



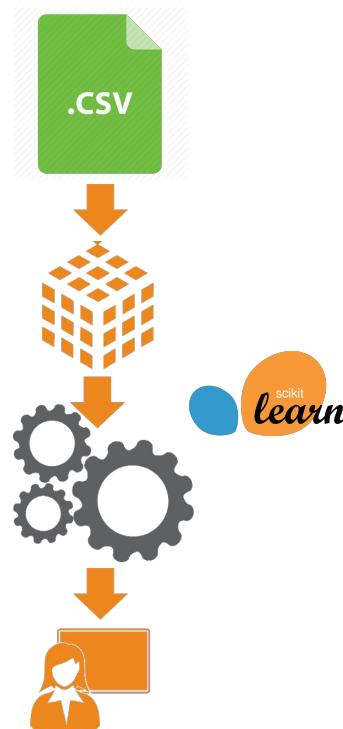
# Fairness Label for ML Pipeline

Team 4: Ke Yang, Tin Luu, and Hong Gong

12/05/19

# Motivation

- No system-level support for **end-to-end responsible data analysis** for users' own dataset.
- No system-level support for evaluating the **fitness of fairness measures and interventions** for users' own dataset.
- No system-level support for evaluating **intersectional effects** caused by interaction among multiple sensitive attributes.





## Problem Statement - Goals

- Build a system to support standard operations in a ML pipeline
  - earlier life stages of data → *scikit-learn preprocessing*
  - ML models → *scikit-learn classification* and *model selection*
  - fairness interventions → *AI Fairness 360*
- Design a component to track important fairness-related statistics → **Fairness Labels**
- Visualize high-dimensional fairness labels



# Approach - Fairness Labels

**Static label:** the distributions of *target variable* (+ or -) of all the sub-populations defined by a pair of sensitive attributes, like gender and race.

**Performance label:** the distributions of *performance metric* (TPR, FPR, TNR, FNR, ACC, ..., etc) of all the sub-populations defined by a pair of sensitive attributes.

Adult Income: +: >50K, -: <=50K

*race*

$\{(white) : \{+ : 0.95, - : 0.88\},$

$(black) : \{+ : 0.05, - : 0.11\}\}$

*race, gender*

$\{(white, men) : \{+ : 0.86, - : 0.63\},$

$(white, women) : \{+ : 0.14, - : 0.37\},$

$(Black, men) : \{+ : 0.77, - : 0.46\},$

$(Black, women) : \{+ : 0.23, - : 0.54\}\}$

# Implementation

- Pipeline
  - Integrated most of **scikit-learn preprocessing functionality**
  - Integrated fairness interventions in **AI Fairness 360**
  - Developed components: **Filter, Balance Sampler, Balance Splitter**
- Fairness labels
  - Extract static and performance label
  - Linear combination in the order of sensitive attributes
    - white, women, young, ...
- Visualization
  - Seaborn barplots
  - Tree-structure representation for high-dimensions \*

## Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ... — Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug prices, stock prices.

**Algorithms:** SVM, Linear regression, Lasso, ... — Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, ... — Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization. — Example

## Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics. — Example

## Preprocessing

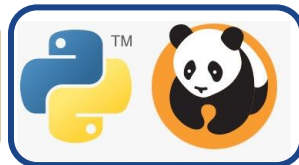
Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction. — Examples

IBM Research Trusted AI

AI Fairness 360 Open Source Toolkit





# Experimental Setup

- **Robustness**
  - various input pipelines
  - new dataset
- **Intersectional effects**
  - use 2 sensitive attributes
  - evaluate existing fairness interventions
  - on benchmark datasets and new dataset

