Actualité de la Fondation du risque

Assurance : discrimination, biais et équité

Arthur Charpentier

Professeur, Université du Québec à Montréal

Ce texte est le résumé de l'article publié dans Opinions et débats n° 25 de juillet 2022 par l'Institut Louis Bachelier, téléchargeable à l'adresse https://www.institutlouisbachelier.org/assurance-discrimination-biais-et-equite/

Les données massives et les performances obtenues par les algorithmes d'apprentissage automatique ont chamboulé l'assurance et l'actuariat. Les questions soulevées par ces nouveaux outils dans d'autres contextes – que ce soit la justice prédictive (ou justice « actuarielle » comme l'appelle Harcourt [2008]) ou les débats sur les « *fake news* », en passant par les véhicules autonomes et la médecine prédictive – poussent les actuaires au doute, et à la méfiance. Kranzberg [1986] affirmait que « *technology is neither good nor bad; nor is it neutral* », mettant en avant que, même sans mauvaises intentions, les algorithmes d'apprentissage pouvaient être injustes. Et corriger ces possibles injustices n'est pas simple. Pour Nielsen [2020] « *technology does not necessarily self-regulate, via either market or social pressures* » (la main invisible des marchés ou de la pression sociale ne suffira peut-être pas). C'est dans ce contexte que nous allons revenir ici sur les problématiques de biais, de discrimination et d'équité, des modèles prédictifs utilisés en assurance. Ces changements, tant sur les données que sur les modèles, que l'on observe depuis une petite dizaine d'années, avaient déjà questionné l'existence même de l'assurance. Pour Löffler *et al.* [2016] « *this leads to demutualization and a focus on predicting and managing individual risks rather than communities* », l'individualisation de plus en plus grande des primes force à s'interroger sur l'avenir de la mutualisation et de la solidarité entre assurés. Les problèmes de discrimination sont alors à envisager dans ce contexte de perte de solidarité. Car paradoxalement, la discrimination n'a de sens qu'en

voyant l'individu en tant que membre d'un groupe caractérisé par un trait partagé (les femmes, les personnes d'origine étrangère, les personnes âgées, etc.).

Ce principe de mutualisation des risques se traduit par le fait que l'assurance est « *the contribution of the many to the misfortune of the few* ». Du fait de l'inversion du cycle de production, l'assureur vend au souscripteur une promesse d'indemnisation, dans le futur, d'un risque aléatoire, en échange du paiement d'une contribution « juste » ou « équitable », a priori proportionnelle au risque de l'assuré (Thiery et Van Schoubrœck [2006] parleront d'équité actuarielle). Le vrai facteur de risque sous-jacent étant une information non observable lors de la signature du contrat, l'assureur va construire des algorithmes prédictifs, à partir d'informations disponibles, pour prédire la fréquence de sinistre, le coût des sinistres, mais aussi la probabilité de frauder, ou la probabilité de souscrire une garantie supplémentaire par exemple. En ne voyant plus un groupe d'assurés comme une mutualité parfaitement homogène, les actuaires ont utilisé des algorithmes de plus en plus fins pour créer des sous-groupes davantage homogènes. Avec le développement des techniques d'apprentissage machine, l'idée de personnalisation, d'individualisation (très présente dans la communauté informatique depuis plusieurs années, comme le soulignaient Adomavicius et Tuzhilin [2005] avec des « profils » individualisés) fait son chemin, et pousse les assureurs à démutualiser de plus en plus. « *At the core of insurance business lies discrimination between risky and non-risky insureds* » avait affirmé Avraham [2017]. Aussi, d'un côté, l'opération d'assurance relève de la technique et a fondamentalement une dimension collective, reposant sur la mutualisation des risques au sein de groupes de risques homogènes. Les systèmes de classification des assurances reposent sur l'hypothèse que les individus répondent aux caractéristiques moyennes (stéréotypées d'une certaine manière) d'un groupe auquel ils appartiennent. C'est la discrimination au sens statistique (mise en œuvre par des outils statistiques puis économétriques). De plus, le contrat d'assurance relève du droit, et a une dimension individuelle. En ce sens, un individu ne peut être traité différemment en raison de son appartenance à tel ou tel groupe, en particulier à un groupe auquel il n'a pas choisi d'appartenir, sinon c'est de la discrimination, au sens légal du terme. Et dans le contexte de données de plus en plus massives, et d'algorithmes prédictifs de plus en plus complexes (pour ne pas utiliser le terme de « boîte noire »), il est devenu de plus en plus difficile de garantir que les assureurs demandent une contribution « juste » aux souscripteurs de polices d'assurance.

