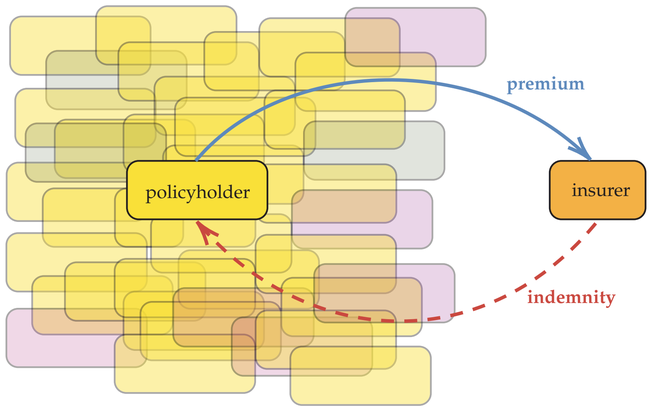
Fairness of Machine Learning in Insurance
New Rags for an Old Man?

Arthur Charpentier[a] (@freakonometrics) & Laurence Barry[b]

[a]Université du Québec à Montréal (charpentier.arthur@uqam.ca) [b]Chaire PARI

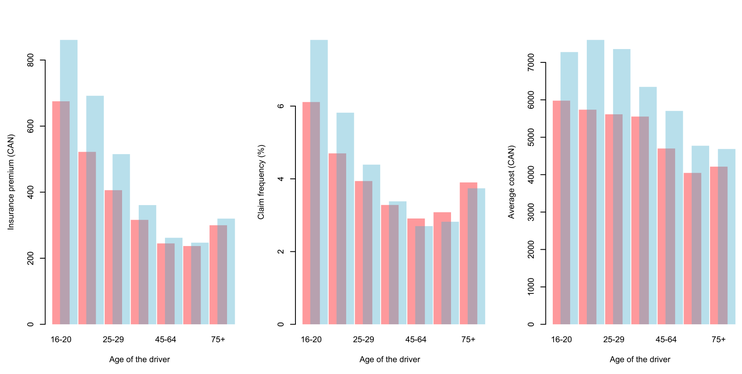
Introduction

"*Insurance is the contribution of the many to the misfortune of the few*", Bénéplanc et al. (2022). "*at the core of insurance business lies discrimination between risky and non-risky insureds*", Avraham (2017), to ask for an **actuarially fair premium**.



By looking into earlier debates on discrimination, we show that some algorithmic biases are a renewed version of older ones, while others show a reversal of the previous order. Paradoxically, while the insurance practice has not deeply changed nor are most of these biases new, the machine learning era still deeply shakes the conception of insurance fairness.

Discrimination & Insurance



	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	x	x	●	x	●	x	x	●	●	●	●	x	x	●
Age	x	x	●	x*	●	x	●	●	●	●*	●	x	x	●
Driving experience	●	x	●	●	●	●	●	●	●	●	●	●	●	●
Credit history	x	x	●	●	●	x	●*	●	●	x*	x	●*	x	●
Education	x	x	x	x	x	x	●	●	●	●	●	●	●	●
Profession	x	x	x	●	x	x	●	●	●	●	●	●	●	●
Employment	x	x	x	●	x	x	●	●	●	●	●	●	●	●
Family	●	x	●	●	●	x	●	●	●	●	●	●	●	●
Housing	x	x	●	●	●	x	●	●	●	x	x	●	●	●
Address/ZIP code	●	●	●	●	●	●	●	●	●	x	x	●	●	●

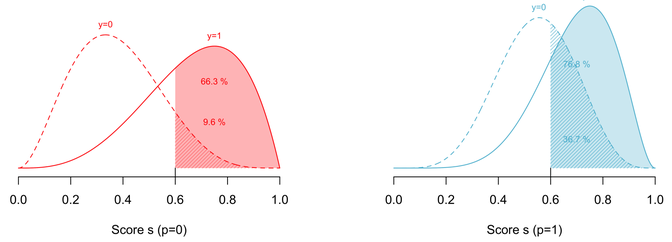
But not much on **proxy based discrimination** (standard in high dimension).

Group fairness

Classically, y is the variable of interest, e.g. accident occurrence ($\{0,1\}$), frequency ($\mathbb{N}$) or annual cost , time before invalidity recovery ($\mathbb{R}_+$), etc), $\boldsymbol{x}$ some covariates and p is a protected (sensitive) attribute (in $\{0,1\}$).
Dwork et al. (2012),Hardt et al. (2016),Corbett-Davies et al. (2017), Berk et al. (2017) introduced various concepts of fairness

- **Demographic parity**, $\mathbb{P}[\widehat{Y}=1|P=p]=\text{cst},\ \forall p$
- **Equalized odds**, $\mathbb{P}[\widehat{Y}=1|P=p,Y=y]=\text{cst}_y,\ \forall p,y$
- **Equalized opportunity**, $\mathbb{P}[\widehat{Y}=1|P=p,Y=1]=\text{cst},\ \forall p$
- **Predictive equality**, $\mathbb{P}[\widehat{Y}=1|P=p,Y=0]=\text{cst},\ \forall p$
- **conditional accuracy equality** $\mathbb{P}[Y=y|P=p,\widehat{Y}=y]=\text{cst}_y,\ \forall p,y$

Those notions are simply related to independence $\widehat{Y}\perp\!\!\!\perp P$, or conditional independence $\widehat{Y}\perp\!\!\!\perp P\mid Y$ or $Y\perp\!\!\!\perp P\mid\widehat{Y}$. Let s denote the score function, $s(\boldsymbol{x},p)=\mathbb{P}(Y=1|\boldsymbol{x},p)$ so that $\widehat{Y}=\mathbf{1}(s>\text{threshold})$.



