Adaptation de l'algorithme CART pour la tarification des risques en assurance non-vie

Antoine Paglia[*,***], Martial Phélippé-Guinvarc'h[**], Philippe Lenca[***,****]

[*]EURO Institut d'actuariat EURIA, antoine.paglia@gmail.com
[**]Actuaire, Direction des assurances Agricoles et Professionnelles, GROUPAMA SA
martialphelippeguinvarch@sfr.fr, martial.phelippe-guinvarc-h@groupama.com
[***]Institut Télécom; Télécom Bretagne; UMR CNRS 3192 Lab-STICC,
philippe.lenca@telecom-bretagne.eu
[****]Université européenne de Bretagne

Résumé. Les développements récents en tarification de l'assurance non-vie se concentrent majoritairement sur la maîtrise et l'amélioration des Modèles Linéaires Généralisés. Performants, ces modèles imposent cependant à la fois des contraintes sur la structure du risque modélisé et sur les interactions entre variables explicatives du risque. Ces restrictions peuvent conduire, dans certaines sous-populations d'assurés, à une estimation biaisée de la prime d'assurance. Les arbres de régression permettent de s'affranchir de ces contraintes et, de plus, augmentent la lisibilité des résultats de la tarification. Nous présentons une modification de l'algorithme CART pour prendre en compte les spécificités des données d'assurance non-vie. Nous comparons alors notre proposition aux modèles linéaires généralisés sur un portefeuille réel de véhicules. Notre proposition réduit les mesures d'erreur entre le risque mesuré et le risque modélisé, et permet ainsi une meilleure tarification.

1 Introduction

Les compagnies d'assurances utilisent quotidiennement des modèles statistiques pour évaluer les risques auxquels elles doivent faire face. L'objectif actuariel est d'estimer l'espérance de sinistre de chaque risque souscrit. Les trente dernières années ont été marquées par la sophistication des modèles de régression utilisés pour quantifier ces risques. La régression linéaire simple qui permettait de modéliser par une droite les variations d'une variable cible –le risque étudié–, a été remplacée à partir des années 1980 par les Modèles Linéaires Généralisés (McCullagh et Nelder, 1989), notés GLM par la suite. Ces modèles permettent à la fois de modéliser des comportements non linéaires et des distributions de résidus non gaussiens. Cela est particulièrement utile en assurance non-vie où les coûts des sinistres, quand ils se concrétisent, suivent une densité très asymétrique clairement non gaussienne. Ils ont permis d'améliorer nettement la qualité des modèles de prédiction du risque et sont aujourd'hui largement utilisés par les compagnies d'assurance. Cependant, bien que très sophistiqués, les modèles GLM présentent des limites en terme de modélisation des interactions entre variables explicatives et en

terme de modélisation de la structure des risques. Ces limites nous conduisent à tester d'autres outils d'apprentissage statistiques ayant démontré des capacités à extraire des structures de dépendances et des particularités entre les données qui restaient jusque là non détectées par les outils de régression classiques. Parmi ceux-ci, mentionnons les réseaux de neurones, les arbres de décision ou encore les *support vector machines* dont l'efficacité a été prouvée dans de nombreux domaines (Wu et al., 2008). Leur utilisation en assurance est cependant moins répandue et/ou confidentielle, notamment en assurance de véhicule, notre domaine applicatif. La littérature est ainsi peu abondante. Nous renvoyons cependant le lecteur intéressé aux études de Apte et al. (1999), Dugas et al. (2003) et Christmann (2004).

Parmi les algorithmes de référence, nous avons testé en pré-étude les arbres de décision simples, les arbres de décision boostés et les réseaux de neurones. Nous avons constaté la meilleure performance en terme de *mean square error* des arbres de décision boostés et l'excellente lisibilité des arbres de décision simples. C'est la lecture visuelle associée au principe de l'algorithme -qui est de créer des groupes de risque homogènes- et aux bonnes performances en terme de *mean square error*, notée MSE par la suite, de l'arbre de régression simple qui nous ont invités à retenir ce modèle. Notons que les tous les véhicules ne sont pas assurés sur l'année complète, par exemple quand ils sont achetés ou vendus en cours d'année. Si la durée moyenne d'exposition d'un groupe de véhicules est inférieure à la durée d'exposition moyenne de la base [1] l'arbre identifie ce groupe de véhicules comme étant moins risqué. Cela conduit à une sous-estimation de la prime pour ces groupes de véhicules (Christmann, 2004). Nous proposons ainsi CART-ANV une modification de l'algorithme CART (Breiman et al., 1984) pour prendre en compte cette spécificité assurance non-vie.