Réfléchir aux égalités de traitement des assurés revient à s'interroger sur la possibilité même de souscrire un contrat, en vue d'une couverture, mais aussi à l'idée de demander une prime non prohibitive, et non dissuasive. Car contrairement à ce que nous apprennent les mathématiques financières (et l'hypothèse de marchés complets, Froot *et al.* [1995]), il n'existe pas en assurance de loi du prix unique, le prix d'un risque étant vu au travers d'une mutualité d'assurés et d'un modèle de tarification. De plus, les souscripteurs n'achètent pas « une assurance », mais une garantie de couverture contre certains risques. Si certaines garanties sont souscrites majoritairement par certaines populations, et pas par d'autres, la différence de prix ne correspond pas forcément à une discrimination, *stricto sensu*. C'est dans ce contexte que nous allons discuter des biais, des discriminations et de l'équité en assurance.

Les données, de plus en plus massives, posent de nombreux défis. Tout d'abord, la réglementation cherche à protéger des informations dites « sensibles » ou « protégées », interdisant parfois de collecter et de stocker certaines variables. Le principal danger est qu'il devient alors difficile d'assurer qu'un modèle ne discrimine pas suivant un critère, si ce critère n'est pas observé. Poser un voile d'ignorance sur certaines caractéristiques ne suffit pas pour imposer l'équité d'un modèle, et ne sert qu'à masquer un potentiel problème (ou comme l'affirmaient Kearns et Roth [2019], « *machine learning won't give you anything like gender neutrality "for free" that you didn't explicitely ask for* »). Un autre défi est celui des innombrables biais des données collectées à travers toutes sortes de sources (questionnaires, objets connectés, données

obtenues via différentes sources, etc.). Parmi ces derniers, on peut mentionner les biais de variable manquante, les biais de définition ou d'interprétation, les biais de mesure, les biais de survie, les biais de rétroaction, etc. Ces « *dark data* » (pour reprendre le terme utilisé par Hand [2020]) forcent à s'interroger sur la pertinence d'une classification des risques, certaines discriminations étant parfois perçues sur la base d'informations biaisées ou mal interprétées. Si le genre du conducteur principal a longtemps été utilisé par les assureurs, on peut s'interroger sur sa signification dans un couple (hétérosexuel) partageant une voiture.

On retrouve ici la difficulté de la définition des variables, bien connue par les statisticiens. On reviendra ainsi sur le paradoxe de Simpson et l'inférence écologique (en anglais, on parle d'« *ecological fallacy* ») où l'absence de certaines variables peut donner une interprétation fausse, fallacieuse, sur le sens d'une potentielle discrimination. Et dans le contexte d'assurance, les données télématiques, et les mécanismes incitatifs de type « *gamification* » posent des questions sur les biais de rétroaction, les assureurs ayant la possibilité d'influencer directement les comportements de tel ou tel assuré, sur la base de données arrivant en temps réel. On retrouve ici une forme de biais de sélection, ce dernier signifiant simplement que les données historiques ont été collectées sur des personnes qui ont choisi de souscrire un contrat et qui ont été acceptées par un assureur au préalable (potentiellement sur la base d'un précédent modèle). Tout comme l'analyse de la fraude ne peut pas se faire de la même manière, si les enquêtes en lien avec la fraude sont menées de manière aléatoire ou si elles reposent sur un modèle préalable de détection de fraude. On retrouve les débats classiques entre les données d'expériences (souvent randomisées, pour reprendre le terme anglais) et les données administratives ou observationnelles.

