Les modèles prédictifs peuvent-il être loyaux et justes ?

Arthur Charpentier

Professeur, Université de Rennes 1

Dans Nosedive *(traduit par le titre* Chute libre *en France), le premier épisode de la saison 3 de la série télévisée* Black Mirror*, on découvre la dystopie d'une société régie par une « cote personnelle », une note, un score allant de 0 à 5. Dans ce monde, chaque personne note les autres, les mieux notés ayant accès à de meilleurs services (priorité dans les services, meilleurs taux, meilleurs prix, etc.). Cette tendance à construire des scores dans toutes sortes de domaines (historiquement sur les crédits mais aujourd'hui sur des aspects criminels, voire civiques dans certains pays) ne va-t-elle pas déboucher sur un monde qui serait un concours de popularité sans fin ? Et comment serait-elle conciliable avec une justice sociale, a priori souhaitable ?*

Les scores de crédit et les réseaux sociaux

Un score de crédit est, d'un point de vue actuariel, une grandeur proportionnelle à la probabilité de ne pas honorer ses engagements en tant que créditeur. Cela peut être de ne plus pouvoir payer les échéances pendant trois mois consécutifs, ou juste d'avoir un retard. Dans la vraie vie, comme toujours, c'est un peu plus compliqué. Aux États-Unis ou en Grande-Bretagne, il n'est pas rare que les étudiants s'endettent sur des dizaines d'années pour avoir l'opportunité de suivre les cours qui les intéressent (même si la motivation est surtout d'obtenir un diplôme en fin de parcours). Mais surtout, dès qu'ils atteignent l'âge de dix-huit ans, des sociétés de notation de crédit vont surveiller tous leurs déplacements. Souvent à leur insu. Et si un jour, un crédit consommation ou hypothécaire est refusé, les raisons ne sont jamais motivées. Est-ce dû à un retard dans un paiement de loyer ? à des amendes de bibliothèque oubliées ? à une facture d'eau impayée, vieille de plusieurs années ?

Les sociétés de notation de crédit aux États-Unis, mais aussi en Chine, commencent à étudier la possibilité d'utiliser des données provenant de médias sociaux pour améliorer le score de crédit. Compter le nombre de fois où un utilisateur utilise le mot « gaspillé » (« *wasted* » en anglais) dans ce qu'il poste en ligne ne peut-il pas révéler une information quant au remboursement de dettes ? C'est en tout cas ce que prétend l'analyste de crédit américain Fico : « *If you look at how many times a person says "wasted" in their profile, it has some value in predicting whether they're going to repay their debt (...) It's not much, but it's more than zero* » [McLannahan, 2015]. En Chine, le prêteur *peer-to-peer* Jubao a révélé qu'il était plus susceptible de donner des « bonus » aux emprunteurs s'ils étaient des amis Facebook avec des célébrités, tel que le raconte Botsman [2017].

Pour l'instant, les sociétés de notation de crédit utilisent encore les données qu'elles connaissent bien (factures de services publics et cartes de crédit), mais elles imaginent que bien des informations intéressantes doivent être accessibles (d'une manière ou d'une autre) sur les réseaux sociaux. Mais les données sont encore rares, et difficiles à analyser. Quid de la composante sarcastique ou humoristique (1) dans un tweet utilisant le mot « *wasted* » ? Comme souvent, la difficulté est que les données réellement pertinentes sont difficiles à obtenir. S'il est possible d'avoir des informations sur le paiement du loyer quand un locataire passe par une agence, que faire pour les transactions entre deux particuliers ? Et si c'était possible, comment traiter le cas de colocataires ? Ne pas obtenir de crédit parce qu'un ancien colocataire n'a pas payé dans les temps devient dérangeant. D'autant plus s'il s'agit peut-être d'une facture de téléphone cellulaire réclamée abusivement par la compagnie de téléphonie, alors que l'abonnement avait été résilié.

