté par la variable *chgt_ groupe*

On récupère la différencre entre l'ancienne valeur du groupe du véhicule de l'assuré et la valeur actuelle. Une différence positive indiquera une amélioration du véhicule. Au contraire, une valeur négative est un signe d'une baisse de la puissance du véhicule.

Un changement de l'âge du conducteur principal et de son ancienneté de permis (*chgt_ naiss_ conducP* et *chgt_ date_ perm*)

Dans le cas d'un avenant du conducteur principal, on récupère la différence entre l'âge du nouveau conducteur et celui de l'ancien conducteur (ou bien la différence entre la date d'obtention de permis du nouveau conducteur et celle de l'ancien conducteur). Ces deux variables serviront pour détecter l'influence du passage à un conducteur plus jeune sur la sinistralité.

Un changement de zone représenté par la variable *chgt_ Zone*

Cette variable est construite à partir de la variable numérique zone_maj_RC qui attribue pour chaque code Insee un coefficient décrivant le risque lié à la commune en question. Plus ce coefficient est élevé, plus le risque de sinistralité est élevé.
La variable construit *chgt_ Zone* aura ainsi trois modalités différentes :

- "-1" qui correspond à un déménagement vers une zone moins risqué.
- "1" qui correspond à un déménagement vers une zone plus risquée
- "0" qui correspond à l'absence d'un changement de zone ou bien à un déménagement dans une zone avec le même niveau de risque

La base *avenant* finale sera donc la concaténation des 5 bases construites regroupant les différents mouvements annuels pour chacun des assurés. Le schéma ci-dessous montre un exemple représentatif de la structure de la base *avenant* finale obtenue. Ici, on a pris en compte uniquement les colonnes des avenants *chgt_ naiss_ conducP* et *chgt_ Zone*.

POLICE	Chgt_naiss_conducP	Chgt_Zone	date_avenant
Police1	-5	-1	10/03/2018
Police1	-3	0	11/12/2017
Police2	4	0	12/06/2016

4.2.3 Méthodologie de création de la base de modélisation

Après avoir présenté les différentes variables *avenant* qui nous intéressent pour cette étude, on s'intéresse dans la suite à la méthode de construction de la base de modélisation. Cette procédure s'effectue en trois étapes.

Etape 1 : Création des blocs des avenants

Cette première étape consiste à transformer la structure de la base *avenant* afin de faciliter sa jointure plus tard avec la base tarifaire. Le but est de garder une seule ligne par *POLICE* dans la base *avenant* en créant un bloc par avenant par police. En considérant l'exemple ci-dessus, la base finale transformée aura la structure suivante :

POLICE	Chgt_naiss_conducP_0	Chgt_Zone_0	date_avenant_0	Chgt_naiss_conducP_1	Chgt_Zone_1	date_avenant_1
Police1	-5	-1	10/03/2018	-3	0	11/12/2017
Police2	4	0	12/06/2016	0	0	-

Etape 2 : Jointure entre la base avenant et la base tarifaire

Après avoir restructuré la base *avenant*, on regroupe les données des deux bases : *tarifaire* et *avenant* dans une nouvelle base temporaire. Cette dernière gardera la même structure de la base *tarifaire* :

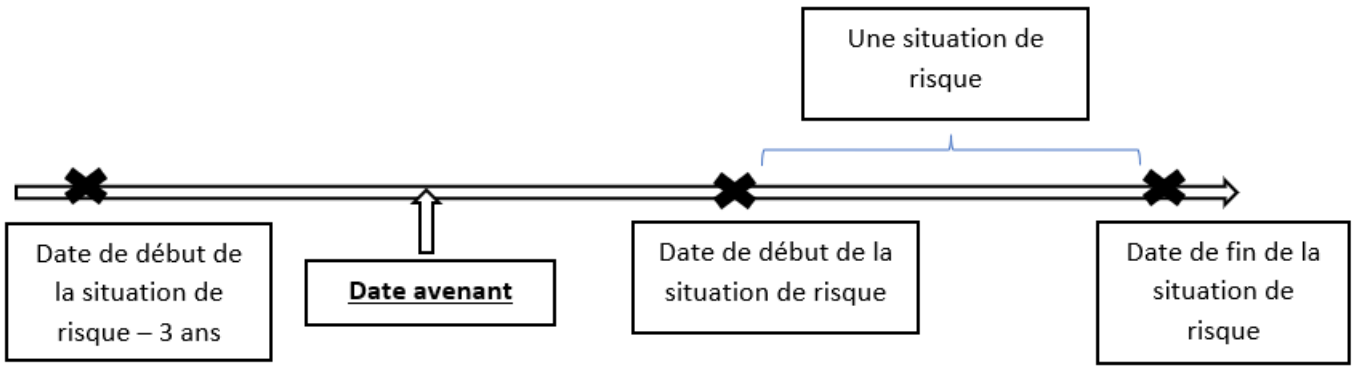
Le schéma simplifié ci-dessous explicite ce mécanisme de jointure.

POLICE	variable1	variable2	date_début	date_fin
police1				
police2				
police3				
police4				
police5				

POLICE	BLOC_0	BLOC_1
police1		
police2		

POLICE	variable1	variable2	date_début	date_fin	BLOC_0	BLOC_1
police1						
police2						
police3						
police4						
police5						

Cette jointure ne se fait pas de manière aléatoire. On rappelle que chaque ligne de la base tarifaire correspond à une situation de risque où toutes les variables sont stables. L'objectif est d'ajouter pour chaque situation de risque de l'information sur le comportement antérieur de l'assuré. Plus précisément, on cherche à détecter les avenants effectués pendant les 3 ans antérieurs à la date de début de la situation de risque. Le schéma ci-dessous explicite cette condition.



Etape 3 : Création des variables avenant

La troisième étape consiste à traiter la base de modélisation ainsi construite. Le but de ce traitement est d'avoir une structure plus adaptée aux modèles GLM. On crée ainsi les nouvelles variables explicatives liées aux avenants antérieurs à partir des différents blocs précédemment construits. La colonne de la variable *avenant* correspondra à la somme des colonnes présentes dans chaque bloc.

4.3 Construction de la base de modélisataion

Une fois la base de modélisation résultante de la jointure entre la base tarifaire et la base *avenant* est construite, le but est de créer un modèle de régression logistique pour expliquer la variable cible target_SNBSIN2GRAV. On cherche notamment à détecter le pouvoir explicatif des variables *avenant*. Pour ce faire, plusieurs pistes ont été explorées, on les détaillera dans la partie suivante.

4.3.1 Première piste et Résultats obtenus

On rappelle que l'objectif de l'étude est d'enrichir la base tarifaire avec de nouvelles variables liées au comportement antérieur de l'assuré pour être capable de mieux expliquer les sinistres graves.

On a commencé par créer un modèle de régression logistique en utilisant comme base d'entrée toute la base de modélisation (base tarifaire à laquelle on a ajouté les variables *avenant*).

Chaque point coloré du graphique ci-dessous correspond à une valeur du coefficient AUC et un nombre bien déterminé de variables. Ces deux valeurs définissent ensemble un modèle sélectionné par l'outil. Le graphique ci-dessous 4.2 résume ainsi l'ensemble des modèles obtenus.

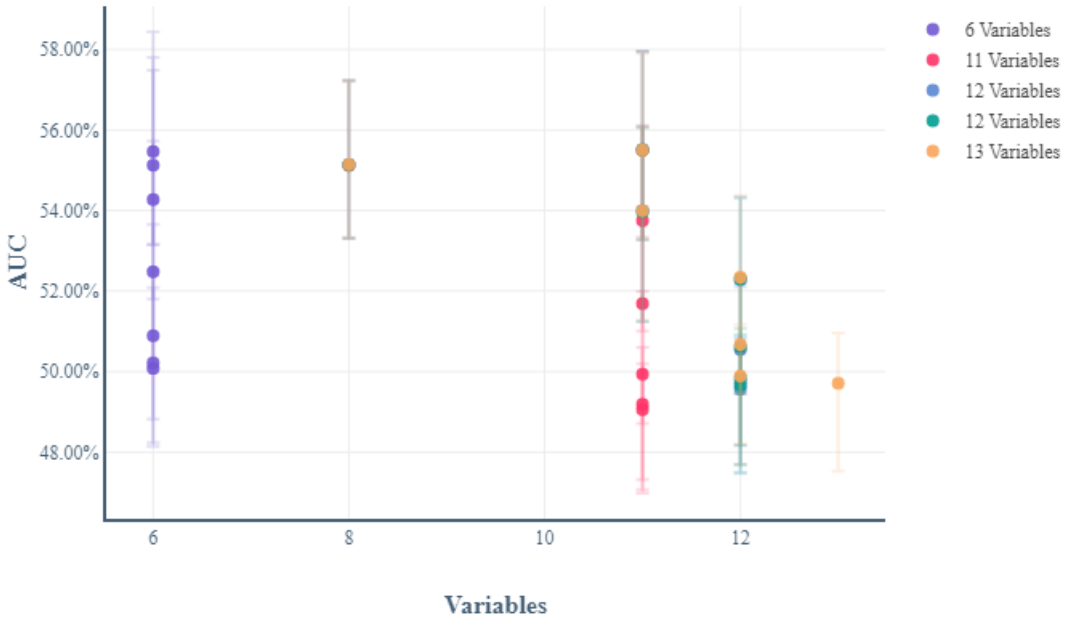


FIGURE 4.2 – Graphique représentatif de l'ensemble des modèles sélectionnés

L'ensemble des modèles obtenus possèdent un coefficient AUC assez faible majoritairement inférieur à 0.7. Ceci est dû principalement à une faible exposition de la variable cible vu la rareté du phénomène qu'on cherche à détecter. Les contrats ayant aucun sinistre grave sont plus présents dans la base que les contrats ayant des sinistres graves. Certes ce constat est très rassurant pour un assureur, mais pour arriver à construire un modèle robuste explicatif des sinistres graves, il faudra gérer ce déséquilibre présent dans la base.

On présentera dans la partie suivante les différentes méthodes de traitement des données déséquilibrées ainsi que les avantages et les inconvénients de chacune d'entre elles.

4.3.2 Solution proposée : Gestion des données déséquilibrées

La gestion des données déséquilibrés est l'un des enjeux majeurs que l'on peut rencontrer dans l'apprentissage automatique. En effet, on s'intéresse très souvent à des phénomènes certes rares, qui ne sont pas dominants dans la base de modélisation, mais qui peuvent avoir un impact très important.

Ce déséquilibre dans la distribution des données est couramment rencontré dans la détection des fraudes, détection d'anomalies, reconnaissance faciale, etc. C'est le cas également de la base de modélisation utilisée pour cette étude. En effet, La première piste précédemment exposée a permis de se rendre compte de l'impact la distribution inégales des classes de la

variable cible binaire sur les résultats de la modélisation : la classe des assurés ayant eu un sinistre grave est très défavorisée par rapport à celle qui n'ont pas eu de sinistre grave.