Le premier enjeu de l'assureur est la bonne mesure du risque. L'article vise donc premièrement à comparer la performance globale des GLM (le modèle de référence en actuariat) par rapport à l'algorithme CART-ANV. L'amélioration de la segmentation d'un portefeuille d'assurés constitue un enjeu économique et stratégique majeure. En effet, l'assureur cherche à développer sa part de marché sur les segments qui conduisent à la fois à un avantage concurrentiel et à un profit. La qualité d'une segmentation par groupe de risques peut se mesurer selon quatre critères majeurs que sont l'équité, l'homogénéité, le caractère réalisable et le caractère incitatif (Feldblum, 2006). L'absence de biais entre le risque mesuré et le risque prédit correspond au critère d'équité et stipule que les primes payées par le groupe doivent refléter les pertes occasionnées par ce groupe. Le critère d'homogénéité exprime le fait que les risques au sein d'un groupe sont homogènes et qu'il n'est pas possible de subdiviser ce groupe en plusieurs sous groupes ayant des primes significativement différentes. Nous comparons donc également la performance des deux approches par segment sur les deux principaux critères d'équité et d'homogénéité. Enfin, nous comparons également les deux approches dans leurs aspects pratiques *i.e.* dans la préparation des données, la mise en œuvre des outils, la fiabilisation des résultats, leurs lectures et dans leurs communications interne et externe.

La tarification en assurance se fait généralement sur la somme de modèles par risque (Responsabilité Civile, incendie, bris de glace, tout accident...). Cela se comprend statistiquement par le fait que les densités des sinistres et les variables explicatives sont différentes sur chacun de ces risques. De plus, dans les systèmes informatiques, les primes de chaque risque sont souvent paramétrées indépendamment les unes des autres. Néanmoins, évaluer la prime

[1] C'est le cas par exemple des véhicules neufs qui rentrent en base systématiquement quand ils sont achetés en cours d'année.

pure au niveau du véhicule permet d'augmenter la performance du modèle global et d'intégrer les dépendances éventuelles entre les risques souscrits. C'est pourquoi, tant d'un point de vue technique que commercial, nous modélisons le sinistre total du véhicule plutôt que de construire un modèle par risque.

Dans la section 2, nous présentons les données réelles ayant servi à comparer les modèles CART-ANV et GLM. La section 3 présente les options retenues pour le modèle GLM et pour le modèle CART-ANV. La section 4 présente les résultats de la comparaison entre GLM et CART-ANV. Enfin nous concluons en section 5.

2 Les données

Nous avons suivi une méthodologie proche de CRISP-DM[2] (Shearer, 2000), dont l'une des étapes les plus importantes consiste à préparer les données. Nous décrivons ci-dessous les principaux éléments de la base de données brute et les transformations opérées.

Nous précisons que nous utilisons une base de données réelle, brute et volumineuse d'un assureur. Pour des raisons -évidentes- de confidentialité, nous avons dépersonnalisé la base. Par exemple, nous n'avons pas explicité des variables comme l'usage des véhicules (noté usage1, usage2 ...) et avons effectué des homothéties des variables quantitatives comme le montant du sinistre.

Par ailleurs, l'assureur est en charge de la mesure de son propre risque et donc de valider les modèles sur sa propre base de données correspondant à son *business*. Nous ne comparons donc les modèles que sur une seule base de données. Une comparaison sur d'autres bases, dont les caractéristiques seraient différentes car par exemple issues de processus métier différents, ne ferait pas sens pour l'objectif que nous poursuivons.

2.1 Description de la base de données brute

La base de données rassemble plusieurs exercices, soit un total de plus de trois millions de contrats. Les contrats d'assurance sont décrits par 45 variables explicatives (âge du véhicule, puissance, montant de la franchise ...) dont la majorité sont discrètes et comportent de nombreuses modalités (code postal, Catégories Socio-Professionnelles -CSP-, marque du véhicule ...). Un souscripteur assure en moyenne 1,9 véhicule, mais cela va de 1,2 pour les salariés et retraités à presque 4 pour la catégorie socio-professionnelle la plus équipée. La durée moyenne de présence d'un contrat dans le portefeuille est de 0,84 an. L'assuré paye une prime au *prorata* de la durée assurée. Si l'assureur n'intègre pas cette période d'exposition dans la modélisation de la prime pure, il ne récoltera que 84% du montant des primes. Nous proposerons dans cet article une méthode pour intégrer cette composante dans l'algorithme CART et qui sera présentée section 3.2.2.

