Discrimination et IA : comment limiter les risques en matière de crédit bancaire

Les discriminations humaines relayées par les intelligences artificielles sont largement documentées. Mais les solutions pour éviter ces biais restent, elles, encore à perfectionner.

Temps de lecture : minute

4 octobre 2021

L'intelligence artificielle (IA), cet ensemble de technologies visant à reproduire les capacités cognitives et affectives humaines, a envahi notre quotidien. Si certaines applications médiatisées de l'IA inquiètent, telles que la reconnaissance faciale ou les drones de combat autonomes, une application moins sensationnelle prend néanmoins une place croissante dans nos vies : l'IA comme aide à la décision.

Couplé au *big data* – ce concept qui fait référence à de grandes masses de données – les algorithmes de *machine learning* apprennent à prédire des phénomènes sur le fondement des relations (corrélations mathématiques), reliant le phénomène considéré à une grande masse de données appelée " jeu d'apprentissage ".

Ces modèles contribuent par exemple à prédire la météorologie du lendemain, sur la base de nombreuses données (température, pression, densité de l'air…). Mais ce type d'aide à la décision permet aussi de prédire si tel candidat à une formation est susceptible d'obtenir son

diplôme, ou bien si tel autre candidat à l'embauche sera performant à l'avenir dans son nouveau poste, ou encore si tel emprunteur, sollicitant sa banque dans le cadre d'un crédit immobilier sera, au final, en mesure de le rembourser. On comprend dès lors l'importance sociale de ces décisions guidées par les données.

Le risque de discrimination

L'un des risques de dérive éthique majeure de ces modèles d'aide à la décision concerne la discrimination d'une personne physique sur la base d'un attribut protégé par la loi, c'est-à-dire une caractéristique de la personne qui ne peut en aucun cas être utilisée par respect des valeurs de justice et d'égalité. Par exemple, pour la France, le genre ne peut être un critère recevable pour toute prise de décision regardant la personne.

Aussi, le risque de discrimination algorithmique existe dès lors que les données du jeu d'apprentissage présentent des corrélations entre le phénomène que l'on cherche à prédire et l'attribut protégé, tel le genre. Pour expliquer ce risque discriminatoire, plaçons-nous dans le cadre de la distribution de crédit bancaire. Supposons, à des fins d'illustration, que les emprunteurs femmes, clientes historiques d'une banque, aient en moyenne et par le passé, moins remboursé leur crédit que les hommes : alors le lien entre le genre et la prédiction du défaut (ne pas avoir remboursé l'intégralité de son crédit) sera " appris " par le modèle de *machine learning*, qui s'en servira pour ses prédictions futures.

En clair, les emprunteuses seront automatiquement moins bien notées (note appelée *credit score*) que leurs homologues masculins, toutes choses égales par ailleurs, ce qui constitue une discrimination d'après le genre, prohibée par la loi.

En première analyse, la solution à ce problème semble triviale, il suffit de supprimer tous les attributs protégés par la loi du jeu de données

d’apprentissage. Pourtant, par le truchement des corrélations entre attributs protégés et non protégés, le problème demeure sous une forme indirecte, plus subtile à identifier et nécessitant des analyses statistiques.

Illustrons ce cas : en droit français, il est interdit de discriminer un individu d’après son âge. En revanche, rien ne semble interdire, a priori, de fonder la prédiction de défaut d’un client sur son ancienneté dans la banque. Or, cette variable est potentiellement liée (corrélée) à l’âge de l’individu, ce qui entraîne de facto une discrimination indirecte, bannie également en droit français.

Une proposition de règlement européen pour limiter les risques

Consciente de l’importance prise par l’IA dans la vie des citoyens européens et des risques associés, la Commission européenne a proposé un premier cadre légal pour l’IA le 21 avril 2021 dernier.

L’approche retenue est fondée sur les risques, avec une gradation des exigences en quatre niveaux selon l’activité considérée. Le *credit scoring* bancaire est classé dans la catégorie des risques élevés, ce qui implique que l’IA satisfasse aux exigences définies dans le titre III, chapitre 2 du règlement, préalablement à toute mise sur le marché, afin de réduire au minimum les risques, jusqu’à un niveau résiduel jugé acceptable.