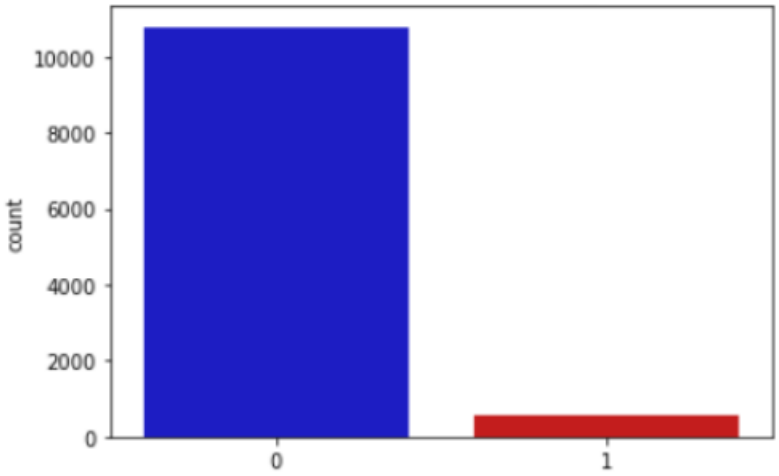


FIGURE 4.3 – Distribution de la variable cible target_SNBSIN2GRAV dans la base initiale

Le graphique ci-dessus (9.1) montre la répartition des classes de la variable cible binaire target_SNBSIN2GRAV. On remarque bien que les contrats ayant eu un sinistre grave sont très minoritaires dans la base de tous les contrats sinistrés présents dans notre portefeuille les années 2018, 2019 et 2020. Ce déséquilibre affecte négativement les performances des modèles. Ces derniers auront un biais en faveur de la classe majoritaire tout en ignorant la classe minoritaire. Pour remédier à ce problème, plusieurs techniques existent à savoir : Le sur-échantillonnage, Le sous-échantillonnage, La combinaison entre le sous-échantillonnage et le sur-échantillonnage.

Le sur-échantillonnage

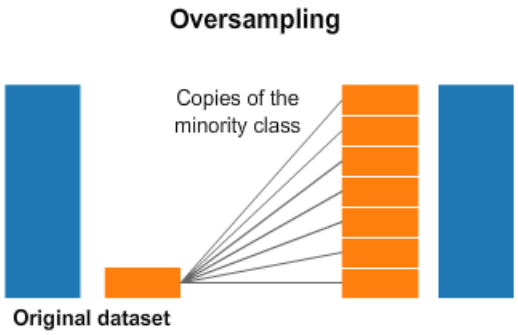


FIGURE 4.4 – Schéma explicatif de la méthode de sur-échantillonnage

Le sur-échantillonnage (4.4) consiste à augmenter la taille de la sous-population rare dans la base de données. Pour ce faire, deux approches peuvent être considérées :

Le sur-échantillonnage aléatoire (Random Oversampling)

Le sur-échantillonnage aléatoire augmente la taille de la classe minoritaire en dupliquant les lignes de la base originale. Certes, cette méthode est la stratégie la plus simple à implémenter mais elle favorise le risque de surapprentissage. On utilisera plutôt la technique SMOTE.

La technique de sur-échantillonnage synthétique minoritaire (SMOTE)

A l'instar de la précédente méthode, le sur-échantillonnage à l'aide de SMOTE (4.5) augmente la taille de l'ensemble de données d'apprentissage. Cependant, ces nouvelles lignes créées ne résultent pas d'une duplication des lignes de la base d'origine. Cette approche vise à générer de nouvelles instances de la classe minoritaire sans créer des lignes qui soient identiquement égales à celles déjà présentes. La première étape consiste à repérer un individu de la classe sous-représentée que l'on cherche à augmenter. On cherche ensuite ses k-plus proches voisins. On crée ainsi les nouvelles valeurs en appliquant un coefficient généré aléatoirement à ses k-plus proches voisins déjà sélectionnés.

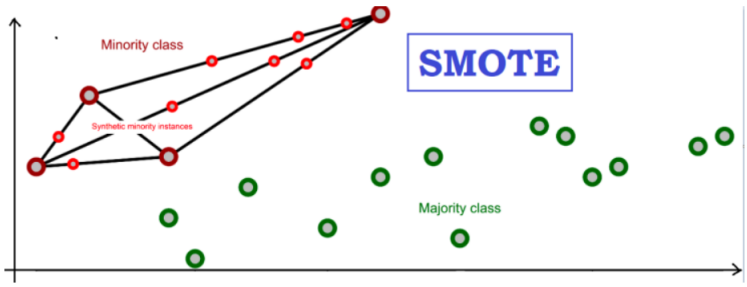


FIGURE 4.5 – Schéma explicatif de la technique SMOTE

Le sous-échantillonnage

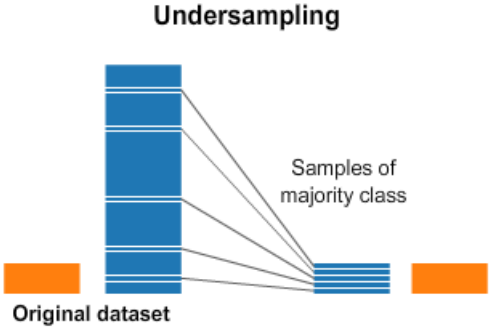


FIGURE 4.6 – Schéma explicatif de la méthode de sous-échantillonnage

Contrairement à l'approche précédente, la méthode de sous-échantillonnage (4.6) consiste à réduire le nombre d'échantillons de la classe majoritaire pour qu'il corresponde au nombre d'échantillons de la classe minoritaire. Pour ce faire, deux approches peuvent être considérées :

Le sous-échantillonnage aléatoire : Random Undersampling

Cette approche s'appuie sur le même principe que la méthode de Random Oversampling. L'algorithme sélectionne aléatoirement des lignes de la classe majoritaire et les élimine de la base. Cette méthode est simple à implémenter. Cependant, elle présente un fort risque de perte des données essentielles.

Le sous-échantillonnage en utilisant l'algorithme Edited Nearest Neighbor (ENN)

L'algorithme ENN permet de réduire la taille de la base majoritaire en utilisant l'algorithme des plus proches voisins. On commence par trouver le K-plus proche voisin de chaque observation. On vérifie ensuite si la classe majoritaire du K-plus proche voisin de l'observation est la même que la classe de l'observation ou non. Si la classe majoritaire du K-plus proche voisin de l'observation et la classe de l'observation sont différentes, alors l'observation et son K-plus proche voisin sont supprimés de l'ensemble de données. Par défaut, le nombre de plus proches voisins utilisés dans ENN est K=3.

Il est également possible de combiner des techniques de sur-échantillonnage et de sous-échantillonnage pour avoir des résultats plus robustes. Les deux algorithmes couramment utilisés sont l'algorithme SMOTE pour le sur-échantillonnage l'algorithme Edited Nearest Neighbours (ENN) pour le sous-échantillonnage.

4.4 Analyse descriptive

La méthode qui sera retenue pour la suite pour gérer le déséquilibre des données est la combinaison des deux stratégies récemment décrites. Cette partie sera consacrée à une première analyse descriptive de la base de modélisation.
Le graphique ci-dessous (4.7) montre la répartition de la variable cible target_SNBSIN2GRAV dans la nouvelle base ainsi construite. les classes de cette variable sont mieux équilibrées qu'avant.

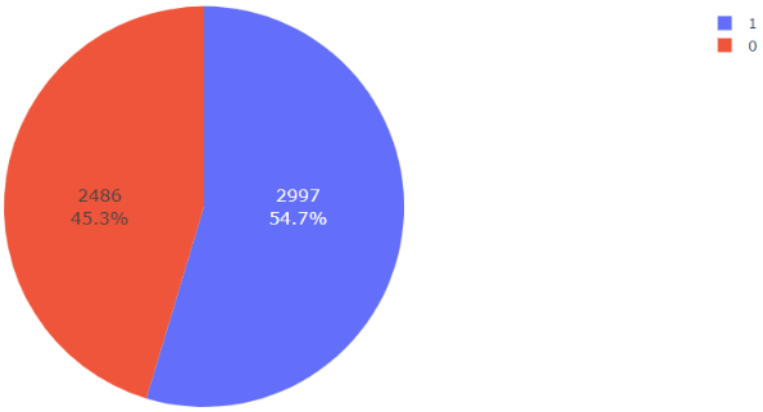


FIGURE 4.7 – Distribution des sinistres graves dans la base de modélisation

Afin d'assurer une bonne compréhension de notre portefeuille, on a tracé les distributions des variables explicatives suivantes (4.8) (4.9) : *Age_conducteur*, *Age_vehicule*, *Ancienete_contrat*, *Anciennete_permis*. Le profil moyen d'un assuré de notre portefeuille possède les caractéristiques suivantes :

Age_conducteur : 50 ans
Anciennete_permis : 30 ans
Age_obtention_permis : 20 ans
Age_conducteur_secondaire : 41 ans
Anciennete_permis_secondaire : 20 ans
Age_vehicule : 9 ans
Age_acquisition_vehicule : 3 ans
Anciennete_contrat : 4 ans
Coef_BonusMalus : 1

Les valeurs ci-dessus montre qu'on a plutôt un portefeuille d'assurés âgés ayant une bonne expérience de conduite. Ce qui explique grossièrement en partie la sous-représentation de la classe des sinistres graves dans notre base initiale.

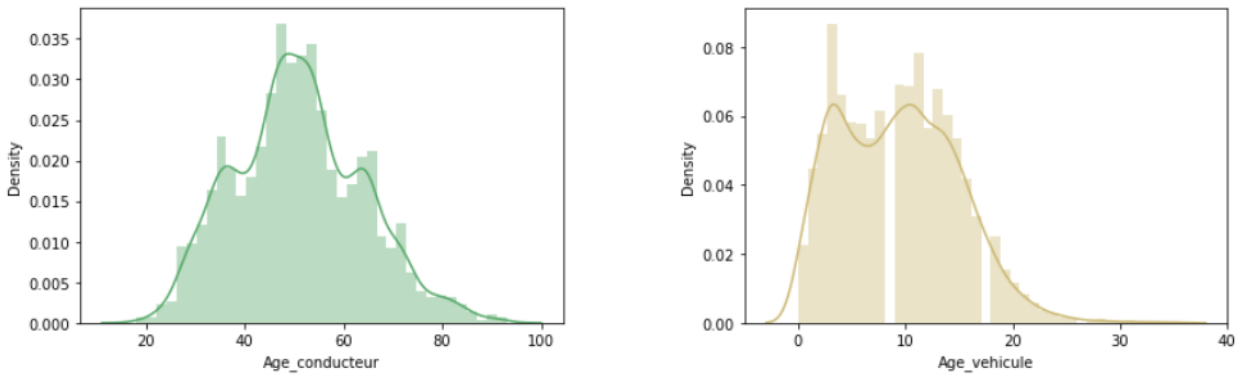


FIGURE 4.8 – Distribution des variables Age_conducteur et Age_vehicule

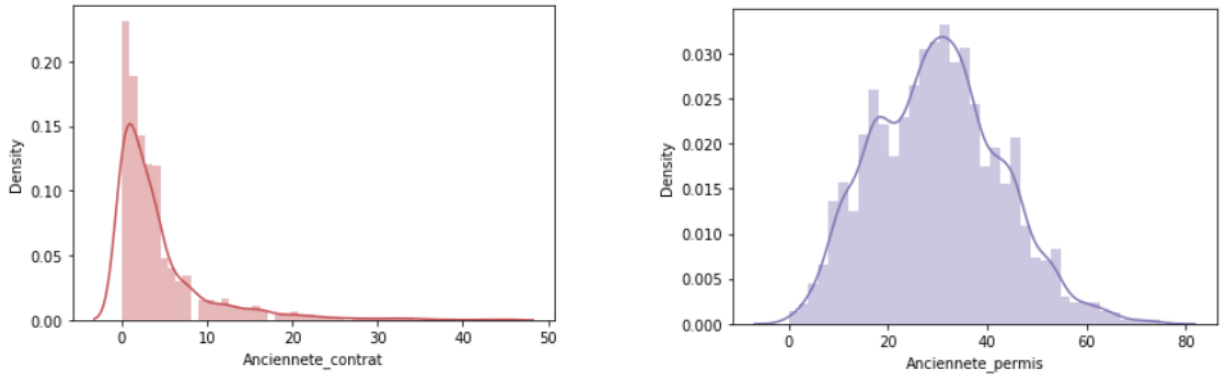


FIGURE 4.9 – Distribution des variables Anciennete_contrat et Anciennete_permis

On s'intéresse ensuite à une première étude de la relation entre les variables *avenant* construites et la variable cible target_SNBSIN2GRAV. On a tracé les histogramme de répartition des variables chgt_garage, chgt_naiss_conducP, chgt_date_perm et chgt_groupe en fonction de l'occurrence ou non d'un sinistre grave.