Environ 150 000 garanties sinistrées sont enregistrées dans la base. Le montant enregistré est le montant d'indemnisation de l'assureur sur une garantie, ce qui correspond à la valeur totale du sinistre diminuée de la franchise. La fréquence moyenne des sinistres est de 7,49% pour un montant moyen de 1237 euros, soit une prime pure en ne tenant compte d'aucunes variables explicatives de 276 euros. La table 1 détaille la répartition des sinistres. Ainsi, 6,48%

[2] `http://www.crisp-dm.org`

des assurés ont eu un sinistre compris entre 1 et 4500 euros. Ces sinistres de faible montant contribuent à 39,30% de la charge totale des sinistres. Il est intéressant de noter l'importance prise par les sinistres extrêmes : seulement 0,01% des assurés ont eu un sinistre extrême mais la somme de ces sinistres contribue à 19% du montant total des sinistres. Le montant maximum de sinistre est de 7,8 millions d'euros.

Charge de sinistre	% obs.	% du montant total	Moyenne	Médiane
0	92,51	0	0	0
]0,4500]	6,48	39,30	1680	1266
]4500,30000]	0,92	28,00	8394	6762
]30000,150000]	0,06	13,78	60705	50682
[150000,max]	0,01	19,10	471627	276012

TAB. 1 – *Répartition de la charge de sinistre. Cette répartition montre l'importance des valeurs extrêmes : 0,01% des assurés contribuent à 19% du montant total des sinistres.*

L'écart entre le montant moyen et la médiane dans une tranche de sinistre montre que la distribution des sinistres est très asymétrique. Ceci est confirmé par le coefficient de dissymétrie - skewness - qui est de 3,2 et l'histogramme de la distribution des sinistres par contrats présenté dans la figure 1.

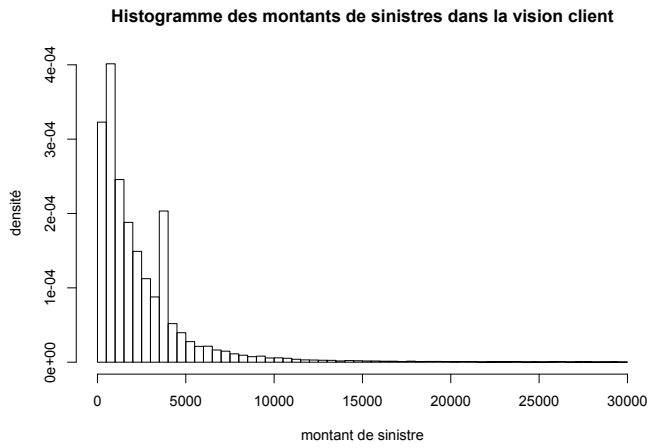

FIG. 1 – *L'histogramme restreint aux sinistres compris entre 0 et 30 000 euros montre une distribution très asymétrique et la présence d'un pic*

2.2 Préparation des données

La plupart des variables de la base de données brutes ont fait l'objet d'un traitement préalable. Nous présentons dans cette section les traitements les plus importants et de leurs conséquences éventuelles sur les résultats du modèle.

Les valeurs non-cohérentes et extrêmes pour les variables explicatives : les valeurs non-cohérentes ont été détectées notamment à l'aide d'un expert sur le risque assuré (particulièrement pour les variables puissances et des âges des véhicules). Une fois identifiées, ces valeurs ont été, soit remplacées par la valeur la plus probable (après une régression sur la variable à remplacer), soit définies comme valeurs manquantes. Les assureurs étant soumis à des contrôles externes (IFRS, Solvency), ils s'imposent une rigueur de traitement des données. Ainsi même si cela surprend, les règles de contrôles de cohérences établies par des experts (qui connaissent les véhicules, les contrats d'assurance et les processus d'alimentation de la base de données) sont jugées plus fiables qu'une procédure automatique non-experte de traitement des incohérences.

Regroupements : des regroupements de modalités ont été opérés notamment pour les variables aux modalités très nombreuses afin d'éviter, comme c'est le cas avec la méthode GLM, la création d'un nombre très important de variables binaires pour ces variables. Ces regroupements sont réalisés, soit en utilisant des outils de classification statistique, soit en utilisant l'avis d'experts. Par exemple les 100 modalités d'origine de la variable CSP ont été regroupées en 12 modalités ; la variable $Genreduvehicule$ est issue d'un regroupement d'une cinquantaine de variables en 6 groupes à dire d'experts métier ; la variable $Usage$ a également été créée à partir d'un regroupement d'une cinquantaine de catégories et possède 8 niveaux (Usage1, Usage2 ...). Bien que les regroupements des variables qualitatives ne soient pas nécessaire pour CART-ANV, ils sont utiles dans le cadre d'une comparaison avec la GLM.

