Mémoire présenté le : 9 mai 2023

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Mathieu NUNES
Titre : **Modélisation du risque arrêt de travail pour un portefeuille de TNS par méthodes Data Science**

Confidentialité : Non ☑ Oui (Durée : 1 an ☑ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :
Sylvain Caraco
Anaelle Le Berre
Cécile Paradis

Membres présents du Jury du Master Actuariat de l'ISFA :
Pierre Ribereau

Entreprise :
Nom : ENTORIA
Signature :

ENTORIA
SAS au capital de 2 000 000 €
166 rue Jules Guesde
92300 LEVALLOIS-PERRET
RCS NANTERRE [illegible] 391

Directeur de Mémoire en entreprise :
Nom : Brice IEMMI
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)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise

Signature du candidat

Résumé

L'investigation développée dans ce mémoire est réalisée dans le but d'améliorer le modèle de tarification d'un produit prévoyance, et en particulier du risque arrêt de travail qui lui est lié. Des axes d'amélioration ayant été décelés et non traités, une étude de l'incidence en arrêt de travail pourrait répondre à ce besoin de combler les lacunes existantes. L'objectif premier est donc de parvenir à une meilleure segmentation du modèle, qui demande donc une plus grande compréhension du risque associé. Plusieurs approches sont étudiées ici, en modélisation statistique et apprentissage automatique. L'étude détaille les démarches et les choix effectués dans la construction du modèle de fréquence.

Un autre enjeu de ce mémoire est d'exploiter au mieux le pouvoir prédictif des variables en jeu, en ayant recours à de l'enrichissement et du retraitement de données de manière pertinente. Cela passera notamment par la construction de zoniers selon différentes méthodes, afin d'optimiser le traitement de l'information géographique.

Mots-clés : *GLM*, *Arbres CART*, *Variables tarifaires*, *Arrêt de travail*, *Forêt aléatoire*

Abstract

The investigation developed in this dissertation is carried out with the aim of improving the pricing model of a provident product, and in particular of the work stoppage risk linked to it. As areas for improvement have been identified and not addressed, a study of the incidence of work stoppage could meet this need to fill existing gaps. The primary objective is therefore to achieve a better segmentation of the model, which requires a better understanding of the associated risk. Several approaches are studied here, in statistical modeling and machine learning. The study details the steps and choices made in the construction of the frequency model.

Another challenge of this thesis is to make the best use of the predictive power of the variables involved, by using data enrichment and reprocessing in a relevant way. This will be done through the construction of zoners according to different methods, in order to optimize the processing of geographic information.

This study is being conducted on historical data within the 2016-2021 observation window. The goal of this study is to propose a solution to improve the current rates of Entoria's individual benefit product.

Mots-clés: *GLM*, *CART algorithm*,*Pricing variables*, *Work stoppage*, *Random Forest*

Note de Synthèse

Ce mémoire a été réalisé au sein de la Direction Technique en assurance de personnes du courtier grossiste Entoria. Cette équipe gère des problématiques liées à la Prévoyance et à la Santé. L'objet d'un contrat de prévoyance est d'assurer sa protection contre les aléas de la vie humaine. Au sein de ces aléas se trouve l'arrêt de travail qui est le thème lié à cette étude.

La compréhension d'un risque est une étape importante dans un processus de tarification. Ainsi, l'étude du risque arrêt de travail, et plus précisément de l'incidence d'un arrêt de travail, en prévoyance individuelle est une étape cruciale qui permet de mieux appréhender les différents profils de risque afin de proposer des prix plus fins. Certaines informations clefs comme la profession ou la zone de vie de l'assuré n'étant pas parfaitement maîtrisées aujourd'hui au sein de l'entreprise, ou partiellement connues, il semble naturel de chercher à combler ce manque. La figure 3 permet par exemple d'apprécier le gain en termes de répartitions qu'a permis l'analyse data science. La population du portefeuille est désormais segmentée en groupes significativement différents quant à leur exposition au risque arrêt de travail.

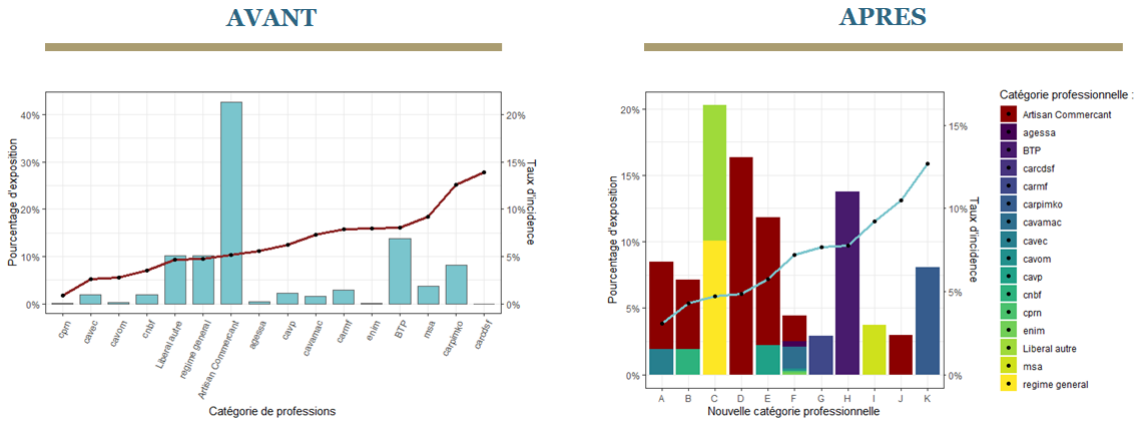

FIGURE 1 – Évolution de la variable liée à la profession des assurées avant et après étude data science

Pour cela, la base de données a été construite à partir de données de gestions enrichies de données externes, qui ont été agrégées à la maille du code Siret et du département du TNS. Dans une optique de préparation des études Data Science qui vont être menées, un traitement des données est réalisé, permettant directement d'écarter les informations non pertinentes et de discrétiser les variables intéressantes.

La première phase d'étude correspond à la construction d'un modèle linéaire généralisé de loi de Poisson. Le point de départ repose sur une approche de sélection automatique de variables par algorithme forward basé sur l'indice AIC des modèles testés. S'en suit un raffinement du modèle sélectionné afin de le rendre davantage robuste, avec des variables non corrélées et suffisamment significatives.

Dans la phase suivante, une approche fondamentalement différente est exploitée, l'apprentissage par forêt aléatoire. L'objectif est ici de challenger la méthode actuelle de tarification utilisée par Entoria, en la confrontant avec une approche non-paramétrique.

Dans chacune des deux phases précédentes, l'information géographique n'est pas encore exploitée dans les modèles. Deux méthodes d'étude des résidus sont mise en place, d'une part analytique, et d'autre part prédictive par arbre de décision CART. L'approche prédictive, qui semble être la plus pertinente, utilise les données externes aux données de gestion pour essayer d'apprendre et prédire quels sont les départements qui nécessitent d'être regroupés ou séparés. La figure 4 présente le zonier ainsi créé lors de la création du modèle par forêt aléatoire.

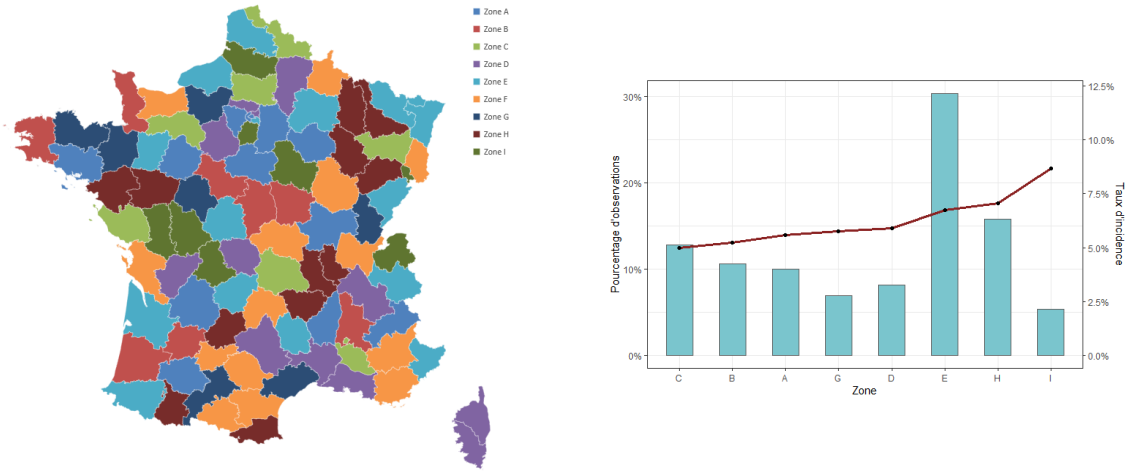

FIGURE 2 – Représentation du zonier créé par arbre CART sur une carte de France, et étude de sa répartition et des taux d'incidence par zone

Enfin, une comparaison finale des deux approches est réalisée. Celle-ci s'appuie à la fois sur des arguments théoriques en fonction des caractéristiques inhérentes à chacune, mais aussi par une vision davantage métier en s'intéressant aux effets par agrégations de populations cibles.

Synthesis note

This thesis was carried out within the Technical Department of the Entoria insurance wholesaler. This team manages issues related to provident insurance and health insurance. The purpose of a provident contract is to ensure protection against the hazards of human life. Among these hazards, there is the work stoppage which is the theme of this study.

Understanding a risk is an important step in the underwriting process. Thus, the study of the risk of work stoppage, and more precisely of the incidence of work stoppage, in individual provident insurance is a crucial step that allows to better understand the different risk profiles in order to propose more accurate prices. Since certain key information such as the insured's profession or living area is not perfectly mastered today within the company, or is only partially known, it seems natural to try to fill this gap. For example, the figure 3 illustrates the gains in terms of distribution that data science analysis has made possible. The portfolio population is now segmented into groups that are significantly different in terms of their exposure to work stoppage risk.

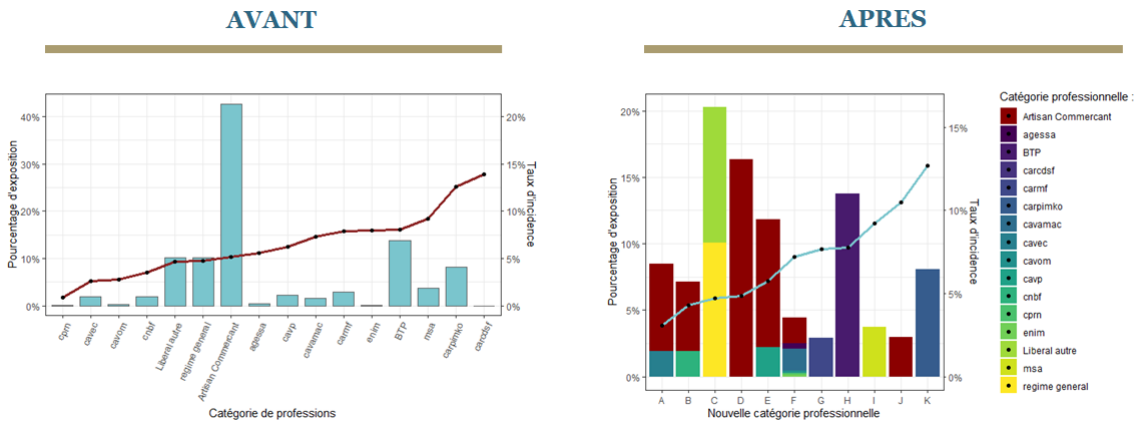

Figure 3 – Changes in the occupation variable of insured women before and after the data science study

For this purpose, the database was built from management data and open data, which were aggregated at the level of the Siret code and the department of the TNS. In order to prepare the Data Science studies that will be conducted, a data processing is carried out, allowing the direct elimination of useless information and the clean discretization of the interesting variables.

The first phase of the study corresponds to the construction of a generalized linear Poisson

model. The starting point is based on an approach of automatic selection of variables by forward algorithm based on the AIC index of the tested models. This is followed by a refinement of the selected model to make it more robust, with uncorrelated and sufficiently significant variables.

In the next phase, a fundamentally different approach is exploited, the random forest learning. The objective here is to challenge the current pricing method used by Entoria, by comparing it with a non-parametric approach.

In each of the two previous phases, the geographical information is not yet exploited in the models. Two methods of studying the residuals are implemented, on the one hand analytical, and on the other hand predictive by CART decision tree. The predictive approach, which seems to be the most relevant, uses data external to the management data to try to learn and predict which departments need to be grouped or separated. The figure 4 shows the zonier thus created when the random forest model was created.

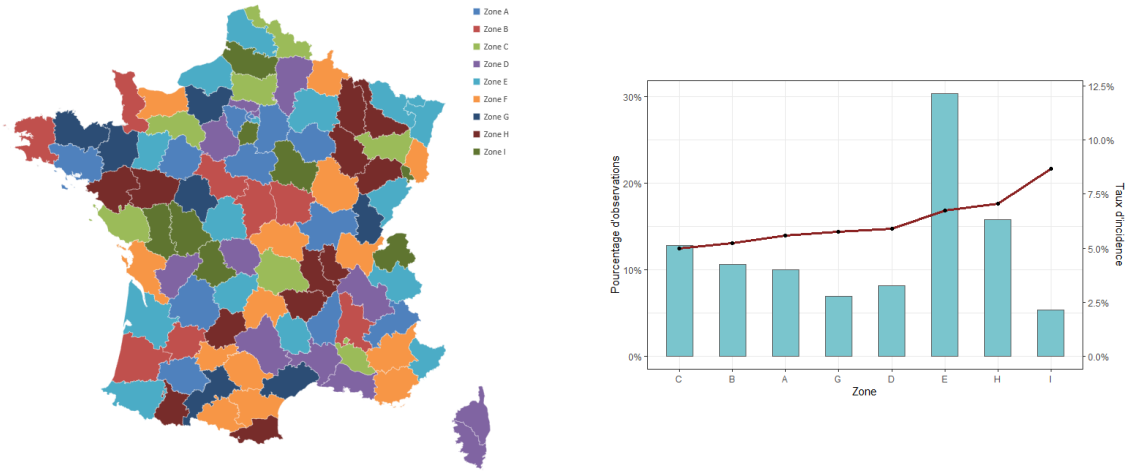

Figure 4 – Representation of the CART tree on a map of France, and study of its distribution and incidence rates by zone

Finally, a final comparison of the two approaches is performed. This comparison is based both on theoretical arguments according to the inherent characteristics of each approach, and on a more business-oriented vision of the effects of aggregating target populations.

Remerciements

Je souhaite en premier lieu adresser mes sincères remerciements à mon tuteur Brice IEMMI, qui m'a accompagné et dont les conseils avisés m'ont aidé à mener à bien cette étude. Je remercie aussi Nataliya VEZHNYAEVA dont le soutien a été très instructif lors de la création de la base de données. Un grand merci à l'ensemble des membres de la direction technique d'Entoria, pour leur disponibilité et leur bonne humeur qui m'ont permis d'évoluer à leur côté. Une note particulière pour Camille LE BOURHIS, André BAHABI, et Nathan VEVERT sans qui cette année d'alternance n'aurait pas eu la même saveur.

Je remercie aussi mon tuteur académique Pierre THEROND pour son encadrement lors de l'établissement du sujet d'étude.

Une pensée également à ma famille pour son soutien, et plus particulièrement à ma mère qui a toujours été présente et m'a encouragé tout le long de ma scolarité, merci à elle.

Je ne pourrais pas terminer ces remerciements sans m'adresser aux amis qui m'ont accompagné durant ces 3 années de formation. Je pense tout d'abord à mes précieux colocataires Gurvan JAOUEN et Quentin EMERY, avec qui nous avons passé deux années mémorables, et avec qui j'ai eu des échanges très constructifs. Sans oublier Guillaume MORISSE, Julie MARZIO, Pierre HERRY, Lucie BULTEAU et Thomas MASSE, mes fidèles compagnons avec qui nous avons traversé chaque étape dans la joie, la fête, et la bonne humeur. J'adresse enfin un clin d'oeil à Clément LAGARDE et Siméon DELYON MANIABLE, pour leur présence sans faille, qui m'a permis de rester toujours plus productif et motivé.

Un grand merci.

Table des matières

Introduction

L'assurance prévoyance est une couverture contre les aléas de la vie, qui sont l'incapacité, l'invalidité, la dépendance et le décès. L'assuré, en contrepartie des cotisations versées à l'assureur, obtient une garantie de protection financière, en cas de réalisation de l'aléa.

Le sujet de ce mémoire concerne plus particulièrement le risque arrêt de travail, défini comme étant une inaptitude physique ou psychologique à exercer une activité professionnelle, de façon partielle ou totale. Les prestations de prévoyance incapacité complètent les indemnisations de la Sécurité Sociale afin de reconstituer le revenu de l'assuré.

Ce mémoire a été réalisé à partir de données concernant des Travailleurs Non Salariés (TNS) pour un produit de prévoyance individuelle. L'objectif étant d'étudier, par des méthodes statistiques et de Data Science, le lien entre les caractéristiques des assurés et l'incidence de leur arrêt de travail entre les années 2016 et 2021.

Un enjeu de ces travaux est de parvenir à enrichir la base de données interne d'Entoria avec des données extérieures. Dans cet optique a été jointe la base de données Sirène via le code Siret des TNS. Aussi, des données issues de la plateforme Open Data data.gouv ont été importées afin d'apporter des informations géographiques plus tarifantes que la position géographique elle-même.

La finalité de ces travaux devra donc répondre à un double objectif : accroître les capacités de prédiction des taux d'incidence en arrêt de travail ; et démontrer l'intérêt d'utilisation d'Open Data dans la tarification en prévoyance TNS.

Première partie

Cadre de l'étude

Chapitre 1

Contexte et présentation du marché de la prévoyance

1.1 Nature de la prévoyance

La prévoyance est définie comme l'assurance qui a pour objectif de protéger les personnes des aléas de la vie humaine. La prévoyance est définie par l'article n°89-10009 de la loi EVIN comme "Les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité ou des risques d'incapacité de travail ou d'invalidité ou du risque chômage". [8]

Dans le cas du décès, les prestations au titre du contrat de prévoyance seront destinées aux bénéficiaires désignés par l'assuré lors de la création du contrat. Les prestations pour le décès peuvent être de deux natures différentes, en capital fixe, ou sous la forme d'une rente. Le capital est comme son nom l'indique une somme fixée au départ, qui sera versée lors du décès de l'assuré. Les rentes, quant à elles, sont des versements réguliers qui seront destinés au conjoint dans le cas d'une « rente conjoint », afin de prendre en compte la perte de revenu de l'assuré, ou destinées au enfants de l'assuré dans le cas d'une « rente éducation »afin de prendre en charge les dépenses et financements relatifs aux études des enfants bénéficiaires jusqu'à leur majorité (ou plus si indiqué dans le contrat lors de la création de celui-ci).

La garantie incapacité a pour rôle de compléter le revenu de l'assuré lorsque celui-ci subit un sinistre qui l'empêcherait de poursuivre son activité pendant une durée temporaire. Les prestations peuvent être forfaitaires, c'est-à-dire fixées par le contrat, ou indemnitaires, afin de toujours compléter l'intégralité du revenu, après versement des prestations du régime obligatoire.

La garantie invalidité quant à elle intervient lorsque l'assuré subit un sinistre qui le met dans un état définitif d'incapacité d'exercer son activité, ou si son statut d'incapacité temporaire atteint une plage de 3 années consécutives. Le principe est le même que pour la garantie incapacité, c'est-à-dire qu'elle vise à compléter la perte de revenu de l'assuré.

La prévoyance peut exister sous forme de contrat collectif ou individuel. Une entreprise peut en effet souscrire une assurance prévoyance collective qui sera adressée à ses salariés. Le marché de la prévoyance est majoritairement composé de contrats collectifs, de par le caractère obligatoire qui s'impose au collège cadre. Quant aux contrats individuels, ils

seront adressés aux salariés n'ayant pas de couverture et qui souhaitent se protéger, ainsi qu'aux travailleurs non-salariés.

Comme indiqué précédemment, l'étude menée dans ce mémoire concerne la prévoyance individuelle, et plus particulièrement sur le risque arrêt de travail pour une population de travailleurs non-salariés.

1.2 Le risque arrêt de travail

Selon le ministère du travail, l'arrêt de travail est une prescription de votre médecin, attestant que votre état de santé ne vous permet pas d'exécuter votre contrat de travail ou de continuer votre activité. Il peut avoir différentes causes :

- **Une maladie** : l'arrêt maladie est accordé lorsque votre état de santé vous empêche d'exercer votre activité.
- **Un accident du travail/accident de trajet** : l'arrêt de travail peut avoir pour origine un accident survenu au cours de votre activité professionnelle.
- **Une maladie professionnelle** : l'arrêt de travail peut résulter d'une maladie due à l'exercice de l'activité professionnelle ou survenue au cours de cette activité.

Dans le cas d'une maladie ou d'accident d'origine professionnelle, l'arrêt de travail est défini comme « professionnel ». De même, si la cause survient dans le cadre de la vie courante, l'arrêt est défini comme « privé ». A ces deux familles sont attribuées deux types d'arrêt :

- Un arrêt de travail professionnel peut être défini comme une incapacité temporaire ou permanente.
- Un arrêt de travail privé peut exister sous la forme d'une incapacité temporaire ou d'une invalidité.

L'invalidité désigne un état irréversible tant au niveau physique que psychique de l'individu.

L'incapacité de travail désigne l'état d'impossibilité temporaire d'exercer une activité professionnelle ou d'effectuer certaines tâches liées à son travail. Après une certaine période, l'individu sort de cet état temporaire et peut reprendre son activité.

L'incapacité permanente de travail est déclarée suite à une maladie ou un accident d'origine professionnelle, qui cause des séquelles lui empêchant de continuer d'exercer un activité professionnelle.

L'invalidité et l'incapacité permanente possèdent tous deux un barème permettant d'évaluer leur intensité.

Chapitre 2

Problématique et démarche

2.1 Problématique et motivations opérationnelles de l'entreprise

Afin d'élaborer ses Tarifs Arrêt de Travail en Prévoyance Individuelle, Entoria dispose à ce jour d'un modèle GLM pour prédire l'incidence. Les variables explicatives sont des variables issues des données de gestion. Cependant, des points d'amélioration et d'optimisation existent mais aucune étude n'a été réalisée à ce jour. Ainsi, l'objectif de ce mémoire est de répondre à ce besoin d'amélioration des performances de prédiction de l'incidence d'arrêt de travail. Dans cette optique, trois évolutions sont envisagées :

- Améliorer la base de données historiques utilisée pour la calibration du modèle actuel, en retenant deux approches :
  - Enrichir le panel de variables à disposition en s'appuyant sur des données externes importées depuis des open data.
  - Exploiter le potentiel prédictif de certaines variables clefs par des retraitements pertinents.
- Envisager d'autre méthode de prédiction en se demandant si la méthode d'apprentissage par GLM est la plus adaptée. D'autres approches seraient-elles plus performantes ?
- Envisager un renouvellement de zonier en étudiant différentes approches.

Afin de mesurer les gains de performance imputables à ces évolutions, des indicateurs peuvent être proposés, afin de comparer la performance des modèles mis en place, et conclure finalement sur les caractéristiques du modèle retenu et du modèle GLM actuel. Par ailleurs, ce mémoire permettra de porter une opinion sur l'intérêt de l'utilisation d'open data pour la tarification d'un produit prévoyance individuelle, dans une optique d'amélioration et/ou d'une refonte d'un modèle de prédiction.

2.2 Choix des méthodes d'apprentissage utilisées

Afin d'estimer l'incidence annuelle en arrêt travail, plusieurs possibilités sont envisageables, que ce soit en termes de concepts ou de techniques.

