

Mémoire présenté le 14/06/2018
pour l'obtention du diplôme de
Statisticien Mention Actuariat de l'ISUP
et l'admission à l'Institut des Actuaires

Par : Neil BELLAGHA

Titre : Modélisation de la sinistralité en incapacité d'un portefeuille de TNS et de salariés par apprentissage supervisé binaire

Confidentialité : oui, deux ans. *Les signataires s'engagent à respecter la confidentialité indiquée.*

Membre(s) présent(s) du jury de l'Institut des Actuaires :

Membre(s) présent(s) du jury de la filière :

Entreprise : Generali France

Signature :

Directeur de mémoire en entreprise :
Aurélie ADELE

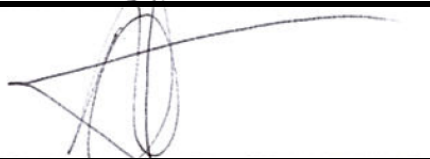
Signature :

Invité :

Signature :


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

Signature du responsable entreprise



Bibliothèque :

Signature du candidat



Secrétariat :

Remerciements

Tout d'abord, je tiens à remercier tout particulièrement ma tutrice en entreprise, Aurélie ADELE, pour ses conseils avisés et sa relecture attentive. Je remercie également mon tuteur académique, Olivier LOPEZ, pour l'orientation technique judicieuse qu'il a su donner à mon mémoire.

J'adresse ensuite mes remerciements chaleureux à Raphaël GUILMIN pour m'avoir fait confiance et m'avoir permis d'effectuer mon alternance au sein de Generali France. Ce fut pour moi une expérience très riche en enseignements, autant sur le plan des connaissances techniques que sur le plan des relations interpersonnelles.

D'une manière plus générale je tiens à remercier tous les membres de l'équipe qui m'a accueilli, aussi bien pour leur disponibilité que pour l'ambiance agréable qu'ils entretiennent et perpétuent au sein de la Direction Technique Vie, ainsi que pour l'intérêt qu'ils m'ont porté tout au long de mon alternance à travers leur aide et leurs précisions.

Je remercie également amis et famille pour leurs encouragements à persévérer, sincères et efficaces.

Résumé

Les méthodes d'apprentissage utilisées dans ce mémoire répondent au besoin de mieux connaître les variables à disposition de l'assureur (sources interne et externe) qui permettent d'expliquer et de prédire pour le risque incapacité les deux événements redoutés par tout assureur, à savoir d'une part l'occurrence d'au moins un sinistre incapacité et d'autre part un ratio prestations/primes supérieur à 100%, et cela sur une fenêtre d'observation allant de 2010 à 2016 (années pleines). Les résultats peuvent constituer des pistes d'optimisation de la tarification ou des majorations annuelles.

Dans une démarche à la fois prédictive et explicative, la performance évaluée sur l'échantillon d'apprentissage ainsi que sur l'échantillon test a été quantifiée pour comparer quatre modèles d'apprentissage automatique issus des arbres CART avec quatre modèles de régression logistique (apprentissage statistique). Le choix de telles méthodes d'apprentissage n'est pas anodin puisqu'elles permettent d'établir pour les variables étudiées des indicateurs de pertinence dans la discrimination des populations étudiées (sinistrés et non sinistrés ; sinistrés rentables et non rentables) comme l'importance, le Mean Decrease Entropie/Gini/Accuracy ou encore les odds ratios.

L'étude révèle que des variables capturant l'écoulement du temps comme l'ancienneté en portefeuille ainsi que des variables liées au cadre de vie général de l'assuré comme sa situation familiale, son domaine d'activité professionnelle et sa zone géographique de résidence sont les variables les plus discriminantes. Le maintien d'une sélection médicale (surprime) semble nécessaire, et la levée de l'interdiction de tarifer selon le sexe mérite d'être discutée.

Mots clés : Incapacité de travail, Apprentissage supervisé binaire, Rentabilité, Arbres CART, Régression logistique, Variables tarifaires.

Abstract

The supervised learning methods used in this study try to satisfy the need to know more about the variables accessible to the insurer (internal and external sources) that can explain and predict the two events feared by any insurer regarding the incapacity for work, namely the occurrence of at least one insurance claim on the one hand, and a claims-to-premiums ratio greater than 100% on the other hand, and this over an observation window from 2010 to 2016 (full years). The results can be used to optimize pricing or annual increases of the premium.

The approach, both predictive and explanatory, consists in quantifying the performance relying on the learning sample as well as on the test sample, in order to compare four machine learning models derived from the CART trees with four logistic regression models (statistical learning). The choice of such learning methods is not insignificant since they make it possible to establish indicators of relevance of the variables in the discrimination of the studied populations (policyholders in a situation of incapacity for work and those who are not ; profitable and unprofitable policyholders in a situation of incapacity for work) like the importance, the Mean Decrease Entropy/Gini/Accuracy criterion or the odds ratios.

The study reveals that variables capturing the passage of time such as contract lifespan as well as variables related to the general living environment of the policyholder as his family situation, his professional activity and his geographical area of residence are most discriminant variables. The maintenance of a medical selection (extra premium) seems necessary, and the lifting of the prohibition to price by sex deserves to be discussed.

Key words : Incapacity for work, Binary supervised learning, Profitability, CART trees, Logistic regression, Pricing variables.

Note de synthèse

La tarification au plus juste est un enjeu important dans l'équilibre financier d'une entreprise d'assurance. Ainsi, l'établissement d'un tarif sur-mesure en fonction du profil des assurés doit se faire après une prise de conscience et une quantification des facteurs porteurs du risque : c'est le fil directeur de cette étude qui porte sur le risque **incapacité de travail**. Par ailleurs, la grande applicabilité des méthodes d'apprentissage, qu'il soit automatique ou statistique, permet une approche originale de détermination de la pertinence des variables caractérisant les assurés. Pour cela, nous avons choisi de modéliser successivement deux aspects binaires de la sinistralité motivés par les deux événements critiques redoutés par tout assureur à savoir :

— la survenance d'au moins un sinistre :

$$Y = \begin{cases} 1 & \text{si l'assuré a été sinistré au moins une fois en incapacité} \\ 0 & \text{si l'assuré n'a jamais été sinistré en incapacité} \end{cases}$$

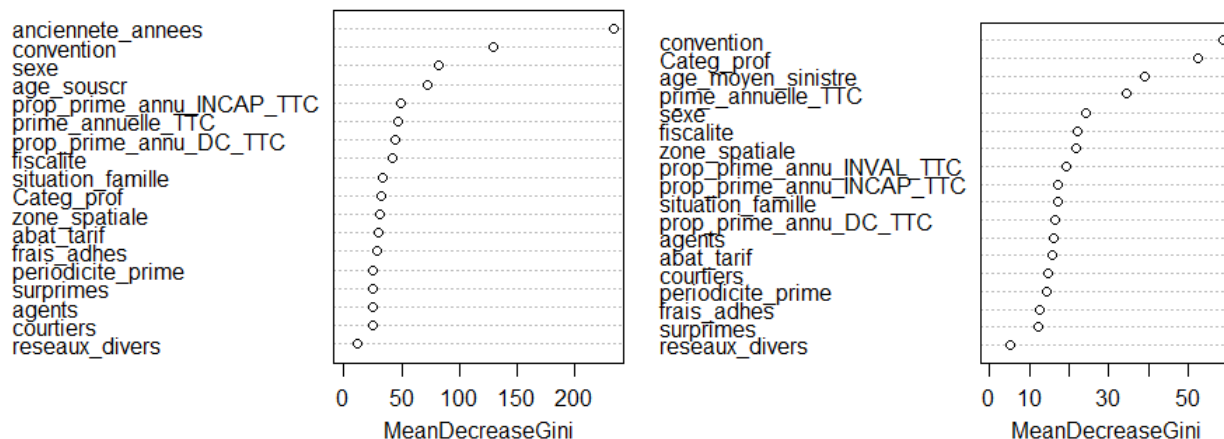
— un ratio prestations/primes en incapacité supérieur à 100% : $Y = \mathbb{1}_{S/P > 100\%}$

Une première idée de l'influence des variables retenues dans la modélisation sur les deux phénomènes étudiés peut être faite en observant le tableau suivant :

VARIABLES	MODALITÉS - portefeuille des assurés	assurés sinistrés total des assurés	MODALITÉS - portefeuille des sinistrés	sinistrés non rentables total des sinistrés
Zone Spatiale	Pôle intérieur/Arc extérieur	14.5% / 17.5%	Pôle intérieur/Arc extérieur	64.3% / 73.1%
Convention (travailleurs non salariés vs salariés)	professions médicales	17.1%	professions médicales	72.6%
	professions paramédicales	23.1%	professions paramédicales	74.4%
	profession libérales/expert/conseil	13%	profession libérales/expert/conseil	78.0%
	artisans - commerçants	13.8%	artisans - commerçants	64.0%
	professions agricoles	23.2%	professions agricoles	53.2%
	salariés	19.1%	salariés	70.1%
Sexe	femme/homme	21.0% / 13.8%	femme/homme	74.9% / 62.4%
Situation de Famille	en couple/seul(e)	17.2% / 16.3%	en couple/seul(e)	65.8% / 72.1%
Âge à la souscription	<=33 ans	18.2%	<=37 ans / >37 ans	76.8% / 58.0%
	33 ; 45 ans	14.9%		
	>45 ans	17.6%		
Âge Moyen Sinistre			<=40 ans / >40 ans	76.5% / 56.9%
Ancienneté (années)	<=3 ans / >3 ans	09.7% / 24.6%		
Abattement tarifaire	oui/non	13.9% / 18%	oui/non	75.8% / 66.5%
Catégorie Professionnelle	1 / 2-3	15.7% / 18.3%	1	75.0%
			2	64.0%
			3	45.6%
Surprimés	oui/non	21.0% / 16%	oui/non	63.7% / 69.2%
Agents généraux	oui (1)/non (0)	16.8% / 16.6%	oui/non	63.3% / 73.1%
Courtiers	oui (1)/non (0)	17.0% / 16.4%	oui/non	73.2% / 63.9%
Réseaux divers	oui (1)/non (0)	12.3% / 16.9%	oui/non	72.2% / 68.8%
Périodicité prime	mensuelle/non mensuelle	16.7% / 17.3%	mensuelle/non mensuelle	69.7% / 62.0%
Frais d'adhésion	oui/non	10.7% / 17.4%	oui/non	67.3% / 69.0%
Fiscalité	Agricole	23.7%	Agricole	52.2%
	Assurance vie	18.1%	Assurance vie	72.4%
	Madelin	16.2%	Madelin	69.4%
Prime annuelle moyenne TTC (€)	<=900 € / >900 €	14.3% / 19.4%	<=722 € / >722 €	80.9% / 62.4%
Proportion (prime) Incapacité	<=46% / >46%	12.5% / 19%	<=48% / >48%	72.9% / 67.0%
Proportion (prime) Décès	<=25% / >25%	19.2% / 13.7%	<=28% / >28%	70.3% / 66.8%
Proportion (prime) Invalidité			<=28% / >28%	65.9% / 75.8%
Taux de travail à temps plein	<=30% / >30%	20.6% / 13.5%		
Espérance de vie à la naissance	<=83.3 ans / >83.3 ans	14.3% / 20.6%	<=83 ans / >83 ans	62.5% / 75.3%
Personnes par ménage	<=2.2 / >2.2	19.5% / 14.5%	<=2 / >2	77.5% / 63.6%

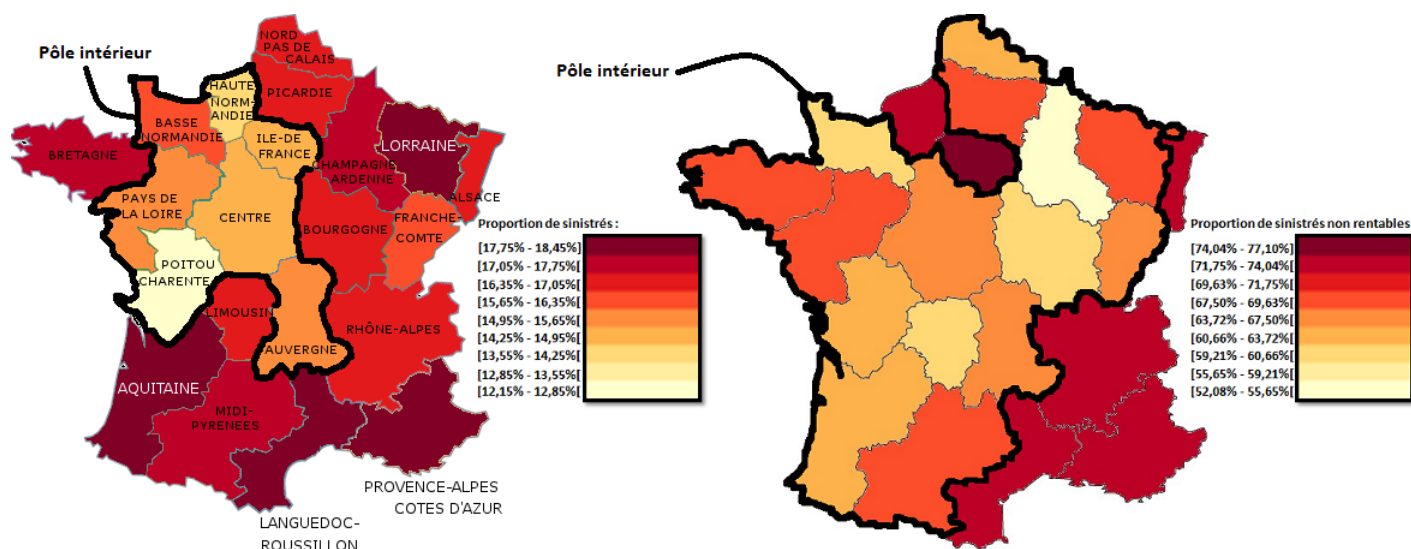
Variables finales utilisées pour la modélisation - La proportion de sinistrés de tout le portefeuille est de 16.7%, et la proportion de sinistrés non rentables de tout le portefeuille est de 68.89%.

Bien que les effectifs de chaque modalité pondérant les proportions de sinistrés et de sinistrés non rentables ne sont pas reportés dans le tableau, il est tout à fait possible de se rendre compte que certaines variables comme l'ancienneté en portefeuille, la convention (i.e. l'activité professionnelle) ou encore la prime annuelle moyenne TTC toutes garanties confondues sont déterminantes dans notre étude puisque l'une et/ou l'autre des deux proportions étudiées varient significativement d'une modalité à l'autre de ces variables. Ce constat est confirmé, à l'issue de la modélisation, par le critère d'importance provenant des forêts aléatoires :



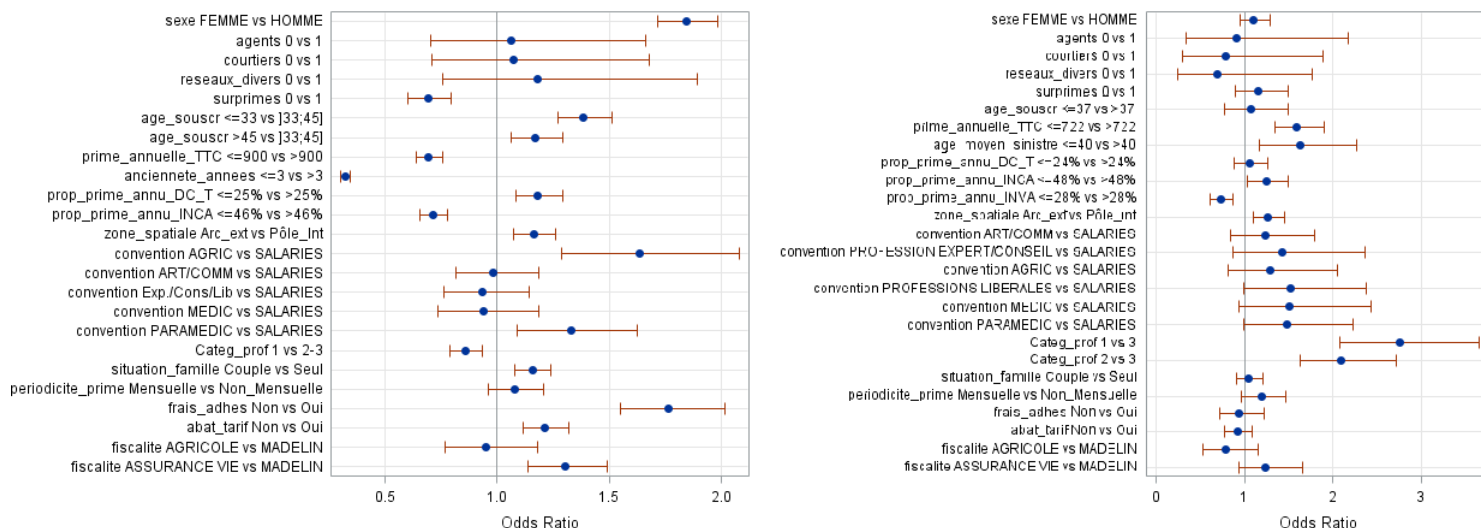
Mean Decrease Gini à l'issue de la procédure des forêts aléatoires - à gauche : modélisation de la proportion de sinistrés ; à droite : modélisation de la proportion de sinistrés non rentables.

La zone géographique présente elle aussi un pouvoir de discrimination de la sinistralité puisqu'elle permet de distinguer deux regroupements de régions (avant la réforme territoriale effective à partir du 1^{er} janvier 2016) : une première zone constituée notamment des régions Centre, Basse-Normandie, Pays-de-la-Loire, Poitou-Charente et Auvergne, et une seconde zone composée entre autres des régions Rhône-Alpes, Languedoc-Roussillon, Alsace, Provence-Alpes-Cotes d'azur et dans une certaine mesure l'Ile-de-France.



Cartographie des 21 régions en fonction de la proportion d'assurés sinistrés (gauche) et de la proportion de sinistrés non rentables (droite) du portefeuille.

Les odds ratios issus de la modélisation par régression logistique permettent de confirmer que, toutes choses égales par ailleurs, la zone extérieure présente un risque plus élevé, à la fois pour l'occurrence d'au moins un sinistre incapacité et pour la non-rentabilité, que le pôle intérieur. Toutes ces disparités géographiques méritent donc d'être prises en compte dans la tarification à la souscription ou lors des majorations tarifaires annuelles.



Odds Ratios à l'issue de la régression logistique - à gauche : modélisation de la proportion de sinistrés ; à droite : modélisation de la proportion de sinistrés non rentables.

Cependant, la variable la plus prépondérante dans toute l'analyse de l'occurrence d'au moins un sinistre incapacité est l'ancienneté en portefeuille, puisqu'un assuré de plus de 3 ans d'ancienneté a significativement beaucoup plus de chances d'avoir un premier sinistre incapacité. Une majoration annuelle en conséquence, passé ce marqueur temporel, permettrait d'optimiser la rentabilité du portefeuille en chargeant la prime des plus anciens assurés, à juste titre puisqu'ils ont un risque aggravé d'être sinistré lié à l'augmentation de leur âge et de la fenêtre d'exposition. Une majoration encore plus ciblée consisterait à majorer d'avantage les assurés qui ont à la fois 3 ans d'ancienneté en portefeuille et qui n'ont pas encore passé le seuil de 40 ans d'âge civil, puisque les plus jeunes, d'après la variable « âge moyen sinistre », semblent être les moins rentables selon la tarification antérieure.

Si l'on s'intéresse à la performance des méthodes d'apprentissage, et que l'on s'appuie sur les 3 scores que sont l'erreur d'apprentissage R_n^{appr} , le taux de succès global (TSG, i.e. le complémentaire à 1 de l'erreur de prévision R_n^{test}) et le taux de vrais positifs (TVP), on constate que les méthodes d'apprentissage automatique autant que les méthodes d'apprentissage statistique sont appropriées. En effet, bien qu'adaboost affiche, dans tous les cas, de bonnes erreurs d'apprentissage (comparativement aux autres méthodes), et qu'elle forme avec les forêts aléatoires un couple de méthodes qui maximisent le TVP respectivement pour la modélisation de l'occurrence d'au moins un sinistre et la modélisation de la non rentabilité, la régression logistique n'est pas en reste.

Modèles d'apprentissage	Apprentissage	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Arbre complet	0.1095	0.1744	0.7929	0.1771	0.0823	0.0949	0.6224
Arbre optimal	0.1634	0.1313	0.8333	0.0480	0.0075	0.0405	0.6617
Forêts aléatoires	0.1666	0.1544	0.8319	0.0048	0.0004	0.0044	0.6311
Adaboost	0.1118	0.1696	0.7840	0.2435	0.1064	0.1370	0.6437
Régression logistique sans interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
modèle complet	0.1650	0.1296	0.8323	0.0310	0.0052	0.0258	0.7053
sélection de variables bi-directionnelle	0.1647	0.1296	0.8325	0.0327	0.0053	0.0274	0.7048
Régression logistique avec interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
modèle complet	0.1618	0.1285	0.8342	0.0781	0.0125	0.0656	0.7117
sélection de variables bi-directionnelle	0.1627	0.1285	0.8353	0.0702	0.0096	0.0606	0.7122

Moyennes (obtenues en réitérant 100 fois l'échantillonnage) des scores de performance des modèles de classification binaire des assurés selon la survenance ou non de sinistres, sans recours aux données externes. BS=Score de Brier ; TSG=Taux de Succès Global ; TVP=Taux de Vrais Positifs ; TFP=Taux de Faux Positifs ; PSS=Score de Pierce ; AUC=Area Under Curve.

Modèles d'apprentissage	Apprentissage	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Arbre complet	0.1133	0.3126	0.6302	0.7364	0.6248	0.1116	0.5702
Arbre optimal	0.2896	0.2013	0.6904	0.8723	0.7459	0.1264	0.6180
Forêts aléatoires	0.2822	0.2243	0.7054	0.9470	0.8744	0.0726	0.6379
Adaboost	0.1152	0.2205	0.6689	0.8206	0.6951	0.1255	0.6105
Régression logistique sans interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
modèle complet	0.2967	0.1973	0.7001	0.9128	0.8102	0.1026	0.6591
sélection de variables bi-directionnelle	0.2969	0.1969	0.6988	0.9115	0.8117	0.0999	0.6557
Régression logistique avec interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
modèle complet	0.5131	0.4584	0.4934	0.4075	0.3004	0.1070	0.5535
sélection de variables bi-directionnelle	0.2756	0.2083	0.6821	0.8480	0.7160	0.1320	0.6467

Moyennes (obtenues en réitérant 100 fois l'échantillonnage) des scores de performance des modèles de classification binaire des sinistrés selon leur rentabilité (S/P inférieur ou supérieur à 100%), sans recours aux données externes.

Ainsi, la prise en compte d'interactions d'ordre deux permet au GLM de maximiser le TSG pour la modélisation de l'occurrence d'au moins un sinistre, tandis que l'absence de prise en compte de ce type d'interactions permet à la régression logistique de performer quasiment aussi bien que les forêts aléatoires pour la modélisation de la non rentabilité, eu égard au TSG et au TVP.

S'agissant des données externes, la prise en compte des deux variables externes les plus liées à la variable cible selon le V de Cramer permet d'améliorer simultanément le taux de vrais positifs (TVP) des quatre méthodes d'apprentissage automatique, dans le cadre d'une modélisation de la non-rentabilité. Cependant, la portée limitée de l'impact de l'apport de données externes à la modélisation provient très vraisemblablement du faible nombre de variables finalement retenues (le contraire aurait accru considérablement la complexité de la régression logistique) et de la grande différence de granularité des données du portefeuille (assurés) et des données externes (régions françaises). Les assureurs gagneraient donc à réfléchir à une manière moins classique de tarifier, basée sur l'environnement et le mode de vie des assurés (alimentation, activité physique,...).

Summary note

Fair pricing is an important issue in the financial stability of an insurance company. Thus, the establishment of a made to measure premium according to the profile of the policyholders must be done after realizing and a quantifying the factors « carrying » the risk : it is the guideline of this study which deals with the **incapacity for work**. Moreover, the great applicability of learning methods, whether machine learning or statistical learning, allows an original approach to determine the relevance of the variables characterizing the policyholders. To this end, we have chosen to model successively the two critical events feared by any insurer :

— the occurrence of at least one insurance claim :

$$Y = \begin{cases} 1 & \text{if the policyholder has been in a situation of incapacity for work at least once} \\ 0 & \text{if the policyholder has never been in a situation of incapacity for work} \end{cases}$$

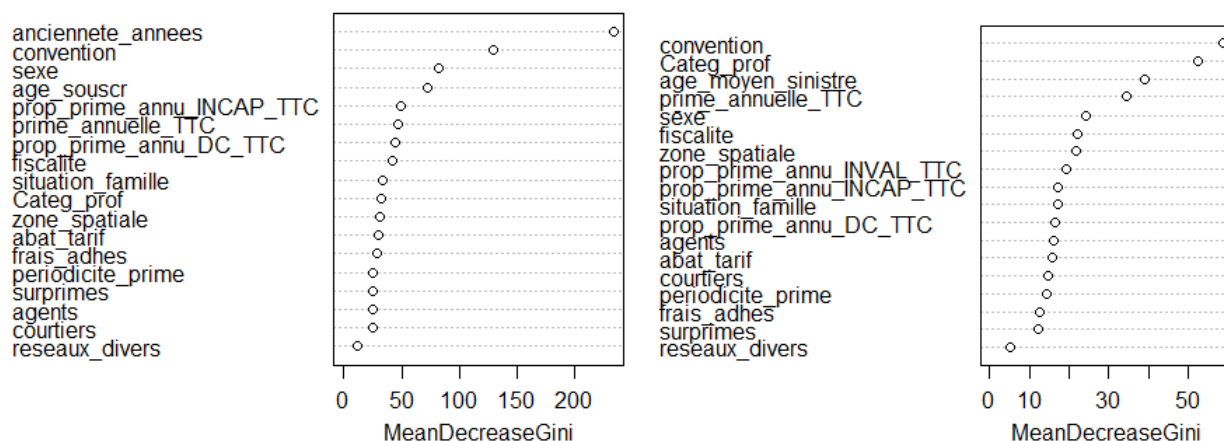
— a claims-to-premiums ratio greater than 100% : $Y = \mathbb{1}_{\text{claims-to-premiums ratio} > 100\%}$

A first idea of the influence of the variables selected in the modeling on the two phenomena studied can be made by observing the following table :

VARIABLES	CATEGORIES - portfolio of all policyholders	policyholders in a situation of incap. for work all policyholders	CATEGORIES - portfolio of all policyholders in a situation of incapacity for work	unprofitable policyholders in a sit. of incapacity for work all policyholders in a sit. of incap. for work
Spatial area	Interior/Exterior	14.5% / 17.5%	Interior/Exterior	64.3% / 73.1%
Profession (self-employed vs employees)	Medical professions	17.1%	Medical profession	72.6%
	Medical professions	23.1%	Paramedical profession	74.4%
	Liberal professions/ Experts advisers	13%	Liberal professions	78.0%
	Experts advisers		Experts advisers	83.0%
	Craftsmen - traders	13.8%	Craftsmen - traders	64.0%
	Agricultural profession	23.2%	Agricultural profession	53.2%
Sexe	Woman/Man	21.0% / 13.8%	Woman/Man	74.9% / 62.4%
Family status	In a relationship/single	17.2% / 16.3%	In a relationship/single	65.8% / 72.1%
Age at entry	<=33 years-old	18.2%	<=37 y.-o. / >37 y.-o.	76.8% / 58.0%
	33 ; 45 years-old	14.9%		
	>45 years-old	17.6%		
Average age of incapacity for work			<=40 y.-o. / >40 y.-o.	76.5% / 56.9%
Lifespan of the contract	<=3 y.-o. / >3 y.-o.	09.7% / 24.6%		
Reduction of the premium	yes/no	13.9% / 18%	yes/no	75.8% / 66.5%
Professional categories	1 / 2-3	15.7% / 18.3%	1	75.0%
			2	64.0%
			3	45.6%
Extra premium	yes /no	21.0% / 16%	yes /no	63.7% / 69.2%
General agents	yes (1)/no (0)	16.8% / 16.6%	yes (1)/no (0)	63.3% / 73.1%
Insurance broker	yes (1)/no (0)	17.0% / 16.4%	yes (1)/no (0)	73.2% / 63.9%
All other business providers	yes (1)/no (0)	12.3% / 16.9%	yes (1)/no (0)	72.2% / 68.8%
Periodicity of premium	by the month/lower frequency	16.7% / 17.3%	by the month/lower frequency	69.7% / 62.0%
Membership fees	yes/no	10.7% / 17.4%	yes/no	67.3% / 69.0%
Tax system	Agricultural	23.7%	Agricultural	52.2%
	Life assurance	18.1%	Life assurance	72.4%
	Madelin	16.2%	Madelin	69.4%
Average annual premium including taxes (€)	<=900 € / >900 €	14.3% / 19.4%	<=722 € / >722 €	80.9% / 62.4%
Incapacity proportion (premium)	<=46% / >46%	12.5% / 19%	<=48% / >48%	72.9% / 67.0%
Death proportion (premium)	<=25% / >25%	19.2% / 13.7%	<=28% / >28%	70.3% / 66.8%
Invalidity proportion (premium)			<=28% / >28%	65.9% / 75.8%
Proportion of full-time work	<=30% / >30%	20.6% / 13.5%		
Life expectancy at birth	<=83.3 y.-o. / >83.3 y.-o.	14.3% / 20.6%	<=83 y.-o. / >83 y.-o.	62.5% / 75.3%
Number of people in household	<=2.2 / >2.2	19.5% / 14.5%	<=2 / >2	77.5% / 63.6%

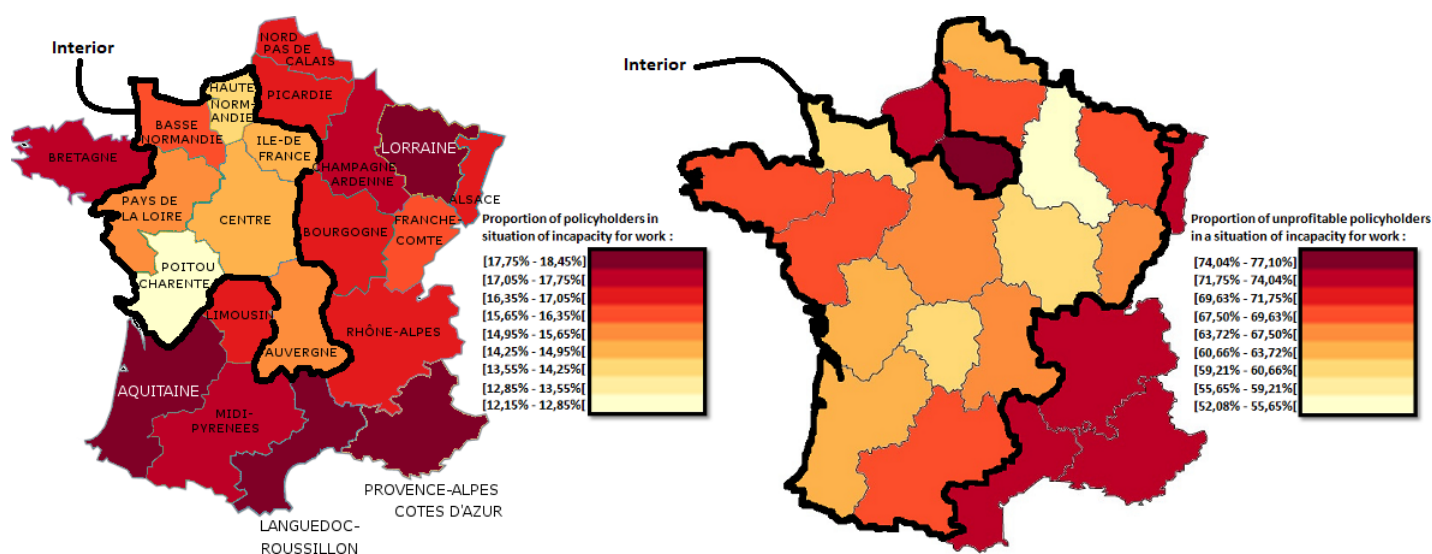
All the variables used - The average proportion of policyholders in a situation of incapacity for work of the entire portfolio is 16.7%, and the average proportion of unprofitable policyholders in a situation of incapacity for work of the entire portfolio is 68.89%.

Although the headcount of each modality weighting the proportions of policyholders in a situation of incapacity for work and those who are unprofitable are not reported in the table, it is possible to realize that certain variables such as contract lifespan, the profession or the average annual premium is decisive in our study since one or other of the two proportions studied vary significantly from one modality to another of these variables. This is confirmed, at the end of the modeling, by the importance criterion coming from random forests :



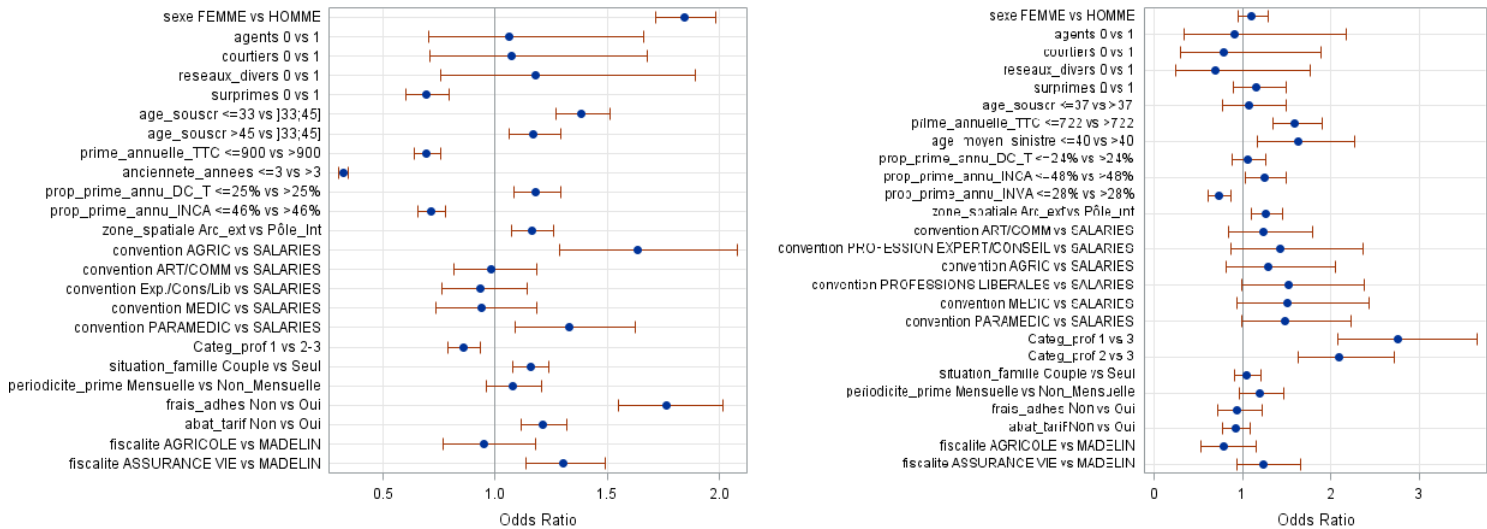
Mean Decrease Gini criterion coming from random forests - left : modeling of the proportion of policyholders in a situation of incapacity for work; right : modeling of the proportion of unprofitable policyholders in a situation of incapacity for work.

The geographical area also is a good variable to discriminate the populations of interest since two groups of regions are distinguished : a first zone composed of the Center, Basse-Normandie, Pays-de-la-Loire, Poitou-Charentes and Auvergne, and a second zone composed of other regions such as Rhône-Alpes, Languedoc-Roussillon, Alsace, Provence-Alpes-Cote d'azur and to a certain extent Ile-de-France.



Mapping of the 21 regions according to the proportion of policyholders in a situation of incapacity for work (left) and the proportion of unprofitable policyholders in a situation of incapacity for work (right) of the portfolio.

The odds ratios coming from the logistic regression confirm that, all other things being equal, the exterior presents a higher risk, both for the occurrence of at least one claim and for the non-profitability, than the interior. All these geographical disparities deserve to be taken into account in subscription pricing or in annual premium increases.



Odds Ratios coming from logistic regression - left : modeling of the proportion of policyholders in a situation of incapacity for work ; right : modeling of the proportion of unprofitable policyholders in a situation of incapacity for work.

However, the most prevalent variable in the entire analysis of the occurrence of at least one claim is the contract lifespan, since a policyholder with more than 3 years of seniority is significantly more likely to have an incapacity for work for the first time. An annual increase, beyond this time marker, would optimize the profitability of the portfolio by charging the premium rightly since they have an increased risk of being in a situation of incapacity for work due to the increase of their age and the exposure window. An even more targeted extra premium would consists in charging the premium of the policyholders who have three years of seniority and who have not yet passed the age of 40, since the youngest are the least profitable according to the variable « average age of incapacity for work ».

Regarding the performance of learning methods, if we rely on the 3 scores that are the learning error R_n^{appr} , the overall success rate (TSG, i.e. equal to 1 minus the prediction error R_n^{test}) and the True Positive Rate (TVP), both machine learning and statistical learning methods are appropriate. Indeed, although adaboost displays, in all cases, good learning errors (compared to other methods), and although it forms with the random forests a couple of methods that maximize the TVP respectively for the modeling of the the occurrence of at least one claim and the modeling of non-profitability, logistic regression is not outdone.

Supervised learning Models	Learning	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Full tree	0.1095	0.1744	0.7929	0.1771	0.0823	0.0949	0.6224
Optimal tree	0.1634	0.1313	0.8333	0.0480	0.0075	0.0405	0.6617
Random Forest	0.1666	0.1544	0.8319	0.0048	0.0004	0.0044	0.6311
Adaboost	0.1118	0.1696	0.7840	0.2435	0.1064	0.1370	0.6437
Logistic regression without interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
full model	0.1650	0.1296	0.8323	0.0310	0.0052	0.0258	0.7053
bidirectional selection of variables	0.1647	0.1296	0.8325	0.0327	0.0053	0.0274	0.7048
Logistic regression with interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
full model	0.1618	0.1285	0.8342	0.0781	0.0125	0.0656	0.7117
bidirectional selection of variables	0.1627	0.1285	0.8353	0.0702	0.0096	0.0606	0.7122

Averages (obtained by repeating the sampling 100 times) performance scores of the binary classification models of policyholders according to the occurrence or not of at least one claim, without recourse to external data. BS=Brier score; TSG=overall success rate; TVP=True Positive Rate; TFP=False Positive Rate; PSS=Pierce score; AUC=Area Under Curve.

Supervised learning models	Learning	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Full tree	0.1133	0.3126	0.6302	0.7364	0.6248	0.1116	0.5702
Optimal tree	0.2896	0.2013	0.6904	0.8723	0.7459	0.1264	0.6180
Random Forest	0.2822	0.2243	0.7054	0.9470	0.8744	0.0726	0.6379
Adaboost	0.1152	0.2205	0.6689	0.8206	0.6951	0.1255	0.6105
Logistic regression without interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
full model	0.2967	0.1973	0.7001	0.9128	0.8102	0.1026	0.6591
bidirectional selection of variables	0.2969	0.1969	0.6988	0.9115	0.8117	0.0999	0.6557
Logistic regression with interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
full model	0.5131	0.4584	0.4934	0.4075	0.3004	0.1070	0.5535
bidirectional selection of variables	0.2756	0.2083	0.6821	0.8480	0.7160	0.1320	0.6467

Averages (obtained by repeating the sampling 100 times) performance scores of the binary classification models of policyholders in a situation of incapacity for work according to their profitability (loss ratio above or under 100%), without recourse to external data.

Thus, taking second-order interactions into account allows the GLM to maximize the TSG for modeling the occurrence of at least one claim, while the absence of second-order interactions allows logistic regression to perform almost as well as random forests for non-profitability modeling, relying on TSG and TVP.

With regard to external data, taking into account the two variables that are most related to the target variable according to Cramer criterion simultaneously improves the True Positive Rate (TVP) of the four machine learning methods when modeling non-profitability. However, the limited extent of the impact of the external data to the modeling is very likely due to the small number of external variables finally selected (the opposite would have considerably increased the complexity of the logistic regression) and the big difference in granularity of the variables of the portfolio (policyholders) and the variables of the external data (french regions). Insurers would therefore benefit from thinking about a less traditional way of pricing, based on the environment and the lifestyle of the policyholders (nutrition, physical activity,...).

Table des matières

Liste des abréviations	3
Introduction	4
I Cadre de l'étude	6
1 Enjeux et démarche statistiques	6
1.1 Contours statistiques de l'objectif actuariel	6
1.2 Choix des méthodes d'apprentissage au regard de l'objectif	6
1.3 Adéquation des méthodes à la topologie des données	7
1.4 Méthodologie statistique	7
2 Présentation du portefeuille étudié	8
2.1 Présentation de l'ancien et du nouveau produit	8
2.2 Les données à disposition et leur pertinence	11
3 Généralités et formalisme du contexte d'apprentissage	21
3.1 Performance théorique	22
3.2 Performance empirique	24
3.3 Pouvoir explicatif et pouvoir prédictif	25
II Analyse exploratoire, transformation et échantillonnage des données . . .	27
1 Distributions des variables explicatives et lien avec la variable cible	27
1.1 Les variables quantitatives et leur discrétisation	27
1.2 Les variables qualitatives et le regroupement de leurs modalités	31
1.3 Vue globale des liaisons après transformations	34
2 Validité et traitement des données externes	39
3 Échantillonnage : Apprentissage et test	42
III Classification binaire des assurés selon la survenance de sinistres	43
1 Arbres CART	43
1.1 Critère d'impureté	45
1.2 Critères d'arrêts	47
1.3 Élagage de l'arbre complet par validation croisée	49
1.4 Importance des variables	52
2 Agrégation d'arbres CART	53
2.1 Motivations	53
2.2 Bagging d'arbres, et forêts aléatoires	54
2.3 Boosting d'arbres binaires	58
3 Comparaison avec la régression logistique	61
3.1 Définitions et Hypothèses	61
3.2 Estimation des paramètres	63
3.3 Modélisation des interactions	68

4	Synthèse du pouvoir explicatif des modèles	69
5	Prévision de l'échantillon test (pouvoir de généralisation)	70
5.1	Scores de prédiction	70
5.2	Courbe ROC et AUC	72
6	Apport des données externes	74
IV	Étude de la rentabilité des sinistrés par classification binaire	75
1	Panorama des données transformées	76
2	Pouvoir explicatif des modèles	81
3	Pouvoir de généralisation des modèles	85
4	Contribution des données externes	86
	Conclusion	88
	Annexes	90
A	- Excès de risque en classification binaire	91
B	- Distribution des variables utilisées pour la modélisation de la rentabilité	92
C	- Graphiques supplémentaires relatifs à la modélisation de la rentabilité	96
D	- Compléments sur la régression logistique	97
E	- Nomenclature des garanties en incapacité	98
	Table des figures	101
	Bibliographie	103

Liste des abréviations

ACP	Analyse en Composantes Principales
AFCM	Analyse Factorielle des Correspondances Multiples
AIC	Akaike Information Criterion
AUC	<i>Area Under Curve</i> (Aire sous la courbe ROC)
CART	<i>Classification And Regression Trees</i>
CP	Catégories Professionnelles
DC	Décès
Éch.	Échantillon
EMV	Estimateur du Maximum de Vraisemblance
FP	Frais Professionnels
GLM	Generalized Linear Model
IdF	Île-de-France
i.i.d.	indépendantes et identiquement distribuées
IJ	Indemnités Journalières
Incap	Incapacité
Inval	Invalidité
LGN	Loi des Grands Nombres
MDE/G	Mean Decrease Entropie/Gini
MDA	Mean Decrease Accuracy
MRE	Minimiseur du Risque Empirique
MSA	Mutualité Sociale Agricole
MV	Maximum de Vraisemblance
OOB	Out-Of-Bag
OR	Odd Ratio
PACA	Provence-Alpes-Côte d'Azur
ROC	<i>Receiver Operating Characteristic</i>
RPS	Risques Psychosociaux
RSI	Régime Social des Indépendants
TDC	Tableau Disjonctif Complet
TNS	Travailleur(s) Non Salarié(s)
TTC	Toutes Taxes Comprises
WMW	Wilcoxon-Mann-Whitney

Introduction

L'état d'incapacité est défini comme l'impossibilité physique ou psychique, partielle ou totale, d'exercer son activité professionnelle, que cela soit dû à une maladie ou un accident. L'assurance en cas d'incapacité de travail consiste donc à verser des prestations financières ayant vocation à recouvrir partiellement la perte de revenus qui est la conséquence de l'arrêt du travail. Il s'agit en France, dans le contexte d'une compagnie d'assurances comme Generali, de la part complémentaire qui vient en relais de celle versée par la branche Assurance Maladie de la Sécurité Sociale pour les salariés, et par exemple du Régime Social des Indépendants (RSI) pour les travailleurs non salariés non agricoles, et de la Mutualité Sociale Agricole (MSA) pour les TNS agricoles.

En 2016, la société Rehalto du groupe SCOR a publié une [enquête sur les arrêts de travail concernant des salariés](#). Selon cette étude, 20% des salariés arrêtés estiment que leur arrêt est imputable à leur travail : 14 % des arrêts sont dus à des tensions liées à l'organisation du travail et 6% surviennent suite à des difficultés liées aux pratiques managériales de l'entreprise. La charge de travail et l'environnement physique sont cités comme les principales sources de difficultés pour les salariés arrêtés suite à des tensions liées au travail. Les troubles psychologiques (14%) constituent le quatrième motif d'arrêts de travail (après notamment les maladies ordinaires et les troubles musculo squelettiques), et sont plus élevés dans les services. L'étude des incapacités de travail est donc intéressante à plus d'un titre, puisque c'est un risque dont la réalisation est fréquente, récurrente, potentiellement coûteuse et ayant plusieurs sources, physiques ou contextuelles.

L'objectif de ce mémoire est, pour un produit de prévoyance concernant les Travailleurs Non Salariés (TNS) ainsi que les salariés, **d'étudier les caractéristiques des assurés qui influent sur leur sinistralité en Incapacité** entre 2010 et 2016, et cela au moyen de méthodes d'apprentissage supervisé. La sinistralité présente ici un double aspect : d'une part, le premier aspect qui permet de partager la population des assurés en sinistrés et en non sinistrés, est la survenance ou non d'au moins un sinistre, et d'autre part, le deuxième aspect qui permet d'apprécier l'intensité des sinistres dès qu'ils surviennent est la rentabilité (sinistrés rentables et non rentables). Le déroulé de ce mémoire sera donc dicté par cette dualité de la sinistralité.

Si, pour le premier aspect de la sinistralité, le caractère binaire de la segmentation qui nous intéresse (sinistrés et non sinistrés) impose une modélisation par apprentissage supervisé binaire, le deuxième aspect relatif à la rentabilité laisse quant à lui la possibilité d'opérer une régression (avec comme variable cible réelle positive le ratio prestations/primes) autant qu'un apprentissage supervisé binaire. Pour plus de robustesse aux valeurs aberrantes et aux anomalies de la base d'étude, ainsi que pour une meilleure préservation de la confidentialité des

chiffres, nous opterons donc pour ce dernier choix, afin d’opposer les sinistrés rentables aux sinistrés non rentables (i.e. un ratio prestations/primes respectivement inférieur et supérieur à 100%), d’autant plus qu’il permet d’estimer une probabilité d’être non rentable, ce qui est l’évènement critique redouté par tout assureur.

Nous tâcherons au cours de ce mémoire à la fois d’étudier la performance des méthodes utilisées, habituellement mises à profit pour établir des tarifs en Non-Vie (modèles coût-fréquence), que nous comparerons au traditionnel GLM, et, en même-temps, de dire dans quelle mesure telle ou telle caractéristique de l’assuré est pertinente pour l’explication de la sinistralité en incapacité et est donc dans une certaine mesure utile pour la tarification de cette garantie. L’intérêt d’une telle démarche d’appréhension et de compréhension de la sinistralité, qui s’inscrit dans un cadre de lancement de nouveau produit, est motivée par un contexte hyperconcurrentiel, où est soulevée la pertinence de la sélection médicale et celle de la tarification géographique.

Aussi, bien que l’intérêt d’une maîtrise de la rentabilité vient notamment de la réglementation en vigueur depuis le 1^{er} janvier 2016 dans le cadre de la réforme Solvabilité II, qui repose sur trois piliers essentiels à une bonne solvabilité des entités assurantielles, il est pourtant ici fait abstraction de toutes contraintes réglementaires, et notamment celle (cf. article [\[Poi\]](#)) interdisant de tarifier selon le sexe par exemple. En effet, le but n’étant pas de produire un tarif par catégories d’assurés mais uniquement de comprendre la structure relative des variables explicatives de la sinistralité du portefeuille d’intérêt.

De toute évidence, pour prétendre l’exhaustivité, cette étude devrait s’appuyer sur plus de caractéristiques relatives aux assurés, comme le fait de savoir s’ils ont ou non une bonne hygiène de vie, par exemple. L’absence de ce genre de variables tarifaires dans notre portefeuille est lacunaire, c’est donc pour cette raison que nous allons recourir aux données externes et ainsi apporter de la richesse à la base d’étude, bien que l’information apportée par les données externes ne soit pas intrinsèque aux assurés. Cet apport externe de données constitue une originalité de ce mémoire, outre l’application de méthodes d’apprentissage supervisé au risque Incapacité.