Les sinistres extrêmes : la table 1 montre l'importance des sinistres extrêmes dans le montant final de la prime puisque ces sinistres qui ne représentent que 0,01 % du nombre total d'assurés contribuent à hauteur de 19 % du montant total des sinistres. Dans notre approche, les classes de risques ne sont pas fixées *a priori*, ce qui nécessite de fixer un seuil d'écrêtement pour l'ensemble du portefeuille. En utilisant la méthode de la fonction moyenne des excès - mean excess loss - présentée dans Embrechts et al. (1997), on obtiendrait un seuil d'écrêtement de 300 000 euros pour le montant des sinistres agrégés au niveau du véhicule. Toutefois, il peut être préférable de fixer le seuil d'écrêtement selon d'autres critères en fonction du modèle statistique utilisé (en particulier pour les modèles minimisant une distance quadratique). Le seuil S est fixé au quantile à 99 % de la distribution des montants des sinistres strictement positifs, soit 30 000 euros (parmi les 2,2 millions d'observations de la base, 1600 observations sont supérieures à ce seuil).

Agrégation au niveau du véhicule : nous proposons de modéliser le tarif global d'un véhicule plutôt que de modéliser le risque de chaque garantie souscrite. Pour cela, il nous faut sommer les coûts de sinistre de chaque garantie et créer une variable qui décrit la formule « théorique » de garantie utilisée par le véhicule. Ces formules sont notées de la « formule 1 »

(Responsabilité Civile uniquement, 25% des cas) à la « formule 4 » (Tout Accident, 45% des cas). D'autres formules existent mais ne figurent plus dans la cible commerciale actuelle. Ces formules atypiques représentent 14 % des cas et ne seront pas utilisées pour la modélisation.

3 La modélisation

Par nature, si les sinistres deviennent prévisibles, le risque ne peut plus faire l'objet d'un contrat d'assurance. L'assureur n'ambitionne donc pas de prévoir chaque sinistre mais vise à estimer l'espérance du sinistre. C'est pourquoi, le Mean Square Error est intrinsèquement élevé quelque soit le modèle mis en œuvre.

Notons également les particularité de la distribution des coûts de sinistre. En plus d'obtenir un coefficient de variation élevé, nous observons une très forte asymétrie ($Kurtosis \simeq 3$) des coûts de sinistre (*cf.* figure 1). Cela s'aggrave par le fait que le coût est forcé à 0 quand il n'y pas de sinistre (ce qui représente souvent plus de 90% des lignes en assurance non-vie). Ainsi, même si la taille de la base paraît importante, elle n'est pas surdimensionnée pour obtenir des résultats significatifs.

La préparation des données (section 2.2) a explicité l'importance des sinistres extrêmes et a retenu un seuil d'écrêtement $S = 30\ 000$€. Dans les modèles mis en œuvre, cela se traduit par : $E[Y|X] = E[Y|X, Y < S] + E[Y|X, Y \geq S]$ où Y désigne la charge de sinistre par police et X les variables explicatives associées à la police. Dans la suite, Y désignera les sinistres écrêtés.

3.1 Régression par modèle linéaire généralisé

Un modèle linéaire généralisé a été ajusté sur la base d'apprentissage. L'hypothèse d'une loi de « poisson » a été retenue pour l'erreur dans le paramétrage de la GLM. La sélection de modèle a été effectuée selon une stratégie de sélection « forward ». Cette stratégie consiste à partir du modèle sans variable explicative, puis à ajouter la variable qui réduit le plus l'erreur sur la base de validation. Cette opération est réitérée jusqu'à ce que l'ajout de variables augmente l'erreur sur la base de validation. Le modèle final est celui dont le choix des variables explicatives minimise l'erreur sur la base de validation. Cette stratégie de sélection de modèle est consistante avec la stratégie utilisée pour paramétrer les autres algorithmes et permet de comparer de manière objective les résultats produits par ces algorithmes et les modèles GLM.

3.2 L'algorithme CART pour les arbres de régression

CART peut être utilisé pour des problèmes de classification ou de régression. Dans le cadre de l'estimation de la prime pure, le problème est lié à une régression sur le montant et la fréquence des sinistres. Nous renvoyons le lecteur à Breiman et al. (1984) et à Hastie et al. (2008) pour une explication complète de l'algorithme, en particulier sa capacité à gérer les valeurs manquantes.

Le résultat de l'algorithme CART qui cherche à prédire la valeur de Y, le sinistre, est une fonction notée $\hat{f}_w(X)$ en fonction des valeurs explicatives X (i.e. $E[Y|X]$) et un paramètre de complexité w qui correspond au nombre de nœuds dans l'algorithme CART. L'ajustement de l'algorithme se fait en deux étapes. La première étape est l'ajustement de la fonction $\hat{f}_w(x)$

sur la base d'apprentissage et la deuxième étape consiste à trouver le nombre de nœuds terminaux w qui minimise l'erreur de généralisation sur la base de test.