En effet, les prédictions tête par tête peuvent être définies sous différentes formes. Peuvent être imaginées aussi bien la prédiction du nombre de sinistres par année que la durée annuelle en arrêt de travail par exemple, ou bien encore la probabilité qu'un persona entre

en arrêt de travail dans l'année. Dans le cadre de ce mémoire, le choix s'est porté sur l'estimation du nombre de sinistres annuels par tête, afin que les résultats soient dans le format utilisé par l'outil de tarification d'Entoria.

Ensuite, un choix de la méthode de prédiction est nécessaire. Il existe différents outils de Data Science qui permettent de traiter ce sujet, chacun ayant des avantages et inconvénients. Trois outils de prédiction sont exploités dans ce mémoire, à savoir les modèles linéaires généralisés, notés GLM, les arbres de décision CART et les forêts aléatoires qui sont une extension des arbres CART. Ces différents outils sont présentés en détail dans la troisième partie, qui est dédiée aux modélisations statistiques et au machine learning.

Deuxième partie

Données et analyse exploratoire

Chapitre 3

Présentation des données

3.1 Création de la base de données

Comme pour toute étude statistique, le point de départ consiste à construire la base de travail, qui permettra de réaliser les différentes modélisations et études techniques nécessaires afin de tirer des conclusions et prédire l'incidence en arrêt de travail.

Les données internes sont stockées dans de nombreuses bibliothèques reliées entre elles par des clefs primaires variées. Une telle architecture induit indéniablement une vigilance durant la manipulation des données, afin de s'assurer de la qualité des jointures effectuées. Dans le cadre de ce mémoire, les données sont importées de trois serveurs principaux dans lesquels sont stockées, dans différentes bibliothèques, les données relatives aux contrats, aux assurés, aux sinistres, aux courtiers, etc.

En pratique, la construction de la base de données de travail se déroule selon les étapes suivantes :

— Il s'agit tout d'abord d'identifier les variables explicatives qui seront étudiées dans la suite de l'étude. Pour cela, une réflexion préliminaire sur les possibilités offertes par les bibliothèques de données internes permet d'identifier les variables principales qui devraient permettre d'expliquer une part importante du risque arrêt de travail. Pour autant, l'intérêt ne doit pas être uniquement porté sur ces seules variables. En effet, il n'est pas rare que certaines informations soient contre-intuitives, c'est-à-dire qu'une variable peut intuitivement paraître très corrélée à la variable cible, mais que dans les faits, l'étude statistique conclue que l'explication apportée n'est pas significative. Le cas contraire est également envisageable. Ainsi, l'idée est de sélectionner volontairement plus de variables que nécessaire, même celles dont il n'est pas spécialement naturel de penser au premier abord, puis s'ensuivra une sélection plus objective par l'étude statistique.
— Une fois les informations cibles bien définies, il est question de joindre, via les clefs primaires, chacune d'elles depuis les bibliothèques internes où elles sont stockées. Comme spécifiée auparavant, cette étape cruciale doit être menée avec une précaution particulière.
— S'ensuit la phase de manipulation des données durant laquelle est vérifiée la qualité des données, et durant laquelle sont excluent les valeurs aberrantes, ainsi que les informations qui sont globalement trop manquantes.
— Enfin, la dernière étape de la création de la base de données consiste à modifier

cette dernière afin que chacune des lignes reflète un risque sur une période d'une année. Ce qui aboutira à une base composée de 278040 lignes, construites à partir de données historiques qui s'étendent sur la plage 2016 à 2021.

Vient ensuite l'étape de réflexion concernant l'enrichissement de la base de données par des données externes. Les prémices de cette tâche ont fait l'objet de discussions sur les possibilités qui s'offraient à une étude concernant un produit de prévoyance pour une population de TNS. De cette phase de recherche sont ressortis deux axes d'études principaux. Tout d'abord, il s'agit d'enrichir les informations concernant les TNS à la maille du code SIRET, en faisant une jointure entre la base de travail et la base SIRENE mise à disposition par l'INSEE. Malheureusement, une présence relativement forte de lignes vides pour l'information du SIRET dans les bibliothèques internes a abouti à un manque d'environ 50 000 lignes dans la base d'étude, ce qui représente près d'un cinquième du nombre d'observations totales. Ainsi se pose la question concernant les méthodologies et approches pouvant être mises en place sur ces données, dont une part non-négligeable n'est pas disponible, afin de parvenir à exploiter leur pouvoir prédictif. Enfin, le deuxième axe consiste à lier la situation géographique à une maille choisie (région, département, code postal, etc) avec des informations géographiques diverses. Une sélection de différentes variables plus ou moins pertinentes est faite : le panel comporte à la fois des données intuitivement importantes comme la densité médicale, et des données plus originales qui peuvent même être considérées plus anecdotiques comme la température moyenne du lieu. Il est en effet intéressant de vérifier si une information purement géographique ou démographique peut avoir un impact significatif sur la sinistralité en prévoyance. Ces données proviennent de la plateforme data.gouv, une plateforme de diffusion de données publiques de l'état français et développée par Etalab, une mission placée sous l'autorité du Premier ministre.

3.2 Variables internes

Les variables internes correspondent aux données historiques d'Entoria, ce sont les informations des assurés stockées dans les serveurs de l'entreprise. Celle-ci sont séparées en trois catégories qui les caractérisent : les informations concernant directement l'assuré, les informations relatives au produit d'assurance rattaché à l'assuré, et les informations relatives au courtier par lequel l'assuré a procédé à l'adhésion du contrat de prévoyance.

Voici une présentation des données internes retenues pour cette étude.
Tout d'abord, les variables donnant des informations qui concernent l'assuré sont les suivantes :

DEP : Cette variable donne le département de l'assuré. Cette information est exploitée de différentes manières dans ce mémoire. Une explication des méthodologies suivies sera présentée dans la partie correspondante. L'objectif est de conserver l'approche qui sera désignée comme étant la plus performante en terme d'explication de l'incidence selon les indicateurs de performance considérés dans cette étude.

REG : Cette variable est similaire à la précédente, la différence étant qu'il s'agit ici des régions françaises et non des départements. Utiliser cette variable et la variable département dans un même modèle n'est pas possible. Un choix devra donc est fait par la suite afin de

ne conserver qu'une unique fois l'information géographique.

RÉGIME RO : Cette variable renseigne le Régime Obligatoire (RO) de l'assuré. Ce régime est le régime légal de prévoyance auquel est soumis l'assuré. Cette information est directement liée à l'activité professionnelle, une utilisation plus ou moins précise peut donc potentiellement apporter une part intéressante d'explication de la sinistralité.

LIGNE TARIFAIRE PRÉVOYANCE : La ligne tarifaire est une variable qualitative comportant plusieurs catégories. Cette variable n'est pas une variable "de base" au sens premier du terme puisqu'elle a été construite par les services d'Entoria à travers une étude des charges des assurés. Ainsi, de part sa construction, cette variable est corrélée à la fréquence d'incidence en arrêt de travail. Pour autant, l'étude préliminaire de cette variable révélera des modalités sous-représentées, c'est pourquoi cette variable devra connaître des modifications avant d'être potentiellement ajoutée dans un modèle.

CATÉGORIE PROFESSIONNELLE : Cette variable n'est pas une variable de base. Elle est le résultat d'un travail de regroupement de professions.
Il est clair que la catégorie professionnelle est une information majeure pour la détermination de la fréquence d'arrêt de travail. En effet, il existe des activités professionnelles qui sont davantage exposées à ce risque que d'autres. On pense notamment aux activités manuelles, et/ou physiques qui sont particulièrement exposées à un risque de blessure.

AGE : L'âge est une variable dont il est connu qu'il existe un lien avec l'incidence d'arrêt de travail. En effet, l'exposition au risque à tendance à augmenter avec l'âge. Dans le cadre de cette étude ne seront considérés que les âges faisant partis de la tranche 18-65 ans, qui correspondent respectivement à l'âge minimal et maximal de souscription.

FAMILLE PRODUIT : Cette variable décrit la famille à laquelle appartient le produit d'assurance. Il existe dans cette base un ensemble de trois produits - AP, PPE, et GPE qui ont chacun une particularité qui leur est propre. En effet, le produit AP est un produit forfaitaire, alors que le produit PPE est un produit indemnitaire qui engage l'assureur a compléter le revenu assuré en cas de sinistre : le montant versé s'adapte au montant versé par le régime RO de telle sorte que :

$$engagement_{assureur} = revenu - montant_{regimeRO}$$

Enfin, le produit GPE a été pensé pour être un mixte entre les deux produits précédents. Si le revenu assuré est supérieur à 1 PASS, alors les prestations sont indemnitaires comme le produit PPE, et dans le cas où le revenu est en-deçà de ce seuil, alors les prestations seront fixées à la signature du contrat, comme pour le produit AP. Une seconde particularité du produit GPE est son absence de questionnaire médical. Cette variable est intéressante dans le sens où il parait raisonnable d'émettre l'hypothèse que des distinctions de l'ordre de la conception du produit ont un impact sur la fréquence d'arrêt de travail, et cette étude permettra de vérifier la conjecture.

SITUATION FAMILIALE : Cette variable indique la statut de l'assuré parmi les modalités suivantes : veuf, divorcé, célibataire ou marié.
L'intérêt de cette variable découle d'une hypothèse sur l'impact psychologique que peut avoir sa situation de famille sur l'assuré.

COLLÈGE & LIBELLE STATUT : Ces deux variables ne seront pas utilisées dans leur état de base. Seule l'information concernant le statut de gérant majoritaire de l'assuré sera conservée dans une nouvelle variable à expliquer. En effet, cela est du au fait qu'il est important pour Entoria de mieux comprendre le risque porté par les gérants majoritaires.

GENRE : Les profils de risques des hommes et des femmes ne sont pas les mêmes. En effet, il n'est pas nécessaire de rappeler que l'espérance de vie des femmes est plus forte que celle des hommes, ou que les femmes connaissant contrairement, aux hommes, l'incapacité en cas de grossesse. De plus, certains secteurs d'activités sont principalement occupés par l'un des genres : c'est par exemple le cas dans le secteur du BPT, où la concentration d'hommes est bien plus forte, ou dans le secteur des services à la personne qui comprend une présence largement féminine [16].

FRANCHISE : La franchise correspond au nombre de jours d'attente nécessaires avant versement des indemnités en cas d'incapacité de travail de l'assuré. Étant donné qu'on considère que l'arrêt de travail est "validé" aux yeux de l'assureur à partir de l'expiration de la franchise, cette variable est donc directement liée à l'incidence car une franchise plus élevée diminue le risque d'arrêt de travail avéré.

Ensuite, voici les variables concernant le courtier ayant procédé à la souscription du contrat auprès de l'assuré :

CLUB : Chez Entoria, les courtiers sont catégorisés par rapport à la relation que ces derniers entretiennent avec Entoria. En effet, les courtiers les plus fortement associés profitent de certains avantages non négligeables, c'est pourquoi on peut imaginer que la clientèle associée à ce genre de courtiers n'est pas la même que celle associée à un courtier qui n'est que peu fidèle à Entoria. L'idée est donc que cette différence de population assurée peut potentiellement engendrer une disparité pour le risque arrêt de travail, avec de plus ou moins bons risques.

TYPOLOGIE : La typologie du contrat donne la nature de l'établissement de courtage. Elle précise si le courtier est un indépendant, ou si celui-ci est est filiale mère ou la filiale fille, ou encore si c'est une filiale affiliée. L'information réellement intéressante à retenir pour cette variable est le statut d'indépendant ou non du courtier.

CODE POSTAL DU COURTIER : Cette variable précise le code postal du courtier. Dans le modèle actuellement en production, Entoria dispose d'un zonier qui permet de classer les courtiers par département et utilise cette information pour tarifer. Dans le cadre d'étude de ce mémoire, le choix a été fait de préférer utiliser l'information géographique relative aux assurés plutôt que celle des courtiers.

CLASSE COURTIER : Cette variable est une information qui précise la qualité globale du risque souscrit par le courtier, mesurée par les résultats techniques passés. Ainsi, il est attendu que cette variable soit évidement très corrélée avec l'arrêt de travail.

3.3 Variables externes

Le choix d'un recours à des données externes à Entoria a été motivé par plusieurs facteurs. Dans un premier temps l'objectif est de parvenir à combler un manque de renseignements concernant les assurés. En effet, certaines informations sont manquantes, ou du moins ne sont pas systématiquement enregistrées dans les tables de données de l'entreprise, ce qui pose par conséquent un réel problème car une modélisation statistique ne peut s'entraîner que sur une base de données complète sans informations manquantes. Ainsi, par exemple, à partir du code Siret des entreprises des TNS, il a été possible de rapatrier l'activité principale de l'établissement afin de contrebalancer la très faible présence de la profession au sein des bases. Ensuite, dans une démarche différente, l'ajout de variables extérieures pourrait permettre d'enrichir les données avec des informations plus originales, plus ou moins liées précisément à l'assuré, afin de se rendre compte si des informations moins "personnelles" ou "assurantielles" peuvent participer à l'explication du risque arrêt de travail.

Concernant la pertinence de ces variables externes, il est nécessaire de s'interroger préalablement sur la finesse des variables importées, ainsi que sur la manière dont celles-ci vont pouvoir être jointes aux données internes.

Les données importées proviennent des sites de l'INSEE et d'Etablab, qui sont respectivement une direction générale du ministère de l'Économie et des Finances, et un département de la direction interministérielle du numérique placé sous la direction du Premier ministre.

Concernant la finesse de jointure, les données de la base SIRENE issue du site de l'INSEE sont d'une maille précise car elles concernent directement les établissements correspondants à chacun des assurés : l'information est donc à la maille du TNS. Quant aux données issues de la plateforme d'open data data.gouv.fr d'Etalab, la jointure est faite selon position géographique de l'assuré, et plus précisément selon sa région ou département. Le choix au niveau de la commune aurait permis une maille plus fine, quasiment au niveau personnel de l'assuré. Cependant, les données au niveau du code postal sont trop peu représentées, et les études et analyses statistiques en auraient donc été impactées. C'est pourquoi le choix s'est tourné vers une maille moins sectorisée, plus riche en termes d'observations, qui permet d'avoir une meilleure stabilité au niveau de la modélisation par méthode d'apprentissage.

La décision de conserver uniquement les variables décrivant le département ou la région n'a pas été prise de prime abord. En effet, il paraissait plus adapté de prendre cette décision après avoir mené une étude, que ce soit en termes de statistiques descriptives ou de modélisation, afin de se persuader de retenir la meilleure option. Cependant, à priori, l'information délivrée à l'échelle régionale bénéficie d'un nombre d'observations par modalité plus important qu'à l'échelle départementale ; mais reste à vérifier si cette maille plus large permettra de différencier les profils de risque, et donc vérifier si le changement de modalité est suffisamment tarifant ou non, compte tenu de la perte de précision en moyenne qu'une telle maille entraîne.

Enfin, il est important de noter que, pour des raisons de corrélation évidentes, l'intégration d'une donnée externe dans un modèle d'apprentissage entraînera le retrait de la variable de jointure associée afin de ne pas entacher la qualité du modèle. De plus, avant toute chose, il est nécessaire de s'assurer que l'utilisation des données de l'INSEE pour enrichir la base

interne ait un sens, c'est-à-dire vérifier que la population décrite par ces données externes est similaire à la population interne. Dans le cas de ce mémoire, il s'agit donc de confirmer le fait que le portefeuille d'assurés étudié est représentatif de la population française en termes de zone géographique (étant donné que cette dernière est le lien entre les données).

Tout d'abord, voici les variables de la base SIRENE.
La jointure avec la base Sirène a été effectuée à l'aide du code SIRET de l'établissement de l'assuré. Grâce à celui-ci, il a été possible de rapatrier diverses informations concernant l'entreprise.

APE : Activité principale de l'établissement : L'intitulé exacte de l'activité n'est pas présent dans la base SIRENE, mais un code de référence permet de récupérer l'information dans la nomenclature en vigueur, la Naf Rév2, recensant la liste complète des activités. L'intérêt de cette variable est qu'elle apporte beaucoup de détails sur l'activité de l'entreprise de l'assuré. Ainsi, celle-ci pourrait potentiellement être plus intéressante que la variable actuelle de profession, ou alors servir à construire une nouvelle variable relative à l'activité professionnelle qui serait davantage prédictive.

EFFECTIF ÉTABLISSEMENT : Cette variable donne une précision sur le nombre d'employés détenus par l'entreprise de l'assuré. L'information est donnée par un code qui désigne une tranche de salariés (ex : "10 à 19 salariés"). Ces données sont basées sur l'année 2016.

CARACTÈRE EMPLOYEUR : Caractère employeur de l'établissement : Cette variable indique si l'établissement est en prise d'emploi ou non.

Variable de la source data.gouv :
PRIX MOYEN DES HABITATIONS - par département et par région : Ces variables renseignent respectivement les prix moyens des habitations par m^2 par département et par région.

TEMPÉRATURE MIN, MAX, MOY PAR REGION : La base de données contient un relevé des températures journalières par région depuis l'année 2016. Pour des raisons pratiques d'utilisation, il a été décidé de ne retenir qu'une valeur moyenne, une valeur maximale, et une valeur minimale sur la totalité de la temporalité donnée. Ainsi, à chaque région est associée trois températures qui les caractérisent.

AIRE D'APPARTENANCE : Cette base de données contient des informations, au niveau du code postal, sur les aires des villes auxquelles les communes appartiennent : elle décrit le nombre de personnes habitant dans la zone de la commune, la catégorie de la commune (commune-centre, pôle principal, pôle secondaire, etc), etc. Ces données sont basées sur l'année 2017.
Malgré le fait que ces données possédaient un bon potentiel pour expliquer une part de l'incidence de l'arrêt de travail, il n'a malheureusement pas été possible de joindre ces données à la base de données interne d'Entoria. En effet, étant donné la maille communale des données, la richesse de la base interne n'était pas suffisante et un grand nombre d'observations ne pouvaient pas obtenir l'information souhaitée. Ainsi, cette variable est inutilisable dans cette étude.

DENSITÉ DE POPULATION : Cette variable donne la densité de population au km^2 agrégée au département et à la région. C'est le rapport entre l'effectif de la population d'une zone géographique et la superficie de cette zone. Le résultat s'exprime en nombre d'habitants par kilomètre carré.

DENSITÉ MÉDICALE : La densité médicale est le ratio qui rapporte les effectifs de médecins à la population d'un territoire donné. Ici, l'information est importée aux mailles département et région. La densité médicale s'exprime en nombre de médecins pour 100 000 habitants.

DENSITÉ MÉDICALE MÉDECINE GÉNÉRALE : Cette donnée est la même mais ne représente que la part des médecins généralistes.

REVENU MÉDIAN : Cette variable renseigne le revenu médian de la zone géographique associée. Elle permet de décrire la population présente dans cette zone.

TAUX DE PAUVRETÉ : Cette variable indique la proportion d'individus dont le revenu est inférieur à 60% du revenu médian dans la zone concernée.

TAUX DE LOGEMENTS SOCIAUX : Le taux de logements sociaux est le nombre de logements sociaux rapporté au nombre de résidences principales.

TAUX DE LOGEMENTS INDIVIDUELS : Le taux de logements individuels est le nombre de logements individuels rapporté au nombre de résidences principales.

PART DE TRANSPORT EN COMMUN : L'indicateur représente la part des actifs se déplaçant principalement en transports en commun pour aller travailler, selon leur département de résidence. Les indicateurs des parts des déplacements domicile-travail permettent de décrire les comportements et de suivre leur évolution au fil du temps, de les mettre en relation avec les politiques mises en œuvre au niveau national et local.

Chapitre 4

Analyse descriptive des données

4.1 Analyse des variables explicatives et lien avec la variable sinistre

L'objectif de cette partie est de présenter les données et d'identifier les variables dont l'analyse bivariée indique que le pouvoir prédictif de ces variables est significatif.

Cette section présente une liste non-exhaustive des retraitements, en mettant en avant les variables qui se démarquent le plus.

Le portefeuille d'étude étant composé exclusivement de Travailleurs Non Salariés, celui-ci se trouve sans surprise être majoritairement constitué d'hommes, à hauteur de 73% comme l'atteste la figure 4.1 ci-dessous :

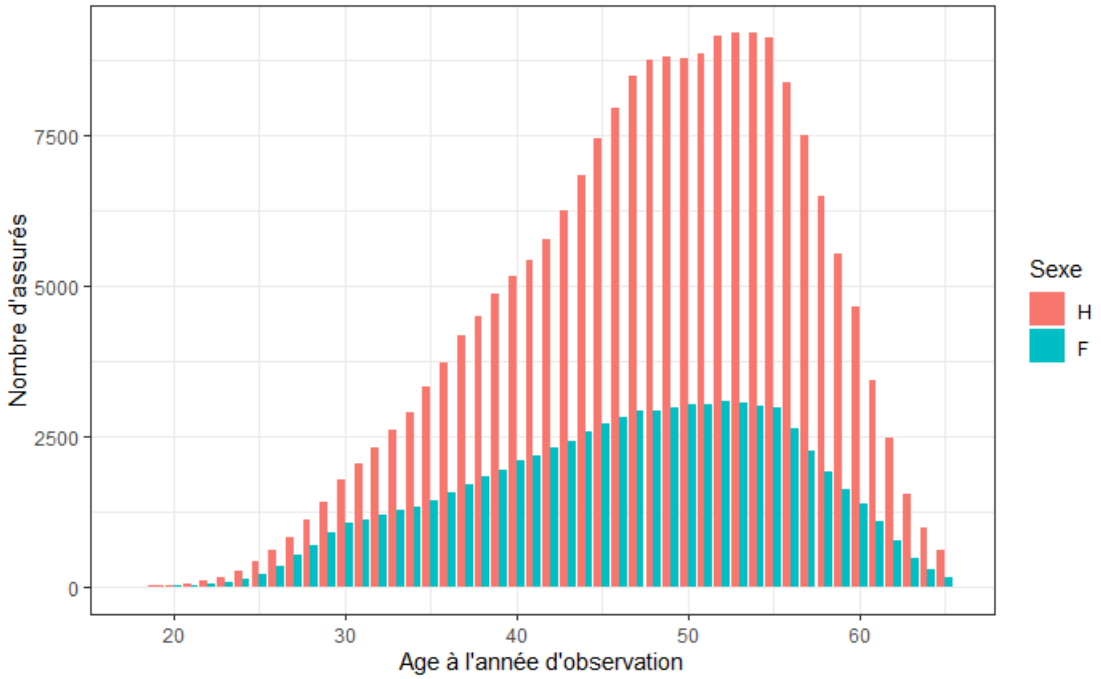

FIGURE 4.1 – Étude de la répartition Hommes/Femmes dans le portefeuille

Le taux d'incidence, i.e. le pourcentage de contrat qui bénéficie d'une indemnisation au titre d'un arrêt de travail, est légèrement supérieur à 6% par année d'exposition sur la plage de 2016 à 2021. Il s'agit de taux d'incidence moyen, compte tenu des différentes franchises appliquées. La figure 4.2 illustre une hausse tendancielle entre 2016 et 2019 qui s'explique par la commercialisation d'un produit prévoyance sans sélection médicale de 2017 à 2019, ainsi qu'un pic du taux d'incidence annuel en 2020, suivi d'un retour dans la moyenne des deux années précédentes en 2021. Ce pic peut en partie s'expliquer par la vague d'arrêt de travail ayant eu lieu en mars 2020 durant la période sanitaire de la COVID19 que la France a connu cette année là.

