Fairness

Marco Scutari[1] and Francesca Panero[2]

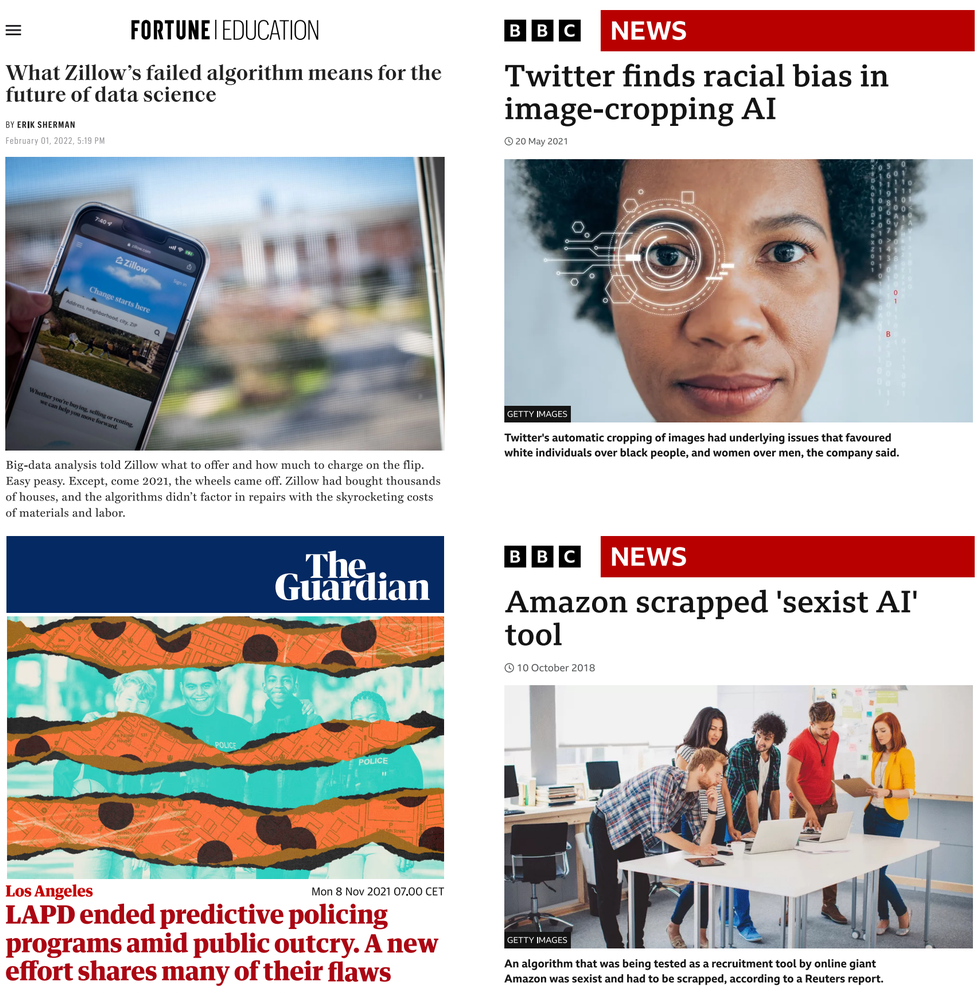
[1] IDSIA [2] Department of Statistics, LSE

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Machine Learning Has No Understanding

Machine learning creates black boxes that use probabilistic associations for prediction. It does not understand the causality that drives reality, and often struggles with uncertainty and bias.

We Need Fairness!



What Zillow's failed algorithm means for the future of data science

BY ERIK SHERMAN

February 01, 2022, 5:19 PM

Big-data analysis told Zillow what to offer and how much to charge on the flip. Easy peasy. Except, come 2021, the wheels came off. Zillow had bought thousands of houses, and the algorithms didn't factor in repairs with the skyrocketing costs of materials and labor.

Twitter finds racial bias in image-cropping AI

20 May 2021

GETTY IMAGES

Twitter's automatic cropping of images had underlying issues that favoured white individuals over black people, and women over men, the company said.