On l'a déjà mentionné, une notion centrale sera celle de discrimination, terme particulièrement ambigu, puisque les actuaires utiliseront la version statistique du terme (on peut penser à l'analyse discriminante introduite par Ronald Fisher), alors que les juristes y voient un traitement inégal et défavorable appliqué à certaines personnes en raison de certains critères. Même s'il existe des différences culturelles, entre les pays, on retrouvera souvent un certain nombre de caractéristiques protégées (par la morale, ou par la loi) comme le genre ou le sexe de la personne, la race ou l'origine nationale ou ethnique, le handicap et toute information génétique, etc. Ces critères sont parfois présentés comme des clubs dans lesquels on tombe à la naissance, pour reprendre l'expression de Macnicol [2006] (qui font aussi écho au concept de « voile d'ignorance » et de « loterie génétique »). D'autres critères comme l'âge sont plus complexes car un assuré traversera tous les âges au cours de sa vie : s'il y a une « discrimination » contre les jeunes, l'assuré en souffrira à vingt ans alors qu'il est dans le groupe défavorisé, avant de passer progressivement dans le groupe privilégié (sans évoquer une possible solidarité intergénérationnelle). Enfin, des critères relèvent davantage de choix, plus ou moins conscients. Une première difficulté est que de nombreuses discriminations ne sont pas intentionnelles. Pire encore, contrairement à ce qui peut exister dans la littérature traditionnelle sur les discriminations (où des proxys sont potentiellement utilisés à la place d'une variable sensible, comme le *redlining* où les quartiers d'une ville sont un proxy d'une information éthique et raciale), en assurance, certaines variables sensibles (comme le genre) ont longtemps été utilisées comme proxy d'informations difficilement accessibles (comme des informations comportementales en matière de conduite automobile).

Une autre difficulté repose sur un problème classique en grande dimension, et sur la multicolinéarité des variables prédictives. Ceci peut donner lieu à une discrimination par proxy (parfois appelée discrimination statistique ou discrimination indirecte dans les directives européennes en lien avec la discrimination), qui consiste à utiliser une variable très corrélée à la variable protégée. L'utilisation intensive de proxys (non détectés) dans le développement de modèles a soulevé des inquiétudes quant à l'équité. Et l'enrichissement de données rajoute de plus en plus de variables pouvant être vues comme engendrant une discrimination indirecte.

La dernière notion que nous décrirons est la notion d'équité d'un modèle prédictif. Après un rapide survol des concepts de justice, nous présenterons les mesures classiques d'équité qu'il est possible d'utiliser pour quantifier l'ampleur d'une possible discrimination. Si on formalise rapidement, on dispose d'un triplet (y, x, p), où y est une variable d'intérêt (nombre de sinistres, coût annuel, nombre de visites chez le médecin, etc.), x un ensemble de variables explicatives admissibles, utilisées pour prédire y, et p une variable sensible, ou protégée (supposée unique ici). Construire un modèle prédictif $\hat{y} = m(x)$ en utilisant seulement les variables x et pas p ne suffit pas à garantir que le modèle ne puisse pas discriminer suivant p, tout simplement car p peut être très corrélée à certaines caractéristiques x (on retrouve l'idée de proxy). Barocas *et al.* [2019] notent que les grands principes associés à l'équité se traduisent : 1. par une indépendance entre $\hat{y}$ et p, autrement dit la prédiction n'a rien à voir avec le groupe de p ; 2. par une notion de séparation : $\hat{y}$ est indépendante de p étant donné y, et 3. une notion de suffisance : y est indépendante de p étant donné $\hat{y}$. Ces principes vont se traduire par différentes notions d'équité de groupe, les plus populaires étant la parité démographique et la notion d'égalité des chances. Ces notions (dites de groupe), très populaires et largement utilisées (par exemple sur le marché du travail, aux Etats-Unis) sont à distinguer des approches individuelles qui émergent dans la littérature scientifique, inspirées des techniques d'inférence causale et visant à chercher à un contrefactuel afin de répondre à la question « que se serait-il passé si l'assuré avait la caractéristique $p = 1$ au lieu de $p = 0$? » (si on suppose que la variable protégée est binaire, $p \in \{0, 1\}$) . C'est une relation causale entre la variable sensible p et la variable de risque y, qui peut légitimer une discrimination statique, comme le suggérait la Commission européenne, qui proposait d'autoriser « des différences proportionnelles dans les primes et les prestations des particuliers lorsque l'utilisation du sexe est un facteur déterminant dans l'évaluation du risque, sur la base de données actuarielles et statistiques pertinentes et précises ». Néanmoins, la présence de proxys pose de nombreux défis, car l'approche contrefactuelle usuelle (consistant à changer la variable protégée p seulement, *ceteris paribus*) n'a pas de sens en grande dimension, en présence de proxys fortement corrélés à la variable sensible : une intervention (conceptuelle et fictive) sur la variable sensible p doit avoir un impact sur une ou plusieurs variables prédictives x, et donc sur la prévision.