Notons que la fonction $\hat{f}_w(x)$ est de la forme :

$$\hat{f}_w(x) = \sum_{j=1}^{w} \bar{Y}_{j,w} \times I\{x \in R_{j,w}\}$$

où w désigne le nombre de nœuds terminaux de l'arbre, $I\{x \in R_{j,w}\}$ est la fonction indicatrice associée au nœud final $R_{j,w}$ et $\bar{Y_{j,w}}$ désigne la moyenne empirique dans le groupe j.

Partant de la base de données initiale, l'algorithme calcule pour chaque variable et pour chaque séparation possible la valeur de la déviance du nœud *parent, fils gauche et fils droit*. Dans le cas d'une minimisation de l'erreur quadratique, cette déviance a pour expression : $D = \sum_{i \in Noeud} (y_i - \bar{y_i})^2$. L'algorithme calcule ensuite, pour chacune des séparations possibles, la valeur R de la réduction de déviance : $R = D_{parent} - (D_{fils\ gauche} + D_{fils\ droit})$. Le nœud finalement retenu est la séparation qui maximise la réduction de déviance R. L'algorithme recommence ensuite la création d'un nouveau nœud jusqu'à ce que le critère d'arrêt sur le nombre minimum d'individu dans un nœud soit rencontré. L'arbre produit de nombreux nœuds. Une deuxième étape, appelée élagage, consiste à retirer tous les nœuds qui résultent du surapprentissage en utilisant une deuxième base, la base de validation.

3.2.1 Paramétrage de l'algorithme sous R

Plusieurs packages existent sous R pour construire des arbres de décision avec l'algorithme CART. Nous avons retenu le package de référence rpart de Therneau et al. (2009) car il nous permet de recoder les modifications à apporter sur l'algorithme pour intégrer le temps d'exposition.

La fonction rpart a 6 paramètres : $xval$, $minbucket$, $maxcompete$, $maxsurrogate$, cp et $maxdepth$.

Le paramètre $xval$ désigne le nombre de validations croisées effectuées par le modèle ($xval = 10$ par défaut). Pour un portefeuille d'assurance qui possède une forte variance dans la variable Y à expliquer, la validation croisée n'est pas très adaptée à moins d'utiliser un faible nombre de validations croisées de façon à ce qu'à chaque étape la proportion de la base utilisée pour la validation contienne un grand nombre d'observations. Notre approche préconise donc de ne pas utiliser cette option ($xval = 0$) et de découper aléatoirement, classiquement, la base de donnée en trois parties : 50% de la base sert à l'apprentissage, 25% servent à la validation et à trouver le paramètre de complexité w optimal du modèle, les 25% restant serviront de base de test permettant de comparer les différents modèles. Les modèles GLM et CART-ANV seront donc comparés objectivement sur les mêmes bases.

Le paramètre $minbucket$ désigne le nombre minimum d'individus dans un groupe de risque final. Ce paramètre est un des deux critères d'arrêt dans la construction de l'arbre. Une valeur trop grande ne permet pas de modéliser les singularités des données. Une valeur trop petite crée des nœuds trop spécifiques qui seront supprimés par le second critère d'arrêt. Dans notre cas, il est fixé à 1000 véhicules. En effet, l'ordre de grandeur de la fréquence de survenance d'un sinistre (<10%) et l'importance de la variance et de l'asymétrie de Y, font qu'une

moyenne sur moins de 1000 véhicules a peu de chance d'être significative. Les temps de traitement étant raisonnables, il n'était pas utile d'optimiser ce paramètre à une valeur supérieure.

$maxcompete$ impacte uniquement l'affichage et pas les résultats (il permet l'affichage des $maxcompete$ meilleurs critères de réduction de déviance R du nœud *parent*).

$maxsurrogate$ est le paramètre qui permet de définir le nombre de variables de substitution pour les variables prédictives ayant des valeurs manquantes dans la base de données. En effet, d'une part, CART ne supprime pas les observations ayant des valeurs manquantes et d'autre part, ne remplace pas les valeurs manquantes par des valeurs estimées. Une fois le nœud créé, l'algorithme sélectionne parmi les variables de substitution celles qui représentent le mieux le nœud créé pour répartir les observations où la variable est manquante. Par exemple, si la séparation du nœud est créée sur l'âge et que la seconde variable qui explique le mieux cette séparation est la puissance l'algorithme va classer les observations où l'âge n'est pas renseigné dans le nœud en fonction de la puissance. Dans notre cas, deux variables de substitutions sont suffisantes.

cp est un critère d'arrêt qui utilise le critère de complexité w et qui vise à optimiser les temps de calculs. Nous l'avons fixé à zéro pour ne pas retenir ce critère d'arrêt.