I - Cadre de l'étude

1 - Enjeux et démarche statistiques

1.1 - Contours statistiques de l'objectif actuariel

Il convient en tout premier lieu de préciser l'objectif de modélisation de cette étude parmi les trois suivants. En effet, la modélisation d'une variable Y à prévoir à partir d'une autre X peut avoir un objectif :

1. **purement explicatif**, c'est-à-dire montrer que le facteur X agit sur la variable Y , ce qui nécessite une démarche exhaustive et très rigoureuse, avec éventuellement la mise en place de plans d'expérience,
2. **de prévision mais explicatif**, i.e. en plus de prévoir, on cherche à savoir quelles sont les variables qui sont importantes pour la prévision et comment elles y aident,
3. **de prévision pure**, où l'on ne s'occupe pas de savoir comment fonctionnent les méthodes, l'essentiel étant d'avoir une bonne qualité de prévision, moyennant une optimisation efficace de leurs paramètres respectifs.

Dans notre cas précis, nous sommes guidés par un objectif du **deuxième type**, puisque ce mémoire a pour but d'aider l'assureur, lors de la **souscription ou du renouvellement annuel** de contrat pour la garantie **incapacité**, à prédire l'**occurrence** et la rentabilité (binaires) de la sinistralité d'un assuré ou d'un assuré, abstraction faite des contraintes réglementaires, et ceci afin d'améliorer la **rentabilité**. Nous avons donc bien un objectif de **prévision**. Cependant, cela ne peut se faire sans une prise de recul sur les facteurs agissant sur les événements d'intérêt (occurrence d'au moins un sinistre d'une part, et non-rentabilité d'autre part), ceci afin de comprendre les facteurs de risques et ainsi mieux appréhender le risque. Il y a donc une dimension **interprétative** à notre objectif.

1.2 - Choix des méthodes d'apprentissage au regard de l'objectif

Pour atteindre cet objectif à deux dimensions, et dans un souci de maîtrise et de compréhension des méthodes utilisées, nous avons décidé au préalable de recourir à des méthodes relativement simples et qui permettent de mesurer d'une manière ou d'une autre l'impact de chaque variable explicative sur la variable cible. Ces méthodes d'apprentissage automatique dérivent des arbres CART, et nous les comparerons au traditionnel GLM issu de l'apprentissage statistique (cf infra section 3 page 21 pour le cadre de l'apprentissage). Outre la simplicité et l'interprétabilité, ce choix a priori se justifie par le fait qu'il n'y a pas de méthode uniformément meilleure. En effet, en fonction de l'échantillon, en fonction de la structure des données et de leur topologie, il est nécessaire d'essayer plusieurs méthodes pour bien sûr avoir une bonne

qualité d'estimation mais surtout et avant tout une bonne qualité de prévision. Ce travail de recherche, parfois laborieux (optimisation de paramètres) peut mener à des résultats décevants, ou à retenir des méthodes dont la compréhension est ardue, au risque d'empiéter sur le troisième type d'objectif (de prévision pure) mentionné ci-dessus.

1.3 - Adéquation des méthodes à la topologie des données

Comme nous venons de l'évoquer, il existe toute une panoplie de méthodes d'apprentissage, diverses et variées, ayant toutes en commun la construction de frontières, avec des propriétés et des formes particulières, de l'espace des variables explicatives, pour modéliser Y . Aussi, le second aspect à considérer lors d'une modélisation, outre l'objectif, et qui va guider le choix d'une méthode plutôt qu'une autre, est le **type des variables** que l'on va étudier : est-on en présence de p variables explicatives toutes **quantitatives**, toutes **qualitatives**, ou un **mélange** des deux ? De même, la variable à modéliser peut être quantitative ou qualitative. En fonction de tout cela, des méthodes fonctionnent et d'autres pas, quand certaines méthodes sont adaptées dans tous les cas.

La nécessité d'avoir une **pratique** et une **expérience** des modèles d'apprentissage est un point crucial, ce qui a été déterminant dans le cadre de ce mémoire puisque cela a conditionné l'utilisation des méthodes utilisées. En effet, celles qui ont été retenues ici et qui constituent le **coeur du mémoire**, dérivent du modèle linéaire général(isé) ou des arbres de décision. Elles ne sont donc pas des « boîtes noires » dont la compréhension et la maîtrise sont opaques, et elles marchent lorsque des variables quantitatives et qualitatives sont mélangées.

1.4 - Méthodologie statistique

La stratégie usuelle lors d'une modélisation commence par une **extraction** des données avec ou sans sondage¹. Dans notre cas, il n'est pas besoin de sondage, puisque le nombre d'assurés avoisine les quarante mille, quant au nombre de mouvements (de prestations, de primes), il n'est que de l'ordre de quelques millions². En revanche, le nombre de variables est conséquent, avec quelques centaines de variables au global des trois tables (cf. figure I.1 page 20). Il a donc fallu procéder à un tri des variables, en écartant, selon leur **signification**, les plus triviales (dates et heures de mise à jour des tables, codes apporteurs,...).

Nous n'abordons pas cet aspect d'extraction pour les données externes dans la mesure où elles proviennent d'institutions dignes de confiance comme l'[INSEE](#) et [Etalab](#).

Ensuite, viennent l'**exploration** (effectifs, modalités, données manquantes,...), le nettoyage, et surtout l'évaluation de la **qualité des données** d'autant plus que la réglementation autour de la réforme Solvabilité II en vigueur depuis le 1^{er} janvier 2016 repose sur trois piliers dont une des problématiques transversales est la nécessité d'avoir des aspirations élevées de ce point

1. Le sondage est utile si les données sont très volumineuses rendant ainsi difficile l'optimisation des paramètres.

2. ce qui est gérable avec le logiciel SAS®.

de vue. Malheureusement, à l'issue de cette étape, le manque de fiabilité avéré³ ou présumé de certaines variables, à priori très intéressantes comme la nature de la pathologie ou encore la probabilité de rechute, nous a amenés à les écarter. Les seules variables retenues sont donc celles présentant un sens actuariel et ne présentant ni valeurs manquantes ni anomalies.

En outre, une fois la base figée, la **transformation** des variables (quantitatives rendues qualitatives par exemple) peut être utile comme nous le verrons lors de l'analyse exploratoire des données au chapitre II page 27.

Enfin, vient la segmentation de l'échantillon entier en échantillons disjoints, d'une part d'apprentissage⁴ et d'autre part de test (respectivement de 66% et 33% ici). En effet, quelle que soit la méthode d'apprentissage considérée, il faut estimer les paramètres des modèles, et les optimiser soit sur un échantillon de validation soit par **validation croisée** sur l'échantillon d'apprentissage. C'est cette dernière que nous avons retenue par défaut ici compte-tenu de la taille de la base (environ 40 000 assurés), puisqu'un échantillon de validation nous amènerait à réduire davantage la taille des échantillons d'apprentissage et de test.

Des scores d'évaluation de la qualité de l'apprentissage et de celle de la prévision sont par la suite calculés, respectivement sur les échantillons d'apprentissage et de test. Pour atténuer la dépendance de ces scores aux échantillons, il est préférable de réitérer l'échantillonnage et la modélisation plusieurs fois, pour ainsi pouvoir retenir les moyennes de chaque score. Le temps de construction des modèles étant longs, nous avons décidé de ne réitérer la démarche que 100 fois.

En conclusion, la meilleure méthode est retenue en cherchant si possible un compromis entre qualité de prévision et interprétabilité des résultats obtenus. En effet, autant retenir, entre deux méthodes relativement équivalentes d'un point de vue de la performance, celle qui donne un modèle interprétable. Une fois le choix fait, on estime *in fine* le modèle sur tout l'échantillon (regroupant échantillons d'apprentissage et de test), sans changer les paramètres, et on applique le modèle retenu en production.

2 - Présentation du portefeuille étudié

2.1 - Présentation de l'ancien et du nouveau produit

Le produit « L », dont la commercialisation début 2018 a motivé ce mémoire, est venu supplanter le produit « A », l'ancien produit du périmètre des TNS (majoritairement) et des salariés. Cet ancien produit se décline chronologiquement en 4 générations qui témoignent des évolutions apportées en termes de tarifs et de clauses contractuelles de manière générale (garanties,...). Grâce à la très forte ressemblance entre le produit L et les deux dernières générations du produit A (A10 et A15, respectivement commercialisées en janvier 2010 et en janvier 2015)

3. d'après le référentiel des services informatiques ou l'incohérence de certaines variables pour un même sinistre, ainsi que la forte proportion de valeurs manquantes.

4. avec éventuellement une partie de validation.

en termes de garanties proposées, de segmentations tarifaires et au regard d'autres caractéristiques comme l'âge de fin de garantie en incapacité ou encore les modalités de souscription (questionnaire de santé,...), il est possible, sans trop prendre de risque, de faire l'hypothèse que la sinistralité future du nouveau produit qui a motivé cette étude sera très vraisemblablement la même que celle de A10 et A15 dont les données servent de base à ce mémoire. Il convient donc à présent de présenter les caractéristiques contractuelles communes aux produits d'intérêt qui vont nous éclairer sur la composition et les contours du portefeuille étudié.

Conditions d'adhésion. Les personnes physiques peuvent être assurées sous condition de :

- ne pas être âgé de plus de 65 ans à la date de l'adhésion,
- justifier d'un état de santé jugé satisfaisant par l'assureur, au moyen d'un questionnaire prévu sur le bulletin d'adhésion,
- pour les garanties applicables en cas d'arrêt de travail :
 - pour les **travailleurs salariés** : exercer une activité professionnelle en contrat à durée indéterminée, hors emploi saisonnier, à plein temps ou à temps partiel d'au moins 70 % (hors raisons médicales) rémunérée, régulière et continue.
 - pour les **TNS** : exercer une activité professionnelle hors emploi saisonnier, rémunérée, régulière et continue dont les revenus sont déclarés à l'administration fiscale. L'assureur pourra à cette occasion soumettre le futur assuré à un examen médical qu'il jugerait nécessaire.

Au vu des déclarations du candidat à l'assurance, l'assureur peut :

- soit accepter le risque soumis,
- soit l'accepter à conditions spéciales,
- soit le refuser.

Étendue territoriale. Les garanties s'exercent dans le monde entier sauf dans les pays faisant l'objet d'une exclusion contractuelle. Ne sont également pas pris en charge les sinistres survenus lors des séjours effectués par l'assuré dans les zones dites « Formellement déconseillées » ou « Déconseillées sauf raison impérative » à la date du départ selon la nomenclature du Ministère des Affaires Étrangères.

A noter que les indemnités sont toujours payées en euro, afin d'éviter un risque lié au taux de change.

Franchise. En cas d'incapacité temporaire totale de travail de l'assuré, l'assureur verse les indemnités journalières après expiration du délai de franchise. Les deux types de franchise qui sont susceptibles de s'appliquer sont la franchise absolue et la franchise relative, cette dernière n'étant pas accordée aux professions salariés.

Indemnités journalières par palier. L'assuré peut choisir un montant différent d'indemnité journalière par période de prestation se décomposant ainsi :

- Palier 1 : du 1^{er} au 90^{ème} jour d'arrêt de travail,
- Palier 2 : du 91^{ème} au 365^{ème} jour d'arrêt de travail,
- Palier 3 : du 366^{ème} au 1095^{ème} jour d'arrêt de travail.

L'assuré ne peut pas couper une période d'indemnisation : s'il prend le premier palier, il doit obligatoirement prendre le deuxième et le troisième palier. S'il prend le deuxième palier, il doit prendre le troisième et dans tous les cas il doit nécessairement prendre le troisième palier. Il ne peut par exemple pas prendre uniquement le premier et le troisième palier. La possibilité de mettre des montants différents par palier est liée au fait que les régimes professionnels offrent des protections différentes selon les secteurs d'activités.

Cette segmentation par palier permet donc de choisir des montants garantis (indemnités journalières) différents pour chacune des trois strates temporelles. Également, un assuré craignant surtout les répercussions financières d'un sinistre long peut par exemple choisir des montants garantis faibles pour le premier palier (afin d'alléger la fraction afférente de sa prime) et des indemnités journalières élevées pour les paliers 2 et 3.

Frais professionnels et paliers. L'assuré doit exercer à la date du sinistre une activité professionnelle non salariée rémunérée pour bénéficier de cette garantie. Cette garantie permet de bénéficier d'indemnités journalières supplémentaires afin de couvrir notamment les frais généraux de la société de l'assuré ou ceux liés à son activité professionnelle (salaire des employés, loyers, charges diverses...).

Deux durées de garantie sont proposées :

- Palier A : jusqu'au 365^{ème} jour d'incapacité de travail,
- Palier B : jusqu'au 1 095^{ème} jour d'incapacité de travail.

L'assuré choisit le délai de franchise absolue⁵, le montant de l'indemnité journalière ainsi que la durée de garantie.

Pour une nomenclature exhaustive des garanties proposées au titre de l'incapacité, se référer à l'[annexe page 98](#).

Cessation des garanties. Les garanties cessent à la date anniversaire du contrat qui suit le 70^{ème} anniversaire de l'assuré. Elles cessent également de plein droit à la retraite de l'assuré et au plus tard à son 70^{ème} anniversaire.

Effet et Durée. Les contrats sont conclus pour une durée indéterminée et sont résiliables chaque année, par lettre recommandée adressée au moins 2 mois avant la date anniversaire du contrat.

5. Il n'y a pas de franchise relative pour les garanties Frais professionnels.

Sauf radiation, l'adhésion se renouvelle annuellement par tacite reconduction au 1^{er} jour du mois anniversaire de la date d'effet de l'adhésion.

Calcul de la prime. À la souscription, une fois accepté le risque du candidat à l'assurance par l'assureur, ce dernier tient compte de certaines caractéristiques de l'assuré pour tarifier :

- la cible professionnelle parmi 7 : les médicaux, les paramédicaux, les experts/conseil, les libéraux, les artisans/commerçants, les agricoles et les salariés,
- la catégorie professionnelle qui indique le niveau de risque correspondant à son activité professionnelle :
 - La catégorie professionnelle 1 (CP1) : assuré dont l'activité ne comporte pas de travail manuel,
 - La catégorie professionnelle 2 (CP2) : assuré dont l'activité comporte un travail manuel réputé non dangereux,
 - La catégorie professionnelle 3 (CP3) : assuré dont l'activité comporte un travail manuel dangereux et/ou un risque spécifique,
- l'âge à la souscription,
- les montant garantis pour chaque garantie souscrite,
- la pratique de certaines activités sportives à risque ou certains antécédents médicaux indiqué dans le questionnaire de santé, qui peuvent entraîner une surprime.

Périodicité de la prime. Les cotisations sont payables d'avance selon la périodicité choisie par l'assuré sur le bulletin d'adhésion : elles peuvent être payées par année, par semestre, par trimestre ou par mois.

2.2 - Les données à disposition et leur pertinence

Le recours aux données externes pour tenter d'apporter plus d'informations à l'ensemble des variables explicatives nous amène à distinguer ces variables « externes » des variables « internes » issues des bases de données de la compagnie d'assurance et propres à chaque assuré. Nous allons dans la suite les passer en revue afin de discuter de leur pertinence et de leur apport potentiel à la modélisation de la sinistralité.

2.2.1 - Les variables explicatives « internes »

Parmi les variables des bases internes, se trouvent trois catégories de variables : celles relatives à la typologie de l'assuré, celles relatives au contrat (garanties, fiscalité,...) et celles relatives aux réseaux commerciaux au travers desquels l'affaire est arrivée en portefeuille. Nous allons dans la suite justifier l'utilité à priori de chacune des variables « internes » pour notre étude.

Région. Il s'agit des 21 régions de France métropolitaine hors Corse d'avant le redécoupage de 2016. L'information du département et l'information de la commune de résidence sont également

disponibles mais elles sont trop fines (parfois pas assez d'assurés par modalité) et engendrent beaucoup trop de modalités pour la modélisation ou même pour un regroupement de modalités qui conserverait une « homogénéité » géographique⁶. L'information sur la région de résidence est donc intéressante dans la mesure où elle peut révéler une différence géographique de sinistralité, probablement due à des facteurs socio-économiques (exemple de la macrocéphalie du système urbain français dû au poids de la métropole parisienne) susceptibles d'être captés par les données externes à la maille départementale présentées en infra, ou due à une concentration différente des métiers à risques (régions agricoles,...).

Convention. Il s'agit ici de la branche professionnelle associée au métier de l'assuré, qui ne correspond pas exactement à une convention collective nationale, mais plutôt à une sorte de secteur d'activité, qui va distinguer les assurés en 7 catégories : les travailleurs non salariés médicaux, paramédicaux, libéraux, experts-conseils, artisans-commerçants, agricoles, et les salariés.

Il existe des différences d'exposition au risque d'incapacité à la fois entre les 7 catégories, et entre les différents métiers d'une même catégorie (cf. paragraphe suivant relatif à la variable « Catégorie professionnelle »). Il peut par exemple sembler logique de prime abord que les médecins soient les moins exposés de par leur formation, leur expérience métier et leur appartenance même à la sphère médicale qui implique un niveau élevé de conscience, de vigilance et de connaissance des déterminants du risque incapacité. Aussi, les experts-conseils exercent majoritairement des métiers qui entraînent, toutes choses égales par ailleurs, une forte sédentarité et une faible exposition aux dangers physiques dans le cadre de leur travail, contrairement aux agriculteurs qui sont amenés à porter des charges lourdes ou manipuler des produits toxiques et qui utilisent des outils à l'occasion de travaux manuels. Ils sont donc plus exposés aux risques de blessures, tout comme certains artisans-commerçants.

Mais l'incapacité n'a pas qu'une dimension physique, elle a aussi une dimension psychique : par exemple, les salariés quel que soit leur secteur (industrie ou services) peuvent être exposés aux risques d'incapacité liée à des troubles psychosociaux autant qu'aux risques d'incapacité physique. En effet, le stress et de manière plus générale la souffrance au travail débouchent souvent, s'ils ne sont pas rectifiés à temps, sur des arrêts de travail. L'institut national de recherche et de sécurité (INRS), dont la vocation est de prévenir les maladies professionnelles et les accidents du travail, recense (cf. [INR](#)) 6 causes de risques psychosociaux (RPS) :

1. **l'intensité élevée du travail** : surcharge, cadences élevées, pression du temps et des délais, horaires imprévisibles, obligation de concentration et de performance, ...
2. **les exigences émotionnelles fortes** : l'exemple le plus probant est celui des professionnels de santé confrontés à la maladie et à la mort.
3. **l'absence d'autonomie** : travail à la chaîne répétitif et rébarbatif, faible participation aux décisions,...

6. L'« inhomogénéité » géographique viendrait par exemple du cas de deux départements, l'un au Nord et l'autre au Sud, ayant une sinistralité élevée et qui contrasterait avec celle des autres départements limitrophes.

4. **les relations et rapports sociaux difficiles** : harcèlement moral, irrespect, manque de reconnaissance, mise au placard, sentiment que l'on n'existe pas, agressivité et incivilité pour les salariés en contact avec le public, absence de soutien de sa hiérarchie ou de ses collègues, mauvaise ambiance,...
5. **les conflits de valeur** : faire un travail que l'on juge inutile, vendre un crédit à des personnes à très faibles revenus, faire la promotion d'une méthode que l'on sait inefficace, sentiment de bâcler et de mal faire son travail par manque de temps,...
6. **Des changements ou un contexte potentiellement négatifs** : incertitude de l'emploi, fusion, restructuration, licenciement,...

Si ces 6 facteurs de risques semblent concerner les salariés, les TNS ne sont pas en reste (cf. [idG](#)) puisqu'ils sont eux aussi sujets à des situations pathogènes et anxiogènes, liés à la charge de travail importante et leur responsabilité d'entrepreneur, mais également à la solitude face à la décision et la possibilité généralement limitée de déléguer. Les conditions des TNS conduirait chaque jour deux d'entre eux au suicide.

Parmi les possibles conséquences de tous ces facteurs du point de vue de l'incapacité, on trouve les dépressions, les maladies psychosomatiques, les problèmes de sommeil, les troubles musculo-squelettiques, les maladies cardio-vasculaires, et des accidents du travail. Le secteur d'activité et le type de métier, regroupés en conventions, sont donc des déterminants du risque incapacité dans la mesure où l'exposition au risque est différente.

Catégorie professionnelle Cette variable peut prendre 3 modalités allant de 1 à 3, qui donnent de façon croissante l'intensité de l'exposition au risque d'incapacité selon le métier donné à une maille plus fine que la convention (infirmières, manutentionnaires, avocats,...). Elle est basée, comme vu plus haut, à l'absence (CP1) ou la présence (CP2 et CP3) de travail manuel à risque. La catégorie professionnelle est donc un déterminant du risque étudié, incontournable dans le cadre de notre étude.

Sexe. Il peut évidemment exister des différences d'expositions au risque incapacité selon que l'on est une femme ou un homme, notamment à cause du fait que certains métiers à risque sont majoritairement exercés par des personnes d'un même sexe (infirmières, manutentionnaires,...). Par ailleurs, les longs arrêts de travail pour cause de maternité ne concernent que les femmes. Le sexe est donc un facteur de risque incapacité.

Situation de famille. Cette variable indique si les assurés de la base sont veufs, divorcés, célibataires ou mariés. La solitude du célibat peut être subie et causer une vulnérabilité psychologique dont la portée n'est pas négligeable, surtout dans une France où près d'un habitant sur 10 est touché par la solitude objective (cf. [dF](#)) : la solitude objective est un constat et concerne une personne qui a peu de contacts sociaux, par opposition à la solitude subjective qui décrit

un sentiment ⁷). La sphère de vie privée peut donc empiéter sur la sphère de vie professionnelle. De même, les engagements et les contraintes qu'imposent une famille nombreuse sur les parents peuvent engendrer un déséquilibre entre les sphères de vie privée et professionnelle, et entraîner des problèmes psychiques ce qui constitue un facteur de risque incapacité.

Âges et ancienneté. L'âge à la souscription, l'âge moyen à la survenance de sinistres et l'ancienneté en portefeuille (vue à fin 2016 pour les assurés encore en portefeuille) sont des variables temporelles déterminantes du risque incapacité puisqu'il n'est plus à démontrer que le risque s'intensifie avec l'âge. Cependant, il reste à déterminer le seuil de basculement du risque, que l'on déterminera par la suite. La tarification du produit « A » sur lequel se base notre étude se fait à l'âge à la souscription et non à l'âge atteint ⁸, c'est pourquoi la modélisation de la [partie III](#), dont la variable cible distingue les assurés sinistrés des assurés non sinistrés, se fera notamment à partir de l'âge à la souscription, mais également sur l'ancienneté en portefeuille afin de prendre en compte l'aggravation du risque lié au temps écoulé depuis la souscription. En revanche, lors de la modélisation binaire de la rentabilité dans la [partie IV](#) portant sur la population sinistrée, seul l'âge moyen du sinistre incapacité sera retenu.

Abattement tarifaire Il s'agit d'une mesure commerciale spécialement dédiée aux créateurs d'entreprises, puisqu'ils bénéficient d'une réduction de prime de 50 % pour la première année d'assurance si l'adhésion intervient au cours de la première année d'activité de l'entreprise, et d'une réduction de prime de 25 % la première année d'assurance si l'adhésion intervient au cours de la deuxième année qui suit la création de l'entreprise.

Indicateur Surprimes. La surprime concerne les assurés ayant eu des antécédents médicaux, consignés dans le questionnaire de santé lors de l'adhésion, révélant un état de santé fragile (hypertension, stent, surcharge pondérale,...), un terrain héréditaire défavorable (myopie forte, asthme,...), ou des pratiques à risque (résultats positifs passés concernant des dépistages sérologiques portant sur les virus des hépatites B et C ou sur celui de l'immunodéficience humaine). Les surprimés peuvent également être des assurés qui conduisent des deux-roues ou qui pratiquent un sport considéré à risque tels que les sports d'altitude, le parapente, le bobsleigh, le rugby, la plongée sous-marine ou même les safaris.

Fiscalité. La loi « Madelin » a pour vocation de permettre aux TNS non agricoles de bénéficier d'avantages fiscaux liés aux garanties de prévoyance et constitués par des cotisations déductibles du bénéfice imposable, à titre compensatoire afin de pallier les lacunes des régimes obligatoires (RSI,...) en termes de niveaux de couverture. Cependant, ce dispositif n'est pas obligatoire, et l'assuré peut préférer un cadre fiscal semblable à celui de l'assurance-vie s'il est plus avantageux. En effet, le régime Madelin ne dispense pas les assurés qui bénéficient de ce cadre

7. Une personne peut être entourée et se sentir seule, ou au contraire fréquenter peu de monde et ne pas en souffrir.

8. Le contrat de prévoyance sur lequel nous basons notre étude ayant une échéance annuelle avec tacite reconduction, l'âge atteint désigne l'âge lors de la reconduction du contrat.

fiscal de déclarer aux impôts les prestations qui leurs sont payées en cas de sinistre, contrairement aux assurés n'ayant pas fait le choix du régime Madelin, qui, eux, en sont dispensés. On peut donc penser que le choix du régime fiscal pour un TNS est important pour expliquer la sinistralité, dans la mesure où la fiscalité des revenus de remplacements perçus en cas de sinistre par les non-bénéficiaires du régime Madelin est avantageuse. Ainsi, le numérateur du S/P peut donc varier, toutes choses égales par ailleurs, entre un TNS Madelin et un TNS non Madelin, du fait de la fréquence des sinistres ou de la durée d'indemnisation notamment. Par conséquent, la tarification devrait donc théoriquement intégrer une majoration tarifaire en contrepartie d'une éventuelle surconsommation de prestations de la part des non-Madelin, ce qui n'est pas le cas actuellement.

Par ailleurs, les salariés et les travailleurs du secteur de l'agriculture n'étant pas concernés, ils ne peuvent respectivement bénéficier que du régime fiscal de l'assurance-vie et celui des travailleurs agricoles.

Périodicité de la prime. Cette variable peut prendre 4 modalités différentes : mensuel, trimestriel, semestriel et annuel. Le raisonnement sous-jacent qui justifie l'intérêt de cette variable dans notre modélisation de la sinistralité consiste à dire que, théoriquement, plus un assuré est en situation de précarité financière, avec une trésorerie à flux tendus, plus il optera pour une fréquence élevée pour les versements de prime afin de répartir dans le temps l'effort financier, et, à l'inverse, plus l'assuré se trouve dans un contexte d'aisance financière, plus il pourra verser sa prime d'un seul tenant. Bien que le lien entre périodicité de prime et sinistralité ne soit pas direct, cette variable peut cependant être révélatrice d'un contexte financier (voire, par extrapolation, d'un contexte général) défavorable à l'assuré, propice à la survenance de sinistre incapacité, notamment sous sa forme psychologique.

Montant moyen de prime annuelle TTC. Le montant moyen de prime annuelle TTC est le montant de prime moyen payé annuellement par chaque assuré (somme des primes/ancienneté en portefeuille), toutes garanties confondues (décès, incapacité, invalidité). Ce montant est censé refléter l'appétence des assurés à se protéger par une assurance prévoyance, l'« appétit » à l'assurance. Ainsi, une consommation de prime importante peut être révélatrice d'un risque accru de survenance de sinistre non capté par d'autres variables, et pourtant bien réel mais dissimulé à l'assureur à cause de l'asymétrie de l'information entre ce dernier et l'assuré.

Une autre variable aurait pu être utilisée en lieu et place du montant moyen de prime annuelle TTC : c'est le niveau de garantie. Cependant, à cause des [différentes combinaisons de franchises et de paliers pour l'incapacité \(Annexe page 98\)](#) d'une part, et la dissemblance des 3 garanties (décès ponctuel, incapacité durant au plus 3 ans, invalidité de durée indéterminée), de leur modalités et leur fréquences de versement (capitaux décès, indemnités journalières, rentes invalidité) non comparables d'autre part, il n'était pas évident de calculer un montant garanti au global des 3 garanties.

Proportions de prime allouées au décès, à l'incapacité et à l'invalidité. La part de prime allouée à chacune des trois garanties principales que sont le décès, l'incapacité et l'invalidité, peut indiquer ce que l'assuré redoute le plus, ce qui peut potentiellement coïncider avec ce qui est le plus susceptible d'arriver. Ainsi, il n'est pas déraisonnable de penser que plus l'assuré va allouer de prime à l'incapacité, plus sa consommation de prestations pour cette garantie est potentiellement grande, si tant est qu'il y ait une certaine asymétrie de l'information entre lui et l'assureur, relativement à son exposition réelle aux risques et à son comportement. De plus, dans la mesure où l'invalidité prolonge l'incapacité si celle-ci dure plus de 3 ans, il n'est pas étonnant qu'une grande proportion de prime allouée à l'incapacité aille de paire avec une part importante également allouée à l'invalidité, au détriment du décès. Par conséquent, la part de prime allouée à chacune des garanties, et notamment à l'incapacité, peut constituer un trio de variables intéressant pour discriminer les sinistrés des non sinistrés en incapacité d'une part, mais aussi les sinistrés rentables des sinistrés non rentables d'autre part.

Indicateur frais d'adhésion. [L'association GPMA \(Groupement de Prévoyance Maladie Accident\)](#) qui compte près de 300 000 adhérents, a notamment pour mission de négocier et souscrire des contrats Prévoyance-Santé auprès des sociétés du groupe Generali, en plus d'un accompagnement solidaire en cas de sinistre. Les frais d'adhésion à cette association sont fixes et s'élèvent à 7 €. Ainsi, les frais d'adhésion, qui viennent charger davantage la prime, indiquent l'adhésion (modalité 7 €) ou non (0 €) à l'association GPMA. Le cadre attrayant et favorable aux sinistrés d'accompagnement solidaire et de soutien au-delà des contrats d'assurance que propose l'association et son fonds d'entraide peut éventuellement révéler une plus grande sinistralité de ses adhérents par rapport au reste du portefeuille, du fait d'une plus grande fréquence de survenance de sinistres ou de durée d'indemnisation.

Apporteurs d'affaires : Agents, Courtiers et Réseaux divers. Si les agents généraux et les courtiers sont deux intermédiaires en assurance, le premier représente la société d'assurance pour le compte de laquelle il vend des contrats d'assurance et dont il engage la responsabilité, tandis que le deuxième agit pour le compte de ses clients, assurables ou assurés, dont il a pour mission de les mettre en relation avec un assureur. Cette distinction entre les parties d'intérêts contraires représentées par ces deux types d'intermédiaires pourrait impliquer une meilleure sélection des « bons » assurés (non sinistrés ou sinistrés rentables) de la part du mandataire de l'assureur qu'est l'agent général, et, à l'inverse, une moins bonne sélection des assurés par le mandataire de l'assuré qu'est le courtier, ce dernier ayant pour vocation de trouver pour ses clients le contrat qui s'adapte le plus à leur profil de risque spécifique, et cela au meilleur tarif.

En effet, le courtier, qui peut être spécialiste d'un type d'assuré, est contraint par une obligation légale de conseil envers ses clients, qui ne connaissent pas forcément bien le domaine de l'assurance (garanties, exclusions, tarifs,...), et qui peuvent pour certains rencontrer des difficultés à s'assurer de par leur profil risqué (âge élevé, profession à risque,...).

Quant à l'agent général, bien qu'il soit mandaté par un assureur, il exerce avant tout une profession libérale rémunérée par des commissions dont le volume est proportionnel au chiffre d'affaires généré, ce qui l'amène vraisemblablement à être guidé par un objectif de maximisation du chiffre d'affaires au risque de desservir les intérêts de l'assureur qu'il représente. Aussi, un agent général ne bénéficie pas toujours de la même capacité d'analyse du marché des risques que les cabinets de courtage, qui peuvent par ailleurs avoir un grand poids économique leur permettant de négocier des prix avantageux (agissant à la hausse sur des ratios S/P par exemple).

Toutes ces différences d'enjeux et d'objectifs, inter et intra apporteurs, qui sous-tendent les professions d'agent général et de courtier font qu'ils peuvent venir faire grossir le portefeuille de l'assureur avec des assurés dont la sinistralité (et en particulier la rentabilité) est très variée. La variable relative à l'apporteur peut donc potentiellement être un facteur déterminant dans notre étude. Quant aux réseaux divers, ils sont constitués de réseaux de distribution diversifiés et complémentaires, comme le réseau [réseau salarié](#) composé de conseillers et chargés de clientèle, ou encore le réseau d'[établissements financiers importants](#) (banques privées et banques de détail, institutions de prévoyance, plateformes de courtage), liés à Generali par des accords de partenariats et qui commercialisent des contrats de l'assureur italien sous leur propre marque.

2.2.2 - Les variables explicatives « externes »

Une discussion sur la pertinence des variables externes ne peut se faire sans une réflexion préalable sur le niveau de finesse des données externes qui vont être raccordées par une jointure aux données internes. Ainsi, se pose la question de savoir comment rattacher ces données aux portefeuille étudié. Autrement dit, quelle variable des bases de données internes de l'entreprise, également commune à la source de données externes peut être utilisée pour permettre la fusion des deux ? Cette variable doit présenter un niveau de granularité suffisant en termes de modalités pour permettre d'étudier l'impact des données externes de la manière la plus fine, et donc la plus fidèle possible à la réalité.

Une grande finesse géographique comme le rend possible la maille communale⁹, bien que permettant une grande précision spatiale, est lourde à mettre en place, à cause du grand nombre de communes, et les données communales n'étant souvent pas constituées à partir de recensements exhaustif, l'INSEE recommande de manier avec précaution les effectifs inférieurs à 200 (d'une variable donnée) ou des zones de moins de 2 000 habitants à cause de l'imprécision des sondages. Les comparaisons entre territoires de petites tailles sont donc à proscrire.

A l'inverse, une maille géographique moins fine comme la maille régionale, entraîne une trop grande réduction de l'information par moyenne par rapport à la maille communale par exemple, et donc une moins bonne représentativité des disparités départementales. Il est donc envisageable de recourir à des données externes déclinées à la maille départementale (96 départements au total) issues d'études très intéressantes, disponibles sur des sites institutionnels

9. puisque l'on dispose du code postal pour chaque assuré de la base.

comme celui de l'[INSEE](#), de la [DREES](#) ou encore de la mission Etalab placée sous l'autorité du Premier ministre et concrétisée par une [plateforme ouverte des données publiques françaises](#). Il est toutefois important de noter que lors de l'inclusion dans la modélisation des données externes retenues, il est important de retirer de l'étude la variable spatiale explicative (à la maille régionale) afin de ne pas créer de redondance ou d'interférences entre l'information (i.e. la dispersion) sur la sinistralité apportée à la modélisation par cette variable spatiale et l'information véhiculée par les variables externes rattachées grâce à une variable spatiale (à la maille départementale).

Afin de faire au mieux pour ne pas manquer de capter tout facteur de risque, ou, à l'inverse tout indicateur de bien-être, liés à un effet géographique ou démographique, à un contexte économique, culturel, médical ou alimentaire pouvant influencer sur la sinistralité, nous allons tirer profit de la diversité des variables externes que nous présentons dans le paragraphe qui suit. Mais la limite majeure d'une telle approche réside dans l'extrapolation des caractéristiques intrinsèques des assurés de Generali à partir de données qui n'ont pas, à l'origine, cette vocation puisque les « individus » statistiques que l'on étudie ne sont pas des zones géographiques (variable la plus commune aux données externes, et donc variable à minima pour une jointure) mais des personnes physiques, qui plus est propres à un portefeuille d'assurance. Ainsi, pour s'assurer qu'au moins le portefeuille étudié est bien représentatif de la population générale française en termes d'effectifs par zone géographique, nous vérifierons en section 2 page 39 qu'il existe bien entre eux un lien suffisamment fort.

Géographie/Démographie : altitude moyenne ; proportion d'actifs ; proportion d'actifs à temps plein ; proportion moyenne de naissances entre 2009 et 2014 ; proportion moyenne de décès entre 2009 et 2014 ; nombre moyen de personnes par ménage ; espérance de vie à la naissance ; espérance de vie à 60 ans ; moyenne des divorces entre 2010 et 2014 ; indice de vieillissement.

Économie : Le salaire annuel moyen ; la moyenne des niveaux de vie ; le taux de pauvreté ; les dépenses régionales brutes pour l'aide sociale en 2013 ; le taux de fraude ; la part {des revenus d'activité ; de la pension de retraite ; des revenus du patrimoine ; des prestations sociales ; des impôts} dans les revenus globaux ; la proportion de personnes travaillant dans {l'administration publique ; l'agriculture ; le bâtiment/travaux publics ; le commerce inter-entreprises ; la conception/recherche ; la culture/loisirs ; la distribution ; l'éducation/formation ; l'entretien/réparation ; la fabrication ; la gestion ; les transports/logistique ; les prestations intellectuelles ; la santé/l'action sociale ; les services de proximité} ; la proportion en 2013 d'établissements {de l'industrie ; de la construction ; des transports et services ; de l'administration publique}.

Climat : ensoleillement horaire total en 2013.

Culture/Divertissement : nombre de chambres d'hôtels par superficie en 2017 ; nombre de campings par superficie en 2017 ; nombre d'aires de jeu par superficie ; nombre de salles de

théâtre et de cinéma par superficie.

Santé/Hygiène de vie : temps moyen pour se rendre chez son généraliste à partir de son domicile ; nombre de médecins {généralistes ; spécialistes} en pourcentage de la population de la région ; proportion de la population marchant à pieds du domicile au lieu de travail ; proportion de décès par {maladies infectieuses ; maladies de l'appareil circulatoire ; tumeurs ; maladies de l'appareil respiratoire ; suicides ; diabète ; accidents de transports} ; proportion de personnes qui {fréquentent les fast foods avec une fréquence supérieure à 1 fois par mois ; ont un intérêt faible pour la qualité de l'alimentation ; fument quotidiennement ; consomment des protéines plus d'une fois par jour ; consomment des fruits et légumes plus d'une fois par jour ; n'ont pas pris de vacances les 12 derniers mois ; ont un statut nutritionnel élevé ; passent quotidiennement un temps élevé devant l'écran} ; proportion de blessés dans des accidents de la route ; consommation quotidienne moyenne d'alcool (en gramme) ; pourcentage de personnes ayant une consommation d'alcool supérieure à la médiane nationale ; proportion de personnes ayant une consommation quotidienne d'alcool ; pourcentage d'usagers d'alcools connaissant des ivresses répétées ; pourcentage d'usagers objet d'une API (alcoolisation ponctuelle importante) au minimum une fois par mois ; pourcentage de personnes consommant du vin au moins une fois par semaine ; pourcentage de personnes consommant de la bière au moins une fois par semaine, pourcentage de personnes consommant un alcool fort au moins une fois par semaine.

Ici, nous mettons l'accent sur l'originalité des données culturelles ou météorologiques. Le raisonnement sous-jacent à l'introduction de ces deux aspects, qui n'ont à priori pas de lien direct avec les caractéristiques médico-sociales supposées logiquement constituer le noyau dur de toute problématique en prévoyance, consiste à tenter de capter un effet psychologique géographique. Par exemple, une zone présentant une densité élevée des centres de divertissement et d'épanouissement intellectuels pourrait également être une zone où les décès, ainsi que les incapacités causées par des maladies, sont peu probables. Un effet psychologique néfaste peut également être provoqué à long terme par un climat maussade ou la solitude (personnes vivant seules).

Par ailleurs, des indicateurs économiques et financiers tels que le revenu médian des ménages (plus robuste aux valeurs extrêmes que le revenu moyen) ou des variables démographiques peuvent être révélateurs également d'un contexte de prospérité économique et d'aisance financière propice à une plus grande longévité ou un meilleur état de santé.

2.2.3 - Préparation des données brutes

La page suivante a pour but de présenter de manière schématique les différentes étapes constituant le passage des données brutes aux données finales.

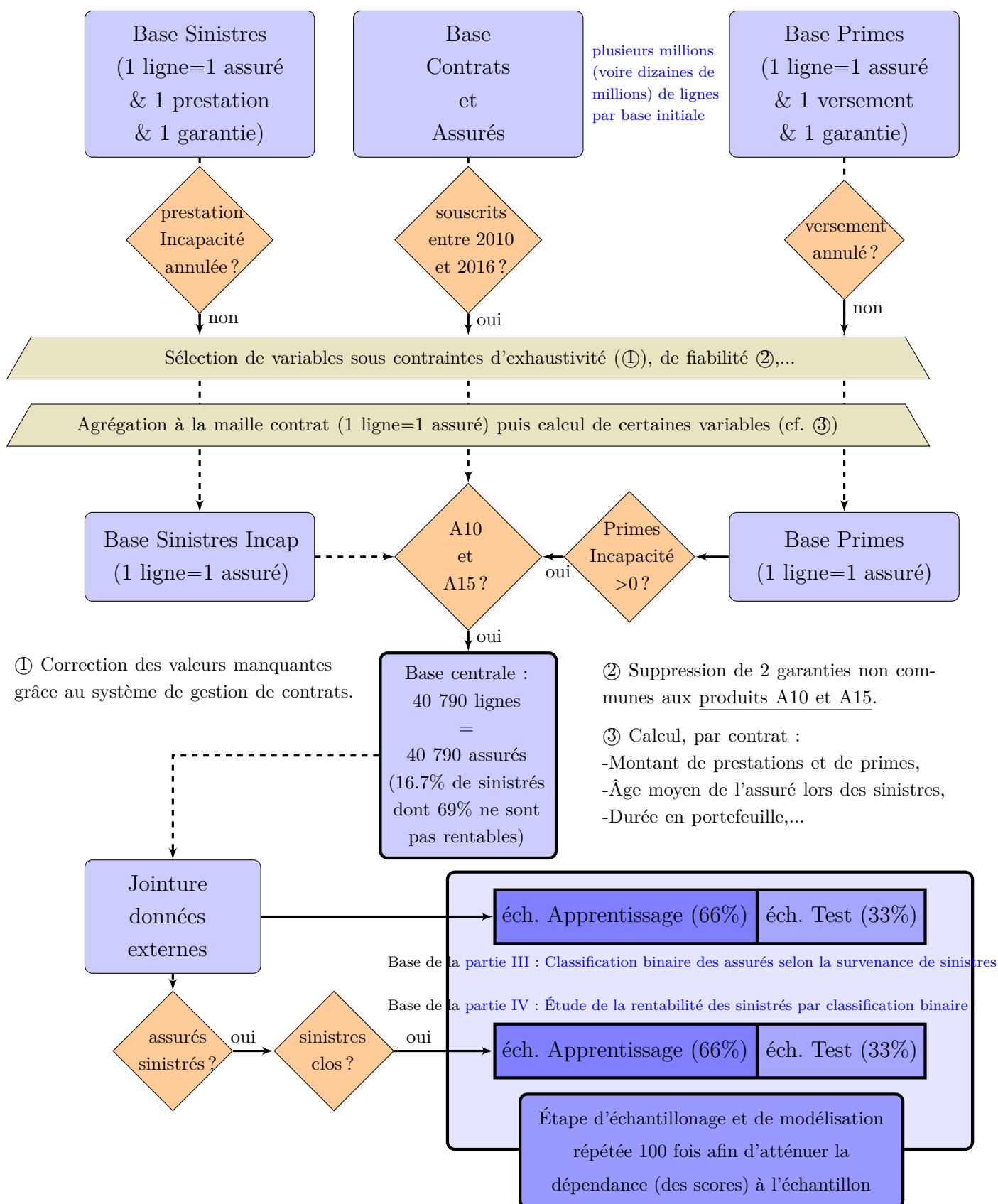


FIGURE I.1 – Schéma explicatif du passage des données brutes aux données finales.

3 - Généralités et formalisme du contexte d'apprentissage

Le principe de l'apprentissage consiste à inférer des règles générales à partir d'exemples. Plus concrètement, supposons l'existence de n observations i.i.d $D_n = (Y_i, X_{1,i}, \dots, X_{m,i})_{i=1}^n \in (\mathcal{Y} \times \mathcal{X})^n$, d'une part d'une variable aléatoire unidimensionnelle $Y \in \mathcal{Y}$, et d'autre part d'un ensemble de m variables aléatoires unidimensionnelles $\mathbb{X} = (X_1, \dots, X_m) \in \mathcal{X}$, dont la loi jointe $L(Y, X_1, \dots, X_m)$ n'est pas connue. Les algorithmes d'apprentissage vont chercher à construire un prédicteur $\hat{\phi} \in \mathcal{Y}^{\mathcal{X}} := \{f|f : \mathcal{X} \rightarrow \mathcal{Y}\}$, censé estimer une fonction ϕ liant la variable réponse Y et les variables explicatives \mathbb{X} , à supposer qu'il y en ait une (i.e. $\hat{\phi}(\mathbb{X}) \approx \phi(\mathbb{X}) = Y$), à partir des n observations de chaque variable ¹⁰.

Ce cas précis d'apprentissage où l'échantillon de données contient de l'information à la fois sur \mathbb{X} et sur Y est appelé **apprentissage supervisé**. Son utilité, une fois la fonction $\hat{\phi}$ déterminée, est de pouvoir prédire la valeur Y d'une nouvelle valeur observée de \mathbb{X} .

Si à présent, l'échantillon de données n'est composé que de n observations de seulement m variables $(X_{1,i}, \dots, X_{m,i})_{i=1}^n$ (pas de variable Y à prédire), il peut néanmoins être utile de faire émerger des schémas qui sont structurants (i.e. des classes d'assurés homogènes au regard des valeurs prises par les variables (X_1, \dots, X_m)) : c'est l'**apprentissage non supervisé**. Le résultat de ce type d'approche, c'est à dire l'étiquetage de chaque assuré $i \in \{1, \dots, n\}$ avec un label désignant sa classe d'appartenance peut être vu comme la création ad-hoc d'une variable catégorielle Y (contrairement à l'apprentissage supervisé où cette variable est préexistante et n'est pas forcément catégorielle), et dont chaque valeur est liée à la combinaison de valeurs prises conjointement par les variables (X_1, \dots, X_m) . Tout l'enjeu de l'apprentissage non supervisé est de donner un sens à chaque classe, eu égard aux valeurs prises par les variables pour les assurés de chaque classe.

Ainsi, l'objectif de ces 2 méthodes d'apprentissage n'est à priori pas le même (prédiction ou segmentation), mais les résultats de ces 2 méthodes nous donnent des informations précieuses sur l'influence des variables et leur inter-connexion. Le coeur statistique de ce mémoire est l'apprentissage supervisé, mais nous utiliserons une méthode (k -means) d'apprentissage non supervisé sur le premier plan de l'AFCM (pages 38 et 79).

Une autre distinction qu'il peut être utile de relever concerne les 2 approches de l'apprentissage supervisé : l'apprentissage supervisé **statistique** et l'apprentissage supervisé **automatique**. Si la première nécessite d'émettre puis de vérifier des hypothèses (loi de probabilité sur tout ou partie des variables, modèle explicite donnant la structure de la liaison supposée entre variables d'entrée et variable de sortie, erreur entre modèle et données...), et s'attache d'avan-

10. Le prédicteur $\hat{\phi}$, dont l'implémentation constitue un algorithme d'apprentissage, est donc une fonction de l'espace $\mathcal{Y}^{\mathcal{X}}$ construite à partir de $(\mathcal{Y} \times \mathcal{X})^n$.

tage au problème de l'inférence des paramètres comme dans le cadre d'un GLM, la deuxième est, quant à elle, tournée entièrement vers la performance des procédures et l'optimisation computationnelle des résultats qui en découlent, et ne nécessite pas forcément de supposer une structure sur les données, ni de recourir à un modèle mathématique, formel et rigoureux. Il s'agit donc de deux faces d'une même pièce dont l'objectif est commun (prédire), mais dont les fondements (statistique et informatique) diffèrent.

A présent intéressons nous à l'évaluation de la performance de notre algorithme d'apprentissage supervisé. Il convient donc d'introduire la fonction de perte, première pierre de l'édifice.

La fonction de perte L'efficacité des procédures d'apprentissage supervisé se mesure en comparant la valeur prédite $\hat{\phi}(\mathbb{X})$ et la valeur à prédire Y . Ce rapprochement peut se faire au moyen de différentes fonctions de perte $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$, qui seront d'autant plus élevées que la prévision est loin de la réalité, et d'autant plus proche de 0 sinon.

Par exemple, on peut citer quelques fonctions de perte :

- perte quadratique (régression) : $\ell[\hat{\phi}(\mathbb{X}), Y] = [\hat{\phi}(\mathbb{X}) - Y]^2$
- perte absolue (régression) : $\ell[\hat{\phi}(\mathbb{X}), Y] = |\hat{\phi}(\mathbb{X}) - Y|$
- perte binaire (classification) : $\ell[\hat{\phi}(\mathbb{X}), Y] = \mathbf{1}_{\hat{\phi}(\mathbb{X}) \neq Y}$

Aussi, on remarque que dans le cadre de la classification binaire ($\mathcal{Y} = \{0, 1\}$), il n'y a pas de différence entre les trois fonctions de perte. Par ailleurs, cette quantité $\ell[\hat{\phi}(\mathbb{X}), Y]$ étant aléatoire, il est nécessaire d'introduire dans le paragraphe suivant la notion de risque d'un prédicteur, qui est homogène à une espérance mathématique.

