info TECH

Actuariat et Data Science : une convergence indispensable

Nabil RACHDI

RESPONSABLE DATA SCIENCE

nabil.rachdi@actuaris.com

Nous avons souhaité, dans cette Infotech, établir un bref état des lieux des nouveaux usages liés à l'utilisation grandissante de la data science dans notre secteur de l'assurance. Nous donnons tout d'abord un bref détour historique de certains termes employés aujourd'hui devenus des « buzzwords » : Big Data, Intelligence Artificielle, Machine Learning, Deep Learning ou encore Statistical Learning. Une certaine confusion demeure encore entre ces termes, notamment auprès du grand public, et est entretenue avec l'emploi de ces différentes appellations (en particulier « Big Data ») par des médias, sans qu'ils précisent ce que ces termes recouvrent réellement…

Très utilisés en assurance, ces concepts datent des années 1990 pour les plus récents, seul le nouveau *branding* de ces techniques laisse penser que ces approches sont totalement novatrices. Nous expliquerons ensuite pourquoi et comment appliquer ces méthodes à l'univers de l'assurance afin de proposer des solutions innovantes.

UN PEU D'HISTOIRE

1940-1950 : Les débuts de l'Intelligence Artificielle (IA) : le mythe du Robot humain

Travaux pionniers d'Alan Turing et premiers modèles neuronaux par W. McCulloch et W. Pitts.

Intelligence Artificielle

1950-1960 : Officialisation de l'Intelligence Artificielle et Naissance du *Machine Learning*

Le Workshop de Dartmouth College (US) en 1956 a scellé la discipline de l'Intelligence Artificielle ou l'apprentissage de « comportements ». Naissance de la sous-discipline *Machine-Learning* pour l'apprentissage à partir de « données ».

Machine Learning

1960-1980 : Essoufflement de L'IA et expansion du *Machine Learning*

L'Intelligence Artificielle montre des limites dans l'apprentissage du comportement humain.

Statistical Learning

1980-2000 : Du *Machine Learning* au *Statistical Learning* [début des GLM en assurance]

Rapprochement du *Machine Learning* avec le monde des probabilités statistiques et développement des modèles par arbres *CART, Gradient Boosting, Random Forest,* etc.

Data Science
Big Data
Deep Learning

2000-Aujourd'hui : *Data Science, Big Data, Deep Learning* et… Intelligence Artificielle

W. Cleveland définit en 2001 la ***Data Science*** comme **« l'extension du domaine de la statistique à divers champs d'application »**. Le *Big Data* et *Deep Learning* sont des termes introduits par les GAFA pour le traitement et l'analyse de données massives.

Intelligence Artificielle : le retour !

L'ACTUAIRE : UN DATA SCIENTIST AUGMENTÉ ?

Aujourd'hui, la plupart des travaux actuariels (de la Tarification au Provisionnement, et à l'évaluation du besoin de capital) utilisent des méthodes d'apprentissage statistique apparentées aux méthodes de Data Science et de *Machine Learning.*

Pourquoi parle-t-on d'apprentissage ?

Le cadre de l'apprentissage enrichit le cadre statistique traditionnel, en proposant des méthodes d'analyse nécessitant moins d'hypothèses. Ces techniques vont extraire un maximum d'information disponible à partir des données elles-mêmes, avec le moins d'a priori possible.

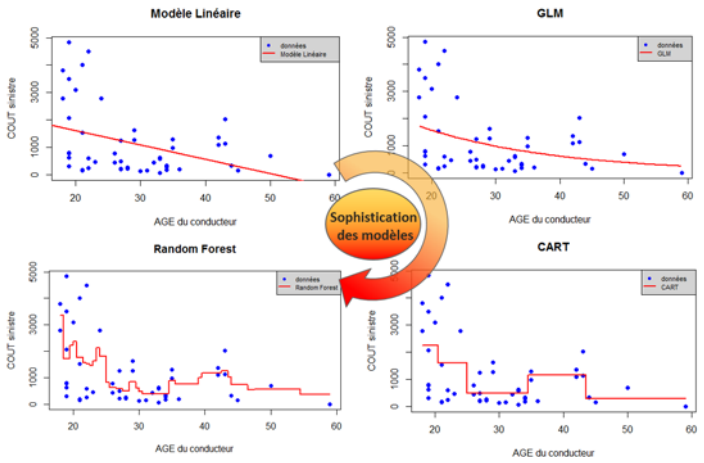
L'actuaire, par nature spécialiste de ces techniques et expert en assurance, est le profil le plus approprié pour être moteur de la transformation numérique qui s'opère aujourd'hui dans notre secteur.

