Méthodes d'apprentissage statistique – « Machine Learning »

Fabrice TAILLIEU, Sébastien DELUCINGE, Rémi BELLINA

Le marché de l'assurance a rarement été marqué par un environnement aussi difficile qu'au cours de ces dernières années. Cette situation met encore davantage en lumière la nécessité au sein des compagnies d'effectuer les diagnostics pouvant aider le management dans sa prise de décisions. L'utilisation de méthodes innovantes et rapides à mettre en œuvre comme les méthodes d'apprentissage statistique (ou « machine learning ») peut être un avantage compétitif déterminant dans ce contexte.

INTRODUCTION

Le marché de l'assurance est extrêmement concurrentiel en France, on parle même souvent d'un marché « saturé », ce qui fait peser une incertitude importante sur les marges des compagnies. Les chiffres publiés par la FFSA[1] montrent que pour les sociétés dommages, le résultat net comptable rapporté aux cotisations acquises a atteint en 2012 son plus bas niveau depuis 10 ans. La FFSA et le GEMA ont par ailleurs récemment annoncé que les résultats des sociétés dommages s'étaient dégradés en 2013. Mais au-delà des statistiques d'une année donnée, c'est la faiblesse des résultats observée depuis plusieurs années qui soulève des interrogations.

Plusieurs facteurs expliquent et peuvent accentuer ce phénomène. Tout d'abord, on observe chez de nombreux acteurs une aggravation des charges de sinistres sur certaines branches, comme l'Automobile (impact des corporels lourds) ou les Dommages aux Biens (événements climatiques récurrents), les deux principales branches d'assurance non-vie. Ensuite, l'environnement économique et financier a fragilisé les résultats de ces dernières années. Enfin et surtout, le marché de l'assurance connait de fortes évolutions : la réforme de l'ANI, la réforme du FGAO[2], la fin de la différenciation liée au critère de genre, la loi Hamon, l'entrée en vigueur de Solvabilité II, etc. Ces évolutions soulèvent de nombreuses interrogations au sein des compagnies d'assurance, notamment autour de la mesure de leurs impacts respectifs sur les résultats techniques futurs, et sur les processus à mettre en place pour le suivi et la gestion des affaires en portefeuille.

Dans ce contexte, il est crucial pour toute compagnie de savoir mener correctement des études sur le niveau de la rentabilité et sur la qualité de la souscription. **En effet, quelle que soit la qualité des processus de décision en place, une décision prise sur un diagnostic erroné ou biaisé s'avérera souvent contre-productive**.

Une des problématiques posées aux compagnies d'assurance est de pouvoir identifier les segments de clientèle fragilisant leurs résultats, ou au contraire ceux qui peuvent être créateurs de richesse : comment identifier les marchés, les portefeuilles, voire les clients, les plus rentables ? Et à l'inverse ceux pour lesquels il n'est pas judicieux d'accorder des réductions commerciales ou d'enclencher des actions marketing ?

Il est également primordial de mettre en place un suivi objectif des affaires en portefeuille, indépendamment de la structure tarifaire utilisée lors de la souscription de ces affaires : en effet analyser le portefeuille sur la base des seules variables tarifaires peut avoir un effet auto-réalisateur et conduire l'analyse à occulter des effets qui pourtant s'avèrent pertinents.

Une fois les décisions prises et les mesures mises en œuvre (en termes de souscription, de résiliations, d'augmentations ou de baisses tarifaires ciblées, etc.), il est en outre important de pouvoir analyser les effets de ces décisions : comment s'assurer « en temps réel » du caractère