3.1 - Performance théorique

Le risque théorique. Le risque d'un prédicteur $\hat{\phi}$ est défini par la relation $R(\hat{\phi}) = \mathbb{E} \left[\ell[\hat{\phi}(\mathbb{X}), Y] \right]$.

Il vérifie 2 propriétés :

- $\forall \hat{\phi}, R(\hat{\phi}) \geq 0$
- si $\hat{\phi}(\mathbb{X}) \stackrel{p.s.}{=} Y$ alors $R(\hat{\phi}) = 0$

Ainsi, la recherche théorique du prédicteur le plus performant revient à rechercher celui dont le risque est le plus faible : c'est le prédicteur de Bayes.

Prédicteur de Bayes. L'estimateur bayésien ϕ^* est défini comme étant celui qui minimise le risque :

$$\phi^* \in \operatorname{argmin}_{\phi \in \mathcal{Y}^{\mathcal{X}}} R(\phi)$$

Par conséquent, si $\forall x \in \mathcal{X}$, l'infimum de $t \mapsto \mathbb{E} \left[\ell[t, Y] \mid \mathbb{X} = x \right]$ est atteint, alors $\phi^*(x) \in$

$\operatorname{argmin}_{t \in \mathcal{Y}} \mathbb{E} [\ell[t, Y] | \mathbb{X} = x]$ est un prédicteur de Bayes, puisque si l'on prend un quelconque $\phi \in \mathcal{Y}^{\mathcal{X}}$, on a que :

$$\begin{aligned} R(\phi) &= \mathbb{E} [\ell[\phi(\mathbb{X}), Y]] \\ &= \mathbb{E} \left[\mathbb{E} [\ell[\phi(\mathbb{X}), Y] | \mathbb{X} = x] \right] \\ &= \mathbb{E} \left[\mathbb{E} [\ell[\phi(x), Y] | \mathbb{X} = x] \right] \\ &\geq \mathbb{E} \left[\mathbb{E} [\ell[\phi^*(x), Y] | \mathbb{X} = x] \right] \\ &= \mathbb{E} [\ell[\phi^*(\mathbb{X}), Y]] := R(\phi^*) \end{aligned}$$

L'inconvénient est que le risque de Bayes $R^* := R(\phi^*)$ est inatteignable car pour ce faire, il faut d'une part connaître la loi de (\mathbb{X}, Y) , et d'autre part pouvoir minimiser sur l'espace fonctionnel $\mathcal{Y}^{\mathcal{X}}$. Il est toutefois intéressant de s'attarder sur l'expression de l'estimateur de Bayes lorsque $\mathcal{Y} = \{0, 1\}$ qui définit la classification (variable réponse qualitative) binaire.

Ainsi, si $\mathcal{Y} = \{0, 1\}$ et que l'on prend $\ell[\hat{\phi}(\mathbb{X}), Y] = \mathbb{1}_{\hat{\phi}(\mathbb{X}) \neq Y}$, alors :

$$\begin{aligned} \phi^* &\in \operatorname{argmin}_{t \in \{0,1\}} \mathbb{E} [\ell[Y, t] | \mathbb{X} = x] \\ &= \operatorname{argmin}_{t \in \{0,1\}} \mathbb{E} [\ell[Y, t] | \mathbb{X} = x] \\ &= \operatorname{argmin}_{t \in \{0,1\}} \mathbb{E} [\mathbb{1}_{Y \neq t} | \mathbb{X} = x] \\ &= \operatorname{argmin}_{t \in \{0,1\}} \mathbb{P} [Y \neq t | \mathbb{X} = x] \\ &= \operatorname{argmax}_{t \in \{0,1\}} \mathbb{P} [Y = t | \mathbb{X} = x] \\ &= \operatorname{argmax}_{t \in \{0,1\}} \left\{ \eta^*(x) \times \mathbb{1}_{t=1} + (1 - \eta^*(x)) \times \mathbb{1}_{t=0} \right\} \\ &\quad \text{avec } \eta^*(x) = \mathbb{P} [Y = 1 | \mathbb{X} = x] \\ &\implies \boxed{\phi^* = \mathbb{1}_{\eta^*(x) > \frac{1}{2}}} \end{aligned}$$

Par conséquent, lorsque nous étudierons l'occurrence ou non d'au moins un sinistre d'une part, et la rentabilité ou non d'autre part, la variable réponse étant binaire ($\mathcal{Y} = \{0, 1\}$), nous procéderons comme suit (approche « plug-in ») :

1. estimer la fonction de régression $\eta^*(x)$ par $\hat{\eta}(x)$
2. contruire $\hat{\phi}$, classifieur tel que $\hat{\phi}(x) = \mathbb{1}_{\hat{\eta}(x) > \frac{1}{2}}$

L'Oracle. L'impossibilité de minimiser le risque sur $\mathcal{Y}^{\mathcal{X}}$ peut nous amener à la recherche d'une alternative au prédicteur de Bayes, qui minimiserait le risque sur un ensemble fini $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ de fonctions de forme connue (par exemple les fonctions linéaires) : c'est la notion d'Oracle $\phi_{\mathcal{F}}^* := \operatorname{argmin}_{\phi \in \mathcal{F}} R(\phi)$. Autrement dit, si ϕ^* appartient à \mathcal{F} alors ϕ^* et $\phi_{\mathcal{F}}^*$ sont confondus, ce dernier étant le meilleur prédicteur local (au sens du risque).

Excès de risque et Compromis biais-variance. L'excès de risque $\varepsilon(\hat{\phi}) := R(\hat{\phi}) - R(\phi^*)$ peut être vu comme l'erreur totale commise par un prédicteur, et peut se décomposer en une erreur d'estimation et une erreur d'approximation :

$$\forall \hat{\phi} \in \mathcal{F}, \quad \varepsilon(\hat{\phi}) = \underbrace{R(\hat{\phi}) - R(\phi_{\mathcal{F}}^*)}_{\text{Variance}} + \underbrace{R(\phi_{\mathcal{F}}^*) - R(\phi^*)}_{\text{Biais}}$$

=erreur estimation =erreur approximation

L'erreur d'approximation, qui est la même pour tout prédicteur dans \mathcal{F} , reflète la performance de l'oracle par rapport au prédicteur de Bayes et soulève l'importance de bien choisir l'espace \mathcal{F} . L'erreur d'estimation, quant à elle, est propre au prédicteur étudié (pour un espace \mathcal{F} fixé), et représente sa qualité comparativement à l'oracle.

Cette décomposition du risque illustre le phénomène du « compromis biais-variance » de l'excès de risque¹¹ :

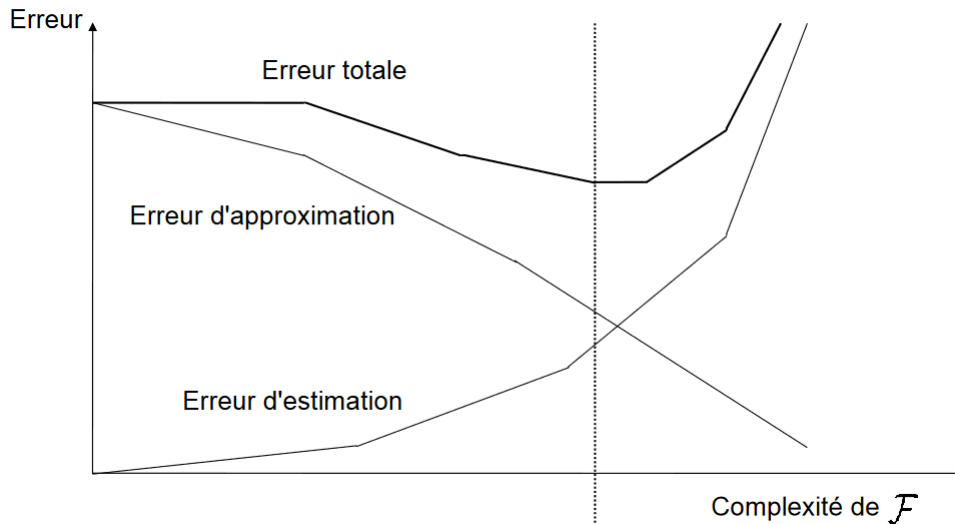


FIGURE I.2 – Illustration de la décomposition de l'excès de risque (erreur totale) en biais (erreur d'approximation) et en variance (erreur d'estimation).

Ainsi, plus la complexité (i.e. la variété des fonctions) de \mathcal{F} se rapproche de celle de $\mathcal{Y}^{\mathcal{X}}$, plus le biais diminue et la variance augmente. Il s'agit donc de choisir un espace de complexité intermédiaire comme dans le cadre du modèle linéaire (et plus généralement du GLM).

3.2 - Performance empirique

Le Risque Empirique. Par ailleurs, en l'absence d'information sur la loi des variables, il n'est pas possible de calculer le risque d'un prédicteur car celui-ci est défini comme une espérance. Ainsi, pour pallier cet obstacle, c'est le risque empirique R_n (moyenne empirique sur les points

11. La décomposition biais-variance de l'excès de risque d'un prédicteur est à distinguer de la décomposition biais-variance usuelle du risque de ce même prédicteur (cf. page 26).

de l'échantillon) qui est calculé en pratique en raison de sa convergence vers le risque, en vertu de la loi des grands nombres :

$$R_n(\hat{\phi}) := \frac{1}{n} \sum_{i=1}^n \ell[\hat{\phi}(\mathbb{X}_i), Y_i] \stackrel{LGN}{\underset{n \rightarrow \infty}{\approx}} \mathbb{E} \left[\ell[\hat{\phi}(\mathbb{X}), Y] \right]$$

Minimiseur du risque empirique. Cette nouvelle définition en amène une autre, celle du minimiseur du risque empirique (MRE), qui n'est autre que le prédicteur appartenant au sous-ensemble \mathcal{F} de $\mathcal{Y}^{\mathcal{X}}$ qui minimise le risque empirique :

$$\hat{\phi}^{MRE} \in \underset{\phi \in \mathcal{F}}{\operatorname{argmin}} R_n(\phi)$$

Ce prédicteur est l'ancêtre commun des algorithmes d'apprentissage supervisé. Le problème de minimisation qui le définit se fait sur \mathcal{F} et non sur $\mathcal{Y}^{\mathcal{X}}$, car sur $\mathcal{Y}^{\mathcal{X}}$ le nombre de minimiseurs n'est pas fini, ce qui rend d'une part impossible la résolution du problème (i.e. la recherche de tous les minimiseurs), et d'autre part cela risque de conduire à un choix de prédicteur calibré sur les données (comme le peigne de Dirac, fonction irrégulière qui met une masse à l'endroit des données), dont le pouvoir de prédiction (que l'on peut constater empiriquement en étudiant la différence entre $\hat{\phi}(X_j)$ et Y_j pour un assuré j qui n'a pas servi à la construction de $\hat{\phi}$) est donc très réduit : c'est le phénomène de **sur-apprentissage** (i.e. fort pouvoir explicatif). Imposer une structure aux éléments de \mathcal{F} (fonctions linéaires par exemple) peut donc améliorer le pouvoir prédictif de $\hat{\phi}^{MRE}$. Aussi, pour finir de poser le cadre de cette étude, nous revenons, dans la section qui suit, plus en détail sur ce dilemme entre pouvoir explicatif et pouvoir prédictif, puisque ces deux pouvoirs vont nous aider à comparer les méthodes d'apprentissage entre elles.

3.3 - Pouvoir explicatif et pouvoir prédictif

La distinction entre pouvoir explicatif et pouvoir prédictif repose sur la décomposition biais-variance, non pas de l'excès de risque (comme en page 24), mais celle du risque :

$$\begin{aligned} R(\hat{\phi}) &= \mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) - Y \right]^2 \right] = \mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) \right]^2 - 2 \times \hat{\phi}(\mathbb{X}) \times Y + Y^2 \right] \\ &= \mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) \right]^2 \right] - 2 \times Y \times \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] + Y^2 \\ &= \mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) \right]^2 \right] - \underbrace{\mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right]^2 + \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right]^2}_{=0} - 2 \times Y \times \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] + Y^2 \\ &= \underbrace{\mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) \right]^2 \right] - \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right]^2}_{\operatorname{Var}[\hat{\phi}(\mathbb{X})]} + \underbrace{\mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right]^2 - 2 \times Y \times \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] + Y^2}_{\left(\mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] - Y \right)^2 = \mathbb{E} \left[\hat{\phi}(\mathbb{X}) - Y \right]^2} \end{aligned}$$

d'où :

$$R(\hat{\phi}) = \mathbb{E} \left[\left[\hat{\phi}(\mathbb{X}) - Y \right]^2 \right] = \underbrace{\operatorname{Var} \left[\hat{\phi}(\mathbb{X}) \right]}_{\text{Variance de } \hat{\phi}(\mathbb{X})} + \underbrace{\mathbb{E} \left[\hat{\phi}(\mathbb{X}) - Y \right]^2}_{\text{Biais}^2 \text{ de } \hat{\phi}(\mathbb{X})}$$

Il ne faut donc pas forcément chercher à construire un prédicteur $\hat{\phi}$ ayant un faible biais (fort pouvoir explicatif) et donc une grande variance (faible pouvoir prédictif) si l'on cherche à bien prédire. Ainsi, cette décomposition illustre le compromis biais-variance, et soulève la nécessité de distinguer et d'évaluer deux niveaux de performances : d'une part le pouvoir explicatif, mesuré sur l'échantillon d'apprentissage par le risque empirique $R_n^{appr} = \frac{1}{n_{appr}} \sum_{i=1}^{n_{appr}} \ell[\hat{\phi}(\mathbb{X}_i), Y_i]$ (erreur d'apprentissage), et d'autre part, le pouvoir prédictif ou pouvoir de généralisation sur un échantillon test disjoint de l'échantillon d'apprentissage, et mesuré soit par $R_n^{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \ell[\hat{\phi}(\mathbb{X}_i), Y_i]$ (erreur de prévision), soit par d'autres indicateurs selon le contexte et les objectifs métier (par exemple l'AUC ou le score de Brier comme nous le verrons plus loin).

II - Analyse exploratoire, transformation et échantillonnage des données

L'étape essentielle et préalable à toute étape de modélisation est l'exploration des données afin d'appréhender la structure conjointe et relative des différentes variables, de détecter d'éventuelles incohérences révélatrices d'irrégularités, ou encore des valeurs aberrantes qui pourraient créer un biais. Cependant, pour ce faire, nous n'allons pas perdre de vue la distinction entre la variable cible d'une part et les variables explicatives d'autre part, et nous veillerons à ce que l'information contenue dans ces dernières soit la moins redondante d'une variable explicative à l'autre.

Aussi, dans cette partie, nous tentons d'avoir une première idée des liaisons entre variable cible, qualitative binaire, et variables explicatives, ce qui pourrait constituer une première idée et une anticipation d'un éventuel lien qu'une modélisation ultérieure serait susceptible de révéler.

Cette partie a pour objectif de décrire la méthodologie utilisée, et afin de ne pas surcharger cette partie, celle-ci ne traite pas de l'analyse exploratoire et des transformations opérées sur la deuxième base (servant à étudier la rentabilité). Pour plus de détails, se référer à l'annexe afférente [page 92](#).

1 - Distributions des variables explicatives et lien avec la variable cible

1.1 - Les variables quantitatives et leur discrétisation

Le découpage en classes des variables quantitatives ne simplifie pas les modèles, au contraire, celui de la régression logistique est plus complexe car présente plus de paramètres à estimer (si le nombre de modalités d'une variable est strictement supérieur à 2 modalités). Il devient cependant plus flexible.

Nous nous limiterons donc à la régression classique, aux arbres et aux forêts car certaines méthodes ne sont pas adaptées à la présence de variables explicatives qualitatives. Les SVM par exemple peuvent les prendre en compte à condition de définir un noyau adapté mais ce n'est pas standard.

L'analyse graphique de la proportion de sinistrés dans les intervalles (bornés par les vingtiles) partitionnant chacune des variables explicatives continues peut révéler des seuils ou des liaisons définissant ainsi des regroupements d'intervalles en modalités. Certaines liaisons comme les liaisons linéaires peuvent néanmoins être conservées telles quelles (sans transformations) à conditions que les effectifs dans chacun des intervalles d'une variable explicative ne soient pas

faibles (d'où le choix d'une segmentation des variables par déciles ou vingtiles : cf infra).

Les six variables continues sont l'âge à la souscription, la durée en portefeuille, la prime annuelle moyenne payée toutes taxes comprises, ainsi que la part (en pourcentage) allouée à chacune des 3 garanties au sein de cette prime.

Âge à la souscription. Dans l'intervalle $]16; 33]$ des quatre premiers déciles, l'âge à la souscription présente une proportion de sinistrés (par rapport à la moyenne de 16,72%) globalement plus élevées que les déciles qui suivent. Le franchissement du seuil de sinistralité que constitue cette moyenne de 16,72% est également visible dans l'intervalle à partir de l'âge de 46 ans (deux derniers déciles). Dans l'entre-deux en revanche, c'est-à-dire entre 34 ans et 45 ans inclus, on observe une inflexion. Nous découperons donc la population en 3 parts afin d'apporter de la simplicité : $]16; 33]$, $]34; 45]$ et enfin $]46; 65]$.

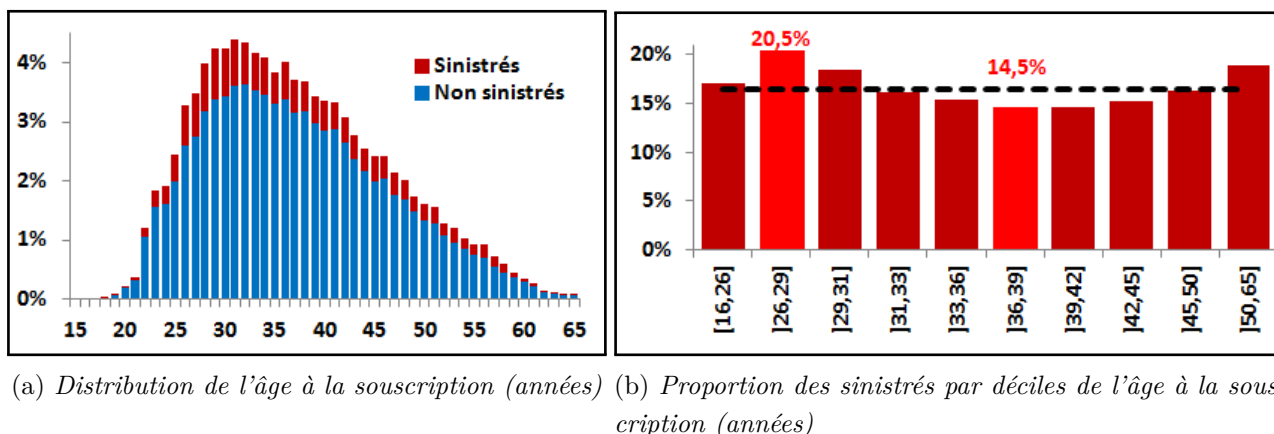
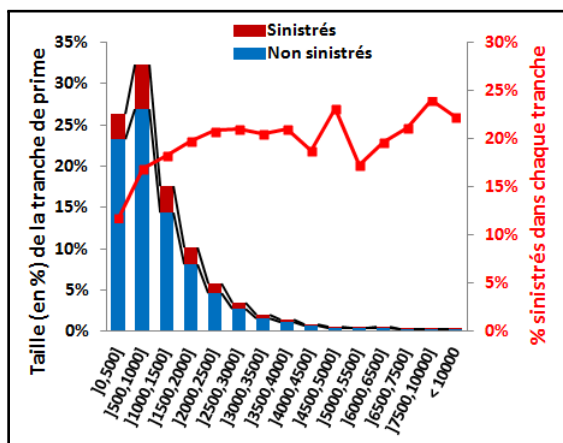


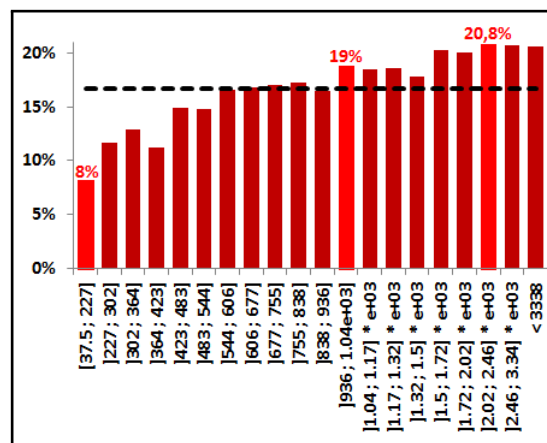
FIGURE II.1 – Histogrammes relatifs à l'âge à la souscription.

Prime annuelle moyenne TTC. Quant aux primes versées par les assurés, on constate sur la figure suivante que le montant a clairement une relation linéaire avec le taux de sinistrés. Le franchissement du seuil que constitue la moyenne du taux de sinistrés se situe au niveau d'une prime de 936 €, nous décidons donc pour plus de simplicité d'arrondir ce montant à 900€, qui va donc partager la population en deux suite à la discrétisation de la variable.

Cependant, il n'est pas évident de déterminer le lien de cause à effet : est-ce qu'un assuré qui, une fois qu'il a payé un montant élevé de prime et bénéficie donc d'une bonne couverture, va plus facilement être en incapacité de travail, ou bien est-ce qu'un assuré qui se sait à risque va payer plus de primes ?



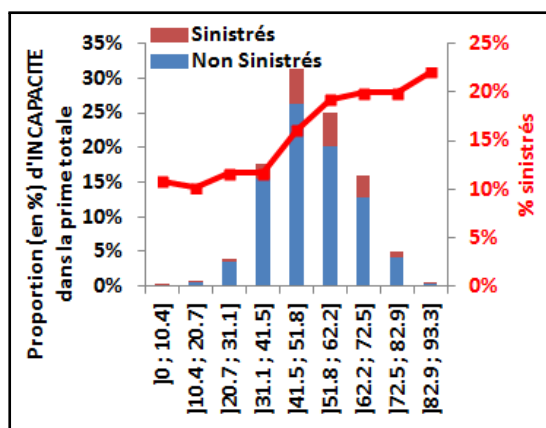
(a) Distribution de la prime annuelle moyenne TTC (€)



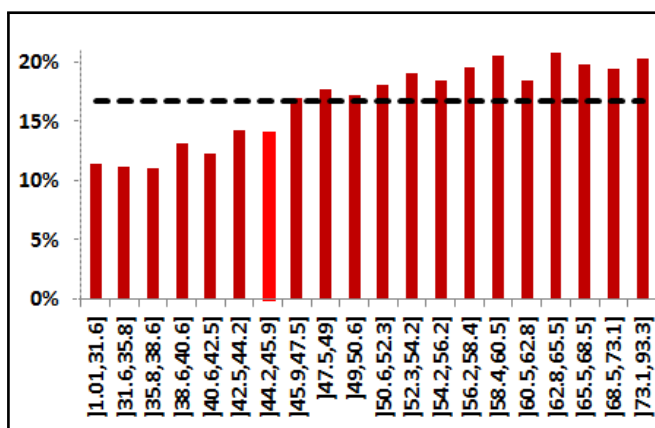
(b) Proportion des sinistrés par vingtiles de la prime annuelle moyenne TTC (€)

FIGURE II.2 – Histogrammes relatifs à la prime annuelle moyenne TTC (€).

Proportions des 3 garanties dans la prime annuelle moyenne. La part allouée à chacune des trois garanties (décès, incapacité, invalidité), et plus particulièrement à l'incapacité, peut véhiculer une information précieuse, à l'instar de la prime annuelle moyenne : le degré de confiance d'un assuré en son état de santé, ce qui peut revenir en une estimation de ses chances d'être sinistré. De plus, il est également intéressant de déterminer un éventuel seuil au-dessus duquel la proportion de sinistrés augmente de manière relativement significative, notamment par rapport à la proportion moyenne de sinistrés qui est de 16,72%.



(a) Distribution des assurés selon la part (en %) de l'Incapacité dans leur prime totale



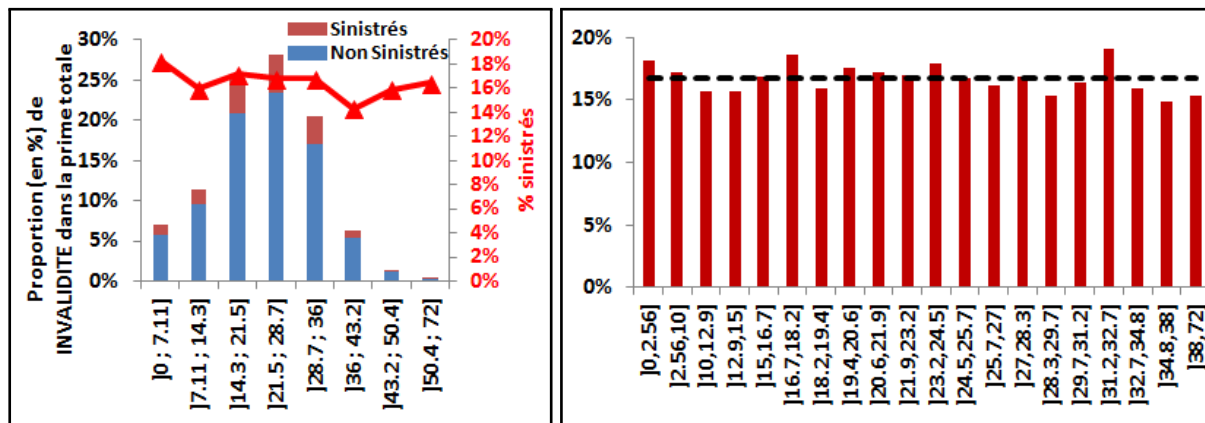
(b) Proportion des sinistrés par vingtiles de la part (en %) de la garantie Incapacité dans leur prime totale

FIGURE II.3 – Histogrammes relatifs à la proportion de prime allouée à la garantie Incapacité (en % de la prime annuelle moyenne TTC).

En effet, au vu de la figure II.3 précédente, il semblerait qu'un tel seuil existe et se situe autour de 46%.

Aussi, si l'on s'intéresse à la garantie invalidité, qui va souvent de paire avec la garantie incapacité à cause du fait qu'au bout de 3 ans d'incapacité un sinistré passe systématiquement

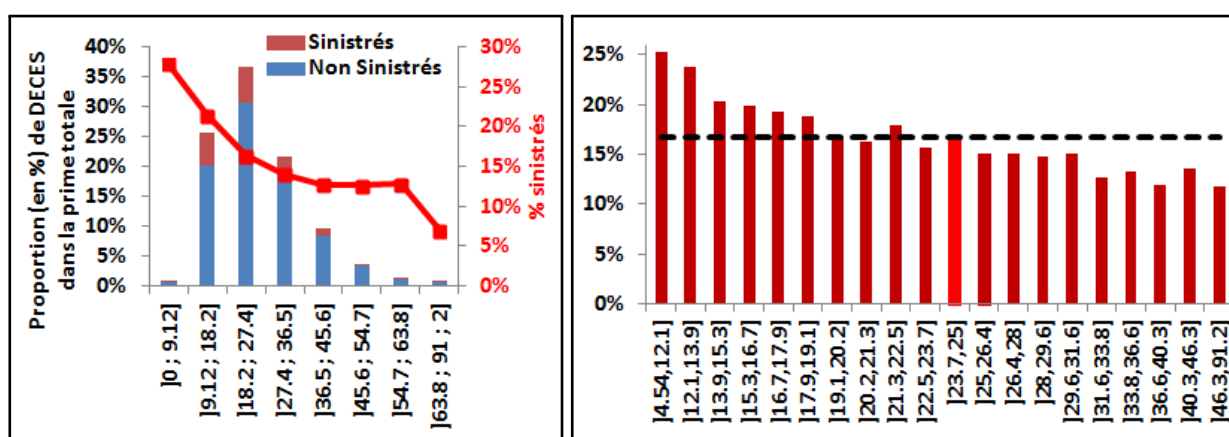
en invalidité si son état ne s'améliore pas, on constate que quelle que soit la proportion de prime allouée à cette garantie invalidité, la proportion de sinistrés est sensiblement la même. Une telle variable ne semble donc à priori pas intéressante pour la suite : nous l'écartérons donc lors de la modélisation de la [partie III](#) au cours de laquelle nous ne garderons donc que la part de prime allouée aux garanties incapacité et décès.



(a) Distribution des assurés selon la part (en %) de l'Invalidité dans leur prime totale (b) Proportion des sinistrés par vingtiles de la part (en %) de la garantie Invalidité dans leur prime totale

FIGURE II.4 – Histogrammes relatifs à la proportion de prime allouée à la garantie Invalidité (en % de la prime annuelle moyenne TTC).

En ce qui concerne la part de prime allouée à la garantie décès, le constat qui est fait est à l'opposé de l'observation faite pour la garantie Incapacité. D'après la figure [II.5](#) suivante, en dessous de 25% de prime attribuée à la garantie décès, un assuré a un risque accru d'avoir un sinistre incapacité.

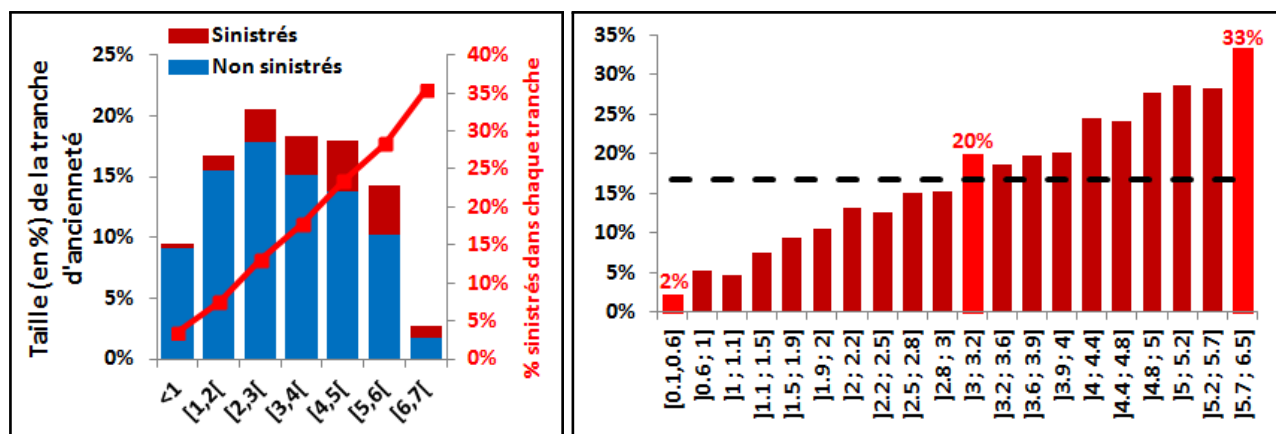


(a) Distribution des assurés selon la part (en %) du Décès dans leur prime totale (b) Proportion des sinistrés par vingtiles de la part (en %) de la garantie Décès dans leur prime totale

FIGURE II.5 – Histogrammes relatifs à la proportion de prime allouée à la garantie Décès (en % de la prime annuelle moyenne TTC).

Ancienneté en Portefeuille. Le constat est le même que pour la prime annuelle moyenne, puisque le taux de sinistralité croît à mesure que l'assuré est présent depuis une longue date,

notamment à partir de 3 années, qui sert donc de valeur de segmentation de la population en deux parties :



(a) Distribution de l'ancienneté en portefeuille (b) Proportion des sinistrés par vingtiles de l'ancienneté en portefeuille (années)

FIGURE II.6 – Histogrammes relatifs à l'ancienneté en portefeuille (années).

A noter que ce constat peut sembler logique à première vue puisqu'une plus grande durée en portefeuille augmente la fenêtre d'exposition, et puisque cela signifie que l'âge des assurés augmente. Nous verrons, d'un point de vue statistique, lors de l'étude du V de Cramer et lors de la modélisation que cette variable a une forte influence sur la variable réponse. Et, puisque la tarification pour le produit d'intérêt se fait en considérant l'âge à la souscription et non pas l'âge atteint, il serait judicieux d'un point de vue opérationnel de suggérer une majoration selon l'ancienneté. Le seul bémol concernant cette variable est qu'elle est une variable temporelle qui n'est pas entièrement observée (présence de censures à droite).

1.2 - Les variables qualitatives et le regroupement de leurs modalités

Une des principales motivations de cette étude est de savoir s'il y a un effet régional sur la sinistralité. Ainsi il est intéressant de remarquer au vu de la figure II.7 suivante que deux zones spatiales s'opposent : une sorte de pôle intérieur, où la proportion de sinistrés est faible, et, à l'inverse, une sorte d'arc extérieur l'entourant, où la proportion de sinistrés y est plus élevée.

Nous simplifierons donc la variable « Région » en une nouvelle variable « **zone spatiale** » ayant ces deux modalités, d'autant plus qu'il peut être utile de regrouper les modalités car les arbres de décision ont tendance à préférer les variables qui ont le plus de modalités pour les scissions.

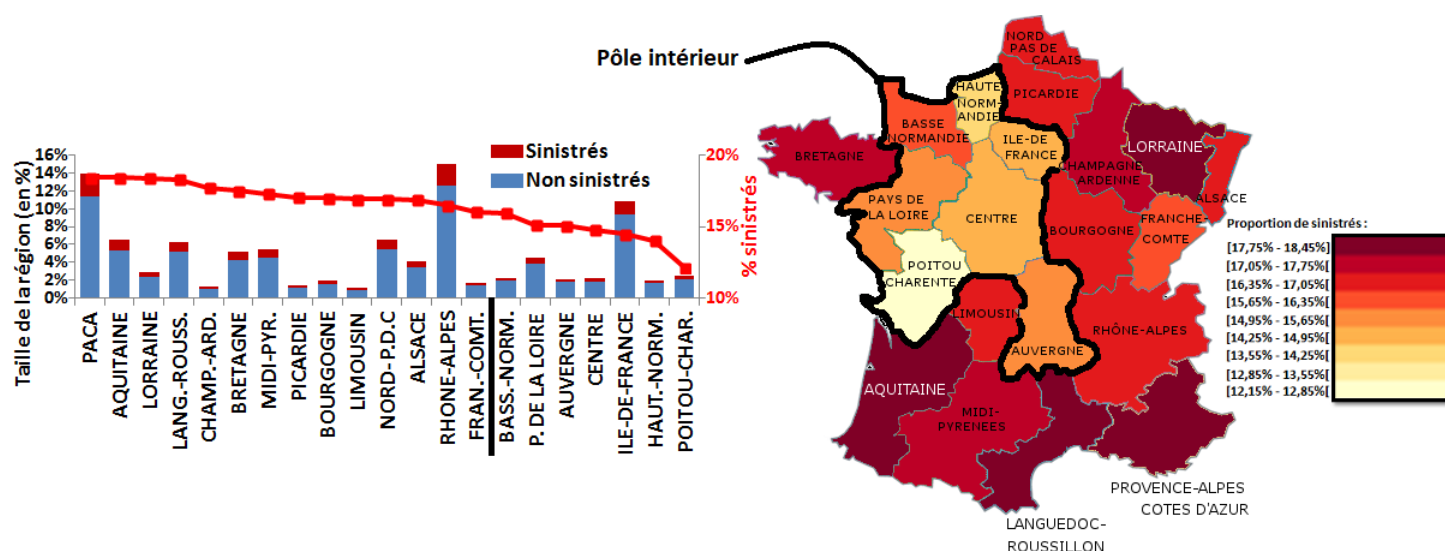


FIGURE II.7 – Histogramme et cartographie des 21 régions en fonction de la proportion de sinistrés.

Quant au **réseau commercial**, la proportion de sinistrés par modalité (Oui/Non) de la figure II.8 suivante ne semble pas indiquer un lien significatif avec la variable cible puisqu'il y a environ 17% de sinistrés pour la majorité des six modalités (Agents-Oui, Agents-Non, Courtiers-Oui,...), et sachant que les assurés ayant eu une fois recours à un réseau « divers », et pour lesquels il y a 12% de sinistrés, sont faibles en proportion, leur différence a donc une portée limitée.

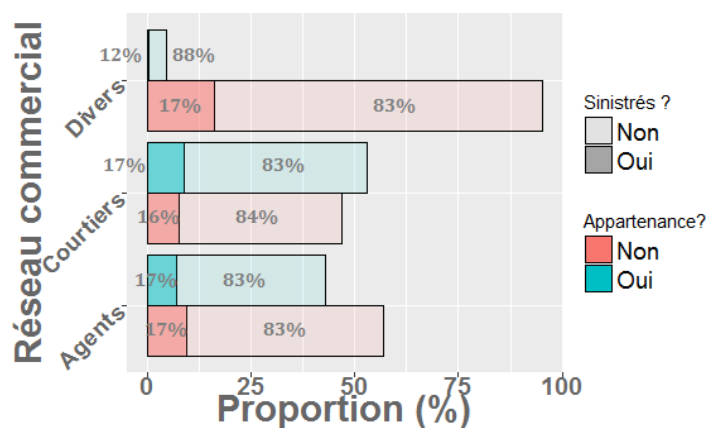


FIGURE II.8 – Proportion des assurés selon leur appartenance aux réseaux commerciaux.

Enfin, concernant le **reste des variables** ci-dessous (figure II.9), une **première partie** composée du sexe, de la convention, des frais d'adhésion et de l'abattement tarifaire présente des proportions de sinistres qui varient de manière relativement significative d'une modalité à l'autre d'une même variable, et avec des effectifs souvent suffisants ou du moins non négligeables pour chacune d'elles, ce qui vient accorder du crédit à ces différences de proportions. En revanche, la **deuxième partie** des variables constituée de la catégorie professionnelle, de la situation de famille, de la périodicité de prime, de la fiscalité et de l'indicatrice des surprimés présente une relativement faible variabilité de la proportion des sinistres d'une modalité à l'autre d'une même variable mais peut tout de même présenter un lien avec la variable cible, par exemple si la

majorité des modalités ont une proportion de sinistrés supérieure à la moyenne du portefeuille (16.7%).

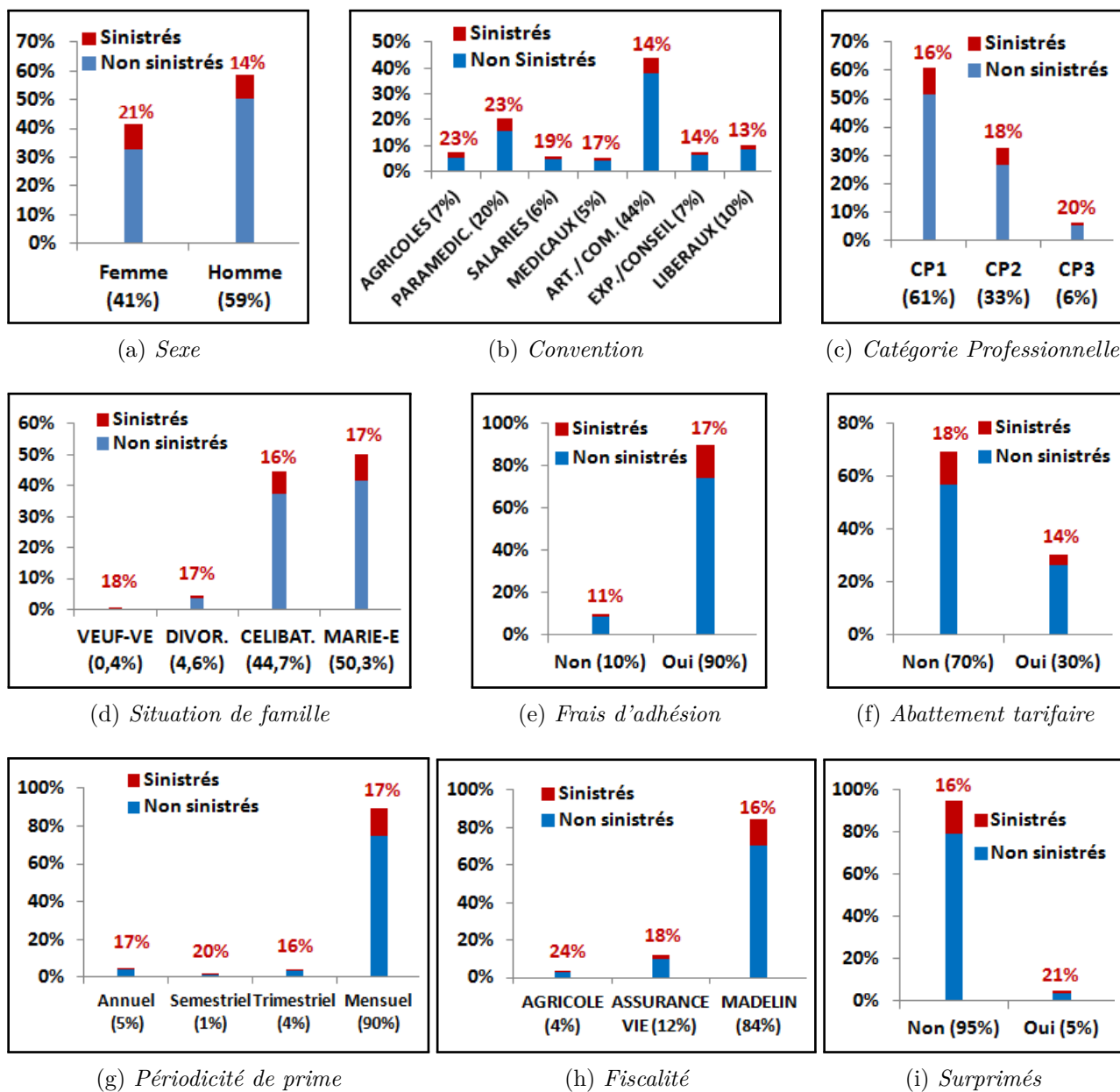


FIGURE II.9 – Distributions de 9 variables explicatives catégorielles selon la proportion de sinistrés dans chaque modalité.

Nous opérons donc des regroupements de modalités, et le tableau de la page suivante donne un vue détaillée des données finales.

Il est intéressant de noter que les assurés supprimés présentent en leur sein une part de sinistrés (21%) qui se trouve être plus grande (en proportions) que celle des assurés sinistrés parmi les non supprimés (16%). Ce constat est rassurant car la différenciation des assurés en fonction de l'exposition à certains risques permet d'isoler et donc de redresser une partie du portefeuille moyennant une surprime. L'utilité de la surprime se fait donc déjà ressentir. Nous verrons à la

partie IV comment est la rentabilité des surprimés sinistrés pour savoir si ce redressement est correctement fait.

En outre, une femme a plus de chances d'être sinistrée qu'un homme (écart de 7.2 points de pourcentage), ce qui est probablement attribuable, dans une certaine mesure, aux congés maternité. Enfin, notons qu'il y a un écart de 14.9 points de pourcentage entre les assurés ayant plus de 3 ans d'ancienneté en portefeuille et ceux ayant moins de 3 ans, en faveur de ces derniers qui sont moins sinistrés en proportion. L'utilité d'une majoration tarifaire en fonction de l'ancienneté en portefeuille (à partir de 3 ans) se fait également ressentir.

VARIABLES	MODALITÉS	EFFECTIF SINISTRÉS			
Zone Spatiale	Pôle intérieur / Arc extérieur	26.5%	14.5%	73.5%	17.5%
Convention	professions médicales	05.3%	17.1%		
	professions paramédicales	20.3%	23.1%		
	profession libérales / expert / conseil	17.3%	13%		
	artisans - commerçants	43.7%	13.8%		
	professions agricoles	07.2%	23.2%		
	salariés	06.1%	19.1%		
Sexe	femme / homme	41.5%	21.0%	58.5%	13.8%
Situation de Famille	en couple / seul(e)	50.3%	17.2%	49.7%	16.3%
Âge à la souscription	<=33 ans	40.2%	18.2%		
]]33 ; 45]] ans	40.3%	14.9%		
	>45 ans	19.5%	17.6%		
Ancienneté (années)	<=3 ans / >3 ans	53.0%	09.7%	47.0%	24.6%
Abattement tarifaire	oui / non	30.5%	13.9%	69.5%	18%
Catégorie Professionnelle	1 / 2-3	61.0%	15.7%	39.0%	18.3%
Surprimés	oui / non	04.9%	21.0%	95.1%	16%
Agents	oui (1) / non (0)	42.9%	16.8%	57.1%	16.6%
Courtiers	oui (1) / non (0)	52.9%	17.0%	47.1%	16.4%
Réseaux divers	oui (1) / non (0)	04.7%	12.3%	95.3%	16.9%
Périodicité prime	mensuelle / non mensuelle	89.7%	16.7%	10.3%	17.3%
Frais d'adhésion	oui / non	09.8%	10.7%	90.2%	17.4%
Fiscalité	Agricole	03.9%	23.7%		
	Assurance vie	11.8%	18.1%		
	Madelin	84.4%	16.2%		
Prime annuelle moyenne TTC (€)	<=900 € / >900 €	53.2%	14.3%	46.8%	19.4%
Proportion (prime) Incapacité	<=46% / >46%	35.3%	12.5%	64.7%	19%
Proportion (prime) Décès	<=25% / >25%	55.1%	19.2%	44.9%	13.7%

TABLE II.1 – Variables finales utilisées pour la modélisation - La proportion de sinistrés du portefeuille est de 16.7%.

1.3 - Vue globale des liaisons après transformations

Le V de Cramer permet d'avoir une idée du pouvoir discriminant (rappelons que l'on cherche notamment à discriminer les sinistrés des non sinistrés) d'une variable explicative. Il est défini par :

$$V_{Cramer} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

où k et r sont respectivement le nombre de modalités de la variable cible (ici binaire donc $k=2$)

et le nombre de modalités de la variable explicative considérée, et χ^2 est la statistique de test du Khi-deux de Pearson (cf. [Nan] pour plus de détail sur l'expression de cette statistique).

D'après la figure II.10 suivante, il semblerait que l'ancienneté en portefeuille ainsi que la convention soient les deux variables qui ont un bon pouvoir discriminant ($V_{Cramer} \in [0.1; 0.2[$). Il n'est donc pas étonnant que la modélisation par apprentissage mette en exergue ces deux variables. Aussi, remarquons le faible pouvoir discriminant de l'indicatrice des surprimés ce qui vient nuancer notre remarque sur l'utilité d'une surprime. La zone spatiale présente également un faible pouvoir discriminant et semble souffrir du regroupement pourtant nécessaire des nombreuses modalités relatives aux 21 régions étudiées et à l'issue duquel un niveau de finesse de l'information a été perdu.

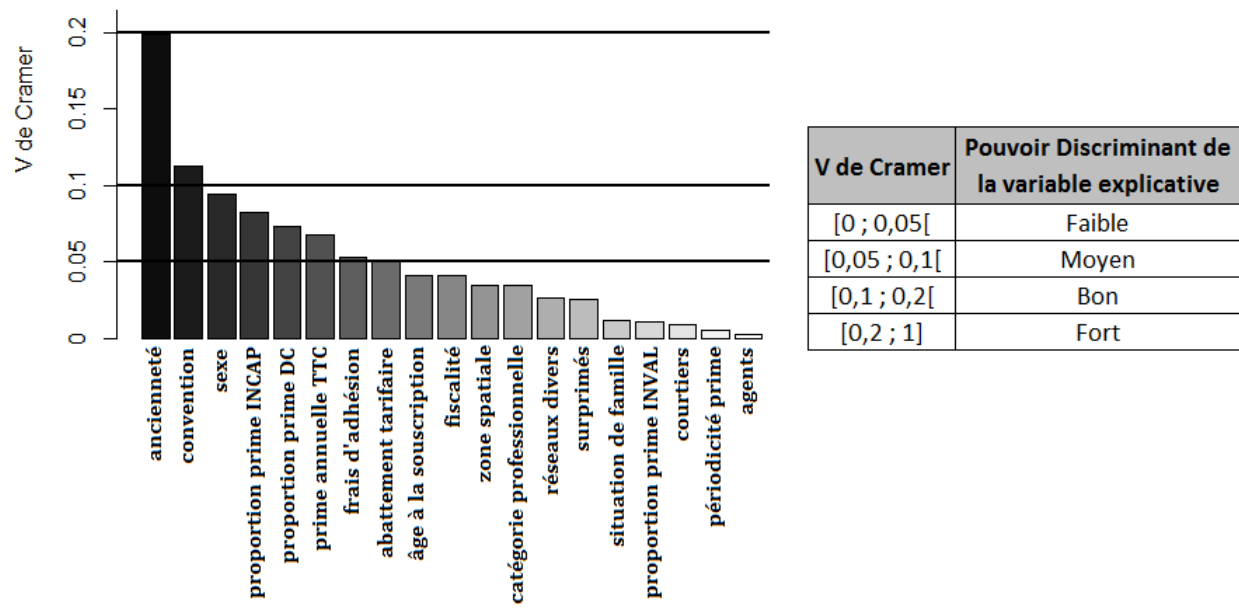


FIGURE II.10 – *V de cramer entre les variables explicatives qualitatives et la variable cible binaire.*

Toutefois, cette figure II.10 ne permet pas de prendre connaissance de la **complexité des données** et de l'interaction des modalités inter-variables dans l'explication de la sinistralité. En effet, si l'on prend l'exemple de l'âge à la souscription (V_{Cramer} inférieur à 5%) et le sexe (V_{Cramer} entre 5% et 10%), on constate que les femmes ayant souscrit jeunes ont apparemment plus de chances d'être sinistrées que les hommes qui ont souscrit jeunes d'après la figure II.11a. Ainsi, l'âge à la souscription, dont la valeur du V_{Cramer} est faible et indique un faible pouvoir discriminant lorsqu'il est considéré seul, permet une plus grande acuité dans la compréhension de la sinistralité lorsqu'il est considéré conjointement avec le sexe. De même, le croisement de l'âge à la souscription avec une variable ayant un bon pouvoir discriminant comme l'ancienneté en portefeuille apporte des nuances supplémentaires (de couleurs sur la figure II.11b) en plus des deux blocs rouge et bleu délimités par une ancienneté autour de 3 ans.

