FairX: A Comprehensive Benchmarking Tool for Model Analysis using Fairness, Utility and eXplainability

Md Fahim Sikder, Resmi Ramachandranpillai, Daniel de Leng, Fredrik Heintz

Reasoning and Learning (ReaL) Lab,
Artificial Intelligence and Integrated Computer Systems (AIICS),
Department of Computer and Information Science (IDA)

LINKÖPING UNIVERSITY

Agenda

- Introduction
- Background
- FairX
  - Overview
  - Explain different Modules
- Future Works
- Summary

Introduction

- Fairness benchmarking tool
  - FairX
    - Data Loading and Pre-processing
      - Tabular and Image dataset
    - Fair Models (Benchmarking)
      - Fair Generative Models Support
    - Evaluation
      - Synthetic Data Evaluation Support

Background

- When I say fairness, what do I mean by it?
- Three types of bias-mitigation techniques:
  - Pre-processing
  - In-processing
    - Most of the Fair Generative Models is here!
  - Post-processing
- Various fairness and utility metrics
  - It would benefit the community if everything is in one place!
- Need ways to evaluate synthetic data

Background (Contd.)

Table 1
Comparison of existing benchmarking tools with FairX over different key areas of interests: Fairness Evaluation; Synthetic Data Evaluation; Model Explainability; and Generative Fair Model Training.

Benchmarking Tools	Fairness Evaluation	Synthetic Data Evaluation	Explainability	Generative Model Training
Fairlearn [4]	✓	✗	✗	✗
AIF360 [5]	✓	✗	✓	✗
Jurity [14]	✓	✗	✗	✗
AEQUITAS [15]	✓	✗	✗	✗
REVISE [17]	✓	✗	✗	✗
FairBench [16]	✓	✗	✗	✗
FairX (ours)	✓	✓	✓	✓

FairX Overview

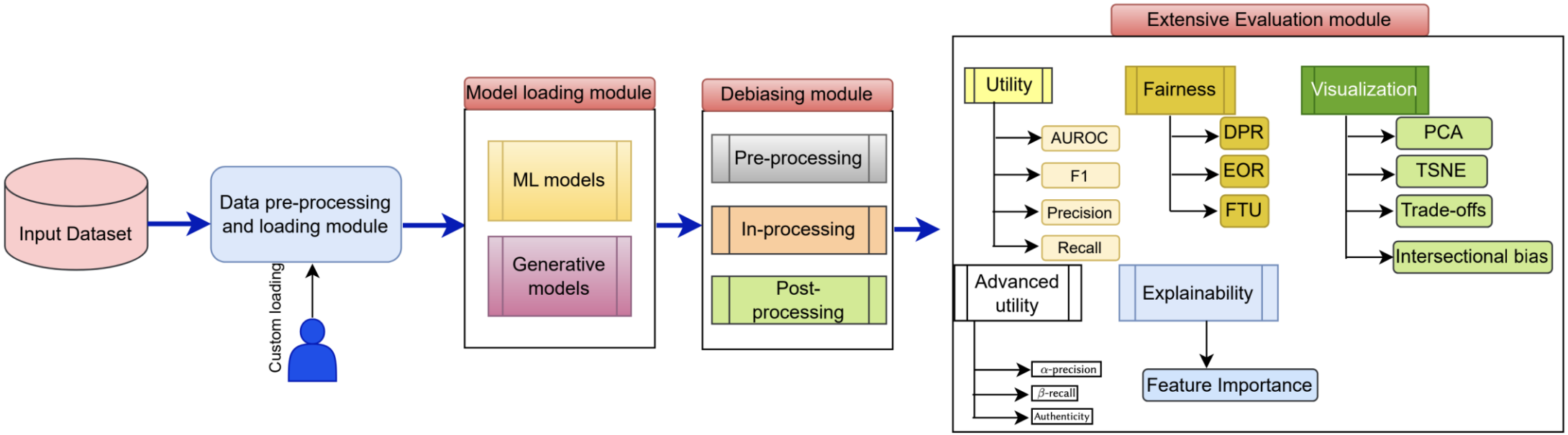

Fig: FairX Workflow

Data Loading Module

- Data pre-processing and Loading
  - Aids to prepare the dataset for the benchmarking as well as standalone use
    - Currently available dataset
      - 4 Tabular datasets including Adult Income, COMPAS, Credit-Card
      - 2 Image datasets including Colored-MNIST, CelebA [1]
    - Option to load custom dataset with the help of **"CustomDataClass"**

[1] Liu, Ziwei, et al. "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.