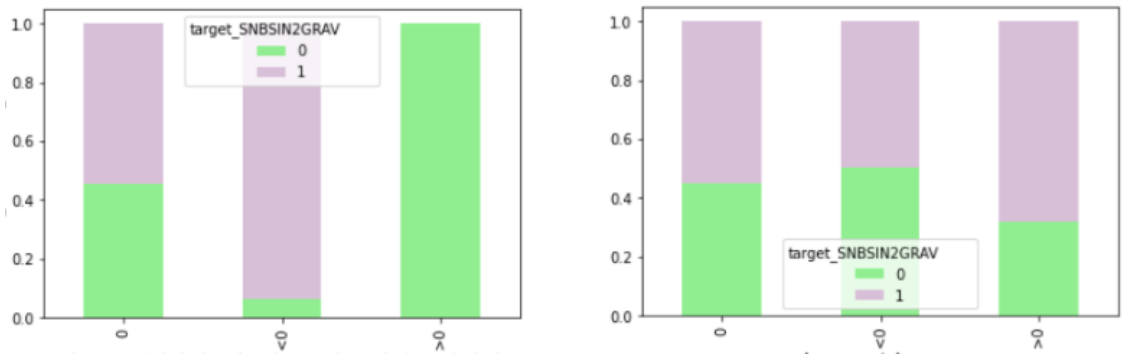


FIGURE 4.10 – Histogrammes de répartition des variables chgt_date_naiss_conducP (à gauche) et chgt_Zone (à droite) en fonction de la variable cible target_SNBSIN2GRAV

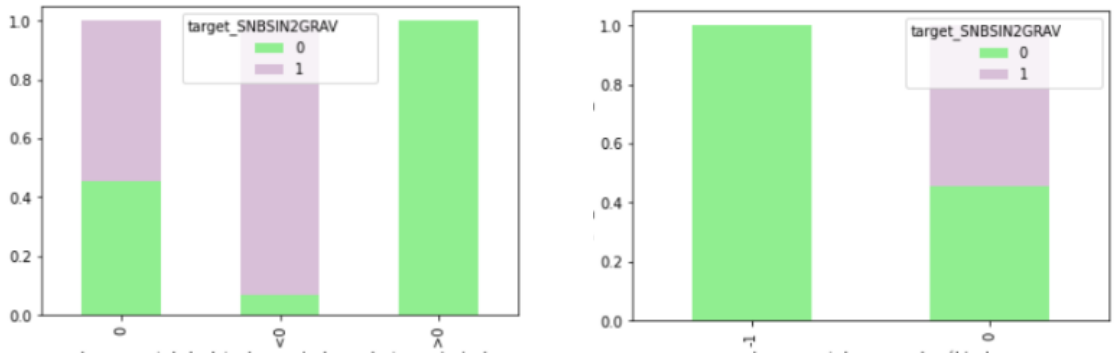


FIGURE 4.11 – Histogrammes de répartition des variables chgt_date_perm (à gauche) et chgt_groupe (à droite) en fonction de la variable cible target_SNBSIN2GRAV

D'après les figures ci-dessus (4.10) (4.11), le changement du groupe de véhicule n'apparait pas comme une éventuelle variable explicative de la sinistralité grave : Tous les assurés

ayant changé de groupe de véhicule n'ont pas de sinistres graves.

Au contraire, l'occurrence ou non d'un sinistre grave varie en fonction des variables décrivant le changement de zone, le changement du conducteur principal (caractérisé par le changement de la date de naissance et la date d'obtention du permis). Les conducteurs les plus jeunes ont plus de probabilité d'avoir un sinistre grave. Par ailleurs, un déménagement vers une zone plus risquée engendre une augmentation de l'occurrence des sinistres graves.

5

Résultats : Modélisation de la fréquence des sinistres graves

Après avoir construit la base de modélisation et traité le problème des données déséquilibrées, on s'intéresse dans cette partie à la modélisation des sinistres graves. On rappelle que notre objectif est de prédire pour les assurés sinistrés si oui ou non ils vont avoir un sinistre grave à partir d'un panel de variables explicatives : tarifaires et *avenant*. Notre cible est la variable target_SNBSIN2GRAV. Cette dernière est binaire. Elle vaut 1 dans le cas d'un sinistre grave et 0 sinon

Le premier modèle qu'on va tester est un modèle GLM plus précisément un modèle de régression logistique.

On commence par sélectionner uniquement les variables tarifaires dans notre modèle de propension. L'outil propose un certain nombre de modèles en fonction du nombre de variables retenues. On choisit ainsi le modèle le plus pertinent et le plus robuste qu'on enrichira à l'aide des variables *avenant* déjà incluses dans la base tarifaire. Le but sera, par la suite, d'étudier l'impact de ces nouvelles variables ainsi que la qualité du modèle.

Les résultats obtenus seront détaillés ci-dessous.

5.1 Modèle de propension retenu

5.1.1 Sélection du modèle

Comme on l'a explicité avant, l'outil utilisé nous propose un panel de modèles en fonction du nombre de variables retenues. Le but étant de choisir le modèle le plus pertinent en termes

d'indicateurs (RMSE, GINI, etc) et qui décrit au mieux la variable cible.

Le paramètre de lissage est un paramètre de la performance du modèle. Il décrit la sensibilité du modèle aux signaux faibles. En effet, un modèle très lissé (coefficient de lissage fort) suivra uniquement la tendance globale alors qu'un modèle peu lissé capturera les variations les plus faibles. Cependant, choisir un modèle très peu lissé peut nuire au niveau de robustesse globale.

Le coefficient AUC (Area Under the Curve), quant à lui, varie entre 0 et 1. On parle d'un très bon modèle si cette valeur dépasse 0.7 et d'un excellent modèle si cette valeur est supérieure à 0.9.

En plus du degré de lissage et du coefficient AUC, on cherche à choisir le modèle qui minimise l'erreur RMSE (Root Mean Square Error) pour avoir les prédictions les plus fiables possibles.

Le modèle retenu doit donc satisfaire au mieux les différentes conditions mentionnées ci-dessus.

Pour remédier à la problématique des données déséquilibrées, on a essayé les trois approches classiques décrites ci-dessus : Le sur-échantillonnage, le sous-échantillonnage ainsi qu'une combinaison des deux. On présentera le panel des modèles proposés par l'outil dans les trois cas ainsi que le modèle sélectionné pour chaque approche.

Dans les trois cas, on applique un modèle de régression logistique avec la variable target binaire target_SNBSIN2GRAV. On sélectionne, dans un premier temps, uniquement les variables tarifaires. On choisit ensuite le modèle le plus robuste parmi ceux suggérés qu'on enrichira, dans un deuxième temps, avec les variables *avenant*.

Méthode de sur-échantillonnage

Le premier graphique ci-dessous (5.1) représente l'ensemble des modèles proposés. Pour un nombre fixe de variables, on a un ensemble de modèles possibles qui diffèrent suivant le coefficient AUC.
On a choisi le modèle avec 14 variables ayant un AUC égale à 0.78.

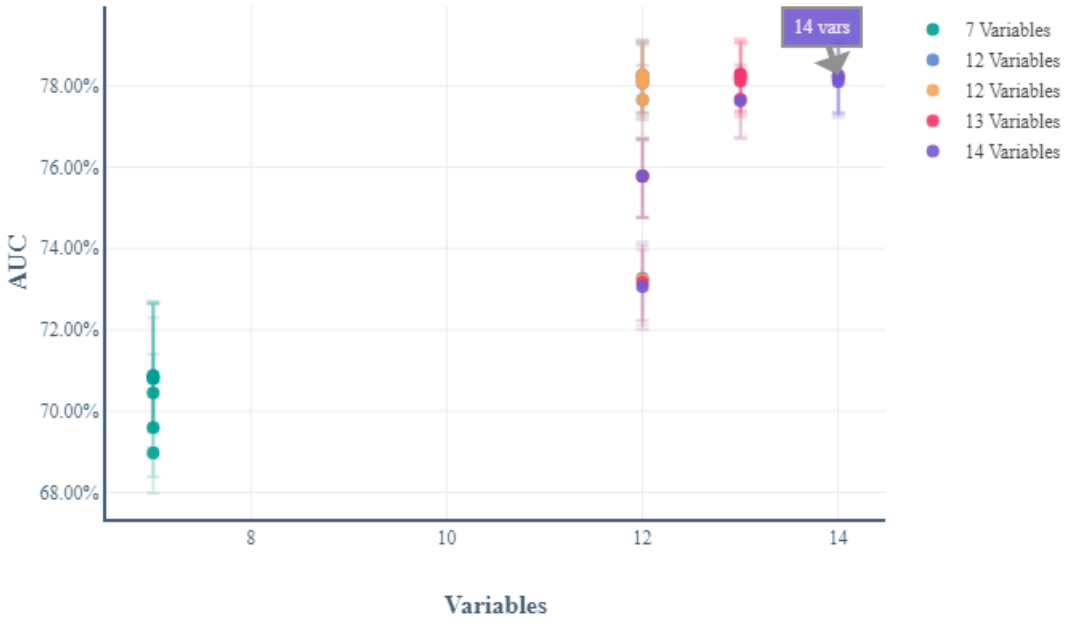


FIGURE 5.1 – Graphique représentatif de l'ensemble des modèles sélectionnés en utilisant l'approche de sur-échantillonnage

On remarque ainsi que le méthode de sur-échantillonnage en utlisant l'algorithme SMOTE permet d'obtenir de bons résultats.

Méthode de sous-échantillonnage

La deuxième méthode testée pour traiter le déséquilibre des données est la méthode de sous-échantillonnage. L'inconvénient majeur de cette approche est la perte importante d'informations qu'elle peut engendrer. Le graphique ci-dessous (5.2) résume l'ensemble des modèles proposés par l'outil.

On remarque une baisse importante du coefficient AUC. Il ne dépasse pas 0.56 pour l'ensemble des modèles sélectionnés. Cette valeur est considérée très faible. Cette approche ne sera donc pas retenue pour la suite de l'étude.

