Fair Bayesian Optimization

V. Perrone, M. Donini, M. B. Zafar, R. Schmucker*, K. Kenthapadi, C. Archambeau

Amazon Web Services | {vperrone, donini, zafamuh}@amazon.com

* = Work done during internship at AWS. Currently at CMU - rschmuck@andrew.cmu.edu

Algorithmic Fairness and Hyperparameters

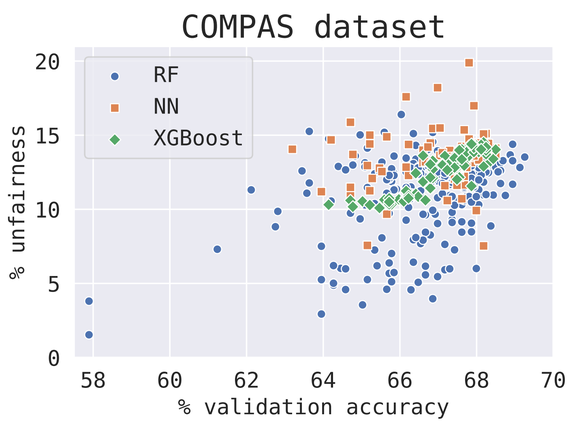
Unfairness in Machine Learning models: with the increasing use of machine learning (ML) in domains such as financial lending, hiring, criminal justice, and college admissions, there has been a major concern for the potential for ML to unintentionally encode societal biases and result in systematic discrimination.

Practical ML settings → optimizing the performance of ML models in a black-box manner while enforcing fairness constraints.

Issue → being tailored to specific models and fairness definitions, most existing algorithmic fairness techniques are inapplicable to these settings.

Intuition → optimizing hyperparameters for both fairness and accuracy makes possible to obtain less biased and still accurate models.

Our solution supports arbitrary fairness definitions, allows for multiple constraints to be enforced simultaneously, and is complementary to existing bias mitigation techniques.



Unfairness-accuracy trade-off by varying the hyperparameters of XGBoost, RF, and NN on a recidivism prediction task. Each dot corresponds to a different hyperparameter configuration. For a given level of accuracy, models with very different levels of unfairness can be generated simply by changing the model hyperparameters.

Statistical definitions of fairness: there is no consensus on a unique definition of fairness, and some of the most common definitions are conflicting.

Equal Opportunity (EO): equal True Positive Rates (TPR) across subgroups;

Equalized Odds (EOdd): equal False Positive Rates (FPR), in addition to EO;

Statistical Parity (SP): positive predictions to be unaffected by the value of the protected attribute, regardless of the actual true label;

A model is ϵ-fair if it violates the fairness definition by at most $\epsilon \geq 0$. In the case of EO, a model is ϵ-fair if the difference in EO (DEO) is at most ϵ, i.e. if the absolute value of difference of the TPRs (across subgroups) is at most ϵ.

Fair Bayesian Optimization

We propose **Fair Bayesian Optimization (FairBO)** to optimize the hyperparameters of a black-box function while satisfying arbitrary fairness constraints. Our goal is to find