The Guardian

Los Angeles Mon 8 Nov 2021 07.00 CET

LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws

Documents show how data-driven policing programs reinforced harmful patterns, fueling the over-policing of Black and brown communities

BBC NEWS

Amazon scrapped 'sexist AI' tool

10 October 2018

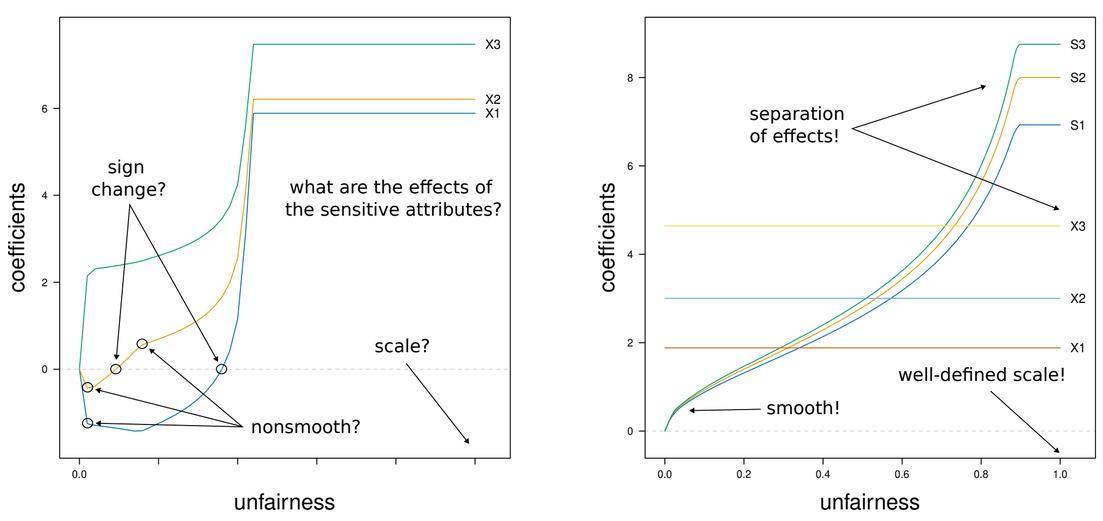
GETTY IMAGES

An algorithm that was being tested as a recruitment tool by online giant Amazon was sexist and had to be scrapped, according to a Reuters report.

What We Want from Fair Models

- Interpretability: What do the parameters mean in the domain the data come from?
- Explainability: Why does the model give specific predictions?
- Best practices: What are the best ways to select and validate models?
- Confidence: Are the effects we measure statistically significant? What is the error margin of predictions?

Optimisation vs Statistics



References

M. Scutari, F. Panero and M. Proissl (2022). Achieving Fairness with a Simple Ridge Penalty. Statistics and Computing, 32, 77. https://cran.r-project.org/web/packages/fairml/

The Fair Ridge Regression Model (FRRM)

Consider a regression model with $\mathbf{X}$ predictors, $\mathbf{S}$ sensitive attributes, response $\mathbf{y}$ and a given level of fairness $r \in [0, 1]$ (0 = complete fairness, 1 = no fairness constraints).

1. Compute $\widehat{\mathbf{U}}$ (the fair predictors) as $\mathbf{X} = \mathbf{S}\widehat{\mathbf{B}}_{\text{OLS}} + \widehat{\mathbf{U}}$.
2. Estimate $\widehat{\boldsymbol{\beta}}_{\text{FRRM}} = \left(\widehat{\mathbf{U}}^{\text{T}}\widehat{\mathbf{U}}\right)^{-1}\widehat{\mathbf{U}}^{\text{T}}\mathbf{y}$.
3. Estimate $\widehat{\boldsymbol{\alpha}}_{\text{OLS}} = \left(\mathbf{S}^{\text{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\text{T}}\mathbf{y}$.