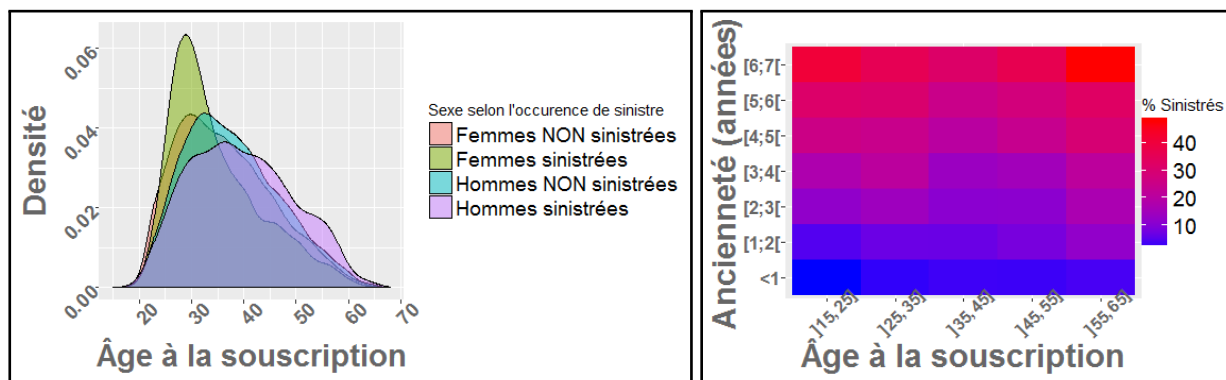
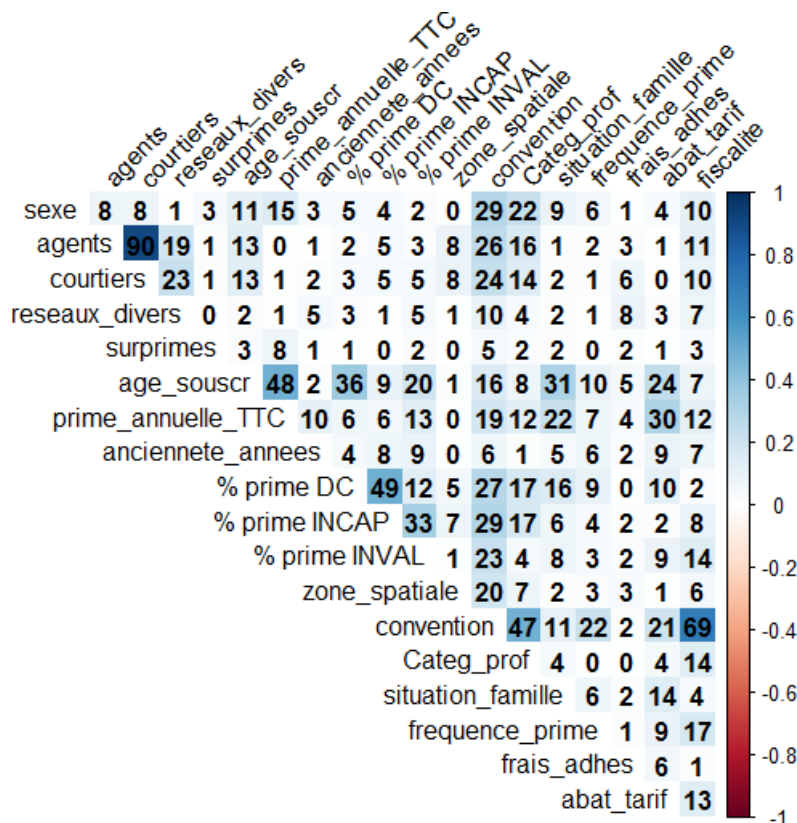
(a) Âge à la souscription \times Sexe(b) Âge à la souscription \times Ancienneté

FIGURE II.11 – Proportion de sinistrés pour le croisement de deux variables avec l'âge à la souscription.

Si l'on s'intéresse cette fois au V_{Cramer} entre variables explicatives, on se rend compte, au vu de la figure II.12, que pour les deux variables ayant le plus fort pouvoir discriminant sur la figure II.10 à savoir l'ancienneté et la convention, seule cette dernière est fortement liée ($V_{Cramer} > 40\%$) à d'autres variables explicatives (la catégorie professionnelle et la fiscalité).

FIGURE II.12 – V de Cramer (en %) entre variables explicatives, toutes qualitatives.

Cependant, nous choisissons de n'éliminer aucune variable explicative (hormis le cas évoqué en page 30 concernant la proportion de prime allouée à l'invalidité), car bien que certaines d'entre elles semblent liées, le croisement de leurs modalités paraît a priori pertinent pour la compréhension de la sinistralité comme en atteste la figure II.13 suivante : par exemple, un

assuré qui alloue plus de 46% de sa prime à l'incapacité et moins de 25% au décès a plus de chances d'être sinistré au moins une fois en incapacité qu'un assuré qui opte pour l'inverse (moins de 46% à l'incapacité et plus de 25% au décès). Quant aux assurés qui augmentent ou diminuent simultanément les parts de prime allouées au décès et à l'incapacité, ils ont des chances intermédiaires d'être au moins une fois en incapacité au regard de la proportion de sinistrés.

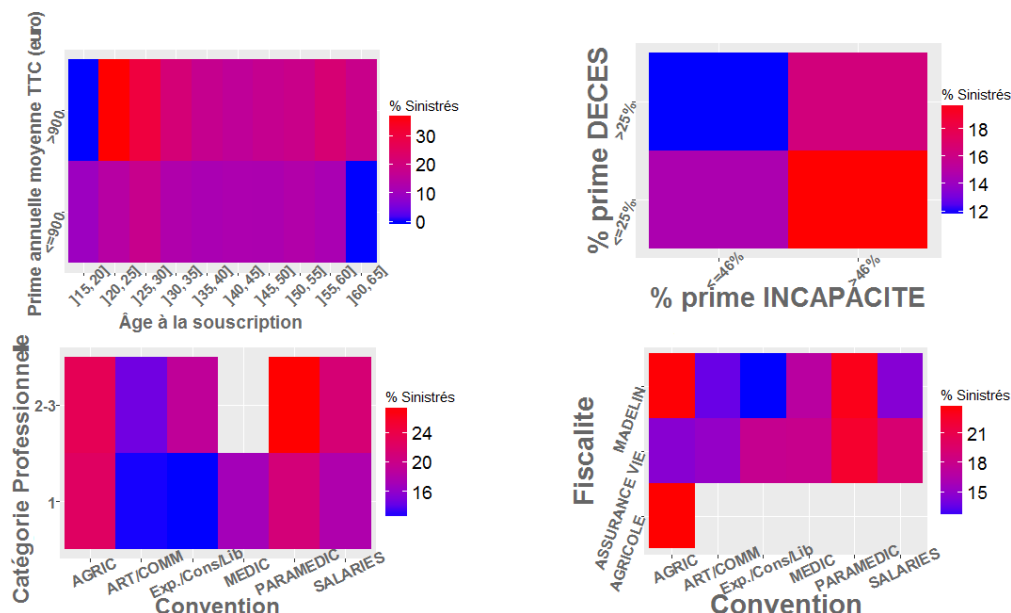


FIGURE II.13 – Proportion de sinistrés pour les croisements de modalités de certaines variables très liées entre elles d'après le V de Cramer ($>40\%$).

La figure II.14 suivante donne la représentation des modalités des variables explicatives « internes » (après discrétisation) sur le premier plan de l'Analyse Factorielle des Correspondances Multiples (AFCM), qui consiste à chercher des facteurs (obtenus en réalisant une Analyse en Composantes Principales¹ (ACP) sur le Tableau Disjonctif Complet² (TDC)) en nombre réduit (au plus 3 généralement) et résumant le mieux possible (en terme d'inertie³) les données.

1. L'ACP consiste à chercher une base $\text{Vect}(F_1, F_2, \dots, F_p)$ de l'espace des assurés, dont les axes orthogonaux (F_1, F_2, \dots, F_q) d'inertie maximum sont obtenus par combinaisons des variables numériques initiales (v_1, v_2, \dots, v_p) . L'ordre des indices $(1, \dots, q)$ indique de manière décroissante les facteurs qui résument le plus l'inertie des assurés : la variance des projections des assurés sur le facteur F_i est plus grande que la variance des projections des assurés sur le facteur F_j , avec $i < j$. Cette façon de procéder qui consiste à chercher tour à tour les facteurs maximisant la variance des projections, permet de concentrer l'inertie sur les premiers facteurs, permettant ainsi de se décharger du reste des facteurs, représentant une faible proportion de l'inertie.

2. Le TDC comporte en lignes les individus et en colonnes les indicatrices des modalités des variables qualitatives : la valeur à l'intersection de la ligne i et de la colonne m (associée à la variable M) vaut 1 si l'individu i possède la modalité m (de la variable M), et 0 sinon.

3. L'inertie est une notion de dispersion qui généralise la variance à au moins deux variables et qui est égale à la somme des variances des variables considérées.

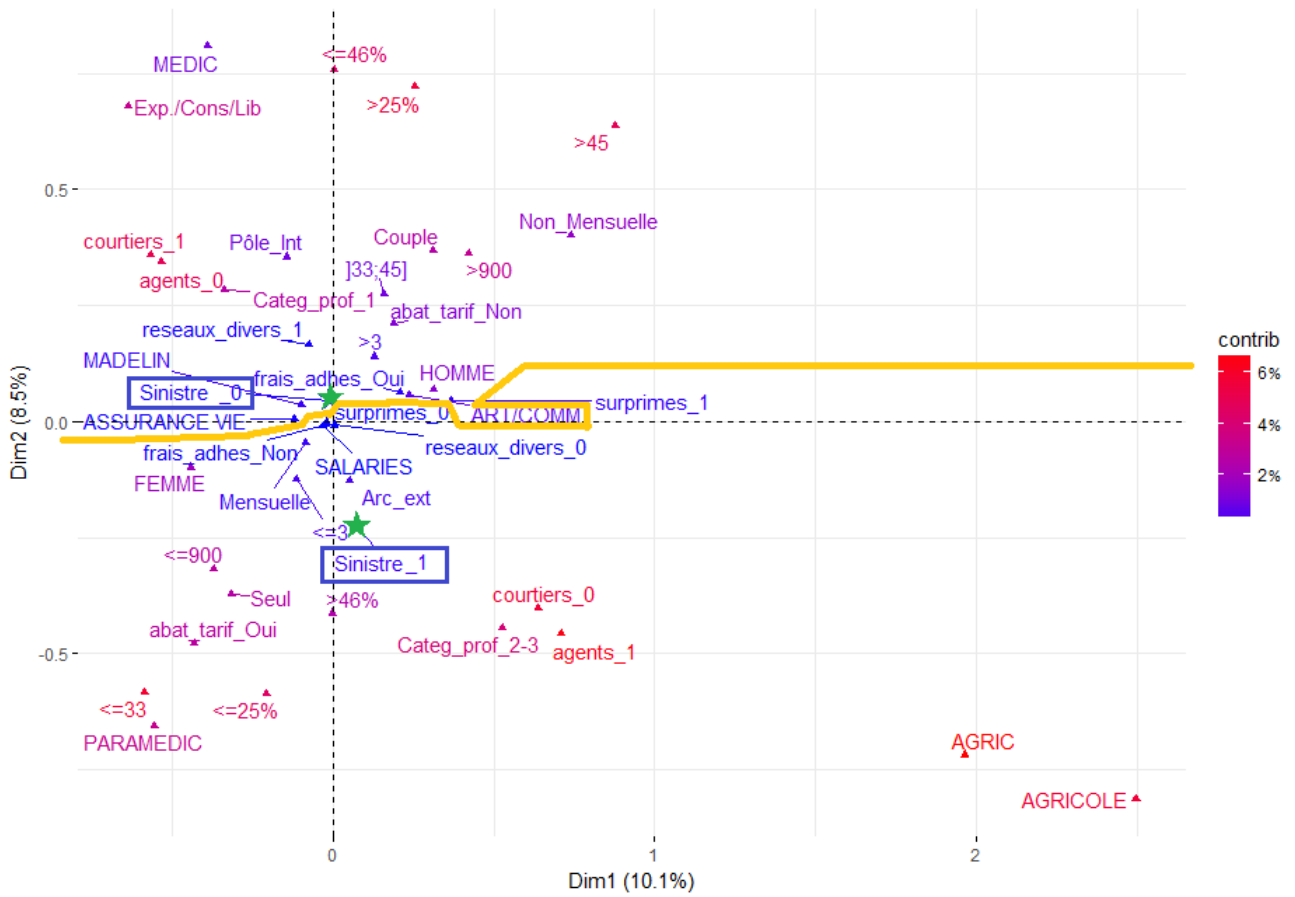


FIGURE II.14 – Représentation des 42 modalités des variables explicatives « internes » et des 2 modalités (★) de la variable cible sur le premier plan de l'AFCM, avec une coloration en fonction de leur contribution (en termes d'inertie) à ce plan. La frontière orange sépare les deux clusters obtenus par k-means.

L'analyse de la figure II.14 repose sur le repérage des modalités ayant des contributions importantes (en terme d'inertie) aux axes, pour pouvoir regarder ensuite leur positionnement sur le graphique, et ainsi donner du sens aux facteurs qui résument ici 18.6% de l'inertie totale. En effet, plus une modalité a une grande contribution, plus elle est prépondérante dans la définition et la caractérisation des axes pour lesquels elle a des coordonnées élevées. Ceci se comprend plus aisément avec la formule suivante donnant la contribution Ctr_l^k de la modalité l de la variable L à l'axe $k \in \{1, 2\}$ de l'AFCM :

$$Ctr_l^k = \frac{\frac{n_{l+}}{n} \times (c_l^k)^2}{\mathcal{I}_k} \text{ avec } \begin{cases} n_{l+} = \text{nombre d'assurés ayant la modalité } l \text{ de la variable } L \\ n = 40\,790 = \text{nombre total d'assurés} \\ c_l^k = \text{coordonnée de la modalité } l \text{ de la variable } L \text{ sur l'axe } k \\ \mathcal{I}_k = \sum_{j=1}^r \frac{n_{j+}}{n} \times (c_j^k)^2 = \text{Inertie de l'axe } k, \\ \text{où } r = 42 = \text{nombre total de modalités} \end{cases}$$

Une vision duale de ce type de graphique repose sur l'idée qu'une modalité est au centre de gravité des individus qui la possèdent. La représentation des modalités sur les plans factoriels

est donc plus lisible et plus synthétique que la représentation des assurés sur ces mêmes plans factoriels, d'autant plus que le nombre de modalités est très inférieur au nombre d'assurés (42 modalités et 40 790 assurés).

De ce que l'on observe, la modalité « Sinistre_1 » désignant l'évènement $\{Y = 1\}$ (« assuré sinistré au moins une fois en incapacité ») de la variable cible se positionne clairement en bas du centre du référentiel (coordonnée négative sur l'axe 2), et s'oppose à l'autre modalité complémentaire « Sinistre_0 » ($\{Y = 0\}$) sur cet axe. Autrement dit, le deuxième facteur de l'AFCM sert à discriminer les assurés sinistrés des assurés non sinistrés.

Une aide à la décision consiste à dessiner une frontière (en orange sur la figure [II.14](#)) par k -means avec $k = 2$ classes : cette méthode d'apprentissage non supervisé sera détaillée à partir de la page [79](#). La frontière obtenue étant horizontale, cela confirme bien que c'est l'axe 2 qui sépare le mieux les sinistrés des non sinistrés puisqu'il est « coupé » en deux par cette frontière orange.

Ainsi, l'AFCM aidée d'un k -means permet de mettre en évidence l'influence des modalités des variables explicatives sur l'évènement d'intérêt (occurrence d'au moins un sinistre) en étudiant leur distance avec les deux modalités de la variable cible sur le deuxième axe. Ainsi, les agricoles, les CP 2-3, les paramédicaux, les assurés de moins de 33 ans, les personnes vivant seules, les assurés ayant intégré le portefeuille par l'intermédiaire d'un agent et non d'un courtier, les assurés payant moins de 900€⁴ de primes annuelles TTC en moyenne toutes garanties confondues, ceux dont la proportion de primes allouée à l'incapacité dépasse 46% de la prime annuelle moyenne TTC et ceux bénéficiant d'un abattement tarifaire créateur sont ceux qui sont le plus susceptibles d'être sinistrés, du fait de la proximité sur l'axe 2 des modalités afférentes avec la modalité « Sinistre_1 » de la variable cible.

Notons toutefois que l'inertie globale de ce premier plan de l'AFCM ne résume que 18.5% de l'inertie totale du nuage des assurés. Or, il est d'usage de considérer que le premier plan résume suffisamment l'inertie totale si cette dernière s'y retrouve à hauteur de 70%. Il faut donc accorder un niveau de confiance modéré à la lecture de la figure [II.14](#).

Aussi, nous allons dans la partie suivante laisser place à la modélisation par apprentissage qui tentera de déterminer la part de chaque variable dans l'explication et la prévision de la variable cible. Mais avant cela, nous nous employons dans les deux sections suivantes, d'une part à vérifier le bien-fondé de la jointure des données externes avec les données du portefeuille, et, d'autre part à séparer la population en deux échantillons, d'apprentissage et de test.

2 - Validité et traitement des données externes

Validité. Il importe, avant d'envisager toute modélisation se basant sur des données externes, de vérifier si la distribution de la population de notre portefeuille est la même que celle de la population générale française, selon le département, au moyen du coefficient de corrélation et

4. Pourtant, il y a 19.4% de sinistrés parmi les assurés qui paient moins de 900€ de prime annuelle en moyenne contre 14.3% chez ceux qui en paient plus, toutes garanties confondues.

du test d'homogénéité du χ^2 .

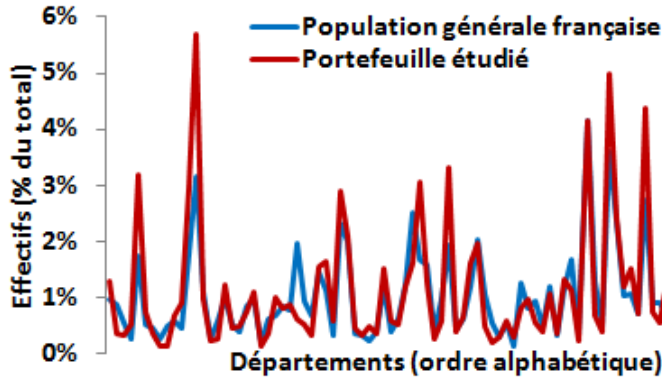


FIGURE II.15 – Répartition départementale des effectifs du portefeuille étudié et de la population générale française.

Aussi, bien que le test du χ^2 nous amène à rejeter l'hypothèse H_0 d'homogénéité des deux populations au risque de première espèce $\alpha = 5\%$ (p-valeur $\approx 0 < 0.05$), le test de corrélation de Pearson (coefficient=0.8016) nous permet de rejeter l'hypothèse H_0 de nullité du coefficient de corrélation afférent, avec le même niveau de confiance de 5% (p-valeur de l'ordre de 10^{-6}). Cette grande conformité des deux populations, qui peut s'apprécier sur le graphique de la figure II.15, légitime le recours aux données externes relatives à la population générale française pour tenter d'apporter plus d'informations susceptibles d'expliquer la sinistralité du portefeuille d'intérêt.

Pré-sélection. Le nombre élevé de variables externes nous conduit irrémédiablement à une sélection, pour n'en garder que les plus pertinentes. Nous allons donc recourir à 2 critères de sélection : le score $S_{Y/X}$ et le test de Wilcoxon-Mann-Whitney.

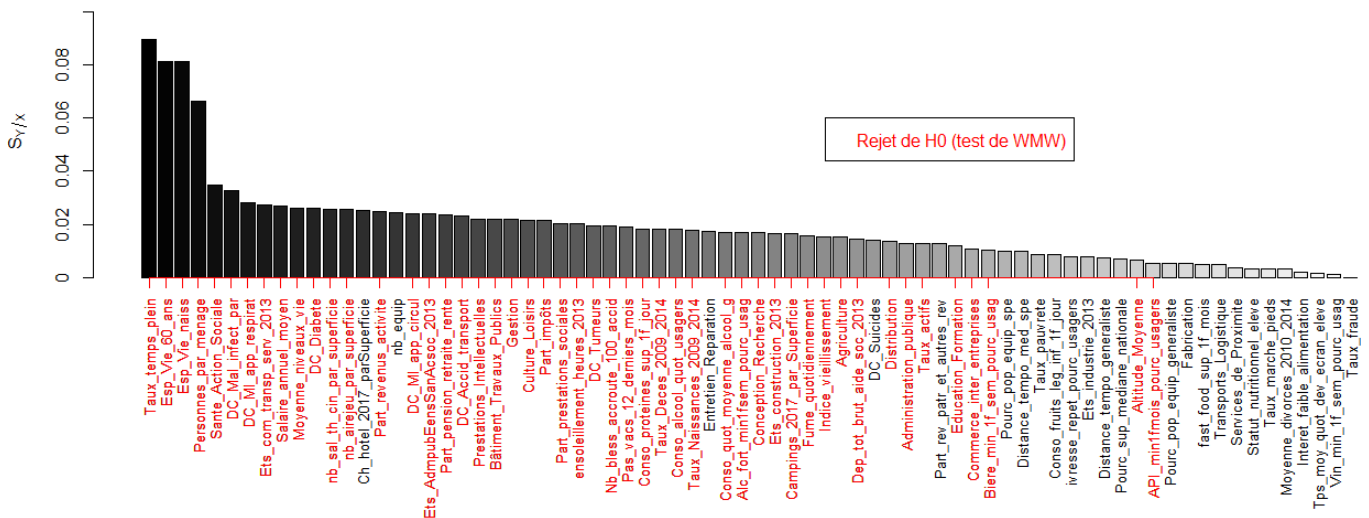


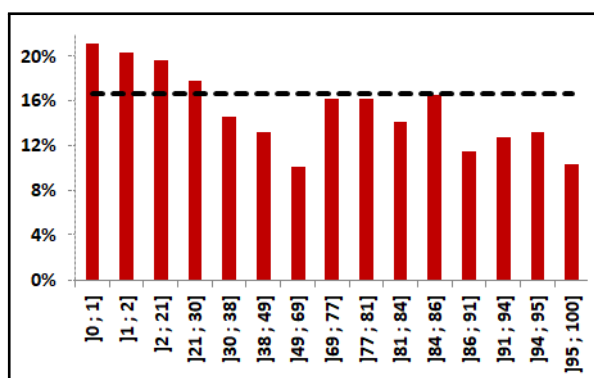
FIGURE II.16 – $S_{Y/X}$ entre les variables externes toutes quantitatives et la variable cible binaire, et mise en évidence en rouge d'une liaison significative (rejet de l'hypothèse H_0 du test de WMW) au risque de première espèce $\alpha = 5\%$.

Le premier est défini par la relation $S_{Y/X} := \sqrt{\frac{\sigma_E^2}{\sigma_E^2 + \sigma_R^2}}$, où $\sigma_E^2 = \frac{1}{n} \sum_{l=1}^2 n_l \times (\bar{y}_l - \bar{y})^2$ est la

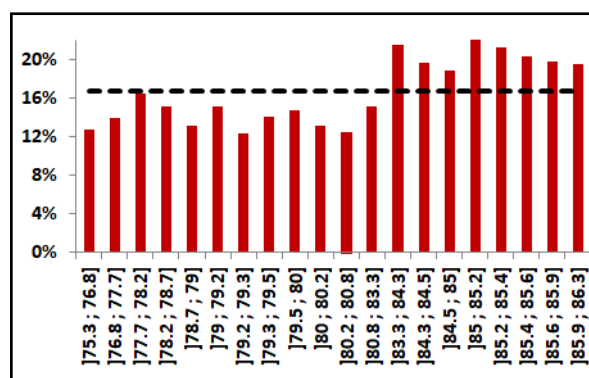
variance inter-classes et $\sigma_R^2 = \frac{1}{n} \sum_{l=1}^2 n_l \times \sigma_l^2$ est la variance intra-classes, les deux classes « l » étant les sinistrés et les non sinistrés⁵. Le second critère est un test d'hypothèses, dont le rejet de l'hypothèse H_0 signifie que la localisation (au sens des rangs moyens) de la variable externe considérée n'est pas la même dans les deux populations (sinistrés et non sinistrés). Pour plus de détails sur la statistique de test, se référer à [Raka].

Nous choisissons donc de ne garder que les trois variables dont le score $S_{Y/X}$ est le plus élevé, et pour lesquelles l'hypothèse H_0 du test de Wilcoxon-Mann-Whitney est rejetée : le taux de travail⁶ à temps plein, l'espérance de vie à la naissance⁷ et le nombre de personnes par ménage.

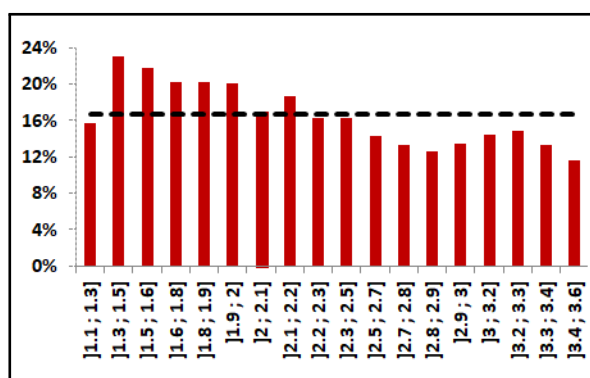
Discrétisation. La discrétisation se fait de la même manière que les variables « internes », puisque l'on va choisir une valeur de la variable externe qui va séparer les valeurs supérieures des valeurs inférieures pour qu'elles soient globalement associées à des proportions de sinistrés plus élevées ou moins élevées que la proportion moyenne de sinistrés (16.7% marqués par un trait en pointillés sur les figures suivantes).



(a) Taux de travail à temps plein



(b) Espérance de vie à la naissance



(c) Personnes par ménage

FIGURE II.17 – Histogrammes de la proportion de sinistrés par intervalles de mêmes effectifs, relativement aux 3 variables externes les plus liées à la variable cible.

5. Les notations n_l , \bar{y}_l et σ_l^2 désignent respectivement l'effectif, la moyenne et la variance des valeurs de Y pour les assurés de la classe $l = \{\text{sinistrés, non sinistrés}\}$.

6. Le taux de travail à temps plein désigne la proportion de la population travaillant à temps plein.

7. L'espérance de vie à 60 ans n'est pas retenue à cause d'une signification proche de celle de l'espérance de vie à la naissance.

On constate ainsi que les assurés qui habitent dans des régions où il y a moins de 30% de la population qui travaille à temps plein, des régions où l'espérance de vie à la naissance est supérieure à 83.3 ans, ou bien des régions où le nombre moyen de personnes par ménage est inférieur à 2.2, sont des assurés qui ont plus de chances d'être sinistrés que le reste du portefeuille.

Le tableau ci-dessous nous conforte dans nos choix de variables externes puisque la proportion de sinistrés varie de 5 à 7 points de pourcentage d'une des deux modalités à l'autre de chaque variable externe discrétisée :

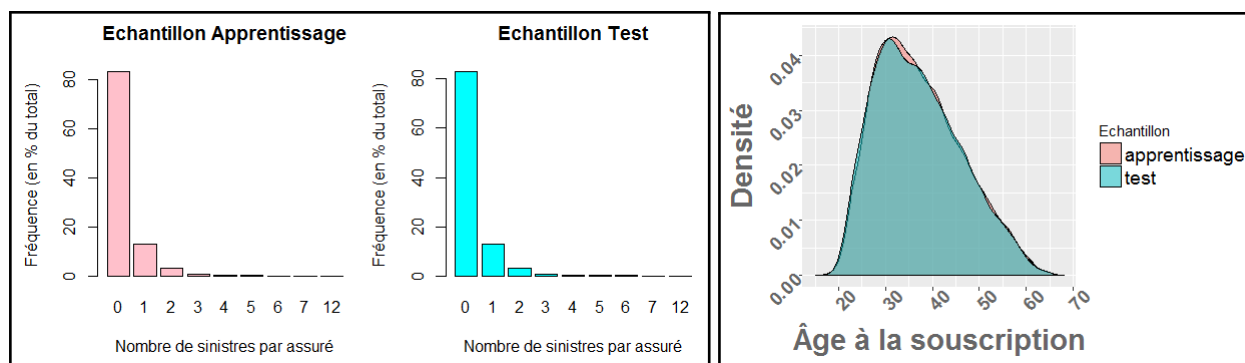
VARIABLES	MODALITÉS	EFFECTIF	SINISTRÉS
Taux de travail à temps plein	$\leq 30\%$ / $> 30\%$	45.3%	20.6% / 54.7% 13.5%
Espérance de vie à la naissance	≤ 83.3 ans / > 83.3 ans	61.6%	14.3% / 38.4% 20.6%
Personnes par ménage	≤ 2.2 / > 2.2	44.6%	19.5% / 55.4% 14.5%

TABLE II.2 – Variables externes finales utilisées pour la modélisation - La proportion de sinistrés du portefeuille est de 16.7%.

3 - Échantillonnage : Apprentissage et test

Nous devons veiller à ce que les échantillons d'apprentissage et de test soient semblables et comparables, afin d'éviter qu'une modalité peu fréquente soit absente de l'échantillon d'apprentissage et ainsi que les modèles ne soient construits sans elle. Une telle situation serait ennuyeuse dans la mesure où si une telle modalité se retrouve dans l'échantillon test, les modèles seront incapables de la prédire.

Nous choisissons à titre d'exemple de comparer les deux sous-échantillons en considérant le nombre de sinistres incapacité par assuré (variable plus fine que la variable cible), dont la valeur la plus élevée (12) se trouve être une valeur peu fréquente et est présente dans l'échantillon d'apprentissage aussi bien que l'échantillon test. Le constat est le même pour les deux queues de la distribution de l'âge à la souscription, qui sont représentées de la même manière dans les deux sous-échantillons. Après vérification pour toutes les variables, toutes les modalités présentes dans l'échantillon d'apprentissage sont bien présentes dans l'échantillon test.



(a) Nombre de sinistre Incapacité par assuré

(b) Âge à la souscription

III - Classification binaire des assurés selon la survenance de sinistres

Dans cette partie, la variable cible est :

$$Y = \begin{cases} 1 & \text{si l'assuré a été sinistré au moins une fois en incapacité} \\ 0 & \text{si l'assuré n'a jamais été sinistré en incapacité} \end{cases}$$

ce qui permet d'obtenir une estimation de la probabilité $\mathbb{P}[Y = 1 | \mathbb{X}]$ d'être sinistré sachant les caractéristiques \mathbb{X} .

1 - Arbres CART

L'utilisation d'arbres de décision présente de nombreux avantages. En effet, ils résistent convenablement à l'absence de relation de linéarité entre variable cible et variables explicatives, ils permettent d'utiliser des variables cibles ou explicatives dissemblables (mélange de variables qualitatives et quantitatives) et le résultat obtenu est très simple à interpréter. Par ailleurs, bien que les données de ce mémoire ne présentent pas de valeurs manquantes, l'algorithme CART implémenté dans R¹ permet de les prendre en compte.

Dans notre cas, où nous souhaitons mettre en relief les variables permettant de classer la population des assurés en 2 groupes prédéfinis (les clients sinistrés indiqués par le signe « + », et ceux qui ne le sont pas, désignés par le signe « - »), l'arbre de décision est un arbre de segmentation binaire.

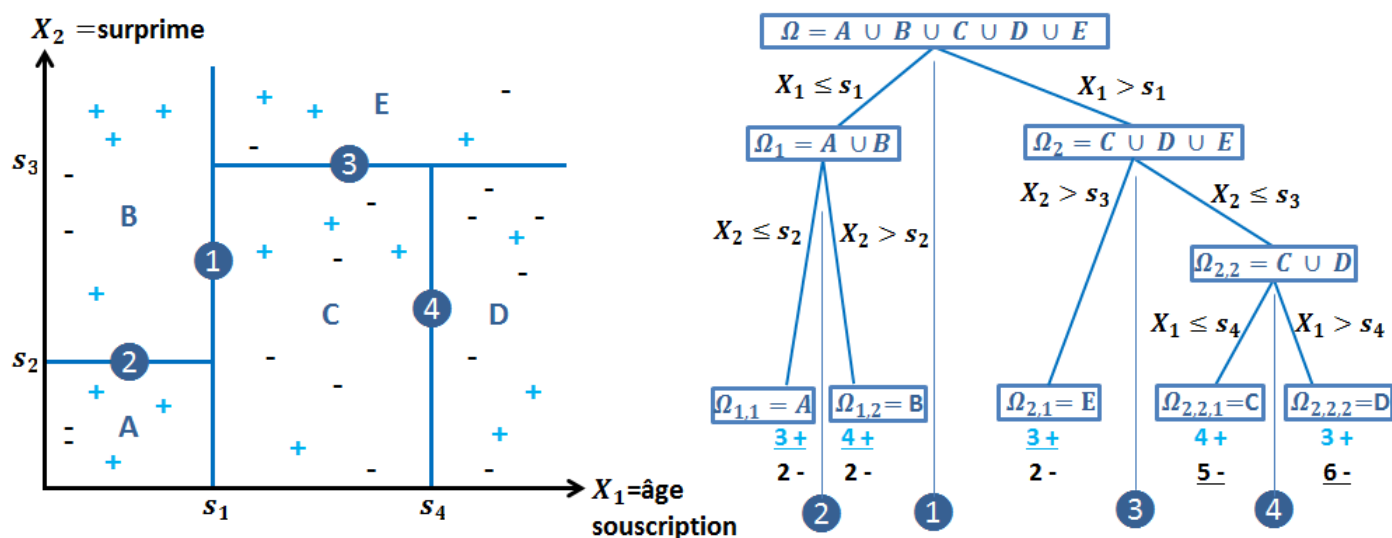


FIGURE III.1 – Exemple fictif illustrant la vision duale du principe d'un arbre CART de segmentation en deux groupes (« + » et « - ») à partir de deux variables explicatives « X_1 » et « X_2 ».

1. package « rpart ».

Il est constitué d'une série de nœuds qui divisent récursivement l'ensemble des assurés (au départ la population totale de l'échantillon d'apprentissage) en deux sous groupes². Chaque division intervient en choisissant la variable qui sépare le mieux (en termes d'homogénéité des assurés eu égard à la variable cible) le groupe, aussi la succession de nœuds induit la prise en compte des interactions entre les variables comme illustré sur la figure III.1 précédente.

Ainsi, on constate que les cloisons numérotées de ① à ④ entre les assurés du schéma de gauche de la figure III.1 définissent tour à tour des bipartitions des valeurs des variables X_j , de part et d'autre des seuils s_i , et correspondent, selon la même numérotation, à une bifurcation sur la structure arborescente à droite de la figure III.1. La récursivité de la dichotomie implique l'inclusion des « boîtes » délimitant les regroupements d'assurés : les assurés dans les cases adjacentes (C et D) délimitées par la cloison ④ sont inclus dans l'ensemble des assurés (E, C et D) délimités par la cloison ③. Sur le schéma de droite, cela se traduit par le fait que les nœuds terminaux (dits feuilles) $\Omega_{2,2,1} = C$ et $\Omega_{2,2,2} = D$ sont inclus dans l'arbre dont la racine est le nœud $\Omega_2 = C \cup D \cup E$. De ce fait, le regroupement des cases d'assurés C et D se traduit sur la figure de gauche par la suppression de la cloison ④, et sur la figure de droite par la section des branches partant du nœuds $\Omega_{2,2} = C \cup D$ vers les nœuds C et D : on parle alors d'**élagage de l'arbre**.

L'arbre illustratif construit à la figure III.1 a donc $\hat{\phi}(X_1, X_2) = \mathbb{1}_{X_1 > s_1, X_2 > s_3} + \mathbb{1}_{X_1 \leq s_1}$ pour expression mathématique, qui est donnée par les classes majoritaires de chaque feuille, et qui vaut 1 quand l'assuré est sinistré, et 0 sinon. Cette expression va donc être figée, et va donner, in fine, lors de la phase de prédiction, la valeur à prédire (sinistré ou non) pour un assuré dont on connaît les caractéristiques X_1 et X_2 .

Aussi, on voit bien sur cet exemple la nécessité de définir un critère d'arrêt de la procédure ainsi qu'un critère d'évaluation de l'**hétérogénéité** (relativement à la répartition des sinistrés « + » et de leur complémentaires « - ») des nœuds, sans quoi beaucoup de cloisons seront établies afin de constituer des feuilles (nœuds terminaux) parfaitement homogènes, isolant les sinistrés des non sinistrés, ce qui constituerait du sur-apprentissage, et se ferait au détriment de la capacité de généralisation du modèle. L'objectif de cette étude étant certes la caractérisation des sinistrés, il est néanmoins primordial que le modèle retenu ne se contente pas de révéler une vérité propre au portefeuille étudié mais puisse se généraliser au portefeuille futur dont la composition est susceptible de changer.

Cependant, l'objectif de l'assureur à travers la mutualisation des risques et la lutte contre l'antisélection, et de faire en sorte que la proportion de non sinistrés soit plus grande que celle des sinistrés. Ces derniers étant relativement peu nombreux, ils deviennent donc difficiles à caractériser de manière robuste avec peu de cloisons s'ils sont éparpillés dans l'espace des variables utilisées (X_1 et X_2 du schéma de gauche de la figure III.1).

2. Une division correspond donc à une valeur séparatrice pour une variable quantitative, ou, dans le cas qualitatif, à un fractionnement en deux parties de ses modalités.

1.1 - Critère d'impureté

De ce fait, l'importance du choix d'un critère qui permet de faire en sorte qu'il y ait dans chaque nœud à constituer une majorité (de sinistrés ou de non sinistrés) la plus forte et la plus marquée possible se fait ressentir. C'est ainsi qu'intervient le critère d'impureté d'un nœud qui est d'autant plus faible qu'une majorité forte se dégage de celui-ci. Ainsi, à chaque itération de la procédure, la constitution de deux nouveaux nœuds par cloisonnement d'un nœud parent se fait en choisissant une variable X_j et un seuil s_i qui vont, parmi tous les choix possibles, permettre un plus grand gain de pureté d'un nœud parent aux nœuds fils³, ou plus exactement, une **plus grande baisse de l'impureté**, cette baisse étant définie dans le cas d'un arbre binaire par la quantité

$$\text{impureté}(\text{nœud parent}) - \sum_{i=1}^2 \frac{\text{effectif}_{i^{\text{ème}} \text{ nœud fils}}}{\text{effectif}_{\text{nœud parent}}} \times \text{impureté}(i^{\text{ème}} \text{ nœud fils})$$

Aussi, l'utilité de la recherche d'une majorité la plus écrasante possible au vote majoritaire dans chaque nœud se comprend aisément dès lors que l'on considère par exemple (en gardant les notations de la figure III.1) le rapport $\frac{\text{effectif}_+}{\text{effectif}_{\text{nœud } \Omega_{2,2}}}$ comme étant une bonne estimation de la probabilité conditionnelle $\mathbb{P}[Y = + | \text{nœud } \Omega_{2,2}] = \mathbb{P}[Y = + | X_1 > s_1, X_2 \leq s_3]$ pour un assuré d'être sinistré conditionnellement au fait que son âge à la souscription (dans notre exemple X_1) et le montant de surprime qui lui est appliquée (dans notre exemple X_2) satisfont respectivement les conditions de seuil s_1 et de plafond s_3 menant au nœud $\Omega_{2,2}$. En effet, une grande majorité permet d'avoir une estimation proche de 0 ou de 1.

Reste à présent à instancier l'« impureté » en lui associant l'expression explicite d'une fonction qui à un nœud lui associe un score normalisé à 1. Pour cela, nous pouvons utiliser l'Entropie, empruntée à la théorie de l'information et reposant sur l'idée que la symétrie des effectifs dans chaque nœud apporte une grande incertitude. En effet, dans notre cas précis, une équirépartition des sinistrés et des non sinistrés dans un nœud ne permet pas d'informer l'assureur sur les chances qu'a un assuré satisfaisant les conditions sur les variables explicatives menant au nœud en question d'être sinistré ou non. L'entropie non normalisée, définie par l'expression

$$\begin{aligned} \text{Entropie}(\text{nœud}_j) &= - \sum_{c \in \{-, +\}} \frac{\overbrace{\text{effectif}_{\text{classe } c \text{ nœud}_j}^{p_c}}}{\text{effectif}_{\text{nœud}_j}} \times \log \left(\frac{\overbrace{\text{effectif}_{\text{classe } c \text{ nœud}_j}^{p_c}}}{\text{effectif}_{\text{nœud}_j}} \right) \\ &= \underbrace{-p_+ \times \log(p_+)}_{E_+} - \underbrace{(1 - p_+) \times \log(1 - p_+)}_{E_-} \end{aligned}$$

peut se lire comme la moyenne, pondérée par les proportions p_c , des incertitudes marginales $-\log(p_c)$ (positives car $p_c \in [0; 1]$, et décroissantes à mesure que les proportions p_c augmentent) apportées par chacune des classes $c \in \{-, +\}$ à l'incertitude globale d'un nœud. Les deux cas

3. On essaie de rendre maximale l'homogénéité à l'intérieure de chacun des deux nœuds fils (intra-nœuds) ou maximale l'hétérogénéité entre les deux nœuds fils (inter-nœuds).

limite $p_+ = 0$ et $p_+ = 1$ annulent l'entropie ($p \log(p) \xrightarrow{p \rightarrow 0^+} 0$ d'où le prolongement par continuité), et la position intermédiaire est réalisée lorsque $p_+ = p_- = \frac{1}{2}$ ce qui entraîne une entropie maximale égale à $-2 \times \frac{1}{2} \times \log \frac{1}{2} = \log 2 (> 0)$. L'**entropie normalisée** par $\log 2$ est donc une fonction concave $p_+ \longrightarrow \{-p_+ \times \log(p_+) - (1 - p_+) \times \log(1 - p_+)\} / \log(2)$, dont la représentation graphique est une courbe en cloche plafonnée par la valeur 1, atteinte pour $p_+ = \frac{1}{2}$. On cherchera donc, si l'on souhaite constituer des nœuds différenciant le plus possible les sinistrés des non sinistrés, à minimiser l'entropie en se rapprochant « le plus vite possible » des valeurs $p_+ = 0$ ou $p_+ = 1$ (de manière équivalente $p_- = 1$ ou $p_- = 0$) où elle est la plus faible.

Alternativement, une fonction d'impureté analogue est l'indice de Gini associant à chaque nœud, dans le cadre de notre classification binaire, à la quantité $2 \times p_+ \times p_- = 2 \times p_+ \times (1 - p_+)$ qui s'annule, à l'instar de l'entropie, lorsque $p_+ = 1$ ou $p_+ = 1 - p_- = 0$, mais est majorée cette fois-ci par $\frac{1}{2}$ lorsque $p_+ = p_- = \frac{1}{2}$ (coefficient de normalisation).

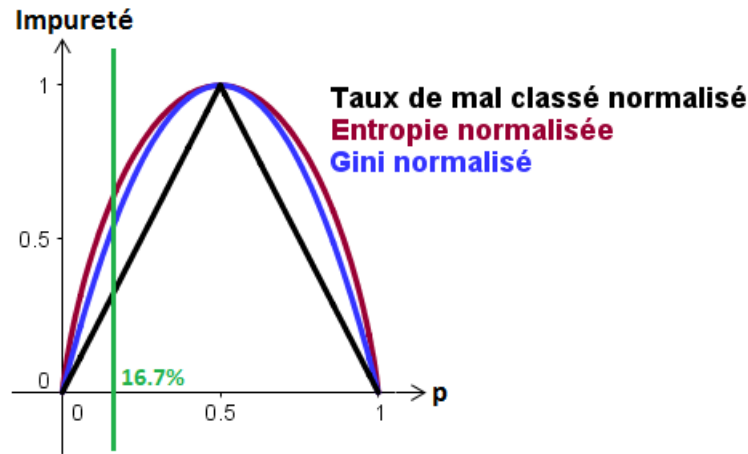


FIGURE III.2 – Entropie, Gini et taux de mal classés normalisés dans le cas d'une classification binaire où p représente la proportion d'une des deux classes d'un nœud donné.

Les flancs des deux courbes en cloche de la figure III.2 permettent une décroissance (en termes de dérivée) plus rapide de l'impureté quand on se rapproche des cas limites recherchés ($p_+ = 0$ ou $p_+ = 1$) comparativement au cas où l'on prend comme critère d'impureté le taux de mal classé (ou erreur d'apprentissage) qui consiste à dénombrer pour un nœud le nombre d'assurés dont l'état (sinistré ou non) diffère de la classe de la majorité des assurés de ce même nœud. Le taux de mal classés est donc égal à la proportion de la classe minoritaire du nœud considéré. Or, dans la mesure où nous cherchons le critère qui engendre la plus grande vitesse de décroissance de l'impureté d'un nœud à l'autre, avec comme répartition initiale de sinistrés $p_+ = 16,7\%$, le critère de Gini ou l'entropie apparaissent comme les plus adaptés.

De plus, puisqu'il peut être intéressant d'isoler sans trop de cloisons les sinistrés, qui sont toutefois en plus faible proportion que les non sinistrés, l'entropie est, parmi les trois fonctions d'impureté, celle qui peut permettre de constituer plus rapidement des nœuds à faibles effectifs (d'où l'utilité de bien choisir le critère d'arrêt « minbucket » évoqué au paragraphe suivant) et de très faible impureté puisque la dérivée de l'entropie est plus pentue que celle, par exemple,

du taux de mal classés au voisinage de $p = 0$ ou $p = 1$.

1.2 - Critères d'arrêts

Par ailleurs, le deuxième point important lors de la construction d'un arbre CART, outre la fonction d'impureté, est la calibration des deux principaux paramètres⁴ de la procédure CART appelés critères d'arrêts, à savoir le nombre minimal d'assurés par nœud terminal (« minbucket »), ainsi que γ la valeur minimale de décroissance du risque empirique R_n^{appr} .

Pour comprendre γ , il est utile d'introduire la notion de coût global C d'un arbre $\hat{\phi}$. L'idée du coût est tout d'abord de dire qu'un arbre volumineux (de grande taille), avec beaucoup de feuilles, est un arbre coûteux en temps de calcul (**coût temporel**) et coûteux du point de vue de l'intégration opérationnelle des résultats puisqu'une **tarification** selon plusieurs variables significatives et selon plusieurs conditions est lourde à mettre en place. Ensuite, le taux de mal-classés, i.e. l'erreur d'apprentissage, qui n'est autre que le risque empirique $R_n^{appr}(\hat{\phi})$ (égal à $\frac{1}{n_{appr}} \sum_{i=1}^{n_{appr}} \mathbb{1}_{\hat{\phi}(x_i) \neq y_i}$ dans notre cas), est un indicateur du pouvoir explicatif, et à ce titre est la deuxième composante du coût global d'un arbre, qui est donc défini comme suit :

$$C(\hat{\phi}) = R_n^{appr}(\hat{\phi}) + \gamma \times \mathcal{T}(\hat{\phi})$$

où $\mathcal{T}(\hat{\phi})$ est la taille de l'arbre, c'est-à-dire le nombre de nœuds terminaux, et γ est un facteur multiplicatif de cette taille. Ainsi, lors de l'expansion de l'arbre, passant d'un arbre de taille \mathcal{T} vers un arbre de taille $\mathcal{T}' = \mathcal{T} + 1$ (une étape de la récursion divisant un nœud en deux), la différence de coût $\Delta C_{\mathcal{T} \rightarrow \mathcal{T}'} := C_{\mathcal{T}} - C_{\mathcal{T}'} = (R_n^{appr|\mathcal{T}} + \gamma \times \mathcal{T}) - (R_n^{appr|\mathcal{T}'} + \gamma \times \mathcal{T}')$ est donc égale à $(R_n^{appr|\mathcal{T}} - R_n^{appr|\mathcal{T}'}) + \gamma \times (\mathcal{T} - (\mathcal{T} + 1))$, autrement dit après simplification $\Delta C_{\mathcal{T} \rightarrow \mathcal{T}'} = \Delta R_n^{appr|\mathcal{T} \rightarrow \mathcal{T}'} - \gamma$.