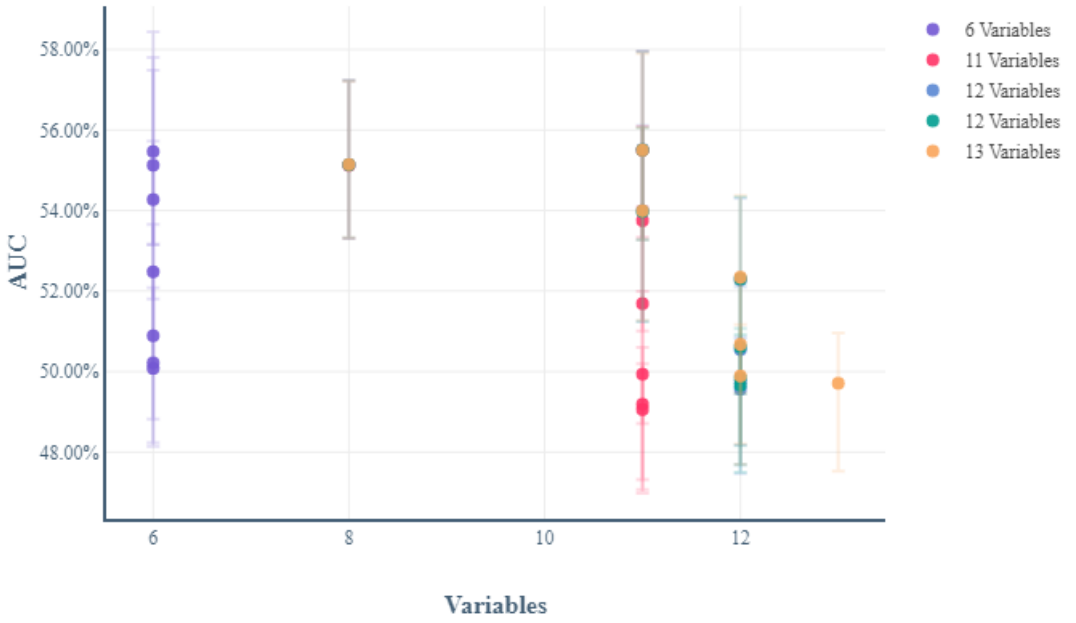


FIGURE 5.2 – Graphique représentatif de l'ensemble des modèles sélectionnés en utilisant l'approche de sous-échantillonnage

Méthode combinant le suréchantillonnage et le souséchantillonnage

Après avoir exploré les résultats obtenus avec la méthode de sur-échantillonnage et de sous-échantillonnage, on s'intéresse dans cette partie à une approche qui combine les deux précédentes pour traiter le déséquilibre des données de la base de modélisation : On augmente la taille de la classe minoritaire tout en diminuant la taille de la classe majoritaire.

Le graphique ci-dessous (5.3) résume l'ensemble des modèles proposés en fonction du nombre de variables sélectionnées et la valeur du coefficient AUC. On note une amélioration des performances des modèles notamment une augmentation du coefficient AUC par rapport à la méthode de sur-échantillonnage. On dépasse 0.78 pour la valeur de l'AUC.

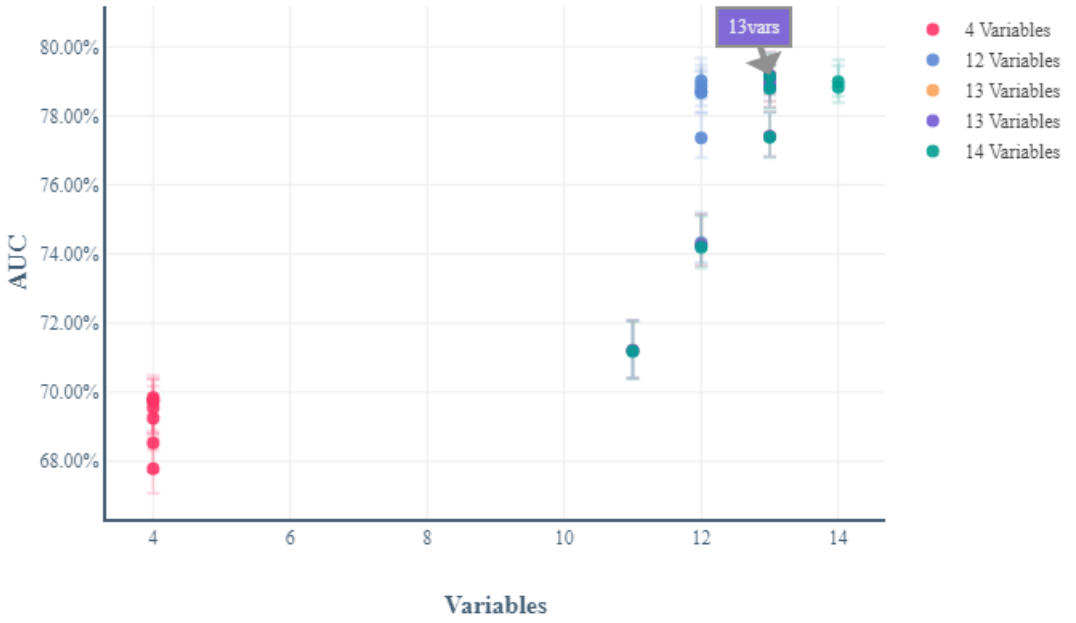


FIGURE 5.3 – Graphique représentatif de l'ensemble des modèles sélectionnés en utilisant l'approche combinant le sur-échantillonnage et le sous-échantillonnage

Cette approche sera donc retenue pour la suite de l'étude.

On s'intéresse dans la partie suivante à une étude plus détaillée de la méthode de sélection du modèle ainsi que les performances de ce dernier.

5.1.2 Analyse du modèle retenu

Sélection du modèle

Après avoir sélectionner l'approche à utiliser pour gérer le problème de déséquilibre des données : une combinaison entre la méthode de sous-échantillonnage et la méthode de sur-échantillonnage. On va étudier dans cette partie le modèle retenu.

Comme on l'a précisé avant, on commence par entrainer la base d'apprentissage sur l'ensemble de la base tarifaire. On sélectionne ensuite le modèle le plus performant parmi ceux suggérés par l'outil et on l'enrichit ensuite avec les variables *avenant*.

Le tableau ci-dessous résume les métriques du modèle retenu pour cette approche.

	modèle sans variables *avenant*	modèle avec variables *avenant*
AUC	0.70	0 .78
RMSE	0.43	0.39

Ces résultats montrent une amélioration du pouvoir explicatif du modèle après ajout des variables avenants due à l'augmentation du coefficient AUC. De plus, la RMSE a légèrement diminué. On étudiera dans la suite les autres indicateurs de performance du modèle.

La courbe Lift

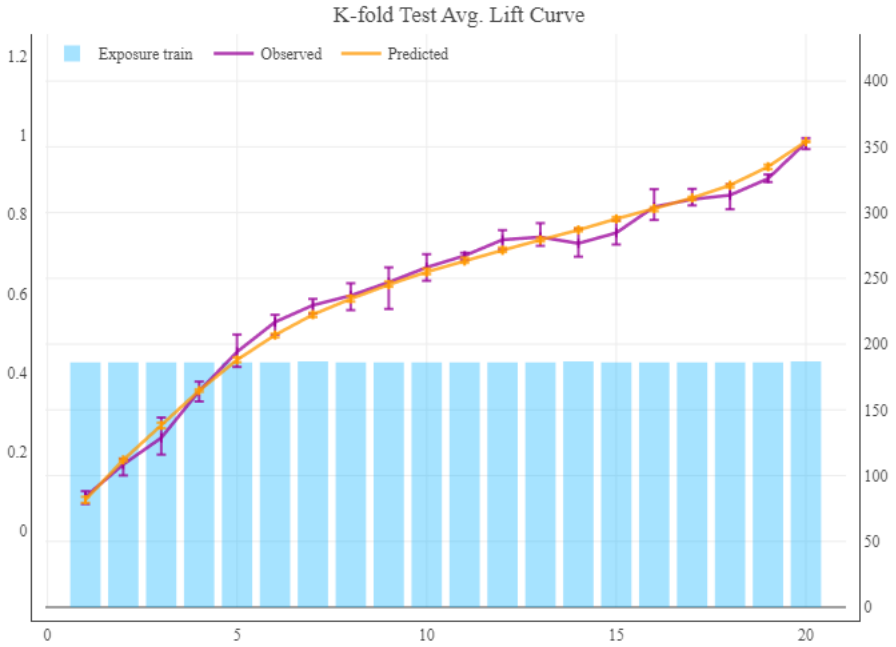


FIGURE 5.4 – La courbe Lift obtenue pour le modèle sélectionné

Globalement, la courbe (5.4) des valeurs observées ainsi que celle des valeurs prédites ont la même tendance et se suivent. Cependant, on peut remarquer quelques légères sur-estimations et sous-estimations. Mais le modèle demeure valide et ajuste bien les données.

Spread

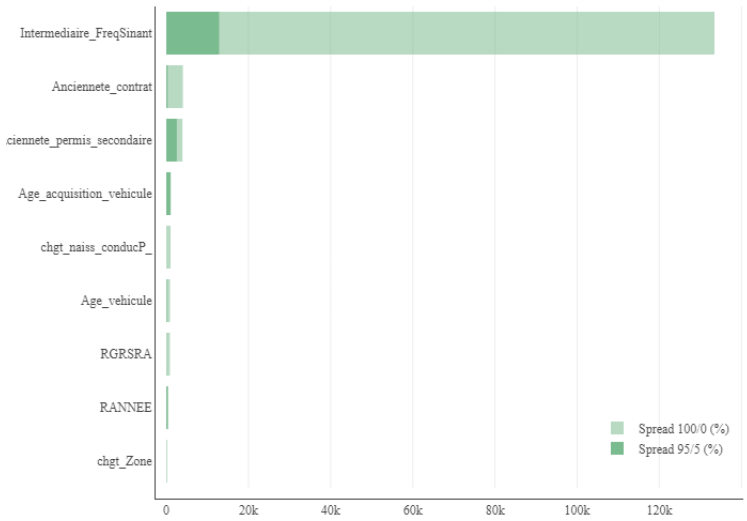


FIGURE 5.5 – Classement des variables suivant le spread 100/0

Le graphique ci-dessus (5.5) représente le classement des variables retenues par le modèle suivant la valeur du spread 100/0.On remarque que seulement deux variables ont été retenues parmi les quatre variables *avenant* sélectionnées dans le modèle : changement de la date de naissance du conducteur principal et le changement de zone.

Les résidus quantiles

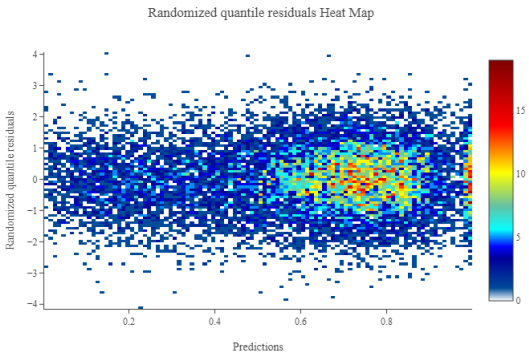


FIGURE 5.6 – Représenation graphique des résidus du modèle

Le graphique des résidus (5.6) montre une symétrie autour de la valeur 0. De plus, la majorité des résidus se situe entre les valeurs 2 et -2. On observe quelques dépassements

de ces deux valeurs. Mais, le modèle est globalement validé. Le modèle retenu semble donc être un modèle pertinent. Il sera retenu pour la suite de l'étude. La partie suivante sera consacrée à la présentation des résultats obtenus par la validation croisée.