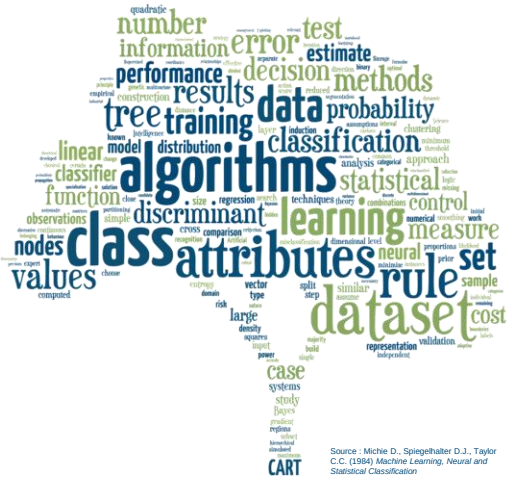
[1] Rapport annuel de la FFSA, 2012
[2] Voir notre article « La réforme FGAO, quels impacts pour l'assurance automobile », Mars 2013

approprié, ou non, des décisions prises et comment identifier dans les meilleurs délais les éventuelles actions correctrices à mettre en œuvre, par exemple auprès d'un réseau de distribution ?

En d'autres termes : comment effectuer les bons diagnostics ? Dans la plupart des compagnies, ces travaux sont aujourd'hui menés en utilisant des techniques et des indicateurs relativement standardisés. L'utilisation de méthodes innovantes et rapides à mettre en œuvre peut constituer un avantage compétitif déterminant.

L'APPRENTISSAGE STATISTIQUE

L'apprentissage statistique consiste à mettre en œuvre des algorithmes permettant de reproduire le processus d'apprentissage par l'exemple, en assimilant des informations afin d'en extraire des connaissances. L'accent est mis ici sur l'apprentissage supervisé – qui vise à attribuer une classe à des entrées – dans le cas d'une sortie quantitative.



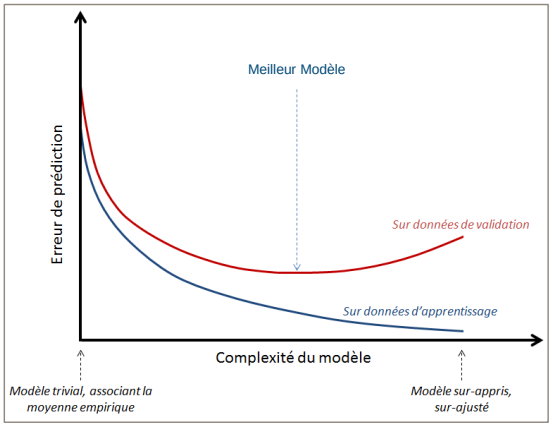
Plus concrètement, il s'agit d'étudier une variable aléatoire Y (par exemple la fréquence de sinistres) dépendant de plusieurs variables explicatives $X \in \mathbb{R}^p$ (âge de l'assuré, type de véhicule, etc.). On cherche une application ϕ qui prédirait cette variable Y sur la base des observations X, l'objectif étant de minimiser l'écart entre les valeurs observées et les valeurs prédites.

On attend de l'algorithme qu'il s'approprie les données disponibles – réalisations du phénomène inconnu à décrire – tout en conservant de bonnes performances sur de nouvelles données. C'est dans cette ambivalence que réside l'enjeu majeur des méthodes d'apprentissage : un bon modèle se doit d'apprendre les interactions entre données tout en gardant une certaine distance par rapport à un échantillon donné afin de s'assurer que celles-ci ne sont pas endogènes à cet échantillon, mais bien généralisables à d'autres. On souhaite éviter ainsi le phénomène de sur-apprentissage.

De plus, le machine learning offre par essence une grande liberté qui est à opposer aux résultats fournis par les méthodes paramétriques (comme les GLM) qui imposent une forme prédéterminée à la fonction de prédiction limitant de fait tout sur-apprentissage, mais confinant aussi le degré de compréhension qu'il est possible d'atteindre.

En lien direct avec ce constat, l'échantillon de travail est classiquement subdivisé en 3 parties :

- Échantillon d'**apprentissage** (~70 %) : c'est l'échantillon sur lequel les modèles sont construits.
- Échantillon de **validation** (~20 %) : il permet de sélectionner le meilleur modèle parmi tous ceux construits.
- Échantillon de **test** (~10 %) : ce dernier est utilisé pour tester l'adéquation du modèle optimal (au sens de la base de validation) et évaluer objectivement l'erreur commise.