Mais le gros « malus » dans le score de crédit est bien souvent le fait de ne jamais avoir eu de carte de crédit. On pourrait penser qu'une personne qui n'a pas eu besoin d'une carte de crédit (et se contentait d'une carte de débit, permettant d'acheter chez un commerçant, comme la majorité des cartes bancaires en France) est le propre d'une personne prudente, qui n'a pas besoin de crédit pour des dépenses quotidiennes. Mais pour les établissements de crédit, cette personne n'est pas fiable car on ne la connaît pas. Et c'est à elle de prouver qu'elle l'est (on revient à la pratique récurrente d'inversion de la charge de la preuve évoquée dans Charpentier [2016]). C'est étrangement ce qui se passe aujourd'hui quand on veut entrer sur le sol américain sans avoir de page Facebook.

Dans un monde de surveillance généralisée

Et si les établissements de crédit n'étaient pas les seuls intéressés par notre vie ? Que serait un monde si, en plus de savoir si je paie mes factures à temps, certains cherchaient à connaître mes réseaux d'amis, à savoir quels journaux je lis, si je préfère acheter du lait entier ou du lait demi-écrémé ? Quand on visite le musée de la Stasi à Berlin, on découvre que ce monde a existé, qu'une personne sur soixante-trois était agent (ou indicateur) de la Stasi (en comptant les indicateurs occasionnels, la proportion peut atteindre une personne sur six). Le musée décrit un panoptisme total, chacun étant observé en permanence, comme le décrit Foucault [1975].

Mais ce cauchemar ne correspond-il pas à notre monde actuel, de surveillance permanente, plus ou moins consentie. Surveillance via les téléphones cellulaires (géolocalisation pour la fonction la plus courante, mais parfois aussi des enregistrements audio à l'insu de l'utilisateur par certaines applications), via les objets connectés, mais aussi les caméras de surveillance couplées à des algorithmes de reconnaissance faciale de plus en plus performants. Fin 2017, en Chine, 170 millions de caméra étaient installées, et le cap des 300 millions devrait être atteint d'ici 2020. Lors d'une expérience tentée par la BBC (2), il a fallu sept minutes pour retrouver le journaliste John Sudworth qui se promenait dans les rues.

Le danger est que l'on ne sait jamais trop qui contrôle. De plus en plus de compagnies privées de sécurité se sont associées aux gouvernements. Les fournisseurs de messageries électroniques lisent nos messages pour détecter du spam, mais aussi pour revendre certaines informations. Par exemple, dans les règles de confidentialité annexées aux conditions générales d'utilisation de Gmail (Google) on lit : « Nos systèmes automatisés analysent vos contenus (y compris les e-mails) afin de vous proposer des fonctionnalités personnalisées sur les produits, telles que (...) des publicités sur mesure ». Les assureurs envisagent de plus de plus l'installation de boîtiers GPS dans les voitures, mais en passant par des prestataires externes. Au-delà de la propriété des données (évoquée dans Charpentier et Suire [2016]), on peut s'interroger sur leur revente et leur utilisation. Savoir que quelqu'un se rend régulièrement dans un centre de transfusion sanguine est une information potentiellement intéressante, surtout couplée à d'autres.

Depuis 2014, le gouvernement chinois travaille sur un système d'évaluation de ses propres citoyens programmé pour être mis en place en 2020, comme le raconte Trujillo [2017]. Ce « système de crédit social » vise à créer un « score citoyen » (pour reprendre l'expression de Galeon et Bergan [2017]), afin de prédire, et prévenir, les dangers potentiels, normalisant les conduites individuelles par des dispositifs panoptiques (par exemple la vidéosurveillance), induisant des réflexes d'autodéfense et d'autocontrôle. Comme le disait Foucault [1975], il s'agit de « faire que la surveillance soit permanente dans ses effets, même si elle est discontinue dans son action ; que la perfection du pouvoir tende à rendre inutile l'actualité de son exercice » (même si souvent, aujourd'hui, elle est en plus continue dans son action). Certains de ces scores sont utilisés par la police pour savoir où patrouiller pour faire baisser la criminalité, comme PredPol. Mais quand on y regarde de plus près, les prédictions disent, en substance, que les crimes auront lieu (en majorité) dans les zones (historiquement) les plus criminogènes de la ville. La frontière entre la banalité et la tautologie est étroite. Et le réel danger est que, bien souvent, les scores transforment les probabilités en quasi-certitudes, et le soupçon devient une preuve, comme le notait Supiot [2015].