Résultats de la validation croisée

la base de modélisation a été initialement découpée en deux parties. 80% de cette base servira pour entraîner le modèle et 20% costituera la base de test.

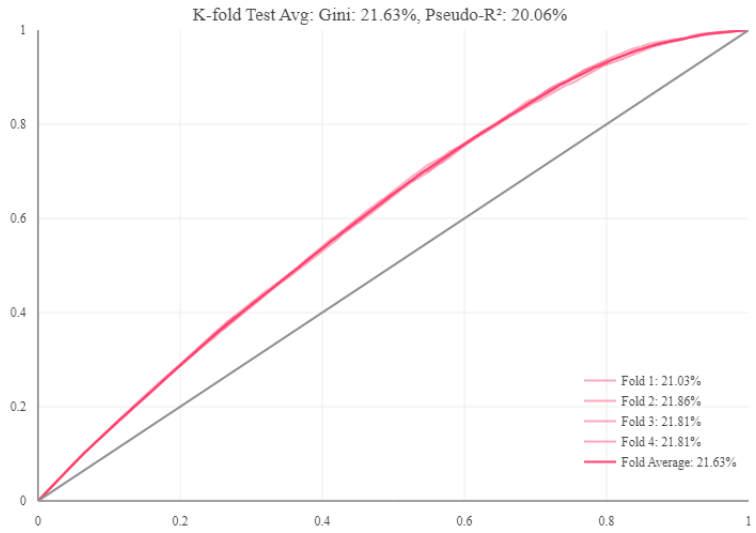


FIGURE 5.7 – La courbe de Lorenz

La courbe de lorenz (5.7) affiche des coefficients de GINI assez proches pour les 4 folds de la base d'apprentissage. Le coefficient de GINI est de l'ordre de 21%. On peut ainsi affirmer que cet indicateur est stable sur l'ensemble de la base d'apprentissage.
Afin de s'assurer que le modèle obtenu est généralisable à d'autres bases de données, on a effectué en dehors de l'outil la procédure de validation croisée sur la base test. Les résultats obtenus sont récapitulés dans le tableau ci-dessous et permettent de confirmer l'absence de sur-apprentissage.

	train K-fold	**test k-fold**
F1	0.81	0.77
F2	0.80	0.78
F3	0.80	0.76
F4	0.79	0.75
Moyenne	0.80	0.77

5.2 Impact des Variables *avenant*

5.2.1 Changement de zone

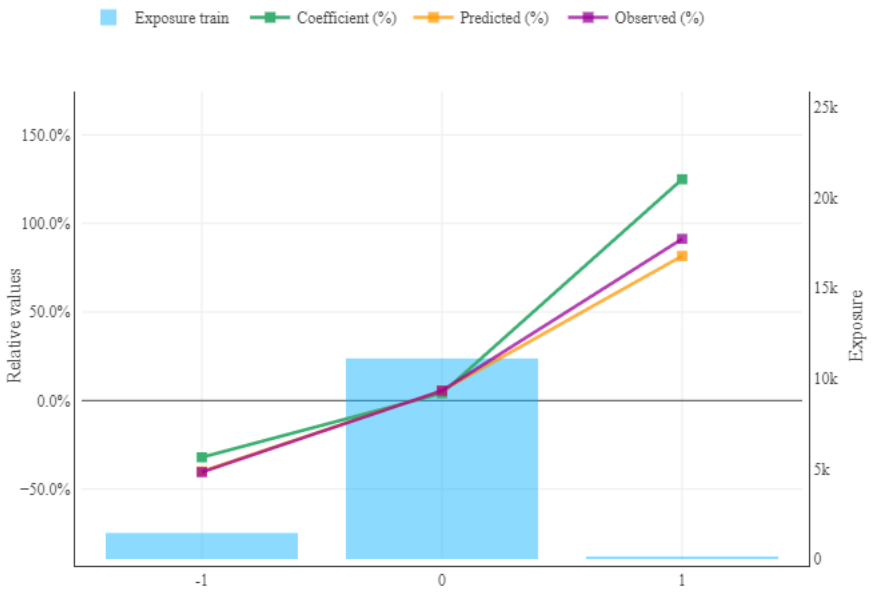


FIGURE 5.8 – Représentation graphique des coefficients de la variable chgt_zone

Le graphique ci-dessus montre les différents coefficients normalisés obtenus pour la variable changement de zone. La modalité « 1 » correspond à un déménagement d'une petite ville à une grande ville. Alors que la modalité « -1 » correspond à un déménagement dans le sens inverse. D'après les coefficients obtenus, on s'aperçoit que le risque augmente dans le cas d'un déménagement à une grande ville. Ce résultat est assez cohérent avec ce qu'on observe. Les grandes villes sont plus peuplées et le risque d'avoir des sinistres graves peut s'accroitre dans ces zones géographiques.

5.2.2 Changement de la date de naissance du conducteur principal

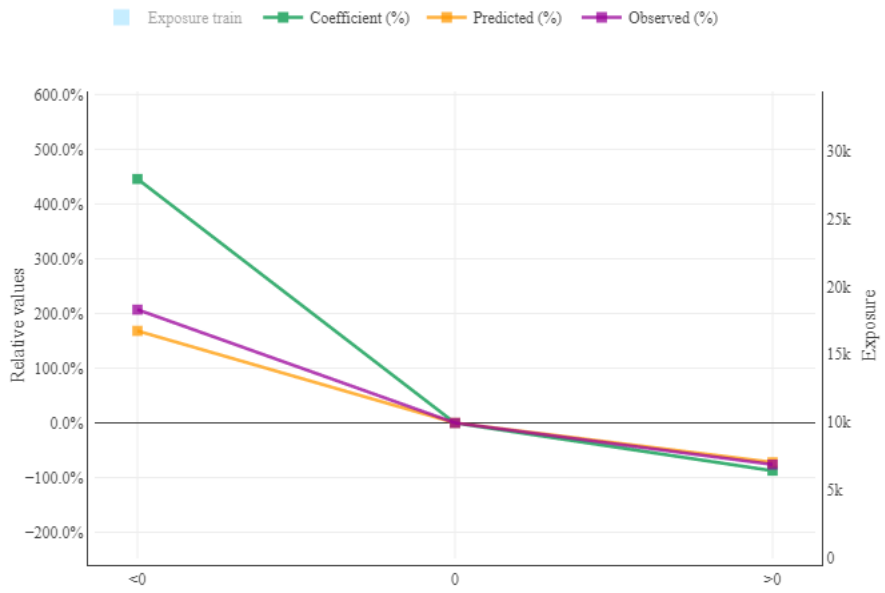


FIGURE 5.9 – Représentation graphique des coefficients de la variable chgt_naiss_conducP

Le graphique ci-dessus montre les différents coefficients normalisés obtenus pour la variable changement de la date de naissance du conducteur principal. La modalité « >0 » signifie que le nouveau conducteur principal est plus âgé que l'ancien. Alors que la modalité « <0 » correspond à un passage à un conducteur plus jeune. D'après les coefficients obtenus, on s'aperçoit qu'un conducteur jeune présente un risque plus fort d'avoir un sinistre grave. Au contraire, le passage à un conducteur âgé diminue ce risque.

Les différents indicateurs détaillés ci-dessus ainsi que la validation croisée effectuée sur la base test permettent de valider le modèle retenu. Certes l'ajout des variables avenant a amélioré le pouvoir explicatif du modèle. Cependant, la valeur du GINI reste faible en la comparant à celle d'un modèle tarifaire classique où on prédit les sinistres ordinaires. Ceci pourrait être expliqué par un problème de volumétrie des variables explicatives notamment les variables avenant. Ceci a fortement joué en défaveur des performances du modèle. On testera dans la suite un deuxième modèle couramment utilisé pour résoudre les problèmes de classification binaire : XGBoost.

5.3 Résultats obtenus avec un modèle XGboost

L'algorithme XGBoost est souvent utilisé pour résoudre des problématiques de classification. Il représente l'un des algorithmes les plus optimisés en terme de performances informatiques. On a implémenté cet algorithme avec notre variable cible target_SNBSIN2GRAV

en utilisant la même base de modélisation utilisée pour le modèle GLM (c'est-à-dire la base ré-échantillonnée). On essaye donc de prédire notre variable cible avec le panel des variables explicatives disponibles regroupant les variables tarifaires ainsi que les variables *avenant* construites.

Le tableau ci-dessous récapitule les différents indicateurs de performance du modèle obtenu :

Indicateurs	**Valeurs**
Accuracy	0.89
Recall	0.96
Precision	0.86
ROC_AUC	0.80
RMSE	0.32

L'accuracy, le recall ainsi que la précision permettent d'évaluer le pouvoir prédictif du modèle.
Ici, l'accuracy vaut 0.89. Cette valeur signifie que 89% des prédictions sont de bonnes prédictions. C'est le pourcentage de la somme des vrais positifs et des vrais négatifs prédits.
Pour compléter l'accuracy, on calcule également le recall et la précision.
D'une part, le recall se concentre sur l'ensemble des assurés qui ont réellement eu un sinistre grave et calcule le pourcentage de ceux qu'on arrive à détecter avec le modèle.
Plus précisément, le recall est donné par la formule suivante :

$$Recall = \frac{vrais\_positifs}{vrais\_positifs + faux\_ngatifs}$$

Avec,

- *vrais_positifs* correspondent aux assurés pour qui on a prédit un sinistre grave parmi ceux qui ont réellement eu des sinistres graves
- *vrais_négatifs* correspondent aux assurés pour qui on n'a pas prédit un sinistre grave parmi ceux qui n'ont pas eu de sinistre grave
- *faux_négatifs* correspondent aux assurés pour qui on n'a pas prédit un sinistre grave parmi ceux qui ont eu des sinistres graves

D'autre part, la précision se concentre sur l'ensemble des assurés pour qui on a prédit l'occurence d'un sinistre grave. On calcule ainsi le taux des prédictions correctes suivant la formule suivante :

$$Precision = \frac{vrais\_positifs}{vrais\_positifs + Faux\_positifs}$$

D'après les valeurs obtenues pour le recall, la précision ainsi que l'accuracy, on peut affirmer que, globalement, on obtient de bonnes prédicitons.

Tout comme les modèles GLM, il est intéressant également d'examiner la valeur du coefficient AUC calculé à partir de la courbe de ROC. On obtient une valeur d'AUC égale 0.8 qui légèrement supérieure à celle obtenue dans le cas d'un modèle GLM. L'erreur quadratique a également baissé. Elle est égale à 0.32 pour le modèle XGboost.

La validation du modèle nécessite également d'appliquer la méthode de validation croisée afin de s'assurer que le modèle pourrait être généralisé à d'autres bases de données. La base d'apprentissage a été découpée en deux : 80% représente la base d'apprentissage et 20% représente la base test. On a effectué ensuite la validation croisée sur 5 échantillons. Les résultats obtenus sont résumés dans le tableau ci-dessous.

train-auc-mean	**test-auc-mean**
0.63	0.62
0.70	0.68
0.75	0.73
0.78	0.76
0.80	0.77

Les valeurs obtenues pour la base test et la base d'apprentissage sont assez proche. On peut donc éliminer l'hypothèse de sur-apprentissage.