Il existe de nombreuses méthodes d'apprentissage statistique : les arbres (CHAID, C4.5, CART…), les réseaux de neurones, les Support Vector Machines, etc. Nous nous concentrons ici sur les arbres de régression binaires CART (Classification and Regression Trees)[3].

Arbres de décision CART

La technique des arbres est fondée sur la classification d'un objet par une suite de tests sur les variables explicatives. Ces tests sont organisés de façon hiérarchique et récursive, puisqu'ils vont chercher à séparer l'échantillon initial en deux sous-échantillons homogènes, mais qui soient les plus distincts possibles ; puis à reproduire cette séparation sur chacun des sous-échantillons. Cette

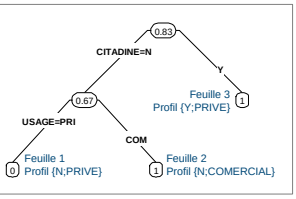
[3] BREIMAN L., FRIEDMAN J., OLSHEN R.A., STONE C.J. (1984) Classification and Regression Trees. Chapman and Hall

capacité à discriminer est basée sur l'évaluation de la réduction d'hétérogénéité, quantifiée par la variance intra-nœud définie par :

$$\widehat{\Delta} = \sum_{i=1,i\in N}^{n_N} \left(y_i - \bar{y}(N)\right)^2 - \left(\sum_{i=1,i\in N_G}^{n_{N_G}} \left(y_i - \bar{y}(N_G)\right)^2 + \sum_{i=1,i\in N_D}^{n_{N_D}} \left(y_i - \bar{y}(N_D)\right)^2 \right)$$

où $\bar{y}(N)$ est l'espérance empirique de Y sachant le nœud N. En d'autres termes, on cherche donc à tester toutes les séparations envisageables en deux nœuds distincts (N_G et N_D) afin de trouver la séparation qui maximise la réduction d'hétérogénéité $\widehat{\Delta}$. On répète cette opération jusqu'à obtenir un profil par feuille (un nœud terminal) : on parle alors d'arbre saturé. Les valeurs associées à ces feuilles sont les moyennes empiriques pour le profil identifié. On présente ci-dessous un exemple d'arbre saturé sur une base d'apprentissage restreinte de 6 individus et 3 profils.

X_1	X_2	Y
CITADINE	USAGE	NOMBRE_SIN
N	PRIVE	0
N	COMMERCIAL	1
Y	PRIVE	1
Y	PRIVE	0
N	COMMERCIAL	1
Y	PRIVE	2



Les performances des arbres dépendent essentiellement de leur taille. Un arbre trop grand et donc très spécialisé conduit à des problèmes de sur-apprentissage. On cherche ainsi à obtenir des arbres plus parcimonieux et robustes. Le principe de l'élagage est de construire l'arbre saturé sur les données d'apprentissage, puis de supprimer les nœuds pour lesquels les divisions ultérieures n'améliorent pas significativement la capacité de prédiction sur la base de validation.

Les arbres CART présentent de nombreux avantages, comme le fait d'être non paramétriques et de posséder une représentation graphique hiérarchisée avec des règles de décision simples. De plus les interactions entre les variables ne relèvent pas d'un jugement d'expert mais sont intégrées au sein même de l'algorithme, elles sont par là même non subjectives.

Méthodes d'agrégation

Afin de gagner en efficacité et en précision, le processus de construction d'un arbre peut être distribué ou répété.

Cette procédure permet de réduire la variance de l'estimation, mais conduit à une structure moins simple et tangible des arbres de décision.

Les méthodes de parallélisation comme le bagging[4] et les forêts aléatoires[5] assemblent des arbres saturés construits indépendamment les uns des autres sur des échantillons bootstrappés des données d'apprentissage. La fonction de prédiction est alors la moyenne de toutes les fonctions obtenues.

