Machine Learning : Du GLM à l'arbre de CART en passant par le Random Forest

Le paysage de l'actuariat IARD est en train de changer sous l'impulsion d'une modélisation actuarielle tournée, non plus vers les modèles statistiques économétriques classiques (Modèles Linéaires Généralisés - GLM) mais vers les modèles de statistiques d'apprentissages ou Machine Learning. Ces algorithmes combinés à l'émergence de bases de données plus riches et mieux structurées multiplient les possibilités de modélisation et d'appréhension du risque. Par ailleurs, et pour exploiter au mieux les résultats de ces algorithmes, il est nécessaire de maitriser parfaitement leurs fonctionnements afin d'éviter notamment les problèmes courants de sur-apprentissage, de causalité ou d'homophilie des données.

1- Pourquoi les Machine Learning sont le facteur clé de succès de la modélisation actuarielle de demain ?

Un marché IARD en mutation

Sur ce marché connu pour son importante intensité concurrentielle, les changements technologiques (digitalisation, simulateurs de tarifs, comparateurs en ligne, etc..) et réglementaires tendent à modifier le comportement des assurés et un marché qui devient de plus en plus arbitré par les prix. Dans ce contexte, les **acteurs doivent adapter leur stratégie à ce nouvel environnement en se différenciant de la concurrence**.

Révolution des données : une richesse à exploiter

La **donnée est le nouvel or noir de l'assurance**. Le **volume et la diversité** de données disponibles, aussi bien en interne et qu'externe, se sont **accrus** notamment grâce au développement de **l'open Data** (données disponibles en ligne).

Cette **émergence de données impacte le métier** des assureurs IARD et des actuaires aussi bien en termes de pricing que de gestion de risque.

L'enjeu est double :

- **Capter et travailler les données afin de les rendre exploitables ;**
- **Choisir l'algorithme qui saura comprendre les données et les faire parler.**

Machine Learning, les algorithmes permettant de revoir les modèles actuariels à l'appui des données émergentes

Les modèles GLM, classiquement utilisés en assurance IARD ont l'avantage de permettre l'utilisation de tests statistiques pour juger de la qualité d'un modèle, mais il nécessite de faire des hypothèses a priori fortes que ce soit sur la loi de la variable à expliquer ou bien sur les interactions entre les variables explicatives. Ainsi **les modélisations statistiques classiques sont restrictives et ne sont pas adaptées à l'exploration des données : les Machine Learning le sont.**

Les Machine Learning vont permettre de **capter et de retranscrire les interactions entre les données et ainsi d'affiner l'appréhension du risque.**

2- Qu'est-ce que les Machine Learning

Le concept de Machine Learning

Principe : Les Machine Learning sont une catégorie de modèle non paramétrique dont le leitmotiv pourrait être résumé de manière schématique à la citation suivante : « la connaissance est la fille de l'expérience » (Simon de Bignicourt).

Ainsi, **le principe de ces algorithmes est de réaliser une tâche sur la base de l'expérience tirée des données**. *Exemple : En assurance non-vie, cette tâche pourrait être le scoring de la résiliation des assurés et les données utilisées pour les caractéristiques assurés.*

Avantages : Ces méthodes **ne font pas d'hypothèses fortes sur la distribution des données à expliquer**. L'unique hypothèse sur les données à expliquer est qu'elles sont identiquement et aléatoirement générées par un processus à partir des variables explicatives. Par ailleurs, elles ont l'avantage de **détecter les interactions entre les variables sans avoir à les spécifier au préalable**.

Multitudes de modèles : Les modèles de Machine Learning, les plus connus et utilisés en assurance sont les **arbres de décision**, les réseaux de neurones et les **Random Forest**.

Focus sur les arbres de décision

Avantages : L'arbre de décision est devenu une méthode très prisée au vu de la **rapidité de ses temps de calcul**, de sa capacité à gérer tous types de variables et à sélectionner les plus pertinentes, ainsi que **la lisibilité et la facilité d'interprétation des résultats**.

