Data Science et .. Machine Learning

Exemples d'application en assurance non vie

Novembre 2016

Rédigé par Julie,
Membre de l'équipe Périclès Actuarial

Machine Learning en IARD : des applications concrètes

Comment les utiliser malgré les contraintes internes

Dans un marché de plus en plus **transparent et arbitré par les prix** les acteurs doivent adapter leur stratégie.

En tarification IARD, il faut pour cela repenser les modèles pour mieux comprendre le risque et affiner la segmentation associée : **l'innovation est alors un enjeu stratégique.**

Si **les Machine Learning offrent aujourd'hui de nombreuses possibilités** et ont montré leur suprématie par rapport aux modèles classiques GLM dans diverses études…

..ils **restent encore peu utilisés en remplacement des classiques modèles tarifaires GLM.** Ceci est notamment dû, à des contraintes informatiques (système de production inadaptée).

Une manière d'utiliser le potentiel des Machine Learning sous contrainte de modèle tarifaire figé **est de les exploiter pour affiner les segmentations des variables tarifaires et identifier de nouvelles variables, etc ..**

Machine Learning en IARD : des applications concrètes

Cas 1- Application dans le cas d'une assurance automobile (1/2)

- Dans le cas d'une assurance automobile, **les Machine Learning sont un outil innovant et performant pour améliorer la segmentation et construire par exemple un véhiculier** (classement des véhicules en groupe de risque homogène).

- Le véhiculier est au cœur des enjeux de segmentation. Il permet en effet à l'assureur de se différencier de la concurrence et peut, selon **la méthode de construction retenue,** présenter un **triple intérêt :**
  - Améliorer le modèle tarifaire, en réduisant, par exemple, la variance inexpliquée ;
  - Capter de l'information sur le risque à partir des caractéristiques techniques du véhicule ;
  - Capter des informations sur le comportement de l'assuré à partir du véhicule qu'il conduit.

- L'un des **enjeux du véhiculier** est également de pouvoir intégrer **les nouveaux véhicules** ou de manière plus large les véhicules absents de l'étude au moment de sa construction.

Machine Learning en IARD : des applications concrètes

Cas 1- Application dans le cas d'une assurance automobile (2/2)

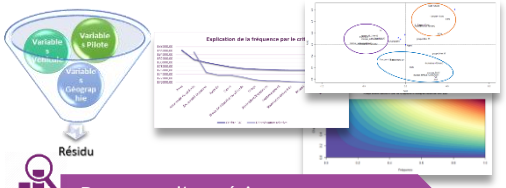
- Exemple de méthodologie développée pour construire un véhiculier réduisant la variance inexpliquée dans le cas d'une base de données réduite et avec une forte hétérogénéité à partir d'un algorithme de Machine Learning : **l'arbre de décision CART (régression)**

PHASE 1 : Approche résiduelle

Isolement de la part du risque expliquée par les autres facteurs que le véhicule

GLM pour le coût et la fréquence

$$g(E[Y|X_1,\dots,X_p]) = \beta_0 + \sum_{k=1}^{I} \beta_k X_k$$



Retours d'expérience

- Les variables explicatives doivent être choisies en tenant compte entre autres du lien entre le conducteur et le véhicule
- Le choix de la loi paramétrique et de la fonction lien doit tenir compte de l'utilisation souhaitée des résidus (notion de distance)
- Un classement en coût x fréquence va permettre d'obtenir un véhiculier pour chaque dimension et de comparer le véhiculier construit aux classements SRA

PHASE 2 : Utilisation de la crédibilité pour gérer l'hétérogénéité des données

Utilisation détourner d'une méthode de crédibilité pour définir les véhicules dont l'information est fiable qui serviront de base d'apprentissage à l'algorithme de Machine Learning

Modèle Bühlmann-Straub

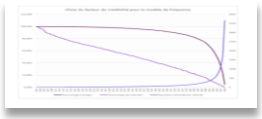
$$\widehat{\mu(\theta_i)} = Z_i \bar{X}_i + (1 - Z_i)\mu_0$$

$$\mu_0 = \sum_{i=1}^{I} \frac{Z_i}{Z_\bullet} X_i$$

$$Z_i = \frac{w_{i\bullet}}{w_{i\bullet} + \frac{\sigma^2}{\eta^2}}$$

$$w_\bullet = \sum_{i=1}^{I} Z_i$$

Détermination d'un seuil du facteur de Bühlmann-Straub à partir duquel les véhicules est considéré comme crédible



Retours d'expérience

- L'étape de crédibilité a permis d'améliorer l'apprentissage des arbres et de manière plus large de construire un véhiculier plus pertinent indépendamment de la maille
- Le seuil pour le facteur de crédibilité et la méthode pour le définir doivent être choisis avec soin selon l'algorithme de Machine Learning a utilisé et les données et tenant compte de la perte d'information sur la base d'apprentissage

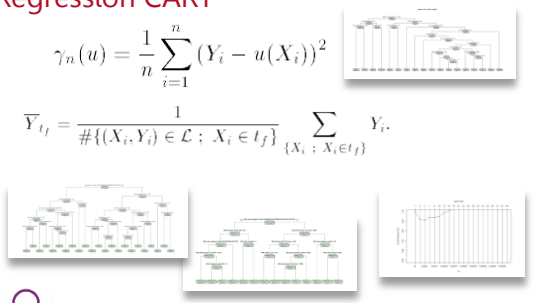
PHASE 3 : Machine Learning pour classer les véhicules en risque homogène