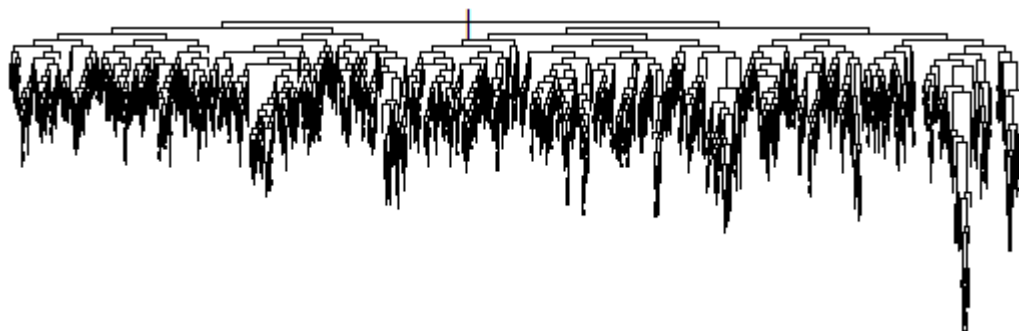
Par conséquent, lors de la quête de l'arbre optimal, il n'est pas illogique de chercher la décroissance du coût global à chaque étape de la récursion (i.e. $C_{\mathcal{T}} > C_{\mathcal{T}'=\mathcal{T}+1}$), autrement dit $\Delta C_{\mathcal{T} \rightarrow \mathcal{T}'} > 0$, ce qui est équivalent à vérifier la condition $\Delta R_n^{appr|\mathcal{T} \rightarrow \mathcal{T}'} > \gamma$. Dans la mesure où le risque empirique R_n^{appr} , en plus d'être une composante du coût global C , peut être utilisé comme critère de cloisonnement d'un nœud⁵, il est, à ce titre, décroissant au global des feuilles d'une étape de la récursion à l'autre. La décroissance de R_n^{appr} entraîne que $\Delta R_n^{appr|\mathcal{T} \rightarrow \mathcal{T}'} \geq 0$ à chaque itération, d'où le fait que la racine de l'arbre (i.e. l'échantillon d'apprentissage entier) présente par essence et par construction un risque empirique plus élevé que le risque empirique au global des feuilles.

4. Les autres paramètres comme « minsplit », « maxsurrogate » et « maxdepth » (respectivement le nombre minimal d'assurés par nœud intermédiaire, le nombre de variable de substitution à utiliser pour classer des assurés présentant des valeurs manquantes, et la profondeur maximal de l'arbre) ne sont pas abordés ici.

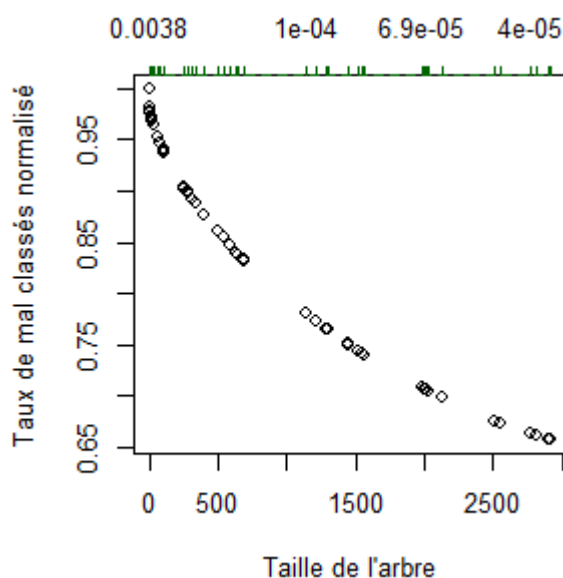
5. C'est le taux de mal classés de la figure III.2 page 46 donnant les 3 fonctions d'impureté généralement utilisées.

Ainsi, γ est une sorte de paramètre de contrôle de la taille maximale de l'arbre à partir duquel on souhaite élaguer par validation croisée (pour l'optimiser d'avantage ; cf. infra). Ce critère d'arrêt, s'il est strictement supérieur à 0, réalise donc un compromis entre taille raisonnable (synonyme de simplicité) propice à une utilisation opérationnelle en assurance et bon pouvoir explicatif de l'arbre (grande taille), puisque sa construction s'arrête dès que $\Delta R_n^{appr}|\mathcal{T} \rightarrow \mathcal{T}' \leq \gamma$.

Nous commençons donc par fixer « minbucket » à 1 et γ à 0 pour se défaire de toute contrainte, et obtenir l'arbre complet (unique pour un échantillon d'apprentissage fixé) suivant dont les feuilles sont les plus pures possible :

FIGURE III.3 – arbre complet ($\gamma=0$)

Il contient 2917 nœuds (soit 2916 scissions), son erreur réelle non normalisée à la racine est de 16,7% (égale à la proportion de sinistrés du portefeuille) et celle au global des feuilles est 10,95%⁶. La décroissance de l'erreur d'apprentissage (ou encore risque empirique R_n^{appr}) normalisée à un 1 (le risque empirique de 16,7% à la racine étant le majorant, c'est lui qui fait office de facteur de normalisation) au fil des scissions présente le profil suivant :

FIGURE III.4 – Taux de mal classés normalisés en fonction de la taille de l'arbre (abscisses inférieures) et du paramètre γ (abscisses supérieures).

6. Elle n'est pas nulle probablement à cause du manque de flexibilité imposé par la discrétisation des variables continues, et par l'absence de variables très explicatives.

1.3 - Élagage de l'arbre complet par validation croisée

Avant d'aborder l'élagage par validation croisée, attardons nous sur l'expression du coût global C d'un arbre $\hat{\phi}$, et remplaçons γ par C_p :

$$C(\hat{\phi}) = R_n^{appr}(\hat{\phi}) + C_p \times \mathcal{T}(\hat{\phi})$$

Si l'on se place au niveau de l'arbre complet où comme nous l'avons vu le risque empirique R_n^{appr} est minimal et la taille \mathcal{T}_{max} maximale, la valeur de C_p qui minimise le coût global C est 0. Si l'on augmente la valeur du coefficient de pénalisation C_p , comme la taille de l'arbre représente un coût, alors l'arbre qui minimise le coût global C a donc une taille plus petite que \mathcal{T}_{max} . Ainsi, on se rend compte qu'en augmentant le C_p , on réduit la taille de l'arbre. De là, on comprend qu'il est possible d'associer à chaque arbre issu d'une étape de la procédure récursive de construction de l'arbre de taille \mathcal{T}_{max} , une valeur de C_p qui va en décroissant à partir de la racine. On retrouve cette idée dans le fait qu'en fixant le critère d'arrêt γ (apparaissant dans l'expression $C(\hat{\phi}) = R_n^{appr}(\hat{\phi}) + \gamma \times \mathcal{T}(\hat{\phi})$ de la section précédente page 47) à 0, on obtient un arbre le plus feuillu possible $\phi^{\mathcal{T}_{max}}$ (de taille \mathcal{T}_{max}), et qu'en fixant γ strictement supérieur à 0, la procédure descendante de cloisonnement de l'échantillon d'apprentissage va s'arrêter en amont de $\phi^{\mathcal{T}_{max}}$ dès que la décroissance de R_n^{appr} est plus faible que γ , permettant ainsi d'obtenir un arbre de taille intermédiaire $\mathcal{T} < \mathcal{T}_{max}$.

Ainsi, chaque C_p associé à un sous-arbre de l'arbre complet $\phi^{\mathcal{T}_{max}}$ peut être vu comme un palier lorsque γ parcourt l'ensemble des réels positifs en décroissant. Autrement dit, l'arbre obtenu en spécifiant γ dans $\{0\}$ ou dans $]C_p^{\mathcal{T}=i+1}; C_p^{\mathcal{T}=i}]$ ($\forall i \in \llbracket 1; \mathcal{T}_{max} - 1 \rrbracket$) est respectivement de taille \mathcal{T}_{max} et de taille i , et donc $C_p^{\mathcal{T}=i}$ désigne la plus grande valeur de γ qui permet d'obtenir un arbre de taille i , avec $C_p^{\mathcal{T}_{max}} = 0$. Ceci s'observe à travers la fonction en escalier de la figure suivante :

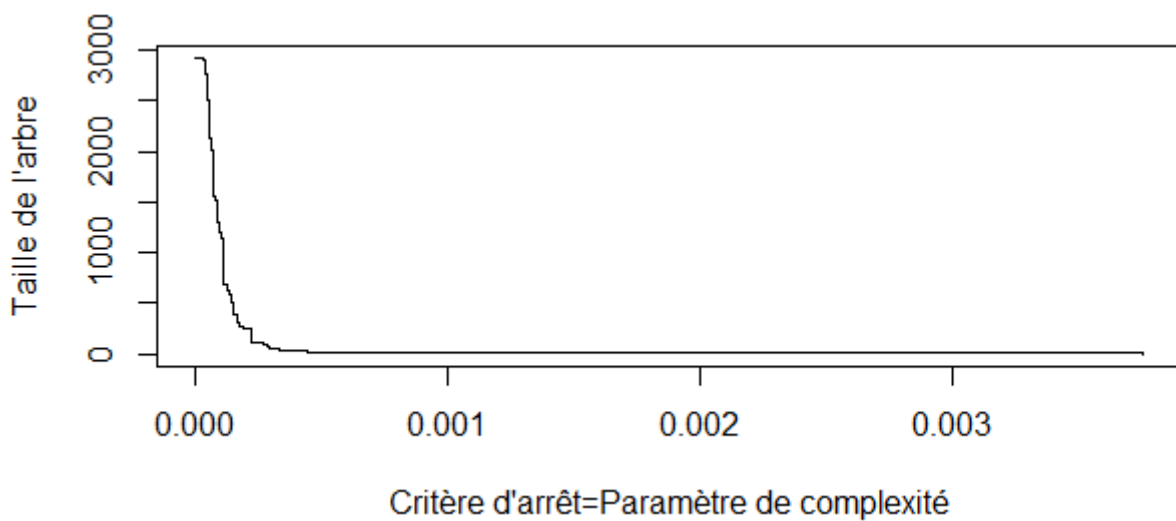


FIGURE III.5 – Taille de l'arbre en fonction du paramètre de complexité γ .

En pratique, nous tenterons de déterminer le γ qui permet d'obtenir l'arbre qui minimise, parmi tous les arbres allant de la racine à l'arbre de taille \mathcal{T}_{max} , non pas le coût global,

mais le risque empirique R_n^{appr} . Or, il est évident que par construction c'est l'arbre de taille \mathcal{T}_{max} ($\gamma = 0$). Ainsi, pour obtenir un arbre plus robuste (efficient sur des données différentes de l'échantillon d'apprentissage), nous allons plutôt choisir le γ qui va **minimiser** R_n^{appr} par « **k-fold cross-validation** », avec $k=10$. Autrement dit, on va segmenter l'échantillon d'apprentissage initial en $k = 10$ sous-échantillons disjoints et de même taille, et, chacun des k sous-échantillons va servir à tour de rôle d'échantillon de validation et les $k - 1$ autres échantillons font office d'échantillon d'apprentissage.

Or, le problème est qu'à chaque fois que l'on va estimer un arbre, pour une valeur fixée du paramètre γ , sur une partie différente de l'échantillon (fraction $\frac{k-1}{k}$ de l'échantillon à chaque fois), on risque d'obtenir un arbre différent d'une itération $j \in \llbracket 1; k = 10 \rrbracket$ à l'autre de la validation croisée, et donc un nombre de feuilles différent. Par conséquent, ce que l'on optimise par validation croisée n'est pas le nombre de feuilles \mathcal{T} mais bien le critère de pénalisation γ .

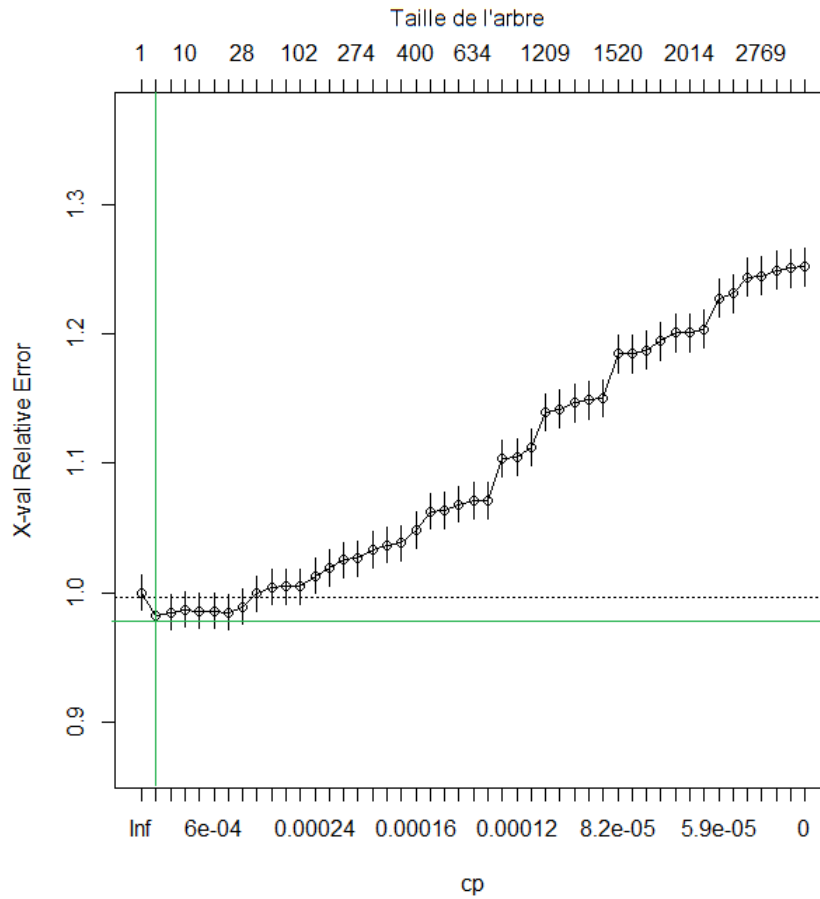


FIGURE III.6 – Taux de mal classés par validation croisée (normalisés par la valeur à la racine) en fonction du couple $(\mathcal{T}_i \mid C_p^{\mathcal{T}_i})$.

On obtient ainsi, pour plusieurs valeurs du critère de pénalisation γ (ou de manière équivalente C_p) une suite de taux moyens de mauvaises prédictions⁷ désignés par « X-val Relative Error » en ordonnées sur la figure III.6 précédente. Les abscisses supérieures de cette figure III.6 désignent la taille \mathcal{T}_i qu'aurait l'arbre entier $\phi^{\mathcal{T}_{max}}$ (estimé sur tout l'échantillon d'appren-

7. Chaque taux moyen est une moyenne de $k=10$ taux.

tissage, sans validation croisée) si on l'élaguait en choisissant le paramètre de pénalisation $C_p^{\mathcal{T}_i}$ correspondant sur l'axe des abscisses inférieures.

On constate une forte croissance de l'erreur par validation croisée à partir d'une certaine valeur du paramètre de complexité. Ceci peut s'expliquer par le fait que lors de la validation croisée, pour une valeur donnée de γ et pour k sous-échantillons d'apprentissage donnés, on peut avoir k arbres assez différents. Ainsi, lorsque γ est très faible, les arbres étant très sensibles et instables (sur-apprentissage), peuvent donc être très différents d'une itération à l'autre de la validation croisée. Là encore, c'est le pouvoir de généralisation qui diminue quand la taille de l'arbre augmente.

Une fois le $\gamma = C_p^{\text{opt}} = 0.001655995$ optimal calibré⁸, et donc une taille d'arbre fixée ($\mathcal{T} = 6$ feuilles), on élague les branches superflues de l'arbre $\phi^{\mathcal{T}_{\max}}$, en se ramenant à l'arbre associé au C_p optimal C_p^{opt} . L'arbre obtenu est le suivant :

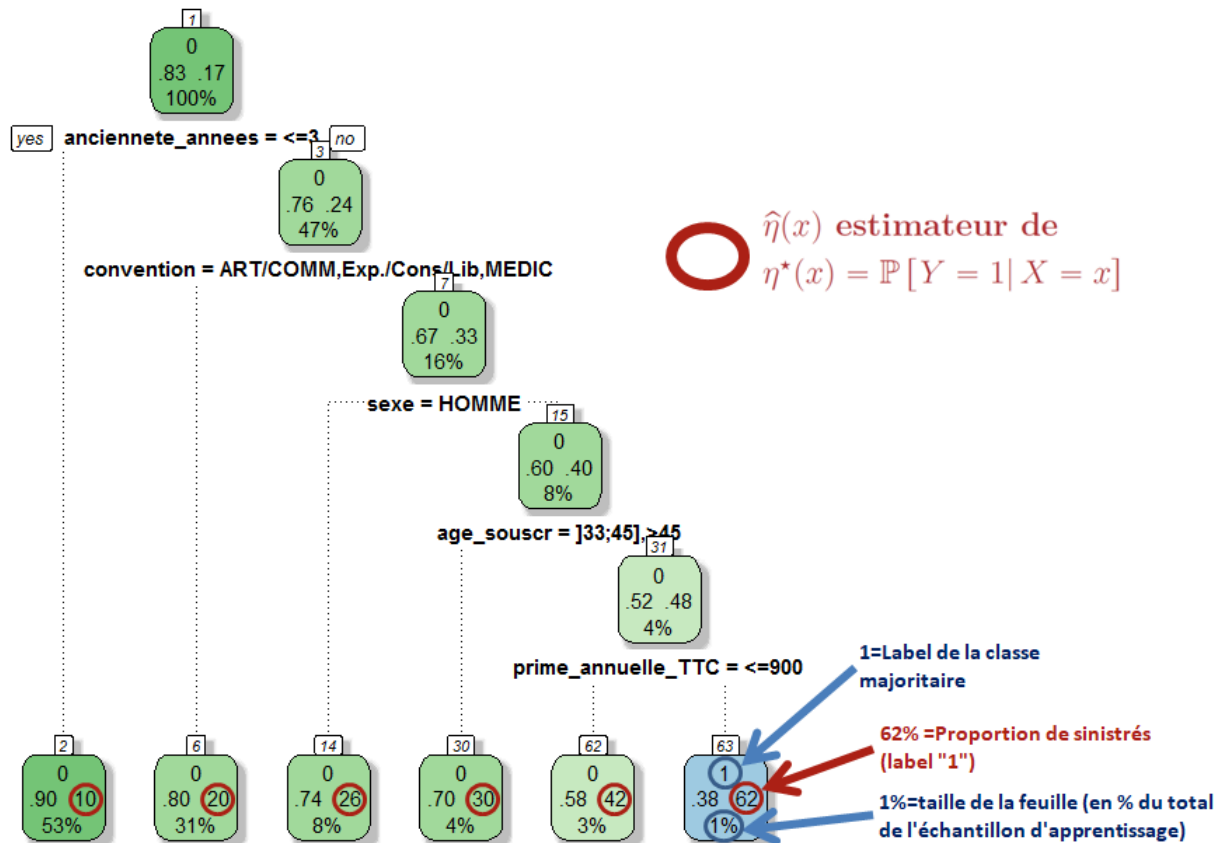


FIGURE III.7 – Arbre optimal pour $\gamma = C_p^{\text{opt}} = 0.001655995$

Mais cet arbre élagué, qui, rappelons-le, dépend de l'échantillon d'apprentissage utilisé, performe à peine mieux que l'échantillon d'apprentissage sans aucune modélisation (le taux de mal classés de l'arbre élagué est de 16,34%, et celui de la racine vaut 16,7%), car les sinistres, qui ne représentent que 16,7% de l'échantillon d'apprentissage, se retrouvent dilués dans les feuilles,

8. C'est la plus grande valeur de γ qui minimise le taux moyen de mauvaises prédictions par validation croisée : cf croisement des droites vertes, horizontale et verticale, sur la figure III.6.

notamment la première, qui contient 53% de l'échantillon d'apprentissage, dont 90% de non sinistrés.

Ainsi, à défaut d'améliorer considérablement le risque empirique par rapport à la racine, cet arbre élagué nous renseigne sur les interactions les plus pertinentes dans l'explication de la survenance de sinistre incapacité : un assuré qui n'est ni un artisan-commerçant, ni un expert-conseil/libéral, ni une personne du secteur médical, qui n'est pas de sexe masculin, qui a plus de 3 ans d'ancienneté en portefeuille, et ayant souscrit avant l'âge de 33 ans a plus de chances d'être sinistré qu'une autre personne n'ayant pas ces caractéristiques, avec un risque aggravé si le montant annuel moyen de prime toutes garanties confondues est supérieur à 900€⁹. A l'inverse, tout assuré ayant moins de 3 ans d'ancienneté en portefeuille, quelles que soient ses autres caractéristiques, a seulement 10% de chances d'être sinistré.

Enfin, on notera la présence de la variable sexe qu'il est, d'un point de vue réglementaire, pourtant interdit d'utiliser pour tarifier.

1.4 - Importance des variables

Par ailleurs, il est important de noter qu'à chaque étape de la récursion le cloisonnement étant fait de telle sorte à augmenter la pureté des nœuds fils, il est aisé de comprendre que la variable retenue pour ce cloisonnement est déterminante pour les assurés des nœuds fils obtenus : cela traduit l'importance d'une variable, qui, au global de l'arbre, est définie comme la somme des hausses de pureté (ou de manière équivalente à la somme des décroissances d'impureté) qu'elle a pu engendrer au fil de la construction de l'arbre, de la racine jusqu'aux feuilles.

Cependant, ne considérer l'importance d'une variable que lorsque celle-ci est retenue pour une scission peut introduire un grand biais dans le calcul de l'importance des variables puisque certaines d'entre elles peuvent induire, pour un même nœud, un partage des assurés semblable à celui engendré par la variable finalement retenue (la plus importante pour ce nœud). Ainsi, le calcul de l'importance d'une variable tient compte de cet aspect.

Aussi, il est utile de remarquer que les six variables les plus importantes pour l'arbre entier ($\gamma = 0$, figure III.3) et apparaissant sur la figure III.8 ne sont pas celles qui apparaissent dans l'arbre optimal de la figure III.7 ($\gamma = C_p^{opt} = 0.002180549$). En effet, l'arbre optimal donne une lecture graphique simplifiée puisqu'il ne nous donne que les variables les plus importantes pour les premières scissions. La figure III.8 nous donne donc une vision complémentaire indispensable pour apprécier l'importance des variables calculée tout le long de la construction de l'arbre

9. $\hat{\eta}(x)$, l'estimateur de $\eta^*(x) = \mathbb{P}[Y = 1 | \mathbb{X} = x]$, vaut 62% > 50% pour la feuille tout à droite sur la figure III.7. L'estimateur plug-in vaut donc 1 pour cette feuille, ce qui correspond au label des sinistrés.

complet.

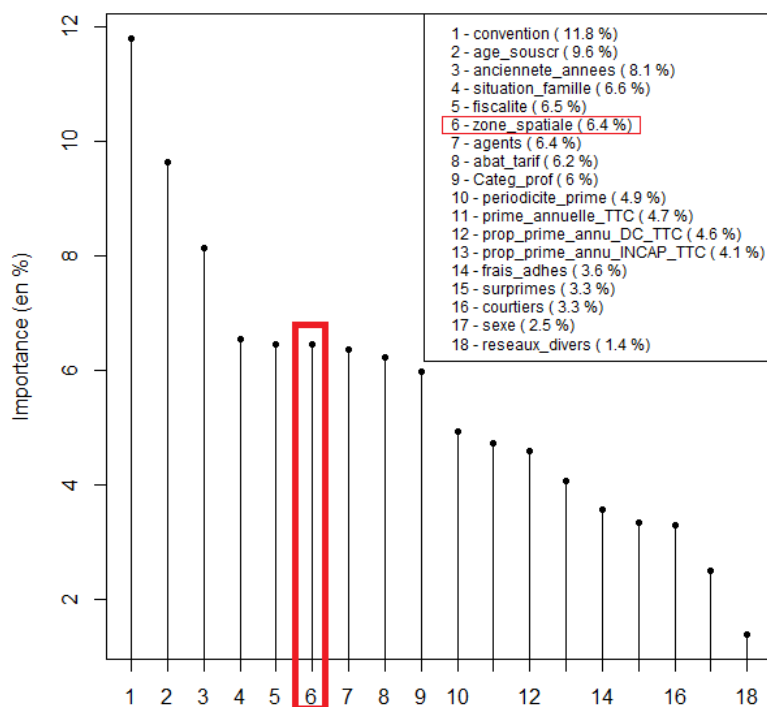


FIGURE III.8 – Importance, pour l'arbre entier ($\gamma = 0$), des variables (en pourcentage)

Nous y constatons que l'ancienneté en portefeuille est la troisième variable la plus importante, derrière la convention et l'âge à la souscription. Viennent ensuite la situation familiale, la fiscalité et la zone spatiale. L'importance d'une variable est donc une manière innovante de hiérarchiser les variables les plus explicatives de l'occurrence d'au moins un sinistre incapacité.

2 - Agrégation d'arbres CART

2.1 - Motivations

Nous avons vu que les arbres CART dessinent des pavés dans l'espace des variables explicatives lors de l'apprentissage, et la classe majoritaire des observations d'un même pavé est celle qui est retenue lors de la prédiction. Aussi, si l'on change l'échantillon d'apprentissage alors ces pavés, et donc les **prévisions** pour un même assuré, peuvent **changer radicalement** d'un arbre à l'autre. Une solution potentielle à cette sensibilité à l'échantillon d'apprentissage est l'agrégation d'arbres CART.

Faire de l'agrégation de modèles consiste à mélanger les prévisions de plusieurs modèles dérivant d'une même famille de modèles, dans notre cas les arbres CART, pour se ramener à un seul modèle résumant le tout, et cela afin d'améliorer la prévision. Ainsi, plutôt que d'estimer un seul arbre CART, nous allons en estimer quelques dizaines voire quelques centaines. Par conséquent, la **complexité** des méthodes peut rapidement augmenter, mais **contre-intuitivement**, cela a tendance à ne pas faire apparaître de phénomène de sur-ajustement, bien au contraire,

cela a plutôt tendance à nous en préserver.

Pour ce faire, nous allons adopter les deux façons de modéliser qui existent : le **Bagging** (contraction de **B**ootstrap **A**GGregat**I**NG) et le **Boosting**. Ces deux méthodes diffèrent dans leurs approches respectives puisque, à échantillon d'apprentissage fixé, la première comporte un **aléa** alors que la deuxième est **déterministe**. En pratique, le Bagging autant que le Boosting n'ont de réelle utilité que sur des familles de modèles qui présentent une grande sensibilité à l'échantillon d'apprentissage comme les arbres CART, ce qui est encore une fois **paradoxal** et contraire au bon sens, puisqu'une **combinaison de modèles instables**¹⁰ donnent un **modèle plus stable**.

2.2 - Bagging d'arbres, et forêts aléatoires

Le principe de l'aggregation d'arbres par bootstrap¹¹, dit Bagging d'arbres, est assez élémentaire puisque l'on va constituer B sous-échantillons de même taille \tilde{n} construits aléatoirement avec remise¹² parmi l'échantillon d'apprentissage. Ensuite, sur chacun des B échantillons bootstrap, on va estimer un arbre CART, et la prévision finalement retenue $\hat{\phi}_i$ de la variable réponse Y est pour chaque assuré $i \in \llbracket 1 ; n \rrbracket$:

- la moyenne des B prédictions $\left(\hat{\phi}_i^b\right)_{b \in \llbracket 1 ; B \rrbracket}$ données par chacun des B arbres si Y est quantitative : $\hat{\phi}_i = \frac{1}{B} \sum_{b=1}^B \hat{\phi}_i^b$
- la modalité la plus fréquente à l'issue des prévisions de chacun des B modèles CART si Y est qualitative : dans notre cas binaire $\hat{\phi}_i = \mathbb{1}_{\frac{1}{B} \sum_{b=1}^B \hat{\phi}_i^b > 0.5}$

Une telle manière de procéder par ré-échantillonnage permet de réduire la variance (comme nous le verrons dans la section suivante page 55) et donc d'améliorer le pouvoir prédictif et de généralisation. Par ailleurs, bien que l'échantillonnage bootstrap entraîne une hausse du temps de calcul à mesure que l'on augmente le paramètre B égal au nombre d'échantillons bootstrap, ceux-ci étant constitués de manière aléatoire et indépendante, il est possible de paralléliser cette procédure et de gagner un temps substantiel de calcul.

2.2.1 - Choix du nombre de variables

Pour un même individu i , la variance associée à la moyenne $\frac{1}{B} \sum_{b=1}^B \hat{\phi}_i^b$ des B prévisions $\left(\hat{\phi}_i^b\right)_{b \in \llbracket 1 ; B \rrbracket}$ de même distribution¹³ et supposées corrélées par paire avec un même coefficient $\rho = \frac{\text{Cov}(\hat{\phi}_i^b, \hat{\phi}_i^{b'})}{\sqrt{\text{Var}(\hat{\phi}_i^b)} \times \sqrt{\text{Var}(\hat{\phi}_i^{b'})}} = \frac{\text{Cov}(\hat{\phi}_i^b, \hat{\phi}_i^{b'})}{\sigma^2} \in [0 ; 1]$ (où σ^2 est la variance commune), dépend beaucoup de ce coefficient de corrélation :

10. Instable=donne des modèles (et donc des prévisions) différents d'un échantillon d'apprentissage à l'autre.

11. Bootstrap=échantillonnage aléatoire avec remise=ré-échantillonnage.

12. Certains assurés peuvent donc faire partie de plusieurs échantillons, d'autres ne feront partie d'aucun.

13. Une distribution identique implique que leur coefficient de corrélation ρ est dans $[0 ; 1]$ et non dans $[-1 ; 1]$, puisque les écarts entre les deux variables et leur moyenne commune ont « tendance » (en espérance) à être de même signe, leur covariance est donc supérieure à 0.

$$\begin{aligned}
 Var \left(\frac{1}{B} \sum_{b=1}^B \hat{\phi}_i^b \right) &= \frac{1}{B^2} \times Var \left(\sum_{b=1}^B \hat{\phi}_i^b \right) \\
 &= \frac{1}{B^2} \times \left(\underbrace{\sum_{b=1}^B Var(\hat{\phi}_i^b)}_{\sigma^2} + \underbrace{\sum_{(b,b') \in \{1, \dots, B\}^2, b \neq b'} Cov(\hat{\phi}_i^b, \hat{\phi}_i^{b'})}_{\rho \times \sigma^2} \right) \\
 &\quad \underbrace{B \times (B-1) \text{ termes}}_{B \times (B-1) \text{ termes}} \\
 &= \frac{B \times \sigma^2 + B \times (B-1) \times \rho \times \sigma^2}{B^2} \\
 &= \frac{\cancel{B} \times \sigma^2 \times [B \times \rho + (1 - \rho)]}{\cancel{B}^2} \\
 &= \sigma^2 \times \rho + \frac{(1 - \rho) \times \sigma^2}{B}
 \end{aligned}$$

Ainsi, le premier constat est que $Var \left(\frac{1}{B} \sum_{b=1}^B \hat{\phi}_i^b \right)$ est une fonction croissante de ρ (de dérivée $\sigma^2 \times \frac{B-1}{B} > 0$) sur $[0; 1]$, de minimum $\frac{\sigma^2}{B} < \sigma^2$ (atteint pour $\rho = 0$) et de maximum σ^2 (atteint pour $\rho = 1$). Le bootstrap permet donc de faire diminuer la variance, qui est au plus égale à la variance commune des arbres (sans bootstrap).

En outre, plus les B modèles élémentaires constituant l'agrégation sont corrélés entre eux (ρ proche de 1) **par bootstrap**, plus les B prévisions sont corrélées entre elles, et moins l'agrégation par moyenne des prévisions entraînera une réduction de variance de la prévision, d'après le premier terme $\sigma^2 \times \rho$ de la formule ci-dessus.

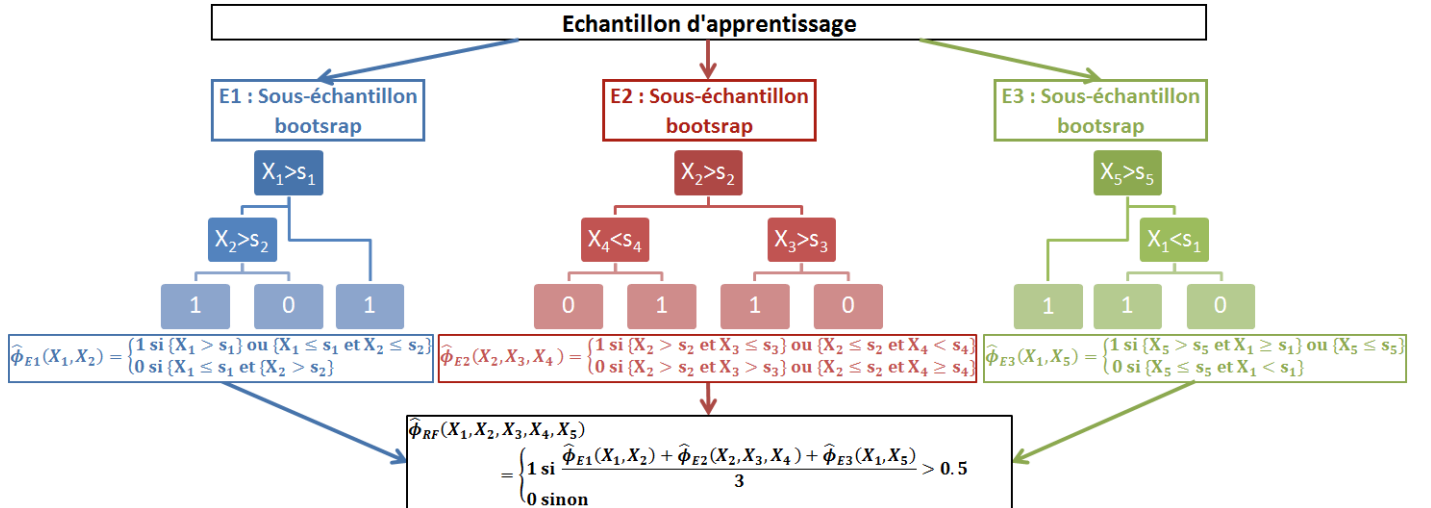


FIGURE III.9 – Illustration de la construction du prédicteur par forêts aléatoires $\hat{\phi}_{RF}(\mathbb{X})$ à partir de $\mathbb{X} = (X_i)_{i \in [1; 5]}$.

Par conséquent, l'aspect bootstrap étant très utile dans le cas de bases de données non volumineuses (car dans ce cas, constituer B échantillons indépendants, avec B grand, contraint la taille de chaque échantillon et donc diminue la qualité d'apprentissage), le concepteur des arbres CART a imaginé l'ajout d'un **aléa** dans le choix des variables lors de la construction

de chaque arbre afin de générer B prévisions **les moins semblables, les plus décorréliées**. Aussi, le choix des nœuds de chaque arbre est sous-optimal (par rapport à l'arbre complet de la partie III figure III.3) car construit à partir d'une sélection de $m < p$ variables tirées aléatoirement parmi les p variables totales comme illustré sur le schéma III.9 ci-dessus.

Le nombre de variables m est donc un des paramètres à optimiser : si $m < p$ on parle de **forêts aléatoires**, et si $m = p$, on retombe sur l'approche Bagging classique. Au vu de la figure suivante III.10 donnant le taux de bons classements en fonction du paramètre m , nous retiendrons $m = 2$ pour lequel le taux de bon classement vaut 0,8334498 :

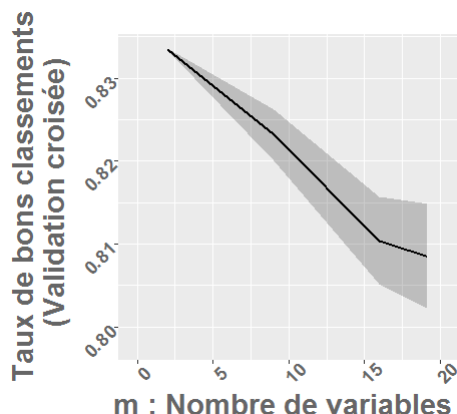


FIGURE III.10 – Forêts aléatoires : Taux de bons classements ($= 1 - R_n^{appr}$) par validation croisée, en fonction du paramètre m , pour un nombre d'itérations=500 (500 arbres construits).

2.2.2 - Élagage des arbres

Ici, contrairement aux arbres CART simples, nous n'élaguerons pas les arbres de l'agrégation par validation croisée comme nous l'avons fait auparavant pour réduire le sur-apprentissage, ce qui nous permet au passage un gain de temps de calcul. En effet, le phénomène de sur-apprentissage, très présent dans les arbres isolés lorsqu'ils sont complets (i.e. $\gamma = 0$: cf. page 48) et qui nous amène à les élaguer, est dans le cas du Bagging estompé par l'échantillonnage bootstrap, par l'aléa dans le choix des variables introduit par les forêts aléatoires, et par conséquent par la multitude d'arbres non corrélés entre eux.

Aussi, chaque arbre complet construit étant de très faible biais mais de grande variance, nous comptons sur la moyenne des prévisions de chaque arbre pour réduire la variance du risque, et non plus sur l'élagage. La **perte d'interprétabilité** due à la complexité de cette approche n'est pas complète puisqu'il nous est toujours possible de déterminer l'importance de chaque variable comme nous le verrons en page 57.

2.2.3 - Erreur out-of-bag

Lors de la procédure d'agrégation par Bagging, on peut, afin de contrôler la qualité de l'agrégation, déterminer un proxy dit « out-of-bag » de l'erreur de prévision. En effet, les B

échantillons complémentaires¹⁴ (dit « out-of-bag ») de chacun des B sous-échantillons bootstrap (dits « in-bag ») peuvent être utilisés pour estimer l'erreur de prévision en calculant la moyenne des B « erreurs out-of-bag » élémentaires. C'est donc une estimation faite sur une partie de l'échantillon d'apprentissage, elle n'est en aucun cas faite sur l'échantillon test qui servira *in fine* à comparer toutes les méthodes entre elles. Pour un paramètre $m = 2$, et un nombre d'itérations égal à 500, l'erreur out-of-bag vaut 16.59%.

Aussi, plus on augmente le nombre des modèles que l'on va agréger, plus on fait la moyenne de beaucoup de prédictions variées, et finalement moins on observe de sur-apprentissage. L'erreur out-of-bag n'augmente donc pas à mesure que l'on étoffe notre agrégation de modèles.

2.2.4 - Importance des variables

A l'instar des arbres CART, deux indices sont utilisés dans le cas du bagging pour quantifier l'importance d'une variable : Le « Mean Decrease Accuracy » (MDA) et le « Mean Decrease Gini/Entropie » (MDG/E).

Le MDG/E considère que l'importance globale d'une variable est la somme des décroissances d'impureté provoquées par cette variable pour chaque nœud de chaque arbre de l'agrégation. L'importance d'une variable est donc ici toujours basée sur l'argument qu'une variable qui est souvent choisie pour une scission ou une variable qui contribue beaucoup à la réduction de l'impureté d'un nœud par cloisonnement est une variable importante.

Quant au MDA, il s'agit d'un critère plus audacieux et plus sophistiqué que le précédent puisqu'il est égal, pour une variable, et pour un arbre, à la différence entre l'erreur out-of-bag avec et sans permutation aléatoire des valeurs de cette variable uniquement, les autres variables n'étant pas permutées.

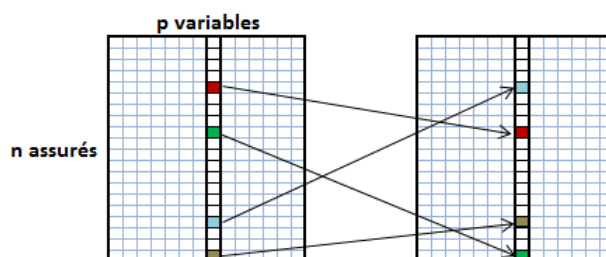


FIGURE III.11 – Principe du MDA : étude de l'influence d'une variable par permutation aléatoire de ses valeurs.

Évidemment, la qualité de l'erreur de prévision out-of-bag est fortement pénalisée suite à une telle permutation aléatoire, et ainsi, l'idée sur laquelle se fonde le mean decrease accuracy est de dire que plus une réorganisation aléatoire des valeurs d'une variable dégrade la qualité de prévision out-of-bag, plus cette variable est importante dans le modèle. Ceci étant

14. Composés des assurés qui n'ont pas été sélectionnés dans un échantillon bootstrap.

fait pour chaque variable et pour chaque arbre où la variable apparaît, le MDA d'une variable au global de la procédure de Bagging est la somme du MDA de cette variable pour chaque arbre.

Le MDG/E étant calculé à partir de la décroissance de l'impureté de chaque nœud, c'est un critère local d'importance, contrairement au MDA qui n'est pas local mais global car calculé une fois que tous les arbres sont construits.

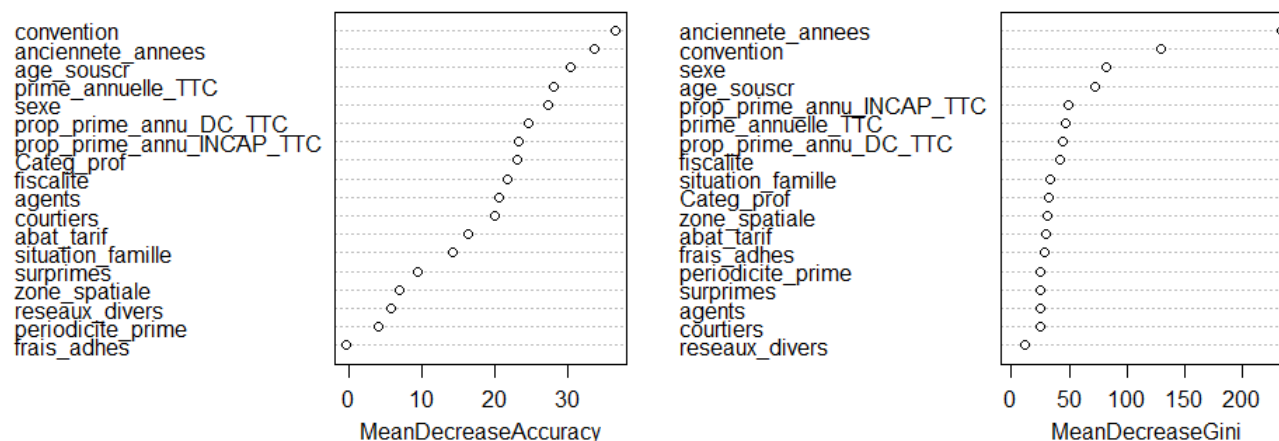


FIGURE III.12 – Eboulis des MDA et MDE obtenus par forêts aléatoires (500 souches=500 arbres à deux feuilles).

L'ancienneté en portefeuille, la convention, l'âge à la souscription, le sexe et la prime annuelle moyenne TTC sont les variables les plus importantes d'après les deux critères.

2.3 - Boosting d'arbres binaires

Le Boosting est également une technique d'agrégation, mais qui est qualifiée d'adaptative. Ainsi, la première technique de Boosting qui a été conçue est adaboost, pour « adaptative Boosting ». Ici aussi l'objectif est de diminuer la variance et le biais de prévision, mais l'avantage par rapport à d'autres méthodes comme les k -plus-proches voisins (très sensibles au paramètre k) est la faible sensibilité aux paramètres. Il n'est donc pas utile de consacrer beaucoup de temps à les optimiser.

2.3.1 - Principe de fonctionnement

L'idée est toujours d'avoir une séquence de modèles, mais la différence avec le Bagging est qu'à chaque itération nous allons essayer de construire un nouvel arbre comme une version mieux ajustée de l'arbre qui le précède dans la séquence ordonnée d'arbres.

Ainsi, plusieurs arbres sont construits par bootstrap afin d'être *in fine* agrégés, et les observations mal ajustées par un arbre donné, à l'origine d'une erreur d'ajustement non nulle, vont

se voir attribuer lors de la construction de l'arbre suivant un poids supérieur aux poids des observations bien ajustées. C'est en ce sens que le Boosting est une méthode d'agrégation dite adaptative.

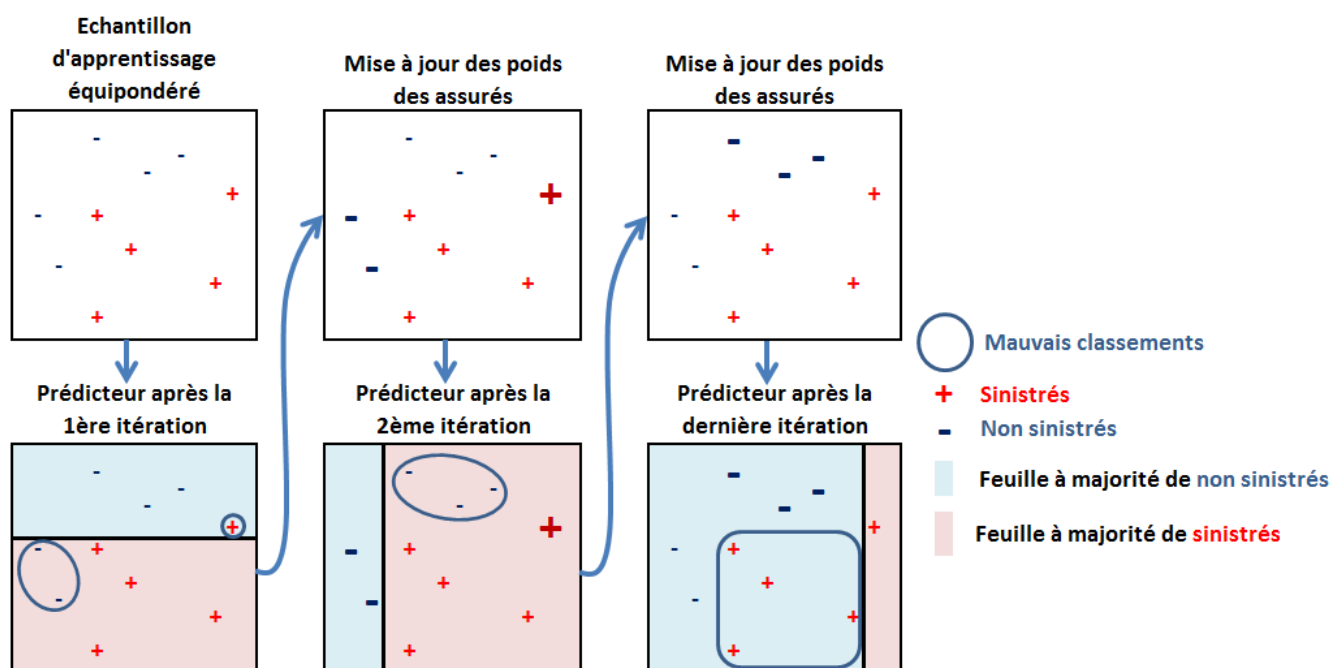


FIGURE III.13 – Principe du Boosting

2.3.2 - Adaboost

La plus connue des méthodes de Boosting est Adaboost. On constate au vu de l'algorithme 1 page 60 que si l'on ne se trompe pas en terme d'ajustement ($\hat{\phi}_{k-1}(x_i) = y_i$) relativement à un assuré et pour un arbre $k - 1$ donné, le poids $W_k(i)$ de cet assuré i ne varie pas beaucoup lors de l'ajustement de l'arbre k suivant, voire accuse une baisse qui peut être d'autant plus considérable (à cause de l'exponentielle $\exp(-\alpha_{k-1})$) que la performance α_{k-1} de l'arbre construit à l'étape $k - 1$ est grande.

A contrario, la procédure accorde plus de poids aux observations mal ajustées, d'autant plus si la performance α_{k-1} de l'arbre à l'étape $k - 1$ est grande, et ainsi le modèle suivant (étape k) va donc faire attention à mieux estimer ces observations. A la fin, pour chaque assurés i de l'échantillon d'apprentissage, on fait voter chaque arbre, c'est à dire que la classe (sinistrés ; non sinistrés) qui lui sera attribuée est la classe dominante, en attribuant un plus fort poids aux arbres qui ont une meilleure qualité d'ajustement.