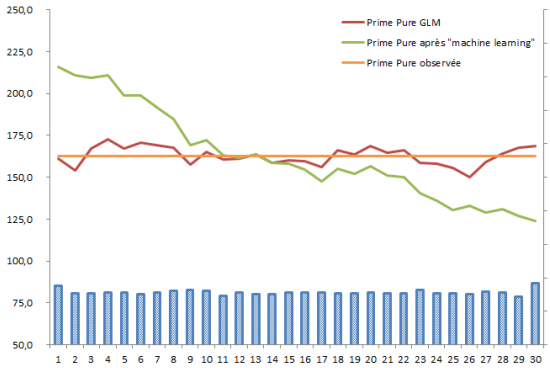
Les méthodes adaptatives comme le gradient boosting[6] ou le stochastic gradient boosting[7] sont des méthodes de descente, où chaque arbre construit dans le modèle améliore la capacité prédictive de l'arbre précédent, l'algorithme s'arrêtant ainsi quand celle-ci est jugée la plus appropriée.

ETUDES DE CAS

Ces méthodes innovantes peuvent être mises en œuvre dans de nombreux domaines. Nous présentons ci-dessous quelques exemples d'applications.

Etude de cas 1 : S'assurer du pouvoir discriminant de la structure tarifaire existante

L'utilisation des méthodes d'apprentissage statistique permet de mettre en évidence de nouveaux effets prédictifs sur une structure tarifaire existante.



L'objectif de cette étude était de vérifier la bonne adéquation de la structure tarifaire avec l'évolution du marché d'une part et des risques souscrits d'autre part.

On a donc cherché à prédire la prime pure résiduelle (l'écart entre la prime pure prédite par le GLM et la charge de sinistres observée) à partir de variables explicatives non retenues dans la

[4] BERK R.A. (2004) An introduction to Ensemble Methods for Data Analysis. Department of Statistics UCLA
[5] BREIMAN L. (2001) Random Forests. Statistics Department, Berkeley
[6] FRIEDMAN J.H. (2001) Greedy function approximation: a gradient boosting machine. The annals of Statistics
[7] FRIEDMAN J.H. (2002) Stochastic gradient boosting. Computational Statistics & Data Analysis

structure tarifaire existante. L'algorithme a identifié une quinzaine de variables explicatives non incluses dans la structure tarifaire existante afin de créer 30 profils.

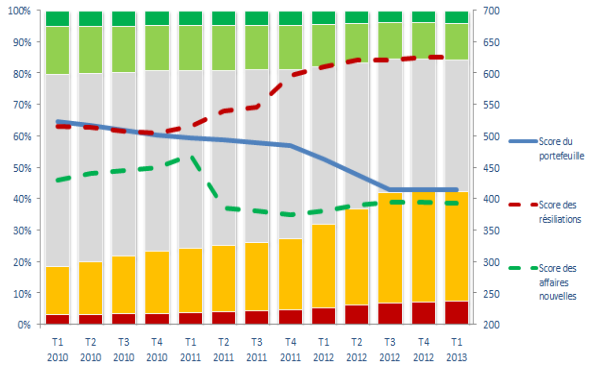
L'étude a démontré la nécessité de faire évoluer la structure tarifaire, en exhibant notamment une sous-tarification de 30% pour le profil 1 (le plus risqué) et une sur-tarification de 30% pour le profil 30 (le moins risqué).

D'un point de vue opérationnel, l'entreprise d'assurance a choisi d'ajouter à son tarif GLM une variable « score » à 30 modalités, synthétisant le pouvoir discriminant des 15 variables explicatives identifiées par l'algorithme.

Etude de cas 2 : Construire un indicateur de pilotage du portefeuille

L'entreprise que nous avons accompagnée avait observé une baisse de la rentabilité technique de son portefeuille habitation en 2012.

L'étude a été réalisée à partir de données de 9 trimestres successifs (de T1 2010 à T1 2013) et a tout d'abord permis de mettre en évidence une baisse significative du score moyen du portefeuille depuis fin 2011.



