



Data Science et Machine Learning

Exemples d'application en assurance non vie

Novembre 2016



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).

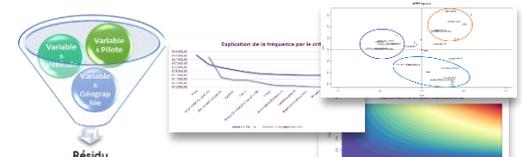
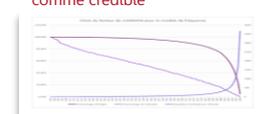
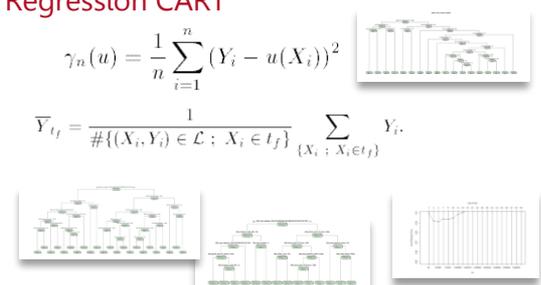
- Le véhiculier est au cœur des enjeux de segmentation. Il permet en effet à l'assureur de se différencier de la concurrence et peut, selon **la méthode de construction retenue**, présenter un **triple intérêt** :
 - ▶ Améliorer le modèle tarifaire, en réduisant, par exemple, la variance inexpliquée ;
 - ▶ Capturer de l'information sur le risque à partir des caractéristiques techniques du véhicule ;
 - ▶ Capturer 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	PHASE 2 : Utilisation de la crédibilité pour gérer l'hétérogénéité des données	PHASE 3 : Machine Learning pour classer les véhicules en risque homogène
<p>Isolement de la part du risque expliquée par les autres facteurs que le véhicule</p>	<p>Utilisation détournée 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</p>	<p>Explication du résidu des véhicules « crédibles » d'après la phase 2 par les variables véhicules à l'aide d'algorithme Machine Learning</p>
<p>GLM pour le coût et la fréquence</p> $g(E[Y X_1, \dots, X_p]) = \beta_0 + \sum_{k=1}^p \beta_k X_k$ 	<p>Modèle Bühlmann-Straub</p> $\widehat{\mu}(\theta_i) = Z_i X_i + (1 - Z_i) \mu_0$ $\mu_0 = \frac{\sum_{i=1}^I Z_i X_i}{\sum_{i=1}^I Z_i}$ $Z_i = \frac{w_i \bullet}{w_i \bullet + \frac{\sigma^2}{\mu^2}}$ $w_i \bullet = \sum_{j=1}^J Z_{ij}$ <p>Détermination d'un seuil du facteur de Bühlmann-Straub à partir duquel les véhicules est considéré comme crédible</p> 	<p>Régression CART</p> $\gamma_n(u) = \frac{1}{n} \sum_{i=1}^n (Y_i - u(X_i))^2$ $\bar{Y}_{I_j} = \frac{1}{\#\{(X_i, Y_i) \in \mathcal{L} : X_i \in I_j\}} \sum_{\{X_i, Y_i \in I_j\}} Y_i$ 
<p>Retours d'expérience</p> <ul style="list-style-type: none"> ▶ 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 	<p>Retours d'expérience</p> <ul style="list-style-type: none"> ▶ 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 	<p>Retours d'expérience</p> <ul style="list-style-type: none"> ▶ 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 en jeu, il faut tenir le degré de segmentation pour éviter l'épineux problème de sur-apprentissage récurrent dans l'étude des Machine Learning mal maîtrisée

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 ;
 - ▶ Capturer de l'information sur le risque à partir de la zone géographique
 - ▶ Capturer 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 routes (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

- 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**

