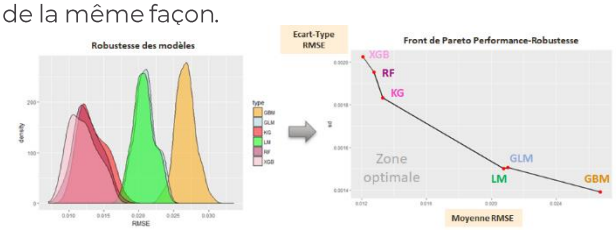
DATA INSIGHTS

PRICING

Exposition et Machine Learning

Comment prendre en compte l'exposition en ML ?

Encore aujourd'hui, une forte majorité d'acteurs utilise des méthodes GLM ou GAM. Les méthodes de Machine Learning réputées pour leur performance et leur adaptivité les concurrencent. Cependant, en Machine Learning, **la question de la prise en compte de l'exposition est rarement traitée**, alors que celle-ci a un impact majeur dans les modélisations. Aussi, en fonction des méthodes, l'exposition ne s'introduit pas de la même façon.



Pourquoi utiliser un offset en GLM ?

Un offset (exposition) est une variable dont le coefficient est arbitrairement fixé à 1 dans les modèles GLM.

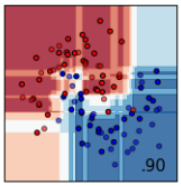
$$\mathbf{E(N)} = g^{-1}(\boldsymbol{\beta}.f(X_i) + \mathbf{1} \times \boldsymbol{offset})$$

Très utilisé dans les modèles GLM log-poisson pour son interprétabilité, son enjeu est **de pondérer par rapport à la durée de couverture d'un contrat pour mettre au même niveau les différentes observations.** Par exemple, d'un point de vue tarification, cela permet de prendre en compte des individus dont la durée dans le portefeuille diffère ou bien de faire des modèles sur des nombres d'individus différents d'une ligne à l'autre. L'objectif est donc d'ajouter une information importante avec un effet fixe **modifiant l'espérance mais aussi la variance**. Cependant le cadre statistique et l'utilisation de coefficients disparaissent ou se complexifient lors de l'utilisation par exemple **de méthodes Deep Learning ou Random Forest**.

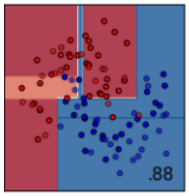
En tarification, pour un usage pratique, les assureurs doivent vérifier que les primes proposées sont annuelles. Ainsi en fixant l'exposition à une année, lors de la souscription, la prime proposée correspond à une couverture d'une année entière. Il est nécessaire de **conserver cette commodité**.

Les arbres CART ou forêt aléatoire

En régression, ajouter l'**exposition en tant que variable** dans un modèle CART ou un Random Forest ne permet que difficilement son utilisation en tant qu'offset. En effet, certaines variables pourraient être corrélées à l'exposition, et conduire ainsi à une interprétation de l'offset différente de celle attendue (effet linéaire).



De plus, **l'utilisation de poids n'est pas adaptée pour un offset**, cela revient dans certains cas à du Weighted Least Square Regressions. Si transformer la variable réponse est une solution simple, elle reste limitée à des expositions à faible variance, sinon elle **biaise fortement les résultats** .



Une alternative serait d'utiliser des **Distributed Random Forest avec l'exposition incorporée dans les fonctions de coûts,** prenant en compte les expositions à chaque split, d'autant plus qu'ils sont naturellement adaptés...

L'exposition avec la méthode GBM

Les **Gradient Boosting Machines** sont des méthodes où l'exposition peut être intégrée dans la modélisation, notamment via des fonctions de bases semblables aux GLMs. Le choix de cette méthode est très pertinente pour conserver **l'usage traditionnel** de l'offset.

$$\widehat{F}(x) = F_0 + \sum_{m=1}^{M} \gamma_m \, h_m(x)$$

Quels résultats avec les Neural Networks ?

La réflexion s'apparente à celle des Forêts de Brieman. Une solution intéressante est **d'introduire l'exposition seulement dans la fonction de coût** et non dans les neurones.
Dans le cas contraire, la rétro propagation des erreurs compterait doublement les expositions. Cependant, il reste une **question ouverte sur l'ajout ou non des faibles expositions** pouvant biaiser les résultats ou les performances de convergence.

$$\frac{\delta C(offset)}{\delta a^L} \times \frac{\delta a^L}{\delta z^L} \times \frac{\delta z^L}{\delta a^{L-1}}$$

Cette publication a été réalisée sous la direction de
Nabil RACHDI, Head of Data Science

avec l'expertise de :
Pierre CHATELAIN

addactis
THE RISKTECH FOR INSURANCE