FairCR – an evaluation and recommendation system for fair classification algorithms

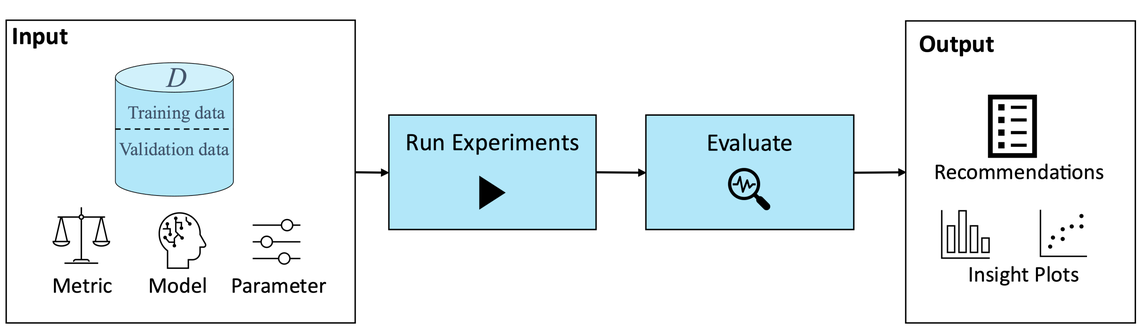
ICDE '24

Nico Lässig
Melanie Herschel
Ole Nies

MOTIVATION

***WHAT** is unfairness?* Biased predictions based on protected attributes (e.g. gender, ethnicity, religion, …).
***WHY** can classifiers be unfair?* Bias in training data is learned by classifiers.
***HOW** can we prevent bias?* Bias mitigation techniques.
***WHICH** bias mitigation techniques are most suited for the current problem?* **Use FairCR.**

SYSTEM OVERVIEW



FAIRNESS METRICS

- **Global (group) fairness:** Similar probability for outcome of protected and unprotected group (group fairness). Assessed over the whole dataset.
- **Individual fairness:** An individual is classified as similar individuals.
- **Local fairness:** Conceptual combination of global and individual fairness. Applies group fairness metrics within local regions.
- FairCR currently includes several metrics from these categories

MODELS

- Bias mitigation algorithms intervene at different phases of the processing pipeline
- FairCR currently includes:

pre-processing	in-processing	post-processing
Modify the training data	*Induce fairness during the learning process*	*Adjust the classifier or output*
Reweighing (2012)	PrejudiceRemover (2012)	RejectOptionClassification (2012)
LFR (2013)	FairnessConstraintModel (2017)	EqOddsPostprocessing (2015)
DisparateImpactRemover (2015)	DisparateMistreatment-Model (2017)	CalibratedEqOdds-Postprocessing (2017)
Fair-SMOTE (2021)	GerryFairClassifier (2018)	JiangNachum (2021)
LTDD (2022)	AdversarialDebiasing (2018)	FaX (2022)
	ExponentiatedGradient-Reduction (2018)	
	GridSearchReduction (2018)	
	MetaFairClassifier (2019)	
	FAGTB (2019)	
	FairGeneralizedLinear-Model (2022)	
	GradualCompatibility (2022)	