Algorithme 1: Adaboost à partir d'arbres CART pour la classification binaire des assurés

Entrées : $\begin{cases} \bullet \text{ échantillon d'apprentissage : } \{(x_{(1)}, y_{(1)}), \dots, (x_{(r)}, y_{(r)})\}, \text{ où } r = \lfloor n \times 66\% \rfloor \text{ est le} \\ \text{nombre d'assurés dans l'échantillon d'apprentissage} \\ \bullet M : \text{Nombre de classifieurs préliminaires (arbres CART)} : \text{ils sont notés } (\hat{\phi}_1, \dots, \hat{\phi}_M) \end{cases}$

Sorties : $\hat{\phi}_{1, \dots, M}^{Adaboost}$

```
// M itérations de l'algorithme, une pour chaque classifieur  $\hat{\phi}_k$  :
for  $k \in \llbracket 1; M \rrbracket$  do
  if  $k=1$  then
     $W_k(i) \leftarrow 1, \forall i \in \llbracket 1; r \rrbracket$  // Initialisation des poids des assurés
  else
    // Mise à jour des poids des assurés selon la qualité de leur ajustement
    // par le précédent prédicteur et selon sa performance globale  $\alpha_{k-1}$ 
     $W_k(i) \leftarrow \begin{cases} W_{k-1}(i) \times \exp(-\alpha_{k-1}) & \text{si } \hat{\phi}_{k-1}(x_i) = y_i \\ W_{k-1}(i) \times \exp(+\alpha_{k-1}) & \text{si } \hat{\phi}_{k-1}(x_i) \neq y_i \end{cases}, \forall i \in \llbracket 1; r \rrbracket$ 
  end
  •  $D_k = \sum_{i=1}^r W_k(i)$  // coefficient de normalisation
  •  $W_k(i) \leftarrow \frac{W_k(i)}{D_k} \quad \forall i \in \llbracket 1; r \rrbracket$  // normalisation des poids
  • estimation de l'arbre  $\hat{\phi}_k$  en tenant compte des pondérations  $W_k(i)$  pour chaque assuré  $i$ 
  •  $\epsilon_k = \sum_{i=1}^r W_k(i) \times \mathbb{1}_{\hat{\phi}_k(x_i) \neq y_i}$  // erreur d'ajustement pondérée du prédicteur  $\hat{\phi}_k$ 
  •  $\alpha_k = \frac{1}{2} \times \log\left(\frac{1 - \epsilon_k}{\epsilon_k}\right)$  // prise en compte des performances du prédicteur  $\hat{\phi}_k$ 
end
```

$$\hat{\phi}_{1, \dots, M}^{Adaboost}(x_i) = \mathbb{1}_{\sum_{k=1}^M \alpha_k \times \hat{\phi}_k(x_i) > 0.5}$$

Parmi les faits contre-intuitifs au profit du Boosting, si on utilise Adaboost sur des souches (i.e. des arbres avec uniquement 2 feuilles), on peut faire mieux qu'un arbre sophistiqué avec autant de feuilles que la somme des feuilles de toutes les souches (i.e. le double du nombre d'arbres). Autrement dit, la qualité de prévision d'Adaboost avec 10 souches peut être meilleures que la qualité de prévision d'un seul arbre avec 20 feuilles. Il peut donc être intéressant, pour ne pas trop augmenter la complexité d'Adaboost, de le pratiquer avec des arbres d'au plus 8 feuilles.

Cependant, l'inconvénient majeur du Boosting se manifeste dans le cas où un grand poids est affecté aux observations erronées (anomalies, données bruitées, erreur d'approximation ou d'arrondi machine,...) ou d'outliers ce qui peut avoir pour conséquence de faire exploser l'erreur, d'où l'importance de bien la contrôler et de ne pas sur-ajuster le modèle. Toutefois, le sur-ajustement n'est pas inquiétant ici car une autre des propriétés paradoxales du Boosting est de pouvoir faire décroître l'erreur d'ajustement du modèle en rajoutant des modèles dans la

séquence qui, si elle s'accroît davantage, permet non seulement de préserver l'erreur d'ajustement mais permet en plus de maintenir une décroissance de l'erreur de prévision, qui finit par se stabiliser.

Le Boosting réduit la variance comme le Bagging mais il peut aussi réduire le biais, tandis que la forêt aléatoire réduit significativement la variance. Ainsi, en fonction du problème posé, parfois c'est le Boosting qui peut donner de meilleurs résultats, des fois ce sont les forêts aléatoires.

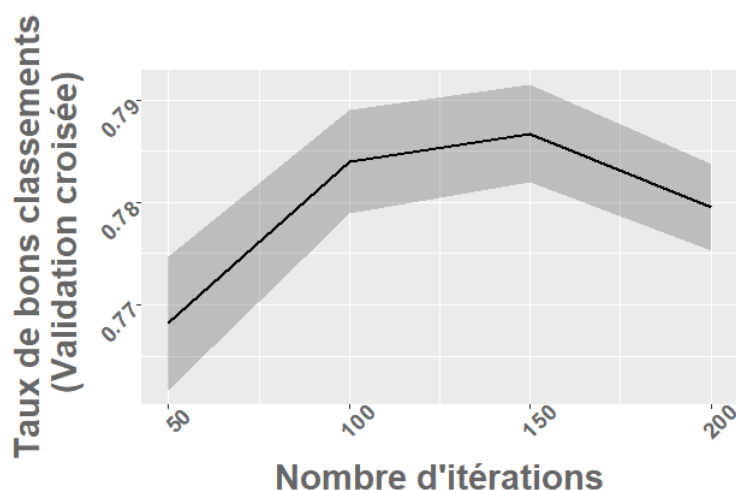


FIGURE III.14 – Adaboost : Taux de bons classements ($= 1 - R_n^{appr}$) par validation croisée, en fonction du paramètre M , le nombre d'itérations.

Le nombre d'itérations retenu est donc $M = 150$ au vu de la figure III.14 ci-dessus puisqu'il maximise le taux de bons classements par validation croisée.

3 - Comparaison avec la régression logistique

3.1 - Définitions et Hypothèses

Un **modèle de régression logistique** est un modèle qui tente de décrire un **supposé lien linéaire indirect** entre notre variable d'intérêt Y qualitative binaire désignant l'occurrence ou non de sinistre (i.e. $Y = 1$ ou $Y = 0$), et les p variables « exogènes » X_1, \dots, X_p , ou, plus exactement, de leur combinaison $\theta_0 + \theta_1 \times X_1 + \dots + \theta_p \times X_p$ **linéaire en les paramètres** $(\theta_j)_j$ que l'on va chercher à estimer dans la phase d'apprentissage.

En effet, il n'est pas possible de décrire un lien supposé linéaire de manière **directe** entre une variable cible à valeur dans $\{0, 1\}$ et des variables explicatives qualitatives (qui peuvent se voir comme une somme de variables indicatrices portant sur chacune de leurs modalités respectives¹⁵) car rien ne garantit que l'espace d'arrivée (espace image) d'une combinaison linéaire d'indicatrices de modalités soit binaire à valeurs dans $\{0, 1\}$ comme l'est la variable cible.

15. Les indicatrices sont des variables discrètes binaires.

Ainsi, la régression logistique se fonde sur l'idée que la variable exogène binaire Y peut se voir comme une variable de Bernoulli, avec une probabilité sous-jacente de valoir 1, notée $\pi_i \in [0; 1]$ pour un assuré i :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{où} \quad Y_i \sim \mathbf{Bernoulli}(\pi_i), \quad \forall i \in \llbracket 0; n \rrbracket \quad \text{et où } n \text{ est le nombre d'assurés (lignes)}$$

Par conséquent, plutôt que de chercher un lien direct entre Y et les régresseurs supposés **déterministes**¹⁶ X_1, \dots, X_p , nous nous contentons de supposer qu'en moyenne (i.e. $\mathbb{E}[Y_i] = \pi_i \in [0; 1]$ pour l'assuré i) la variable cible présente un lien linéaire avec les variables explicatives, et il s'agira ensuite de chercher à caractériser ce lien, afin d'établir une règle de prédiction qui va prédire $Y_i = 1$ si l'estimation de π_i est supérieure à 0.5 et $Y = 0$ sinon. Cependant, il n'est toujours pas possible d'affirmer à priori que $\theta_0 + \theta_1 \times X_1 + \dots + \theta_p \times X_p$ soit à valeur dans $[0; 1]$. Il est donc nécessaire de recourir à une « fonction de lien » bijective $g : \pi \in [0; 1] \mapsto g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \in \mathbb{R}$ dite logistique ou logit, afin de faire correspondre les espaces de définition des deux membres que l'équation de modélisation suivante :

$$\underbrace{\mathbb{R} \ni g(\pi_i) = g(\mathbb{E}[Y_i]) = \log\left(\frac{\pi_i}{1-\pi_i}\right)}_{\substack{\text{membre de gauche de l'équation de} \\ \text{modélisation} = \text{partie à expliquer}}} = \underbrace{\theta_0 + \theta_1 \times X_1^{(i)} + \dots + \theta_p \times X_p^{(i)}}_{\substack{\text{membre de droite de l'équation de} \\ \text{modélisation} = \text{partie explicative}}} \in \mathbb{R}$$

De l'équation précédente, il vient :

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right) \\ \iff \frac{\pi_i}{1-\pi_i} &= \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right] \\ \iff \pi_i &= (1-\pi_i) \times \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right] \\ \iff \pi_i &= \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right] - \pi_i \times \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right] \\ \iff \pi_i \times \left(1 + \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right]\right) &= \exp\left[\theta_0 + \left(\sum_q \theta_q \times X_q^{(i)}\right)\right] \end{aligned}$$

16. Si les régresseurs ne sont pas supposés déterministes, les probabilités conditionnelles et les espérances conditionnelles doivent être utilisées : par exemple, $\mathbb{E}[Y_i] = \pi_i$ si les régresseurs sont supposés déterministes, et $\mathbb{E}[Y_i | \mathbb{X}^{(i)} = x^{(i)}] = \pi_i$ s'ils sont supposés aléatoires. Cependant, par abus de notation et par souci d'allègement, il est possible d'omettre le conditionnement tout en supposant les régresseurs aléatoires.

d'où finalement :

$$\pi_i = \frac{e^{\theta_0 + \theta_1 \times X_1^{(i)} + \dots + \theta_p \times X_p^{(i)}}}{1 + e^{\theta_0 + \theta_1 \times X_1^{(i)} + \dots + \theta_p \times X_p^{(i)}}} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \times X_1^{(i)} + \dots + \theta_p \times X_p^{(i)})}}$$

et l'estimateur « plug-in¹⁷ » de $\eta^*(x) = \mathbb{P}[Y = 1 | \mathbb{X} = x]$ est donc pour l'individu i :

$$\hat{\pi}_i = \hat{\eta}(x^{(i)}) = \frac{1}{1 + e^{-(\hat{\theta}_0 + \hat{\theta}_1 \times x_1^{(i)} + \dots + \hat{\theta}_p \times x_p^{(i)})}}$$

Dans le cas où les variables explicatives sont catégorielles, ce qui est notre cas, ce sont les indicatrices de leurs modalités $m_{\bullet}^{X_{\bullet}}$ qui apparaissent dans l'équation de modélisation, comme suit pour l'individu i :

$$\begin{aligned} g(\pi_i) = \theta_0 + & \left(\theta_1^1 \times \mathbb{1}_{X_1^{(i)} = m_1^{X_1}} + \dots + \theta_1^k \times \mathbb{1}_{X_1^{(i)} = m_k^{X_1}} \right) \\ & + \dots + \\ & \left(\theta_p^1 \times \mathbb{1}_{X_p^{(i)} = m_1^{X_p}} + \dots + \theta_p^j \times \mathbb{1}_{X_p^{(i)} = m_j^{X_p}} \right) \end{aligned}$$

3.2 - Estimation des paramètres

Quant aux paramètres θ , ils sont estimés par Maximum de Vraisemblance, de la manière suivante :

1. Par indépendance des observations, la vraisemblance du n-échantillon (y_1, \dots, y_n) issu de Y s'écrit :

$$\begin{aligned} L_n(Y^{obs}, \theta) &= L_n \left(\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} \right) \\ &= \prod_{i=1}^n \mathbb{P}_{\theta} [Y_i = y_i | \mathbb{X}^{(i)} = x^{(i)}] \\ &= \prod_{i=1}^n \mathbb{P}_{\theta} [Y_i = 1 | \mathbb{X}^{(i)} = x^{(i)}]^{y_i} \times \mathbb{P}_{\theta} [Y_i = 0 | \mathbb{X}^{(i)} = x^{(i)}]^{1-y_i} \\ &= \prod_{i=1}^n (\pi_{\theta,i})^{y_i} \times (1 - \pi_{\theta,i})^{1-y_i} \\ \text{car } \pi_i &= \mathbb{E} [Y_i | \mathbb{X}^{(i)} = x^{(i)}] = \mathbb{P} [Y_i = 1 | \mathbb{X}^{(i)} = x^{(i)}] \\ &= \prod_{i=1}^n \left(\frac{1}{1 + e^{-(\theta_0 + \theta_1 \times x_1^{(i)} + \dots + \theta_p \times x_p^{(i)})}} \right)^{y_i} \times \left(\frac{1}{1 + e^{\theta_0 + \theta_1 \times x_1^{(i)} + \dots + \theta_p \times x_p^{(i)}}} \right)^{1-y_i} \end{aligned}$$

2. l'EMV associé vérifie donc :

$$\hat{\theta}_{MV} = \operatorname{argmax}_{\theta \in \mathbb{R}^p} L_n(Y^{obs}, \theta) = \operatorname{argmax}_{\theta \in \mathbb{R}^p} \underbrace{l_n(Y^{obs}, \theta)}_{\log L_n(Y^{obs}, \theta)}$$

17. On parle d'estimateur plug-in car on remplace le paramètre à estimer θ par son estimateur $\hat{\theta}$ dans l'expression de π_i .

3. Afin d'obtenir une expression de l'EMV, il est donc naturel (« l'extremum s'obtient en annulant la dérivée ») dans un premier temps de s'intéresser au score :

$$S(Y^{obs}, \theta) = \frac{\partial l_n(Y^{obs}, \theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial l_n(Y^{obs}, \theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial l_n(Y^{obs}, \theta)}{\partial \theta_p} \end{pmatrix}$$

car le système de p équations qui nous intéresse et qui donne l'EMV est : $S(Y^{obs}, \hat{\theta}_{MV}) = 0$.

Cependant, dans notre cas de fonction de lien logit, il n'existe pas de formule analytique pour cet estimateur. Le problème associé à la détermination de $\hat{\theta}_{MV}$ étant un problème d'optimisation convexe, un algorithme de type Newton-Raphson (cf [Annexe page 97](#)), adapté à un cadre statistique, peut être utilisé.

4. Pour ce faire, il convient, à chaque itération m de l'algorithme suivant, de mettre à jour la matrice d'information de Fisher $\mathcal{I} = - \left(\mathbb{E} \left[\frac{\partial^2 l_n(Y^{obs}, \theta)}{\partial \theta_i \partial \theta_j} \right] \right)_{1 \leq i, j \leq n}$ car elle dépend de θ :

- Initialisation : $u^{(0)}$
- Pour tout entier m :

$$u^{(m)} = u^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \times S(Y^{obs}, u^{(m-1)})$$

- Arrêt quand $|u^{(m)} - u^{(m-1)}| \leq \Delta$

- On pose $\hat{\theta}_{MV} = \begin{pmatrix} \hat{\theta}_{MV}^1 \\ \vdots \\ \hat{\theta}_{MV}^k \\ \vdots \\ \hat{\theta}_{MV}^p \end{pmatrix} = u^{(m)}$

L'objectif de ce mémoire n'étant pas d'être un exposé nourri sur la régression logistique mais plutôt de présenter l'ossature et l'esprit des méthodes d'apprentissage utilisées, nous orientons, pour plus de détails sur l'algorithme et sur l'expression détaillée et développée de la matrice d'information de Fisher, le lecteur vers le support très pédagogique [\[Rou\]](#) qui constitue une excellente base d'approfondissement.

Nos modèles seront, dans ce qui va juste suivre, de la forme suivante pour un individu i :

$$g(\pi_i) = \theta_0 + \left(\sum_{h,k} \theta^{h,k} \times \mathbb{1}_{X_h^{(i)} = m_k^{X_h}} \right)$$

où $m_k^{X_h}$ est la $k^{\text{ème}}$ modalité de la variable X_h .

Variables	Modalités (indicatrices) associée aux paramètres	Estimations par MV du modèle complet			Estimations par MV du modèle après sélection bi-directionnelle de variables				
		Estimations des paramètres	$\overset{obs}{\chi}_{Wald}^2$ = Valeurs observées de la statistique de test (khi-2 de Wald)	p-valeurs = $\mathbb{P} \left[\chi^2(1) > \overset{obs}{\chi}_{Wald}^2 \right]$	Estimations des paramètres pour les modalités des variables finalement retenues	Intervalles de confiance au niveau 95%		$\overset{obs}{\chi}_{Wald}^2$ = Valeurs observées de la statistique de test (khi-2 de Wald)	p-valeurs = $\mathbb{P} \left[\chi^2(1) > \overset{obs}{\chi}_{Wald}^2 \right]$
Intercept		- 1.92	216	< 0.0001	- 1.8152	- 1.9323	- 1.7000	938.8747	< 0.0001
Sexe	FEMME	0.31	272	< 0.0001	0.3071	0.2709	0.3435	274.6305	< 0.0001
Agents	0	0.03	0	0.7835					
Courtiers	0	0.04	0	0.7482					
Reseaux_divers	0	0.08	0	0.4800					
Surprimes	0	- 0.19	27	< 0.0001	- 0.1862	- 0.2559	- 0.1152	26.9200	< 0.0001
Âge_souscription	<=33	0.16	29	< 0.0001	0.1650	0.1053	0.2248	29.3206	< 0.0001
	>45	- 0.00	0	0.9187	- 0.0041	- 0.0695	0.0610	0.0152	0.9019
Prime_annuelle_TTC	<=900	- 0.18	70	< 0.0001	- 0.1817	- 0.2244	- 0.1391	69.7100	< 0.0001
Anciennete_annees	<=3	- 0.57	989	< 0.0001	- 0.5686	- 0.6041	- 0.5334	992.4581	< 0.0001
Proportion_prime_annu_DC	<=25%	0.08	14	0.0002	0.0850	0.0416	0.1286	14.6572	0.0001
Proportion_prime_annu_INCAP	<=46%	- 0.17	58	< 0.0001	- 0.1704	- 0.2139	- 0.1270	59.0926	< 0.0001
Zone_spatiale	Arc_ext	0.08	14	0.0002	0.0762	0.0364	0.1163	13.9607	0.0002
Convention	AGRIC	0.39	27	< 0.0001	0.3783	0.2336	0.5204	26.7458	< 0.0001
	ART/COMM	- 0.12	11	0.0007	- 0.1214	- 0.1927	- 0.0500	11.1089	0.0009
	Exp./Cons/Lib	- 0.17	13	0.0003	- 0.1702	- 0.2624	- 0.0788	13.2149	0.0003
	MEDIC	- 0.17	6	0.0135	- 0.1709	- 0.3068	- 0.0379	6.2126	0.0127
	PARAMEDIC	0.18	19	< 0.0001	0.1786	0.0974	0.2597	18.6128	< 0.0001
Categorie_professionnelle	1	- 0.08	13	0.0003	- 0.0784	- 0.1194	- 0.0373	14.0068	0.0002
Situation_famille	Couple	0.07	16	< 0.0001	0.0713	0.0358	0.1068	15.5033	< 0.0001
Periodicite_prime	Mensuelle	0.04	2	0.1981					
Frais_adhesion	Non	0.28	71	< 0.0001	0.2859	0.2212	0.3526	72.8045	< 0.0001
Abattement_tarifaire	Non	0.10	21	< 0.0001	0.0947	0.0537	0.1360	20.3499	< 0.0001
Fiscalite	AGRICOLE	- 0.12	3	0.1048	- 0.1233	- 0.2710	0.0247	2.6736	0.1020
	ASSURANCE VIE	0.19	11	0.0007	0.1926	0.0797	0.3042	11.3124	0.0008

FIGURE III.15 – Estimations $(\hat{\theta}_{MV}^k)_k$ par MV des paramètres $(\theta^k)_k$ de la régression logistique sans interactions modélisant l'occurrence d'au moins un sinistre, avant (modèle complet) et après sélection bi-directionnelle de variables.

Une fois la procédure d'estimation des paramètres opérée, nous obtenons le tableau III.15 suivant qui donne, en plus des paramètres estimés par maximum de vraisemblance, le score observé $\overset{obs}{\chi}_{Wald}^2$ du test de nullité de ces paramètres et la p-valeur associée. Cette dernière est définie dans notre cas précis comme étant la probabilité $\mathbb{P} \left[\chi^2(1) > \overset{obs}{\chi}_{Wald}^2 \right]$ qu'une variable aléatoire suivant une loi du Khi-deux à 1 degré de liberté dépasse la valeur observée $\overset{obs}{\chi}_{Wald}^2$ de la statistique de test χ_{Wald}^2 ¹⁸ puisque cette dernière suit elle-même une loi du Khi-deux à 1 degré de liberté sous l'hypothèse H_0 de nullité du paramètre considéré.

Aussi, rejeter l'hypothèse H_0 au risque de première espèce $\alpha = 5\%$ en se basant sur la p-valeur revient identiquement à rejeter H_0 si $\overset{obs}{\chi}_{Wald}^2$ dépasse $q_{1-5\%}^{\chi^2(1)} = 3.8415$ le quantile d'ordre $1 - \alpha$ (avec $\alpha = 5\%$) d'une loi Khi-deux à 1 degré de liberté. En effet :

$$\begin{aligned} \text{Rejet de } H_0 \text{ au risque de première espèce } \alpha = 5\% &\iff \mathbb{P} \left[\chi^2(1) > \overset{obs}{\chi}_{Wald}^2 \right] \leq \alpha \left(= \mathbb{P} \left[\chi^2(1) > q_{1-5\%}^{\chi^2(1)} \right] \right) \\ &\iff q_{1-5\%}^{\chi^2(1)} \leq \overset{obs}{\chi}_{Wald}^2 \end{aligned}$$

18. Pour l'estimation $\hat{\theta}_{MV}^k$ du $k^{\text{ème}}$ paramètre, $\chi_{Wald}^2 \stackrel{H_0}{=} (\hat{\theta}_{MV}^k)^2 / \widehat{Var}(\hat{\theta}_{MV}^k)$, où $\widehat{Var}(\hat{\theta}_{MV}^k)$ est, à un coefficient multiplicatif près égal au nombre d'assurés, le $k^{\text{ème}}$ élément de la diagonale de l'estimation $\hat{\mathcal{I}}$ au point $\hat{\theta}_{MV}$ de la matrice d'information de Fisher \mathcal{I} .

La p-valeur a un lien avec le niveau de confiance que l'on accorde à la décision finale de se prononcer en faveur de la non nullité du paramètre considéré (i.e. rejet de l'hypothèse H_0 au risque de première espèce $\alpha = 5\%$) : une p-valeur inférieure au seuil $\alpha = 5\%$ (en vert sur le tableau [III.15](#)) signifie qu'il y a au plus 5% de chances qu'en se basant sur la valeur observée $^{obs}\chi^2_{Wald}$ de la statistique du test de Wald et son positionnement par rapport au quantile $q_{1-5\%}^{\chi^2(1)}$, on se soit prononcé à tort en faveur de la nullité du paramètre. A l'inverse des valeurs en vert, les valeurs en rouge dans le tableau [III.15](#) indiquent un non-rejet de l'hypothèse H_0 de nullité du paramètre considéré à cause d'une trop grande p-valeur, autrement dit à cause d'une trop grande possibilité de se tromper en se prononçant en faveur de la nullité du paramètre.

L'analyse de la significativité de la non-nullité (au niveau de confiance 95%) des paramètres nous indique que les indicatrices des réseaux commerciaux et celle de la périodicité de prime n'ont pas de réelle influence sur la (probabilité de) survenance d'au moins un sinistre incapacité. C'est d'ailleurs cohérent avec la sélection de variables selon le critère AIC ¹⁹ qui ne retient pas ces variables.

L'interprétation des paramètres repose sur l'idée qu'un coefficient significativement supérieur à 0 indique que la modalité associée est un facteur favorisant la survenance de sinistres incapacité. Par exemple, être une femme augmente le risque d'être sinistré(e) en incapacité ($\hat{\theta}_{MV}^{Femme} \approx 0.3071 > 0$). Ceci peut s'expliquer en partie par les congés maternité. De plus, les indicatrices étant de même amplitude unitaire, plus un coefficient est élevé en valeur absolue, plus l'influence négative ou positive de la modalité associée est forte, et vice-versa. Ainsi, c'est le fait d'avoir une ancienneté en portefeuille d'au plus trois ans qui semble avoir le plus fort impact sur la variable cible, avec un coefficient de -0.5686, ce qui signifie que les assurés qui prennent cette modalité ont moins de chances (signe négatif du coefficient) d'avoir au moins un sinistre incapacité.

Cependant, la limite d'une telle modélisation est de ne pas prendre en compte les interactions entre variables : quid du coefficient associé à l'indicatrice issue du produit de deux indicatrices de modalités ? Autrement dit, quelles chances a un assuré d'être sinistré au moins une fois s'il est caractérisé conjointement par deux variables explicatives ? Ainsi, une variable non significative (i.e. le paramètre associé est considéré nul) dans le cadre d'une modélisation sans interactions peut l'être si elle est couplée à une autre variable explicative (interaction d'ordre deux).

Nous tenterons donc, à partir de la page [69](#), d'opérer une régression logistique en y ajoutant les interactions d'ordre deux, mais avant cela, nous abordons la notion d'odd ratio, qui va apporter une lecture supplémentaire de l'influence des variables.

19. La sélection de variables selon le critère AIC est expliquée à la [partie III](#) en page [84](#).

L'odd ratio

$$OR_X(x_A, x_B) := \frac{odd_X(x_A)}{odd_X(x_B)} := \frac{\frac{\pi_{x_A}^X}{1-\pi_{x_A}^X}}{\frac{\pi_{x_B}^X}{1-\pi_{x_B}^X}}$$

concerne deux individus A et B qui diffèrent uniquement au regard d'une variable explicative X (toutes choses égales par ailleurs) dont ils prennent respectivement la modalité x_A et x_B . L'odd ratio est dans notre cas le rapport des chances (odds) d'être sinistré au moins une fois en incapacité (variable cible Y égale à 1) sachant que la variable X prend la modalité x_A plutôt que x_B puisque π_x^X désigne $\mathbb{P}[Y = 1 | X = x] = \mathbb{P}[Y = 1 | \mathbf{1}_{X=x} = 1]$. Par exemple, si l'on a une probabilité $\mathbb{P}[Y = 1] = \pi$ d'être sinistré au moins une fois en incapacité égale à $\frac{1}{4}$, cela signifie que sur 4 assurés, un est sinistré au moins une fois et les trois autres jamais, soit un rapport de 1 sinistré sur 3 non sinistrés, c'est-à-dire $odd = \pi/(1 - \pi) = 1/3$. Les odds généralisent cela au cas où l'on sait la modalité observée par une variable explicative.

Par conséquent, une règle d'interprétation naturelle consisterait à considérer que l'assuré dont la variable X prend la modalité x_A a plus de chances (respectivement moins de chances) d'être sinistré au moins une fois en incapacité qu'un assuré dont la variable X prend la modalité x_B si l'odd ratio $OR_X(x_A, x_B)$ est strictement supérieur à 1 (respectivement strictement inférieur à 1), toutes les autres variables explicatives étant égales par ailleurs.

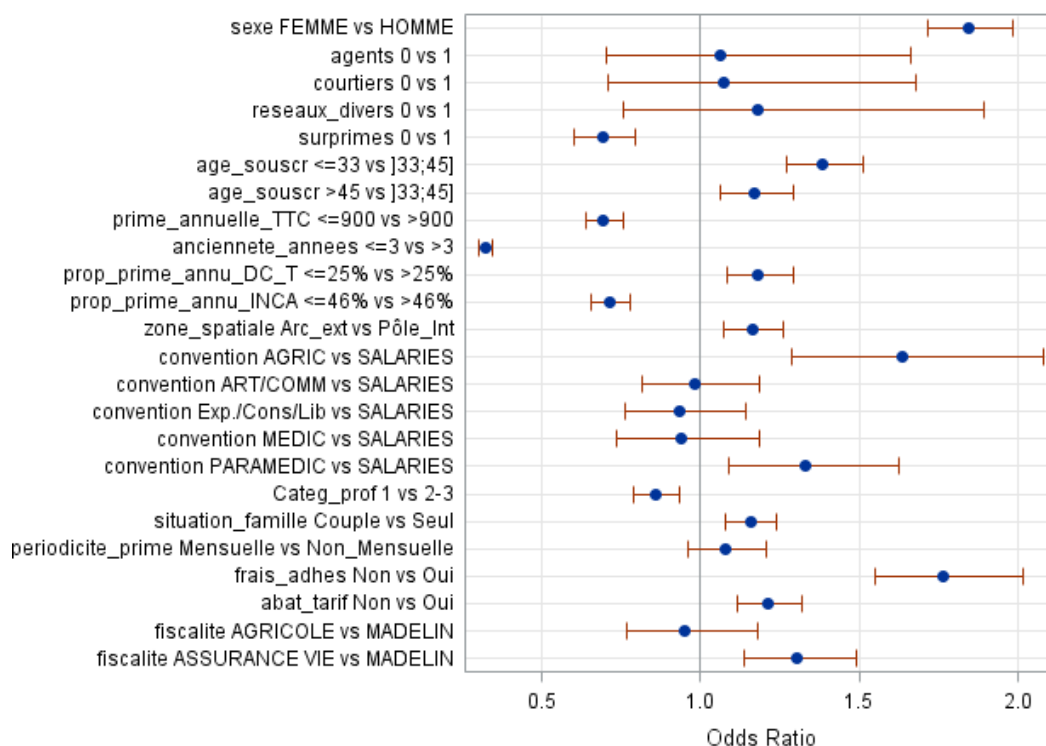


FIGURE III.16 – Le rapport des côtes (odds), appelé odd-ratio (OR), indique l'influence d'une modalité A d'une variable par rapport à une autre modalité B de la même variable (notation « variable modalité A vs modalité B »), toutes choses égales par ailleurs.

La figure III.16 ci-dessus nous indique les différents odd-ratios ainsi que leurs intervalles de confiance de niveau 95% et nous informe que la variable ancienneté en portefeuille, qui a deux modalités, est liée à la fois à l'odd ratio strictement inférieur à 1 le plus faible, le seul l'odd ratio dont l'inverse est supérieur à 2, et l'odd ratio qui est localisé avec le plus de précision puisqu'il a un intervalle de confiance de faible étendue. On est donc certain qu'une ancienneté supérieure à 3 ans constitue la caractéristique, parmi celles considérées, qui augmente le plus les chances d'être sinistré au moins une fois en incapacité.

D'autres odds ratio sont différents de 1, comme ceux qui nous informent qu'une femme et que ceux qui payent une prime annuelle moyenne TTC supérieure à 900€ ont plus de chances d'être sinistrés au moins une fois que leurs complémentaires respectifs, à savoir les hommes d'une part et ceux qui payent moins de 900€ d'autre part, toutes choses étant égales par ailleurs.

A présent intéressons nous à la modélisation des interactions d'ordre deux, que l'on va ajouter au modèle sans interactions.

3.3 - Modélisation des interactions

Nos modèles seront, dans cette partie, de la forme suivante pour un individu i :

$$g(\pi_i) = \underbrace{\theta_0 + \left(\sum_{h,k} \theta_{h,k}^{h,k} \times \mathbb{1}_{X_h^{(i)} = m_k^{X_h}} \right)}_{\text{modèle de la section précédente}} + \underbrace{\left(\sum_{l,h,j,k} \theta_{l,j}^{h,k} \times \mathbb{1}_{X_l^{(i)} = m_j^{X_l}} \times \mathbb{1}_{X_h^{(i)} = m_k^{X_h}} \right)}_{\text{interactions d'ordre 2}}$$

où $m_k^{X_h}$ est la $k^{\text{ème}}$ modalité de la variable X_h , et $m_j^{X_l}$ est la $j^{\text{ème}}$ modalité de la variable X_l .

On remarque (figure III.17 page 69) que sur les 31 variables ou interactions d'ordre 2 significatives du modèle complet (sans sélection de variables), 16 le sont également pour le modèle avec sélection de variables. Les plus élevés en valeur absolue parmi ces 16 coefficients estimés sont notamment ceux associés à l'ancienneté et le sexe pour lesquels l'interprétation est la même que précédemment : lorsqu'un assuré a plus de 3 ans d'ancienneté ou est une femme, il a un risque accru d'être sinistré.

La convention, qui n'est significative sans interactions que pour le modèle avec sélection de variables, présente également des coefficients estimés relativement élevés en valeur absolue qui indiquent par exemple que la profession agricole est un facteur aggravant du risque d'être sinistré au moins une fois en incapacité.

Estimations par MV du modèle avec et sans sélection bi-directionnelle de variables								
Modèles	Variables	Modalités (indicatrices) associée aux paramètres		Estimations des paramètres	Intervalles de confiance au niveau 95%			χ^2_{Wald} $=$ Valeurs observées de la statistique de test (khi-2 de Wald) $\mathbb{P} \left[\chi^2(1) > \chi^2_{Wald} \right]$
Interactions significatives pour le modèle complet	age_souscr*fiscalite	<=33	AGRICOLE	- 0,4093	-0,7204	-0,1019	7	0.0094
	surprimes*convention	0	MEDIC	0,3727	0,0589	0,7205	5	0.0260
	surprimes*fiscalite	0	ASSURANCE VIE	0,3598	0,0724	0,6806	5	0.0194
	age_souscr*fiscalite	>45	AGRICOLE	0,3297	0,0475	0,6139	5	0.0224
	conventio*periodicit	MEDIC	Mensuelle	- 0,2656	-0,4679	-0,0582	6	0.0109
	age_souscr*fiscalite	<=33	ASSURANCE VIE	0,2591	0,0402	0,4794	5	0.0206
	age_souscr*fiscalite	>45	ASSURANCE VIE	- 0,2381	-0,4632	-0,0152	4	0.0370
	surprimes*convention	0	PARAMEDIC	0,2270	0,0281	0,4296	5	0.0265
	prime_annu*fiscalite	<=900	AGRICOLE	0,1962	0,00915	0,3843	4	0.0402
	abat_tarif*fiscalite	Non	ASSURANCE VIE	- 0,1897	-0,3679	-0,0159	4	0.0341
	prime_annu*fiscalite	<=900	ASSURANCE VIE	- 0,1694	-0,3193	-0,0205	5	0.0262
	prop_prim*periodicit	<=25%	Mensuelle	0,1161	0,0433	0,1892	10	0.0018
Interactions significatives à la fois pour le modèle complet avec interactions et celui issu de la sélection bi-directionnelle de variables	age_souscr*abat_tarif	<=33	Non	- 0,0795	-0,1582	-0,0012	4	0.0471
	prime_ann*abat_tarif	<=900	Non	- 0,0683	0,014	0,1224	6	0.0134
	anciennet*periodicit	<=3	Mensuelle	- 0,0655	-0,1281	-0,0018	4	0.0419
	Intercept			- 1,7409	-1,8793	-1,6055	622	< 0.0001
	anciennete_annees	<=3		- 0,6896	-0,5751	-0,4929	12	0.0006
	sexe	FEMME		- 0,3405	0,256	0,3598	4	0.0461
	frais_adhe*fiscalite	Non	AGRICOLE	- 0,2949	-0,4036	-0,0563	4	0.0434
	sexe*age_souscr	FEMME	<=33	- 0,2502	0,1768	0,2914	49	< 0.0001
	sexe*age_souscr	FEMME	>45	- 0,1922	-0,2617	-0,1283	23	< 0.0001
	sexe*convention	FEMME	ART/COMM	- 0,1690	-0,2448	-0,1055	15	0.0001
	age_souscr*Categ_prof	<=33	1	- 0,1305	-0,225	-0,0969	11	0.0009
	age_souscr*Categ_prof	>45	1	0,1051	0,0618	0,2023	5	0.0191
interactions significatives uniquement pour la sélection bi-directionnelle de variables	prime_ann*Categ_prof	<=900	1	0,0806	0,0595	0,1429	8	0.0038
	sexe*prop_prime_annu	FEMME	<=25%	0,0734	0,0249	0,1159	8	0.0040
	prime_ann*prop_prime	<=900	<=25%	- 0,0702	-0,1018	-0,025	5	0.0221
	age_souscr*situation	<=33	Couple	0,0634	0,0286	0,1293	4	0.0490
	anciennet*prop_prime	<=3	<=46%	0,0561	0,0132	0,0942	5	0.0205
	sexe*prop_prime_annu	FEMME	<=46%	0,0523	0,0211	0,112	4	0.0356
	anciennet*Categ_prof	<=3	1	- 0,0512	-0,0943	-0,0219	5	0.0237
	sexe*situation_famil	FEMME	Couple	- 0,0409	0,014	0,0861	4	0.0471
	convention	AGRIC		0,4961	0,3038	0,6842	26	< 0.0001
	age_souscr*convention	<=33	MEDIC	0,2919	0,1011	0,4826	9	0.0027
	age_souscr*convention	<=33	AGRIC	- 0,2040	-0,3744	-0,0364	6	0.0179
	convention	Exp./Cons/Lib		- 0,2007	-0,3114	-0,0917	13	0.0003
	surprimes	0		- 0,1974	-0,2679	-0,1255	30	< 0.0001
	age_souscr	<=33		0,1941	0,12	0,2681	26	< 0.0001
	sexe*convention	FEMME	MEDIC	0,1848	0,0475	0,3247	7	0.0089
	frais_adhes	Non		0,1833	0,0826	0,2882	12	0.0005
	prop_prim*convention	<=46%	AGRIC	0,1774	0,0369	0,3149	6	0.0123
	convention	MEDIC		- 0,1696	-0,3363	-0,009	4	0.0420
	prop_prime_annu_INCA	<=46%		- 0,1552	-0,2129	-0,0982	28	< 0.0001
	prime_annuelle_TTC	<=900		- 0,1470	-0,1967	-0,0974	34	< 0.0001
	convention	PARAMEDIC		0,1341	0,0132	0,2527	5	0.0281
	convention	ART/COMM		- 0,1266	-0,2092	-0,0439	9	0.0027
	age_souscr*convention	<=33	PARAMEDIC	0,1105	0,00091	0,2208	4	0.0488
	age_souscr*convention	<=33	ART/COMM	- 0,1100	-0,2041	-0,016	5	0.0219
	abat_tarif	Non		0,0985	0,0568	0,1405	21	< 0.0001
	sexe*convention	FEMME	Exp./Cons/Lib	0,0980	0,0076	0,189	4	0.0341
	zone_spatiale	Arc_ext		0,0796	0,0392	0,1204	15	0.0001
	prime_ann*zone_spati	<=900	Arc_ext	- 0,0668	-0,1068	-0,0267	11	0.0011
	situation_famille	Couple		0,0649	0,0272	0,1028	11	0.0008
	Categ_prof	1		- 0,0490	-0,0957	-0,0022	4	0.0402
	prop_prim*Categ_prof	<=25%	1	0,0433	0,00045	0,086	4	0.0474

FIGURE III.17 – Estimations par MV des paramètres significativement différents de 0 (au risque de première espèce 5%) de la régression logistique avec interactions, avant (modèle complet) et après sélection bi-directionnelle de variables.

4 - Synthèse du pouvoir explicatif des modèles

En continuant de se baser sur le critère R_n^{app} pour différencier les méthodes d'apprentissage supervisé, on obtient le tableau suivant :

Modèles	Erreur d'apprentissage (R_n^{appr})
Arbre complet	0.1095
Arbre optimal	0.1634
Forêts aléatoires	0.1666
Adaboost	0.1118
Régression logistique sans interactions	Erreur d'apprentissage (R_n^{appr})
complet	0.1650
sélection de variables bi-directionnelle	0.1647
Régression logistique avec interactions	Erreur d'apprentissage (R_n^{appr})
complet	0.1618
sélection de variables bi-directionnelle	0.1627

TABLE III.1 – Moyennes des erreurs d'apprentissage obtenues en réitérant 100 fois l'échantillonnage.

C'est donc la méthode adaboost et l'arbre complet qui présentent de faibles erreurs d'apprentissage avoisinant les 11% pour les deux. A l'inverse, ce sont les forêts aléatoires qui maximisent l'erreur avec 16.66%.

5 - Prédiction de l'échantillon test (pouvoir de généralisation)

Afin de déterminer la qualité de prédiction des différents modèles sur l'échantillon test, nous avons besoin de définir certains taux ou scores définis à partir de la matrice de confusion suivante :

Observés	Prévus	
	Sinistrés	Non Sinistrés
Sinistrés	VP	FN
Non Sinistrés	FP	VN

FIGURE III.18 – Matrice de Confusion.

5.1 - Scores de prédiction

L'évènement critique qui nous intéresse et qui est censé nous alerter étant la prédiction d'au moins un sinistre incapacité, c'est lui que l'on qualifie paradoxalement de « positif » et on note :

- $\underline{\text{TVP}} = \frac{\text{VP}}{\text{VP} + \text{FN}} \in [0; 1]$ le taux de **V**rais **P**ositifs (aussi appelé « **sensibilité** ») que l'on souhaite le plus proche de 1 (sinistrés bien prédits),

$TVN = \frac{VN}{VN + FP} \in [0; 1]$ le taux de Vrais Négatifs (aussi appelé « **spécificité** »), que l'on souhaite également le plus proche de 1 (non sinistrés bien prédits),

- $TFP = \frac{FP}{FP + VN} = 1 - \text{spécificité} \in [0; 1]$ le taux de Faux Positifs (ou taux de fausses alertes), que l'on souhaite le plus proche de 0 (sinistrés mal prédits),

$TFN = \frac{FN}{FN + VP} = 1 - \text{sensibilité} \in [0; 1]$ le taux de Faux Négatifs, que l'on souhaite le plus proche de 0 également (non sinistrés mal prédits),

- $PSS = (TVP - TFP) \in [-1; 1]$ est le score de Pierce, que l'on souhaite au moins supérieur à 0, et autant que possible proche de 1 ce qui signifierait une parfaite prévision de tous les sinistrés ($TVP = 1$) sans fausse alerte ($TFP = 0$),
- $TSG = \frac{VP + VN}{VP + FP + FN + VN} = 1 - R_n^{test} \in [0; 1]$ est le taux de succès global, que l'on souhaite le plus proche de 1, mais qui n'est pas forcément un bon indicateur si l'évènement critique est rare (car dans ce cas VP risque d'être faible et VN risque d'être élevé).

Toutefois, bien que l'évènement critique soit $\{Y = 1\}$, il serait dommageable de mal prédire un non sinistré car si les modèles d'apprentissage ainsi que les modalités ou les seuils de variables pertinents qui émergent de ce mémoire aident à ajuster un tarif incapacité, il est possible qu'en conséquence un assuré ou un assuré que l'on croit à tort être relativement à risque soit amené à payer trop cher sa prime, et risquerait de ne pas souscrire et d'aller voir la concurrence. Ainsi, nous nous intéresseront à maximiser TVP et TVN (ce qui revient à minimiser respectivement TFP et TFN), en plus de veiller à maximiser PSS et TSG.

On introduit également le score de Brier, qui permet de mesurer la précision d'un modèle de classification supervisée, en confrontant les $\hat{\pi}_i$ aux Y_i . Il est défini comme suit :

$$BS = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{\pi}_i - Y_i)^2$$

Ce score est donc compris entre 0 et 1, et on comprend aisément que plus les probabilités estimées $\hat{\pi}_i$ sont grandes ($\hat{\pi}_i > 0.5$) lorsqu'il y a effectivement au moins un sinistre ($Y_i = 1$), ou, plus les $\hat{\pi}_i$ sont petites ($\hat{\pi}_i < 0.5$) lorsqu'il n'y a effectivement jamais eu sinistre ($Y_i = 0$), et plus la quantité $(\hat{\pi}_i - Y_i)^2$ va être proche de 0²⁰. Si l'on généralise ce raisonnement à tous les assurés, plus tous les termes de la somme sont proches de 0, **plus le score de Brier se rapprochera 0, et meilleure sera la prévision.**

On obtient ainsi les résultats suivants, avec un code de couleur qui associe le **vert** à la meilleure valeur et le **rouge foncé** à la valeur la moins optimale :

20. Les quantités $\hat{\pi}_i - Y_i$ peuvent être vues comme des résidus.

Modèles	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS
Arbre complet	0.1744	0.7929	0.1771	0.0823	0.0949
Arbre optimal	0.1313	0.8333	0.0480	0.0075	0.0405
Forêts aléatoires	0.1544	0.8319 ²¹	0.0048	0.0004	0.0044
Adaboost	0.1696	0.7840	0.2435	0.1064	0.1370
Régression logistique sans interactions	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS
complet	0.1296	0.8323	0.0310	0.0052	0.0258
sélection de variables bi-directionnelle	0.1296	0.8325	0.0327	0.0053	0.0274
Régression logistique avec interactions	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS
complet	0.1285	0.8342	0.0781	0.0125	0.0656
sélection de variables bi-directionnelle	0.1285	0.8353	0.0702	0.0096	0.0606

TABLE III.2 – Moyennes des scores de prévisions obtenues en réitérant 100 fois l'échantillonnage.

On remarque que le TSG — qui est à relativiser car même si le phénomène d'occurrence d'au moins un sinistre incapacité n'est pas rare, la proportion de sinistrés du portefeuille (16.7%) reste cependant faible par rapport à l'évènement complémentaire d'absence d'occurrence de sinistre incapacité — est majoré par le modèle logistique avec interactions et sélection bi-directionnelle de variables, qui semble être le meilleur modèle de prévision de ce point de vue là, et est minoré par Adaboost, qui semble être au contraire le pire modèle de prévision eu égard au TSG. En revanche, Adaboost maximise **TVP** comparativement aux autres modèles, contrairement aux forêts aléatoires qui le minimisent.

5.2 - Courbe ROC et AUC

Rappelons que dans le cadre de cette partie, nous cherchons à prédire correctement les valeurs de Y , mais nous voulons également quantifier la prédisposition (mathématiquement la probabilité) d'un assuré à être égal à 1 ou 0 (respectivement occurrence ou non d'au moins un sinistre incapacité). C'est donc dans cette optique que nous commençons par tracer la courbe ROC pour tous les modèles.

Courbe ROC. Précédemment, nous avons comparé $\hat{\pi}_i$ à un seuil pour effectuer une prédiction : si $\hat{\pi}_i > \underline{0.5}$, alors $\hat{Y}_i = 1$, où $\hat{\pi}_i$ est la probabilité a posteriori pour un assuré i de l'échantillon test d'être catégorisé « sinistré ». Nous avons ainsi pu établir la matrice de confusion et ainsi déterminer les 2 scores **TVP** et **TFP**. La courbe ROC étend cela en faisant varier le seuil, précédemment fixé à 0.5, continûment sur $[0; 1]$ et de manière croissante. Ainsi, la courbe ROC représente le taux de vrais sinistrés **TVP** en fonction du taux de fausses alertes **TFP** pour chaque valeur de ce seuil dans $[0; 1]$.

21. $R_n^{test} = 1 - 0.8319 = 0.1681$, et pour rappel l'erreur out-of-bag=0.1659.

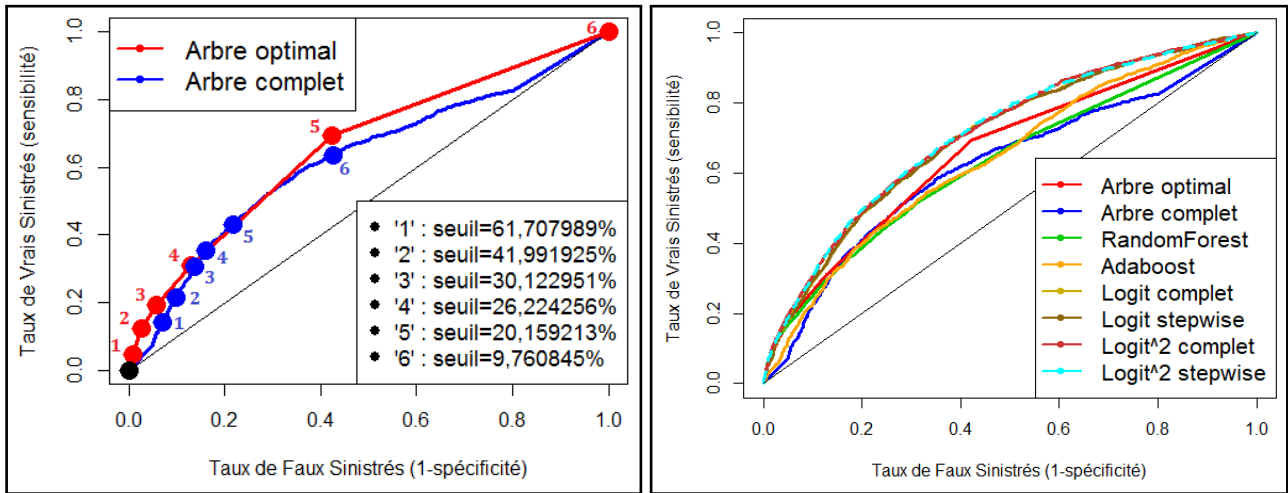


FIGURE III.19 – Courbes ROC.

On comprend par conséquent que si la classification supervisée attribue des valeurs $\{0, 1\}$ de manière totalement aléatoire (équiprobabilité), il y a alors équirépartition dans la table de contingence, et $\mathbf{TV}\mathbf{P} = \mathbf{TF}\mathbf{P}$ dans tous les cas de seuil. Ainsi, la courbe ROC se confond avec la droite $y = x$. À l'opposé, si la classification prédit parfaitement les sinistres quelle que soit la valeur du seuil dans $[0; 1]$ ²², alors $\mathbf{TV}\mathbf{P}$ est, en toute logique, constant et égal à 1, et, dans ce cas, la courbe ROC colle au côté supérieur du carré $[0; 1] \times [0; 1]$.