Inspired by Goodfellow et al. (2018), to avoid un-fairness, penalize according to $\text{HGR}(\widehat{y},p)$ (a distance to independence) e.g.

$$\underset{\boldsymbol{\theta},\boldsymbol{\omega}}{\text{argmin}}\left\{\mathcal{L}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)+\lambda\text{HGR}_{\boldsymbol{\omega}}(\widehat{y},p)\right\},$$

where $\text{HGR}(U,V)$ is $\max\big\{\text{corr}[f(U),g(V)]\big\}=\max_{f\in\mathcal{S}_U,g\in\mathcal{S}_V}\big\{\mathbb{E}[f(U)g(V)]\big\}$ and where $\mathcal{S}_U=\{f:\mathcal{U}\to\mathbb{R}:\mathbb{E}[f(U)]=0\text{ and }\mathbb{E}[f(U)^2]=1\}$ and similarly $\mathcal{S}_V$. But a conditional version, can also be considered

$$\text{HGR}(U,V|Z)=\max_{f\in\mathcal{S}_{U|Z},g\in\mathcal{S}_{V|Z}}\Big\{\mathbb{E}[f(U)g(V)\big|Z]\Big\}.$$

Individual Fairness

Duivesteijn and Feelders (2008), Kusner et al. (2017)

- **Lipschitz property**, $D(\widehat{y}_i,\widehat{y}_j)$ or $D(s_i,s_j)\le d(\boldsymbol{x}_i,\boldsymbol{x}_j),\ \forall i,j$
- **Counterfactual fairness**,

$$\mathbb{P}\big[Y^\star_{P\leftarrow p}=y|\boldsymbol{X}=\boldsymbol{x}\big]=\mathbb{P}\big[Y^\star_{P\leftarrow p'}=y|\boldsymbol{X}=\boldsymbol{x}\big],\ \forall p',\boldsymbol{x},y.$$

If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity.
Black et al. (2020), Torous et al. (2021) or de Lara et al. (2021) suggested

- **Transport-based fairness**, $\mathbb{P}_0[\mathcal{X}_F^-(m,T^\star)]=\mathbb{P}_0[\mathcal{X}_F^+(m,T^\star)]$

$$\begin{cases}\mathcal{X}_F^+(m,T^\star)=\big\{\boldsymbol{x}\in\mathcal{X}:m(\boldsymbol{x},0)>m(T^\star(\boldsymbol{x}),1)\big\}\\ \mathcal{X}_F^-(m,T^\star)=\big\{\boldsymbol{x}\in\mathcal{X}:m(\boldsymbol{x},0)<m(T^\star(\boldsymbol{x}),1)\big\},\end{cases}$$

where $\mathbb{P}_0(\cdot)=\mathbb{P}(\cdot|p=0)$, $\mathbb{P}_1(\cdot)=\mathbb{P}(\cdot|p=1)$, and

$$T^\star\in\underset{T:T_\#\mathbb{P}_0=\mathbb{P}_1}{\text{argmin}}\int_{\mathbb{R}^k}\|\boldsymbol{x}-T(\boldsymbol{x})\|^2d\mathbb{P}(\boldsymbol{x}).$$

To go further

The question we address is

- How to quantify properly price discrimination ?

but then, remaining questions are

- How to identify the source of the discrimination ?
- How to get a ratemaking formula that is "discrimination free" ?

with technical challenges, **continuous protected attribute**, Grari et al. (2022), **non-observed protected attribute**, Grari et al. (2021), Racicot et al. (2021) and connections with **causal-related fairness**, and prevention.

References

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Barry, L. (2020). Insurance, big data and changing conceptions of fairness. *European Journal of Sociology*, 61.

Barry, L. and Charpentier, A. (2022). The Fairness of Machine Learning in Insurance: New Rags for an Old Man? . *ArXiv*, 2205.08112.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.

Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111–121.

Bénéplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.

Charpentier, A. (2022). *Insurance: biases, discrimination and fairness*. ILB, Opinions & Débats, 25.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

de Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021). Transport-based counterfactual models. *arXiv*, 2108.13025.

Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56—66.

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv*, 2202.12008.

Grari, V., Lamprier, S., and Detyniecki, M. (2021). Fairness without the sensitive attribute via causal variational autoencoder. *arXiv*, 2109.04999.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Heras, A. J., Pradier, P.-C., and Teira, D. (2020). What was fair in actuarial fairness? *History of the Human Sciences*, 33(2):91–114.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Racicot, T., Khoury, R., and Pere, C. (2021). Estimation of uncertainty bounds on disparate treatment when using proxies for the protected attribute. In *Canadian Conference on AI*.

Torous, W., Gunsilius, F., and Rigollet, P. (2021). An optimal transport approach to causal inference. *arXiv*, 2108.05858.