Justice prédictive et méthodes actuarielles

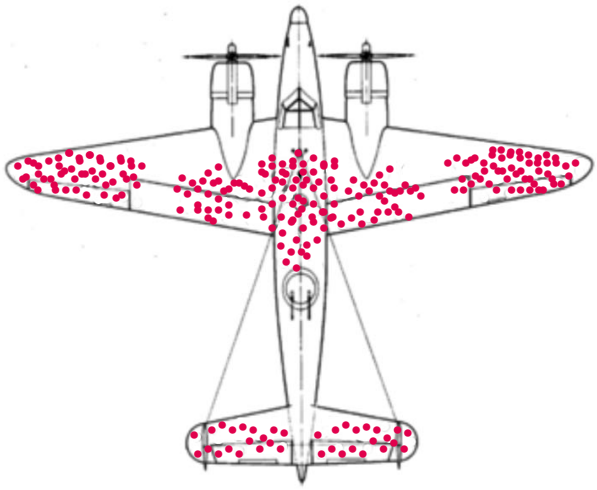
En juin 2010, un rapport de l'Académie de médecine préconisait d'« améliorer la pratique des expertises de dangerosité des criminels sexuels en enseignant et en diffusant les méthodes actuarielles » [Binet, 2010]. Ces méthodes actuarielles sont tout simplement les techniques de *scoring*, de profilage tel que le définit le Règlement européen relatif aux données personnelles du 27 avril 2016 (RGDP). Angèle Christin s'est intéressée aux algorithmes qui estiment la probabilité de récidive dans la justice pénale américaine. Comme elle l'a montré, ces techniques posent nombre de questions, en particulier des biais discriminatoires, l'opacité qui rend difficile les recours, et surtout la difficulté de comprendre ce qui est réellement calculé. Dans l'État de Virginie, un score entre 1 et 10 est renvoyé, convention reprise par Compas (Correctional Offender Management Profiling Alternative Sanctions) qui offre en plus un code couleur qui prédit le risque de récidive violent. Il s'agit alors d'un outil d'aide à la décision, une machine ne pouvant placer une personne en détention seule [Christin *et al.*, 2015].

Les conclusions d'un score prédictif dépendent de deux éléments clés : le modèle utilisé, et les données. Dans la majorité des cas aux États-Unis, les codes des modèles restent opaques (et donc impossibles à attaquer), et rares sont ceux qui ont vu les données utilisées pour calibrer ces modèles. Mais on peut se demander si les décisions de justice ne sont pas elles aussi relativement opaques ? Certes, les juges doivent motiver leurs décisions, ce qui les rend critiquables et attaquables, mais si le processus était si transparent, les issues d'un procès ne devraient-elles pas être alors davantage prévisibles ? Enfin, les différents biais sont assez simples à comprendre. Supposons qu'être riche permet d'avoir un bon avocat, et avoir un bon avocat permet de ne pas avoir certaines condamnations. Dans ce cas, une variable liée à la richesse (le type de véhicule possédé par exemple) sera liée positivement avec le fait de ne pas être coupable (reconnu coupable), et fera baisser le « score de dangerosité ». L'autre danger dans les biais de sélection est qu'ils sont parfois complexes à comprendre, voire paradoxaux. Un exemple classique est celui illustré dans la figure 1 (voir p. 94). Pendant la Seconde Guerre mondiale, il a été demandé à des ingénieurs et des statisticiens comment renforcer les bombardiers qui faisaient face au feu ennemi.

Le statisticien Abraham Wald a commencé à collecter des données sur les impacts sur la carlingue, comme le raconte Mangel et Samaniego [1984]. À la surprise générale, il a recommandé de blinder les endroits des appareils qui présentaient le moins de dommages. En effet, les avions utilisés dans l'échantillon présentaient un biais important : seuls les avions

revenus avaient été pris en compte. S'ils avaient pu revenir avec des trous au bout des ailes, c'est que ces parties étaient suffisamment solides. Et comme aucun avion n'était revenu avec des trous au niveau des moteurs des hélices, c'étaient ces parties qu'il convenait de renforcer.

Figure1 - Endroits endommagés des avions revenus



Source : McGeddon, 2016.