AUC. Cependant, la courbe ROC d'un premier prédicteur peut être au dessus de la courbe ROC d'un second prédicteur dans un intervalle, et en-dessous dans un autre : il n'y a donc pas d'ordre total des courbes ROC. Il est donc d'usage de tenir compte de l'ordre total rendu possible par l'aire sous la courbe ROC, appelée AUC, et qui rend ainsi comparable la performance globale de plusieurs prédicteurs. La grille suivante nous donne des critères d'appréciation de ce pouvoir prédictif :

22. Un seuil égal à 1 est un cas particulier puisque dans cette situation $\hat{Y} = 1 \iff \hat{\pi} > 1$ mais comme $\hat{\pi} \in [0; 1]$ alors $\hat{Y} = 0$ pour tous les individus : les individus dont $Y = 0$ sont donc parfaitement prédits ($\mathbf{TF}\mathbf{P} = 0$), et ceux dont $Y = 1$ ne le sont jamais ($\mathbf{TV}\mathbf{P} = 0$).

AUC	Pouvoir prédictif
[0,5 ; 0,7[Mauvais
[0,7 ; 0,8[Suffisant
[0,8 ; 0,9[Bon
[0,9 ; 0,1]	Très Bon

FIGURE III.20 – Grille d'appréciation de l'aire sous la courbe ROC (AUC).

Modèles	AUC
Arbre complet	0.6224
Arbre optimal	0.6617
Forêts aléatoires	0.6311
Adaboost	0.6437
Régression logistique sans interactions	AUC
complet	0.7053
sélection de variables bi-directionnelle	0.7048
Régression logistique avec interactions	AUC
complet	0.7117
sélection de variables bi-directionnelle	0.7122

TABLE III.3 – Moyennes des AUC obtenues en réitérant 100 fois l'échantillonnage.

D'après l'observation du tableau précédent, il semble que les modèles de régression logistique, notamment ceux avec interactions, permettent une meilleure généralisation de la classification supervisée binaire comparativement aux modèles dérivés des arbres CART.

6 - Apport des données externes

L'apport des 3 variables externes retenues (eu égard à leur score $S_{Y/X}$ et au test de WMW) que sont le taux de travail à temps plein, l'espérance de vie à la naissance et le nombre de personnes par ménage est mitigé voire décevant puisqu'il permet de faire augmenter **TVP** pour l'arbre complet de 17.71% à 19.94% mais il fait diminuer celui d'adaboost de 24.35% à 19.68% (pour le reste des méthodes il est sensiblement le même).

Modèles d'apprentissage	Apprentissage	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Arbre complet	0.0962	0.1865	0.7831	0.1994	0.0985	0.1008	0.6191
Arbre optimal	0.1635	0.1312	0.8333	0.0476	0.0074	0.0401	0.6733
Forêts aléatoires	0.1665	0.1593	0.8314	0.0037	0.0001	0.0036	0.6419
Adaboost	0.0973	0.1623	0.7960	0.1968	0.0825	0.1142	0.6468
Régression logistique sans interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.1648	0.1296	0.8327	0.0305	0.0047	0.0259	0.7060
sélection de variables bi-directionnelle	0.1645	0.1297	0.8328	0.0314	0.0048	0.0266	0.7056
Régression logistique avec interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.1606	0.1284	0.8336	0.0780	0.0132	0.0649	0.7113
sélection de variables bi-directionnelle	0.1624	0.1284	0.8351	0.0710	0.0092	0.0618	0.7123

TABLE III.4 – Moyennes des scores de performance obtenues en réitérant 100 fois l'échantillonnage - Seules 3 variables externes sont incluses.

Nous ne devons toutefois pas conclure que l'information contenue dans des variables relatives à l'environnement des assurés n'est pas utile pour expliquer la survenance d'au moins un sinistre incapacité. En effet, cela est probablement dû au fait que les données externes sont propres à la zone géographique dans laquelle vivent les assurés. Il serait donc intéressant de pouvoir récolter individuellement des informations comme les habitudes alimentaires et la fréquence des pratiques sportives pour chaque assuré.

IV - Étude de la rentabilité des sinistrés par classification binaire

Dans cette partie, contrairement à la [précédente](#), il ne s'agit plus d'étudier la variable cible

$$Y = \begin{cases} 1 & \text{si l'assuré a été sinistré au moins une fois en incapacité} \\ 0 & \text{si l'assuré n'a jamais été sinistré en incapacité} \end{cases}$$

permettant d'obtenir pour chaque feuille d'arbre CART la proportion de sinistrés $\left(\frac{\text{assurés sinistrés}}{\text{total des assurés}}\right)$, ayant vocation à estimer la probabilité $\mathbb{P}[Y = 1 | \mathbb{X}]$ d'être sinistré sachant les caractéristiques \mathbb{X} , mais plutôt la variable cible

$$Y = \mathbb{1}_{S/P > 100\%} \quad \text{avec pour chaque assuré sinistré, avec :}$$
$$S/P = \frac{\sum \text{prestations versées}}{\sum \text{primes versées, hors taxes et nettes de chargements}}$$

débouchant sur la proportion de sinistrés non rentables $\left(\frac{\text{sinistrés non rentables}}{\text{total des sinistrés}}\right)$ censée approcher la probabilité $\mathbb{P}[Y = 1 | \text{sinistré}, \mathbb{X}]$ d'être non rentable sachant que l'assuré est sinistré. Nous ne revenons pas sur la justification du recours à une classification binaire au lieu d'une régression (cas d'une variable cible égale au ratio S/P), reposant principalement sur la simplicité de deux classes, ainsi qu'une plus grande robustesse aux erreurs de valeurs de prestations et de primes, et sur la possibilité d'estimer les probabilités ci-dessus.

Par ailleurs, afin de ne pas introduire de biais dans la détermination du ratio S/P et donc de la variable cible, la population de sinistrés étudiée est réduite aux sinistrés dont l'indemnisation en incapacité est terminée. Par conséquent, la base de données totale passe d'une taille de 40 790 assurés ([partie III](#)) dont 16.7% sont sinistrés, à une taille de 6 800 sinistrés, dont 68,9% sont non rentables. Il n'y a donc pas de provisions au numérateur du ratio S/P, et le coût est donc un coût effectif passé pour chaque sinistré. Enfin, les primes au dénominateur du S/P sont nettes de taxes, contrairement à la variable explicative « prime annuelle moyenne TTC » où les taxes méritent d'y être incluses puisqu'elles constituent un effort financier supplémentaire pour l'assuré et peuvent donc contribuer à refléter un « appétit » à l'assurance, un besoin de couverture élevée, toutes choses étant égales par ailleurs. En revanche, l'assureur étant collecteur de taxes pour le compte de l'État, il n'est pas approprié de les garder au dénominateur du S/P définissant la variable cible et censé estimer la rentabilité du point de vue de l'assureur. De même pour les chargements perçus par l'assureur et censés couvrir ses frais (gestion, commissions apporteurs,...), et n'ont donc pas vocation à couvrir des prestations de sinistres. C'est donc la prime pure qui est au dénominateur du S/P.

Outre la raison évoquée en [introduction](#) d'une étude duale, tenant compte de toute l'épaisseur de la sinistralité (survenance et rentabilité), une autre utilité de cette partie tire sa justification du fait que les résultats de la [partie III](#) sont quantitativement insatisfaisants (score de

prédiction relatif au taux de vrais sinistrés), même si qualitativement, ils ont permis de faire émerger des facteurs de risques et des seuils ou modalités pertinentes. Ainsi, réduire la base aux sinistrés permet de faire passer la proportion d'assuré ayant une variable cible Y égale à 1 de 16,7% à une proportion de 68,9%, permettant ainsi de faire basculer l'équilibre en faveur des assurés ayant subi l'évènement critique ($Y = 1$) quel qu'il soit. Ces derniers ne seront donc pas dilués parmi la population des assurés ayant une cible Y égale à 0 dans les classes constituées par l'arbre CART, ce qui aura pour conséquence d'améliorer l'acuité des modèles à discerner l'évènement critique $Y = 1$, ce qui devrait se traduire par une augmentation du taux de vrais positifs.

Enfin, il est nécessaire de reproduire toute l'analyse exploratoire dont la démarche est donnée dans la [partie II](#), et qui passe par une discrétisation des variables continues et un regroupement des modalités des variables discrètes en tenant compte de la variable cible. A cet égard, la proportion de prime allouée à l'invalidité est, au préalable, réintégrée comme variable explicative. Pour plus de clarté, une grande partie de l'analyse exploratoire se trouve en [annexe page 92](#).

1 - Panorama des données transformées

La figure suivante — mettant face à face la cartographie des régions en fonction des proportions de sinistrés non rentables (carte de gauche) et celle issue de la [partie III](#) (carte de droite) présentant les régions abritant le plus d'actifs sinistrés (TNS ou salariés) en pourcentage sur la base des données en portefeuille — permet de distinguer à chaque fois deux classes de régions : celles dont le risque est aggravé et celle dont le risque est modéré, que l'on désigne respectivement par « arc extérieur » et « pôle intérieur », par allusion aux formes délimitées par une frontière noire sur la figure de droite ci-après :

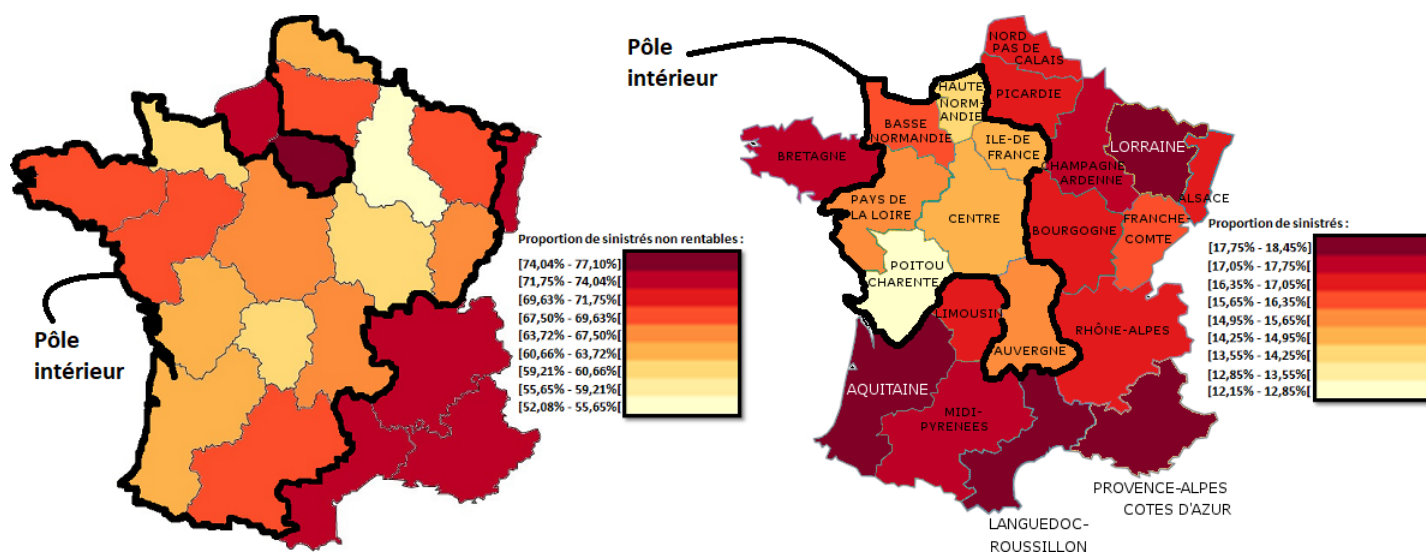


FIGURE IV.1 – Cartographie des 21 régions en fonction de la proportion de sinistrés non rentables (gauche) et, à titre de rappel ([partie III](#)), en fonction de la proportion de sinistrés (droite).

Il est important de relever que certaines régions (IdF, Alsace, PACA, Languedoc-Roussillon,

Rhône-Alpes, Midi-Pyrénées, Lorraine et Bretagne) sont des zones qui méritent une attention toute particulière à cause d’une conjonction de risque — risque élevé de survenance d’au moins un sinistre incapacité, et risque accru de prestations élevées relativement aux primes — néfastes pour l’assureur, et ceci couplé à une population généralement nombreuse (effet volume surtout pour IdF, PACA et Rhône-Alpes).

Mais la zone spatiale n’est pas la variable explicative dont l’écart en points de pourcentage (8.8 points) entre les sinistrés rentables et ceux non rentables est le plus probant. En effet, si l’on se réfère au tableau IV.4 suivant, l’âge moyen au sinistre est la variable à deux modalités qui sépare le mieux les deux populations (environ 19.6 points de pourcentage d’écart).

VARIABLES	MODALITÉS	EFFECTIF	NON RENTABLES
Zone Spatiale	Pôle intérieur/Arc extérieur	47.7%	64.3% / 52.3% 73.1%
Convention	professions médicales	05.5%	72.6%
	professions paramédicales	28.0%	74.4%
	profession libérales	07.4%	78.0%
	artisans - commerçants	36.2%	64.0%
	professions agricoles	10.0%	53.2%
	experts - conseils	06.0%	83.0%
	salariés	06.9%	70.1%
Sexe	femme/homme	51.8%	74.9% / 48.2% 62.4%
Situation de Famille	en couple/seul(e)	51.6%	65.8% / 48.4% 72.1%
Âge à la souscription	<=37 ans	58.0%	76.8%
	>37 ans	42.0%	58.0%
Âge Moyen Sinistre	<=40 ans / >40 ans	61.3%	76.5% / 38.7% 56.9%
Abattement tarifaire	oui/non	25.3%	75.8% / 74.7% 66.5%
Catégorie Professionnelle	1	57.2%	75.0%
	2	35.3%	64.0%
	3	07.5%	45.6%
Surprimés	oui/non	06.1%	63.7% / 93.9% 69.2%
Agents	oui (1)/non (0)	43.2%	63.3% / 56.8% 73.1%
Courtiers	oui (1)/non (0)	53.9%	73.2% / 46.1% 63.9%
Réseaux divers	oui (1)/non (0)	03.5%	72.2% / 96.5% 68.8%
Périodicité prime	mensuelle/non mensuelle	89.3%	69.7% / 10.7% 62.0%
Frais d’adhésion	oui/non	06.3%	67.3% / 93.7% 69.0%
Fiscalité	Agricole	05.5%	52.2%
	Assurance vie	12.7%	72.4%
	Madelin	81.8%	69.4%
Prime annuelle moyenne TTC (€)	<=722 € / >722 €	35.0%	80.9% / 65.0% 62.4%
Proportion (prime) Incapacité	<=48% / >48%	32.8%	72.9% / 67.2% 67.0%
Proportion (prime) Décès	<=24% / >24%	59.8%	70.3% / 40.2% 66.8%
Proportion (prime) Invalidité	<=28% / >28%	69.7%	65.9% / 30.3% 75.8%

TABLE IV.1 – Variables finales utilisées pour la modélisation - La proportion de sinistrés non rentables du portefeuille de sinistrés est de 68.89%.

D’ailleurs, l’âge moyen à la survenance d’un sinistre est la variable la plus liée à la variable cible en se basant sur le V de Cramer de la figure IV.2, puisque ce dernier vaut légèrement plus de 20%, ce qui constitue un fort pouvoir discriminant. De même pour l’âge à la souscription, et dans une moindre mesure le montant annuel moyen de prime toutes garanties confondues, la CP, la convention et le sexe, ces quatre dernières variables ayant un bon pouvoir discriminant

(V de Cramer entre 10% et 20%).

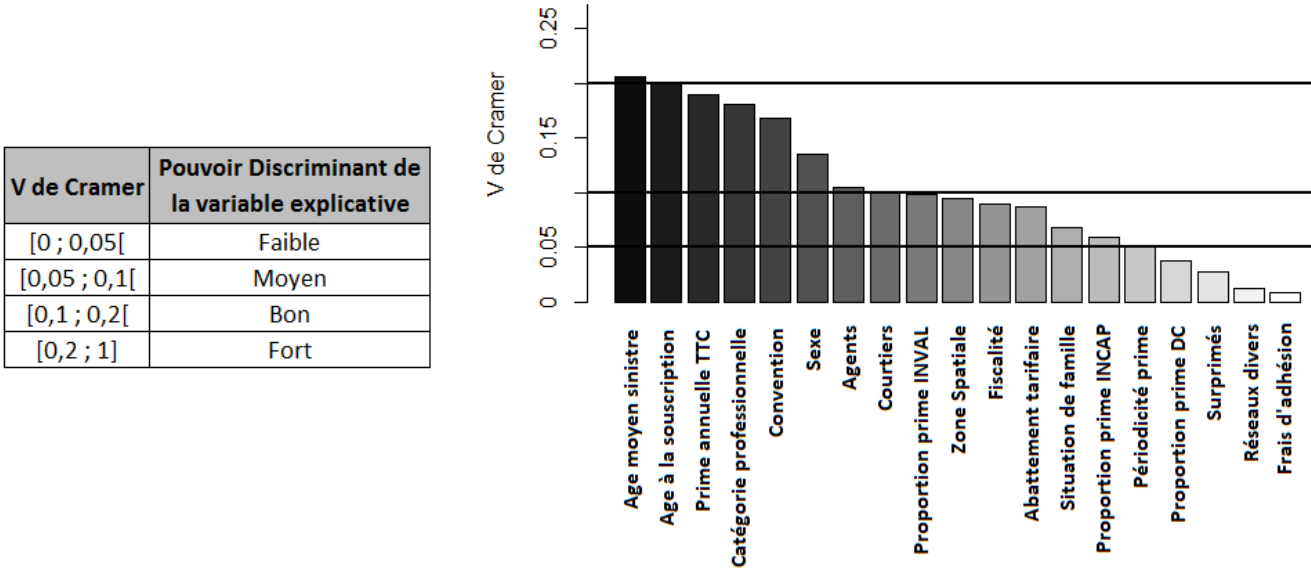


FIGURE IV.2 – *V de Cramer entre les variables explicatives et la variable cible binaire (rentabilité).*

Il est toutefois intéressant de constater que la population des surprimés présente une proportion de sinistrés non rentables inférieure à celle des non surprimés. Cela signifie que leur tarification est correctement (sur-)ajustée à leur exposition au risque, avec une marge de prudence.

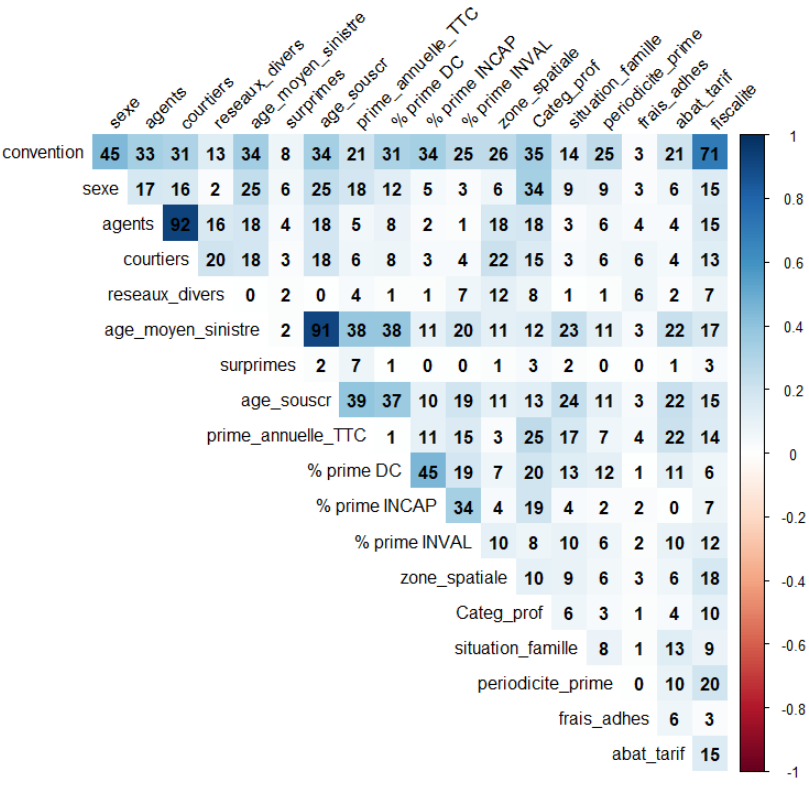


FIGURE IV.3 – *V de Cramer (en %) entre variables explicatives de la rentabilité, toutes qualitatives.*

S'agissant du V de Cramer de la figure IV.3 entre couples de variables explicatives, trois d'entre eux ont un score dépassant les 70% : ceux des couples de variables {agents, courtiers}, {âge moyen sinistre, âge à la souscription} et {fiscalité, convention} valent respectivement 92%, 91% et 71%.

Au vu de la forte liaison entre l'âge moyen à la survenance du sinistre et l'âge à la souscription, nous décidons donc de retirer de l'étude la variable relative à l'âge à la souscription et de conserver l'âge moyen lors d'un sinistre. En revanche, en l'absence d'interprétation valide de la liaison entre les variables « agents » et « courtiers », toutes les indicatrices des réseaux commerciaux sont conservées. Quant à la fiscalité et la convention, elles sont également maintenues pour la modélisation bien qu'une valeur du V de Cramer de 71% soit élevée.

Afin de restituer sous forme visuelle la liaison entre toutes les variables explicatives internes retenues et la variable cible, nous opérons l'AFCM suivante :

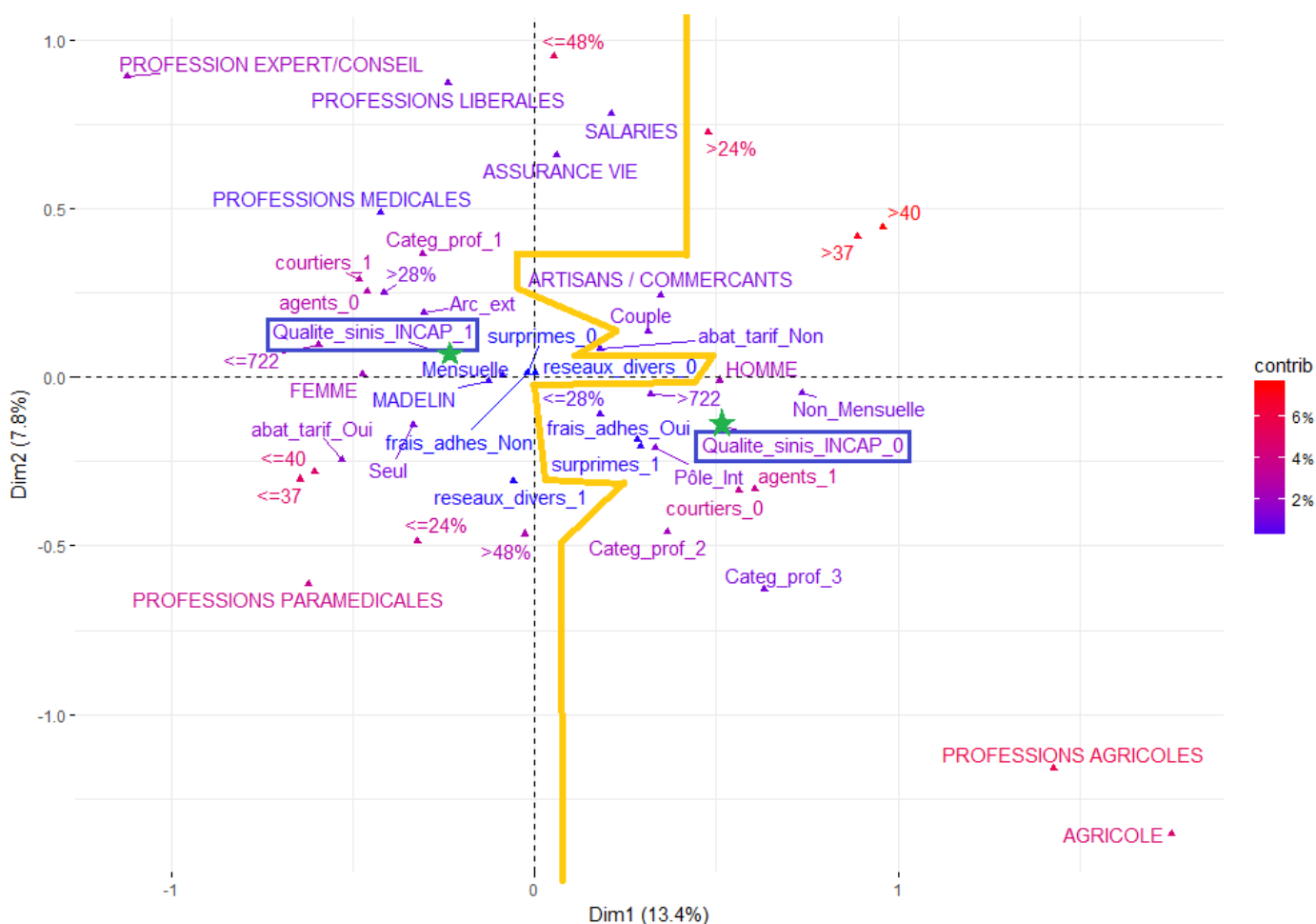


FIGURE IV.4 – Représentation des 45 modalités des variables explicatives internes et des 2 modalités (★) de la variable cible sur le premier plan de l'AFCM, avec une coloration en fonction de leur contribution (en termes d'inertie) à ce plan. La frontière orange sépare les deux clusters obtenus par k-means.

Le premier plan factoriel résume 21.2% de l'inertie totale du nuage des sinistrés, ce qui n'est

pas assez élevé pour donner du poids à l'interprétation que nous allons en faire. Ainsi, les deux modalités encadrées sur la figure IV.4 précédente et dont l'emplacement exact est signalé par une étoile verte, s'opposent sur le premier axe : ce facteur permet donc à priori de discriminer le phénomène d'intérêt qui est la rentabilité ou non des sinistrés. Ainsi, les experts-conseils sont par exemple ceux dont l'abscisse négative est la plus grande en valeur absolue. Ils sont donc les plus caractéristiques des non rentables qui y occupent 83% de l'effectif de la profession : c'est la plus grande proportion de sinistrés non rentables du tableau IV.1 de la page 77. Notons tout de même qu'avec un effectif de 6% de la population des sinistrés, la modalité experts-conseils souffre d'une contribution inertielle relativement modérée.

De plus, une meilleure tarification des assurés de moins de 40 ans serait souhaitable puisque la modalité relative à un âge moyen au sinistre inférieur à 40 ans est proche de la modalité $\{Y = 1\}$ de la variable cible. Ce constat est confirmé par la modalité de l'âge à la souscription inférieur à 37 ans (rappelons que cette variable a été retirée de l'étude, cf. supra). Enfin, les sinistrés non rentables semblent également caractérisés entre autres par une prime annuelle moyenne TTC inférieure à 722 €, par un sexe féminin, une souscription au travers d'un courtier, une profession médicale ou paramédicale, un célibat et un abattement tarifaire créateur. Enfin, notons que les sinistrés CP 2 et 3 ainsi que les agricoles ont des modalités respectives qui ont des coordonnées positives sur l'axe 1 tout comme la modalité $\{Y = 0\}$ de la variable cible, ce qui est paradoxal à première vue mais qui témoigne d'une tarification appropriée pour tenir compte de l'exposition élevée aux causes de survenance d'incapacité.

La frontière en orange sur la figure IV.4, obtenue par la méthode d'apprentissage non supervisé k -means avec $k = 2$ classes, est verticale et sépare les deux modalités de la variable cible ce qui semble donc confirmer que c'est l'axe 1 qui sépare le mieux les sinistrés rentables des sinistrés non rentables. La méthode k -means consiste d'abord à fixer le nombre k de groupes souhaités : dans notre cas $k = 2$ car nous souhaitons polariser les modalités des variables explicatives autour des 2 modalités de la variable cible.

Ensuite, il s'agit d'initialiser les noyaux des deux classes qui sont naturels dans notre cas : les deux modalités de la variable cible. C'est ensuite que commence l'algorithme de type « réallocation dynamique » puisqu'il va consister, pour chaque itération, en premier lieu à affecter toutes les modalités des variables explicatives à la classe la plus proche (celle dont le noyau est le plus proche de la modalité à classer), et en second lieu à mettre à jour les noyaux des classes comme étant leurs barycentres respectifs. Ces noyaux de classes étant mis à jour à chaque itération, il est donc possible que d'une itération à l'autre les modalités ne soient pas affectées à la même classe dans la mesure où elles vont itérativement plus ou moins s'éloigner des barycentres mouvants (d'où l'idée de « réallocation dynamique »). L'algorithme s'arrête lorsque la classification (i.e. l'assignation de chaque modalité à une des deux classes) n'est plus modifiée.

2 - Pouvoir explicatif des modèles

La figure IV.5 suivante de l'arbre optimal nous indique que l'âge moyen atteint lors des sinistres incapacité qui permet la meilleure dichotomie des sinistrés est de 40 ans, et peut être interprété à la lumière de la figure B.4a en annexe à l'origine de la discrétisation de l'âge à la souscription et au vu de la figure B.5 relative à l'ancienneté (variable non utilisée dans la modélisation de cette partie IV), comme étant 37 ans d'âge à la souscription et de 3 ans d'ancienneté, puisque ces deux valeurs sont les seuils qui séparent le mieux les sinistrés rentables des non rentables. Bien que l'ancienneté n'a pas retenue dans cette partie (l'âge moyen au sinistre fait office de variable temporelle), une ancienneté de 3 ans semble se confirmer ici en tant que marqueur temporel d'une dégradation globale (survenance et rentabilité) de la sinistralité.

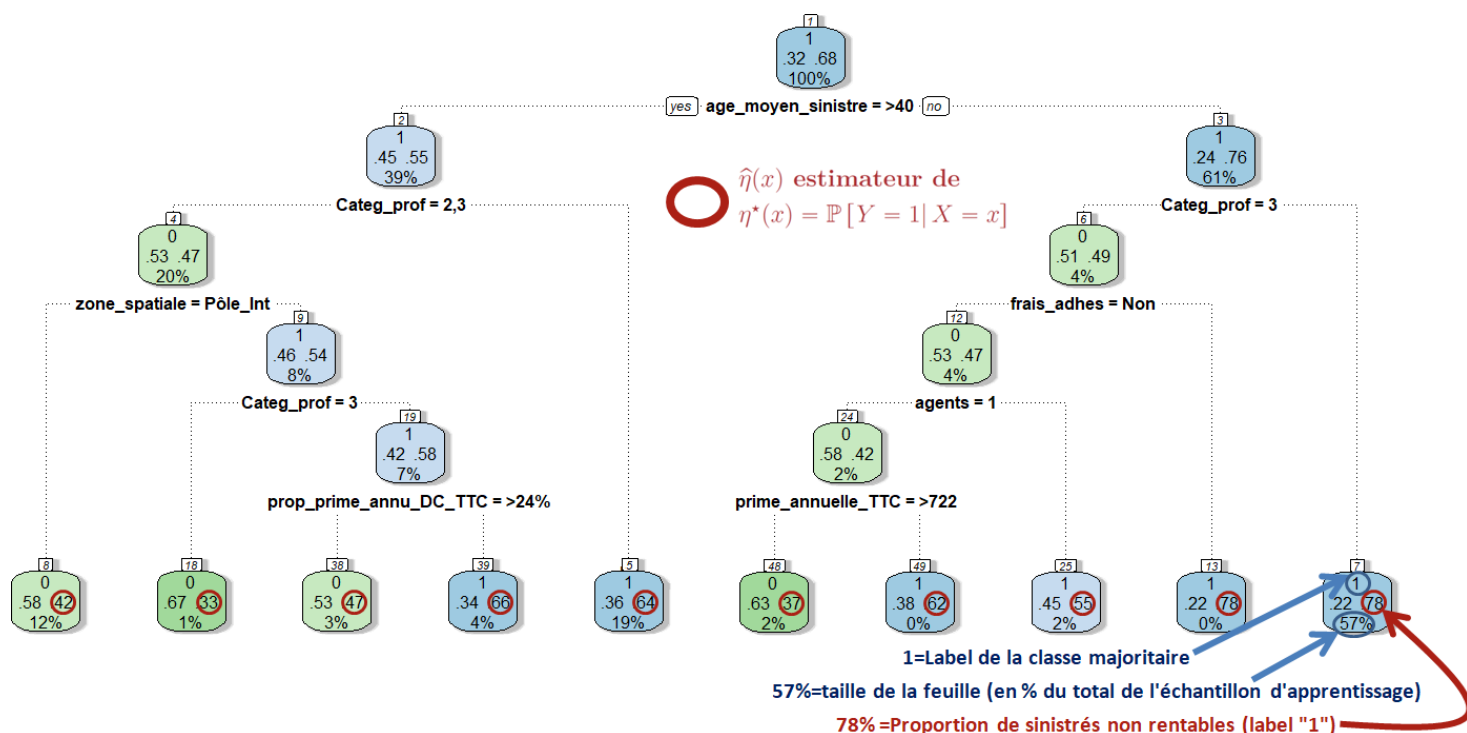


FIGURE IV.5 – Arbre optimal (9 scissions) prédisant la rentabilité.

Notons l'obtention d'un arbre plus étoffé qu'en partie III, avec l'apparition de la catégorie professionnelle qui est donc importante dans l'explication du coût élevé des sinistres incapacité. Ainsi, la feuille la plus volumineuse (57% des sinistrés du portefeuille) contient majoritairement des sinistrés non rentables (78%) caractérisés par un âge moyen lors des sinistres en deçà de 40 ans et une CP de niveau 1 ou 2 (les moins risquées). Comme nous l'avons déjà dit, ce constat paradoxal (CP 1 et 2 moins rentables que CP3) signifie que la tarification de la CP 3 tient correctement compte de l'exposition aggravée au risque incapacité et à un coût élevé de celle-ci. Le même constat peut être fait pour les sinistrés d'âge moyen de sinistre supérieur à 40 ans et qui sont en CP1 (19% des sinistrés du portefeuille) puisqu'ils comptent 64% de sinistrés non rentables, alors que ceux qui sont en CP 2 ou 3 ont environ 47% de leur effectif qui est non rentable.

Cependant, cet arbre optimal n'est pas celui qui minimise l'erreur d'apprentissage, dont le minimiseur est l'arbre complet. La méthode adaboost reste ici aussi une très bonne méthode d'apprentissage, avec une erreur $R_n^{appr}=11.52\%$:

Modèles	Erreur d'apprentissage (R_n^{appr})
Arbre complet	0.1133
Arbre optimal	0.2896
Forêts aléatoires	0.2822
Adaboost	0.1152
Régression logistique sans interactions	Erreur d'apprentissage (R_n^{appr})
complet	0.2967
sélection de variables bi-directionnelle	0.2969
Régression logistique avec interactions	Erreur d'apprentissage (R_n^{appr})
complet	0.5131
sélection de variables bi-directionnelle	0.2756

TABLE IV.2 – Moyennes des erreurs d'apprentissage obtenues en réitérant 100 fois l'échantillonnage.

Pour l'arbre complet, parmi les variables dont l'importance est la plus grande, on trouve la convention (19.8%), la situation de famille (10.8%) et la catégorie professionnelle (10.4%) :

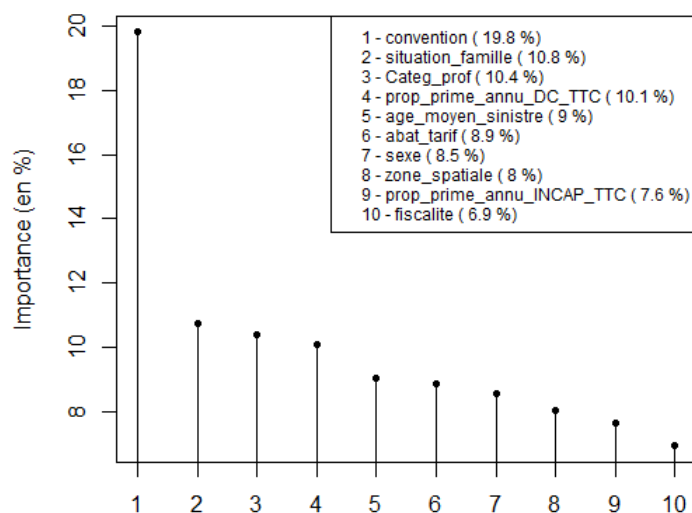


FIGURE IV.6 – Importance des variables (les 10 les plus importantes) expliquant la rentabilité dans le cadre d'une modélisation binaire par arbre CART pour l'arbre entier ($\gamma = 0$).

Au vu de l'amplitude du MDA et du MDE issus des forêts aléatoires, l'importance de la catégorie professionnelle et la convention semble se confirmer. On y trouve également en bonne place l'âge moyen de survenance de sinistres, le sexe ainsi que le montant de prime annuelle TTC :

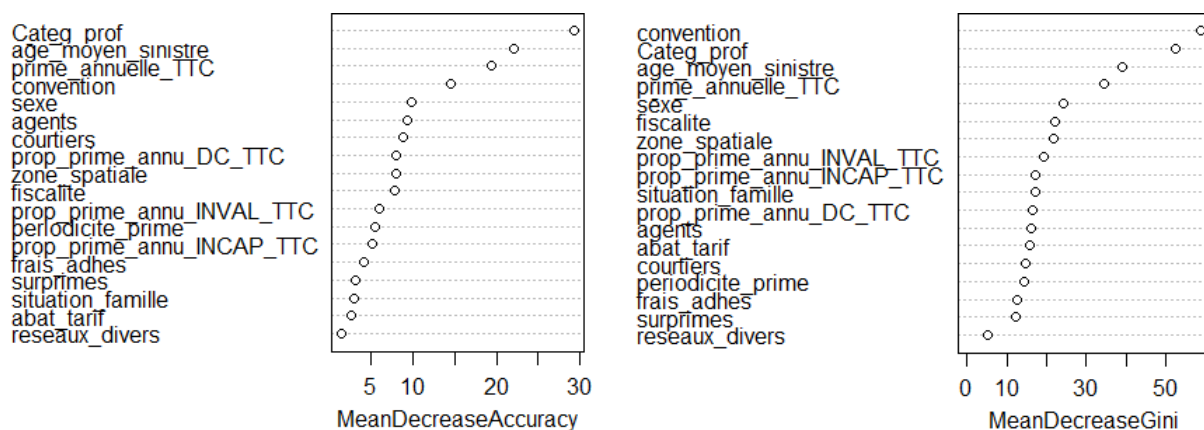


FIGURE IV.7 – Eboulis des MDA et MDE obtenus par forêts aléatoires (500 souches).

Le graphique suivant nous indique quant à lui que la rentabilité est plutôt bien équilibrée entre les différentes modalités d'une même variable, avec des odds ratios compris entre 0.5 et 2 sauf pour la catégorie professionnelle puisque les sinistrés CP 1 présentent une probabilité de non-rentabilité nettement plus grande que les sinistrés CP 2 ou 3 (odd ratios entre 2 et 3). Mis à part cette variable, les autres odds ratios sont proches de 1 et leurs intervalles de confiances contiennent la valeur 1 sauf pour la variable relative à la prime annuelle moyenne TTC et celle relative à l'âge moyen au sinistre qui nous informant qu'en dessous de 722 € de prime toutes garanties confondues ou qu'en dessous de 40 ans en moyenne, les sinistrés ont plus de chances d'être moins rentables que leurs complémentaires respectifs (i.e. plus de 722 € de prime annuelle moyenne TTC d'une part et plus de 40 ans d'âge moyen d'autre part).

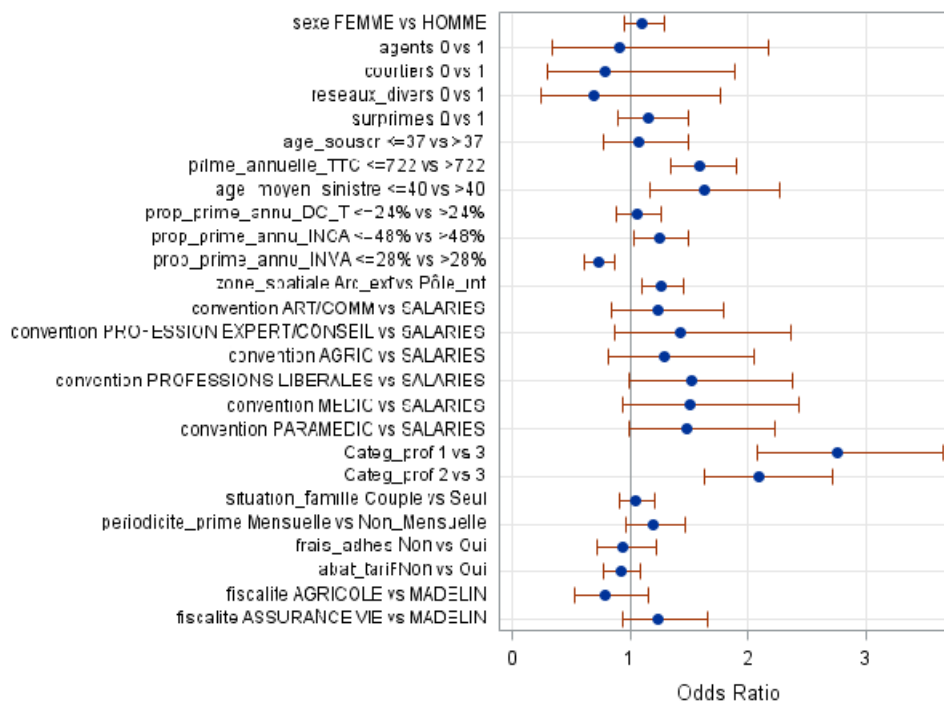


FIGURE IV.8 – Odd Ratios pour les 20 variables issues de la sélection de variables dans le cadre d'une modélisation sans interactions.

Estimations par MV du modèle après sélection bi-directionnelle de variables							
Variables	Modalités (indicatrices) associées aux paramètres		Estimations des paramètres pour les modalités des variables finalement retenues	Intervalle de confiance au niveau 95%		$\chi^2_{Wald} = \text{Valeurs observées de la statistique de test (khi-2 de Wald)}$	$\mathbb{P} \left[\chi^2(1) > \chi^2_{Wald} \right]$
Intercept			0,5117	0,3428	0,6839	35	<.0001
agents*zone_spatiale	0	Arc_ext	0,0992	0,0320	0,1663	8	0.0038
age_moyen*Categ_prof	<=40	1	0,0801	-0,1869	0,0264	2	0.1411
age_moyen*Categ_prof	<=40	2	0,1760	0,0646	0,2876	10	0.0020
age_moyen*fiscalite	<=40	AGRICOLE	0,1413	-0,3911	0,1079	1	0.2652
age_moyen*fiscalite	<=40	ASSURANCE VIE	0,2106	0,0272	0,3958	5	0.0250
prime_ann*Categ_prof	<=722	1	0,1414	-0,0042	0,2858	4	0.0553
prime_ann*Categ_prof	<=722	2	0,1524	-0,3077	0,0019	4	0.0531
prime_annu*fiscalite	<=722	AGRICOLE	0,3186	0,0518	0,5946	5	0.0209
prime_annu*fiscalite	<=722	ASSURANCE VIE	0,3179	-0,5110	-0,1274	11	0.0011
agents	0		0,1022	0,0332	0,1711	8	0.0037
age_moyen_sinistre	<=40		0,3187	0,1681	0,4695	17	<.0001
prime_annuelle TTC	<=722		0,2208	0,0418	0,4025	6	0.0162
prop_prime_annu_INVA	<=28%		0,1925	-0,2708	-0,1150	24	<.0001
zone_spatiale	Arc_ext		0,1091	0,0409	0,1773	10	0.0017
Categ_prof	1		0,5963	0,4506	0,7412	65	<.0001
Categ_prof	2		0,0837	-0,0687	0,2356	1	0.2800
fiscalite	AGRICOLE		0,1630	-0,3934	0,0781	2	0.1740
fiscalite	ASSURANCE VIE		0,0858	-0,0872	0,2582	1	0.3294

FIGURE IV.9 – Estimations $(\hat{\theta}_{MV}^k)_k$ par MV des paramètres $(\theta^k)_k$ de la régression logistique avec interactions modélisant la rentabilité binaire de la sinistralité après sélection bi-directionnelle de variables.

Le tableau IV.9 ci-dessus permet l'analyse des paramètres estimés du modèle avec interactions d'ordre 2 et avec sélection bi-directionnelle de variable selon le critère AIC. La **sélection de variables** trouve son utilité lorsque l'actuaire éprouve la nécessité d'écrêter les variables explicatives, afin de diminuer les combinaisons possibles d'interactions d'ordre 2. Ainsi, le critère de choix de modèles que l'on désigne par AIC (pour Akaike Information Criterion) est fondé sur le raisonnement suivant : l'ajustement du modèle est d'autant meilleur que la vraisemblance est grande. Or, la vraisemblance, et donc l'ajustement, augmente avec la complexité du modèle (i.e. le nombre de paramètres, i.e. le nombre de variables explicatives/de modalités/d'interactions). Par conséquent, choisir le modèle qui maximise la vraisemblance revient à choisir le modèle complet, mais ce sur-ajustement est néfaste pour le pouvoir prédictif (pouvoir de généralisation) du modèle.

Une stratégie pour choisir un modèle plus parcimonieux consiste à pénaliser la vraisemblance L_n par une fonction du nombre p de paramètres. L'AIC d'un modèle à p paramètres est par exemple défini par $AIC = -2L_n + 2p$. Il ne s'agit donc plus de choisir le modèle dont les paramètres maximisent la vraisemblance L_n mais plutôt celui dont les paramètres maximisent « -AIC ».

Quant à l'utilité du caractère bi-directionnel de la sélection, elle réside dans le fait qu'il faudrait en pratique, rien que pour une modélisation sans interactions, calculer 2^p critères AIC en partant de p variables explicatives puisque le nombre de modèles possibles à partir de p variables est 2^p . Cela afin de pouvoir ensuite choisir le modèle maximisant « -AIC ». Dans notre

cas d'environ 18 variables explicatives, il faudrait donc logiquement calculer $2^{18} = 262\,144$ AIC ! Le coût temporel d'une telle manière de procéder nous amène à privilégier une autre méthode pas à pas, qui va consister, à chaque pas, d'une part à ajouter (partie ascendante de la procédure de sélection bi-directionnelle) une variable au modèle dont l'ajout conduit à la plus grande maximisation du critère « -AIC » parmi tous les ajouts possibles de variables, et, d'autre part à éventuellement retirer (partie descendante de la procédure de sélection bi-directionnelle) du modèle des variables déjà introduites, pour peu que leur élimination augmente le critère cible « -AIC ». L'arrêt de la procédure a lieu lorsque toutes les variables sont intégrées ou lorsque qu'aucune variable ne permet d'augmenter davantage le critère cible « -AIC ».

Si l'on souhaite analyser le tableau IV.9 de la page 84, on constate que l'interaction la plus significativement non nulle (au risque de première espèce 5%) lors de l'explication de la non-rentabilité est celle mettant en jeu à la fois la prime annuelle moyenne TTC (inférieure à 722 €) et une fiscalité non agricole non madelin (« Assurance vie ») puisque la p-valeur est de 0.0011. L'estimation par MV du paramètre associé est -0.3179, ce qui signifie que les sinistrés vérifiant simultanément ces deux caractéristiques, financière et fiscale, ont moins de chances d'être non rentables pour l'assureur, contrairement par exemple à ceux qui paient également moins de 722 € de prime annuelle TTC mais qui, en revanche, ont une fiscalité agricole, puisque le paramètre estimé afférent est positif, et égal à 0.3186.

3 - Pouvoir de généralisation des modèles

Si Adaboost et l'arbre CART complet sont les méthodes qui optimisent l'erreur d'apprentissage R_n^{appr} , ce sont en revanche les Forêts aléatoires qui maximisent le Taux de Succès Global (TSG) à 70,54%, ou de manière équivalente, qui minimise l'erreur de prévision R_n^{test} à 29,46%.

Modèles	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Arbre complet	0.3126	0.6302	0.7364	0.6248	0.1116	0.5702
Arbre optimal	0.2013	0.6904	0.8723	0.7459	0.1264	0.6180
Forêts aléatoires	0.2243	0.7054	0.9470	0.8744	0.0726	0.6379
Adaboost	0.2205	0.6689	0.8206	0.6951	0.1255	0.6105
Régression logistique sans interactions	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.1973	0.7001	0.9128	0.8102	0.1026	0.6591
sélection de variables bi-directionnelle	0.1969	0.6988	0.9115	0.8117	0.0999	0.6557
Régression logistique avec interactions	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.4584	0.4934	0.4075	0.3004	0.1070	0.5535
sélection de variables bi-directionnelle	0.2083	0.6821	0.8480	0.7160	0.1320	0.6467

TABLE IV.3 – Moyennes des scores de prédiction obtenues en réitérant 100 fois l'échantillonnage.

C'est également par une modélisation par forêts aléatoires que le **TVP** (taux de vrais non rentables) est optimisé, culminant à 94,70%. Ces bonnes performances prédictives ne se font pas aux dépens des performances évaluées sur la base d'apprentissage puisque les forêts aléatoires,

malgré une erreur d'apprentissage de 28,22%, sont la quatrième méthode la plus efficace de ce point de vue. Toutefois, le Taux de Faux Non rentables (**TFP**) est très mauvais : 87,44% de fausses alertes.