$maxdepth$ désigne la profondeur maximum de l'arbre, fixée à 11 dans notre estimation car l'arbre, une fois élagué, a une longueur maximale de 10 nœuds. Ce paramètre n'a donc pas d'impact sur les résultats mais sur le temps de calcul.

3.2.2 Adaptation de l'algorithme au problème de l'estimation de la prime pure

Dans son papier sur l'application des algorithmes de support vector machine à l'estimation de la prime pure, Christmann (2004) propose de diviser les sinistres dans la base de données par leur période d'exposition. Mathématiquement, l'effet de cette transformation est présenté dans l'inéquation suivante :

$$\sum_{i=1}^{M} t_i * (\frac{1}{M} * \sum_{i=1}^{M} \frac{Y_i}{t_i}) \geq \sum_{i=1}^{M} Y_i \quad soit \quad \sum_{i=1}^{M} t_i * \tilde{Y} \geq \sum_{i=1}^{M} Y_i$$

où M désigne le nombre d'individus dans un groupe d'assurés payant la même prime, Y_i désigne le montant de sinistre et t_i la période d'exposition. La démonstration de cette inéquation se fait par récurrence. Cette méthode conduit à une surestimation de la prime *i.e.* que la somme des primes estimées dans un groupe de M assurés est supérieure à la somme des sinistres dans ce groupe.

Afin d'obtenir l'égalité entre les primes pures actuarielles et la sommes des sinistres réels $\left(\sum_{i=1}^{M} t_i * \tilde{Y} = \sum_{i=1}^{M} Y_i\right)$ il vient algébriquement que la quantité $\tilde{Y}$ estimée par l'algorithme dans un nœud et la fonction de déviance $\tilde{D}$ deviennent :

$$\tilde{Y} = \frac{\sum_{i=1}^{M} Y_i}{\sum_{i=1}^{M} t_i} \quad et \quad \tilde{D} = \sum_{i \in Noeud} (y_i - \tilde{y_i} * t_i)^2$$

Nous avons donc intégré cette modification directement dans l'algorithme CART. Cette approche permet de prendre en compte l'effet de la période d'exposition sur la sinistralité du portefeuille et est équivalente mathématiquement à l'introduction d'un offset dans les modèles GLM qui possèdent une fonction de lien logarithmique.

4 Résultat de la comparaison CART-ANV/GLM

Nous comparons dans cette section, les performances générales des deux modèles, leurs performances par segment sur les deux principaux critères d'équité et d'homogénéité. Nous discuterons enfin des aspects pratiques de ces deux approches dans le cadre actuariel.

4.1 Le Mean Square Error

Nous comparons d'abord les modèles à l'aide du Mean Square Error (MSE), un critère usuel de performance d'un modèle. Nous constatons clairement que sur la base de test, l'algorithme CART-ANV sur-performe le modèle GLM.

Modèle	MSEapprentissage	MSEtest
Régression GLM (poisson)	1148103	1177830
CART-ANV	1144881	1176777

TAB. 2 – *MSE calculé sur la base d'apprentissage et sur la base de test*

Les faibles différences indiquées dans le tableau 2 peuvent sembler minimes. Rappelons que le Mean Square Error, est intrinsèquement très élevé en assurance non-vie. Ainsi, réduire de 1/1 000 la valeur du MSE constitue une réelle réduction de l'erreur sur l'estimation de la prime. Cette réduction du MSE est réellement utile à l'assureur qui cherche, sur son portefeuille, le meilleur résultat. [3]

4.2 Le critère d'équité

Nous cherchons à illustrer graphiquement la performance des deux approches en terme d'équité en projetant les résultats par segment (*i.e.* groupe de véhicules homogènes).

Réaliser cette analyse pose le problème du choix de la taille du segment dans lequel on mesure les écarts entre les modèles. Si le segment est trop grand, les écarts tendent à s'effacer quelques soient leurs performances. Si au contraire, on cherche à mesurer les écarts dans des segments trop petits, le caractère aléatoire de la sinistralité ne permet plus de comparer les écarts produits par les modèles.

Ainsi que le montre la figure 2, dans sa partie en haut à gauche, l'algorithme CART-ANV et la GLM montrent des performances comparables en terme de biais lors d'une projection des résultats sur une seule variable explicative (segments-unidimensionnels). En effet, les courbes CART-ANV et GLM épousent toutes les deux l'histogramme des sinistres.