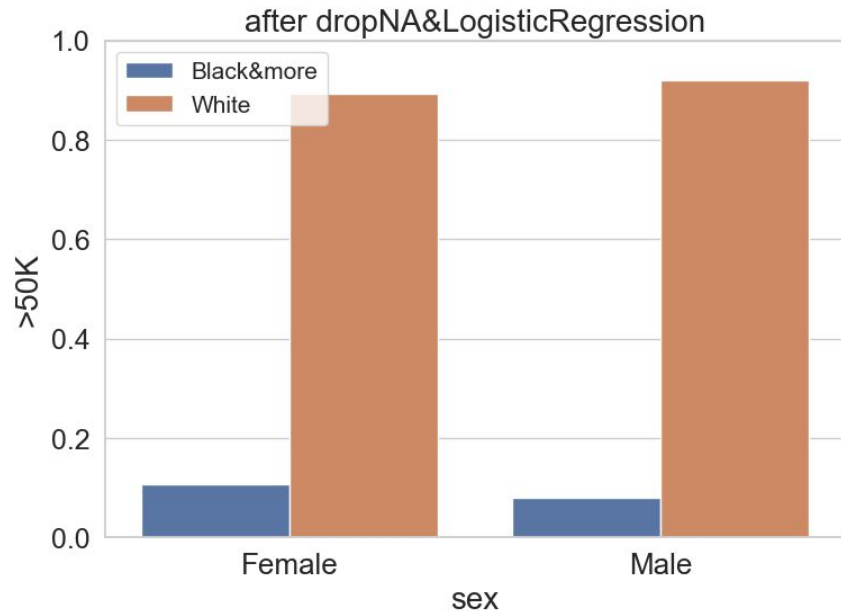
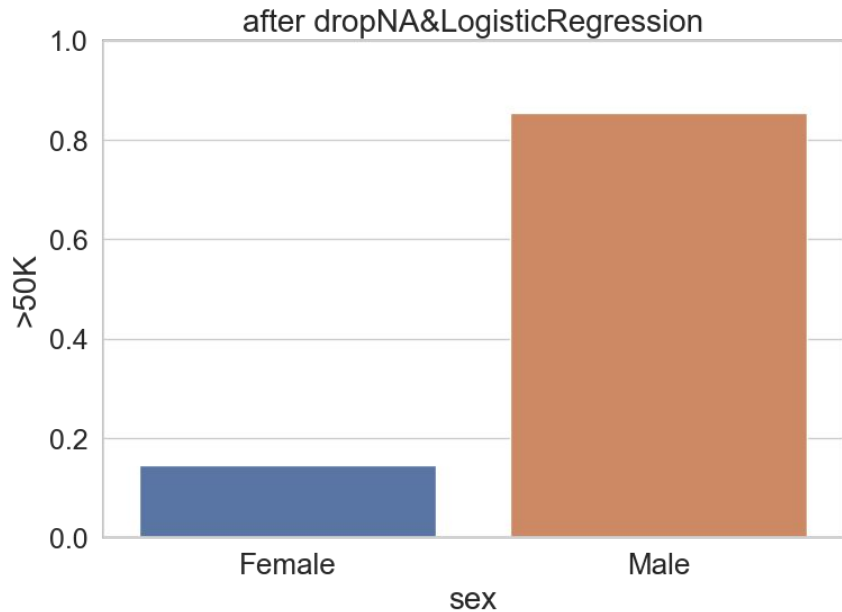


# Datasets

- Benchmark datasets:
  - **Adult Income:** [the fitness of fairness preprocessing interventions](#)
  - **German Credit:** [intersectional effects](#)
  - **COMPAS Score:** [intersectional effects](#)
- Explore new dataset:
  - **Law School Admissions:** [robustness](#)