Test de robustesse

Après avoir testé la validation croisée sur les 5 échantillons de la base de test et d'apprentissage, on tire 100 échantillons aléatoires de la base d'apprentissage ainsi que la base de test et on mesure l'AUC correspondant à chaque échantillon.
Les résultats obtenus sont résumés ci-dessous :

quantiles	**AUC train**	**AUC test**
2.5	0.681	0.703
25	0.732	0.712
50	0.764	0.753
97.5	0.821	0.780

Ces résultats permettent de déterminer un intervalle de confiance pour l'AUC de la base d'apprentissage à 95% allant de 0.681 à 0.821

Après avoir exploré certaines métriques de performance du modèle, on s'intéresse aux résultats obtenus avec le modèle XGBoost. Le graphique ci-dessous montre les variables les plus importantes retenues par le modèle XGBoost.

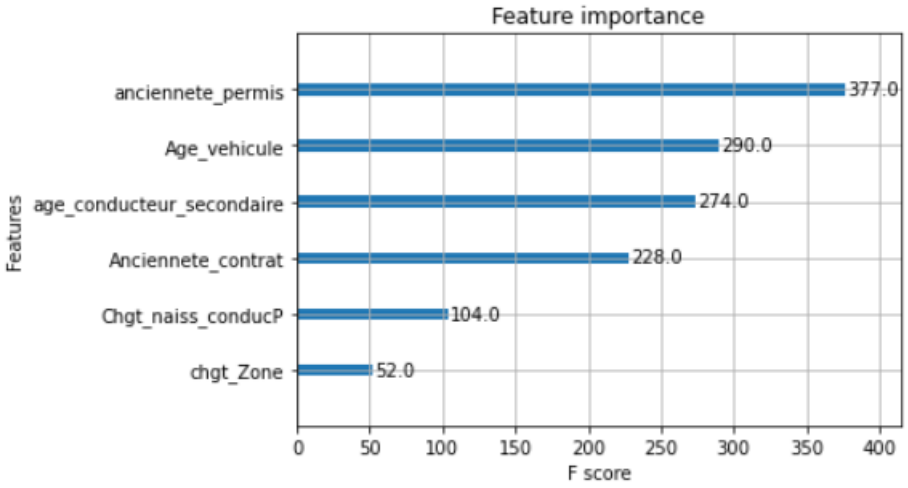


FIGURE 5.10 – Classement des variables explicatives suivant leurs importances

Deux variables uniquement parmi les variables *avenant* construites ont été retenues : *chgt_ Zone* et *chgt_ naiss_ conducP*. Les variables les plus importantes selon le modèle demeurent, cependant, des variables tarifaires : *Age_ conducteur_ secondaire*, *Age_ vehicule*, *Anciennete_ permis* et *Anciennete_ contrat*.

Les résultats obtenus diffèrent de ceux qu'on avait avec le modèle GLM. Cependant, les variables retenues par les deux modèles pour expliquer la sinistralité grave sont les mêmes. L'ordre d'importance de ces variables diffèrent.

On rappelle que le but de cette étude était de chercher de nouvelles variables, principalement des variables liées aux avenants antérieurs, qui pourraientt expliquer les sinistres graves. Les deux modèles confirment la pertinence des deux variables chgt_zone et chgt_naiss_conducP. Il parait ainsi pertinent de les ajouter aux critères de surveillance du portefeuille Automobile. Détecter un changement de zone ou bien du conducteur principal pourrait être un signe d'alerte. L'assuré devrait ainsi être mis sous surveillance.

6

Cas d'application

Après avoir déterminé deux variables *avenant* pertinentes pour décrire les sinistres graves. On s'intéresse dans cette partie à une application pratique de ces résultats notamment pour le processus de surveillance. Le but est d'arriver à définir une nouvelle règle de mise sous-surveillance qui, en plus des critères actuels, permettra d'avoir un meilleur contrôle du portefeuille.
On commencera par une présentation générale du processus de surveillance à Generali ainsi que des différents critères utilisés par l'équipe de surveillance. Ensuite, on proposera une nouvelle règle de mise sous surveillance qui sera comparée à celles actuellement utilisées.

6.1 Présentation du système de surveillance

L'activité de surveillance s'inscrit dans un cadre de contrôle et de meilleure gestion du portefeuille de l'assureur. Elle permet de maintenir la rentabilité du portefeuille voire de l'améliorer. Ce processus est basé sur un panel de critères qui sont propres à chaque branches d'activité permettant ainsi l'homogénéisation des risques au sein du portefeuille du produit concerné, un contrôle sur les différents comportements frauduleux ainsi que le maintien de la rentabilité et du pouvoir compétitif de l'assureur sur le marché.

Ce processus se base sur la définition d'un certain nombre de critères qui seront utilisés pour segmenter les assureurs et détecter ceux présentant ou ayant un potentiel d'avoir un comportement anormal. Ces critères diffèrent d'un assureur à un autre et sont propres aux caractéristiques du portefeuille. Ils se reposent, d'une part, sur les caractéristiques des sinistres antérieurs : la fréquence, le coût et la nature du sinistre, et, d'autre part, les caractéristiques liées au conducteur.

Les critères de surveillance du portefeuile Auto chez Generali

Pour des raisons de confidentialité, les critères de mise sous surveillance utilisés à Generali ne seront pas détaillés dans ce paragraphe. On présentera cependant quelques exemples de règles utilisées. Les critères de surveillance actuellement exploités par l'équipe de surveillance se basent sur la fréquence, la nature du sinistre ainsi que quelques caractéristiques du conducteur principal.

Le tableau ci-dessous résume quelques exemples :

Nature du sinistre	Nombre de sinistres	Période d'observation
Corporel responsable	1	-
Vol	2	24 derniers mois

Ces différents critères ont pour objectif de détecter les contrats susceptibles d'avoir une sinistralité anormale (fréquence élevée ou bien coût élevé) afin de pourvoir prendre des actions à l'avance. Trois cas de figures sont envisagés par l'équipe surveillance de Generali :

- La résiliation du contrat
- La modification du contrat : une augmentation de la prime, une majoration des franchises
- Le classement du contrat : Ceci correspond à une mise sous surveillance du contrat sans prendre de décision particulière c'est à dire tout en le gardant dans le portefeuille

6.2 Insuffisance des critères de surveillance actuels

Après avoir présenté brièvement les critères utilisés pour le processus de mise sous surveillance à Generali, on s'intéresse dans cette partie à l'étude de l'efficacité de ces différents critères.
Pour ce faire, la base de modélisation a été enrichie avec la base de contrôle de l'équipe de surveillance. Ainsi, pour chaque contrat, s'il a été mis sous surveillance, on récupère le critère utilisé.
Le graphique ci-dessous montre la répartition des critères de mise sous-surveillance dans la base des contrats ayant un sinistre grave (dont le montant dépasse le seuil retenu).

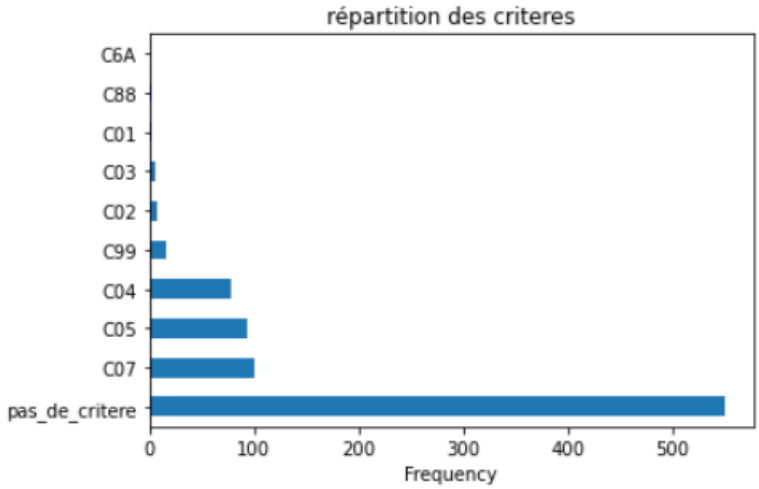


FIGURE 6.1 – Distribution des critères de surveillance dans la base des contrats ayant des sinistres graves

La répartition des critères utilisés actuellement pour la mise sous surveillance montre que la majorité des contrats présentant des sinistres graves parmi ceux présents dans notre base de modélisation ne sont pas mis sous surveillance. Ceci montre que les critères actuels nécessitent une amélioration pour qu'ils s'adaptent mieux aux caractéristiques actuelles du portefeuille Generali.
Afin de mieux mesurer l'impact des variables avenant qui ont été retenues : le changement de zone lié au déménagement ainsi que le changement de la date de naissance du conducteur principal, le tableau ci-dessous montrent la distribution des contrats avec des sinistres dépassant le seuil parmi les différents critères d'avenant et de surveillance.

Critère d'avenant	Critère de surveillance	
OUI	OUI	7.3%
	NON	3.8%
NON	OUI	16.9%
	NON	72%

On constate qu'une partie des contrats ayant une sinistralité grave possèdent des critères liés aux variables *avenant* et n'obéissent à aucun des critères utilisés par l'équipe de surveillance (3.8%). Certes ce pourcentage n'est pas énorme mais prouve la pertinence des variables *avenant* qui ont été construites et qui pourraient contribuer à l'amélioration du processus de mise sous surveillance.

6.3 Définition d'une nouvelle règle de surveillance

Afin de rendre opérationnels les résultats déterminés dans les parties précédentes, on s'intéresse dans cette partie à la construction d'une nouvelle règle de mise sous surveillance.

Cette dernière va inclure les nouvelles variables explicatives de la sinistralité graves qui ont été construites et étudiées tout au long de cette étude : la variable liée au changement de zone ainsi que la variable lié au changement de la date de naissance du conducteur principal. Ces dernières doivent être également combinée aux variables tarifaires.
Pour ce faire, on a choisi de sélectionner la variable tarifaire la plus pertinente d'après les résultats du modèle XGBoost anciennete_permis ainsi que les deux variables *avenant* chgt_zone et chgt_naiss_conducP.

La variabe Anciennete_permis

Afin de pourvoir exploiter la variable anciennete_permis pour définir une nouvelle règle opérationnelle de mise sous surveillance, il faut définir un seuil qui sépare les assurés ayant un risque important d'avoir un sinistre grave et ceux qui ne l'ont pas.
D'aprés le graphique ci-dessous 6.2, on peut retenir le seuil de 13 ans d'ancienneté de permis. En effet, les assurés ayant un permis récent (ancienneté de permis inférieurs à 13 ans) présente un risque plus élevé d'avoir un sinistre grave.