D'autres concepts seront aussi évoqués ici, sans pour autant faire l'objet de chapitres spécifiques, comme la responsabilité. En effet, si un algorithme reproduit ce qu'il observe dans les données, peut-il être jugé responsable de reproduire les biais sociaux ? Sous un angle épistémologique, on demandait historiquement aux modèles de bien « décrire le réel » (ou disons le réel tel qu'il apparaît dans les données, on parlera d'*accuracy* en apprentissage statistique), c'est-à-dire « ce qui est », alors qu'en introduisant une dimension morale et éthique, on demande que le modèle soit en accord avec ce qui « devrait être », suivant une norme éthique (la fameuse opposition « *is ought* » de Hume [1739]), ou entre la « normalité » statistique opposée à la norme morale. L'autre souci est que pour quantifier l'équité, il convient d'avoir accès à ces données personnelles, privées et sensibles, ce qui renvoie aux discussions sur la vie privée (ou *privacy*) et la conformité (ou *compliance*). Finalement, on le verra tout au long du rapport, ces discussions autour de la discrimination, des biais et de l'équité sont très proches de celles portant sur l'interprétation des modèles prédictifs et de la notion d'explicabilité.

Cet aspect narratif de la construction de modèle est important, en particulier lorsque l'on cherche à créer des graphs causaux dirigés afin de comprendre les liens entre la variable protégée p, les possibles variables prédictives x et la variable d'intérêt y. Mais en grande dimension, cet exercice devient vite impossible. En affirmant que « *all models are wrong but some models are useful* », Georges Box insistait sur l'aspect narratif de la modélisation, sur l'interprétation qui en découle. Une compréhension fine des données et des modèles est aujourd'hui indispensable, l'époque des calculs froids et objectifs (ou supposés objectifs) des actuaires semblant révolue.

Bibliographie

ADOMAVICIUS G. ; TUZHILIN A., "Personalization Technologies: a Process-Oriented Perspective", *Communications of the ACM*, vol. 48, n° 10, 2005, pp. 83-90.

AVRAHAM R., "Discrimination and Insurance", *in* Kasper Lippert-Rasmussen (dir), *Handbook of the Ethics of Discrimination*, Routledge, 2017, pp. 335-347.

BAROCAS S. ; HARDT M. ; NARAYANAN A., *Fairness and Machine Learning*, fairmlbook.org, 2019.

FROOT K. A. ; KIM M. ; ROGOFF K. S., "The Law of One Price over 700 Years", National Bureau of Economic Research (NBER), n° 5132, 1995.

HAND D. J., *Dark Data: Why What You don't Know Matters*, Princeton University Press, 2020.

HARCOURT B. E., *Against Prediction*, The University of Chicago Press, 2008.

HUME D., *A Treatise of Human Nature*, Cambridge University Press Archive, 1739.

KEARNS M. ; ROTH A., *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford University Press, 2019.

KRANZBERG M., "Technology and History: 'Kranzberg's Laws' ", *in Technology and Culture*, vol. 27, n° 3, 1986, pp. 544-560.

LÖFFLER M. ; MÜNSTERMANN B. ; SCHUMACHER T. ; MOKWA C. ; BEHM S., "Insurers Need to Plug into the Internet of Things - or Risk Falling Behind", *in European Insurance*, 2016.

MACNICOL J., *Age Discrimination: An Historical and Contemporary Analysis*, Cambridge University Press, 2006.

NIELSEN A., *Practical Fairness*, 0'Reilly Media, 2020.

THIERY Y. ; VAN SCHOUBROECK C., "Fairness and Equality in Insurance Classification", *The Geneva Papers on Risk and Insurance - Issues and Practice*, vol. 31, n° 2, 2006, pp. 190-211.