GLM ET TARIFICATION EN ASSURANCE NON-VIE

Nouvelles approches
à partir d'arbres et de forêts GLM

Les modèles linéaires généralisés (GLM) constituent l'outil de référence utilisé par les actuaires pour la tarification des produits d'assurance non-vie. Ces dernières années, ce socle de base s'est largement enrichi avec la démocratisation des techniques de *machine learning*. Malgré le gain en performance de ces dernières, force est de constater que les GLM n'ont pas disparu du quotidien des actuaires, puisqu'ils répondent directement à un besoin d'interprétabilité de la profession. Améliorer les performances des modèles GLM pour les associer efficacement à des techniques de *machine learning* constitue donc un enjeu important pour garantir la compétitivité et la transparence des modèles de tarification en assurance non-vie.

Introduction

Ces dernières années, l'émergence du big data et la popularisation des techniques de *machine learning* ont offert de nouvelles perspectives pour accroître les performances des processus de souscription et de gestion des risques en assurance. Cette recherche de performance peut en effet s'opérer en exploitant la masse des informations individuelles collectées par les assureurs et en sélectionnant les modèles les plus efficaces pour représenter la variété des profils de risques assurés. Cette logique s'applique notamment à la tarification des contrats d'assurance non-vie où la fourniture de tarifs individualisés est un enjeu important de compétitivité. Ces évolutions récentes s'effectuent néanmoins en parallèle d'un renforcement de la règlementation, que ce soit en matière de protection des données avec la règlementation RGPD, ou par un besoin de transparence des décisions prises par les modèles de *machine learning* (OCDE, 2020). Aussi, une des questions centrales de l'actuaire est de concilier l'exigence de performance, *i.e.* la capacité des modèles à représenter au mieux et avec robustesse les phénomènes individuels observés, avec la capacité à expliquer les prédictions des modèles. Depuis les années 1990, les modèles linéaires généralisés constituent le point de repère utilisé par les actuaires pour la tarification des contrats d'assurance non-vie. Leur simplicité d'utilisation et leur conception transparente permettent en effet d'expliquer facilement la logique de segmentation retenue dans l'offre tarifaire de l'assureur, là où des modèles de *machine learning* plus performants peuvent nécessiter des éclaircissements par le biais de techniques d'interprétation. Dans ce contexte, nous proposons d'aborder dans cet article différentes améliorations apportées aux modèles GLM pouvant être utilisées dans le cadre d'approches reposant sur des techniques de *machine learning*.

GLM et utilisation de formules fermées

Traditionnellement, les modèles GLM sont estimés par maximum de vraisemblance. Pour rappel, la log-vraisemblance d'un modèle GLM s'exprime, pour les variables réponses Y_i, $i \in I = \{1,\dots,n\}$ indépendantes appartenant à la famille exponentielle, sous la forme :

$$log\,(L(\theta|y_i) = \frac{\lambda_i(\theta)y_i - b\,(\lambda_i(\theta))}{a\,(\phi)} + c(y_i,\phi), \quad y_i \in \Upsilon \subset \mathrm{R},$$

et $-\infty$ si $y_i \notin \Upsilon$, avec $a: \mathrm{R} \to \mathrm{R}$, $b: \Lambda \to \mathrm{R}$ et $c: \Upsilon \times \mathrm{R} \to \mathrm{R}$ des fonctions connues à valeurs réelles, θ un vecteur de paramètres à estimer et ϕ un paramètre de dispersion. Le modèle établit alors un lien entre l'espérance de la variable réponse $E(Y_i)$ et un vecteur de variables explicatives x_i relatives à chaque observation $i \in I$ tel que

$$g\,(b'(\lambda_i(\theta))) = < x_i, \theta >.$$

En dehors du cas où la distribution de la variable réponse est gaussienne, l'estimateur du maximum de vraisemblance de θ s'obtient en résolvant numériquement l'équation du score par le biais d'un algorithme de type Newton comme l'algorithme IWLS *(Iteratively re-Weighted Least Square)*.