Un autre danger est celui où les relations causales sont inversées. Que penser de ce médecin qui prescrit un puissant neuroleptique à un patient mis en examen, de peur que la justice lui reproche de ne pas avoir vu la dangerosité de son patient, et qu'inversement, la justice s'appuie sur cette prescription pour prouver que le patient est dangereux ? Un algorithme mal conçu pourrait comprendre de travers le sens des relations causales.

Mais les modèles prédictifs en matière judiciaire ne sont pas que du côté des juges. Lors d'un accident corporel sur la route, la loi Badinter (du 5 juillet 1985) prévoit un droit à indemnisation pour toute victime d'un accident de la circulation dans lequel est impliqué un véhicule terrestre à moteur. Lorsque la société d'assurance du conducteur propose une indemnité, la victime fait une rapide analyse coût/bénéfice pour savoir s'il va au tribunal. Si elle ne construit pas formellement un modèle prédictif, elle tente de voir, à partir de quelques éléments à sa connaissance, les coûts de demander à un juge de trancher sur le montant de l'indemnité, et ses bénéfices (potentiels).

Autre point important, les juristes appellent ces modèles « prédictifs » des modèles « actuariels ». Or la première fonction des actuaires est, entre autres, d'actualiser, de calculer la valeur du temps. Et le temps judiciaire a des conséquences souvent désastreuses. En quoi une décision humaine, imparfaite, prise au bout de cinq ans de procédure serait meilleure qu'une décision automatique prise en quinze jours par une machine ? Nombre de personnes qui ont connu des procédures de plusieurs années, aboutissant à un non-lieu, rêvent de procédures accélérées. Car le « temps perdu » a une valeur, les actuaires le savent bien.

Que penser alors de cette efficacité des modèles algorithmiques ? La justice se doit d'être efficace, mais cette contrainte ne doit pas faire oublier l'objectif central qui est celui de rendre la justice. Que se passe-t-il si l'efficacité devient un objectif, pour ne pas dire le principal objectif ? Car c'est bien la question que posent les modèles prédictifs : quel est l'objectif que l'on cherche à maximiser ? Et comment le formule-t-on de manière simple ?

Aide à la décision ou justification d'une prise de décision

Aux États-Unis, nombre de juges se sont vu reprocher de motiver un jugement à l'aide d'outils d'aide à la décision, ce qui laisse planer un doute sur la fonction réelle de ces outils. L'idée était initialement d'apporter une aide. Récemment, plusieurs systèmes mis en place dans les années passées ont été remis en question. Par exemple en Australie, le STMP (Suspect Targeting Management Plan) proposait d'identifier si des préadolescents devaient être surveillés, ou pas. Ce modèle ressemble à s'y méprendre à n'importe quel modèle actuariel, c'est-à-dire un outil d'évaluation et de prédiction des risques, en se concentrant soit sur les récidivistes, soit sur les personnes suspectées de

commettre un futur crime. Or un rapport récent montrait que son utilisation n'avait eu « aucun impact observable sur la prévention du crime » (3). Parallèlement, aux États-Unis, l'outil Compas (Correctional Offender Management Profiling Alternative Sanctions) a été critiqué [Dressel et Farid, 2018] : « Les défenseurs de ces systèmes soutiennent que les données et l'apprentissage automatique avancé rendent ces analyses plus précises et moins biaisées que celles des humains. Cependant, nous montrons que le logiciel d'évaluation des risques Compas, largement utilisé, n'est pas plus précis ou juste que des prédictions faites par des personnes qui ont peu ou pas du tout d'expertise en matière de justice pénale ». En questionnant des personnes recrutées sur Internet, sans compétences en droit, il s'agissait de prévoir si des personnes allaient, ou pas, commettre un autre crime dans les deux ans à venir. Compas s'est trompé dans 34,8 % des cas, et les internautes dans 33 % des cas. Cela dit, on peut se demander ici ce que « se tromper » signifie. En l'occurrence, on ne mesure pas ici la récidive des personnes, mais la condamnation pour récidive des personnes. Et si les modèles (ou les gens) ne s'étaient pas trompés, mais que les juges, en revanche, oui ?