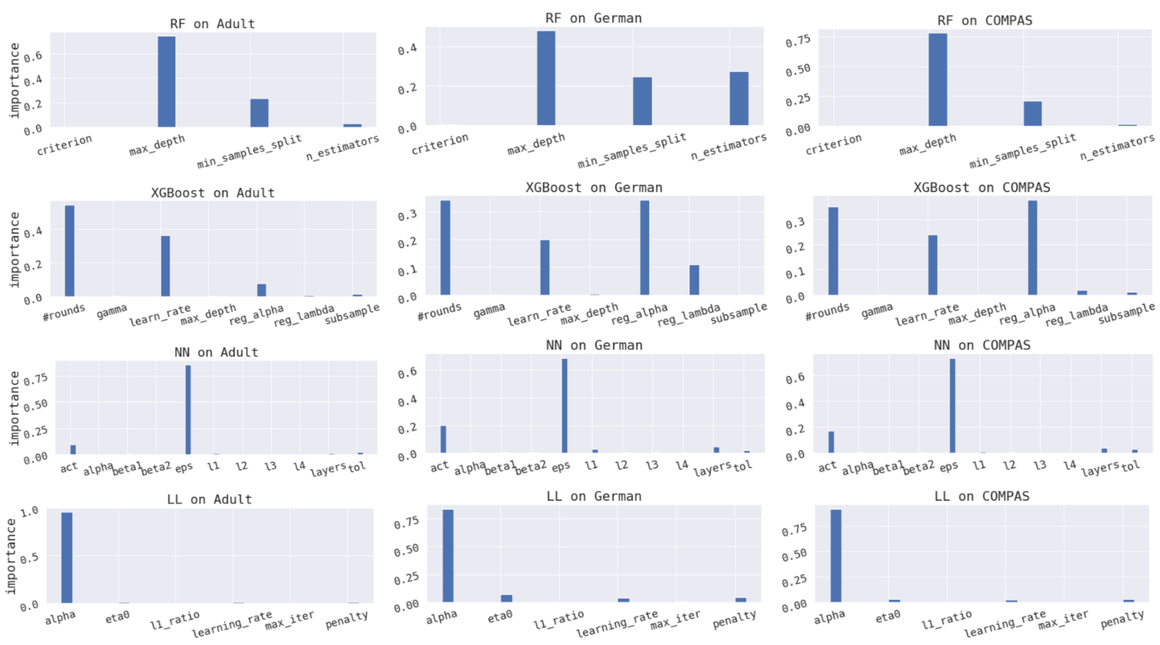
$$\min_{\mathbf{x}\in\mathcal{X}} \{y(\mathbf{x}) \parallel c(\mathbf{x}) \leq \epsilon\},$$

where $y(x)$ is the main objective, $c(\mathbf{x})$ a fairness constraint, and $\epsilon \in \mathbb{R}^+$ an unfairness upper bound. The idea is to place one model on the objective and one on the fairness constraint (e.g., two independent Gaussian processes), and encode the probability of satisfying the fairness definition in the acquisition.

We leverage the **constrained expected improvement (cEI)**, an established acquisition function to extend BO to the constrained case:

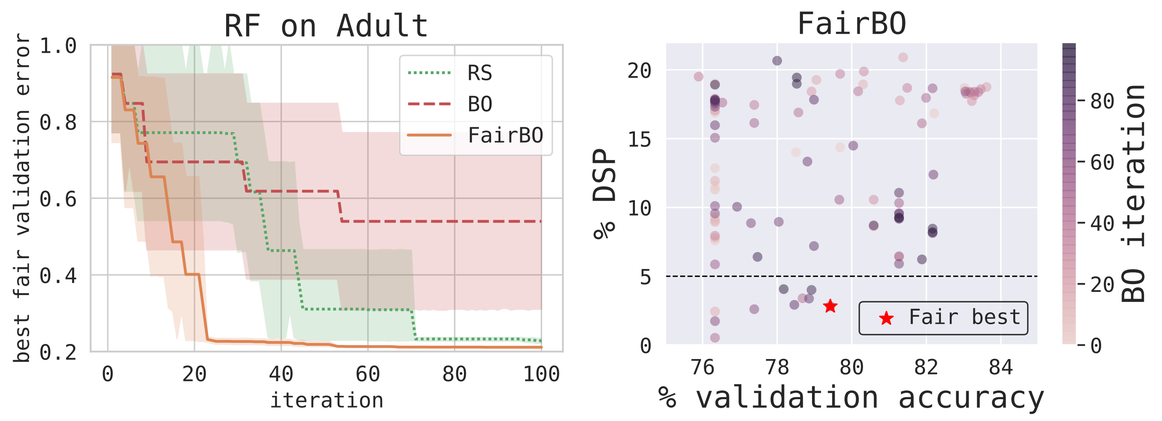
$$cEI(\mathbf{x}) = P(c(\mathbf{x}) \leq \epsilon)EI(\mathbf{x}), \text{ penalizing unfair hyperparameter.}$$