Explication du résidu des véhicules « crédibles » d'après la phase 2 par les variables véhicules à l'aide d'algorithme Machine Learning

Régression CART

$$\gamma_n(u) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - u(X_i))^2$$

$$\overline{Y}_{t_f} = \frac{1}{\#\{(X_i, Y_i) \in \mathcal{L} \,;\, X_i \in t_f\}} \sum_{\{X_i \,;\, X_i \in t_f\}} Y_i.$$



Retours d'expérience

- La régression CART permet d'isoler le signal du bruit et de créer directement des classes de véhicules
- Les arbres de décisions ont l'avantage de créer des règles explicites qui pourront être utilisées directement pour classer les « futurs » véhicules
- Le nombre de classes est enjeu, il faut tenir le degré de segmentation pour éviter l'épineux problème de sur-apprentissage récurrent dans l'étude des Machine Learning mal maitrisée

Machine Learning en IARD : des applications concrètes

Cas 2 - Application dans le cas d'une assurance MRH ou Auto (1/2)

- Dans le cas d'une assurance MRH et automobile, **les Machine Learning sont un outil innovant et performant pour améliorer la segmentation par exemple le zonier** (classement des zones en groupe de risque homogène).

- Le zonier est également au cœur des enjeux de la segmentation. Selon le même principe que le véhiculier, il permet de se différencier et de la concurrence et selon **la méthode de construction** il présentera également un **triple intérêt :**
  - Améliorer le modèle tarifaire, par exemple en réduisant la variance inexpliquée ;
  - Capter de l'information sur le risque à partir de la zone géographique
  - Capter des informations sur le comportement de l'assuré selon ses trajets ou son lieu d'habitation

- Avec le développement d'Open Data (sociologique, météorologique, open BAAC, …) et d'applications récoltant de l'information sur l'état des route (Waze), les possibilités d'amélioration du zonier sont multiples :
  - Développement de modèles pour pricer au trajet
  - Création d'un zonier pour les sinistres graves à l'aide de données externes

Machine Learning en IARD : des applications concrètes

Cas 2 - Application dans le cas d'une assurance MRH ou Auto (2/2)

- Exemple de méthodologie développée qui permet de construire un zonier réduisant la variance inexpliquée dans le cas d'une base de données réduite avec une forte hétérogénéité à partir d'un algorithme de Machine Learning : **le Random Forest**

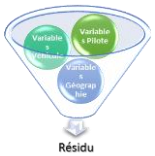
PHASE 1 : Approche résiduelle

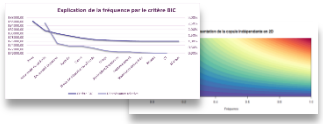
Phase similaire au cas 1

Isolement de la part du risque expliquée par les autres facteurs que le géographie

GLM pour le coût et la fréquence

$$g(E[Y|X_1,\dots..X_p]) = \beta_o + \sum_{k=1}^{1} \beta_k X_k$$





Retours d'expérience

- Une attention particulière doit être portée à la définition de la maille géographique utilisée pour l'étude
- Le choix de la loi paramétrique et de la fonction lien doit tenir compte de l'utilisation souhaitée des résidus (notion de distance)

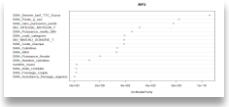
PHASE 2 : Machine Learning pour expliquer la partie résiduelle

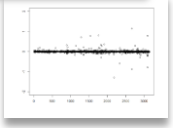
Explication du résidu par les variables explicatives à l'aide d'un algorithme Machine Learning

Random Forest

FORMULE SUR L'IMPORTANCE DES VARIABLES

$$VI(X^j) = \frac{1}{q}\sum_{l=1}^{q}\left(\widetilde{errOOB_l}^{j} - errOOB_l\right)$$





Retours d'expérience

- Le Random Forest permet d'isoler le signal du bruit
- La maille d'analyse doit être choisie avec soin selon les données, tout particulièrement dans le cas de données avec une forte hétérogénéité
- L'importance des variables donne une indication sur les variables géographiques les plus significatives
- Il est peut-être intéressant de sélectionner les variables explicatives dans le cas où de très nombreuses variables sont testées

PHASE 3 : PHASE 3 : Lissage spatial et clustering

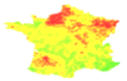
Utilisation de la crédibilité pour gérer le problème d'hétérogénéité et récupérer l'information pour les zones sous-représentés

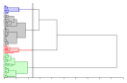
Lissage spatiale et clustering

METHODE DE LISSAGE Distance(Credibilité Weighted)

$$R_i = V_i r_i + (1 - V_i)\overline{r_i}$$

$$\overline{r_i} = \frac{\sum_{k\neq i} r_j d_{ik}^{-P} w_k}{\sum_{k\neq i} d_{ik}^{-P} w_k}$$





METHODE WARD

$$I_t = \frac{1}{n}\sum^{n} d^2(e_i, g)$$

$$I_\varepsilon = \frac{1}{n}\sum_{i=1}^{k} n_i \times d^2(g_i, g)$$

Retours d'expérience

- L'étape de lissage spatial permet de construire un zonier plus pertinent
- Le degré de lissage et la méthode pour le définir doivent être choisis avec soin selon les données
- La méthode de clustering doit être définie de manière à ce que la composition des groupes soit peu sensible à l'ajout d'un groupe supplémentaire