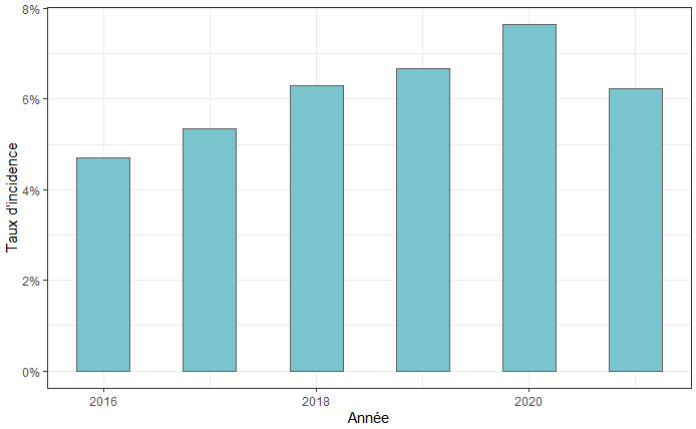

FIGURE 4.2 – Évolution du taux d'entrée en AT marquant un pic d'incidences en 2020

4.1.1 Discrétisation des variables quantitatives

Conserver les variables continues permet de ne pas perdre en finesse lors de la modélisation, car les prédictions sont davantage individualisées en fonction de la caractéristique exacte de chaque individu. Cependant, une telle finesse nécessite une quantité significative et une homogénéité des données sur l'ensemble des valeurs prises par les variables continues. Le choix a été fait ici de discrétiser les variables quantitatives qui suivent.

Étude de la variable relative à l'âge
Deux méthodologies se démarquent pour effectuer une classification. En effet, il est possible de créer des tranches d'âges fixes (exemple : par tranche de 5 ans), mais il est également possible d'étudier la sinistralité par âge afin d'en déduire des regroupements. La seconde solution a été retenue afin de privilégier des regroupements homogènes des risques par âge.

Ainsi, l'évolution de l'âge de l'assuré ayant une influence sur sa fréquence de sinistres, comme le souligne le 1^{er} graphique de la figure 4.3 dont la ligne horizontale symbolise un changement de comportement potentiel vis-à-vis du risque, il pourrait s'avérer judicieux de discrétiser la variable d'âge afin d'obtenir une variable qualitative à trois modalités qui rassemblent les individus ayant un comportement semblable : ainsi est créée la variable de classe d'âge dont la sinistralité des groupes est celle du 2^{eme} graphique.

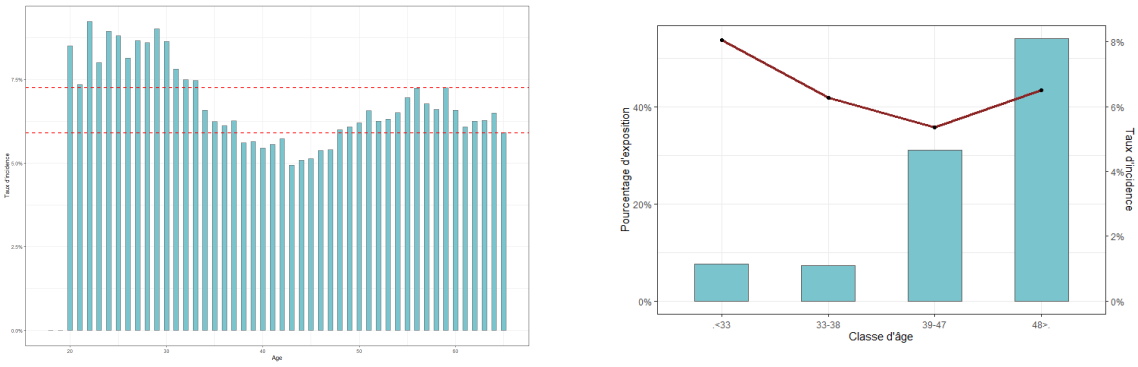

FIGURE 4.3 – Représentation des taux d'incidence de sinistres par âge et pour les classes d'âge

Étude de la variable relative au revenu annuel

La majorité des assurés du portefeuille se situe dans la tranche salariale de 0.5 à 1.5 PASS. Par ailleurs, la population ayant un revenu inférieur à 1 PASS connaît une sinistralité en arrêt nettement plus forte que celle des assurés ayant un revenu annuel supérieur. Ainsi, la variable des revenus est discrétisée afin d'obtenir une variable qualitative composée de deux modalités.

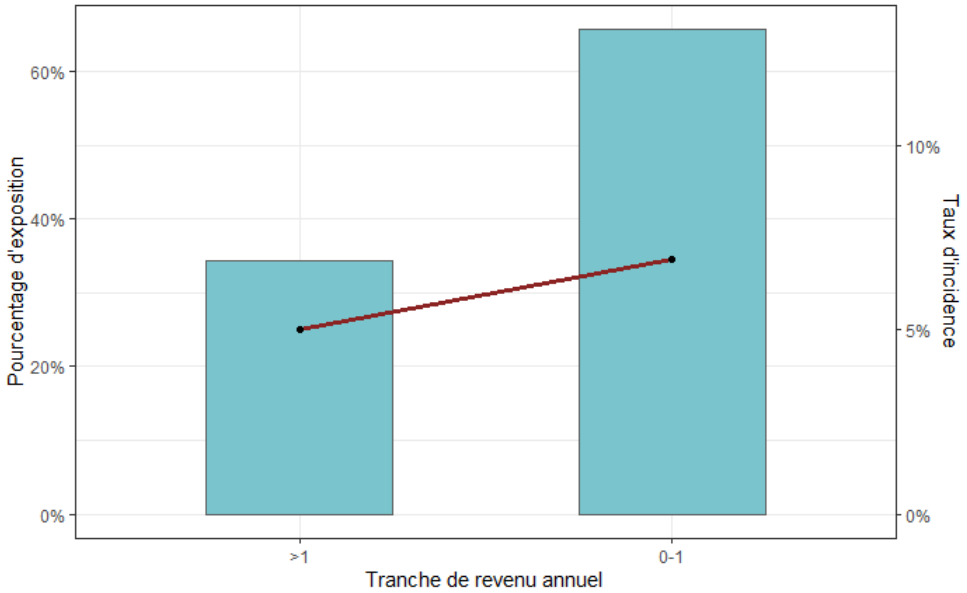

FIGURE 4.4 – Représentation des taux d'incidence de sinistres par tranche de revenu

4.1.2 Regroupement des variables qualitatives

Catégorie professionnelle

L'information concernant la profession des assurés n'est renseignée qu'à hauteur de 40%. Ce faible pourcentage est expliqué par le fait que cette information n'est demandée à la souscription que depuis mars 2017. Certains secteurs d'activités sont naturellement plus exposés au risque arrêt de travail que d'autres, comme par exemple :

— les métiers manuels pour lesquels les assurés sont davantage exposés à un risque de blessures ou d'accident de travail ;
— des professions, comme les auxiliaires médicaux, soumises à des pressions psychosociales fortes.

Il est donc nécessaire d'effectuer un retraitement de la donnée afin de combler ce manque. Il va de soi qu'il est impossible de retracer la profession exacte étant donné que la maille de la profession est particulièrement fine, mais l'idée est de créer des classes de catégories professionnelles dont les expositions au risque arrêt de travail sont significativement différentes.

Ainsi, le regroupement de catégories professionnelles actuellement utilisé par le modèle de prédiction du produit prévoyance est construit par l'utilisation conjointe des informations concernant le régime obligatoire et la ligne tarifaire associée à chaque assuré. Cependant, ce regroupement est loin d'être suffisant puisque certaines classes sont sous représentées, avec moins de 1000 assurés adhérants, ce qui ne permet pas de tirer des conclusions tangibles concernant leur sinistralité, comme le souligne la figure 4.5. De plus, par manque de données internes plus fines permettant de mieux sectoriser ce secteur d'activité, la classe *Artisan Commerçant* représente près de 43% du portefeuille. Pourtant, il est probable que chacun des personas contenus dans cette classe n'est pas exposé au même risque d'arrêt de travail. Dès lors, la conclusion est de nouveau la même : il est nécessaire d'aller plus loin dans le retraitement de cette variable.

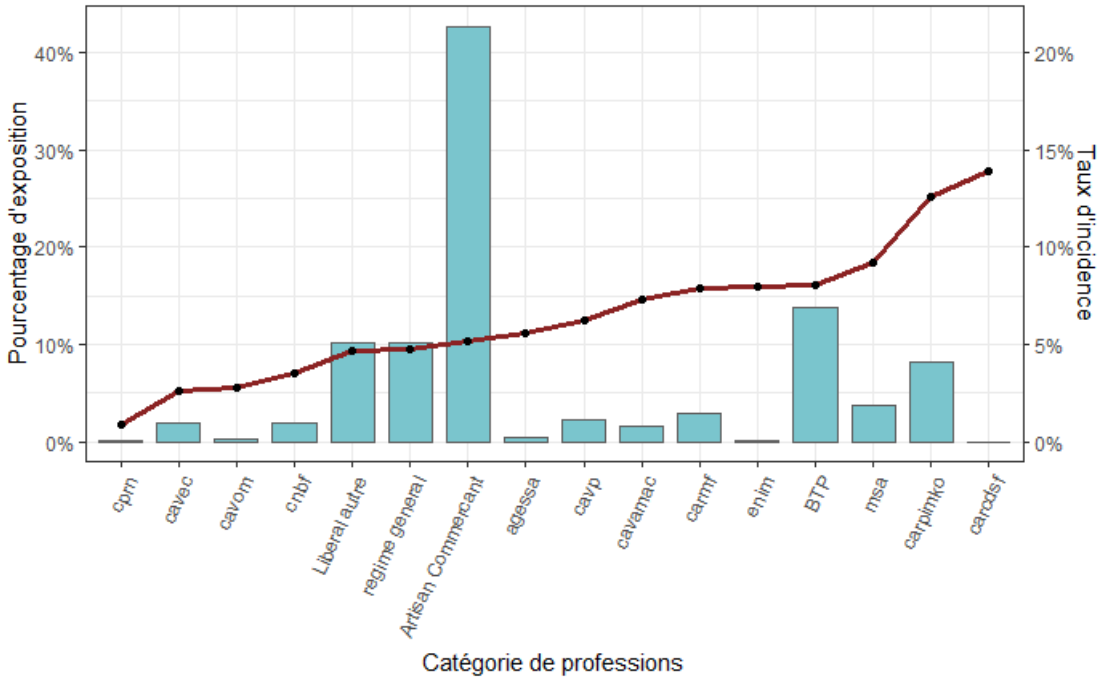

FIGURE 4.5 – Graphique mettant en lumière la présence de classes nécessitant un retraitement

Avant de détailler davantage les étapes suivantes, il est important de noter que les données externes importées de la base SIRENE sont incomplètes. En effet, une quantité importante des identifiants SIRET ne sont pas renseignés dans les bases internes de l'entreprise. Cependant, dans ces bases internes sont stockés, encore une fois de manière incomplète, les codes APE permettant de récupérer l'Activité Principale de l'Établissement grâce à

la nomenclature d'Activités Françaises fournie par l'INSEE. Ainsi, en joignant ces deux variables, il est possible de trouver l'activité principale d'environ 60% de la population totale. Il est clair qu'un manque de 40% des données implique une impossibilité d'utiliser cette variable directement dans une modélisation. Néanmoins, ces quelques lignes ont su se montrer utiles dans le retraitement de l'activité professionnelle. En effet, l'activité professionnelle, la profession, ou encore l'activité principale sont intimement liées, regrouper ces informations ensemble afin de créer une nouvelle variable plus complète fait donc sens.

Le retraitement de la variable catégorie professionnelle se déroule selon trois grandes étapes.

Tout d'abord, il s'agit d'isoler les artisans commerçants, et de conserver inchangées les autres classes de professions. Sont utilisés ici non pas les intitulés exactes des activités principales de l'entreprise de l'assuré, mais plutôt la famille d'activités à laquelle appartient cette APE. Huit familles d'activités ressortent, les autres n'étant pas suffisamment représentées pour être prises en compte. A ce stade, la classe des artisans commerçants a été divisée en huit groupes, auxquels s'ajoute un dernier groupe composé des 15% de lignes restantes qui n'ont pas pu recevoir d'APE (que ce soit pour raison de sous-représentation, ou tout simplement pas absence de donnée).
Cette perte d'information n'étant pas négligeable, la manoeuvre est une seconde fois effectuée, mais en ayant recours cette fois-ci à la variable *secteur d'activité*, déterminée par la profession de l'assuré, qui permet de combler une part du manque, en menant le proportion de lignes non renseignées à 6% des artisans commerçants.

Une fois cette étape passée, la classe des artisans commerçants a été segmentée. Le second point de retraitement consiste à fusionner les autres modalités de la variable de profession dont l'exposition a été jugée trop faible : cinq classes sont regroupées pour n'en former qu'une seule avec davantage d'exposition.

Enfin, l'objectif de la troisième et dernière étape est d'étudier la significativité de cette variable *Nouvelle catégorie professionnelle*. En effet, comme dans la suite de ce mémoire, une attention particulière sera portée sur le fait de rendre significative chacune des modalités des variables explicatives qualitatives. En effet, ceci est un critère important concernant la qualité des variables utilisées, et concernant la qualité prédictive du modèle. Par ailleurs, il est sous-entendu ici qu'une modalité est significative si sa p-value est sous le seuil des 5%, puisqu'aucune théorie n'atteste d'un gain de performance notoire entre une forte (i.e. $< 5\%$) et une très forte (i.e. $< 1\%$) présomption de rejet de l'hypothèse de nullité du coefficient. Ainsi, les regroupements effectués soulignent la décision de préférer regrouper toutes les activités ayant un comportement similaire vis-à-vis de l'arrêt de travail, quitte à rassembler des secteurs d'activité qui n'ont pas de liens directs entre eux : cette vision sert un résultat sectorisé et non un résultat mutualisé.
La figure 4.6 illustre les modifications apportées en termes de répartition et d'incidence, tout en soulignant en rouge l'éclatement de la classe *Artisan Commerçant* qui a été effectué : on constate une nette amélioration dans l'homogénéité de la distribution de la population dans les classes de risque.
Pour aller plus loin, la figure 4.6 illustre que le risque porté par certains profils d' artisans commerçants est le même que celui d'autre catégories professionnelles. Ainsi, il est tout à fait possible d'imaginer un raisonnement similaire pour découper d'autres classes possédant une population suffisante en plusieurs sous-groupes dont le profil de risque est différent, afin de bousculer une nouvelle fois la répartition des classes avec une vision toujours plus

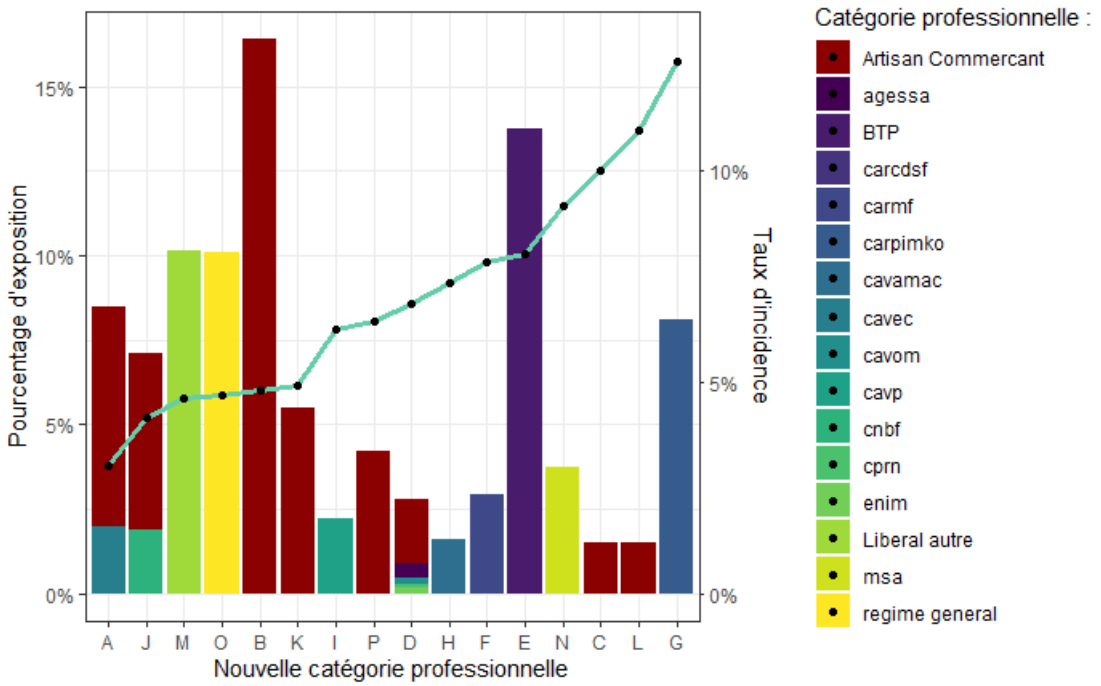

FIGURE 4.6 – Répartition et fréquence des nouvelles catégories professionnelles

tournée vers un regroupement par risques semblables. Malheureusement, cette idée ne peut être exploitée dans cette étude étant donné la proportion trop importante d'observations ne comportant pas l'information SIRET.

Franchise

La franchise est par construction un élément lié à la fréquence d'arrêt de travail. En effet,la probabilité d'entrer en arrêt de travail augmente lorsque la franchise diminue.

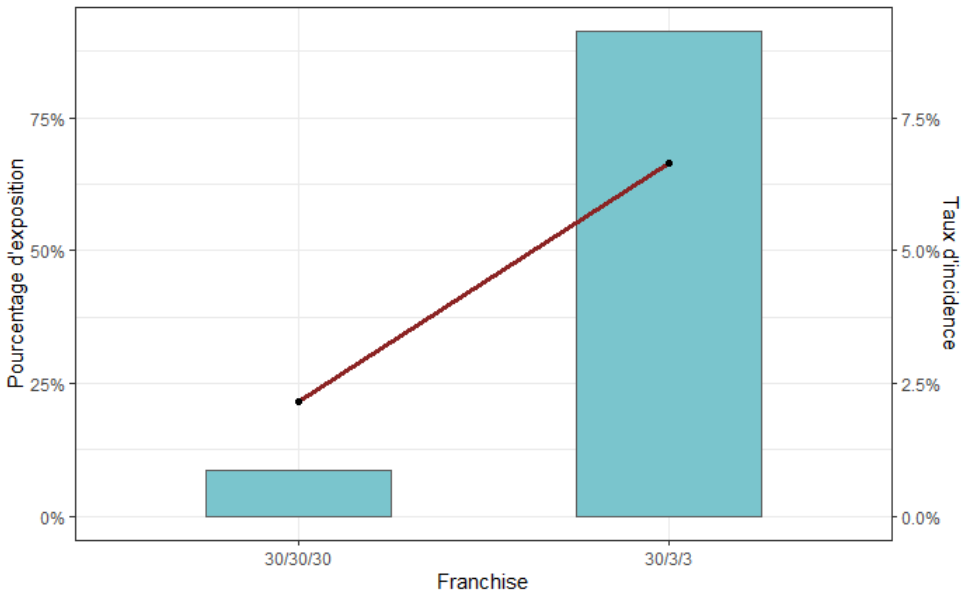

FIGURE 4.7 – Sinistralité par franchise

Famille du produit

Comme cela a déjà été expliqué, le portefeuille de clientèle est séparé en trois familles de produits de prévoyance. Le produit PPE, qui représente un peu moins de 60% du portefeuille, connaît une sinistralité plus faible que celle des autres produits. De son côté, le produit GPE possède à priori davantage de mauvais risques en termes de fréquence. Il n'y a priori pas de surprise à cela, puisque le produit GPE était dénué de questionnaire médical, ce qui explique une présence plus forte de mauvais risques, et par implication d'un plus fort taux d'incidence de sinistres.

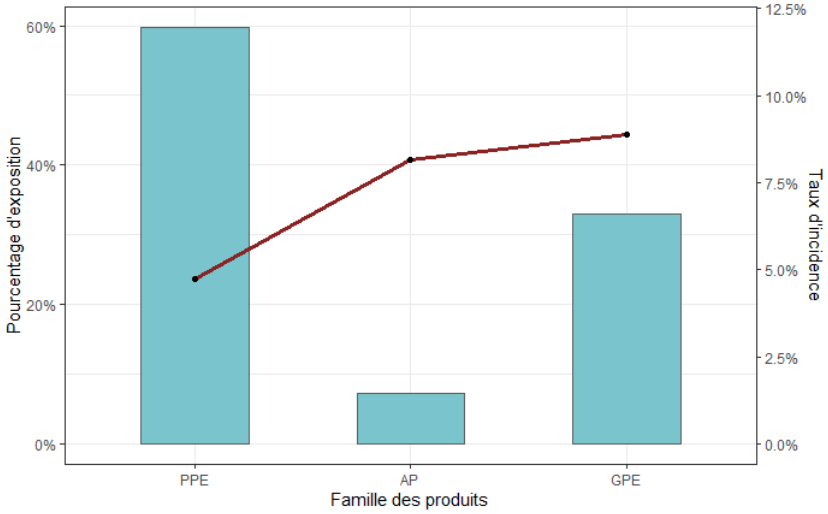

FIGURE 4.8 – Répartition et sinistralité de la base d'assurés parmi les trois familles de produits de prévoyance

Sélection médicale

Comme exprimée dans le paragraphe précédent, l'idée première qui ressort d'une sélection médicale assouplie est la possibilité d'attirer davantage de mauvais risques, étant donné l'accès simplifié au produit de prévoyance. La figure 4.9 confirme cette idée. En effet, les assurés ayant souscrit en remplissant un questionnaire médicale classique, le DSS/QM, ont une sinistralité légèrement en deçà de celle des assurés qui ont répondu à un questionnaire simplifié, le DAT. Enfin, les contrats ayant une absence de questionnaire médical se démarquent en ayant un taux d'incidence de sinistres plus fort.

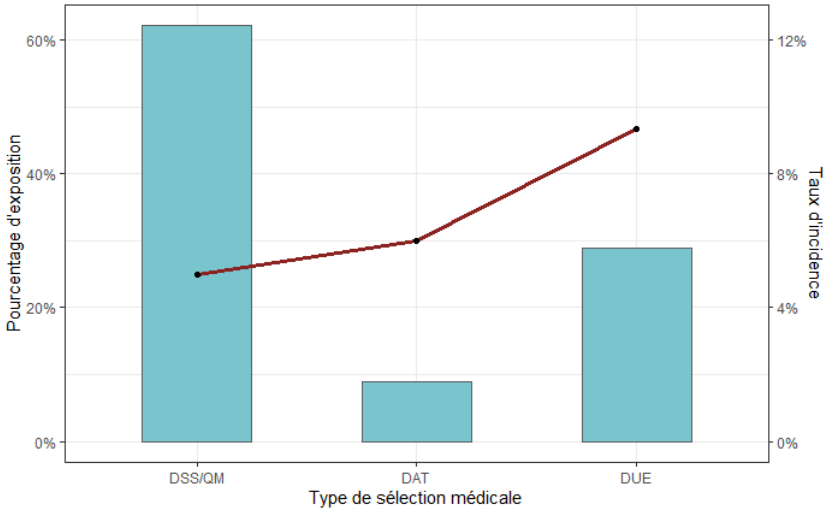

FIGURE 4.9 – Répartition et sinistralité en fonction du type de sélection médicale du produit de prévoyance

Gérant majoritaire

D'après les données de la base, les gérants majoritaires sont moins fréquemment en arrêt de travail. Cette différence peut s'expliquer par le fait que le gérant majoritaire doit effectuer davantage de tâches administratives. Ainsi, les missions quotidiennes du gérant majoritaire sont globalement moins exposées à un risque d'accident.