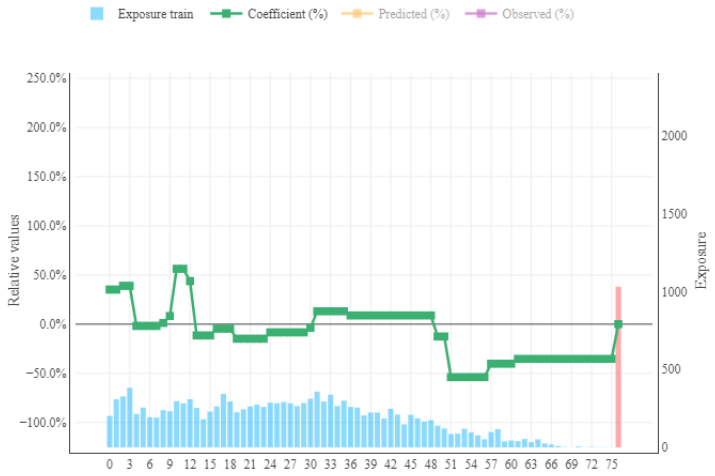


FIGURE 6.2 – Représentaion graphique de la variable anciennete_permis

Nouvelle règle de mise sous surveillance

On a déterminé dans la partie précédente le seuil à utiliser pour la variable tarifaire Anciennete_permis.

A cette dernière, on ajoutera les variables *avenant* qui ont été sélectionnées par les modèles GLM et XGboost pour définir la règle suivante de mise sous-surveillance pour les assurés ayant déjà un sinistre corporel :

Anciennte_permis<13 ans ET (chgt_zone=1 ou chgt_naiss_conducP=-1)
Cette règle a été testée sur l'ensemble de la base de modélisation pour calculer la probabilité moyenne d'avoir un sinistre grave dans le cas où on obéit à la règle et dans le cas contraire.

Les résultats sont présentés ci-dessous :

Critère d'avenant	Moyenne des probabilités d'avoir un sinistre grave
OUI	0.66
NON	0.34

D'après les valeurs de probabilité obtenues, on s'aperçoit que les assurés satisfaisant à la nouvelle règle de mise sous surveillance ont une probabilité élevée d'avoir un sinistre grave. Ce résultat permet de s'assurer que l'on détecte bien le bon segment d'assurés à travers la nouvelle règle établie.

Cette nouvelle règle de mise sous-surveillance sera étudiée, à la suite de ce mémoire, par l'équipe de surveillance. Elle pourra ainsi être incluse au processus de mise sous-surveillance. Une étude à long terme de son impact sur la profitabilité di portefeuille pourra ainsi être envisagée.

7

Conclusion

Les sinistres graves sont caractérisés par une faible fréquence et un coût très élevé. Trouver les variables pertinentes pour les prédire est l'un des enjeux majeurs pour les assureurs.

Le but de ce mémoire était la construction de nouvelles variables qui pourraient, en plus des variables tarifaires classiques, expliquer la fréquence des sinistres graves. Une étape importante lors de cette étude était la création de la base de modélisation et la gestion des données déséquilibrées. En effet, les sinistres graves sont des phénomènes qu'on n'observe pas souvent dans le portefeuille. La variable cible qu'on cherche à expliquer présentait ainsi un déséquilibre net entre la classe des assurés ayant eu un sinistre grave et ceux qui n'ont pas eu de sinistre grave. Pour résoudre ce problème, une technique de rééchantillonnage a été appliquée à notre base. Une fois l'étape de création de la base de modélisation est achevée, on a commencé la modélisation de la variable cible.

Deux approches ont été adoptées. On a commencé par un modèle GLM couramment utilisé dans les processus de tarification. La première étape consiste à modéliser la variable cible en fonction des variables tarifaires. On sélectionne ensuite le modèle le plus pertinent parmi ceux proposés par l'outil utilisé. La deuxième étape de la modélisation consiste à enrichir le modèle obtenu avec les nouvelles variables construites pour étudier leurs impacts. La deuxième approche consiste à utiliser l'algorithme XGBoost très utilisé pour les problématiques de classification et assez reconnu pour ces performances.

On s'aperçoit que les variables construites jugées pertinentes par les deux modèles sont les mêmes : dans les deux cas on retient le changement de zone et le changement de date de naissance du conducteur principal parmi les quatre variables construites. Cependant l'importance explicative accordée à ces variables n'est pas la même pour les deux modèles.

Les résultats obtenus par les différents modèles testés permettent de répondre partiellement à la problématique initiale. On rappelle que l'objectif de cette étude était la construction de

nouvelles variables explicatives basées sur le comportement antérieur de l'assuré. Ces dernières permettent d'améliorer le pouvoir explicatif des modèles de prédiction des sinistres graves basés uniquement sur les variables tarifaires. Afin de mieux comprendre l'apport de ces nouvelles variables, on a testé leur impact sur le processus de surveillance du produit « L'Auto Generali ».

On a défini une nouvelle règle de mise sous-surveillance qui a été établie à partir des résultats obtenus. Cette dernière inclut à la fois la variable tarifaire la plus importante ainsi que les deux variables *avenant* qui ont été sélectionnées par le modèle.

Certes la nouvelle règle ne permet pas de détecter en avance tous les contrats qui ont eu des sinistres graves. Cependant, l'ajout de cette dernière au processus de surveillance aurait pu permettre de détecter environ 4% plus de contrats.

Encore aujourd'hui, la prédiction des sinistres graves reste une problématique ouverte. Les modèles testés lors de cette étude ont certes permis d'apporter de nouvelles pistes pour maitriser mieux les sinistres graves au sein du portefeuille de Generali. Mais, d'autres pistes pourraient encore améliorées ces résultats.

8

Note de synthèse

8.1 Contexte de l'étude et objectif

Les sinistres graves se caractérisent des sinistres ordinaires par leurs coûts élevés. Malgré qu'ils soient des phénomènes rares, leur présence dans le portefeuille affecte fortement la rentabilité de ces derniers. On s'aperçoit ainsi de l'importance de la prédiction de ces sinistres. La problématique de ce mémoire s'inscrit dans ce même cadre. On se propose de construire de nouvelles variables explicatives de la fréquence des sinistres graves. Ces dernières seraient construites à partir des avenants antérieurs effectués par l'assuré.

On s'intéresse au produit « L'Auto Generali » destiné pour les véhicules 4 roues « pur » : Hors camping-car, remorque et caravane, et hors usages spéciaux, plus précisément la garantie RCC (Responsabilité civile pour les sinistres corporels).

8.2 Construction des nouvelles variables explicatives

Une étape cruciale de la modélisation est la construction des nouvelles variables explicatives décrivant le comportement antérieur de l'assuré. Ces dernières seront construites à partir des bases des avenants. La base de modélisation est la base tarifaire qui regroupe les contrats présents dans le portefeuille Generali en 2018, 2019 et 2020. Pour ces contrats, on s'intéresse aux avenants effectués depuis 2015. Le choix de cette date était imposé par des contraintes informatiques.

La première étape consiste à construire la base contenant tous les avenants effectués depuis 2015 jusqu'à 2020. On focalisera notre étude notamment sur 3 types d'avenant.

- Avenant du conducteur principal engendrant un changement de la date de naissance

ainsi que de la date de permis du conducteur principal. On récupère dans ces deux cas, la différence entre l'ancienne et la nouvelle valeur.

- Avenant lié à un déménagement engendrant un changement de la zone d'habitation et donc un changement du risque. En effet, à chaque zone, on attribue un coefficient de risque se basant sur le modèle de sinistralité antérieure. Ainsi, on récupère, pour chaque contrat ayant un avenant lié à un déménagement, la différence entre les coefficients de risque de la nouvelle et de l'ancienne zone.
- Avenant de véhicule engendrant dans certains cas un changement du groupe de véhicule. En effet, le groupe est un indicateur synthétique de la puissance, de la vitesse et du poids. Il renseigne l'assureur sur la puissance du véhicule. Cette variable est numérique et prend des valeurs qui varient entre 20 et 50. 20 signifie que le véhicule est de très faible puissance alors que 50 est un indicateur d'une forte puissance. Dans le cas d'un avenant de véhicule, on récupère la différence entre la valeur du groupe du nouveau véhicule et celle de l'ancien.

Ces différentes variables construites seront ajoutées à la base tarifaire pour obtenir la base de modélisation totale.

8.3 Préparation de la base de modélisation et gestion des données déséquilibrées

Une fois la base tarifaire est enrichie avec les variables *avenant*, on s'intéresse à l'étude de la variable cible. Pour cela, il s'avère essentiel de définir un seuil pour les sinistres graves. Ce dernier nous permettra, par conséquent, de définir notre variable cible.

On utilise dans cette étude une approche statistique basée sur la méthode de qq-plot. Cette méthode nous a permis de déterminer le seuil 50 000 euros. Autrement-dit, tout sinistre dont le montant dépasse cette valeur sera considéré comme un sinistre grave.

Le graphique ci-dessous montre la répartition de la variable cible.

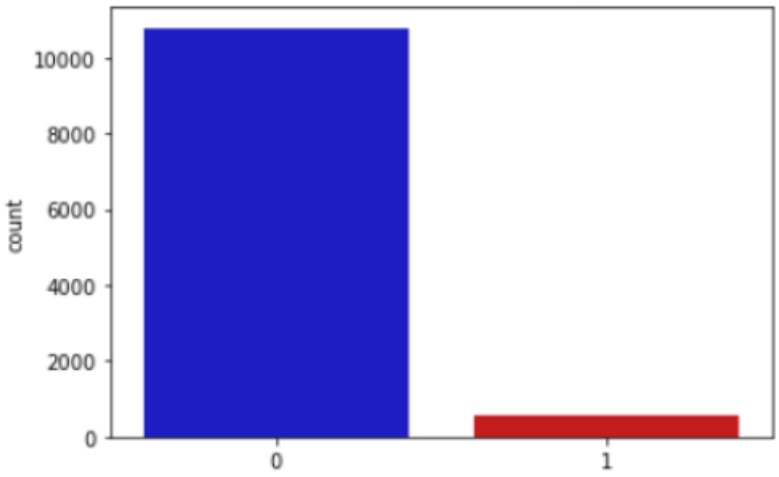


FIGURE 8.1 – Distribution de la variable cible target_SNBSIN2GRAV dans la base initiale

On s'aperçoit que les données présentent un déséquilibre entre la classe des assurés ayant eu un sinistre grave et ceux qui n'ont pas eu de sinistre grave. Plusieurs méthodes de Rééquilibrage de données peuvent être envisagées : un sous-échantillonnage, un sur-échantillonnage, une combinaison des deux méthodes précédentes. Ces trois méthodes ont été testées sur la base de modélisation. On Compare ensuite les différents modèles obtenus. La méthode retenue sera celle qui permet d'avoir le modèle le plus robuste.

En utilisant la méthode de sous-échantillonnage, on obtient des résultats assez faibles : la valeur de l'AUC ne dépasse pas 56%. La méthode de sur-échantillonnage, quant à elle, permet d'améliorer les performances du modèle. On obtient un coefficient AUC supérieur à 70%. Ce dernier est légèrement plus élevé lorsqu'on combine les deux précédentes méthodes. Cette dernière approche sera donc retenue pour la suite de l'étude.

8.4 Résultats obtenus

Après avoir présenté la méthode qui sera retenue pour traiter le déséquilibre des données dans la base, on s'intéresse dans la suite à la modélisation. Deux modèles seront implémentés : le modèle linéaire généralisé GLM et le modèle d'extreme gradient boosting XGBoost.

Modèle GLM

Le modèle GLM est classiquement utilisé dans le processus de tarification au sein de Generali. On l'utilisera dans cette étude pour évaluer la pertinence des variables *avenant* construites.