Model Loading Module

- Bias-Mitigating Models
  - Pre-processing
    - Disparate Impact Remover [1]
  - In-processing
    - FairDisco [2]
    - FLDGMs [3]
    - TabFairGAN [4]
    - Decaf [5]
  - Post-processing
    - On fairness & Calibration [6]

[1] Feldman, Michael, et al. "Certifying and removing disparate impact." *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.
[2] Liu, Ji, et al. "Fair representation learning: An alternative to mutual information." *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022.
[3] Ramachandranpillai, Resmi, Md Fahim Sikder, and Fredrik Heintz. "Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation." *ECAI 2023*. IOS Press, 2023. 1938-1945.
[4] Rajabi, Amirarsalan, and Ozlem Ozmen Garibay. "Tabfairgan: Fair tabular data generation with generative adversarial networks." Machine Learning and Knowledge Extraction 4.2 (2022): 488-501.
[5] Van Breugel, Boris, et al. "Decaf: Generating fair synthetic data using causally-aware generative networks." Advances in Neural Information Processing Systems 34 (2021): 22221-22233.
[6] Pleiss, Geoff, et al. "On fairness and calibration." *Advances in neural information processing systems* 30 (2017).

Evaluation Module

- Fairness Evaluation
  - Demographic Parity Ratio
  - Equal Opportunity Ratio
- Data Utility
  - ACC, Recall
  - F1-Score, AUC
- Advanced Data Utility
  - $\alpha - precision, \beta - recall$ [1]
  - Authenticity [1]

[1] Alaa, Ahmed, et al. "How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models." *International Conference on Machine Learning*. PMLR, 2022.

Evaluation Module (Contd.)

- Visual Evaluation
  - PCA & t-SNE
  - Performance trade-off between Fairness and Accuracy of the data
  - Synthetic Image Quality (only for image domain)
  - Intersectional Bias representation
  - Explainability Analysis
    - SHAP [1]