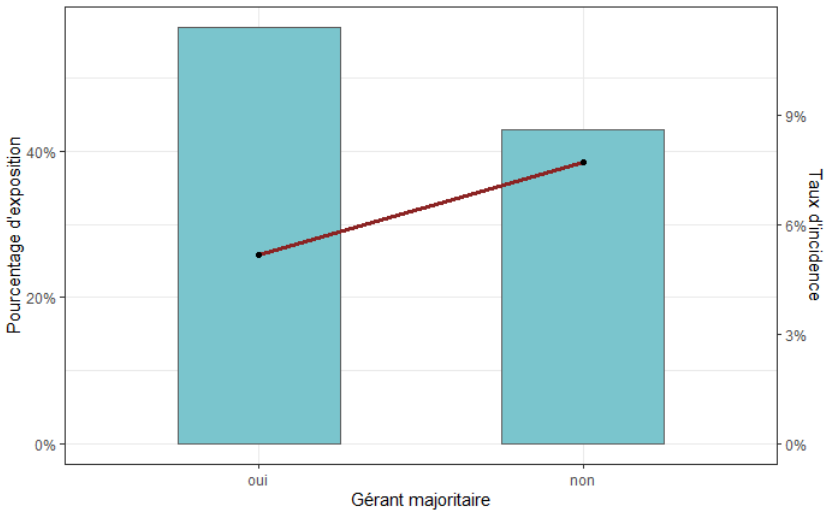

FIGURE 4.10 – Répartition et sinistralité des TNS gérants majoritaires

Zonier actuel
Le zonier actuel est appliqué sur la situation géographique du courtier ayant souscrit le contrat de l'assuré. Celui-ci permet d'expliquer une part de l'incidence d'arrêt de travail, cependant il est utilisé et n'a pas été mis à jour depuis plusieurs années. Un autre point important de ce mémoire est de mettre au point un nouveau zonier à partir des résidus du modèle de fréquence afin de challenger le modèle actuel. La finalité étant de conserver le meilleur zonier, c'est-à-dire celui qui possède le plus grand pouvoir prédictif selon les indicateurs de performance qui seront considérés. Le graphique 4.11 ci-après démontre de plus une présence forte, aux alentours des 60%, de la population du portefeuille dans l'une des classes, ce qui soutient l'idée qu'il est important de vérifier si une amélioration n'est pas envisageable.

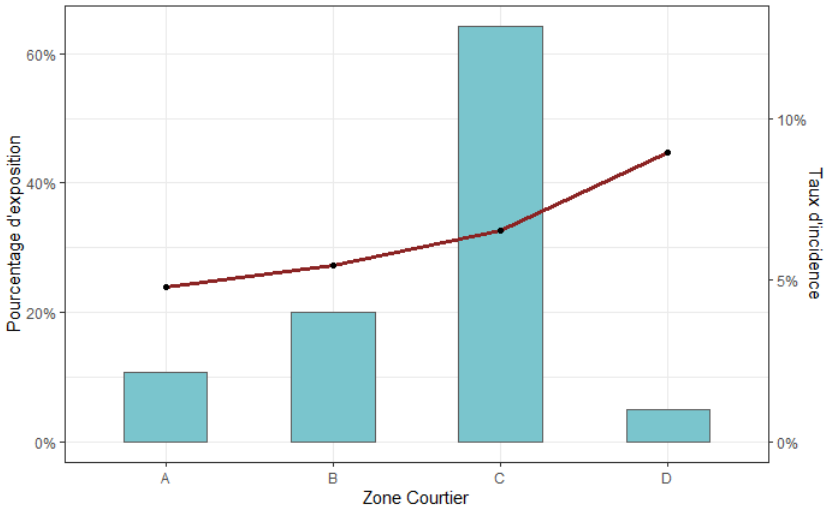

FIGURE 4.11 – Répartition des assurés au sein des différentes zones

4.2 Échantillonnage

Pour le bon déroulement des modélisations qui vont être effectuées, une base d'apprentissage, nommée *Train*, et une base de validation, nommée *Test* sont créées. La méthode de création est assez classique, il suffit de diviser la base de données complète en deux bases distinctes, dont une base dite "Train" qui contient 80% des observations, et une base dite de "Test" qui contient les 20% restants.

Il est nécessaire d'être vigilant sur le fait que les bases d'apprentissage et de validation soient effectivement comparables, c'est à dire que pour chaque variable présente dans la base de test, toutes les modalités sont aussi présentes dans la base d'apprentissage. En effet, dans la situation où cette condition ne serait pas vérifiée, le modèle serait dans l'incapacité de prédire les observations concernées par l'absence d'entraînement.
La figure 4.12 confirme que les deux échantillons de travail possèdent une population d'assurés distribuée d'une manière équivalente.

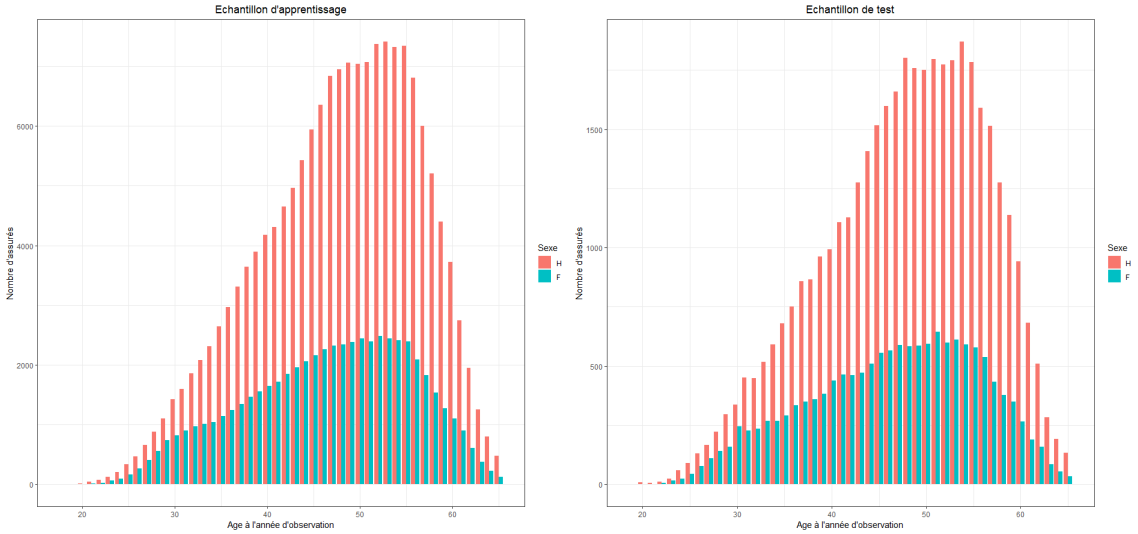

FIGURE 4.12 – Répartition de la population entre les bases d'apprentissage et de validation

Troisième partie

Modélisation statistique et Machine learning

Chapitre 5

Les indicateurs de performances

Il existe plusieurs indicateurs de mesure de performance d'un modèle. Par continuité, ces indicateurs peuvent aussi servir à comparer deux modèles entre eux.

5.1 Le critère AIC - Akaike Information Criterion

Cet indicateur permet de mesurer la qualité d'ajustement du modèle, c'est-à-dire au sens de l'adéquation aux données. Le critère de comparaison entre deux modèles estime qu'un modèle qui diminue la mesure AIC est le plus performant. Cette mesure prend en compte l'erreur d'entraînement du modèle et la pénalise par le nombre de paramètres.

Soit $L_n(Y, \beta)$ la log-vraisemblance du modèle, avec $\widehat{\beta}$ l'estimateur du paramètre β qui maximise la vraisemblance du modèle et p le nombre de paramètres, alors la valeur de l'indice est :

$$AIC = -2.L_n(Y, \widehat{\beta}) + 2.p$$

Le principe étant de minimiser l'AIC, il découle donc de cette formule qu'il est question de dosage entre l'apport d'informations du modèle (par la log-vraisemblance) et le nombre de paramètre du modèle. En effet, la mesure AIC est davantage pénalisée quand le nombre de paramètres augmente.

5.2 Le critère BIC - Bayesian Information Criterion

Ce critère se différencie du critère AIC par le fait qu'il pénalise davantage le nombre de paramètres par la taille de la base de données. La formule devient :

$$BIC = -2.L_n(Y, \widehat{\beta}) + log(n).p$$

où n est le nombre d'observations de la base.

De la même manière, et pour les mêmes raisons, la comparaison entre deux modèles donnera raison au modèle qui minimisera le BIC.

5.3 Les indicateurs d'écarts

Les métriques présentées dans cette partie sont relatives à des écarts entre les valeurs cibles et les prédictions. Ainsi, plus leurs valeurs sont élevées, moins le modèle est performant.

5.3.1 MAE - Moyenne Absolue des Erreurs

Le MAE est sûrement la métrique de régression la plus interprétable. Néanmoins, le recours à la valeur absolue des erreurs crée un manque de régularité.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i \breve{} \hat{y}_i|$$

Cet indicateur donne une idée de la qualité globale des prédictions en considérant les petits écarts et les grands écarts de la même façon, mais ne permet cependant pas de savoir si le modèle à tendance à sous ou sur-estimer les prédictions.

5.3.2 MSE et RMSE

Le MSE, Mean Square Error (moyenne des carrés des résidus en français), est la moyenne arithmétique des carrés des écarts entre les prévisions du modèle et les observations de l'échantillon.

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

où $\hat{y}_i$ est l'estimation de y_i, avec Y la variable à estimer.

Contrairement au MSE, le RMSE s'exprime dans la même unité que la variable cible, et est par conséquent plus facile à interpréter.

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

Contrairement au MAE, ces deux métriques pénalisent davantage les grandes erreurs que les petites. En effet, le recours au carré des erreurs amplifie l'impact d'une erreur plus élevée sur la moyenne. Cette propriété leur permet d'être sensibles aux valeurs aberrantes qui s'écartent des valeurs moyennes.
Ainsi, ces deux métriques sont particulièrement utiles pour éviter de faire de grandes erreurs de prédiction.

5.3.3 MAPE - Mean Absolute Percentage Error

La MAPE est une métrique de régression utilisée lorsque les erreurs du modèle sont considérées en proportion de la valeur prédite. L'idée de cette mesure est de ne pas considérer à la même hauteur l'effet d'une observation faible face à une observation plus forte.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i \breve{} \hat{y}_i}{y_i}\right|$$

La MAPE s'exprime en pourcentage d'écart entre la prédiction et la valeur réelle, ce qui la rend particulièrement facile à interpréter. Cependant, dans le cadre de cette étude, son utilisation est impossible car, par définition, elle impose la condition de non nullité des valeurs réelles.

5.4 L'indice de Gini

L'indice de Gini est un indicateur mesurant la segmentation du modèle, ce qui en fait un indicateur de performance.

A l'origine, le coefficient de Gini est un indicateur qui a été développé pour mesurer l'inégalité des revenus de la population dans un pays. L'indice de Gini est lié à la courbe de Lorenz, sur laquelle est représentée en abscisse la part cumulée de population et en ordonnée la part cumulée des richesses détenues. Selon, l'INSEE, l'indice de Gini est un indicateur synthétique d'inégalités de revenus. Ce coefficient varie entre 0 et 1. Il est égal à 0 dans une situation d'égalité parfaite où tous les revenus seraient égaux. À l'autre extrême, il est égal à 1 dans la situation la plus inégalitaire possible, celle où tous les revenus seraient nuls, à l'exception d'un seul.

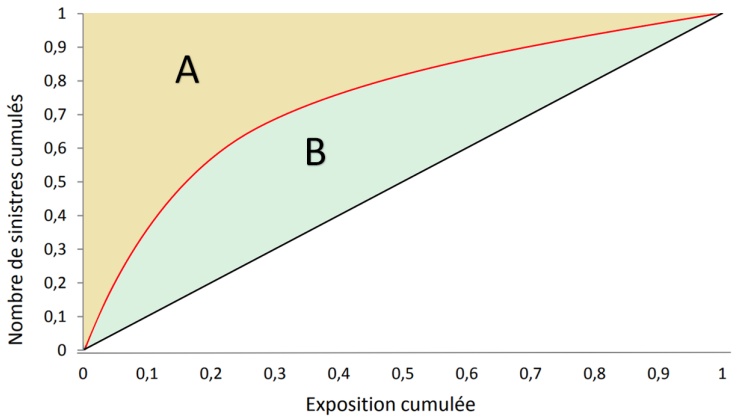

FIGURE 5.1 – Courbe de Lorenz

Sur la figure 5.1, le coefficient de Gini est défini par la part de l'aire au-dessus de la $1^{ère}$ bissectrice occupée par l'aire sous la courbe de Lorenz. En considérant les aires A et B introduites par la figure, le coefficient de Gini est défini par la formule suivante :

$$Gini = \frac{B}{A+B} = \frac{B}{0.5} = 2B$$

Son utilisation peut être étendue aux modélisations statistiques étudiées dans ce mémoire. Pour cela, il est question de modifier les données représentées par les axes de la courbe de Lorenz. Après avoir trié les données par ordre croissant en fonction de nombre de sinistres sera représenté :

— en abscisse, la part cumulée de l'exposition totale ;
— en ordonnée, la part cumulée des sinistres.

De plus, l'indice de Gini ne vaut plus 1 lorsque 1% de l'exposition contient 100% des sinistres. En effet, le maximum est atteint avec le modèle saturé qui prédit parfaitement les sinistres historiques. Ces sinistres historiques sont représentés sur la figure la figure 5.2 par la courbe bleue : d'où le terme "Gini normalisé" qui induit un changement de référence.

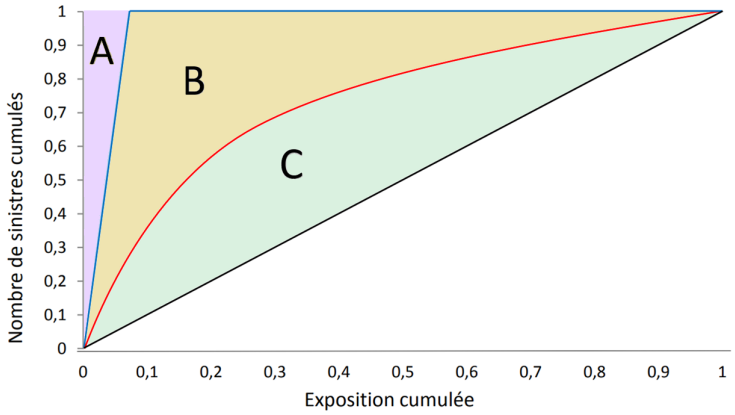

FIGURE 5.2 – Représentation des aires mises en jeu dans le calcul du coefficient de Gini normalisé

La droite noire est la droite représentant la situation de parfaite égalité où chaque point d'exposition contient la même part du risque total : la situation de mutualisation parfaite. Dans la situation réelle du portefeuille de cette étude, moins de 7% de l'exposition contient la totalité des sinistres, comme représenté par la courbe bleue. La dernière courbe, la rouge, représente l'estimation des sinistres par le modèle considéré.

Le coefficient de Gini est égal à l'aire entre la droite de mutualisation parfaite et la courbe de Lorenz, i.e. l'aire C. En effet, l'idée est que lorsque la courbe rouge des prédictions du modèle tend vers la courbe bleue des données observées, alors l'indice de Gini tend vers 1.

Ainsi, l'indice de Gini du modèle saturé est l'aire sous la courbe bleue :

$$Gini_{saturé} = \frac{B+C}{A+B+C} = 2.(B+C)$$

L'indice de Gini du modèle considéré est :

$$Gini_{modèle} = \frac{C}{A+B+C} = 2C$$

D'où la déduction du coefficient de Gini normalisé, qui est la part de l'aire sous la courbe des sinistres occupée par l'aire sous la courbe de Lorenz :

$$Gini_{normalisé} = \frac{Gini_{modèle}}{Gini_{saturé}} = \frac{C}{B+C}$$

Ce coefficient peut à la fois servir à vérifier si l'ajout ou le retrait d'une variable explicative est pertinent, et à comparer la segmentation de deux modèles : la segmentation du modèle s'améliore lorsque l'indice de Gini s'approche de 1.

Une dernière remarque importante, et qui est visuellement représentée sur la courbe de Lorenz, est que la sinistralité prédite représentée par la courbe rouge mutualise les risques. Cette mutualisation diminue lorsque la courbe se rapproche de la courbe des sinistres réels.

5.5 Sélection et limites des indicateurs retenus

Dans le cadre de cette étude, tous ces indicateurs ne seront pas utilisés. Les mesures et comparaisons de performance seront faites en s'appuyant principalement sur le coefficient de Gini, le critère AIC (pour les modèles GLM), et sur le RMSE. Le MAE sera aussi pris en compte mais dans une moindre mesure.

Il est tout de même important de notifier les limites de ces indicateurs.
Le MAE et le RMSE se basent sur des calculs prenant en compte les résidus du modèle. Cependant, dans le cas de prédiction de fréquence d'arrêt de travail les prédictions seront très petites et cela implique que les différences d'écarts se liront jusqu'à cinq chiffres après la virgule : ce qui peut sérieusement remettre en question la pertinence de conclusions tirées d'un écart si faible.
L'AIC, de son côté, ne prend pas forcément en compte la performance de prédiction réelle du modèle. En effet, la valeur de l'indice AIC n'est pas un score qui juge la qualité intrinsèque du modèle. Ce score permet uniquement de comparer l'ajustement de deux modèles : si les deux modèles comparés sont mal ajustés, ce critère ne permet pas de s'en rendre compte.
Enfin, les lacunes de l'indice de Gini sont dues au fait qu'il mesure la performance par une vision "globale". En effet, deux distributions peuvent être estimées avoir un même niveau d'inégalité mesuré par le coefficient de Gini, alors que la segmentation n'est pas la même.Cela se traduit graphiquement par le fait que plusieurs courbes de Lorenz peuvent avoir la même aire, comme par exemple les courbes rouge et verte de la figure 5.3.

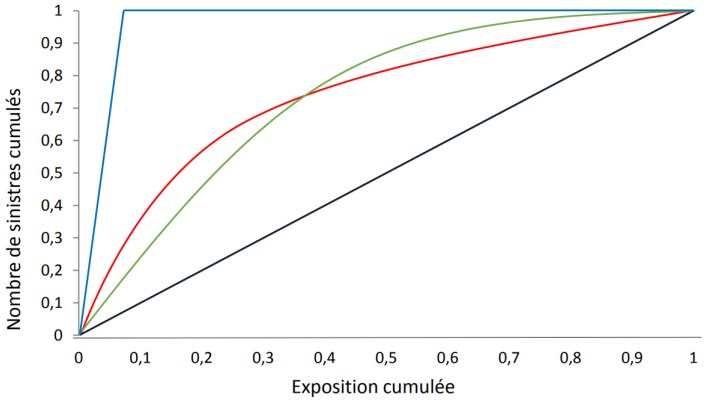

FIGURE 5.3 – Deux courbes de Lorenz ayant le même coefficient de Gini

De plus, il est important de se rendre compte que la variable cible est composée de valeurs discrètes allant de 0 à 4, alors que les prédictions en sortie du GLM ne sont pas des entiers. En effet, le choix de sortie s'est porté sur un taux d'incidence pour des raisons métier, puisque ces taux ont l'ambition d'être implémentés et approchés par des taux d'incidence. Ceci explique les faibles valeurs de l'indice de Gini qui seront rencontrées lors des phases de comparaison, puisque la courbe de Lorenz associée ne pourra jamais vraiment s'écarter de la $1^{ère}$ bissectrice.
Il faudra donc avoir une vision critique vis-à-vis de ces mesures, et surtout avoir une utilisation combinée de ces différents indicateurs pour s'assurer d'un maximum de pertinence pour les conclusions tirées.

Chapitre 6

Les modèles linéaires généralisés - GLM

6.1 Présentation théorique

Les modèles linéaires généralisés sont, comme l'indique leur nom, une extension des modèles linéaires gaussiens. Cette extension a été nécessaire étant donné que le modèle linéaire gaussien n'est pas adapté lorsque la variable à expliquer ne suit pas une loi gaussienne.

6.1.1 Le modèle linéaire gaussien

Voici une brève présentation des modèles linéaires, qui sert de préliminaires avant d'introduire les notions relatives aux GLM (Generalized Linear Models).

Dans tout ce chapitre sont considérées n observations et p variables explicatives. Soit Y la variable à expliquer et $X_1, ..., X_p$ les variables explicatives.
Ce modèle suppose que la variable à expliquer est une combinaison linéaire des variables explicatives multipliées par les paramètres, telle que :

$$Y = X.\beta + u$$

Avec :

- Y est une matrice de taille n x 1 ;
- X est une matrice de taille n x $(p+1)$ dont la $1^{ère}$ colonne est uniquement composée de 1 afin de créer le profil de référence ;
- β est le vecteur des paramètres inconnus à estimer de taille $(p+1)$ x 1 ;
- u est le vecteur des résidus de taille n x 1, qui correspondent à l'écart entre les prédictions et les valeurs réelles.

Autrement dit,

$$\forall i \in [1, ..., n], \quad Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j.X_{i,j} + u_i$$

La moyenne des résidus étant nulle par hypothèse :

$$\forall i \in [1, ..., n], \quad E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j . X_{i,j}$$

Le modèle linéaire estime β par la méthode des moindres carrés ordinaire afin d'obtenir la meilleure prédiction de $\hat{Y}$. Les hypothèses posées par cette approche sont les suivantes :

- la valeur moyenne du terme d'erreur est nulle ;
- homoscédasticité des résidus : $\forall i \in [1, ..., n]$, $Var(u_i) = \sigma^2$;
- absence d'autocorrélation des erreurs ;
- covariance nulle entre u et X ;
- nombre d'observations supérieur au nombre de paramètres ;
- les X sont bornés dans leur ensemble ;
- absence de colinéarité parfaite.

Le modèle est dit "gaussien" quand les erreurs suivent une loi normale $\mathcal{N}(0, \sigma^2)$.

6.1.2 Le modèle linéaire généralisé

Généralités

Un modèle linéaire généralisé est caractérisé par ces trois composantes :

- une composante aléatoire Y, qui est la variable à expliquer dont les observations sont considérées indépendantes et suivant une loi de probabilité dérivant d'une structure exponentielle ;
- une composante déterministe qui est une combinaison linéaire des variables explicatives ;
- une fonction de lien, notée g qui décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_1, ..., X_p$ et l'espérance de la variable cible.

Dans un modèle GLM, l'objectif est de déterminer les valeurs des coefficients $\beta_0, ..., \beta_p$ tels que :

$$\forall i \in [1, ..., n], \quad g(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j . X_{i,j} = X_i^t . \beta$$

Il est commun de poser η, dit le "prédicteur linéaire" comme :

$$\forall i \in [1, ..., n], \quad \eta_i = X_i^t . \beta$$

La différence avec le modèle linéaire classique se trouve dans le fait qu'il s'agit ici d'une modélisation de la transformation de la variable cible par la fonction lien.

Hypothèses sous-jacentes

Pour appliquer une modélisation par GLM, la loi de probabilité de la réponse Y doit appartenir à la famille exponentielle, ce qui signifie que sa densité doit s'écrire de la manière suivante :

$$f_{\theta,\phi}(x) = exp\left(\frac{x.\theta - b(\theta}{a(\phi)} + c(x, \phi)\right)$$

avec :

— $a(.)$ et $c(.)$ des fonctions dérivables sur $\mathbb{R}$;
— $b(.)$ est est une fonction de classe C^1 et de dérivée inversible ;
— θ est le paramètre appelé "naturel" :
— ϕ est le paramètre de dispersion.

Deux propriétés fondamentales découlent de la forme de cette écriture :

$$E[Y] = b'(\theta) = \mu$$

$$V[Y] = b''(\theta).a(\phi)$$

Cette condition sur l'appartenance à cette famille est moins forte que celle du modèle linéaire classique puisqu'elle contient non seulement la loi normale, mais aussi un large panel de lois courantes comme :

— la loi Binomiale ;
— la loi Gamma ;
— la loi de Poisson ;
— la loi Inverse Gaussienne.

Ainsi, en pratique, le choix est porté sur la loi qui est la plus proche de la structure des données d'étude. Par exemple, dans le cas de ce mémoire, l'étude porte sur le phénomène de fréquence d'un événement, et donc sur une loi de comptage : la loi de Poisson peut à priori être considérée comme celle qui sera la plus adaptée.