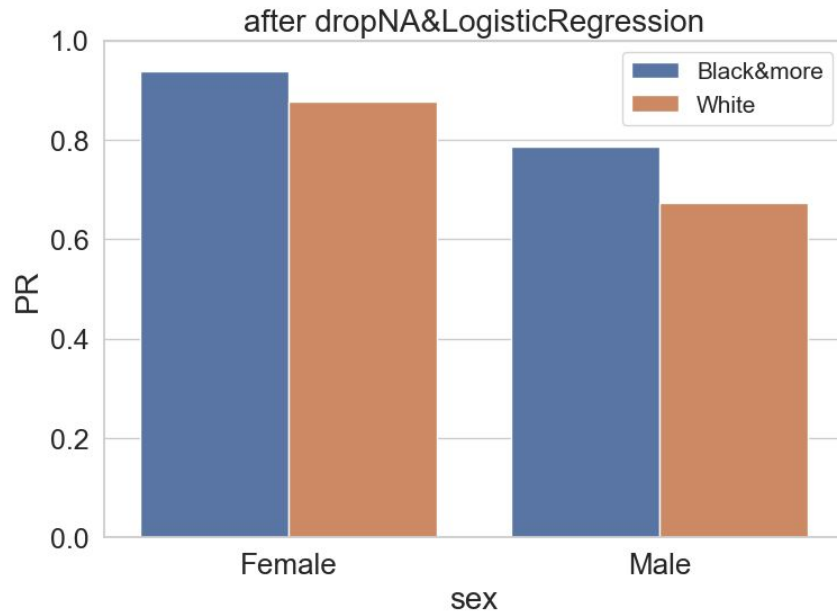
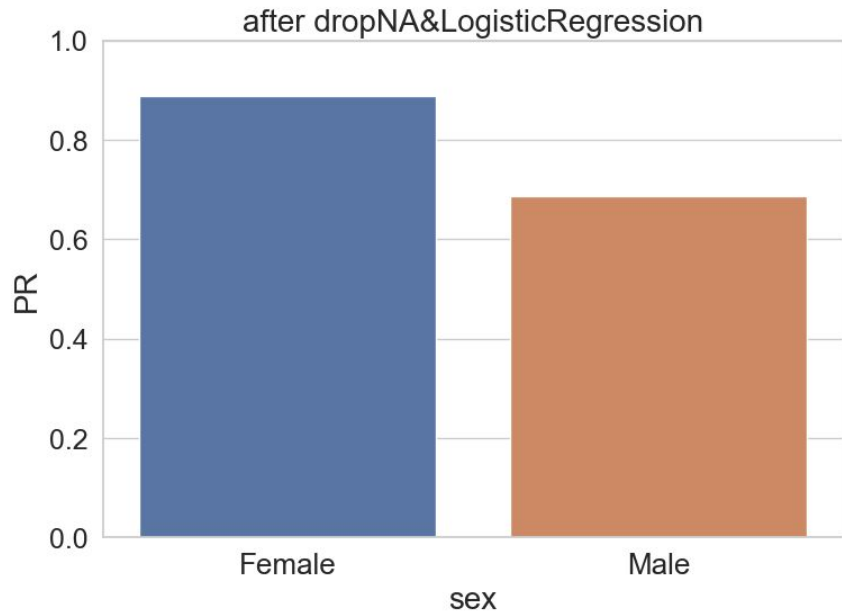
dataset	size	target	1st sensitive att	2nd sensitive att	used in other papers
Adult Income	32,561	income per year	race	gender	Yes
German Credit	1000	credit worthiness	binned age	gender	Yes
COMPAS Score	7,214	whether recidivated	race	gender	Yes
Law School Admissions	438,487	application status	race	gender	No