Notons également les bonnes réalisations prédictives de la régression logistique sans interactions, dont quasiment tous les feux sont au vert : score de Brier, TSG, TVP et AUC. C'est d'ailleurs ce dernier score qui peut être apprécié sur la figure IV.10 suivante :

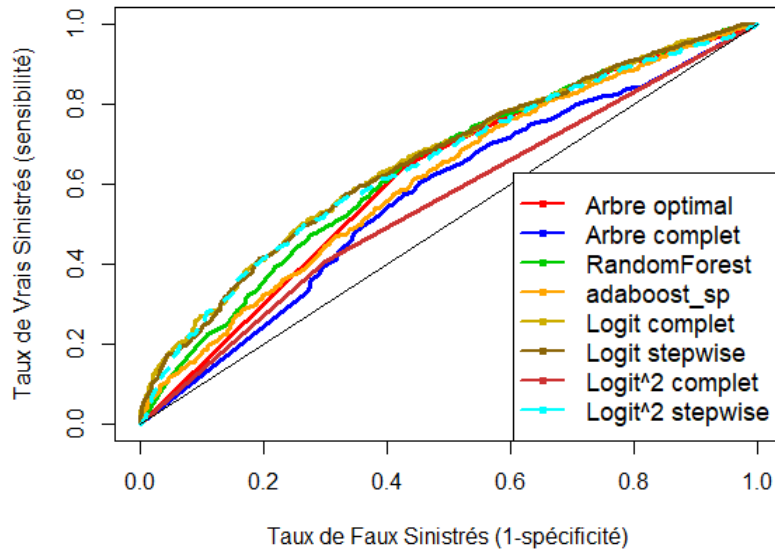


FIGURE IV.10 – Courbes ROC - Modélisation de la (non-)rentabilité.

4 - Contribution des données externes

Toujours dans une tentative d'apport d'informations supplémentaires à travers des sources externes, la figure IV.11 suivante met en évidence les variables externes dont la liaison avec la variable cible, quantifiée par $S_{Y/X}$, est forte. Rappelons que $S_{Y/X} := \sqrt{\frac{\sigma_E^2}{\sigma_E^2 + \sigma_R^2}}$, où $\sigma_E^2 = \frac{1}{n} \sum_{l=1}^2 n_l \times (\bar{y}_l - \bar{y})^2$ est la variance inter-classes et $\sigma_R^2 = \frac{1}{n} \sum_{l=1}^2 n_l \times \sigma_l^2$ est la variance intra-classes, les deux classes « l » étant les sinistrés rentables et les sinistrés non rentables¹.

Il en sort que les espérances de vie à la naissance et à 60 ans, ainsi que le nombre de personnes par ménage, sont les 3 variables les plus fortement liées à la variable à expliquer. Ces liaisons sont confirmées par le test de WMW, qui est un test non paramétrique de localisation basé sur les rangs, dont l'hypothèse H_0 d'égalité des rangs moyens dans chacune des deux classes de sinistrés $\{Y = 0\}$ et $\{Y = 1\}$ est rejetée au risque de première espèce 5%.

Il y a donc bien une différence significative de position/localisation de la distribution des trois variables externes considérées dans chacune des deux populations.

1. Les notations n_l , \bar{y}_l et σ_l^2 désignent respectivement l'effectif, la moyenne et la variance des valeurs de Y pour les sinistrés de la classe $l = \{\text{rentables, non rentables}\}$.

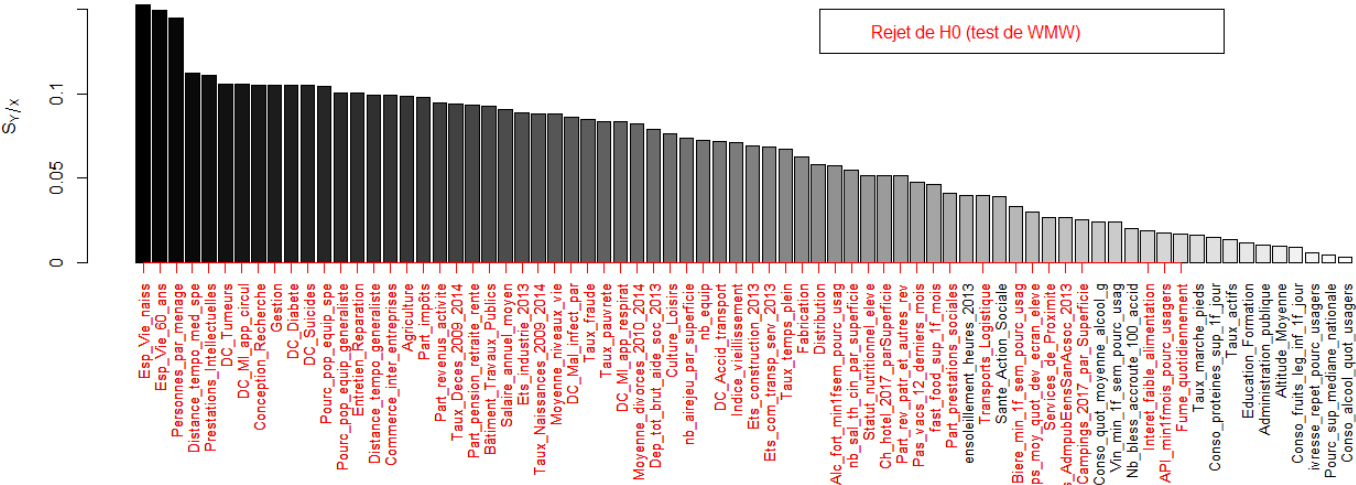


FIGURE IV.11 – $S_{Y/X}$ entre les variables externes toutes quantitatives et la variable cible binaire, et mise en évidence en rouge d’une liaison significative (rejet de l’hypothèse H_0 du test de WMW) au risque de première espèce $\alpha = 5\%$.

Les deux espérances de vie étant relativement proches en signification, nous décidons de n’en garder qu’une seule : l’espérance de vie à la naissance. Les graphiques B.6a et B.6b en annexe page 95 aident à savoir comment ont été fixés les seuils suivants ayant permis un découpage en deux classes des deux variables externes retenues :

VARIABLES	MODALITÉS	EFFECTIF	NON RENTABLES
Espérance de vie à la naissance	≤ 83 ans / > 83 ans	50.1%	62.5% / 49.9% / 75.3%
Personnes par ménage	≤ 2 / > 2	37.9%	77.5% / 62.1% / 63.6%

TABLE IV.4 – Variables externes utilisées pour la modélisation - La proportion de sinistrés non rentables du portefeuille de sinistrés est de 68.89%.

Comparativement à la partie III, l’apport des données externes est moins décevant puisqu’elles permettent de faire simultanément augmenter le taux de vrais non rentables **TVP** des quatre méthodes issues des arbres CART (arbre complet, arbre optimal, forêts aléatoires et Adaboost) comme on peut l’observer sur le tableau suivant :

Modèles d'apprentissage	Apprentissage	Test					
	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
Arbre complet	0.0985	0.3137	0.6381	0.7458	0.6203	0.1254	0.6224
Arbre optimal	0.2967	0.2017	0.6996	0.9097	0.8042	0.1055	0.6617
Forêts aléatoires	0.2808	0.2214	0.7040	0.9514	0.8894	0.0620	0.6311
Adaboost	0.1082	0.2173	0.6777	0.8555	0.7489	0.1066	0.6437
Régression logistique sans interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.2949	0.1971	0.7027	0.9128	0.8012	0.1116	0.7053
sélection de variables bi-directionnelle	0.2936	0.1970	0.7023	0.9184	0.8161	0.1022	0.7048
Régression logistique avec interactions	R_n^{appr}	BS	TSG ($= 1 - R_n^{test}$)	TVP	TFP	PSS	AUC
complet	0.2624	0.2079	0.6759	0.8243	0.6801	0.1442	0.7117
sélection de variables bi-directionnelle	0.3609	0.3108	0.6574	0.8798	0.8759	0.0038	0.7122

TABLE IV.5 – Moyennes des scores de performance obtenues en réitérant 100 fois l’échantillonnage - Seules 2 variables externes sont incluses.

Conclusion

L'incapacité est un risque central en prévoyance. En effet, le décès présente un aléa connu et plutôt maîtrisé (tables de mortalité), dont la tarification contrôlée permet donc d'avoir un S/P inférieur à 100%. Quant à l'invalidité, l'incidence est beaucoup plus faible que l'incapacité, ce qui fait de cette dernière un risque avec un enjeu financier moindre. L'incapacité présente donc un fort enjeu en prévoyance, dont la tarification doit également tenir compte du positionnement concurrentiel.

Ainsi, afin d'optimiser la tarification, il est indispensable d'avoir conscience des variables qui « portent » le risque, ou, plus précisément, qui permettent de discriminer d'une part les assurés à plus fort risque des assurés à moins fort risque, et d'autre part les sinistrés dont la rentabilité est dégradée des sinistrés rentables. Cette prise de conscience ne peut que passer par des tentatives de recherche et d'exploration de variables, qui peuvent aboutir à des variables discriminantes de la sinistralité. Cependant, dans le cas où certaines variables étudiées ne sont pas discriminantes, elles sont tout de même informatives puisqu'elles nous invitent à parcourir d'autres voies et à chercher la réponse ailleurs.

C'est en tout cas l'enseignement de ce mémoire puisqu'il a permis de se rendre compte qu'une tarification qui se fait à l'âge à la souscription suivie de majorations annuelles, uniquement selon la convention (i.e. le secteur d'activité) et les 3 niveaux de risques de la catégorie professionnelle, ne permet pas de prendre finement en considération l'aggravation du risque incapacité liée à l'écoulement du temps et l'accroissement des âges. Par conséquent, rajouter l'ancienneté en portefeuille dans la grille de majoration permet de pallier ce problème, sans toutefois alourdir les règles de majoration et démultiplier les grilles de tarifs, puisqu'une ancienneté en portefeuille de 3 ans semble être un marqueur temporel discriminant les « bons » des « mauvais » assurés. Ajouter un supplément de prime une fois cette durée passée permettra donc d'améliorer la rentabilité du portefeuille en faisant baisser la proportion de sinistrés. Notons toutefois que la durée en portefeuille vue à fin 2016 fait l'objet d'une censure à droite : c'est une variable dont on ne connaît pas la valeur pour tous les assurés puisque certains sont encore en cours à fin 2016. Citons également un autre marqueur temporel pour la rentabilité des sinistrés : il s'agit d'un âge moyen atteint à la survenance des sinistres incapacité égal à 40 ans, qui discrimine donc les plus jeunes, qui sont aussi les moins rentables, des moins jeunes, pour lesquels la prime couvre correctement les prestations. Une plus grande vigilance et une prise en compte de cet aspect dans la tarification signifiera une amélioration de la rentabilité du portefeuille.

Également, l'ajout de critères comme la zone géographique et la situation familiale semblent judicieux dans l'optimisation de la rentabilité. Aussi, le maintien de la sélection médicale et

de la prise en compte d'activités sportives à risque (faisant l'objet d'une surprime) semblent s'imposer, alors même que se pose actuellement la question de l'utilité de telles sélections et de leurs coûts (édition des questionnaires,...). En ce qui concerne le montant de prime investi par l'assuré, le constat d'une proportion plus élevée de sinistrés chez les assurés qui investissent le plus nous amènent à suggérer à l'assureur de faire en sorte que le montant de prime ne soit pas une fonction linéaire du montant garanti à partir d'un certain seuil de prime, afin de dissuader les « mauvais » risques de souscrire, notamment ceux qui ont un besoin justifié d'avoir un haut niveau de couverture du fait de leur exposition, et les amener à se tourner vers la concurrence. Une proportion de prime annuelle (toutes garanties prévoyance confondues) allouée à l'incapacité supérieure à 46% doit notamment alerter l'assureur sur une augmentation de la probabilité de sinistre, quelles qu'en soient les motivations ou la cause (blessures physiques invalidantes, aléa moral,...) surtout si cette prime annuelle (toutes garanties confondues) dépasse environ 900€. Quant à la discrimination selon le sexe, bien qu'interdite par la réglementation, elle semble être une revendication légitime des assureurs puisqu'une différence significative sépare en proportions les sinistrés des non sinistrés et les sinistrés non rentables des sinistrés rentables selon les deux sexes (en faveur des assurés de sexe masculin). En revanche, certaines variables comme le réseau commercial ayant permis l'affaire ne sont pas pertinentes dans la discrimination de la sinistralité incapacité sous les deux formes étudiées (occurrence d'au moins un sinistre, et rentabilité binaire).

Enfin, l'apport des variables externes que sont l'espérance de vie à la naissance et le nombre de personnes par ménage a permis d'améliorer légèrement mais simultanément le pouvoir prédictif des modèles d'apprentissage issus des arbres CART et modélisant la rentabilité binaire, ce qui est assez encourageant puisque ces résultats suggèrent qu'une meilleure compréhension de la sinistralité semble bien résider dans l'environnement des assurés. Le choix qui a été fait de ne retenir que quelques variables externes afin de ne pas accroître considérablement le temps de calcul et afin de permettre une comparaison des pouvoirs explicatif et prédictif avec la régression logistique dont la complexité augmente avec le nombre de variables, ainsi que le choix d'une maille de jointure imposée par les données externes qui est plus grossière que celle des bases de données de l'entreprise, font que la portée de l'apport de ces données externes est limité en intensité. Toutefois, cela ouvre la voie à une nouvelle façon de tarifier, moins classique, basée dans une certaine mesure sur des caractéristiques environnementales larges (économiques, sociales,...) ou des habitudes comportementales (alimentation, activité physique,...) à recueillir par l'assureur notamment par le biais des nouvelles technologies, et pouvant par exemple prendre la forme d'abattements tarifaires et de surprimes. Les seuls freins à ce type d'innovation sont d'ordres éthiques (confidentialité des données personnelles recueillies,...) et sociétaux (mentalités,...).

En guise d'approfondissement de ce mémoire, une étude de rentabilité (non binaire) de sinistre et de fréquence sont nécessaires pour avoir l'ensemble des éléments nécessaires à une tarification sur le modèle coût-fréquence usuels en assurance de biens.

ANNEXES

A - Excès de risque en classification binaire

Comme vu en page 23, $\phi^\star = \mathbb{1}_{\eta^\star(x) > \frac{1}{2}}$, et comme défini en page 24 : $\varepsilon(\hat{\phi}) := R(\hat{\phi}) - R(\phi^\star)$.
On a : $\ell[\hat{\phi}(\mathbb{X}), Y] = \mathbb{1}_{Y \neq \hat{\phi}(\mathbb{X})} = \left(Y - \hat{\phi}(\mathbb{X})\right)^2$ dans le cas binaire.

$$\begin{aligned}\forall \hat{\phi} \in \{0, 1\}^{\mathcal{X}}, \quad R(\hat{\phi}) &= \mathbb{E} \left[\ell[\hat{\phi}(\mathbb{X}), Y] \right] \\ &= \mathbb{E} \left[\left(\hat{\phi}(\mathbb{X}) - Y \right)^2 \right] \\ &= \mathbb{E} \left[\hat{\phi}(\mathbb{X})^2 \right] + \mathbb{E} \left[Y^2 \right] - 2 \times \mathbb{E} \left[Y \times \hat{\phi}(\mathbb{X}) \right] \\ &= \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] + \mathbb{E} \left[Y \right] - 2 \times \mathbb{E} \left[\mathbb{E} \left[Y \times \hat{\phi}(\mathbb{X}) \right] \middle| \mathbb{X} \right] \\ &= \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \right] + \mathbb{E} \left[Y \right] - 2 \times \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \times \mathbb{E} \left[Y | \mathbb{X} \right] \right] \\ &= \mathbb{E} \left[Y \right] + \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \times (1 - 2 \times \mathbb{E} \left[Y | \mathbb{X} \right]) \right]\end{aligned}$$

On a : $Y = \mathbb{1}_{Y=1}$ donc $\mathbb{E} \left[Y | \mathbb{X} \right] = \mathbb{P} \left[Y = 1 | \mathbb{X} \right] = \eta^\star(\mathbb{X})$
donc

$$\begin{cases} R(\hat{\phi}) = \mathbb{E} \left[Y \right] + \mathbb{E} \left[\hat{\phi}(\mathbb{X}) \times (1 - 2 \times \eta^\star(\mathbb{X})) \right] \\ R(\phi^\star) = \mathbb{E} \left[Y \right] + \mathbb{E} \left[\phi^\star(\mathbb{X}) \times (1 - 2 \times \eta^\star(\mathbb{X})) \right] \end{cases}$$

donc $\boxed{\varepsilon(\hat{\phi}) = \mathbb{E} \left[\left(\hat{\phi}(\mathbb{X}) - \phi^\star(\mathbb{X}) \right) \times (1 - 2 \times \eta^\star(\mathbb{X})) \right]}$

B - Distribution des variables utilisées pour la modélisation de la rentabilité

1 - Variables discrètes

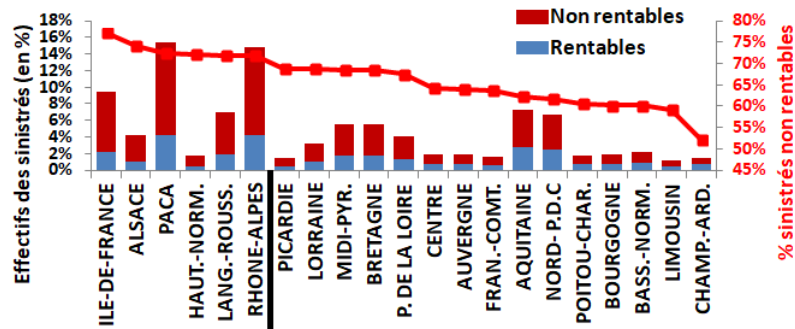


FIGURE B.1 – Histogramme des 21 régions en fonction de la proportion de sinistrés non rentables

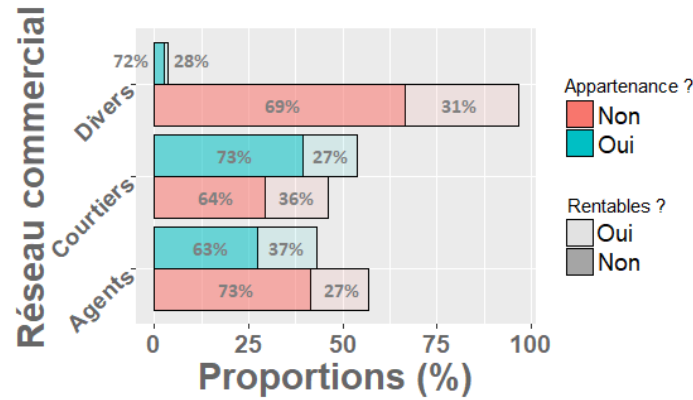
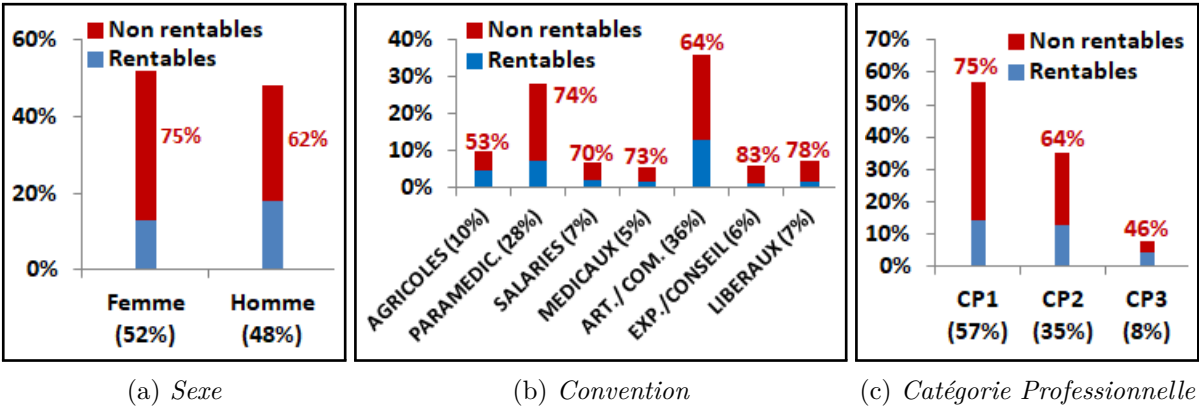
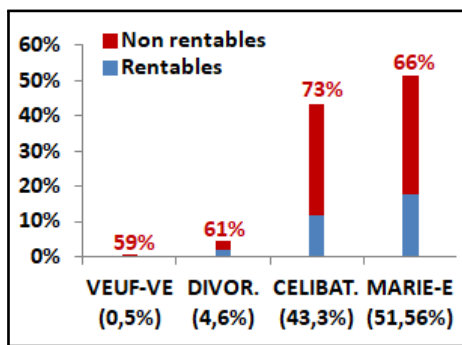
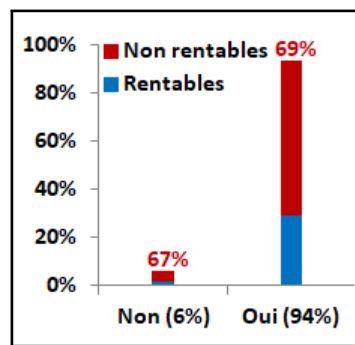


FIGURE B.2 – Proportion de sinistrés non rentables selon leur appartenance aux réseaux commerciaux

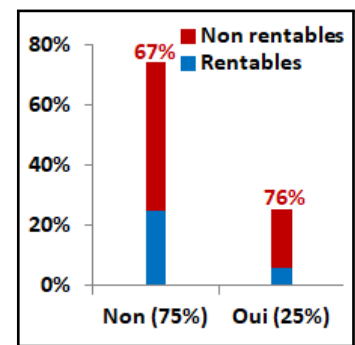




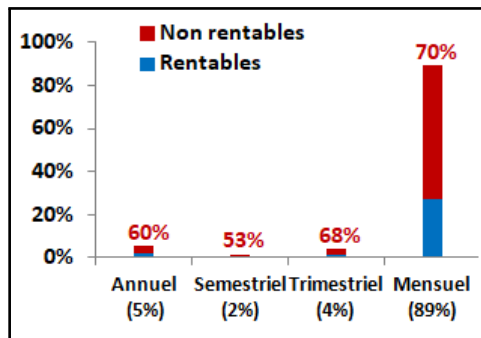
(d) Situation de famille



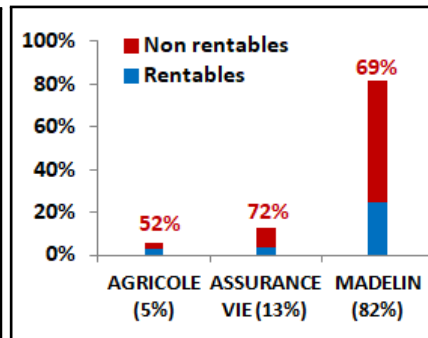
(e) Frais d'adhésion



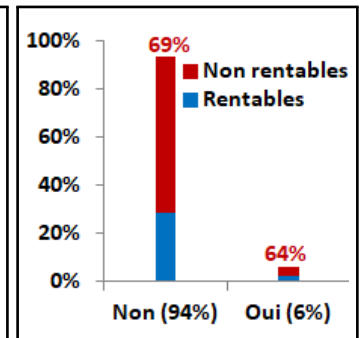
(f) Abattement tarifaire



(g) Périodicité de prime



(h) Fiscalité

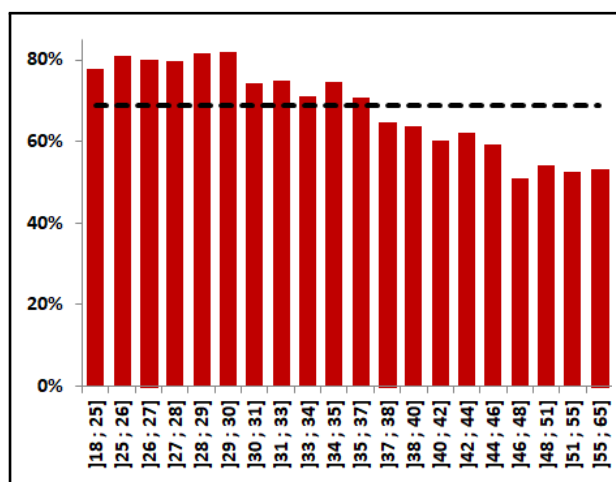


(i) Surprimés

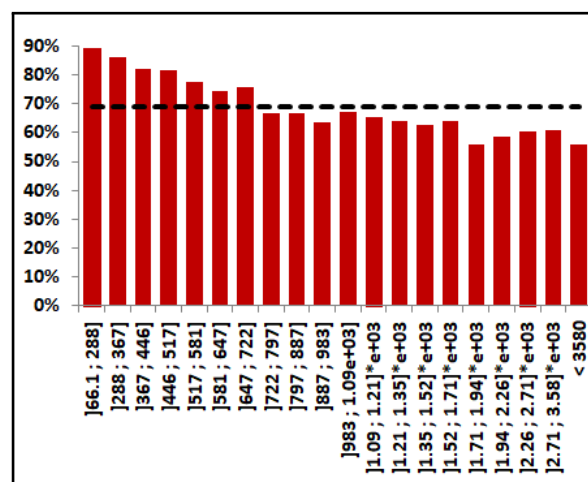
FIGURE B.3 – Distributions de 9 variables explicatives catégorielles selon la proportion de sinistrés non rentables dans chaque modalité

2 - Variables continues

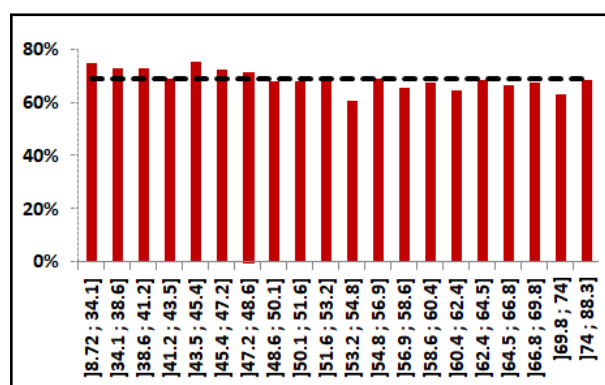
2.1 - Variables internes



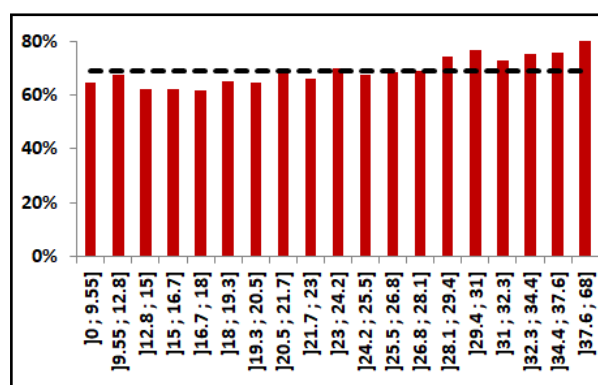
(a) Âge à la souscription



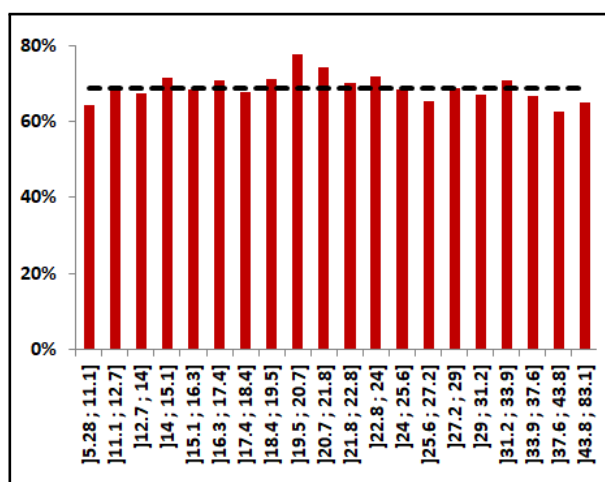
(b) prime annuelle moyenne (€)



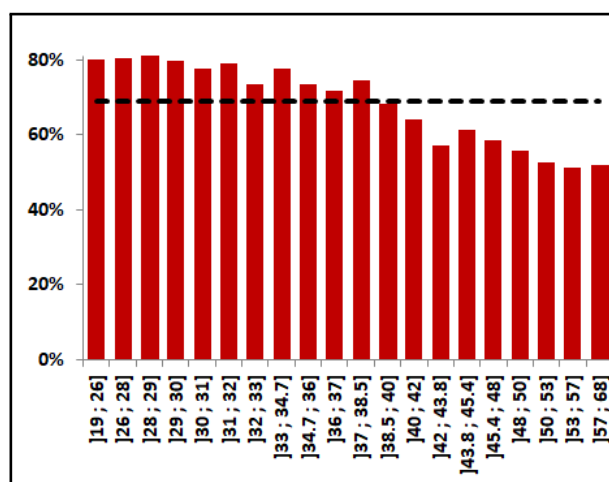
(c) Pourcentage de prime allouée à la garantie Incapacité



(d) Pourcentage de prime allouée à la garantie Invalidité

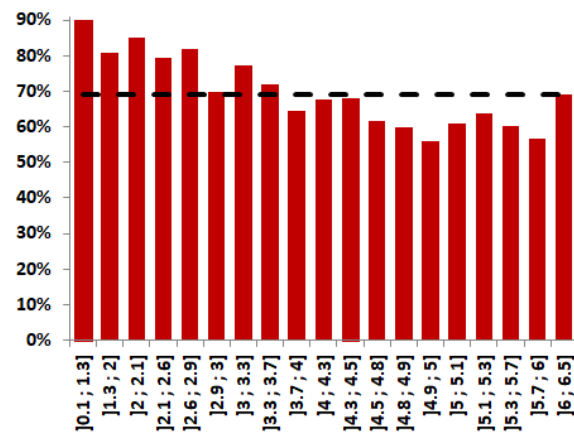


(e) Pourcentage de prime allouée à la garantie Décès

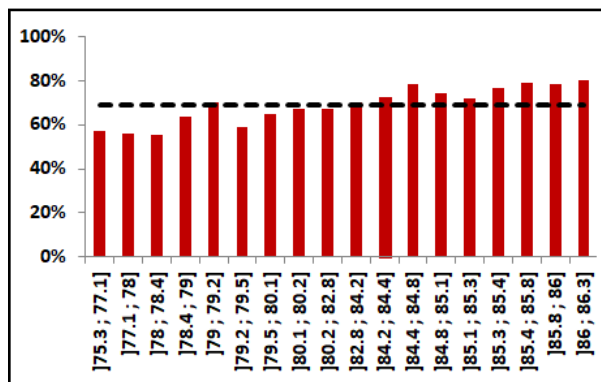
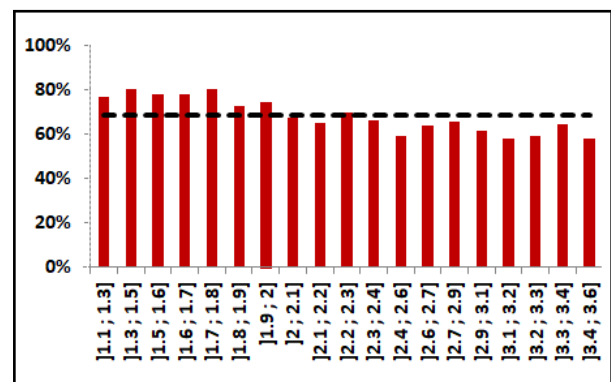


(f) Âge moyen lors de la survenance de sinistre

FIGURE B.4 – Histogrammes de la proportion de sinistrés non rentables par intervalles de mêmes effectifs, relativement aux variables internes

FIGURE B.5 – *Proportion des sinistrés non rentables selon leur ancienneté en portefeuille*

2.2 - Variables externes

(a) *Espérance de vie à la naissance*(b) *Personnes par ménage*FIGURE B.6 – *Histogrammes de la proportion de sinistrés non rentables par intervalles de mêmes effectifs, relativement aux 2 variables externes les plus liées à la variable cible*

C - Graphiques supplémentaires relatifs à la modélisation de la rentabilité

1 - Arbre Complet

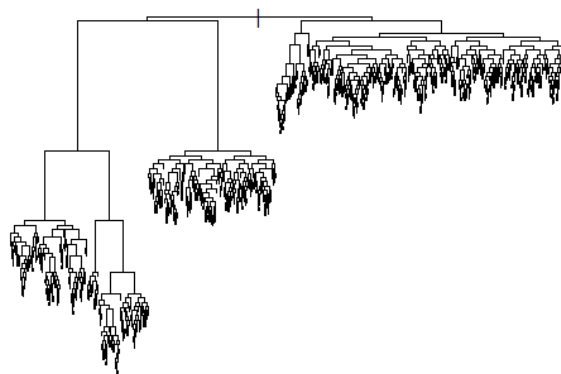


FIGURE C.1 – *Arbre complet (modélisation de la rentabilité)*

2 - Élagage

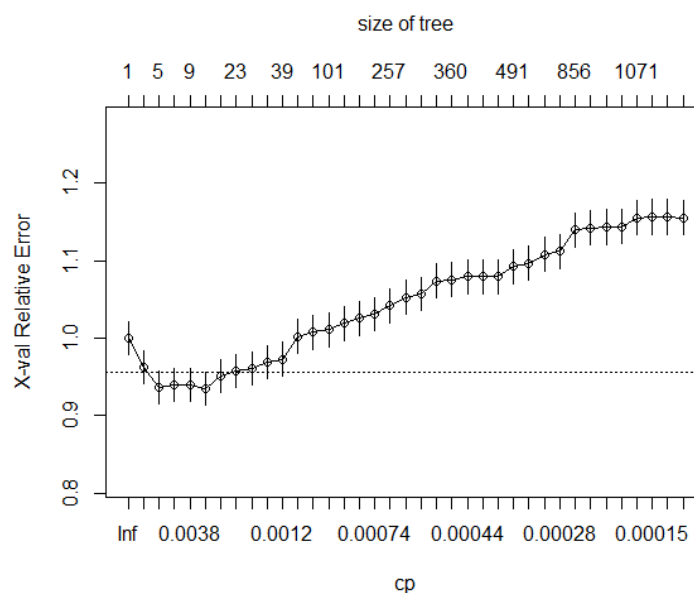


FIGURE C.2 – *Élagage Arbre complet (modélisation de la rentabilité) : 9 scissions retenues, $cp=0.002064694$*

D - Compléments sur la régression logistique

Algorithme de type Newton-Raphson. Le principe général de la méthode de Newton-Raphson est le suivant.

Soit $t : \mathbb{R} \rightarrow \mathbb{R}$ une fonction \mathcal{C}^1 donnée (c'est le score S dans la section 3.1 page 61). La problématique consiste à trouver x^* tel que $t(x^*) = 0$. Par définition de la dérivée, on a :

$$t'(x^*) = \lim_{h \rightarrow 0} \frac{t(x^* + h) - t(x^*)}{h}.$$

La méthode de Newton-Raphson est basée sur l'heuristique suivante : si x est suffisamment « proche » de x^* , alors, par définition de x^* :

$$t'(x) \approx \frac{t(x) - t(x^*)}{x - x^*} \iff x - x^* \approx \frac{t(x)}{t'(x)}.$$

On va utiliser cette méthode itérative en initialisant un x_0 puis en posant, pour tout $n \in \mathbb{N}$,

$$x_n = x_{n-1} - \frac{t(x_{n-1})}{t'(x_{n-1})}.$$

Sous des hypothèses assez souples (fonction t deux fois différentiable au voisinage de x^* par exemple), on a que $x_n \rightarrow x^*$ quand $n \rightarrow +\infty$.

E - Nomenclature des garanties

Pour les garanties en incapacité, il existe 3 causes de franchise absolue :

- Franchise maladie (M) : arrêt de travail lié à une maladie hors hospitalisation.
- Franchise accident (A) : arrêt de travail lié à un accident avec ou sans hospitalisation.
- Franchise hospitalisation (H) : arrêt de travail lié à une hospitalisation nécessitée par une maladie.

Par exemple, la garantie élémentaire IJ 90J 3/0/7 fait référence à des indemnités journalières (IJ) associées à un palier de 90 jours (90J) et une combinaison de franchises H/A/M (Hospitalisation/Accident/Maladie) égale à 3/0/7. Pour une explication des paliers, se référer à la page 10.

Garantie niveau I	Garantie niveau II	Garantie niveau III
Revenu de remplacement (I.J. 90 J)	IT 90 J	IJ 90J 3/0/7
		IJ 90J 3/0/15
		IJ 90J 3/0/30
		IJ 90J 3/15/30
		IJ 365 3/30/90
Revenu de remplacement (I.J. 365 J)	IT 365 J	IJ 365 3/30/90
		IJ 365 90/90/90
		IJ 365 180/180/180
Revenu de remplacement (I.J. 1095 J)	IT 1095 J	IJ 3A 365/365/365
	ITT 1095 J	IJ 3A FR 365-365-365
Frais professionnels 1 an	FP 1 an	FP 365 3/0/7
		FP 365 FR. 3/0/15
		FP 365 FR. 3/0/30
		FP 365 3/15/30
		FP 365 3/30/90
		FP 365 90/90/90
		FP 365 180/180/180
Frais professionnels 3 ans	FP 3 ans	FP 3A 3/0/7
		FP 3A 3/0/15
		FP 3A 3/0/30
		FP 3A 3/15/30
		FP 3A 3/30/90
		FP 3A 90/90/90
		FP 3A 180/180/180
		FP 3A 365/365/365

Table des figures

I.1	Schéma explicatif du passage des données brutes aux données finales.	20
I.2	Illustration de la décomposition de l'excès de risque (erreur totale) en biais (erreur d'approximation) et en variance (erreur d'estimation).	24
II.1	Histogrammes relatifs à l'âge à la souscription.	28
II.2	Histogrammes relatifs à la prime annuelle moyenne TTC (€).	29
II.3	Histogrammes relatifs à la proportion de prime allouée à la garantie Incapacité (en % de la prime annuelle moyenne TTC).	29
II.4	Histogrammes relatifs à la proportion de prime allouée à la garantie Invalidité (en % de la prime annuelle moyenne TTC).	30
II.5	Histogrammes relatifs à la proportion de prime allouée à la garantie Décès (en % de la prime annuelle moyenne TTC).	30
II.6	Histogrammes relatifs à l'ancienneté en portefeuille (années).	31
II.7	Histogramme et cartographie des 21 régions en fonction de la proportion de sinistrés.	32
II.8	Proportion des assurés selon leur appartenance aux réseaux commerciaux.	32
II.9	Distributions de 9 variables explicatives catégorielles selon la proportion de sinistrés dans chaque modalité.	33
II.10	V de cramer entre les variables explicatives qualitatives et la variable cible binaire.	35
II.11	Proportion de sinistrés pour le croisement de deux variables avec l'âge à la souscription.	36
II.12	V de Cramer (en %) entre variables explicatives, toutes qualitatives.	36
II.13	Proportion de sinistrés pour les croisements de modalités de certaines variables très liées entre elles d'après le V de cramer (>40%).	37
II.14	Représentation des 42 modalités des variables explicatives « internes » et des 2 modalités (★) de la variable cible sur le premier plan de l'AFCM, avec une coloration en fonction de leur contribution (en termes d'inertie) à ce plan. La frontière orange sépare les deux clusters obtenus par k-means.	38
II.15	Répartition départementale des effectifs du portefeuille étudié et de la population générale française.	40
II.16	$S_{Y/X}$ entre les variables externes toutes quantitatives et la variable cible binaire, et mise en évidence en rouge d'une liaison significative (rejet de l'hypothèse H_0 du test de WMW) au risque de première espèce $\alpha = 5\%$	40
II.17	Histogrammes de la proportion de sinistrés par intervalles de mêmes effectifs, relativement aux 3 variables externes les plus liées à la variable cible.	41
III.1	Exemple fictif illustrant la vision duale du principe d'un arbre CART de segmentation en deux groupes (« + » et « - ») à partir de deux variables explicatives « X_1 » et « X_2 ».	43

III.2 Entropie, Gini et taux de mal classés normalisés dans le cas d'une classification binaire où p représente la proportion d'une des deux classes d'un nœud donné.	46
III.3 arbre complet ($\gamma=0$)	48
III.4 Taux de mal classés normalisés en fonction de la taille de l'arbre (abscisses inférieures) et du paramètre γ (abscisses supérieures).	48
III.5 Taille de l'arbre en fonction du paramètre de complexité γ	49
III.6 Taux de mal classés par validation croisée (normalisés par la valeur à la racine) en fonction du couple $(\mathcal{T}_i \mid C_p^{\mathcal{T}_i})$	50
III.7 Arbre optimal pour $\gamma = C_p^{opt} = 0.001655995$	51
III.8 Importance, pour l'arbre entier ($\gamma = 0$), des variables (en pourcentage)	53
III.9 Illustration de la construction du prédicteur par forêts aléatoires $\hat{\phi}_{RF}(\mathbb{X})$ à partir de $\mathbb{X} = (X_i)_{i \in [1;5]}$	55
III.10 Forêts aléatoires : Taux de bons classements ($= 1 - R_n^{appr}$) par validation croisée, en fonction du paramètre m , pour un nombre d'itérations=500 (500 arbres construits).	56
III.11 Principe du MDA : étude de l'influence d'une variable par permutation aléatoire de ses valeurs.	57
III.12 Eboulis des MDA et MDE obtenus par forêts aléatoires (500 souches=500 arbres à deux feuilles).	58
III.13 Principe du Boosting	59
III.14 Adaboost : Taux de bons classements ($= 1 - R_n^{appr}$) par validation croisée, en fonction du paramètre M , le nombre d'itérations.	61
III.15 Estimations $(\hat{\theta}_{MV}^k)_k$ par MV des paramètres $(\theta^k)_k$ de la régression logistique sans interactions modélisant l'occurrence d'au moins un sinistre, avant (modèle complet) et après sélection bi-directionnelle de variables.	65
III.16 Le rapport des côtes (odds), appelé odd-ratio (OR), indique l'influence d'une modalité A d'une variable par rapport à une autre modalité B de la même variable (notation « variable modalité A vs modalité B »), toutes choses égales par ailleurs.	67
III.17 Estimations par MV des paramètres significativement différents de 0 (au risque de première espèce 5%) de la régression logistique avec interactions, avant (modèle complet) et après sélection bi-directionnelle de variables.	69
III.18 Matrice de Confusion.	70
III.19 Courbes ROC.	73
III.20 Grille d'appréciation de l'aire sous la courbe ROC (AUC).	74
IV.1 Cartographie des 21 régions en fonction de la proportion de sinistrés non rentables (gauche) et, à titre de rappel (partie III), en fonction de la proportion de sinistrés (droite).	76
IV.2 V de Cramer entre les variables explicatives et la variable cible binaire (rentabilité).	78
IV.3 V de Cramer (en %) entre variables explicatives de la rentabilité, toutes qualitatives.	78

IV.4	Représentation des 45 modalités des variables explicatives internes et des 2 modalités (★) de la variable cible sur le premier plan de l'AFCM, avec une coloration en fonction de leur contribution (en termes d'inertie) à ce plan. La frontière orange sépare les deux clusters obtenus par k-means.	79
IV.5	Arbre optimal (9 scissions) prédisant la rentabilité.	81
IV.6	Importance des variables (les 10 les plus importantes) expliquant la rentabilité dans le cadre d'une modélisation binaire par arbre CART pour l'arbre entier ($\gamma = 0$).	82
IV.7	Eboulis des MDA et MDE obtenus par forêts aléatoires (500 souches).	83
IV.8	Odd Ratios pour les 20 variables issues de la sélection de variables dans le cadre d'une modélisation sans interactions.	83
IV.9	Estimations $(\hat{\theta}_{MV}^k)_k$ par MV des paramètres $(\theta^k)_k$ de la régression logistique avec interactions modélisant la rentabilité binaire de la sinistralité après sélection bidirectionnelle de variables.	84
IV.10	Courbes ROC - Modélisation de la (non-)rentabilité.	86
IV.11	$S_{Y/X}$ entre les variables externes toutes quantitatives et la variable cible binaire, et mise en évidence en rouge d'une liaison significative (rejet de l'hypothèse H_0 du test de WMW) au risque de première espèce $\alpha = 5\%$	87
B.1	Histogramme des 21 régions en fonction de la proportion de sinistrés non rentables	92
B.2	Proportion de sinistrés non rentables selon leur appartenance aux réseaux commerciaux	92
B.3	Distributions de 9 variables explicatives catégorielles selon la proportion de sinistrés non rentables dans chaque modalité	93
B.4	Histogrammes de la proportion de sinistrés non rentables par intervalles de mêmes effectifs, relativement aux variables internes	94
B.5	Proportion des sinistrés non rentables selon leur ancienneté en portefeuille	95
B.6	Histogrammes de la proportion de sinistrés non rentables par intervalles de mêmes effectifs, relativement aux 2 variables externes les plus liées à la variable cible	95
C.1	Arbre complet (modélisation de la rentabilité)	96
C.2	Élagage Arbre complet (modélisation de la rentabilité) : 9 scissions retenues, $cp=0.002064694$	96

Bibliographie

- [Arl04] Sylvain Arlot. *Introduction à la classification*, https://www.math.ens.fr/enseignement/telecharger_fichier.php?fichier=627 , 13 Octobre 2004.
- [Bac] Alain Baccini. *Statistique Descriptive Multidimensionnelle (publications de l'Institut de Mathématiques de Toulouse)*, <https://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf>.
- [Bel] Rémi Bellina. *Méthodes d'apprentissage appliquées à la tarification non-vie*, <http://www.ressources-actuarielles.net/C12574E200674F5B/0/72CE8393E53CE218C1257C39006711AE>.
- [Bes] Philippe Besse. *Exploration Statistique Multidimensionnelle (Data Mining)*, https://www.math.univ-toulouse.fr/%7Ebesse/pub/Explo_stat.pdf.
- [Bou13] Bruno Bouzy. *Risque réel et risque empirique*, <http://www.math-info.univ-paris5.fr/~bouzy/Doc/AA1/RisqueReel.pdf> , 6 février 2013.
- [dF] Fondation de France. *Les solitudes en france – 2016*. https://www.fondationdefrance.org/sites/default/files/atoms/files/les_solitudes_en_france_2016_-_synthese.pdf.
- [DRE] *Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)*, <http://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/>.
- [Eta] *Mission Etalab (Plateforme ouverte des données publiques françaises)*, <https://www.data.gouv.fr/fr/>.
- [Gir] Christophe Giraud. *Classification et apprentissage statistique*, <http://www.cmap.polytechnique.fr/~giraud/MAP553/sec3.4-3.6correction.pdf>.
- [Gue] Benjamin Guedj. *Cours d'apprentissage Statistique*. ISUP, 2016/2017.
- [idG] Site internet de Generali. *Soutien psychologique : pour les entrepreneurs aussi!* <https://www.generalif.fr/professionnel/actu/soutien-psychologique-pour-les-entrepreneurs-aussi/>.
- [INR] INRS. *Les facteurs de risques psychosociaux. Institut national de recherche et de sécurité*, <http://www.inrs.fr/risques/psychosociaux/facteurs-risques.html>.
- [Ins] *Institut national de la statistique et des études économiques (INSEE)*, <https://www.insee.fr/fr/accueil>.
- [LB] Béatrice Laurent and Philippe Besse. *Apprentissage Statistique : modélisation, prévision, data mining*. INSA Toulouse, https://www.math.univ-toulouse.fr/~besse/pub/Appren_stat.pdf.
- [Nan] UFR SEGMI Université Paris Nanterre. *Tests du Khi-deux d'indépendance et d'homogénéité*, https://ufr-segmi.parisnanterre.fr/medias/fichier/section6__1137079084183.pdf.
- [Pag] Jérôme Pagès. *Analyse factorielle multiple avec R*.

- [Poi] Franck Poindessault. La cjeue va-t-elle castrer l'assurance? *L'Argus de l'Assurance.com*, <http://www.argusdelassurance.com/metiers/la-cjeue-va-t-elle-castrer-l-assurance.48730>.
- [Raka] Ricco Rakotomalala. *Comparaison de populations - Tests non paramétriques*, https://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Nonparametriques.pdf.
- [Rakb] Ricco Rakotomalala. *Pratique de la Régression Logistique*, https://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf.
- [Rou] Laurent Rouvière. *Régression logistique avec R*. Université Rennes 2, UFR Sciences Sociales, https://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf.
- [Sal] Hélène Clement Salavera. *La sélection des risques de santé en Prévoyance individuelle : Comment concilier le point de vue des assureurs et des consommateurs ? Des innovations sont-elles possibles ?*, http://www.mba-enass-alumni.org/uploads/3/2/3/1/3231071/mba_enass_2013_salavera_risques-de-sante-prevoyance.pdf.
- [Sap] Gilbert Saporta. *Sensibilité, spécificité, courbe ROC etc.*, http://cedric.cnam.fr/~saporta/Sensibilite_specificiteSTA201.pdf.
- [SAS] SAS Institute Inc. *Guide d'utilisation de la procédure de régression logistique avec le logiciel SAS*, <https://support.sas.com/documentation/cdl/en/statuglogistic/61802/PDF/default/statuglogistic.pdf>.
- [Sep16] Catalina Sepulveda. *Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial.*, <http://www.ressources-actuarielles.net/C12574E200674F5B/0/68C1024B306D64A8C12580B4005FB59C> , 2016.
- [Tuf] Stéphane Tufféry. *Modélisation prédictive et Apprentissage statistique avec R*.
- [Wika] Wikistat. *Scénario : Patrimoine et score d'appétence de l'assurance vie*, <http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-scenar-app-patrimoine.pdf>.
- [Wikb] Wikistat. *Statistique descriptive bidimensionnelle*, <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-des-bi.pdf>.