En revanche, la projection des résultats sur deux variables explicatives (segments multidimensionnels) tend à montrer que l'algorithme CART-ANV est moins biaisé que la GLM, tendance qu'il s'agirait de confirmer dans des travaux futurs, comme cela est illustré dans les trois autres parties de la figure 2. Elles montrent la projection sur la puissance du véhicule pour différentes tranches d'âge. Dans ces trois cas, nous observons que la courbe CART-ANV épouse beaucoup mieux l'histogramme des sinistres. Considérons par exemple les assurés dont

[3] Nous ne cherchons pas à prouver que l'algorithme CART-ANV sur-performe la GLM de manière systématique en assurance non-vie.

l'âge des véhicules est compris entre 0 et 15 ans (partie en haut à droite de la figure 2) et dont la puissance se situe entre 50 et 75 ch. Nous remarquons un écart significatif de la GLM qui inciterait ces assurés à souscrire au juste prix chez un autre assureur.

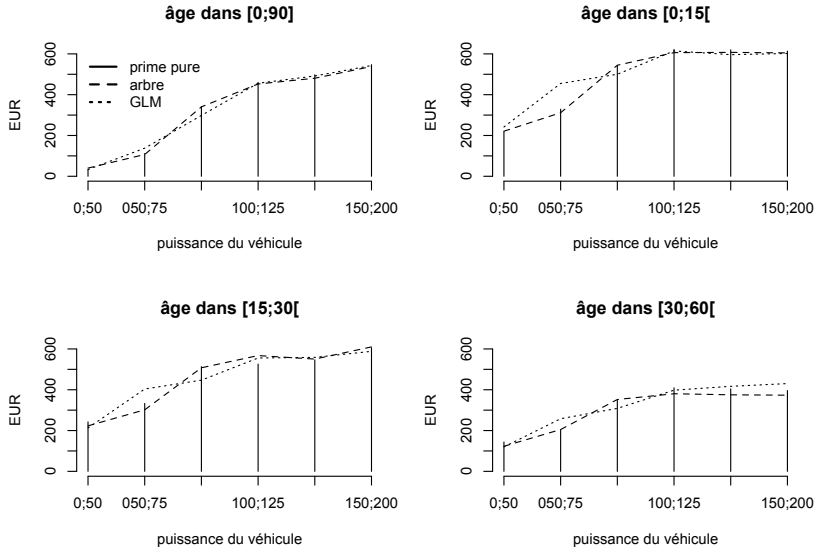

FIG. 2 – *Montant de sinistres en fonction de la puissance du véhicule pour le portefeuille global puis restreint à différentes tranches d'âge*

4.3 Lecture des résultats produits par l'arbre

L'arbre ajusté sur la base d'apprentissage puis élagué sur la base de validation possède un total de 65 nœuds finaux. Le premier résultat qu'il convient de remarquer est la possibilité d'avoir une vision exhaustive du modèle à la fois sur le montant des primes en fonction des variables explicatives et le nombre d'assurés concernés par ce montant de prime.

Cette lisibilité de la tarification permet ainsi de réunir à une même table techniciens, responsables marketing et actuaires pour discuter des stratégies tarifaires à mettre en place. Concernant cette stratégie, il peut par exemple être décidé de supprimer une séparation finale en deux nœuds si celle-ci segmente le risque d'une façon incompatible avec les prix de marché.

De plus, l'arbre permet de montrer que certains critères de risque sont plus importants dans certaines sous-populations que d'autres. Ainsi, la seconde variable d'influence chez les véhicules récents (âge inférieur à 5,6 ans) est la puissance alors que pour les véhicules anciens, la variable puissance intervient beaucoup plus bas dans l'arbre pour discriminer les risques.

Un autre avantage des arbres de régression tient au fait que l'algorithme cherche à chaque étape à créer le nœud qui engendre la réduction d'erreur quadratique la plus importante et ne fournis pas la même profondeur de segmentation sur tout le portefeuille. Par exemple, pour les

véhicules d'âge supérieur à 24,5 ans qui représentent la moitié du portefeuille, seuls six tarifs sont produits par l'arbre sur une profondeur de deux ou trois nœuds seulement (figure 3). La segmentation des risques sur cette partie du portefeuille n'ayant pas un grand intérêt pour l'assuré (les majorations/minorations de tarif étant faibles), cette propriété de l'algorithme apparaît comme un avantage face aux méthodes statistiques qui segmentent uniformément le risque sur le portefeuille.