# Adult Income - static label - positive outcome

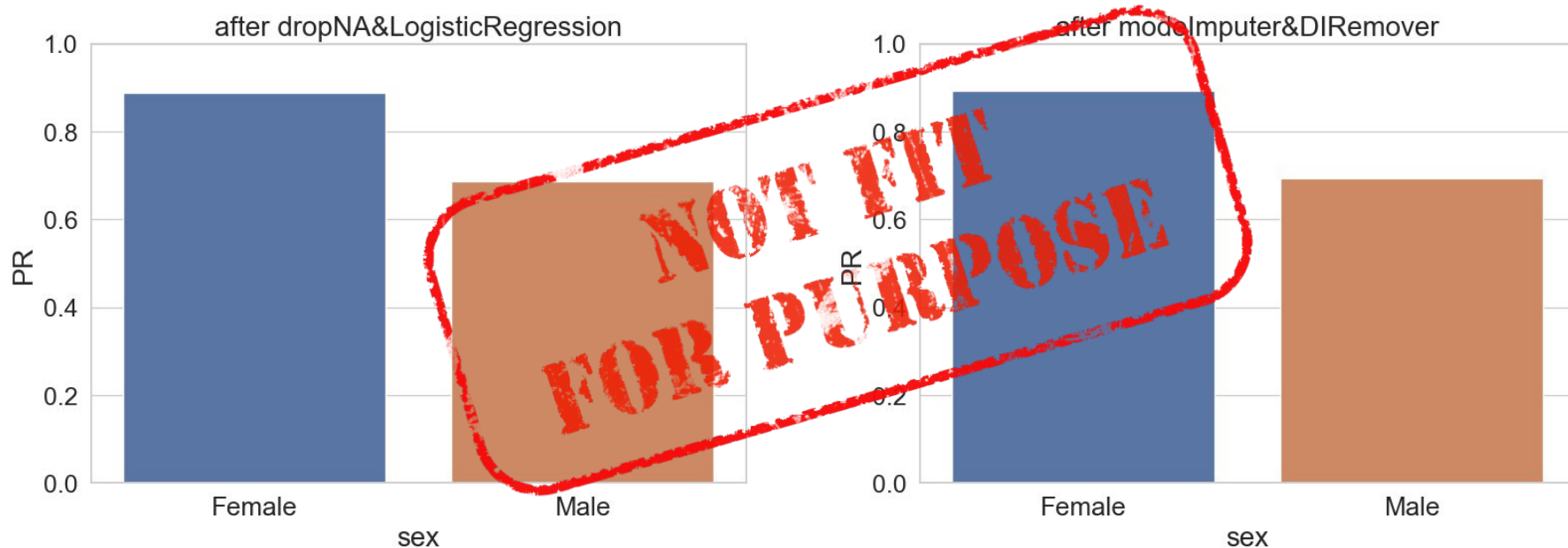




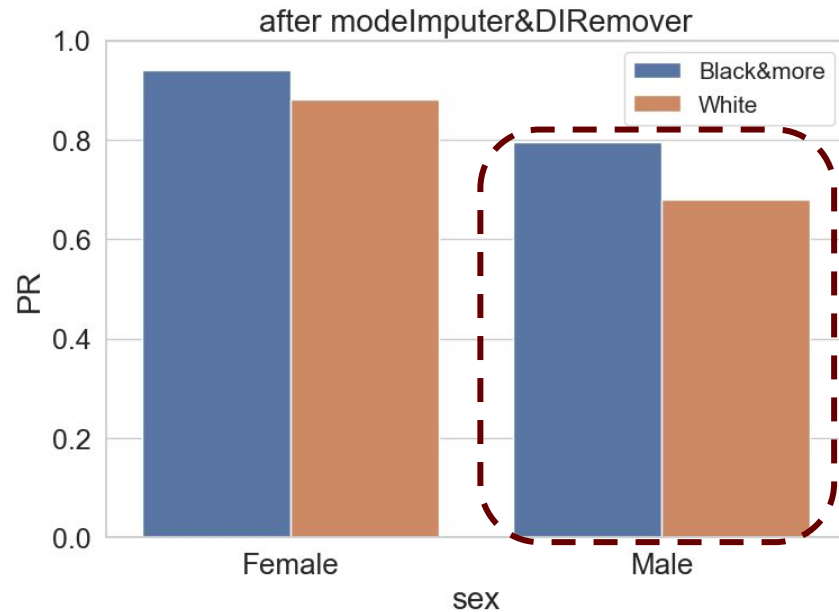
# Adult Income - performance label - selection rate



# Change of performance apply fair intervention



# Intersectional effects - selection rate





## Discussion

- ★ **Build a general system that supports responsibly ML pipeline**
  - \*Integrate more modules
  - \*Support flexible pipelines
- ★ **Track two types of fairness-related statistics through our system**
  - \*High-dimensional fairness label
- ★ **Experiments on benchmark datasets and new dataset**
  - \*Explore the unknown intersectional effects



**Thank you & QAs**