[1] https://shap.readthedocs.io/en/latest/index.html

Visual Evaluation

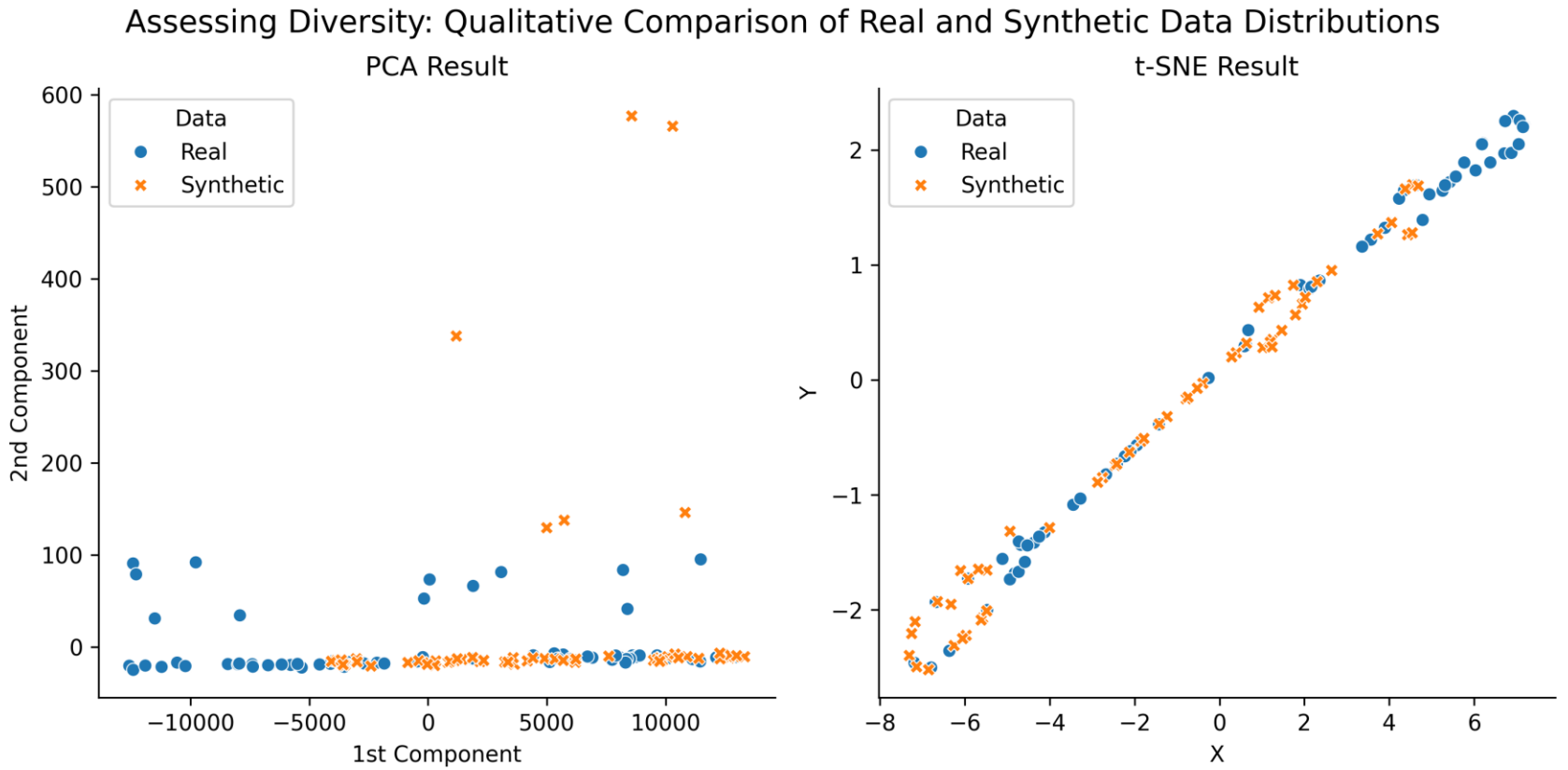

Fig: PCA and t-SNE plots of Original and Synthetic data by TabFairGAN

Visual Evaluation (Contd.)

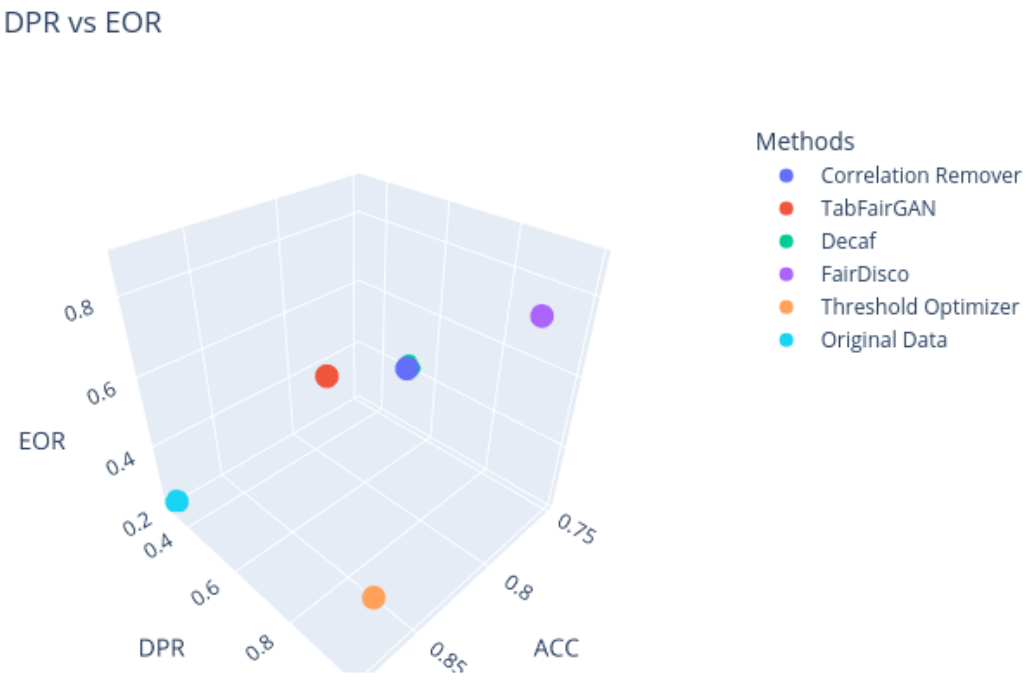

Fig: Model's Performance on Data Utility vs Fairness

Visual Evaluation (Contd.)