Then:

4. If $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\text{OLS}}, \widehat{\boldsymbol{\beta}}_{\text{OLS}}) \leqslant r$, set $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}} = \widehat{\boldsymbol{\alpha}}_{\text{OLS}}$.
5. Otherwise, find the value of $\lambda(r)$ that satisfies

$$R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\text{VAR}(\mathbf{S}\boldsymbol{\alpha})}{\text{VAR}(\mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\text{FRRM}})} = r$$

and estimate the associated $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ in the process.

FRRM solves $\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{argmin}}\, \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \widehat{\mathbf{U}}\boldsymbol{\beta}\|_2^2 + \lambda(r)\|\boldsymbol{\alpha}\|_2^2$.

- ✔ Single solution, computationally inexpensive.
- ✔ Pluggable fairness constraints.
- ✔ Works for all generalised linear models.
- ✔ Interpretable and explainable.
- ✔ A statistical model we know very well.

Confidence and Uncertainty

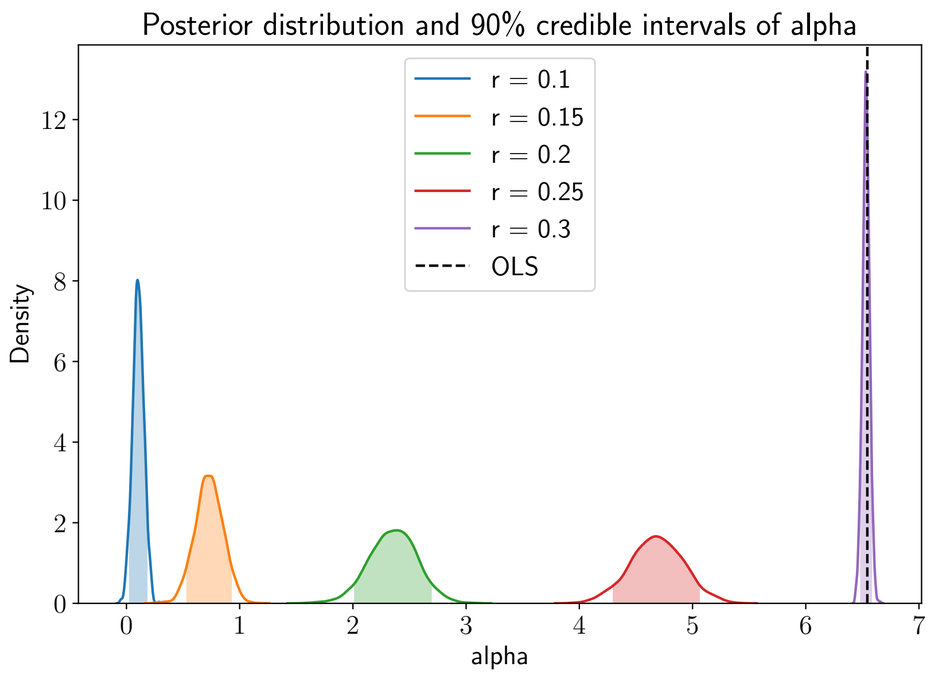
Bayesian statistics allows us to quantify uncertainty in our estimates in an easy and interpretable way. To design the Bayesian version of FRRM, we define the likelihood

$$\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \widehat{\mathbf{U}}, \mathbf{S} \sim \text{MVN}\left(\widehat{\mathbf{U}}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\alpha}, \frac{1}{\tau}I\right)$$

and the priors on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, the penalty parameter λ and the precision τ:

$$\begin{aligned}
\boldsymbol{\beta} \mid \tau &\sim \text{MVN}\left(\mathbf{0}, \frac{1}{\tau}I\right) \\
\boldsymbol{\alpha} \mid \tau, \lambda &\sim \text{MVN}\left(\mathbf{0}, \frac{1}{\tau\lambda}I\right) \\
\lambda &\sim \text{Gamma}\left(\frac{1}{r^5} - 1, 1\right) \\
\tau &\sim \text{Gamma}\,(1, 1)
\end{aligned}$$

Using this scheme, we can design a Gibbs sampler or Hamiltonian Monte Carlo algorithm to sample from the posterior of the parameters. This allows us to easily estimate the posterior distribution of the parameters and investigate their credible intervals.



email: scutari@bnlearn.com