

NOUVEAUX ENVIRONNEMENTS

DATA SCIENCE ET RÉGULATION FINANCIÈRE : UNE FERRARI CONTRE UN VÉLO ?



Bertrand Hassani

Global Head
of Research
and Innovation

Banco Santander

Membre
Labex Réfi

Le Labex Réfi, composé de l'Université Paris 1, de l'ENA et de l'ESCP, a créé un nouvel axe de recherche sur le rapport entre data science et régulation. Comme le montrent ces premiers axes d'analyse, l'objectif des travaux de recherche du Labex est de permettre de profiter des avantages offerts par ces nouveaux environnements, tout en essayant d'en comprendre et d'en limiter les risques.

Comme l'écrivit François Rabelais dans *Pantagruel*, « science sans conscience n'est que ruine de l'âme »... Cette citation résume les questions entourant l'exploitation du Big Data et de sa rapide transition vers l'Intelligence artificielle (A.I.) au sein des institutions financières. L'environnement Big Data nous entraîne vers l'utilisation de nouveaux flux de données, vers l'automatisation des systèmes et naturellement vers la création de systèmes intelligents, et cette évolution se traduit (ou devrait se traduire) par l'arrivée sur le marché de nouveaux produits et de nouveaux services, notamment la mise en place de robo

advisors, de Smart contracts^[1], d'offres personnalisées, de modèles statistiques et mathématiques se basant sur des stratégies d'apprentissage (*machine learning*) plus avancées et utilisant une combinaison de données structurées et non structurées, ainsi que la mise en place de stratégie de *transfer learning*, c'est-à-dire le glissement vers des méthodologies intelligentes pour améliorer l'expérience client d'une part, la gestion des risques d'autre part et bien entendu améliorer au final les résultats de la banque (augmenter les profits). Mais cette évolution (on ne parle pas ici de révolution !) s'opère-t-elle trop rapidement ? Doit-on en avoir peur ? La législation et la régulation sont-elles à la hauteur ?

UN CADRE RÉGLEMENTAIRE ENCORE PEU DÉVELOPPÉ...

Cette dernière question est probablement la plus importante : en réalité, elle conditionne les réponses aux deux précédentes. En effet, si la législation et la régulation sont à la hauteur des enjeux, il n'y a plus de raison d'être effrayé, et quant à savoir si cela va trop vite, cette question n'a plus lieu d'être non plus, car la vitesse acceptable est fonction de l'évolution du cadre légal et réglementaire. Qu'en est-il de ce

[1] Les smart contracts s'exécutent automatiquement sur une blockchain, en prenant en compte l'ensemble des conditions et des limitations programmés dans le contrat à l'origine.

cadre ? À l'heure actuelle, peu de pays ou de juridictions possèdent un cadre réglementaire développé et la législation est encore embryonnaire par rapport à l'enjeu. Quel est cet enjeu ? Selon le Professeur Stephen Hawking, l'AI peut s'affranchir de notre contrôle et se redéfinir à un rythme toujours croissant. En résumé, dans le cas de l'AI, la matérialisation du risque implicite du modèle pourrait aboutir à notre extinction. Bien que la plupart des institutions travaillent activement sur ce sujet, notamment sur des textes de lois traitant de la protection des données, de leur sécurité et de leur traitement, en revanche, en ce qui concerne les modèles, le cadre actuel ne semble pas évoluer suffisamment rapidement. Celui-ci est pourtant très jeune (le premier document réglementaire date de 2011), mais déjà obsolète, car il ne permet pas de capter l'évolution dynamique et potentiellement en temps réel des modèles et de ce qu'ils sous-tendent. En effet, il est important de garder à l'esprit qu'un modèle n'est rien d'autre qu'une représentation de la réalité à un moment donné, et que si cette réalité change (la réalité est ici représentée par les données), les modèles changent aussi. La modélisation Big Data n'est pas une réelle nouveauté : en réalité, les premiers modèles ont été développés dans les années 1950 (pour les plus récents, et même 1810 pour les méthodes de régression ou

bayésiennes) et sont utilisés depuis de nombreuses années notamment en marketing pour la segmentation clients. Par ailleurs, les modèles de régression, de réseaux et autres sont utilisés depuis des années au sein des départements responsables de la gestion des risques. Ce qui a réellement changé, c'est la capacité de traitement des bases de données de grande taille et l'amélioration des temps de calcul.

...ET QUI DOIT POUVOIR ÉVOLUER PLUS RAPIDEMENT

En résumé, le problème du cadre réglementaire et législatif actuel se situe à trois niveaux :

- au niveau des données utilisables ;
- au niveau des infrastructures nécessaires pour garantir la sécurité de l'ensemble ;
- enfin, au niveau des modèles utilisés.