On commence par implémenter le modèle GLM avec uniquement les variables tarifaires. On sélectionne ensuite le modèle qui a le pouvoir prédictif le plus élevé auquel on ajoute les variables construites. On s'aperçoit que l'ajout de variables liées aux avenants antérieurs effectués par l'assuré améliore le pouvoir prédictif du modèle.

Le tableau ci-dessous résume les résultats obtenus.

	modèle sans variables *avenant*	modèle avec variables *avenant*
AUC	0.70	0 .78
RMSE	0.43	0.39

Modèle XGBoost

Outre le modèle GLM, un modèle couramment utilisé pour résoudre des problématiques de classification a été testé : le modèle XGBoost. Le but est de prédire la variable cible, celle qui indique l'occurrence ou non d'un sinistre grave en fonction du panel de variables à disposition. On obtient des performances du modèle légèrement plus élevées que celles du modèle GLM.

Le tableau ci-dessous résume les différents résultats obtenus :

	XGboost	GLM
AUC	0.80	0.78
RMSE	0.32	0.39

Dans les deux cas, deux variables *avenant* ont été sélectionnées : la variable indiquant le changement de la date de naissance du conducteur principal et celle indiquant le changement de zone lié à un déménagement.

Plus précisément, on déduit que le passage à un conducteur principal plus jeune augmente la probabilité d'avoir un sinistre grave. En outre, le changement de zone vers une zone plus risquée joue également en faveur de l'occurrence d'un sinistre grave.

8.5 Cas d'application pratique

Après avoir déterminer les deux variables *avenant* les plus pertinentes pour décrire la sinistralité grave, on s'intéresse à un cas pratique d'application de ses résultats : La surveillance. En d'autres termes, on cherche à définir une nouvelle règle de mise sous-surveillance qui inclura les deux nouvelles variables explicatives. On s'aperçoit que cette nouvelle règle permet de détecter un certain nombre de contrats qui n'auraient pas pu être détecté uniquement en se basant sur les critères actuels de mise sous-surveillance.

8.6 Conclusion

L'objectif initial de ce mémoire était la construction de nouvelles variables explicatives qui pourraient, en plus des variables tarifaires, expliquer au mieux la sinistralité grave. On a commencé par considérer quatre variables qui pourraient potentiellement être pertinente pour la prédiction des sinistres graves. Les modèles implémentés ont permis de garder uniquement deux variables parmi celles ajoutées à la base tarifaire : le changement de zone et le changement de la date de naissance du conducteur principal.

Ces résultats permettent de renforcer le système de surveillance actuel de Generali. En effet, une nouvelle règle de mise sous-surveillance a été construite à partir de ces deux variables. Cette dernière permet de détecter environ 4% de plus de contrats ayant des sinistres graves (comparé au cas où on utilisait uniquement les critères actuels).

Pour conclure, on peut dire qu'on a réussi partiellement à atteindre l'objectif initial de l'étude et à définir un cas d'application opérationnel qui pourra être mis en place facilement. Ce mémoire a également révélé la difficulté d'étudier les sinistres graves qui restent des phénomènes assez rares mais très coûteux. Les résultats obtenus lors de cette étude permettent d'améliorer les modèles prédictifs des sinistres graves. Cependant, déterminer de manière très exacte l'occurrence d'un sinistre grave reste encore un problème ouvert.

9

Executive summary

9.1 Context of the study and objective

Severe claims are characterised by their high costs compared to ordinary claims. Although they are rare phenomena, their presence in the portfolio strongly affects its profitability. This shows the importance of predicting these claims. The problematic of this thesis falls within this same framework. We propose to construct new explanatory variables for the frequency of severe claims. These would be constructed from previous amendments made by the insured.

We are interested in the product "L'Auto Generali" intended for "pure" 4-wheeled vehicles : excluding motorhomes, trailers and caravans, and excluding special uses, more precisely the RCC cover (Civil Liability for bodily injury).

9.2 Construction of new explanatory variables

A crucial step in the modelling is the construction of new explanatory variables describing the past behaviour of the insured. These will be constructed from the amendment bases. The modelling base is the tariff base which includes the contracts present in the Generali portfolio in 2018, 2019 and 2020. For these contracts, we are interested in the endorsements made since 2015. The choice of this date was imposed by IT constraints.

The first step is to build the database containing all the amendments made from 2015 to 2020. We will focus our study on three types of amendment in particular.

- amendment of the main driver resulting in a change of the date of birth and the date

of licence of the main driver. In both cases, the difference between the old and new values is recovered.

- Amendment linked to a move resulting in a change in the zone of residence and therefore a change in risk. In fact, each zone is assigned a risk coefficient based on the previous claims pattern. Thus, for each policy with an amendment due to a move, the difference between the risk coefficients of the new and old zones is calculated.
- A vehicle amendment that in some cases results in a change of the vehicle group. The group is a synthetic indicator of power, speed and weight. It informs the insurer about the power of the vehicle. This variable is numerical and takes values that vary between 20 and 50. 20 means that the vehicle is of very low power while 50 is an indicator of high power. In the case of a vehicle endorsement, the difference between the value of the group of the new vehicle and that of the old vehicle is calculated.

These different constructed variables will be added to the tariff base to obtain the total modelling base.

9.3 Preparation of the modelling base and management of unbalanced data

Once the tariff base is enriched with the variables *avenant*, we are interested in studying the target variable. For this, it is essential to define a threshold for serious claims. The latter will therefore allow us to define our target variable.

In this study, a statistical approach based on the qq-plot method is used. This method allowed us to determine a threshold of 50,000 euros. In other words, any claim exceeding this value will be considered a serious claim.

The graph below shows the distribution of the target variable.

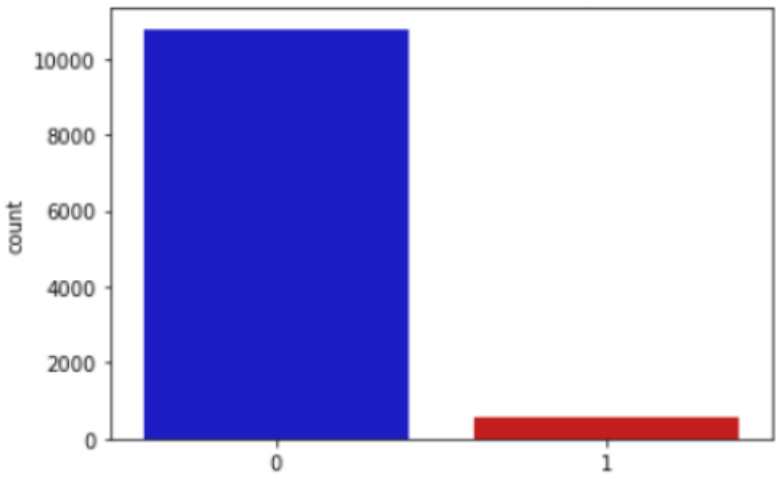


FIGURE 9.1 – Distribution of the target variable target_SNBSIN2GRAV in the initial database

It is found that there is an imbalance in the data between the class of policyholders who have had a serious claim and those who have not had a serious claim. Several methods of rebalancing the data can be considered : sub-sampling, over-sampling, and a combination of the two previous methods. These three methods were tested on the modelling base. The different models obtained are then compared. The method chosen will be the one that provides the most robust model.

Using the sub-sampling method, the results are quite low : the AUC value does not exceed 56%. The oversampling method, on the other hand, improves the performance of the model. The AUC coefficient is higher than 70

9.4 Results

After having presented the method that will be used to treat the imbalance of the data in the database, we will focus on the modelling. Two models will be implemented : the generalized linear model GLM and the extreme gradient boosting model XGBoost.

GLM model

The GLM model is classically used in the pricing process within Generali. It will be used in this study to assess the relevance of the variables constructed.

First, the GLM model is implemented with only the tariff variables. We then select the model with the highest predictive power and add the constructed variables. It is found that adding variables related to previous amendments made by the insured improves the predictive power of the model.

The table below summarises the results obtained.

	model without the variables *avenant*	model without variables *avenant*
AUC	0.70	0 .78
RMSE	0.43	0.39

XGBoost model

In addition to the GLM model, a model commonly used to solve classification problems was tested : the XGBoost model. The aim is to predict the target variable, which indicates the occurrence or non-occurrence of a serious claim, according to the panel of variables at hand. The performance of the model is slightly higher than that of the GLM model.

The table below summarises the different results obtained :

	XGboost	GLM
AUC	0.80	0.78
RMSE	0.32	0.39

In both cases, two variables were selected : the variable indicating the change in the date of birth of the main driver and the variable indicating the change of zone due to a move.

More specifically, it is deduced that switching to a younger main driver increases the probability of a severe claim. In addition, moving to a riskier area also increases the likelihood of a severe claim.

9.5 Practical application cases

After having determined the two most relevant variables to describe serious accidents, we are interested in a practical case of application of these results : surveillance. In other words, we seek to define a new rule of sub-surveillance which will include the two new explanatory variables. It is found that this new rule allows the detection of a certain number of contracts that could not have been detected solely on the basis of the current monitoring criteria.

9.6 Conclusion

The initial objective of this paper was the construction of new explanatory variables that could, in addition to the tariff variables, best explain serious claims. We started by considering four variables that could potentially be relevant for the prediction of serious claims. The models implemented retained only two of the variables added to the tariff base : change of zone and change of date of birth of the main driver.

These results make it possible to strengthen Generali's current monitoring system. Indeed, a new rule of under-surveillance has been built from these two variables. The latter allows to detect about 4% more contracts with serious claims (compared to the case where only the current criteria were used).

In conclusion, it can be said that we have partially succeeded in achieving the initial objective of the study and in defining an operational application case that can be easily implemented. This thesis has also revealed the difficulty of studying serious disasters, which remain fairly rare but very costly phenomena. The results obtained in this study allow for the improvement of predictive models for severe losses. However, it is still an open problem to determine very accurately the occurrence of a severe claim.

Bibliographie

[BEKKAR M., ALITOUCHE T., 2013] *Imbalanced data learning approaches review. International Journal of Data Mining Knowledge Management Process Vol.3, No.4.*

[BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., 1984] *Classication and Regression Trees, Chapman and Hall.*

[Esbjörn Ohlsson, Björn Johansson (2010)] *Non-life Insurance Pricing with Generalized Linear Models. Springer*

[FFSA, 2019] *[6] FFSA (2015). Rapport annuel FFSA 2019*

[Molnar, 2020] *Interpretable Machine Learning : A guide for making black box models explainable*

[Noureddine Benlagha, Michel Grun-Réhomme, Olga Vasechko (2009)] *Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes. Revue Modulad, 47-80*

[Oufella, 2008] *Evolution du concept de Front ROC et combinaison de classifieur*

[Peter McCullagh, John Ashworth Nelder (1999)] *Generalized Linear Models. Chapman Hall*

[Rakatomalala, 2017] *Pratique de la régression logistique, Régression logistique binaire et polytomique.*

[Rakatomalala, 2018] *Descente de gradient, Principe de la descente de gradient pour apprentissage supervisé*

[YEN S., LEE Y., 2009] *Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications 36 (2009) 57185727.*

Sites internet de référence

- Site de l'Argus de l'Assurance : www.argusdelassurance.com
- Site de la Fédération Française des Sociétés d'Assurances : www.franceassureurs.fr
- Site Légifrance : www.legifrance.gouv.fr

Table des figures