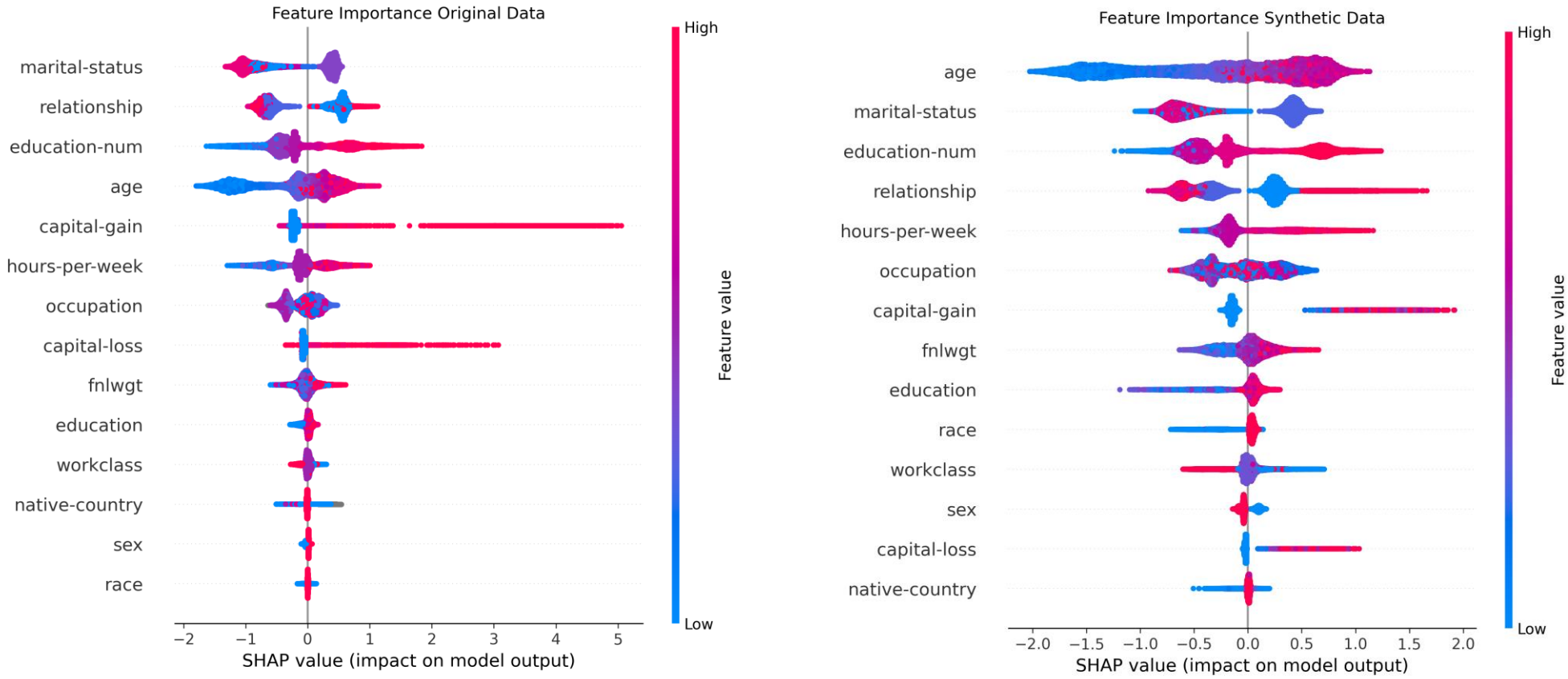

Fig: Feature importance by SHAP of Original (left) and Synthetic (right) data

Visual Evaluation (Contd.)