Ce cadre évolue en particulier en ce qui concerne les données et les infrastructures. L'arrivée de nouvelles directives qui mettent les consommateurs aux commandes telles que le Règlement général sur la protection des données[2], la DSP 2[3] ou le droit à l'oubli est déjà une avancée importante, mais cela est-il suffisant ? Par ailleurs, tout n'est pas encore clair : par exemple, peut-on faire de la collecte de données de clients potentiels (et cela, même si ceux-ci nous en donnent l'autorisation) sur Internet et en particulier sur les réseaux sociaux (*web-scraping*) ? À qui les données non structurées que nous trouvons sur la toile appartiennent-elles ? Les normes telles que l'ISO 27001[4] ou la réglementation publiée par la Réserve fédérale américaine sur la sécurité de l'information[5] sont aussi une

avancée en termes de sécurisation des infrastructures nécessaire à la collecte, la conservation et le traitement de ces données. En revanche, les pratiques et les schémas de gouvernance liés au cadre actuel de la gestion du risque de modèle sont très lourds et non adaptés à la gestion de modèles changeant et s'ajustant en temps réel. Illustrer ce fait est relativement simple ; lorsque nous créons un nouveau modèle, quel qu'il soit, le schéma de gouvernance suivant est relativement classique : une revue de la part d'un collègue de votre propre département, une validation par le département en charge, un audit interne, un audit externe, une revue de la banque centrale... c'est-à-dire un processus qui peut prendre plusieurs mois... Alors que le changement de modèle dû à une modification de l'information sous-jacente demande une réaction quasi instantanée.

AMÉLIORER LES COMPÉTENCES ET LA COMPRÉHENSION DES PROFESSIONNELS

Il est d'ores et déjà nécessaire de modifier les schémas de gouvernance des modèles de façon à les rendre plus dynamiques et flexibles pour qu'ils puissent s'adapter aux nouvelles stratégies de modélisation qui, pour un même objectif, peuvent vous faire utiliser un réseau de neurones[6] un premier jour, une régression logistique[7] le lendemain, une « *random forest* »[8] le suivant, etc. La mise en place de stratégies de modélisation provenant du domaine de la science des données fait partie du progrès et essayer d'arrêter le progrès serait comme essayer d'arrêter un tsunami en mettant les mains devant soi lorsque la vague

se situe à quelques mètres de vous. L'interdiction ou l'« hyper-réglementation » de ces approches, n'est en aucun cas la solution ; en effet, il y aura toujours des entreprises innovantes non assujetties à ces règles qui trouveront un moyen de venir jouer sur ce terrain. Plus simplement, l'apport de ces stratégies est non négligeable en termes d'amélioration des processus et de catalyse de l'activité. La meilleure façon de survivre serait plutôt d'améliorer les compétences de tout un chacun pour pouvoir surfer sur cette vague. Par conséquent, en plus de travailler éthiquement sur le sujet, il me semble nécessaire d'améliorer la compréhension, les connaissances et les compétences quantitatives de la part des professionnels pour être en mesure de tirer le meilleur parti et d'éviter de souffrir des problèmes nécessairement liés à l'utilisation du *Big Data* et de l'intelligence artificielle les yeux bandés (*lifelong learning*). C'est pourquoi les entreprises ont besoin de s'appuyer autant que possible sur des ressources internes appropriées ayant un niveau de compétences quantitatives conséquent, car celles-ci seront les plus à même d'éclairer les autres.

ÉVITER L'APPARITION DE RISQUES NOUVEAUX

En conclusion, en plus de l'adoption d'une approche éthique, seul un cadre réglementaire intelligent et une supervision dynamique permettront de tirer profit des nouvelles technologies et stratégies de modélisation et ainsi d'éviter que l'utilisation de tous ces nouveaux flux de données dans les modèles de segmentation clientèle, de risques ou autres, n'engendre de nouveaux soucis liés aux risques de vente inappropriée (*mis-selling*), de défaut de conseil (dans le cadre de l'utilisation d'un *robo advisor* par exemple), de cyber-sécurité ou d'erreurs... sans pour autant tomber dans l'excès dogmatique inverse et penser que ce nouvel environnement est la panacée. ■

[2] Le règlement européen sur la protection des données personnelles a été publié au JO le 27 avril 2016 et sera applicable le 25 mai 2018.

[3] La directive des services de paiement révisée a été adoptée par le Parlement européen le 8 octobre 2015.

[4] La norme ISO/CEI 27001 concerne la gestion de la sécurité de l'information.

[5] https://www.federalreserve.gov/supervisionreg/topics/info_security.htm ; https://www.federalreserve.gov/supervisionreg/topics/it_exam_guidance.htm ; https://www.federalreserve.gov/publications/supervision_bhc.htm.

[6] Un réseau de neurones est un ensemble d'algorithmes dont la conception est à l'origine très schématiquement inspirée du fonctionnement des neurones biologiques.

[7] La régression logistique vise à construire un modèle permettant de prédire ou expliquer les valeurs prises par une variable cible qualitative.

[8] Les forêts d'arbres de décision font partie des techniques d'apprentissage automatique.