L'enjeu de l'actuaire – data scientist est d'utiliser simultanément sa connaissance métier et son expertise en Data Science afin de dégager des approches innovantes avec des modèles traditionnels augmentés par des modèles de *Machine Learning.*

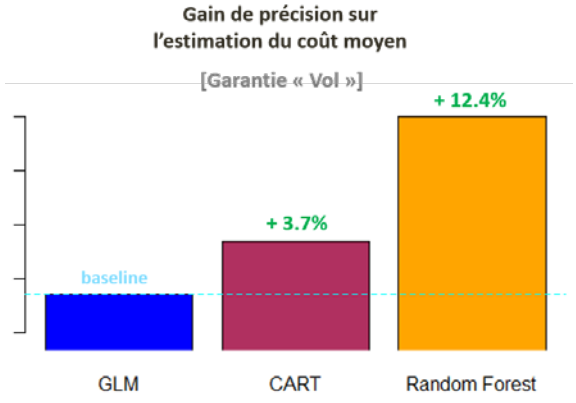
L'ACTUARIAT AUGMENTÉ PAR L'EXEMPLE

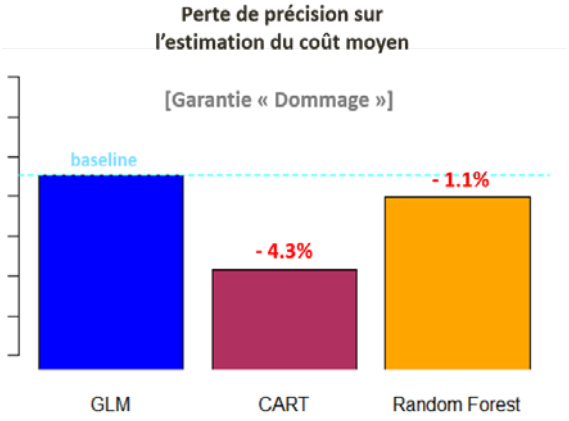
En Tarification des risques de masse non-vie, la méthode traditionnellement utilisée par les actuaires se base sur les GLM *(Generalized Linear Models)*. **L'actuaire augmenté dispose de modèles d'apprentissages supplémentaires (arbres *CART, Random Forest, Gradient Boosting, XGBoost,* etc.) qui pourront alimenter et compléter son analyse de la sinistralité.**

Le niveau de sophistication des modèles de *Machine Learning* peut permettre une explication plus fine de la sinistralité. En particulier, les modèles les plus avancés sont capables de capter une certaine complexité (interactions entre variables, non-linéarités, etc.), ce qui est impossible pour les modèles linéaires classiques. En revanche, les modèles de *Machine Learning*, plus sophistiqués, sont aussi plus flexibles et ne font pas ou très peu d'hypothèses de modélisation (arbres *CART, Random Forest, XGBoost, etc.*). Ces modèles sont capables d'apprendre des relations très complexes entre les données.



Pour la tarification de certaines garanties (« Vol », « Incendie », etc.), les modèles de *Machine Learning* peuvent améliorer assez significativement l'estimation du coût moyen des sinistres faite par les techniques classiques *(GLM)*. Il s'agit en particulier des garanties pour lesquelles la distribution des coûts est assez complexe (non standard), et la relation entre les variables explicatives et le coût comporte de fortes variabilités : c'est par exemple le cas de la garantie « Vol » en assurance automobile comme illustré. Nous constatons une nette amélioration de la précision de l'estimation du coût moyen des sinistres avec des modèles de *Machine Learning* (un gain de précision de plus de 12% comparé à une méthode *GLM*). Ainsi, une méthode type *GLM* n'est pas adaptée pour une problématique de cette nature.



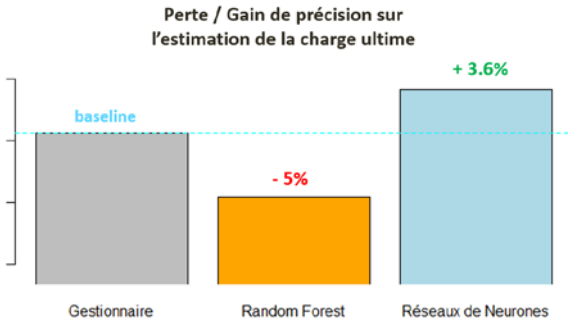


Toutefois, les méthodes *GLM* ne s'avouent pas vaincues pour autant. Sur d'autres types de garanties (« Dommage », « Dégât des Eaux », etc.), les modèles de *Machine Learning* n'améliorent pas l'estimation du coût moyen. C'est souvent le cas pour la garantie « Dommage » en assurance automobile par exemple, où nous pouvons constater des résultats nettement différents de ceux de la garantie « Vol » : en effet, les modèles de *Machine Learning CART* et *Random Forest* sont moins précis que le modèle obtenu par *GLM*. Ce constat s'explique en grande partie par le fait que, pour la garantie « Dommage », les coûts sont distribués selon une loi facilement ajustable (souvent Gamma) permettant de construire une modélisation fine directement à partir d'un modèle *GLM*.

