The role of Bayesian uncertainty in assessing the fairness of machine learning models

Francesca Panero[1], Ernst Wit[2] and Marco Scutari[3]

[1] MEMOTEF, Sapienza University of Rome [2] Università della Svizzera Italiana [3] IDSIA

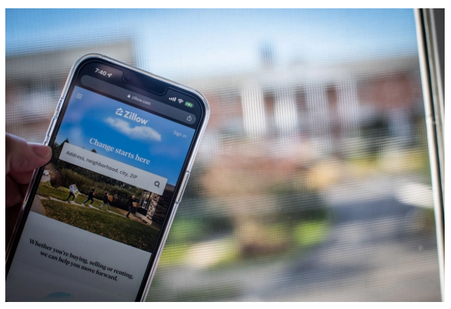
We Need Fairness!

FORTUNE | EDUCATION

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

BY ERIK SHERMAN

February 01, 2022, 5:19 PM

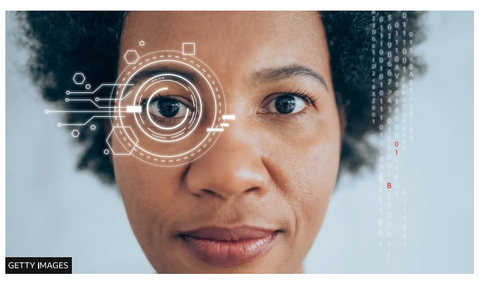


Big-data analysis told Zillow what to offer and how much to charge on the flip. Easy peasy. Except, come 2021, the wheels came off. Zillow had bought thousands of houses, and the algorithms didn't factor in repairs with the skyrocketing costs of materials and labor.

BBC NEWS

Twitter finds racial bias in image-cropping AI

20 May 2021



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

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?
- Confidence: Are the effects we measure statistically significant? What is the error margin of predictions?

(Frequentist) Fair Ridge Reg 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).

To bound the dependence between $\hat{\mathbf{y}}$ and $\mathbf{S}$ by a fairness budget r, Scutari et al. (2022) propose:

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. To find $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$, solve

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

where $\lambda(r)$ satisfies

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

- ✓ Single solution, computationally inexpensive.
- ✓ Pluggable fairness constraints, interpretable and explainable.
- ✓ Works for all generalised linear models.
- ✗ Lacking uncertainty quantification.

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 Bayesian alternative (BFRRM)

To design the Bayesian version of FRRM, we define the likelihood

$$\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \widehat{\mathbf{U}}, \mathbf{S} \sim \mathcal{N}\left(\widehat{\mathbf{U}}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\alpha}, \sigma^2 I_n\right)$$

and the priors on $\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2$:

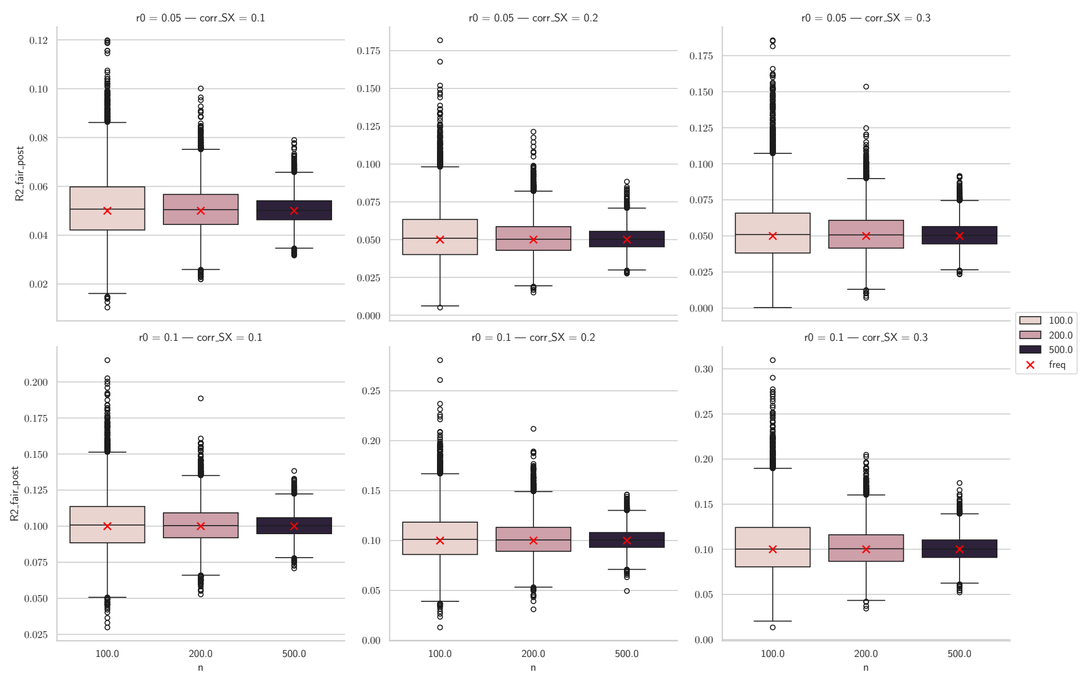
$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}\left(0, g\sigma^2(\widehat{\mathbf{U}}^{\text{T}}\widehat{\mathbf{U}})^{-1}\right), \; g > 0$$