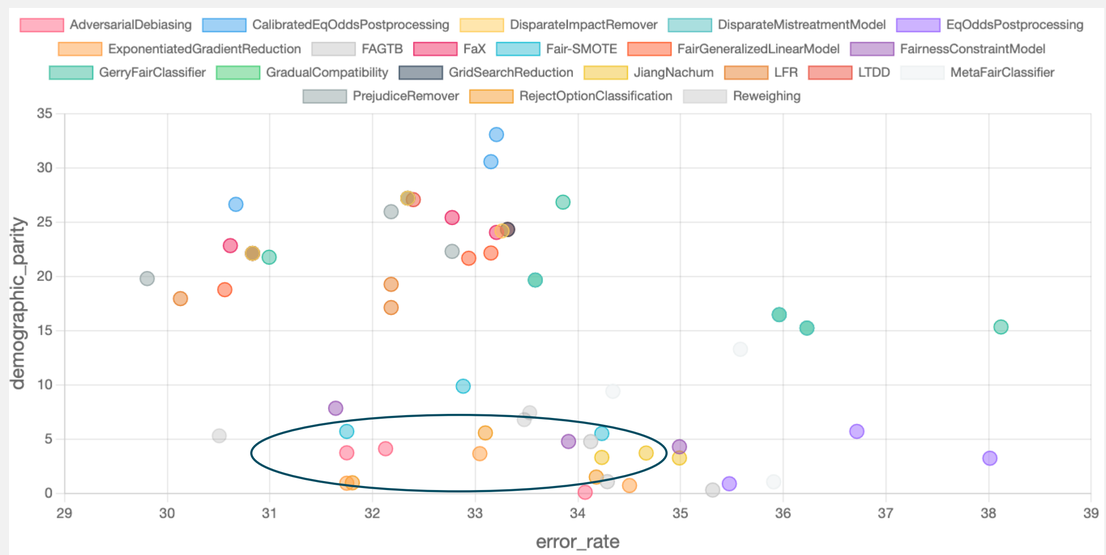
OTHER PARAMETERS

- Select (or upload) a **dataset**, specify the label and protected attributes
- **Lambda:** Weighs the error rate and bias
- **Local Lambda:** Weighs the global and local bias
- **Hyperparameter tuning:** If activated, grid search is applied
- Recommendation-specific parameters:
  - **Runtime-importance**
  - **Allowed memory usage**
- Size of the test data split; fixed seeds (if wanted); number of iterations

RUNNING EXAMPLE

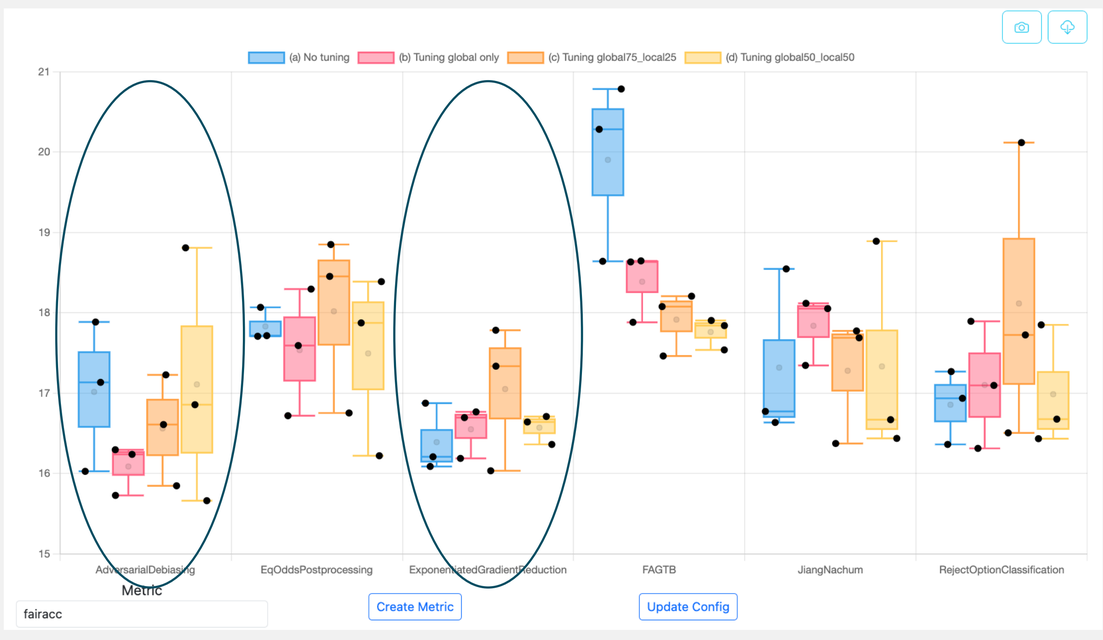
A user wants to determine the best models for the **compas** dataset. **Demographic parity** is the chosen metric. The user wants to evaluate the results with **hyperparameter tuning** turned **on.** The main goal is to achieve global fairness, while staying accurate, so **lambda** is set to **0.5**. Three iterations will be run.

EXPLORE SCATTERPLOTS



- RejectOptionClassification, AdversarialDebiasing, JiangNachum ExponentiatedGradientReduction, FAGTB & EqOddsPostprocessing are outperforming the rest in this setting.
- While local fairness is less important in this setting, the user wants to determine on how other tuning configurations affect the results.

EXPLORE BOXPLOTS



- The optimal algorithms and configurations are:
  - **AdversarialDebiasing** with lambda = 0.5 and local lambda = 0
  - **ExponentiatedGradientReduction** without hyperparameter tuning.

OTHER VISUALIZATIONS AND INTERACTIONS

- Different options to **visualize** the **dataset**
- Show **bar charts** to compare models and certain configurations
- Show **recommendation ranking**
- Return **counterfactuals**
- Start **new predictions**
- Create **new metrics**

IntCDC CLUSTER OF EXCELLENCE
DFG Deutsche Forschungsgemeinschaft
University of Stuttgart Germany
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS