Journée de la recherche en sciences

Equité algorithmique et discrimination

Agathe Fernandes Machado

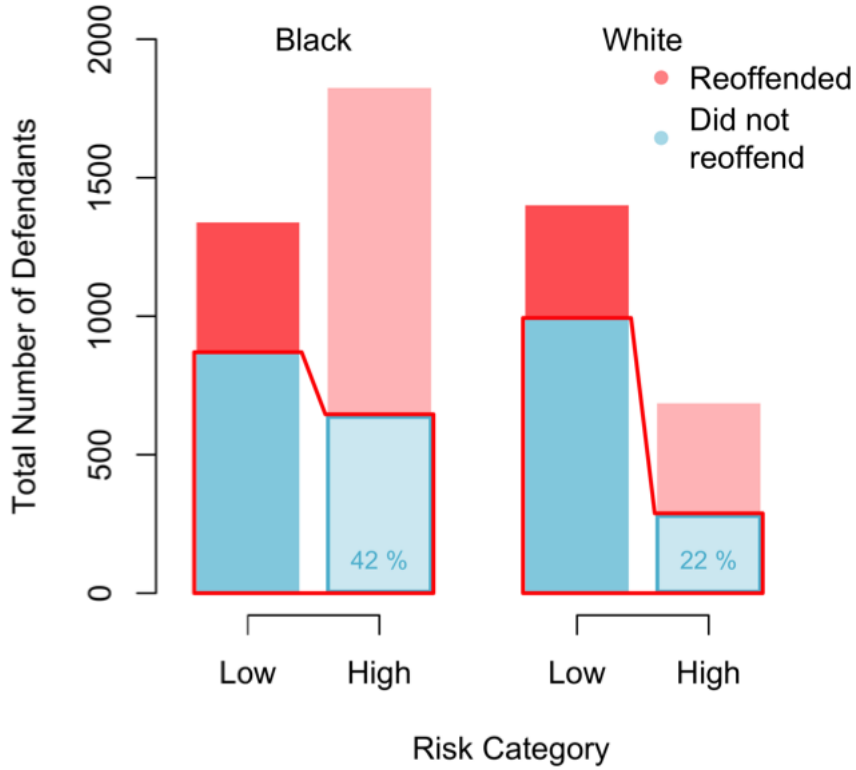
Doctorat en Mathématiques

04/04/2024, UQAM

Discrimination des modèles prédictifs

Correctional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**ection

- Outil d'aide à la décision pour la justice aux Etats-Unis
- Algorithme de prédiction du score de récidive d'un prévenu
- Basé sur 137 variables et le casier judiciaire du prévenu : **adresse postale**, GPA du lycée, etc.



https://github.com/freakonometrics/MAT998X.git

La **variable sensible** origine ethnique n'est pas prise en compte dans le modèle

Origine de la discrimination

Biais statistique dans les données

- Reproduction des injustices du passé
- Minorité sous-représentée dans un jeu de données déséquilibré

X Variables explicatives du modèle

Variables proxy : corrélation entre un attribut sensible et d'autres variables explicatives

➢ Retirer la variable sensible ne suffit pas à éliminer la discrimination

Biais intentionnel

Le biais peut être le résultat de choix délibérés, pouvant être bienveillants ou malveillants.

Evaluation de l’équité algorithmique

Réponse à la **législation** : AI Act (Europe) vise à interdire ou limiter les systèmes d'IA en production présentant un « niveau de risque inacceptable »

« Group Fairness »

- ✓ Sexe
- ✓ Origine ethnique
- ✓ Age
- ✓ Score de crédit

« Demographic Parity »
Indépendance des prédictions avec la/les variable(s) sensible(s)

Calibration
Correspondance entre les scores prédits par un **modèle de classification binaire** et la **fréquence empirique** des évènement observés

Méthodes de correction des biais

x Pre-processing

Déformation des données de l'échantillon d'entraînement pour garantir un modèle « juste »

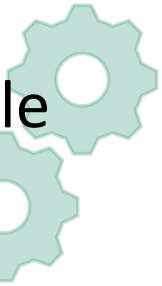
$f(x)$ In-processing

Ajout d'une pénalité portant sur l'équité du modèle dans la fonction objectif

$\hat{y}$ Post-processing

Transformation des prédictions obtenues grâce au modèle afin de les rendre « justes »

Déploiement d'un modèle de **Machine Learning**



EquiPy
Correction des biais liés à plusieurs variables sensibles

Références

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*, 23-05.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Arthur Charpentier (2024). Insurance Biases, Discrimination and Fairness.

François Hu, Philipp Ratz, and Arthur Charpentier (2023). A sequentially fair mechanism for multiple sensitive attributes. *AAAI-2024*.

A. Fernandes Machado, A. Charpentier, E. Flachaire, E. Gallic, F. Hu (2024). From Uncertainty to Precision: Enhancing Binary Classifier Performance through Calibration.

A. Fernandes Machado, F. Hu, P. Ratz, S. Grondin and A. Charpentier (2024). Documentation du package Python EquiPy: https://equilibration.github.io/equipy/.

Agathe Fernandes Machado
https://fer-agathe.github.io
fernandes_machado.agathe@courrier.uqam.ca
UQAM, Pavillon PK, 5ème étage