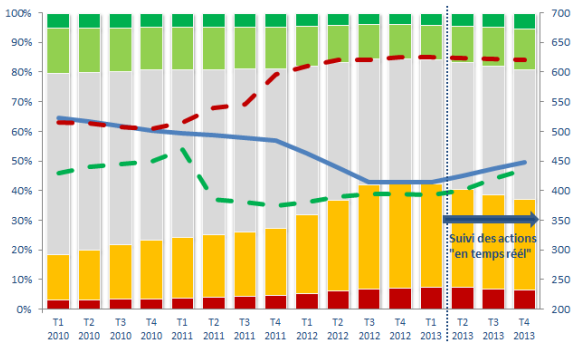
L'utilisation des méthodes d'apprentissage statistique a ainsi permis de construire un nouvel indicateur au sein de l'entreprise, sous la forme d'un score prenant des valeurs comprises entre 1 et 999 (par ordre croissant de rentabilité prédite). La rentabilité a été modélisée à la maille client en tenant compte de l'ensemble des polices détenues au sein de l'entreprise (et notamment autres que les polices habitation).

L'étude a également montré que la baisse du score moyen s'explique par une baisse de la qualité des souscriptions d'une part (dégradation du score des affaires nouvelles depuis mi-2011) et par une incapacité à retenir les clients ayant un bon score d'autre part (constatation du score des affaires résiliées). Enfin, l'étude a montré que la dégradation du score des affaires nouvelles et l'amélioration du score des résiliations ont été le signe avant-coureur de la baisse de la rentabilité observée en 2012.

Face à ce constat, l'entreprise a mis en œuvre des actions visant à sensibiliser son réseau de distribution sur la qualité de la souscription et à fidéliser les meilleurs segments de clientèle.

Nous avons par ailleurs aidé cette entreprise à mettre en place un suivi de ces actions, afin de mesurer leur effet et de s'assurer de leur pertinence.

L'introduction de l'indicateur score et l'application de l'algorithme obtenu grâce aux méthodes d'apprentissage a ainsi permis au management de suivre l'effet des actions « en temps réel » (selon un pas trimestriel). L'entreprise a observé un redressement de la qualité de la souscription au cours de l'année 2013 qui devrait se matérialiser par une amélioration de la rentabilité technique en 2014.



Cette étude peut être vue comme une première approche pour l'introduction d'un concept de valeur client au sein de l'entreprise.

CONCLUSION

Les méthodes paramétriques classiques (par exemple GLM) présupposent une forme particulière des risques et de leurs interactions, ce qui de fait limite la compréhension des subtilités du portefeuille.

Les méthodes d'apprentissage statistique, déjà largement utilisées en marketing ou en détection de fraude, apportent donc une vision nouvelle dans les domaines de la tarification et du pilotage de portefeuille.

Elles offrent ainsi, notamment dans le contexte concurrentiel actuel, une réponse aux besoins croissants de modélisation et d'analyse de données (essor du « Big data »).

A PROPOS DE MILLIMAN

Milliman figure parmi les principaux cabinets de conseil en actuariat au monde. La société intervient dans les secteurs de l'assurance-vie et des services financiers, de l'assurance non-vie, de la prévoyance et de la santé. Fondé en 1947, Milliman est une société indépendante qui compte des bureaux dans les principaux centres économiques dans le monde.

MILLIMAN EN EUROPE

Milliman s'est fortement développé en Europe et compte plus de 250 consultants au travers de ses bureaux d'Amsterdam, Bruxelles, Bucarest, Dublin, Dusseldorf, Londres, Madrid, Milan, Munich, Paris, Stockholm, Varsovie et Zurich.

milliman.com

CONTACTS

Pour toute question ou commentaire sur cet article vous pouvez vous adresser à :

Fabrice Taillieu, Principal
fabrice.taillieu@milliman.com

Sébastien Delucinge, Senior Consultant
sebastien.delucinge@milliman.com

Rémi Bellina, Consultant
remi.bellina@milliman.com