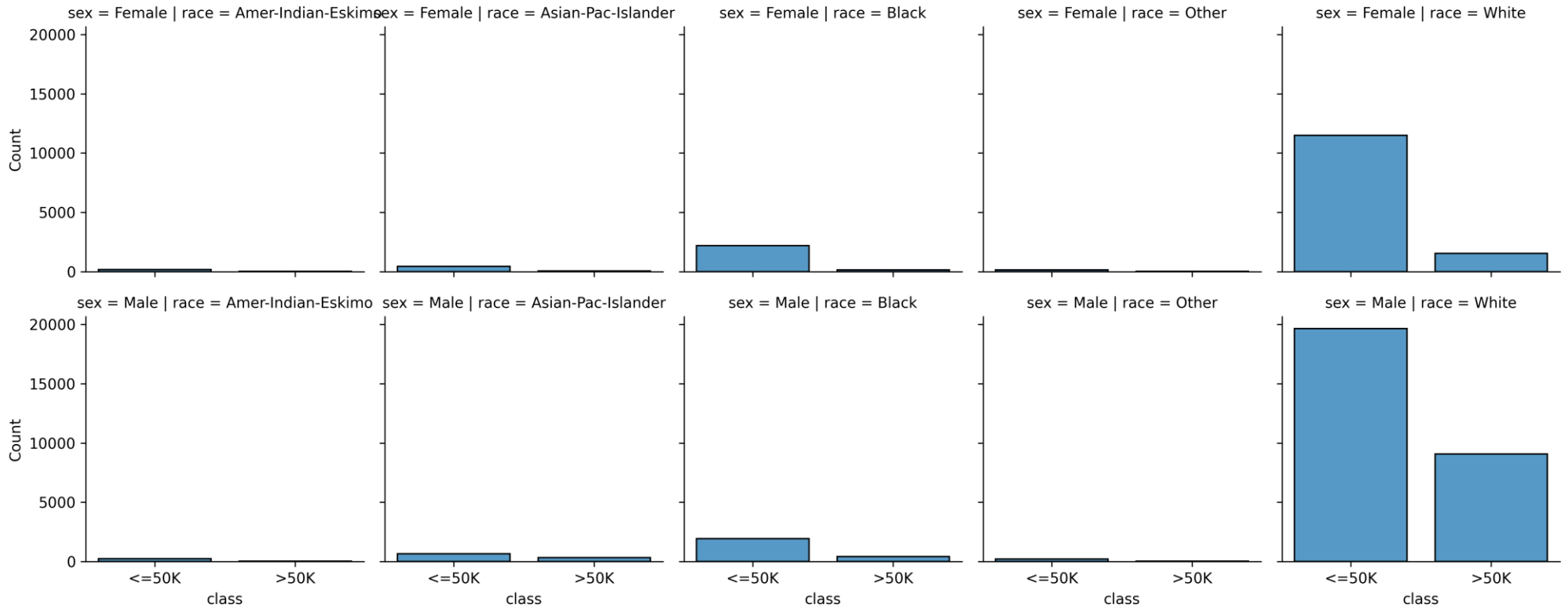

Fig: Intersectional Bias plotting, Representation of ‘sex’ and ‘race’ features on the target class, here we can see the dataset is heavily in favor of white people, dataset: Adult-Income

FairX (Dev-OPS)

- Open Source
  - https://github.com/fahim-sikder/FairX
- Python Support
  - 3.8 – 3.11
- Added CI/CD pipeline
  - Check modules after every commit

GitHub Link

Future Works

- Framework
  - Add new models
  - Add new evaluation
  - Add report support
  - Make a GUI interface (preferably web-based)
- Add support evaluating LLMs
  - Thinking different ways to evaluate the generated content regarding Fairness!
- Push the package to pypi

Summary

- We present FairX: a fairness benchmark with the support of synthetic data evaluation and interpretability.
- FairX is open-source and has capability to train fair generative models and evaluate fair synthetic data.
- Works for both Tabular and Image Modalities.
- We aim to extend the framework by adding more state-of-the-art fairness models as well as by supporting different data modalities.