Prévoir, et se tromper

Et si un des soucis ne venait pas de ce que l'on demande à un outil prédictif ? Prévoir, c'est (fondamentalement) établir une probabilité pour un fait futur. Comme cela avait été rappelé dans un débat sur les sondages et les élections, est-ce qu'on peut dire qu'on se trompe si on annonce qu'un événement peut se produire avec 5 % de chance, et qu'il se produit effectivement ? Pour savoir si une technique de prévision est bonne, il faut collecter un ensemble de prévisions, et les comparer aux observations. C'est ce que font les météorologues depuis une quinzaine d'années, et qui a été formalisé par Gneiting *et al.* [2007]. Leur idée est qu'un ensemble de distributions prédictives est obtenu par un modèle $\{\hat{F}_t, \hat{F}_{t+1}, \hat{F}_{t+2}, \dots, \hat{F}_{t+h}\}$ et qu'il convient de comparer ces distributions aux observations $\{y_t, y_{t+1}, y_{t+2}, \dots, y_{t+h}\}$ – et non pas $\{\hat{y}_t, \hat{y}_{t+1}, \hat{y}_{t+2}, \dots, \hat{y}_{t+h}\}$. Il faut alors introduire une distance entre les distributions prédictives, et les observations. Dans un système physique, il est possible d'imaginer comprendre les différentes relations causales, et donc de prévoir. Mais dans les relations humaines (et la justice en est un exemple parfait), rien n'est aussi simple, aussi automatique que les lois de mécanique des fluides qui permettent de modéliser des phénomènes météorologiques.

Notes

1. En argot (slang) « wasted » *ne signifie pas « gaspillé » mais « ivre ».*

2. Décrit dans « In Your Face: China's all-seeing state », BBC*, 10 décembre 2017. http://www.bbc.com/news/av/world-asia-china-42248056/in-your-face-china-s-all-seeing-state*

3. https://www.numerama.com/politique/300907-un-algorithme-teste-par-la-police-pour-anticiper-les-crimes-des-jeunes-inquiete-laustralie.html

Bibliographie

BINET J.-L., « La prévention médicale de la récidive chez les délinquants sexuels », Académie de médecine, 2010. *http://www.aihus.fr/prod/data/news/rapport_recidive_delinq_sex.pdf*

BOTSMAN R., *Who Can You Trust?: How Technology Brought Us Together – and Why It Could Drive Us Apart*, Portfolio Penguin, 2017.

CHARPENTIER A. ; SUIRE R., « Données et santé : valeurs, acteurs et enjeux », *Risques*, n° 107, septembre 2016.

CHARPENTIER A., « Les dérives du principe de précaution », *Risques*, n° 108, décembre 2016.

CHRISTIN A. ; ROSENBLAT A.; BOYD D., "Courts and Predictive Algorithms", Data & Civil Rights Conference, 2015. *http://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf*

DRESSEL J. ; FARID H., "The Accuracy, Fairness, and Limits of Predicting Recidivism", Science Advances, 2018. *http://advances.sciencemag.org/content/4/1/eaao5580.full*

FOUCAULT M., *Surveiller et punir. Naissance de la prison*. Gallimard, 1975.

GALEON D. ; BERGAN B., "China's 'Social Credit System' Will Rate How Valuable You Are as a Human". *futurism.com*, 2017. *https://futurism.com/china-social-credit-system-rate-human-value/*

GNEITING T. ; BALABDAOUI F. ; RAFTERY A., "Probabilistic Forecasts, Calibration and Sharpness". *Journal of the Royal Statistical Society* (JRRS-B), vol. 69, issue 2, 2007, pp. 243-268.

MANGEL M. ; SAMANIEGO F., « Abraham Wald's Work on Aircraft Survivability », *Journal of the American Statistical Association*, vol. 79, n° 386, 1984, pp. 259-267.

MCLANNAHAN B., "Being 'Wasted' on Facebook May Damage Your Credit Score", *Financial Times*, octobre 2015. *https://www.ft.com/content/d6daedee-706a-11e5-9b9e-690fdae72044*

SUPIOT A., *La gouvernance par les nombres. Cours au Collège de France (2012-2014)*, Fayard, 2015.

TRUJILLO E., « La Chine met en place un système de notation de ses citoyens pour 2020 ». *Le Figaro*, 27 décembre 2017. *http://www.lefigaro.fr/secteur/high-tech/2017/12/27/32001-20171227ARTFIG00197-la-chine-met-en-place-un-systeme-de-notation-de-ses-citoyens-pour-2020.php*