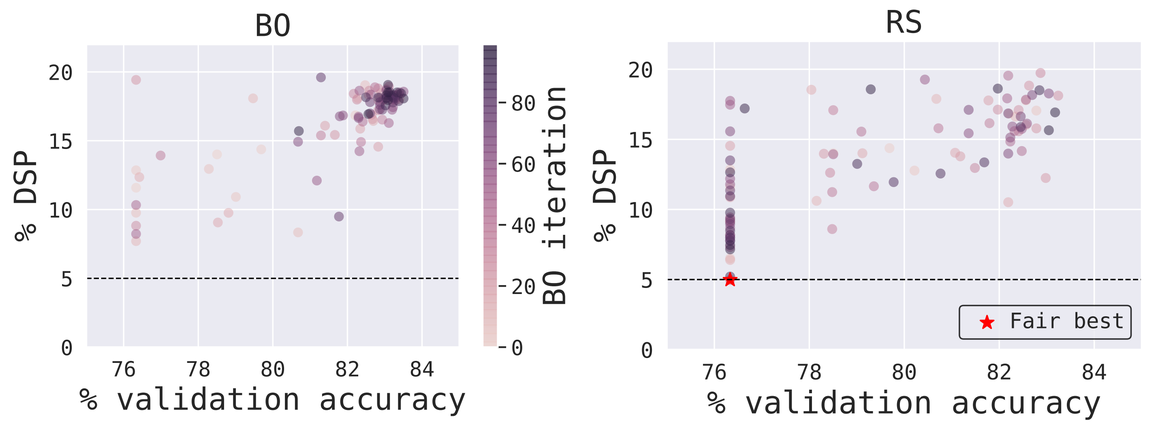
Hyperparameter importance on fairness



We tune four algorithms: **Random Forest (RF), XGBoost, neural network (NN), and linear learner (LL)**. The role of each tuned hyperparameter on unfairness (DSP) is evaluated via fANOVA: the **regularization hyperparameters** impact fairness the most (e.g., in LL this is precisely the regularization factor `alpha`).

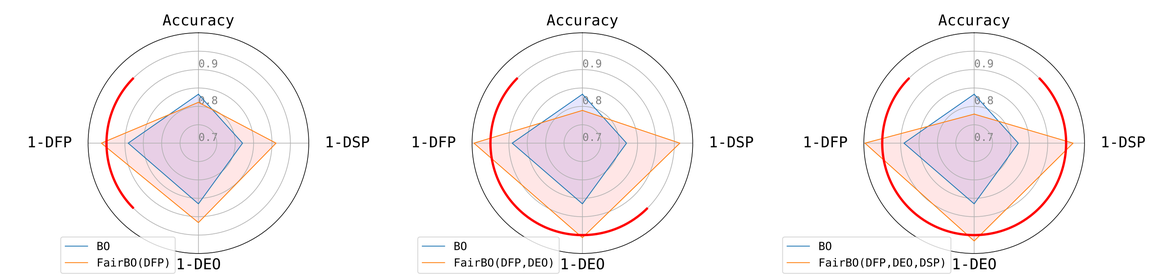
FairBO performance





FairBO finds a fair accurate solution more quickly than random search (RS) and standard BO, which can get stuck in accurate yet unfair regions and fail to return a feasible solution. The horizontal black line represents the fairness constraint DSP ≤ 0.05.

Multiple fairness constraints



- In contrast to most algorithmic fairness techniques, FairBO can seamlessly handle multiple fairness definitions simultaneously.
- Tuning RF on Adult, progressively adding more fairness constraints, represented by the red arches (DFP ≤ 0.05, DEO ≤ 0.05, DSP ≤ 0.05).
- FairBO allows us to trade off relatively little accuracy for a more fair solution, which gets progressively more fair as we add more constraints.

Model-agnostic and model-specific techniques

Validation error of the best fair models for model-specific (first three rows) and model-agnostic fairness methods. We use the fairness constraint, DSP ≤ 0.1.

Method	Adult	German	COMPAS
FERM [1]	0.164 ± 0.010	0.185 ± 0.012	0.285 ± 0.009
Zafar [2]	0.187 ± 0.001	0.272 ± 0.004	0.411 ± 0.063
Adversarial [3]	0.237 ± 0.001	0.227 ± 0.008	0.327 ± 0.002
FERM preprocess [1]	0.228 ± 0.013	0.231 ± 0.015	0.343 ± 0.002
SMOTE [4]	0.178 ± 0.005	0.206 ± 0.004	0.321 ± 0.002
FairBO (ours)	0.175 ± 0.007	0.196 ± 0.005	0.307 ± 0.001

- FERM, Zafar, and Adversarial are model-specific techniques for algorithmic fairness.
- FERM preprocessing and SMOTE are model-agnostic preprocessing on the data.
- FairBO emerges as a surprisingly competitive baseline that can outperform or compete against these highly specialized techniques.
- FairBO acts on the hyperparameters → it can be even used *on top* of model-specific fairness techniques, which come with their own hyperparameters.

References

[1] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *NeurIPS*, 2018.

[2] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *JMLR*, 2019.

[3] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *AIES*, 2018.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.