Pour s'assurer que la loi de Poisson fait effectivement partie de cette famille exponentielle, il faut vérifier que la densité d'une loi de Poisson de paramètre λ satisfait le critère :

$$f_\lambda(k) = \frac{\lambda^k.e^{-\lambda}}{k!} = e^{(k.ln(\lambda)-\lambda-ln(k!)} = exp\left(\frac{k.\theta - e^\theta}{a(\phi)} - ln(k!)\right) = exp\left(\frac{k.\theta - b(\theta)}{a(\phi)} + c(k,\phi)\right)$$

En ayant posé :

— $\theta = ln(\lambda)$;
— $a(\phi) = 1$;
— $b(\theta) = e^\theta$;
— $c(k,\phi) = c(k) = -ln(k!)$

De manières équivalentes, il est possible de démontrer l'appartenance des autres lois à la famille exponentielle.

Fonction de lien

La fonction de lien représente le lien entre la moyenne de la variable cible, et le prédicteur linéaire η. Pour chaque loi appartenant à la famille exponentielle, il existe une fonction de lien dit "canonique". Ces fonctions, ainsi que paramètres associés aux lois usuelles, sont résumés dans la figure 6.1 :

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$	$c(y_i, \phi)$
Normale(μ_i ;σ^2)	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\}$
Poisson(μ_i)	$\log(\mu_i)$	1	ϕ	$\exp(\theta_i)$	$-\log y!$
Binomiale$\frac{1}{m_i}$(m_i ;μ_i)	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	ϕ	$\log(1+\exp\theta_i)$	$\log\binom{m_i}{m_i y_i}$
Gamma(μ_i ;α)	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-\log(-\theta)$	$\alpha\log(\alpha y) - \log y - \log\Gamma(\alpha)$
Inverse Gaussienne(μ_i ;σ^2)	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$

FIGURE 6.1 – Valeurs des paramètres et liens canoniques des lois usuelles

D'après la ligne de la loi normale, la fonction lien est la fonction identité, ce qui est cohérent puisque cela revient au modèle linéaire gaussien avec la moyenne de la variable cible qui est égale au prédicteur linéaire.

Le recours à une fonction lien de type logarithmique offre un confort dans l'application du modèle. En effet, la fonction inverse du logarithme étant la fonction exponentielle, cela permet d'obtenir un modèle multiplicatif simple à lire, à interpréter, et à calibrer.

$$\forall i \in [1, ..., n], \quad g(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j . X_{i,j}$$

$$\Leftrightarrow E[Y_i] = e^{\beta_0} . \prod_{j=1}^{p} e^{\beta_j . X_{i,j}}$$

Estimation des paramètres

Une fois la loi de distribution de la variable à expliquer identifiée, les variables explicatives sélectionnées pour créer le prédicteur linéaire, et la fonction de lien choisie, il est question d'estimer les paramètres $\beta_0, ..., \beta_p$ présentés précédemment. Pour cela, plusieurs approches sont possibles, mais ici ne sera présentée que la méthode par maximum de vraisemblance, qui est la méthode principalement suivie pour une modélisation linéaire généralisée.

Par hypothèse, les réalisations de la variable à expliquer Y sont indépendantes, la vraisemblance du modèle est donc définie par la densité conjointe :

$$L_n(\beta_1, ..., \beta_p) = f_{\theta,\phi}(y_1, ...y_p) = \prod_{i=1}^{n} f_{\theta,\phi}(y_i)$$

Les paramètres, qui paraissent absents à première vue, sont en réalité contenus dans chacune des densités par le relation $b'(\theta) = g^{-1}(X.\beta)$ qui a été présentée dans la précédente sous-section.

Ainsi, l'idée est de déterminer les paramètres $\beta_1, ..., \beta_p$ qui permettent de maximiser la vraisemblance du modèle.

Il est commun de préférer considérer la log-vraisemblance en lieu et place de la vraisemblance. En effet, la fonction logarithme est une fonction strictement croissante, la vrai-

semblance et la log-vraisemblance sont maximisées par les mêmes valeurs des estimateurs. L'avantage de cette variante est de faciliter les calculs puisque le produit des densités devient une somme :

$$ln(L_n(\beta_1, ..., \beta_p)) = ln(\prod_{i=1}^{p} f_{\theta,\phi}(y_i)) = \sum_{i=1}^{p} ln(f_{\theta,\phi}(y_i))) = \sum_{i=1}^{p} \left[\frac{y_i.\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right]$$

Afin de déterminer les estimateurs $\hat{\beta}_1, ..., \hat{\beta}_p$ en lesquels la vraisemblance est minimisée, il faut calculer les dérivées partielles pour chaque paramètre, et trouver les points en lesquels elles s'annulent. Pour $j \in [1, ..., p]$, la dérivée partielle s'écrit :

$$\frac{\partial ln(L_n)}{\partial \beta_j} = \sum_{i=1}^{p} \frac{\partial}{\partial \beta_j} \left[\frac{y_i.\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right]$$

De manière générale, aucune formule exacte ne permet de déterminer les estimateurs qui annulent ces dérivées, et donc qui maximisent la vraisemblance. C'est pourquoi la recherche de ces points d'annulation se fait par méthode itérative. Peut être citée la méthode de Newton-Raphson qui est un algorithme efficace pour trouver numériquement une approximation précise d'un zéro. L'étude des méthodes permettant de résoudre ce problème d'optimisation ne sera pas davantage détaillée dans ce mémoire.

6.2 Modèle de fréquence : estimation de l'incidence des sinistres arrêt de travail

6.2.1 Identification de la loi des sinistres :

La figure 6.2 permet d'apprécier la forme poissonienne de la fréquence des sinistres du portefeuille. Les données représentées en bleu sur le graphique sont les incidences d'une loi de Poisson de paramètre le nombre de sinistres moyen par année. Ainsi, et tout comme c'est le cas pour le modèle actuellement utilisé par Entoria, la modélisation des sinistres est faite par un modèle linéaire généralisé de loi de poisson.

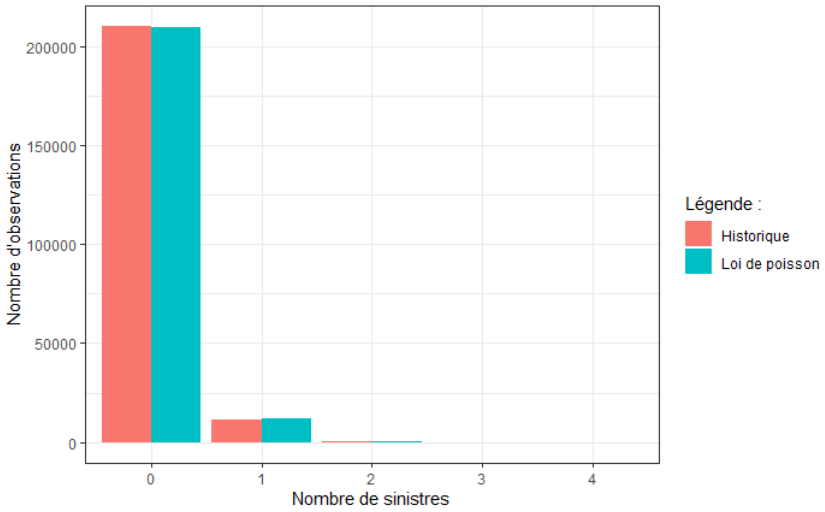

FIGURE 6.2 – La distribution des sinistres du portefeuille (en rouge) est sensiblement la même que la distribution d'une loi de Poisson de paramètre la moyenne des sinistres du portefeuille (en bleu)

6.2.2 Modélisation et significativité des variables explicatives :

Le panel de variables est composé de variables catégorielles. Ainsi, afin de créer un assuré de référence ayant un sens métier, chacun des niveaux de référence associé aux variables correspond à la modalité la plus représentée dans le portefeuille. Une fois le traitement de l'ensemble des variables explicatives terminé, il s'agit d'écarter toutes celles dont l'étude descriptive et graphique permet de soutenir un manque de colinéarité avec la variable de sinistres. A l'issue de cette analyse préliminaire est défini le terme "modèle complet". Ce modèle complet désigne le modèle composé de l'ensemble des variables explicatives qui semblent pertinentes aux yeux des analyses, i.e. toutes celles n'étant pas écartées lors de ce premier tri qui pourrait être qualifié de "tri qualitatif". Cependant, une variable peut sembler avoir un potentiel significatif intéressant sans que ce ne soit le cas dans la réalité de la modélisation. C'est pourquoi il est nécessaire d'effectuer une seconde étape de tri, qui sera cette fois-ci effectuée avec davantage de précision et de critères quantitatifs. Le recours à une procédure de sélection automatique permet de conserver uniquement les variables les plus pertinentes pour expliquer l'incidence de sinistre arrêt de travail. Le critère de sélection suit une méthode *forward* par critère AIC. L'idée est que la machine débute ses calculs d'AIC en partant du modèle nul (i.e. composé d'aucune variable explicative, c'est le modèle moyen), puis exécute une itération de tests afin de déterminer la variable qui minimise le plus l'AIC. Cette étape d'ajout est effectuée jusqu'à atteindre le modèle complet (i.e. composé de toutes les variables explicatives à disposition), ou jusqu'à atteindre un modèle dont l'ajout d'une variable supplémentaire ne peut que diminuer le score AIC. Ainsi, la procédure aboutit au modèle censé être le plus performant au sens de l'AIC. Toutefois ceci n'est qu'une première étape, le facteur humain reste important, c'est pourquoi d'autres optimisations du modèle et des variables sont effectuées dans la suite de cette étude.

A l'issue de la sélection forward, dix variables sont retenues dans cet ordre d'importance (au sens de la procédure décrite ci-dessus), dans le modèle :

— Nouvelle catégorie professionnelle
— Famille du produit

— Franchise du contrat
— Classe d'âge
— Classe du courtier
— Type de sélection médicale
— Sexe de l'assuré
— Gérant majoritaire
— Revenu
— Indépendance du courtier

Afin de comparer des modèles comparables, puisque le modèle construit jusqu'alors ne contient pas encore de variable capable de capter des informations géographiques, il est nécessaire de prendre comme référence le modèle actuel auquel a été retirée la composante géographique. L'étude des zones géographiques est traitée à posteriori dans la suite de ce mémoire.

La figure 6.3 présente les indicateurs de performance appliqués aux deux modèles considérés. En l'état actuel des choses, le modèle nouvellement créé s'en sort mieux sur l'ensemble des critères de comparaison, ceci est valable pour l'étape d'entraînement et de validation. L'AIC semble indiquer une amélioration de l'ajustement du modèle, tandis que l'augmentation du coefficient de Gini annonce une meilleure segmentation. L'indicateur d'écart des résidus souligne une très légère diminution des écarts de prédiction.

	AIC		Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation	Entraînement	Validation
GLM Actuel	97 319	24 160	0,299108	0,298791	0,246765	0,244811
GLM forward	96 700	23 985	0,328302	0,327551	0,246391	0,244322

FIGURE 6.3 – Comparaison du modèle construit par procédure forward avec le modèle actuel

Les résultats de la modélisation sont présentés par la figure 6.4 ci-après. Les modalités de variables non présentes dans ce tableau sont comprises dans l'intercept du modèle, c'est à dire que ce sont les modalités de référence pour chacune des variables. Ce tableau expose à la fois les estimations des paramètres, les valeurs de la statistique de test pour chacune des variables, ainsi que les p-value associées. Le test effectué est un test de Wald qui a pour but de vérifier si l'hypothèse H_0 de nullité du coefficient est rejetée ou non. La p-value est définie comme $P[X^2(1) > X^2_{wald}]$: c'est la probabilité qu'une variable aléatoire de loi de Khi-deux à 1 degré de liberté dépasse la valeur observée X^2_{wald} de la statistique de test, qui elle-même suit cette loi avec le même degré de liberté sous l'hypothèse H_0.

Variable	Paramètre estimé	Khi-2 wald	p-value	Significativité
New_Cat_Prof[T.A]	-0.39169	-7,871	3.51e-15	***
New_Cat_Prof[T.C]	0.19622	2,801	0.005093	**
New_Cat_Prof[T.D]	0.13761	2,43	0.015108	*
New_Cat_Prof[T.E]	0.34520	10,18	< 2e-16	***
New_Cat_Prof[T.F]	0.26272	4,585	4.54e-06	***
New_Cat_Prof[T.G]	0.50676	12,302	< 2e-16	***
New_Cat_Prof[T.H]	0.29990	4,325	1.53e-05	***
New_Cat_Prof[T.I]	0.16413	2,468	0.013597	*
New_Cat_Prof[T.J]	-0.14178	-3,029	0.002451	**
New_Cat_Prof[T.K]	-0.05746	-1,182	0.237008	
New_Cat_Prof[T.L]	0.22357	3,34	0.000838	***
New_Cat_Prof[T.M]	-0.18343	-4,226	2.38e-05	***
New_Cat_Prof[T.N]	0.30946	6,44	1.19e-10	***
New_Cat_Prof[T.O]	-0.02889	-0,693	0.488508	
New_Cat_Prof[T.P]	0.15918	3,203	0.001361	**
Famille_Produit[T.ap]	0.26137	7,799	6.24e-15	***
Famille_Produit[T.gpe]	0.05068	1,043	0.296727	
Franchise[T.30/30/30]	-0.88897	-17,25	< 2e-16	***
ClasseAge[T.33-38]	-0.08207	-1,844	0.065169	.
ClasseAge[T.39-47]	-0.07204	-2,097	0.036030	*
ClasseAge[T.48-]	0.20541	6,285	3.28e-10	***
ClasseCourtier[T.1]	-0.23693	-11,042	< 2e-16	***
Select_Med[T.DAT]	0.11118	3,392	0.000693	***
Select_Med[T.DUE]	0.39836	8,257	< 2e-16	***
Sexe[T.F]	0.19694	9,101	< 2e-16	***
GerantMaj[T.oui]	-0.18518	-7,954	1.80e-15	***
Revenu[T.>1]	-0.10896	-4,966	6.84e-07	***
Independant[T.INDEP]	-0.04017	-1,957	0.050386	.

FIGURE 6.4 – Significativité et interprétation des résultats du modèle construit par procédure forward

La p-value permet de trancher sur la confiance que l'on accorde à la certitude de pouvoir intégrer la variable ou non dans le modèle. A titre indicatif, pour un seuil à $\alpha = 5\%$, vérifier que la p-value est inférieure à 5% est équivalent à vérifier si la statistique de test est inférieure à X^2_{wald}. Lorsque cette inégalité est vérifiée, cela signifie qu'il y a au plus 5% de chance pour que l'hypothèse H_0 de nullité du coefficient soit rejetée alors que ce n'est pas le cas en réalité. Ainsi, le fait de ne pas rejeter l'hypothèse signifie que les chances pour que le paramètre soit en réalité nul sont trop importantes. Sur la figure 6.4, les modalités ne vérifiant pas cette condition de significativité sont surlignées afin de les mettre en évidence.

Dans cette étude, un point d'intérêt particulier est porté sur la significativité de l'ensemble des modalités d'une variable, en gage de robustesse du modèle. Même si une significativité la plus forte possible est davantage appréciée, une modalité est jugée significative si sa p-value est inférieure à 5% (une étoile sur le tableau de la figure 6.4). Toutefois, dans le cas de variables qualitatives à plus de deux modalités, il est toléré une significativité au seuil de 10% sur l'une des modalités.

A partir des paramètres estimés, on peut interpréter l'effet qu'à une variable significative sur la variable cible. Le principe est le suivant : un coefficient significativement supérieur

à zéro indique que le facteur associé favorise la survenance d'un sinistre d'arrêt de travail, tandis qu'un coefficient significativement négatif indique que la modalité connaît une survenance plus faible. Cet effet d'augmentation ou de diminution du risque est considéré comparativement au profil de référence (i.e. celui pour lequel toutes les modalités du tableau de la figure 6.4 valent 0). Par exemple, selon la modélisation de la figure 6.4, un contrat dont la franchise est 30/30/30 est moins susceptible de connaître une survenance de sinistre qu'un contrat 30/03/03 : ce qui correspond bien à l'analyse qui avait été faite concernant la variable franchise dans le chapitre précédent. A cela peut être ajouté que la grandeur en valeur absolue de l'estimation du paramètre donne une précision sur la puissance de l'effet de la modalité : plus la valeur absolue de la valeur est élevée, plus l'influence de la variable sur l'incidence est forte, que ce soit positivement ou négativement.

6.2.3 Corrections apportées au modèle :

Nouvelle catégorie professionnelle :
Afin de rendre la variable *Nouvelle Catégorie professionnelle* significative selon chacune de ses modalités, un travail sur la conception de cette variable est à effectuer. Pour cela, des regroupements on été réalisés afin de lisser la courbe des taux d'incidence, tout en limitant au maximum le nombre de fusions de classes afin de conserver un panel de classes large. La forme finale des classes professionnelles, dont les modalités ont été renommées par ordre de risque croissant, est illustrée par la figure 6.5.

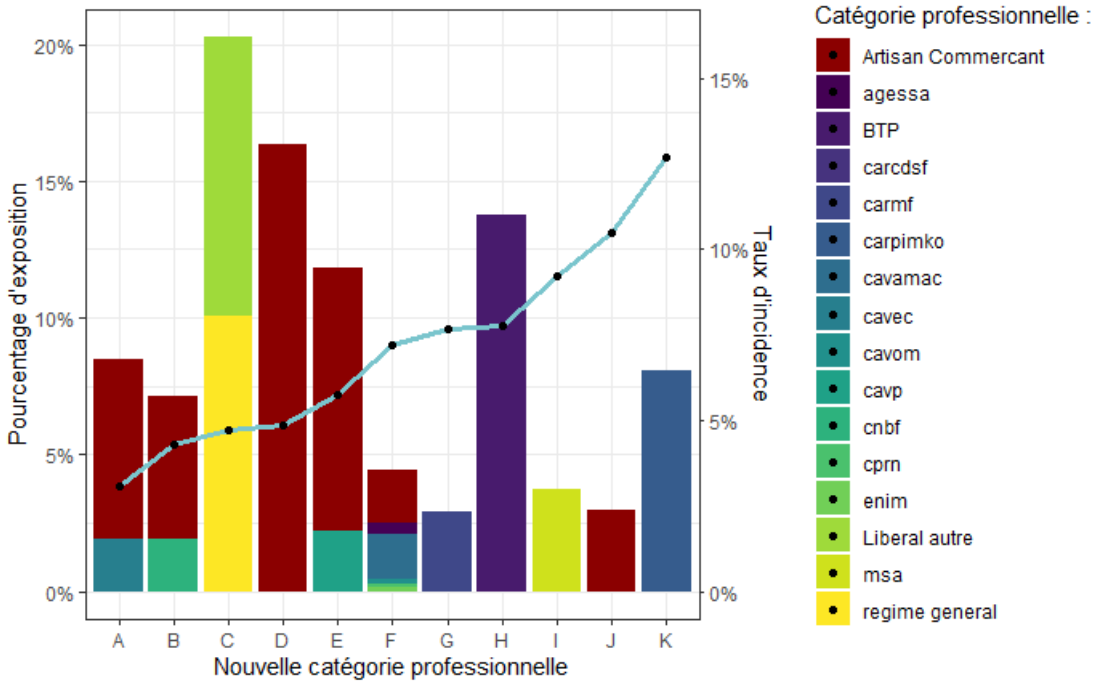

FIGURE 6.5 – Détails de la répartition des artisans commerçants dans la dernière version de la variable catégorie professionnelle

Indépendance du courtier :
Cette variable étant une variable binaire dont l'une des deux modalités n'est pas significative, elle est écartée du modèle afin de respecter la démarche choisie concernant la significativité des variables.

Résultats et performance :
Un aperçu des améliorations obtenues sur la variable de catégories professionnelles à l'issue des modifications effectuées est retranscrit dans la figure 6.6. A ce stade, le variable dispose déjà d'une certaine robustesse construite sur un ensemble de variables explicatives jugées totalement significatives par les tests considérés dans cette étude.

Variable	p-value	Significativité
New_Cat_Prof[T.A]	2.54e-15	***
New_Cat_Prof[T.J]	7.42e-05	***
New_Cat_Prof[T.F]	2.92e-05	***
New_Cat_Prof[T.H]	< 2e-16	***
New_Cat_Prof[T.G]	4.79e-06	***
New_Cat_Prof[T.K]	< 2e-16	***
New_Cat_Prof[T.E]	0.059707	.
New_Cat_Prof[T.B]	0.002336	**
New_Cat_Prof[T.C]	0.002146	**
New_Cat_Prof[T.I]	7.51e-11	***

FIGURE 6.6 – Significativité des modalités de la variable *Nouvelle catégorie professionnelle*

Malgré le gain de significativité observé, les indicateurs de performance de la figure 6.7 ne semblent pas indiquer d'amélioration du modèle. Cela n'est pas nécessairement signe d'une régression des performances, il faut aussi se rappeler que ces indicateurs ne captent pas toute l'information, et que le fait d'agir sur le nombre de modalités impacte directement le calcul de l'AIC par exemple. Ainsi, même si les indicateurs donnent l'impression d'indiquer un recul très léger de performance, il a été préféré de conserver ces modifications et continuer l'étude avec des variables significatives sur l'ensemble de leurs modalités.

	AIC		Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation	Entraînement	Validation
GLM forward	96 700	23 985	0,328302	0,327551	0,246391	0,244322
GLM retraitement 1	96 723	23 999	0,326863	0,326181	0,246414	0,244367

FIGURE 6.7 – Indicateurs de performance

A ce stade, il nécessaire de s'assurer que le modèle ne souffre pas de surdispersion, auquel cas il serait question de se diriger vers une modélisation quasipoisson ou binomiale-négative. Pour cela, un test portant sur la déviance et les degrés de liberté du modèle est effectué. Il en ressort que $\frac{Déviance}{nddl} > 1$. Ainsi, aucun effet de surdispersion n'est à noter. Cependant, afin de s'assurer que c'est bien le cas, un second test est réalisé. Ce dernier consiste à représenter les estimateurs de la modélisation effectuée (i.e. régression de Poisson) jusqu'ici

avec les estimateurs d'une régression Quasipoisson. Le graphique de la figure 6.8 permet de constater qu'aucune différence significative n'existe entre les deux modélisations.

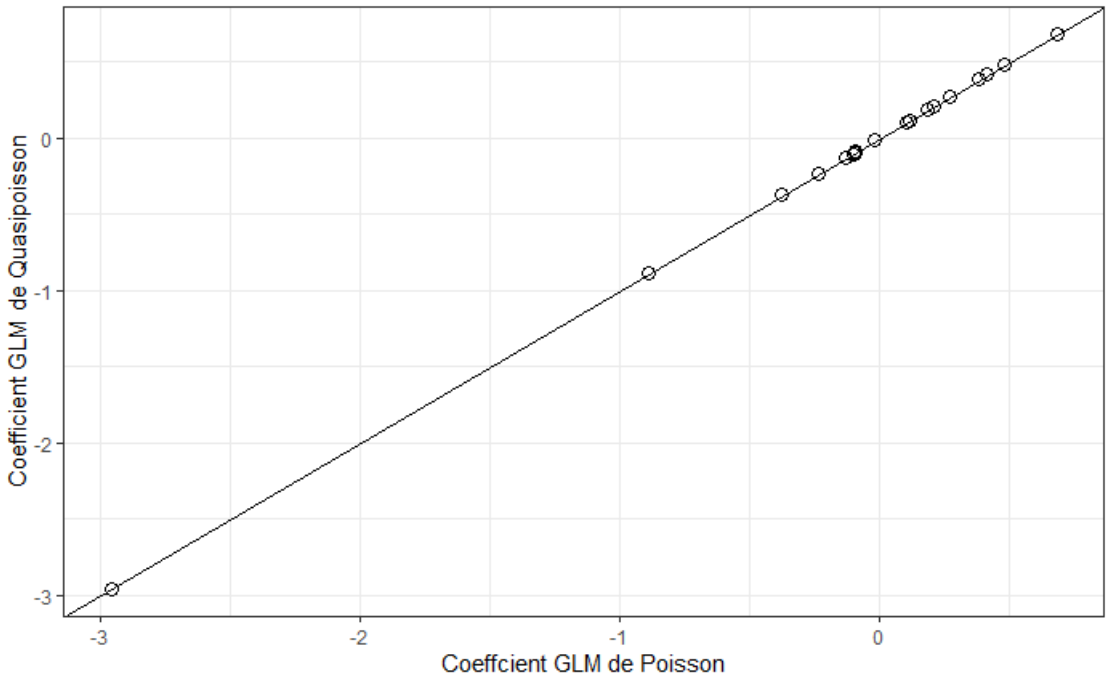

FIGURE 6.8 – Comparaison des coefficients par régression de Poisson et Quasipoisson

6.2.4 Matrice des corrélations

Lors de la conception d'un modèle, l'étude de corrélation entre les variables explicatives permet de faire un tri et de ne conserver qu'une unique variable pour expliquer une information. En effet, des écarts peuvent être créés dans le cas ou des variables très corrélées apparaissent dans un même modèle.

Dans le cas de cette étude, les variables sélectionnées dans le modèle sont toutes des variables qualitatives. De ce fait, un recours au V de Cramer permet d'étudier la force de corrélation entre deux variables. La figure 6.9 exprime les V de Cramer calculés entre toutes les variables significatives. Une relation est considérée forte lorsqu'elle dépasse un certain seuil. Dans cette étude, ce seuil est fixé à $0,5$.

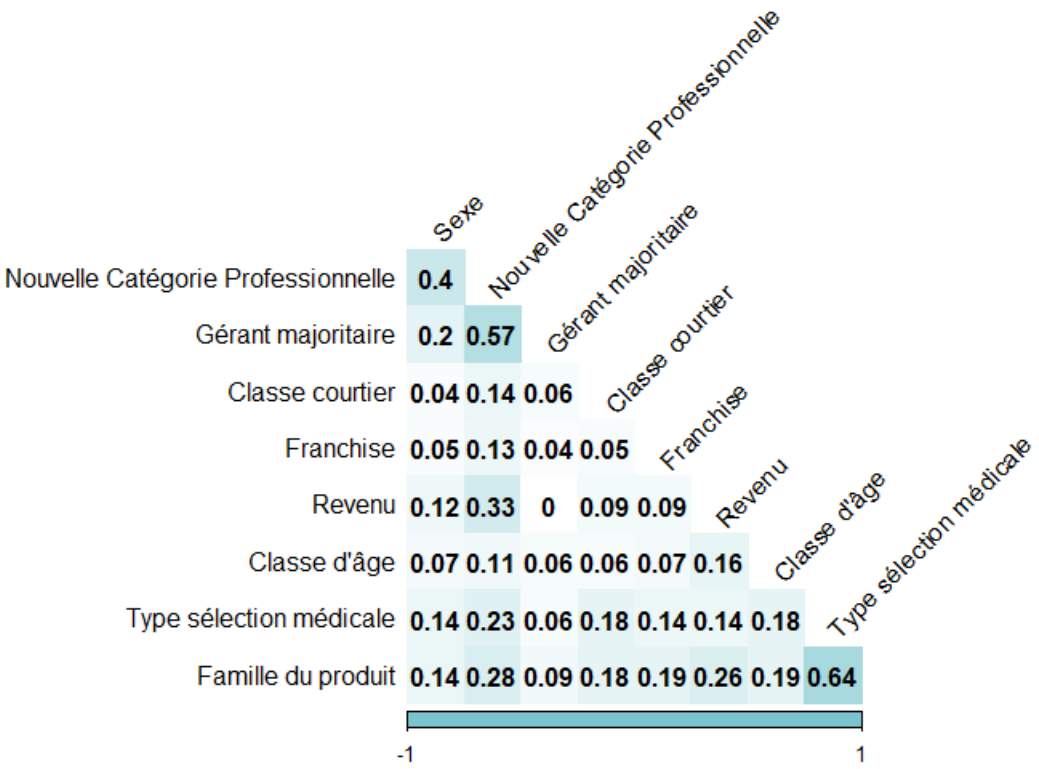

FIGURE 6.9 – V de Cramer illustrant la force du lien entre les variables qualitatives

Cette figure souligne le lien entre les variables *Famille du Produit* et *Type de sélection médicale*. Ce lien était prévisible étant donné la présence quasi-complète d'assurés ayant souscrit sans sélection médicale pour le produit GPE.
Un lien fort entre les variables *Gérant majoritaire* et *Nouvelle Catégorie Professionnelle* est aussi mis en lumière. Un tel lien n'existait pas avec la version initiale des catégories professionnelles, ceci est probablement le fruit d'un regroupement plus fin des assurés dans les nouvelles classes.
Une dernière relation semble se démarquer entre les variables *Nouvelle Catégorie Professionnelle* et *Sexe*, même si la valeur du V de Cramer reste en-dessous du seuil, ce qui n'implique pas de devoir faire un choix entre ces variables. Une fois encore cette valeur du V de Cramer est probablement en lien avec un regroupement plus fin qui rassemble des activités qui sont genrisées.

En réponse à ces découvertes, les variables *Gérant majoritaire* et *Famille Produit* sont écartées du modèle. La figure 6.10 rapporte de nouveau une régression globale du modèle. Cependant, ce dernier ne souffre plus de dépendance entre ses variables explicatives, ce qui est une condition d'application de la méthode GLM. Ainsi, les estimations des paramètres ne sont plus impactées par ce problème de conception du modèle.

	AIC		Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation	Entraînement	Validation
GLM retraitement 1	96 723	23 999	0,326863	0,326181	0,246414	0,244367
GLM retraitement 2	96 851	24 020	0,320837	0,320078	0,246516	0,244437

FIGURE 6.10 – Indices de performance de la modélisation sans variable Gérant Majoritaire

A ce niveau de l'étude peut être constatée une hausse des performances, comparativement au modèle actuel. L'amélioration de l'ajustement du modèle est soulignée par une diminution de l'AIC, tandis qu'une augmentation du coefficient de Gini atteste d'une meilleure segmentation. Une légère diminution de l'erreur quadratique peut aussi être notifiée. Les prédictions obtenues sont davantage en phase avec le profil de l'assuré.

Reste à vérifier la stabilité du modèle en évaluant le modèle avec une procédure par cross validation. Il a été décidé de recourir à un 4-Folds validation.

Cette méthode permet de valider ou non la stabilité d'un modèle, que ce soit en termes de stabilité des estimateurs, ou bien des indicateurs de performance. Cette méthodologie permet de lutter contre le sur-apprentissage et d'éviter le biais d'échantillonnage en ayant recours à la totalité de la base de travail pour entraîner le modèle.

Le principe du 4-Folds consiste à découper la base de données, composée de n lignes, en quatre sous-ensembles homogènes en termes de lignes (i.e. $\frac{n}{k}$ lignes) et de constitution. La figure 6.11 illustre la répartition des différents folds. Ensuite, un de ces sous-ensembles est conservé en tant qu'échantillon test, tandis que les trois autres servent d'échantillons d'apprentissage pour le modèle. Ce processus est réalisé quatre fois, afin d'utiliser une fois chacun des sous-ensembles en tant que base de test. Enfin, les coefficients estimés pour chacune des quatres modélisations, les erreurs, ou encore les indices de Gini, sont stockés et permettent à posteriori d'évaluer la stabilité et la performance du modèle.
Sur la figure ci-après sont représentés les différents indices de Gini relevés lors des modélisations sur les quatres sous-échantillons du 4-Folds. Ce graphique semble vraisemblablement valider la stabilité des résultats du modèle.
Peuvent aussi être appréciés les visuels des coefficients des variables explicatives. Seules deux variables sont affichées, l'idée étant simplement d'illustrer les propos. Les coefficients associés à une sélection partielle des modalités de la variable *Nouvelle catégorie professionnelle* et de la modalité *30/30/30* de la variable *Franchise* ont été recueillis afin d'être introduits dans la figure 6.13 : les valeurs sont relativement stables sur l'ensemble des échantillons.

6.3 Construction du Zonier

Dans cette partie, il est question d'étudier l'impact de la situation géographique sur l'incidence d'arrêt de travail. Aucune variable géographique n'a encore été introduite dans les modèles, et ceci a été fait intentionnellement dans le but de créer une variable "Zone" à partir des résidus du modèle GLM le plus performant. En effet, les résidus du modèle correspondent à la part non-expliquée de la variable Y par l'ensemble des variables expli-

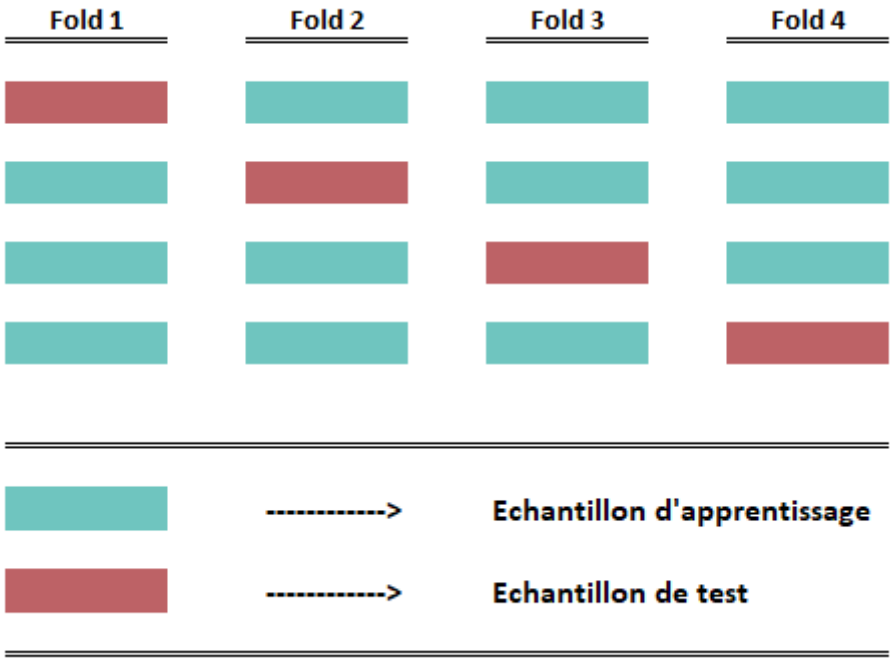

FIGURE 6.11 – Illustration de la répartition des échantillons pour un 4-Folds

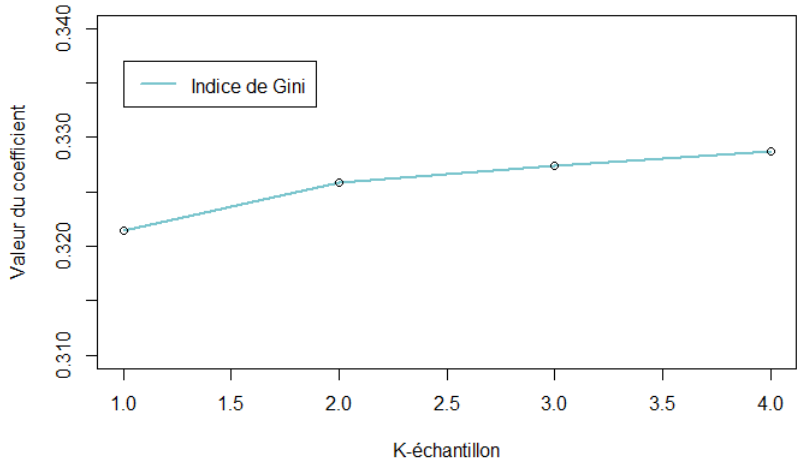

FIGURE 6.12 – Stabilité de l'indice de Gini lors de la cross-validation

catives. Ainsi, le fait de n'avoir aucunement implémenté de variable décrivant directement la situation géographique induit le fait que les informations liées à cette dernière devrait logiquement se trouver dans la part non-expliquée du modèle, c'est à dire les résidus.

Bien sûr, il est évident que certaines variables à priori non relative à la géographie puissent tout de même porter une part d'information la concernant. En effet, par exemple, certains départements (voire même certaines régions) sont davantage représentés par une certaine catégorie professionnelle : peuvent être citées des régions agricoles comme le Grand-Est ou la Nouvelle-Aquitaine. Pour autant, la majeure partie des informations géographiques

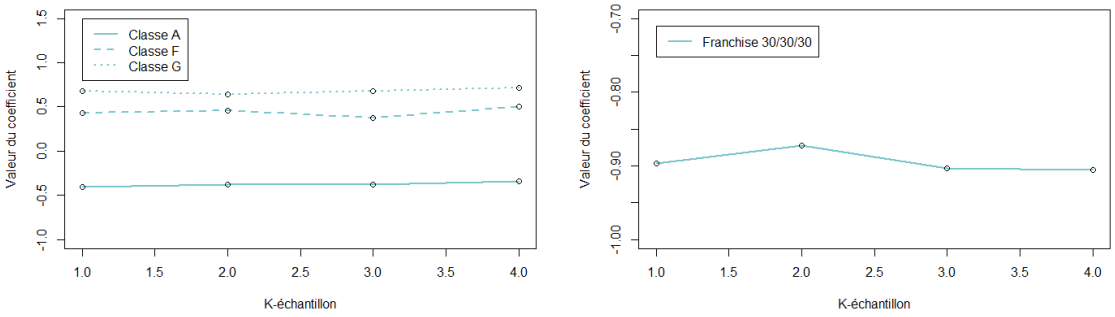

FIGURE 6.13 – Stabilité des coefficients prédits

devraient rester dans les résidus. C'est pourquoi, dans une optique d'optimisation, le choix a été fait de préférer créer un zonier à partir d'une étude des résidus, plutôt que d'ajouter directement des variables géographiques, comme le département, la région, ou encore des variables externes comme celles qui ont été présentées plus tôt, dans le modèle. Cela donne la possibilité de mettre au point une classification en zones plus juste et davantage maîtrisée.

Avant de se lancer dans cette procédure, il est nécessaire de s'assurer qu'il est réellement pertinent d'étudier les effets géographiques. Pour cela, il suffit de vérifier si la simple prise en compte de la variable département dans le modèle apporte davantage d'informations ou non. La figure 6.14 illustre parfaitement les disparités géographiques en termes de fréquence de sinistre. Il faut cependant interpréter ce graphique en prenant en compte les écarts d'exposition par département : certains départements ont une population d'assurés trop faible pour que la fréquence soit réellement pertinente. Un regroupement de cette population sous forme de zonier bien calibré peut répondre à ce problème.

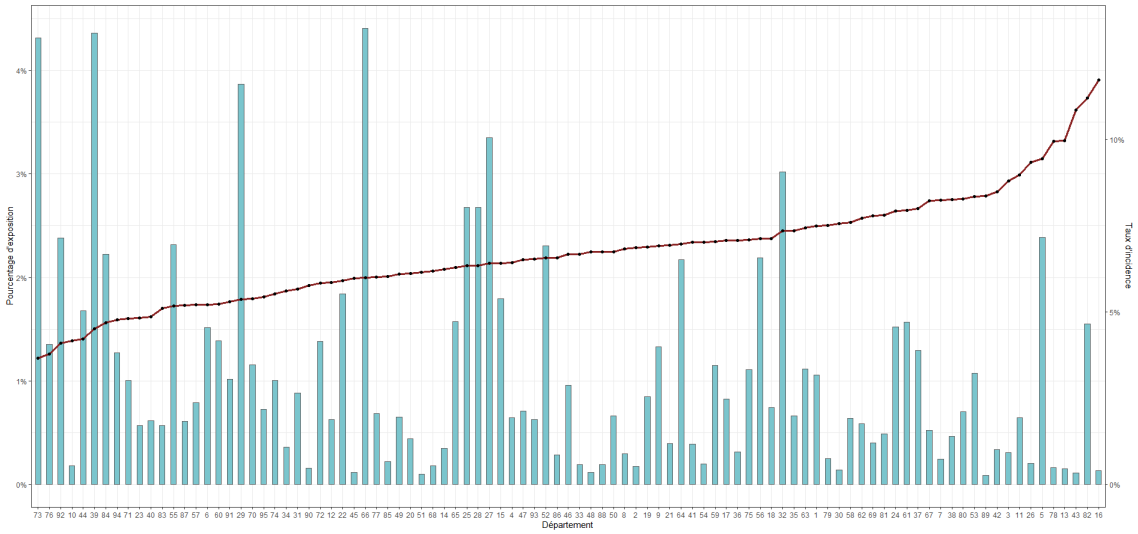

FIGURE 6.14 – Visuel des taux d'incidence par département

Pour ce faire, les résidus issus de la régression de Poisson sont agrégés par département et étudiés à part. Les résidus de Pearson sont utilisés. Ceux-ci sont définis comme l'écart entre les valeurs prédites $\widehat{y_i}$ et les valeurs réelles y_i pondérées par l'écart type de $\widehat{y_i}$, noté

s_i, pour l'observation $i \in [1, ..., n]$:

$$e_{i_P} = \frac{y_i - \widehat{y_i}}{s_i}$$

6.3.1 Classification suivant la taille des résidus

Dans cette partie, les zones géographiques sont créées manuellement, grâce à un certain nombre de conditions comme la taille des résidus, la taille minimale d'exposition par zone etc. Aucun recours a un système d'apprentissage, paramétrique ou non paramétrique n'est utilisé.

Comme annoncé, les résidus sont agrégés par département. A partir de ceux-ci sont calculés et stockés les résidus de Pearson, plus pratiques d'utilisation pour cette approche, car normalisés. L'exposition et le nombre de sinistres par département sont aussi importés.

Par souci de manipulation, les résidus de Pearson par département ayant des valeurs assez faibles, le traitement des résidus n'est pas fait directement sur la valeur brute des résidus, mais sur l'exponentielle de leur valeur. En effet, cette approche ne modifie pas l'ordre des résidus, car l'exponentielle est une fonction croissante, mais permet de les représenter artificiellement sur une plus grande échelle, et ainsi faciliter l'usage. En effet, il est difficile de distinguer des groupes distincts lorsque les valeurs sont trop resserrées : cela pose problème pour créer un classement.

Avant de classifier les résidus selon leur taille, une étude de l'exposition des départements est effectuée. L'objectif est de créer la zone de référence en regroupant les départements ayant moins d'exposition qu'un certain niveau seuil qui restera confidentiel. Ce seuil d'exposition est déterminé par plusieurs critères, tout en veillant à garder une certaine homogénéité du partage de l'exposition totale parmi l'ensemble des zones du zonier.

Ensuite, les départements restants sont classés selon la valeur de leurs résidus, en respectant des tranches de niveaux. Ces bornes permettent de créer cinq zones dont les résidus sont suffisamment écartés pour être considérés comme faisant partie d'une classe différente. Celles-ci sont représentées par la figure 6.15. La *Zone 5* correspond à la zone neutre, dans laquelle sont regroupés les départements ayant trop peu d'exposition.
Un inconvénient de cette approche réside dans la difficulté à considérer les bonnes bornes de séparation entre chacune des classes. Un axe d'amélioration pourrait être de mettre en place un protocole plus complexe qui permettrait de réellement distinguer des disparités entre les résidus. Aussi, la pertinence de regrouper ensemble les départements avec trop peu d'exposition peut être discutée.
Cependant, cette méthode a pour avantage d'être facilement mise en place, car elle ne nécessite aucun outil particulier, aucun paramètre, ni aucune donnée autre que le nombre de sinistres et de l'exposition par département.
De plus, par cette approche, deux types de résultats peuvent être exploités :

— une distribution des départements en différentes zones (comme c'est le cas ici) qui permet de créer une nouvelle variable significative afin de l'ajouter directement dans le modèle de prédiction.
— pour chaque zone créée, il est aussi possible de créer un coefficient indexé sur la taille de ses résidus. Ce coefficient permet d'introduire l'effet géographique à posteriori de la modélisation, sans avoir à ajouter la variable *Zonier* dans le modèle.

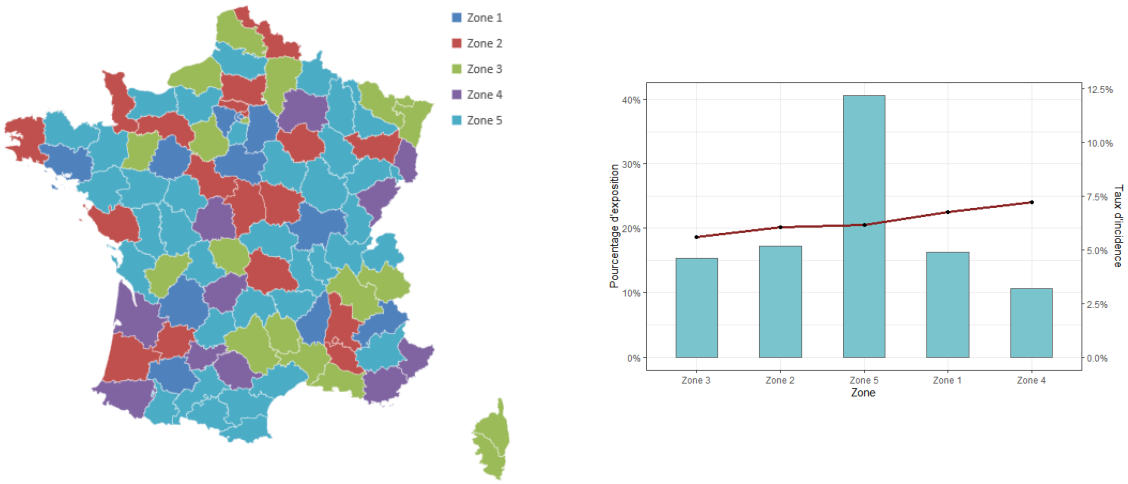

FIGURE 6.15 – Représentation du zonier créé sur une carte de France, et étude de sa répartition et des taux d'incidence par zone

Dans le cadre de ce mémoire, l'attention est portée uniquement sur l'exploitation des résultats sous forme de variable explicative à intégrer dans la modélisation.

6.3.2 Étude par arbre de décision

Une autre approche de construction de zonier, basée sur les arbres de décision, est exploitée. L'intérêt est de pouvoir comparer les résultats aboutissant des deux méthodes afin de conserver le zonier ayant de meilleures performances prédictives.

Les arbres de décisions possèdent des qualités indéniables. En effet, ils sont tout d'abord très facilement lisibles et interprétables, ce qui permet de présenter les résultats simplement à un public plus large. Cette méthode peut, tout comme le GLM, utiliser des variables quantitatives et qualitatives dans le même modèle afin de prédire la variable à estimer. De plus, en cas de valeurs manquantes dans la base de travail, les arbres de décision peuvent gérer cette situation par exemple en utilisant la médiane des valeurs de la variable. Cet atout n'est pas important dans le cadre de l'étude puisque la base a déjà été traitée auparavant afin qu'elle soit complète et de qualité. Enfin, contrairement aux GLM, aucune hypothèse sur les sinistres n'est imposée, c'est-à-dire que la structure de lien entre Y et les variables $X_1, ..., X_p$ ne doit pas nécessairement être linéaire.

Plusieurs méthodes d'induction d'arbres existent, mais seuls sera présentée dans ce mémoire les arbres CART, qui sont des arbres binaires (i.e. chaque noeud ne peut avoir que 0 ou deux branches). Les arbres de décision peuvent être utilisés sous forme d'arbres de régression, afin de prédire des variables quantitatives, ou d'arbres de classification, afin de prédire des variables qualitatives. Ici, la variable à estimer, les résidus, est une variable quantitative, c'est pourquoi cette partie est concernée uniquement par les arbres de régression. Peut aussi être nommée la méthode d'agrégation d'arbres *Forêt aléatoire* qui, comme son nom l'indique, utilise les résultats de nombreux arbres afin de fournir de meilleures performances plus stables et moins sujettes au sur-entraînement, cependant les résultats obtenus sont moins lisibles. Le concept de forêt aléatoire est l'objet du chapitre 7.

Les arbres de décision sont des méthodologies d'apprentissage non paramétriques. C'est

à dire que la formation des branches et des feuilles (et donc des différentes classes de risque) est obtenue par un apprentissage sur une base de données historiques. Le caractère non-paramétrique est dû au fait qu'aucun paramètre n'intervient dans les prédictions, contrairement aux GLM avec les paramètres β.

Théorie des arbres de décision et de l'algorithme CART

Le principe des arbres de régression binaires est de séparer la base de données en deux sous-ensembles appelés *noeuds*. Pour cela est choisie la variable qui permet de répondre au mieux à un certain critère de séparation. L'une des deux sous-populations répond au critère, et l'autre non. Pour chaque noeud ainsi créé est répétée cette étape de séparation en deux sous-ensemble jusqu'à atteinte d'un critère d'arrêt. Une même variable peut être utilisée pour créer une condition de coupure sur plusieurs noeuds. Un noeud ayant atteint ce critère d'arrêt est appelé *feuille* et constitue une classe de risque. L'algorithme assigne à cette classe la valeur qui correspond à la moyenne des observations faisant partie de cette classe. Cet algorithme est illustré par la figure 6.16, qui souligne que le critère d'arrêt peut intervenir à des niveaux d'arbre différents.

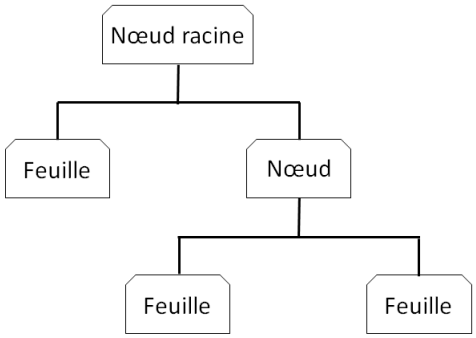

FIGURE 6.16 – Représentation simple d'un arbre de décision CART

La construction des conditions de coupures et du critère d'arrêt des arbres de régression vise à minimiser la déviance, qui est ici égale à la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles :

$$D = \sum_{i \in Ni} (y_i - \widehat{y_i})^2 \quad \text{, avec } N_i \text{ le noeud } i$$

Voici la vision mathématique qui induit la condition de coupure pour un noeud donné, en supposant que les variables sont quantitatives.

Soit $i \in [1, ..., n]$ et $k \in \mathbb{R}$. Avec n le nombre de variables explicatives et $(X_j)_{j \in [1,...,n]}$ les variables explicatives. La variable à expliquer est notée Y. Ainsi, les données sont représentées sous la forme :

	Y	X_1	X_2	...	X_n
Observation 1	Y_1	X_1^1	X_2^1	...	X_n^1
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
Observation p	Y_p	X_1^p	X_2^p	...	X_n^p

FIGURE 6.17 – Tableau des données utilisées dans la démonstration

Posons, pour le noeud en question :

- $\bar{y}_{i,k,gauche}$ la moyenne des valeurs de Y pour les observations qui vérifient $X_i < k$.
- $\bar{y}_{i,k,droit}$ la moyenne des valeurs de Y pour les observations qui vérifient $X_i \geq k$.
- $SE_{gauche}(i,k)$ la somme des carrés des écarts entre les valeurs de Y et $\bar{y}_{i,k,gauche}$ pour les observations vérifiant $X_i < k$
- $SE_{droit}(i,k)$ la somme des carrés des écarts entre les valeurs de Y et $\bar{y}_{i,k,droit}$ pour les observations vérifiant $X_i \geq k$

Ainsi, on peut noter l'erreur globale commise en séparant les observations selon que $X_i < k$ ou $X_i \geq k$ par la somme des erreurs à gauche et à droite, c'est à dire :

$$E(i,k) = SE_{gauche}(i,k) + SE_{droit}(i,k)$$

Ainsi, l'idée derrière la sélection de la variable et de la condition est de minimiser l'erreur induite par le choix effectué. Cela revient à utiliser la variable explicative, ainsi que le seuil les plus adéquats.
On peut alors noter la condition de coupure optimale comme suit :

$$X_{i^*} \geq k^*$$

de sorte que :

$$\forall i \in [1,...,n], \forall k \in \mathbb{R}, \quad E(i^*,k^*) \geq E(i,k)$$

Ainsi sont mis en place X_{i^*}, le caractère de coupure et k^* la valeur seuil du noeud. Les observations au niveau du noeud sont donc séparées en fonction de l'inégalité qu'elles satisfont.
De cela peut être induit que le caractère le plus plus influent est le caractère de coupure se situant au noeud de la racine de l'arbre.

Dans le cas où certaines variables sont qualitatives, le raisonnement reste le même, mais la condition devient :

$$X_i = m$$

avec m qui est une modalité de la variable.

Plusieurs conditions de coupure peuvent être considérées, et même combinées, pour les noeuds de l'arbre. Ces critères sont modulables via des hyperparamètres, ils diffèrent des paramètres des méthodes paramétriques de part leur utilité. En effet, ces hyperparamètres servent uniquement à contrôler le processus d'apprentissage en ajustant les critères :

- Le nombre minimum d'observations rassemblées au niveau du noeud pour pouvoir envisager de mettre en place une coupure ;
- Le nombre minimum d'observations présentes au niveau d'un noeud engendré par la coupure considérée. Ainsi, si le nombre d'observations est insuffisant, alors aucune coupure n'a lieu et le noeud devient une feuille ;
- La profondeur de l'arbre, i.e. ne pas dépasser un certain nombre de noeuds par branche ;
- La valeur d'un paramètre de complexité, noté *cp*. Ce paramètre exerce une influence sur la taille de l'arbre ;
- Le nombre de variables comparées dans un noeud afin de créer une condition de coupure ;
- D'autres hyperparamètres non exploités dans ce mémoire existent.

L'idée est que plus un arbre est profond, plus son biais est faible, mais plus sa variance est élevée. A contrario, un arbre peu profond aura un biais plus fort, mais une plus petite variance. Ainsi, l'optimisation des hyperparamètres vise à trouver le compromis entre biais et variance afin d'obtenir la meilleure performance de prédiction. C'est pourquoi le recours à l'élagage est communément utilisé afin de ne conserver que les feuilles nécessaires.

Modélisation des résidus

Au contraire de la méthode construction du zonier par taille des résidus, la méthodologie présentée dans cette partie se repose sur une prédiction par apprentissage par arbre de décision CART : ce zonier est nommé "zonier prédictif". Tout comme pour le zonier construit dans le chapitre précédent, la maille retenue est la maille du département. Un avantage de cette méthode de construction de zonier est qu'elle permet de classifier des situations géographiques ne possédant que peu d'exposition dans le portefeuille, et dont aucun historique de sinistralité n'est enregistré, grâce à l'utilisation de caractéristiques des départements présents dans l'historique, et qui sont semblables.

Ainsi, l'enjeu est de parvenir à identifier la composante géographique du risque de la manière la plus complète possible. Pour cela, il est nécessaire de connaître en globalité les caractéristiques des départements. Dans cette optique, il est nécessaire d'enrichir la base de données interne par de l'open data. Dans cette optique, huit variables sociaux-culturelles sont jointes à la base de données. Le choix des variables s'est porté sur des informations décrivant les habitants du département, les installations et aménagements du département, ou encore des caractéristiques plus originales du département dont il n'est pas naturel de vouloir les lier aux incidences en arrêt de travail. Voici un rappel des variables en question, importées depuis le site de l'INSEE, qui sont agrégées à la maille département :

- La densité médicale du département ;
- La part de densité médicale représentée par la médecine générale dans le département ;
- Les températures minimum, maximum, et moyenne dans le département (moyenne depuis 2016) ;

— Le revenu médian de la population du département ;
— La densité de population au km^2 du département ;
— Le taux de pauvreté dans le département ;
— Le taux de logements sociaux dans le département ;
— Le taux de logements individuels dans le département ;
— La part de transport en commun dans le département.

Puisque ces variables sont de nature géographique, il est toujours enrichissant d'apprécier un rendu graphique de leur répartition en ayant recours à une cartographie par département.

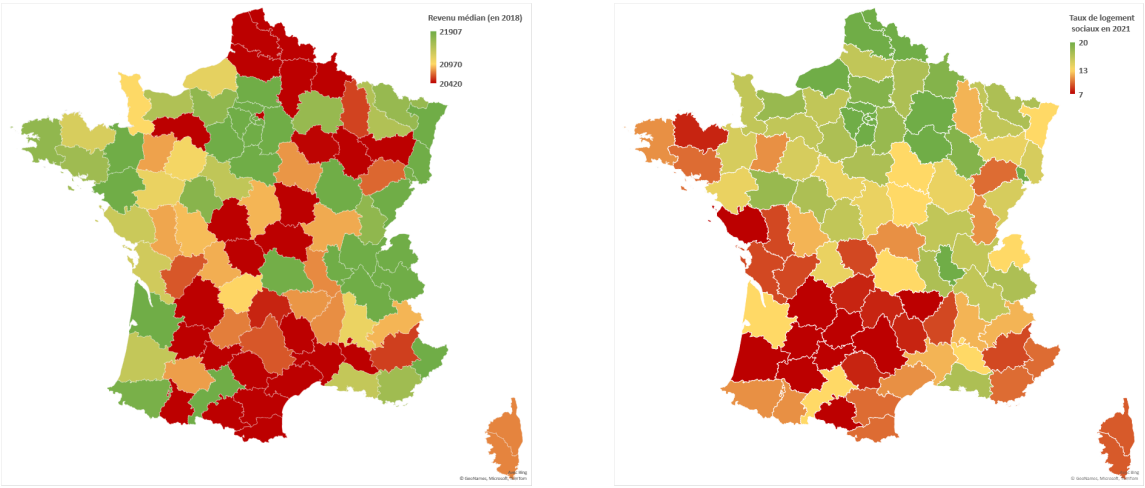

FIGURE 6.18 – Cartographie des revenus médians et du taux de logements sociaux dans les départements français

La figure 6.18 souligne des disparités départementales vis-à-vis du revenu médian. La partie Nord de la France semble avoir davantage de logements sociaux que la partie Sud, la Bretagne faisant exception.

Après une étude descriptive, les variables relatives aux températures maximales et minimales des départements ont été écartées de l'étude car leur corrélation avec les résidus apparaît comme trop peu suffisante.

Dans un premier temps, l'objectif est de former un arbre de décision comportant un nombre convenable de zones pour segmenter suffisamment les départements. Pour cela, un jeu d'optimisation à la fois automatique, grâce à des outils d'aide à la décision informatiques, et manuel, en ajoutant le facteur de décision humain. Dans cette optique est entamée la détermination de l'hyperparamètre qui décide du nombre de variables qui sont comparées dans chaque noeud pour mettre en place la condition de coupure. Cette étape, réalisée grâce à la librairie *Caret* de R, se conclut sur un choix de trois variables à comparer, parmi les neuf. En effet, cela permet de ne pas systématiquement utiliser une condition trop forte sur un noeud d'un niveau trop élevé et qui résulterait sur une coupure trop marquée sur l'un des groupes. Aucune valeur maximale sur la profondeur de l'arbre n'est ici imposée au modèle.
Ensuite la phase d'élagage permet de déterminer le paramètre de complexité qui minimise les erreurs.

Ainsi, avec cette optimisation, il résulte un arbre délivrant huit feuilles sur 6 niveaux, c'est-à-dire huit zones distinctes. La répartition est telle que chaque zone contient au moins de 5% de la population, ce qui représente une sectorisation suffisamment large.

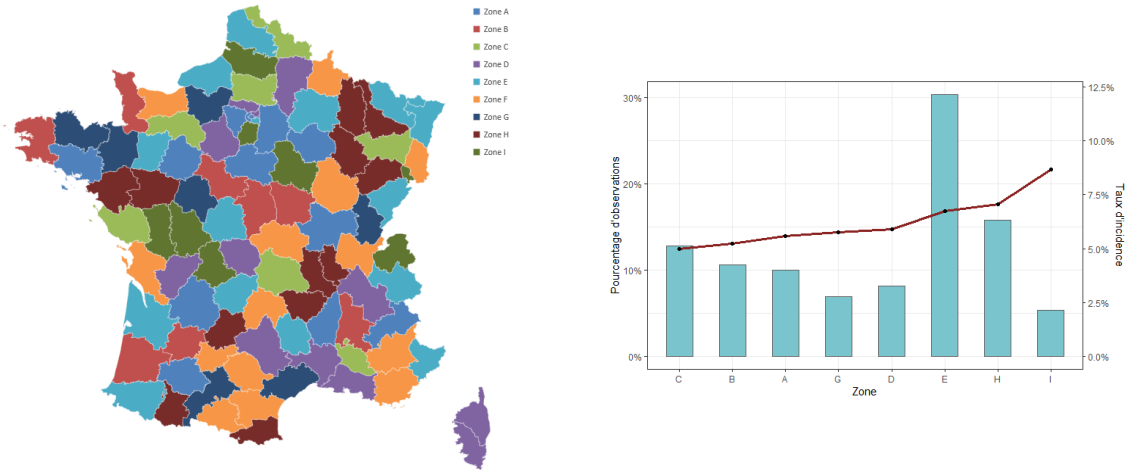

FIGURE 6.19 – Représentation du zonier créé par arbre CART sur une carte de France, et étude de sa répartition et des taux d'incidence par zone

La figure 6.19 permet de constater la présence d'une zone particulièrement risquée. De plus, en comparant ce graphique avec celui de la méthode précédente, il apparaît que les zones sont calibrées sur une échelle plus large de taux d'incidence, ce qui laisse penser que davantage d'informations géographiques sont captées par la régression, et que cela permet une meilleure compréhension des résidus. Reste à vérifier si ce premier avis qualitatif se confirme lors de la comparaison par les indicateurs de performance.

6.3.3 Choix du modèle GLM retenu

Jusqu'ici, un corps commun de modèle a été mis en place. Ensuite, à partir de la partie 6.2, deux méthodes de construction de zonier ont été menées à terme. Dès lors, il est nécessaire de sélectionner un unique modèle qui servira de référence pour représenter la modélisation par la théorie des modèles linéaires généralisés. Ainsi, il s'agit de retenir le modèle considéré comme le plus performant.

	AIC		Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation	Entraînement	Validation
Zonier Taille Résidus	96 818	24 010	0,322372	0,321814	0,246485	0,244422
GLM + Zonier CART	96 701	23 986	0,327090	0,326738	0,246430	0,244354

FIGURE 6.20 – Indices de performances pour les différents modèles après application des zoniers

Après analyse de la figure 6.20, il ressort que le zonier construit à l'aide de l'arbre de décision offre davantage de performance, tant en termes de segmentation que de calibration ou d'erreurs, que ce soit pour les échantillons d'entraînement ou de validation : c'est donc ce modèle qui est retenu pour la suite de l'étude.

Dans la partie suivante, lors de la création du zonier du modèle suivant, la méthode de création par taille des résidus ne sera pas réitérée. En effet, les résultats de cette comparaison laissent penser que cette méthode n'exploite pas suffisamment le pouvoir prédictif des informations géographiques.

Chapitre 7

Modélisation par forêt aléatoire

Plusieurs types de forêts d'arbres de régression existent, mais la méthode par forêt aléatoire (*Random forest* en anglais) est l'approche la plus répandue.
La forêt aléatoire est une méthode ensembliste qui se base, comme son nom le laisse penser, sur un grand nombre d'arbres de décision. Ces arbres de décision sont appelés "classifieurs faibles", car il permettent d'expliquer une part de la variable cible, mais ne permettent pas individuellement d'obtenir les performances souhaitées.

7.1 Théorie des forêts aléatoires

7.1.1 Intérêt et fonctionnement des forêts aléatoires

L'idée derrière les forêts aléatoires est de corriger les défauts des arbres de décision, en améliorant par exemple la stabilité (les arbres de décisions sont fortement impactés par la modification des modalités d'une seule variable par exemple) ou en diminuant le risque de sur-apprentissage. Les forêts aléatoires sont développées, tout comme les algorithmes CART pour les arbres de décision, par BREIMAN.

Les forêts aléatoires utilisent le système de *bagging*, pour *Bootstrap aggregating*, adapté aux arbres CART.
Le *bagging* est une méthode qui consiste à sous-échantillonner les données, par tirages aléatoires avec remise d'observations de la base complète, afin de créer un modèle sur chacun des sous-ensembles. Du fait du tirage aléatoire, les partitions sont indépendantes et légèrement différentes. Le résultat final est la moyenne de prédiction de tous les arbres construits. Les échantillons d'apprentissage étant tous légèrement différents, la moyenne permet de corriger l'instabilité que connaissent les arbres de décision individuels.

La différence entre le *bagging* et les forêts aléatoires se trouvent dans la manière de créer les modèles. En effet, les sous-ensembles de données sont créés de manière équivalente, mais les modèles mis en place sont construits à partir d'une sélection aléatoire des variables explicatives (i.e. un sous-ensemble des variables explicatives), ce qui permet de créer une forêt d'arbres dont les noeuds ne sont globalement pas construits par les mêmes variables. Comme pour le *bagging*, l'ensemble des arbres de la forêt sont agrégés afin de retourner la moyenne des prédictions en cas de forêt de régression, ou le vote majoritaire, en cas de forêt de classification.

Ce renforcement de l'aléa en associant une sélection des variables explicatives aux observa-

tions déjà elles-même aléatoires tend à diminuer significativement la corrélation entre les modèles, ce qui offre un gain de performance conséquent et diminue la variance.
Cependant, en contrepartie de cette augmentation de performance, cette méthode n'offre pas la lisibilité d'un arbre CART. En effet, il est plus difficile de comprendre totalement quelles sont les variables qui impactent les prédictions, et à quel niveau, puisqu'une forêt de 500 arbres, par exemple, ne peut pas être représentée. Les forêts aléatoires entrent dans le groupe des méthodes dites "boite noire". De plus, l'exécution de cette procédure nécessite un temps de calcul bien plus conséquent, et encore davantage pour définir les hyperparamètres optimisés.

7.1.2 Éléments d'interprétation

OOB - Out-Of-Bag

L'algorithme de construction des *Random Forest* permet de calculer l'erreur de généralisation de son modèle, appelée Out-Of-Bag. Dans le cas d'une régression, cette erreur revient à calculer l'erreur quadratique moyenne des prédictions qui n'ont pas été entraînées. Ici, le terme "non entraîné" ne signifie pas que les observations ne font pas partie de l'échantillon d'apprentissage de la forêt. En effet, la procédure de bagging suivie fait que la base d'apprentissage est elle-même séparée en une base *train* et une base *test* : il est donc question ici des observations de cette base *test*, et ce pour chaque classifieur faible. C'est ainsi que cette erreur a été nommée "erreur de généralisation", car elle tend à calculer l'erreur lorsque le modèle est étendu à des données "hors du bagging", qui sont uniquement des données prédites. Cette erreur est centrale dans la théorie des forêts aléatoires, car elle permet de comparer la performance des modèles, et est aussi en lien avec l'estimation de l'importance des variables.

Importance des variables

Certes, la forêt aléatoire fait partie de la famille des "boîtes noires", cependant cela ne signifie pas qu'il est impossible d'obtenir des informations participant à une interprétation du modèle. Il est en effet possible d'établir un classement des variables explicatives basées sur leur importance dans le modèle.

L'importance d'une variable est ici définie par rapport à l'accroissement moyen de l'erreur d'un arbre après permutation aléatoire des valeurs de cette variable dans les échantillons OOB. Plus l'augmentation d'erreur est forte, plus la variable est considérée importante.

Voici le détail du calcul de l'importance de la variable X_j, $j \in [1, ..., p]$. Soit L_t, $t \in [1, ..., T]$ un des T échantillons boostrap effectués dans la modélisation de la forêt aléatoire, contenant k observations. A cet échantillon est associé l'échantillon OOB_t qui contient les observations non incluses dans L_t, et à partir desquelles est calculée l'erreur Out-Of-Bag, notée Err_t. Ainsi, :

$$Err_t = \frac{1}{k} \sum_{i=1}^{k} (Y_i - \widehat{Y}_i)^2$$

Ensuite, une copie de l'échantillon OBB_t, notée OBB_{t_C} est faite, dans laquelle les valeurs des observations de la $j - ième$ variable sont permutées entre elles de manière aléatoire, comme l'illustre la figure 7.1. Est ensuite calculée l'erreur Out-Of-Bag de cet échantillon OBB_{t_C}, notée Err_{t_C}.

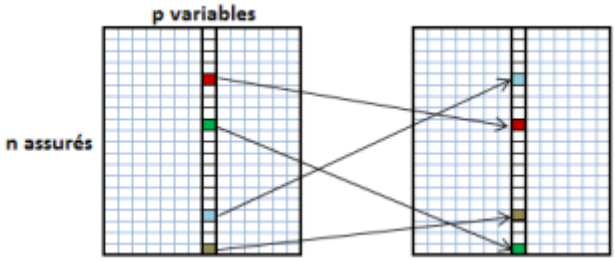

FIGURE 7.1 – Illustration d'une permutation aléatoire de l'une des colonnes de l'échantillon boostrap

Ceci est effectué sur l'ensemble des échantillons boostrap. Ainsi, l'importance de la variable X_j est définie par la formule :

$$Imp(X_j) = \frac{1}{T}\sum_{t=1}^{T}[Err_{t_C} - Err_t]$$

Par cette définition de l'importance, les variables dont les permutations augmentent fortement l'erreur sont considérées importantes. A contrario, si l'écart entre l'erreur après permutation et l'erreur initial faible, voire même si la différence est négative, alors la variable est statuée non importante.

7.2 Estimation de l'incidence des sinistres arrêt de travail

De nouveau, il est question de créer un modèle ne tenant pas compte de la situation géographique, afin de traiter à part les erreurs de prédiction dans le but d'établir un zonier.

Les modélisations par forêts aléatoires sont effectuées à l'aide du package "Ranger" de R. Celui-ci offre une vitesse de calculs performante, et permet l'utilisation de package Caret afin d'aider à l'optimisation des hyperparamètres.

Le premier point d'optimisation est le nombre de variables par noeuds qui sont comparées afin de trouver la condition de coupure qui diminue le plus l'erreur. Pour cela, une procédure informatique automatique construit les modèles en partant d'une unique variable jusqu'à la totalité des variables qui sont comparées. Il se trouve que le modèle est le plus performant en comparant 4 variables.
L'étape suivante est consacrée au choix du nombre d'arbres à créer pour former la forêt. Pour cela, un enchaînement pas à pas de modélisations en augmentant successivement le nombre d'arbres par arbres est effectué. Ceci est effectué de manière récursive jusqu'à ce que l'erreur Out-Of-Bag cesse de diminuer et se stabilise. La figure 7.2 illustre le fait qu'à partir de 300 arbres l'erreur commence à se stabiliser, mais c'est uniquement à partir de 800 arbres que l'erreur OBB devient presque constante : la valeur de l'hyperparamètre est donc fixée à 800 arbres.