Provisionnement

Dans le cadre de provisionnement de sinistres, les algorithmes de *Machine Learning* peuvent s'avérer pertinents pour l'estimation de la Provision dossier par dossier à l'ouverture d'un sinistre, estimation établie en pratique par les gestionnaires. Il est en effet possible de construire des modèles d'apprentissage qui prédisent la charge ultime, ce qui contribue à **donner des éléments quantitatifs supplémentaires aux gestionnaires pour affiner leurs prévisions.**

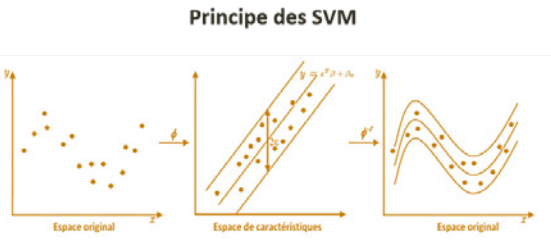
En provisionnement, tous les modèles de *Machine Learning* ne présentent pas les mêmes performances de prédiction. En effet, le contexte de la provision dossier/dossier induit une structure particulière des données, notamment sur la relation entre les variables explicatives (type de sinistre, cause, temps de survenance/déclaration, etc.) et la charge ultime. Les Réseaux de Neurones multicouches donnent des résultats satisfaisants pour ce type de problématique.



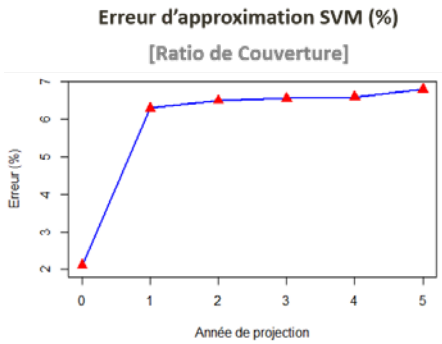
Evaluation du besoin de capital

Le temps de calcul d'un ratio de couverture (capital économique sur fonds propres) en assurance vie peut prendre plusieurs jours, ce qui rend l'évaluation de stress tests très couteux en temps de calcul. Les méthodes de *Machine Learning* permettent de contourner cette difficulté en approximant les différentes briques de risque qui composent le ratio de couverture.

En effet, à partir d'un jeu de simulations sur plusieurs configurations bien spécifiées, il est possible de construire une interpolation du capital économique grâce au *Machine Learning*. Les *Support Vector Machine (SVM)* peuvent par exemple être employés à cette fin, ce qui réduit significativement les temps de calcul : **on dispose alors d'une approximation du capital économique en quelques heures au lieu de plusieurs jours.**



Grâce à cette optimisation du temps de calcul, au prix d'une approximation maîtrisée, on obtient des premières tendances du besoin en capital selon plusieurs scenarii de stress, en un temps raisonnable, ce qui permet de rejouer des simulations en changeant d'hypothèses.



Le cadre de l'apprentissage est un cadre flexible qui permet de concevoir des algorithmes alternatifs aux techniques classiques, en donnant la possibilité de mieux les adapter aux besoins du problème, et en espérant une meilleure prédictivité (meilleure estimation de la sinistralité, du parcours client, etc.) ou gain d'information (identification des caractéristiques les plus influentes d'un sinistre, etc.).

DATA SCIENCE POUR L'ASSURANCE : RÉSOUDRE LE DILEMME PRÉDICTIVITÉ/INTERPRÉTATION

Machine Learning : doit-on privilégier un modèle de prédiction précis avec une interprétation limitée, ou plutôt lui préférer un modèle prédictif moins performant mais plus transparent, interprétable et plus largement exploitable ?