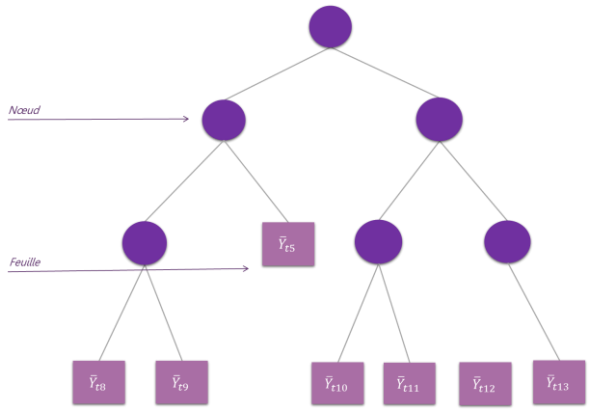
Les typologies : Il convient de distinguer deux types d'arbres de décision :

- Les arbres de régression où la variable à expliquer est quantitative. La technique de l'arbre de régression est utilisée pour répartir les individus d'une population en k sous-populations et prédire la valeur cible pour chaque population.
- Les arbres de classement où la variable à expliquer est qualitative.

Il existe de **nombreuses variantes d'algorithmes** : CART (Classification And Regression Trees), CHAID (Chi-squared Automatic Interaction Detector), les algorithmes de classification supervisée (C5, C4.5, ID3).

Fonctionnement de l'arbre CART :

L'arbre de régression CART construit des estimateurs constants par morceaux sur des partitions créées, à partir des données, par un découpage binaire récursif de l'ensemble des variables explicatives.



Exemple d'arbre de décision

L'algorithme commence par choisir la variable explicative, qui par ses modalités, découpe le mieux la population en deux groupes (nommés noeuds) en maximisant la variance inter-groupe. L'opération est réitérée jusqu'à ce qu'il n'y ait plus qu'un individu par groupe ou bien selon un critère d'arrêt à définir, obtenant les nœuds finaux appelés feuilles. Le coût prédit pour chaque feuille est la moyenne des valeurs cibles de chaque individu de la feuille.

La seconde étape consiste à minimiser une fonction prenant en compte l'erreur quadratique moyenne et le nombre de feuilles. Cette fonction permet d'optimiser le niveau de complexité de l'arbre de manière à prévenir le sur-apprentissage.

Lors de la troisième étape, l'arbre optimal est obtenu par élagage selon le paramètre de complexité.

Inconvénients : L'inconvénient principal des arbres de décision est que la classification dépend fortement de l'ordre des variables choisies ce qui peut nuire au pouvoir prédictif du modèle. Cette limite peut être rectifiée par des techniques de boosting ou bagging.

Random Forest

Les forêts aléatoires ou Random Forest sont des cas particuliers de bagging pour les arbres de décision.

Principe :

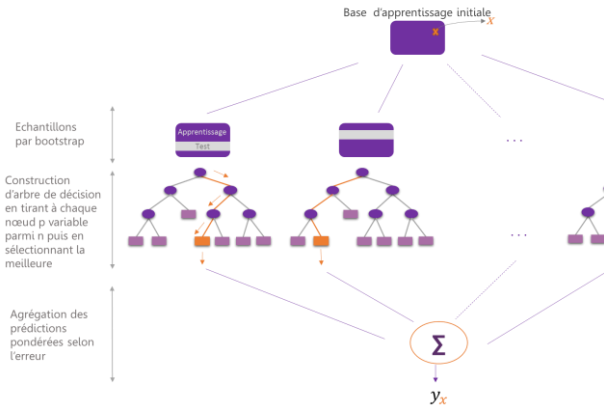
L'algorithme consiste à construire une famille d'arbres de décision sur des échantillons boostrap, suivi de l'agrégation des prédictions des modèles. Le principe de l'algorithme est de chercher, pour chaque scission, non plus la meilleure scission parmi toutes variables explicatives (n), mais la meilleure scission pour p variables explicatives tirées aléatoirement parmi n. Cette double randomisation a été introduite par L. BREIMAN.

Paramètres à optimiser :

Plusieurs paramètres sont à optimiser afin d'obtenir les meilleurs résultats :

- Le nombre de variables sélectionnées par scissions. *La littérature préconise, dans le cas de problèmes de régressions, que le nombre de variables sélectionnées aléatoirement soit égal à la partie entière du tiers du nombre de variables explicatives et ne soit pas inférieur à 5.*
- Le nombre de feuilles de chaque arbre. *Contrairement au bagging simple, il a été démontré que le Random Forest peut être appliqué avec succès à des arbres limités à deux feuilles.*
- Le nombre d'arbres dans la forêt. *Pour le nombre de modèles agrégés, la convergence de performance est*

atteinte avec un nombre de modèles agrégés à p parmi n. Ceci sous-entend que le nombre d'arbres dans la forêt croît avec le nombre de variables.