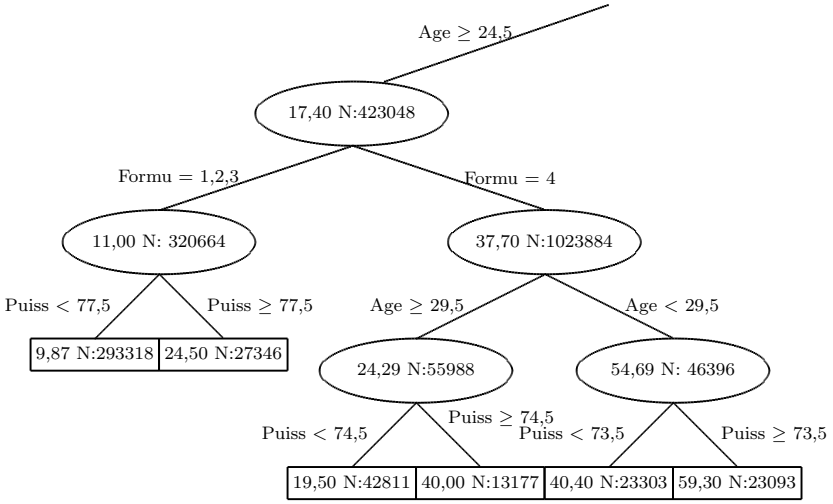

FIG. 3 – *Arbre de régression pour la branche concernant les véhicules anciens.*

5 Conclusion

Cette étude met en œuvre une approche innovante pour la tarification des risques d'assurance non-vie. Alors que les développements récents en actuariat de l'assurance dommage se sont focalisés sur la maîtrise et l'amélioration des Modèles Linéaires Généralisés, nous proposons une version modifiée de l'algorithme CART pour la régression. L'algorithme CART, modifié pour prendre en compte des spécificités assurancielles, a permis de faire ressortir des informations nouvelles sur le risque tout en améliorant les mesures d'erreur entre le risque mesuré et le risque modélisé. L'assureur trouve également une réelle plus-value dans la segmentation produite par l'algorithme. Les véhicules anciens, très nombreux, mais qui ont un risque faible et peu d'enjeux commerciaux, sont modélisés très simplement sur six classes. Par contre, la segmentation est beaucoup plus fine pour les véhicules récents, tout en évitant de paramétrer le modèle sur des classes de risque dont le nombre de personnes assurées n'est pas significatif (évite donc le phénomène de surapprentissage). De plus, les tests réalisés tendent à montrer que l'algorithme CART est moins biaisé que la GLM sur les différents segments de la base. Enfin, la lisibilité des arbres de décision permet de réunir à une même table techniciens, responsables marketing et statisticiens pour discuter des stratégies tarifaires à mettre en place.

Références

Apte, C., E. Grossman, E. Pednault, B. Rosen, F. Tipu, et B. White (1999). Probabilistic estimation based data mining for discovering insurance risks. *IEEE Intelligent Systems 14*, 49–58.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Christmann, A. (2004). An approach to model complex high–dimensional insurance data. *Allgemeines Statistisches Archiv 88*(4), 375–396.

Dugas, C., N. Chapados, Y. Bengio, P. Vincent, G. Denoncourt, et C. Fournier (2003). Statistical learning algorithms applied to automobile insurance ratemaking. In *Casualty Actuarial Society Forum-Arlington*, pp. 179–213.

Embrechts, P., C. Kluppelberg, et T. Mikosch (1997). *Modelling extremal events*. Springer Berlin.

Feldblum, S. (2006). *Risk Classifications, Pricing Aspects*. Encyclopedia of Actuarial Science. John Wiley and Sons.

Hastie, T. M., R. Tibshirani, et J. Friedman (2008). *The Elements of Statistical Learning*. Springer Series in Statistics.

McCullagh, P. et J. Nelder (1989). *Generalized linear models* (2 ed.). UK : Chapman and Hall.

Shearer, C. (2000). The crisp-dm model : The new blueprint for data mining. *Journal of data Warehousing 5*(4), 13–22.

Therneau, T. M., B. Atkinson, et B. Ripley. (2009). *Rpart : Recursive Partitioning*. CRAN. R package version 3.1-45.

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems 14*(1), 1–37.

Summary

Non-life actuarial researches mainly focusses on improving Generalized Linear Models. Nevertheless, this type of model sets constraints on the risk structure and on the interactions between explanatory variables. Then, a bias between the real risk and the predicted risk by the model is often observed on a part of data. Nonparametric tools such as machine learning algorithms are more efficient to explain the singularity of the policyholder.

Among these models, decision trees offer the benefit of both reducing the bias and improving the readability of the results of the pricing estimation. Our study introduces a modification of the Classification And Regression Tree (CART) algorithm to take into account the specificities of insurance data-sets and compares the results produced by this algorithm to the prices obtained using Generalized Linear Models. These two approaches are then applied to the pricing of a vehicle insurance portfolio.