Lors de l'utilisation des méthodes de Machine Learning, **se pose systématiquement le dilemme prédictivité/interprétation.**

En reprenant l'exemple des méthodes *GLM* illustrées plus haut, une méthode *GLM* peut s'avérer moins performante en termes de prédictivité que des méthodes plus sophistiquées telles que les méthodes *Random Forest* et *XGBoost* par exemple. Toutefois, l'actuaire reconnait bien l'aisance et la facilité d'interprétation d'un tarif établi par *GLM*, il s'agit d'une boîte « blanche » : on connait parfaitement la structure du modèle de prédiction, la structure tarifaire est « multiplicative », le modèle est exportable dans les systèmes d'information, etc. De telles interprétations ne sont pas possibles avec des modèles de *Machine Learning avancés*.

Passer de

« Méthodes actuarielles OU Machine Learning avancé ? »

à

« Méthodes actuarielles ET Machine Learning avancé »

Une manière d'appréhender ce dilemme prédictivité/interprétation en assurance est d'adopter **une vision collaborative des modèles de *Machine Learning* avancés avec les modèles traditionnels**. En d'autres termes, l'idée n'est pas a priori de « remplacer » les méthodes existantes par ces nouvelles techniques mais d'abord de bénéficier **d'apport intelligent d'information grâce au *Machine Learning*.**

Le développement de ces modèles traditionnels « augmentés » fait apparaitre de nouveaux défis aux actuaires : celui du choix et de l'adaptation des modèles de *Machine Learning*. En effet, un principe bien connu dans le domaine de l'optimisation prévaut également en *Machine Learning* : le **« No Free Lunch Theorem »**. Ce principe stipule **qu'il n'y a aucune méthode ou aucun algorithme universel permettant une résolution optimale de tous les problèmes.** Autrement dit, un algorithme peut s'avérer très performant sur un problème particulier donné, et avoir de moins bonnes performances sur un autre type de problème. C'est le cas par exemple pour la modélisation des différentes garanties illustrées plus haut, qui induisent des problèmes d'apprentissage différents.

CONCLUSION

Une récente enquête menée par *Kaggle*, une plateforme web de compétition sur des sujets en Data Science, qui recense les techniques de *Machine Learning* les plus utilisées, donne en tête du classement la Régression Logistique. Ensuite viennent les Arbres de Décision, les Forêts Aléatoires et les Réseaux de Neurones. La Régression Logistique est une technique utilisée depuis plus de 20 ans par les actuaires : par exemple en optimisation tarifaire, pour la segmentation de portefeuilles, etc.

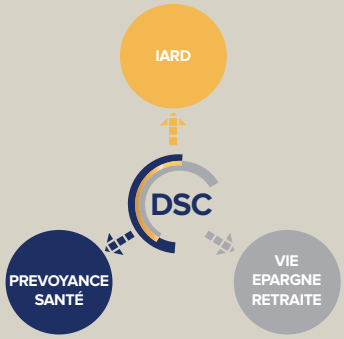
Les techniques et principes liés à la Data Science ne sont donc pas une révolution en soi : les concepts mathématiques utilisés les plus récents datent des années 90.

La véritable innovation est sans doute dans l'imagination* et dans l'adaptation de ces techniques dans les différents champs d'application (notamment ceux qui « ignoraient » de telles approches), en profitant de l'abondance de l'Open Data accessible au public. Le domaine de l'assurance contribue en particulier à la Data Science avec sa propre vision de la donnée, de sa modélisation, ce qui démultiplie les usages et interactions des différents algorithmes de *Machine Learning*, contribuant ainsi à favoriser l'innovation des futurs produits d'assurance.

** **N'hésitez pas à contacter notre Responsable Data Science, Nabil RACHDI, pour imaginer avec vous des approches Data Science innovantes adaptées à vos besoins : nabil.rachdi@actuaris.com***

Le Data Science Center d'ACTUARIS

Le Data Science Center d'ACTUARIS, transverse aux différents pôles métier du cabinet, est le centre de compétence dédié aux techniques innovantes d'analyse de données assurantielles. Nous soutenons un effort important en Recherche et Développement avec une équipe multidisciplinaire composée d'actuaires-data scientists, de data scientists et d'experts informatiques.



ACTUARIS ADDACTIS Group Risk & insurance solutions
www.actuaris.com

ACTUARIS PARIS | 13/15 boulevard de La Madeleine | 75001 Paris | Tél +33(0)1 56 89 07 70
ACTUARIS LYON | 46 bis chemin du Vieux Moulin | 69160 Tassin | Tél +33 (0)4 72 18 58 58
Siège social
SAS au capital de 100 000 euros – NAF 7022 Z – RCS LYON 413 611 344
ACTUARIS est une société membre d'ADDACTIS Group
www.actuaris.com – contact@actuaris.com