3- Machine Learning dans la pratique

Avec l'utilisation des modèles de Machine Learning, les approches deviennent itératives et adaptatives. La nouvelle culture de l'algorithme déployé s'accompagne de tout un écosystème actuariel qui se développe autour.

Dimension juridique

Les projets de Data Science sont contraints par la réglementation sur la Data Privacy qui définit et régit l'utilisation des données personnelles, notamment avec la nouvelle directive européenne (UE) 2016/680 entrée en vigueur le 24 mai 2016 qui sera en application à partir du 25 mai 2018.

Dimensions Managériales

Les Machine Learning sont des modèles complexes qui utilisent des données complexes, et offrent de nombreuses possibilités. Plus encore que lors d'études GLM ou autres, il convient de définir en amont les objectifs, des KPI pertinents afin d'obtenir des résultats cohérents et maitrisés.

Dans un tel projet, il est préférable de définir, en même temps, un cadrage stratégique, fonctionnel et SI alignant les différentes parties sur une même ambition, un même objectif, et de mêmes cas d'usage. Il est recommandé de notamment clarifier l'ambition du projet, quelles sont les données qui ont de la valeur, et qu'est-ce qui peut être fait pour que ces données répondent à la qualité attendue.

Au niveau de l'entreprise, les projets Data Science nécessitent une transformation des pratiques et des processus de l'entreprise pour tendre vers plus de transversalité. Le développement de projet Data Science, et particulièrement les projets Big Data, tend à un travail sous forme de projet en équipe pluridisciplinaire et donc décloisonner les différents utilisateurs des différentes données. Cela sous-entend également une multiplication des interactions entre les services et l'utilisation des différents métiers ayant des visions différentes dans les données. Ceci passe par la transformation de l'organisation du travail qui tend vers une synergie et un partage des compétences

Dimension opérationnelle

L'utilisation des algorithmes Machine Learning nécessite, certes une maitrise des mathématiques et techniques des méthodes utilisés, mais également des logiciels associés (R, Python).

L'approche Machine Learning est différente des approches classiques usuelles et à ce titre est associée à des interrogations qui lui sont propres :

- Une attention particulière doit être apportée à la qualité des données et au sens de ces données. **Le travail des bases de données est une étape décisive** aux cœurs des projets de Data Science. Si avant l'attention s'est beaucoup portée sur le volume, l'enjeu des assureurs concerne désormais la variété des données et leur sens. Les projets de Machine Learning ne nécessitent pas le recours systématique à des données externes, les données internes, structurées ou non, sont une source d'information riche, et constituent avec la question de leur qualité un sujet à part entière.

- Un autre point de réflexion concerne le **problème de sur-apprentissage des données**. Plusieurs méthodes existent pour le limiter, par exemple la cross validation et le tunning des critères des algorithmes.

  Exemple : un paramètre à optimiser pour l'arbre de décision est le critère d'arrêt.

- Plus les algorithmes sont complexes, plus le risque opérationnel associé est important. Si la tendance actuelle, particulièrement en tarification, est une segmentation toujours plus fine, **il convient de définir le degré de segmentation adaptée**.
- Ces modèles, plus complexes que les méthodes statistiques, sont souvent associés à des boites noires de par la subtilité des algorithmes et de leurs paramétrages. **L'enjeu de compréhension des Machine Learning est alors doublé d'une**

problématique d'interprétations et de partage des résultats.

Enfin, **l'approche Machine Learning n'est pas toujours simple à mettre en œuvre d'un point de vue opérationnel** (contraintes informatiques, gestion des modèles dans le temps, etc...). Dans ce contexte, pour répondre aux contraintes opérationnelles, les Machine Learning peuvent être utilisés de manière croisée aux approches classiques pour affiner les hypothèses et choix a priori des approches historiques.

Rédigé par Julie,

Membre de l'équipe Périclès Actuarial