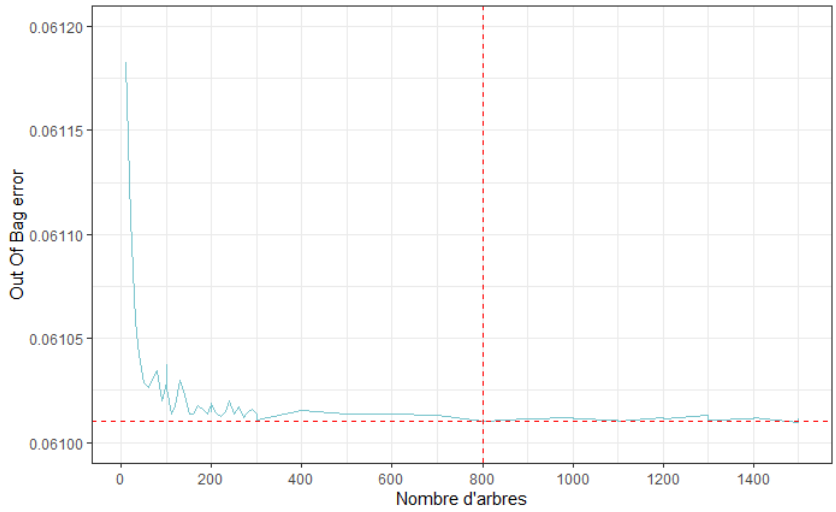

FIGURE 7.2 – Diminution de l'erreur de généralisation jusqu'à atteinte d'une stabilité à partir de 800 arbres

La troisième et dernière étape d'optimisation réalisée concerne la profondeur des arbres. Là encore, la détermination de l'hyperparamètre associé nécessite de tracer la courbe de l'erreur de généralisation de la figure 7.3 qui permet de voir à partir de quelle valeur l'erreur se stabilise. Le choix s'est donc porté sur une profondeur de 12 niveaux.

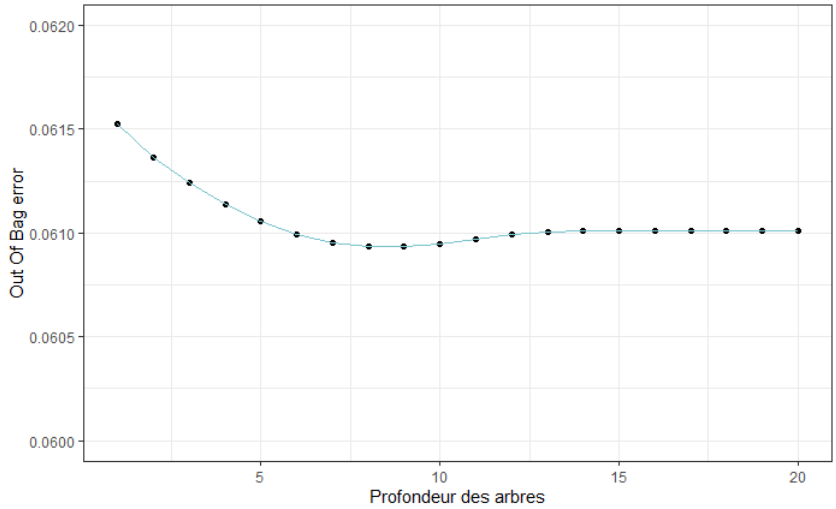

FIGURE 7.3 – A partir de 12 niveaux, l'erreur de généralisation semble constante

A ce stade de la constitution du modèle par méthode de forêt aléatoire, le modèle sans prise en compte de l'information géographique est mis au point. Les résidus sont donc calculés et mis de côté afin d'entreprendre la construction du zonier.
En termes de performances, les résultats obtenus par cette forêt aléatoire sont pour le moment en dessous de ceux obtenus par les modèles linéaires généralisés, selon les indicateurs de la figure 7.4

	Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation
GLM retraitement 2	0,320837	0,320078	0,246516	0,244437
RF	0,313754	0,314602	0,246824	0,244933

FIGURE 7.4 – Les premiers résultats du modèle Random Forest sont moins encourageants que ceux du GLM, tant en termes d'erreurs qu'en termes de segmentation

7.3 Construction du Zonier

La procédure de construction de l'arbre CART est répétée pour aboutir à un nouveau zonier, adapté au modèle Random Forest qui vient d'être mis en place. Le zonier, dévoilé par la figure 8.1, connaît une répartition des assurés relativement homogène, et la pente quasi linéaire rassure sur la pertinence des classes de risque créée par ce zonier.

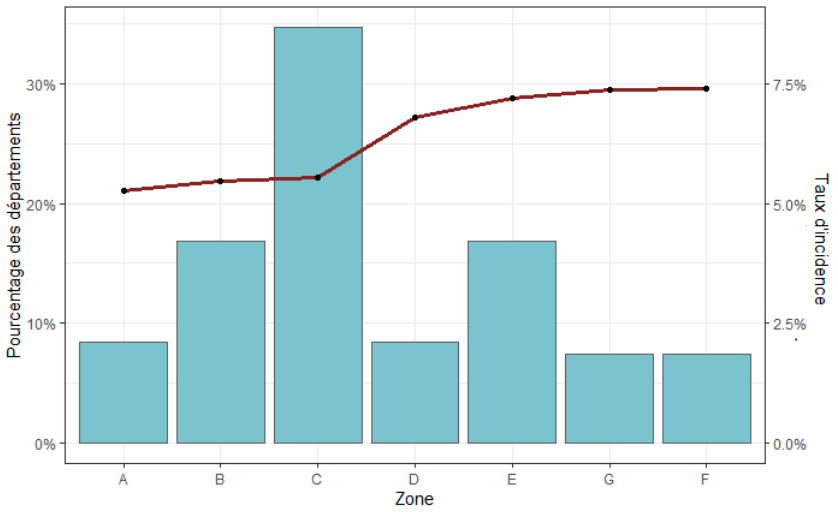

FIGURE 7.5 – Zonier créé à partir des erreurs de prédiction du modèle par forêt aléatoire, dont les zones sont triées par ordre croissant de risque

Cette disparité de risque entre chaque zone permet d'améliorer significativement l'indice de Gini (+0.6 points) sur les échantillons d'apprentissage et de validation, il n'y a semble-t-il pas de perte de performance lors de la généralisation même si il y a une légère augmentation de l'erreur quadratique (+0.01 point) sur les estimations fittées sur l'échantillon d'apprentissage.

	Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation
GLM + Zonier CART	0,327090	0.3267377	0,246430	0,244354
RF + Zonier CART	0,371000	0,371193	0,246770	0,245221

FIGURE 7.6 – L'apport du zonier entraîne une hausse significative de la segmentation du modèle par forêt aléatoire

Chapitre 8

Évaluation de fin d'étude

Le travail de ce mémoire vise à augmenter la compréhension et à expliquer le risque arrêt de travail d'un portefeuille de Travailleurs Non Salariés. Dans cette optique, deux possibilités d'étude ont été opposées, l'une s'appuyant sur des modélisations statistiques, et l'autre se basant sur l'apprentissage supervisé d'un historique. Même si l'objectif de prédiction des incidences leur est commun, ces deux méthodes diffèrent fortement vis-à-vis de leur vision de la variable à expliquer. En effet, les modèles linéaires généralisés ont l'ambition de retrouver la loi des sinistres sachant les variables explicatives, ce qui permet de calculer le nombre de sinistres annuels par profil de risque. Tandis que les forêts aléatoires n'imposent pas de grandes hypothèses et visent à apprendre quelle sinistralité est attachée à un profil de risque, afin de pouvoir « deviner », ou plutôt « répliquer », la sinistralité d'un portefeuille en supposant que les sinistres futurs se produiront de la même manière que les sinistres passés.

Dans ce dernier chapitre, l'objectif n'est pas de désigner le modèle qui est surpassé en tout point les performances du second, mais plutôt de mettre en perspective les qualités et défauts de chacun en se basant à la fois sur les caractéristiques intrinsèques et les résultats réels obtenus lors d'une phase d'évaluation. Cette vision de la comparaison est importante étant donné que le choix final de l'outil qui sera utilisé lors de la tarification se fera non-seulement en fonction des performances relevées par les indicateurs, mais aussi en fonction d'une étude « métier »qui devrait permettre de révéler la fidélité des prédictions par rapport à la réalité des sinistres.

8.1 Interprétation des critères de performance

Tout d'abord, de manière assez linéaire, il peut être intéressant de vérifier quelles sont les conclusions si le seul critère de sélection est porté par les indicateurs de performance. Les méthodes de modélisations n'étant pas les mêmes, tous les indicateurs de performance utilisés jusqu'ici ne sont pas exploitables. Ainsi, seul le RMSE et l'indice de Gini servent de point de comparaison.

	Gini		RMSE	
	Entraînement	Validation	Entraînement	Validation
GLM + Zonier CART	0,327090	0.3267377	0,246430	0,244354
RF + Zonier CART	0,371000	0,371193	0,246770	0,245221

FIGURE 8.1 – L'apport du zonier entraîne une hausse significative de la segmentation du modèle par forêt aléatoire

Ainsi, en exploitant seulement les indicateurs de Gini et d'écarts RMSE, le modèle par forêt aléatoire semble se démarquer des modèles linéaires généralisés par une segmentation sensiblement plus forte. Ce modèle donne donc l'impression de mieux comprendre la structure des profils de risque. Cependant, l'erreur moyenne des prédictions est légèrement plus haute. Sachant que l'estimation globale est aussi supérieure au nombre de sinistres totaux, une première conclusion semble être que ce modèle sur-estime davantage les sinistres que les modèles linéaires généralisés.

8.2 Prédictions pour des profils de risque cibles

Les critères de performance sont le premier point d'appui afin de se faire un avis sur les modèles. Cependant, comme le suggère leur nom, les résultats délivrés ne sont qu'indicatifs et ne reflètent pas toujours la réalité de manière satisfaisante. Ainsi, il est nécessaire d'effectuer des backtestings sur les outils créés. Pour cela, certains profils de risque, qui correspondent à des profils cibles types du produit de prévoyance d'Entoria, ont été sélectionnés et étudiés. Deux graphiques sont présentés ci-après. Ils sont construits de sorte à représenter l'évolution des prédictions de sinistres et de l'exposition par classe de risque.
L'objectif ici est de comparer les prédictions des différents modèles aux sinistres réels.

La figure 8.2 présente les caractéristiques de 13 classes de risque créées à partir de la nouvelle classe professionnelle et de l'âge des assurés. L'axe des abscisses affiche les classes par ordre croissant de risque : cette évaluation des risques découle d'un calcul du nombre sinistres agrégés selon la classe de risque. Ainsi, les assurés faisant partis de la catégorie Professionnelle "A" et ayant plus de 48 ans ont un taux d'incidence plus faible que les assurés de plus de 48 ans de la classe "K".

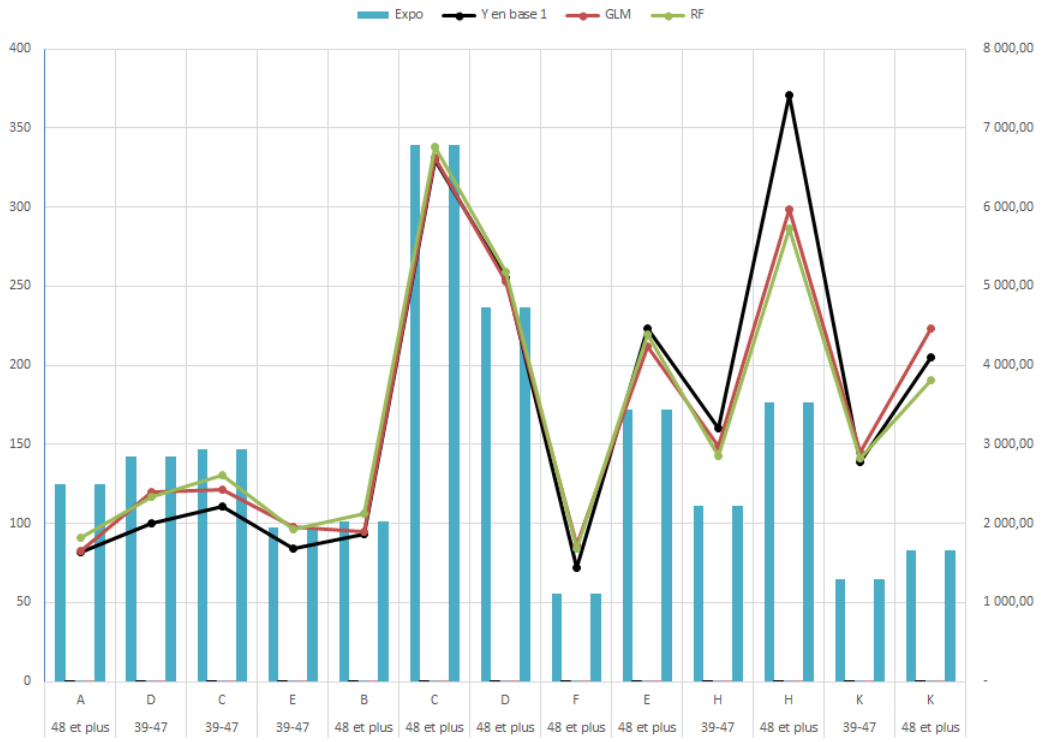

FIGURE 8.2 – Prédictions des sinistres en base 1 de la base test pour des classes d'âge et des classes professionnelles cibles

La vision de ce graphique donne l'impression de s'opposer à l'interprétation des indices de Gini. En effet, comparativement aux prédictions par forêt aléatoire, les prédictions agrégées par classes du modèle GLM semblent être globalement plus proches des nombres réels de sinistres. Un second point semble émerger, les deux approches donnent l'impression d'une tendance de sur-évaluation des risques faibles et de sous-évaluation des incidences lorsque le risque devient plus fort ; tandis que les deux modèles gagnent en précision sur les classes ayant le plus d'exposition.

Le second graphique ci-dessous permet de constater l'impact des modifications effectuées sur la répartition des assurés en classes professionnelles tout âge confondu. Sur ce graphique est ajouté la courbe des prédictions du modèle actuel, ainsi la courbe de sinistralité réelle est représentée en base 1 afin d'améliorer la compréhension graphique, et servira de courbe de comparaison pour les prédictions. Plus la prédiction est proche de cette courbe, plus elle est fidèle à la réalité, et donc plus la méthode de prédiction est performante. Par suite de l'étude menée, il est logiquement attendu une amélioration de la précision de prédiction pour les modèles s'appuyant sur les nouvelles classes professionnelles.

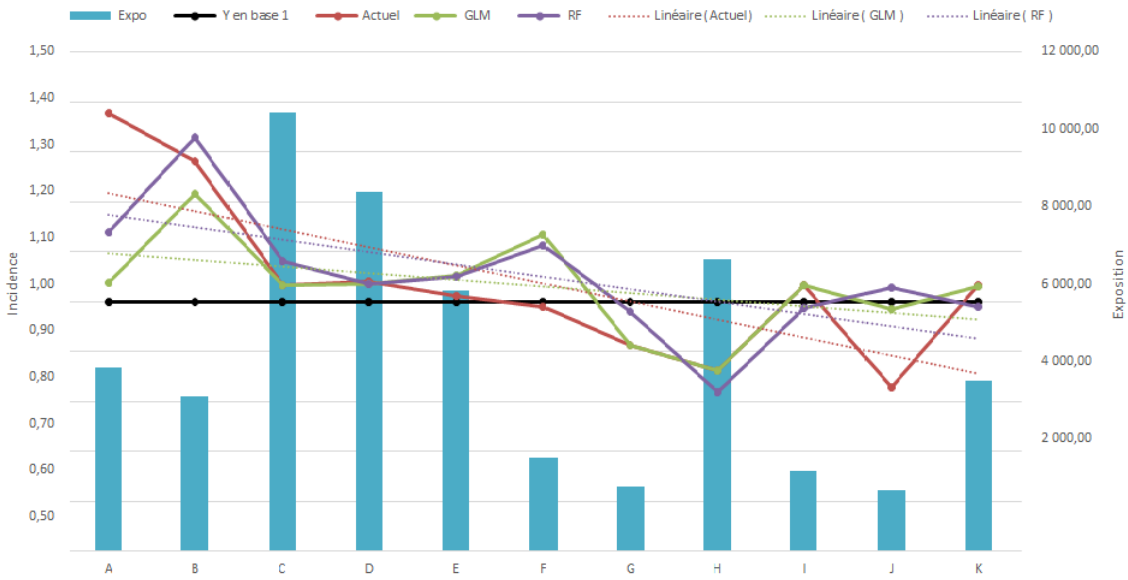

FIGURE 8.3 – Prédictions des trois modèles sur l'échantillon test en fonction des nouvelles catégories professionnelles

Les prédictions agrégées par classes professionnelles du modèle GLM semblent les plus proches des sinistres historiques, notamment lorsque l'exposition est moindre, avec un comportement proche du modèle actuel probablement expliqué par la même approche des modèles linéaires généralisés. Les courbes de tendance permettent de conforter cet avis. La précision des prédictions par forêt aléatoire donne quant à elle l'impression d'être en-deçà de celle par GLM.
A noter que certaines étapes de construction des nouvelles catégories professionnelles sont liées à la construction du modèle GLM, il n'est donc pas étonnant que la compréhension de ce dernier vis-à-vis des classes de risque soit plus approfondie.
Enfin, la tendance de sur-évaluation des risques plus faibles, et sous-évaluations des risques plus forts est une fois encore visible.

Conclusion

Les travaux présentés dans ce mémoire ont été menés dans l'optique de répondre aux trois grands questionnements de la problématique rencontrée par Entoria. Plus largement, l'objectif était de mieux comprendre le risque arrêt de travail pour son portefeuille de contrats prévoyance TNS.

Le contenu de cette étude a débuté par une brève présentation de l'assurance prévoyance, afin de comprendre l'enjeu et les risques associés. La problématique a ensuite clairement été définie, ainsi que les méthodes statistiques qui sont entrées en jeu.

Ensuite, une vaste présentation des données a été faite, ce qui a permis de bien appréhender l'intérêt de chacune des variables explicatives de l'étude. Certaines variables ont dès lors été écartées de par leur manque de lien avec la variable des sinistres. D'autres ont quant à elles été modifiées afin d'exploiter au mieux leur potentiel prédictif.

Une fois les données bien comprises et maîtrisées, les différentes modélisations sélectionnées ont été entraînées et calibrées sur celles-ci. Le développement des-dits modèles a nécessité plusieurs étapes d'optimisation, que ce soit en termes de variables, mais aussi de paramètres ou d'hyperparamètres. Le choix des méthodes de modélisation a été fait dans un objectif d'unification des résultats de ces derniers, afin qu'ils soient comparables et exploitables facilement par les outils d'Entoria.

La comparaison des résultats par les différents indices de performance à disposition a montré ses limites. En effet, ces indices donnent un ordre d'idée mais ne se suffisent pas à eux-même. Ils sont cependant très importants lors du développement et l'optimisation des modélisations pour se rendre compte des évolutions avant et après modification.
Ainsi, pour compléter l'avis « qualitatif » qu'offre les indices de performance, une étude comparative avec une vision davantage métier et « quantitative » a été menée. Le but était de visualiser l'agrégation des prédictions sur différents profils de risque. Ces prédictions ont été effectuées sur une base de test extraite des données de gestion, ce qui permet d'avoir une référence historique à comparer aux prédictions.

Cette phase de comparaison semble désigner que le modèle GLM construit dans cette étude est le plus performant au niveau global. Les indicateurs indiquent une performance au global légèrement moins bonne que le modèle par forêt aléatoire en termes de segmentation, mais l'étude par classe de risque atteste d'une plus grande robustesse au non-global.

En conclusion concernant l'utilisation de données externes pour la tarification d'un produit prévoyance, trois points semblent s'être démarqués dans le cadre de ce mémoire et des données à sa disposition :

— Il existe un réel potentiel d'amélioration d'une base de données lorsque celle-ci est liée à des bases externes. Ce potentiel peut être exploité en créant de nouvelles variables dont les valeurs prises sont le résultat d'un travail d'assemblage entre une ou plusieurs variables internes et externes, ou tout simplement en ajoutant des variables purement externes à la base interne.
— Il est possible de construire des zoniers ayant de bons résultats à partir de données géographiques. La difficulté réside dans la recherche et le choix des variables, qui peuvent être très riches et dont il faut absolument vérifier l'exactitude de information.
— Toutefois, malgré le grand nombre de variables externes, les injecter directement dans un modèle n'offrait que peu de résultats. Cela peut s'expliquer tout d'abord par le fait que les variables externes sont parfois trop générales, et donc pas assez proches de l'individu étudié, mais aussi par le fait que la base de données de cette étude n'avait pas de clef primaire pour chaque individu de la population.

Bibliographie

[1] Aliche Ali. Invalidité et arrêt de travail en Australie. Master's thesis, ENSAE,Institut Polytechnique de Paris, Paris, 2016.

[2] Manon BEAUPOIL. Comment définir une stratégie de hausses tarifaires à l'échéance anniversaire en assurance habitation ? Master's thesis, ISUP, Sorbonne Université, Paris, 2017.

[3] Neil Bellagha. Modélisation de la sinistralité en incapacité d'un portefeuille de TNS et de salariés par apprentissage supervisé binaire. Master's thesis, ISUP, Sorbonne Université, Paris, 2018.

[4] Rémi BELLINA. Méthodes d'apprentissage appliquées à la tarification non-vie. Master's thesis, ISFA, Univ. Clauder Bernard Lyon 1, Lyon, 2014.

[5] Myriam BERTRAND, Frédéric MAUMY. Choix du modèle. Master's thesis, Université Louis Pasteur Strasbourg, France, 2008.

[6] Ghita BOUCHTA. Mise en œuvre de méthodes innovantes de tarification. Master's thesis, ENSAE, Univ. Clauder Bernard Lyon 1, Institut Polytechnique de Paris, Paris.

[7] Damien DOMECQ, Gwendoline LANGJAHR. Assurances collectives, risque arrêt de travail : Mise en place d'un indicateur d'évolution du risque à court terme. Master's thesis, CEA, Centre d'Etudes Actuarielles, 2013.

[8] Etat Francais. Loi evin. <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000709057#:~:text=Loi%20n%C2%B0%2089%2D1009,%2D%20L%C3%A9gifrance>, 31 décembre 1989.

[9] Mohamed Halimi. Réactualisation des méthodes classiques de tarification IARD. Master's thesis, ENSAE, ENSAE Paris, Paris, 2017.

[10] Julie. Machine learning : Du GLM à l'arbre de CART en passant par le Random Forest. <https://www.pericles-group.com/wp-content/uploads/2016/11/ceipa110_datas_ml_enbrefduglmalarbrecartenpassantparlerandomforest-1.pdf>, 2016.

[11] Ozlem KARATEKIN. Tarification et mesure de l'antisélection en assurance santé collective. Master's thesis, DUAS, Université de Strasbourg, Strasbourg, 2014.

[12] Kevin LECOMPTE. Automatiser la comparaison de modèles : Application sur l'amélioration d'un modèle de fréquence par des techniques de machine learning. Master's thesis, ENSAE, Université Clauder Bernard Lyon 1, Institut Polytechnique de Paris, Paris, 2018.

[13] Damien LOUREIRO. Utilisation de la dsn et de l'open data pour élaborer et expliquer un zonier incapacité. Master's thesis, ENSAE,Institut Polytechnique de Paris, Paris, 2021.

[14] Issam MEZRAG. Construction d'un zonier en assurance MRH. Master's thesis, ISUP, Sorbonne Université, Paris, 2018.

[15] Marjorie MIETTON. Construction de tables d'incidence et de maintien en Arrêt de Travail dans le cadre de l'assurance emprunteur. Master's thesis, ISFA, Université Clauder Bernard Lyon 1, Lyon, 2013.

[16] Le Moniteur. Le btp, irréductible bastion masculin. <https://www.lemoniteur.fr/article/le-btp-irreductible-bastion-masculin.974594>, 22 Décembre 2015.

[17] Antoine PAGLIA. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. Master's thesis, EURIA, Université de Bretagne Occidentale, Brest, 2010.

[18] Chloë Vestri. Elaboration d'une loi d'entrée en arrêt de travail et étude de l'impact de l'hétérogénéité sur le coefficient de sécurité. Master's thesis, ISFA, Université Clauder Bernard Lyon 1, Lyon, 2015.

Table des figures