# Fairness Machine Learning

Sujet de thèse

October 24, 2019

## 1 Contexte

Les modèles d'apprentissage automatique ont un rôle de plus en plus important dans la prise de décisions, celles-ci peuvent avoir des impacts significatifs. Par exemple, dans le cas des assurances, ces algorithmes sont appliqués pour établir l'approbation de prêt ou la détection de fraudes.

Ces modèles sont pour la plupart utilisés comme des boîtes noires qui produisent une prédiction. Il est difficile d'acquérir une bonne compréhension du comportement du modèle, et en particulier de l'influence des différentes caractéristiques sur la prédiction du modèle. La potentielle discrimination engendrée par les modèles devient un sujet majeur. La Réglementation Générale de l'Union Européenne sur la Protection des Données (RGPD) constitue un des changements les plus important en matière de réglementation de la confidentialité des données depuis plus de 20 ans. En particulier, les algorithmes entraînés sur des données biaisées ont le potentiel d'apprendre et de perpétuer ces biais. L'injustice est caractérisée lorsqu'il y a une discrimination à l'égard d'un groupe particulier de personnes en raison de caractéristiques sensibles telles que la religion, l'origine ou le sexe. Ces dernières années, de nombreux cas ont été recensés : comme le cas des modèles algorithmiques utilisés pour générer des prédictions de récidive criminelle aux États-Unis (COMPAS) ou des publicités ciblées et automatisées en ligne sur les opportunités d'emploi où il a été démontré une discrimination sur des caractéristiques protégées (l'origine ou le sexe) [Angwin et al., 2016],[Lambrecht and E. Tucker, 2016]. Le comportement discriminatoire n'est pas une caractéristique délibérée de ces algorithmes mais plutôt le résultat de biais présents dans les données d'entrée utilisées pour former les systèmes [Calders and Žliobaitė, 2013].

L'un des problèmes majeurs rencontrés est qu'il ne s'agit pas simplement de supprimer les attributs protégés de la base d'entraînement pour obtenir un modèle non-discriminatoire. En effet, une combinaison complexe d'attributs peut fournir un lien inattendu avec des informations sensibles et ainsi servir de substituts d'un attribut protégé. Par exemple, les informations telles que le modèle ou la couleur de voiture peuvent être corrélées avec le sexe du propriétaire. L'un des nouveaux défis pour les Data Scientist est donc de déterminer si les modèles présentent des biais

discriminatoires et s'ils peuvent les atténuer le plus possible, il ne s'agit donc plus d'un unique objectif de performance de modèle.

# 2 État de l'art, la littérature :

L'un des premiers enjeux avant l'entraînement des modèles d'apprentissage consiste à déterminer si les données brutes utilisées pour former le modèle contiennent des biais discriminatoires. Il existe de nombreuses façons de quantifier les biais et de nombreuses subtilités à prendre en compte pour interpréter les résultats de ces mesures. On dénombre plus de 71 mesures de détection de biais [Bellamy et al., 2018], des chercheurs ont démontré qu'il est impossible de satisfaire à toutes les définitions de l'équité en même temps [Kleinberg et al., 2016]. A noter qu'il ne s'agit pas seulement de différences théoriques sur la façon de mesurer l'équité, le concept est complexe car il dépend du contexte et de la culture.

Il y a deux notions principales d'équité :

- l'équité de groupe : divise une population en groupes définis par des attributs protégés (comme la religion, le sexe, l'origine) et cherche à ce qu'une certaine mesure statistique soit égale pour tous les groupes (deux métriques sont très utilisés : impact disparate, Equal Opportunity).
- L'équité individuelle : cherche à ce que des personnes semblables soient traitées de la même manière.

Le second enjeu consiste à pouvoir atténuer les biais de discrimination contenus dans une data biaisée. Pour traiter ce problème, trois approches différentes sont proposées :

- Prétraitement équitable : Il s'agit de transformer les données d'entraînement en une représentation équitable.
- Traitement en cours équitable : Modification de l'entraînement de l'algorithme en gardant les données biaisées.
- Post-traitement équitable : L'entraînement est déjà réalisé, il s'agit de modifier les prédictions biaisées en sortie pour réduire la discrimination.

# 3 Approche méthodologique :

L'approche qui semble privilégiée est de modifier/adapter les algorithmes d'apprentissage automatiques utilisés habituellement en assurance afin d'atténuer les biais de discrimination. En effet, la littérature sur le sujet montre une application sur des algorithmes très peu applicable en production assurantielle.

La première idée serait d'apporter une modification des algorithmes classiquement utilisés en assurance à l'aide de technique d'adversaire. Il s'agit donc d'une approche en "traitement en cours équitable", c'est à dire, la modification de l'entraînement durant l'apprentissage du modèle. Un algorithme adversaire serait entrainé à prédire l'attribut sensible et permettrait de pénaliser, pendant la phase d'apprentissage, l'algorithme de prédiction assurantiel.

D'une autre part, apporter une approche en "Post-traitement équitable" serait très intéressante car cela permettrait de ne pas ré-entrainer les modèles déjà en production. En effet, cela est très coûteux et cela prendrait un temps considérable. Un second algorithme viendrait modifier les prédictions en sortie de l'algorithme en production de manière à diminuer significativement la discrimination.

Enfin, il serait intéressant, outre le fait d'utiliser des mesures de détection de biais, d'apporter une part d'explication simple et claire sur les biais engendrés.

# References

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. ProPublica, May 23, 2016.

[Bellamy et al., 2018] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

[Calders and Žliobaitė, 2013] Calders, T. and Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Studies in Applied Philosophy, Epistemology and Rational Ethics*, volume 3, pages 43–57. Springer.

[Kleinberg et al., 2016] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. pages 1–23.

[Lambrecht and E. Tucker, 2016] Lambrecht, A. and E. Tucker, C. (2016). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*.