Or, si le jugement de l’acceptabilité d’un niveau de risque constitue déjà une contrainte floue, la commission accorde en outre aux fournisseurs de systèmes d’IA une flexibilité en matière de solution technique de mise en conformité.

L’équité algorithmique

Ainsi, pour minimiser autant que possible le risque discriminatoire, il faut disposer d’un indicateur de mesure approprié. Or la discrimination est associée au concept protéiforme d’équité algorithmique, développé dans le champ du *fair machine learning*.

On distingue trois formes d’équité algorithmique : individuelle, de groupe et contrefactuelle. La première correspond à la discrimination telle que définie dans les textes légaux (chaque individu est évalué indépendamment des attributs protégés) ; la seconde se situe au niveau du groupe et exige une classification identique pour les individus appartenant à même groupe (par exemple le groupe des femmes) ; la troisième forme impose que les résultats de classification soient insensibles à la modification des valeurs des attributs protégés.

Pour un motif de complexité, l’équité de groupe est privilégiée par les chercheurs comme moyen opérationnel de mesure de l’intensité discriminatoire d’un modèle prédictif. Mais là encore, plusieurs indicateurs entrent en concurrence.

Quelle mesure pertinente du risque de discrimination ?

Considérons les deux principaux indicateurs, pertinents dans le cadre du " credit scoring " :

- L’indicateur d’indépendance impose une prédiction identique pour les groupes définis par l’attribut protégé, ce qui revient à dire, avec l’exemple du genre, que les proportions de femmes et d’hommes obtenant un crédit devraient être strictement égales.
- Au contraire, l’indicateur de séparation autorise des proportions différentes de crédits alloués entre hommes et femmes, mais exige

des proportions d’erreurs de prédiction identiques pour les femmes et les hommes.

Ce dernier indicateur semble davantage adapté au cas qui nous intéresse, car le fait d’imposer une stricte égalité dans les proportions de crédits alloués entre hommes et femmes (indicateur d’indépendance parfait) n’est ni une conséquence nécessaire ni une conséquence souhaitable de l’équité individuelle.

Cette idée contre-intuitive s’explique ainsi : si une corrélation empirique réelle existe entre le genre et le défaut, alors ne pas en tenir compte conduirait à allouer des crédits en excès à des individus qui ne pourraient honorer leur dette, les faisant tomber dans la spirale du surendettement, ou bien à ne pas prêter à des personnes pourtant solvables, les menant à une situation d’exclusion bancaire, ces deux résultats étant coûteux socialement.

On comprend dès lors que, dans le cadre du crédit bancaire, le choix d’un indicateur de mesure de la discrimination n’est ni évident, ni neutre.

Un cadre légal qui nécessite davantage de précisions

Aussi, si le cadre légal proposé par la Commission européenne représente une avancée précisant les lignes directrices du futur cadre réglementaire, des conflits d’interprétation inévitables demeurent.

Et certains points, tels les indicateurs de mesure de la discrimination admissibles ainsi que les seuils qui leur sont associés, mériteraient davantage de précisions.

Notons qu’une fois ce cadre réglementaire finalisé et le calendrier d’entrée en vigueur fixé, le secteur bancaire devra se mettre en

conformité urgemment car l'APCR notait en juin 2020 que très peu d'institutions financières s'étaient engagées jusque-là dans l'identification et la remédiation des biais de leurs modèles d'IA.

Si l'enjeu de la discrimination par les IA est peu médiatisé, il est pourtant crucial tant les décisions de ces modèles affectent les citoyens à des moments clés de leur vie, déterminants pour leur intégration à la société comme pour l'amélioration de leur niveau de vie : accès à un établissement d'enseignement, accès à l'embauche, distribution de crédit…

Article rédigé par Christian Goglin,
Professeur en Finance et Intelligence Artificielle, ICD Business School

---

Article écrit par Christian Goglin