$$\boldsymbol{\alpha} \mid \sigma^2, \lambda(r) \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(r)} I_q\right)$$

$$\sigma^2 \mid a_0, b_0 \sim IG(a_0, b_0)$$

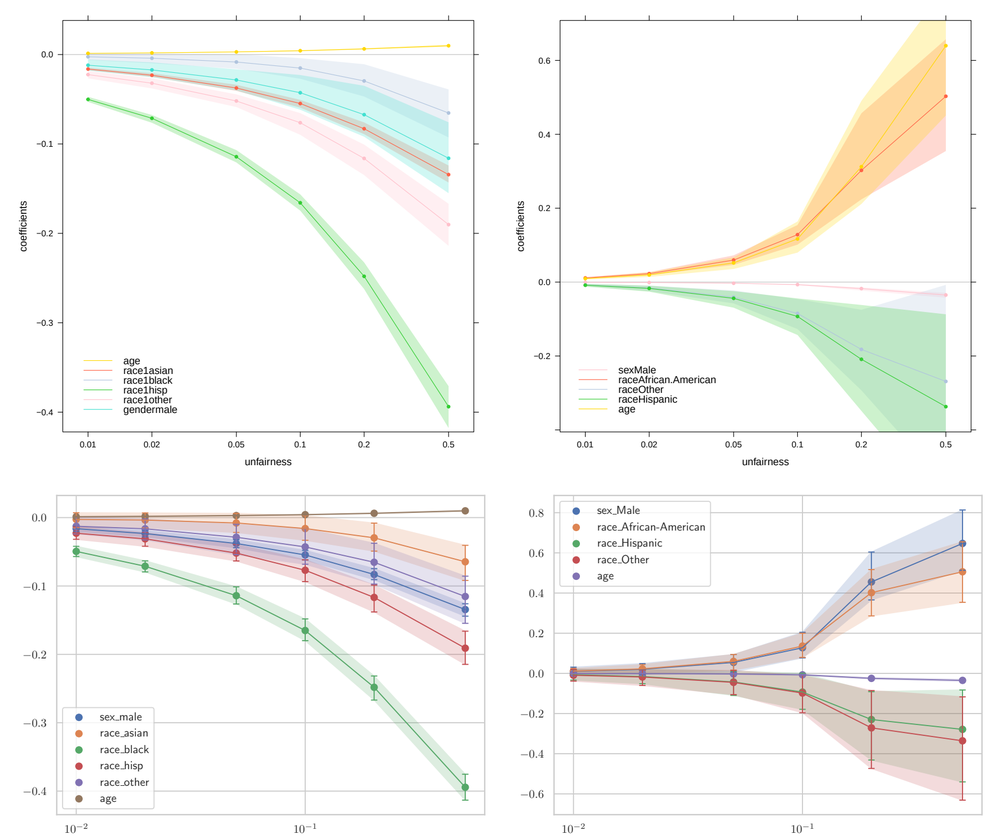
This formulation gives closed form posteriors for all parameters. Therefore, we can easily obtain posterior means and credible intervals for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2$. From these, we derive uncertainty measures for the unfairness measure $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Posterior calibration of BFRRM



Real data

LSAC is a survey among US law school students ($n = 5$K, 9 predictors): we take GPA as response, and race, age, gender as sensitive attributes. COMPAS ($n = 6$K, 13 predictors) comprises demographic and criminal records of offenders in Florida: we take recidivating as response and offender's age, gender, race as sensitive attributes.



email: scutari@bnlearn.com; francesca.panero@uniroma1.it