Fair Classification with Adversarial Debiasing

Advanced Machine Learning

Academic Year 2019/2020

Can algorithms take decisions ?

Should algorithms take decisions ?

Should algorithms take decisions
in sensible domains?

PROJECT OVERVIEW

THE RESEARCH QUESTION

Is an accurate classifier also fair?

Is it possible to develop a fair classifier?

PROJECT OVERVIEW

THE WORKFLOW

Data
Adult Census Income Dataset

↓

Explanatory Data Analysis

↓

Preprocessing

PROJECT OVERVIEW

THE WORKFLOW

Data
Adult Census Income Dataset

↓

Explanatory Data Analysis

↓

Preprocessing

↓

Baseline classifier

PROJECT OVERVIEW

THE WORKFLOW

Data
Adult Census Income Dataset

↓

Explanatory Data Analysis

↓

Preprocessing

↓

Baseline classifier

↓

Fairness Metrics

PROJECT OVERVIEW

THE WORKFLOW

Data
Adult Census Income Dataset

↓

Explanatory Data Analysis

↓

Preprocessing

↓

Baseline classifier

↓

Fairness Metrics

↓

Evaluation

PROJECT OVERVIEW

THE WORKFLOW

Data
Adult Census Income Dataset

↓

Explanatory Data Analysis

↓

Preprocessing

↓

Baseline classifier → **Fair classifier**

↓

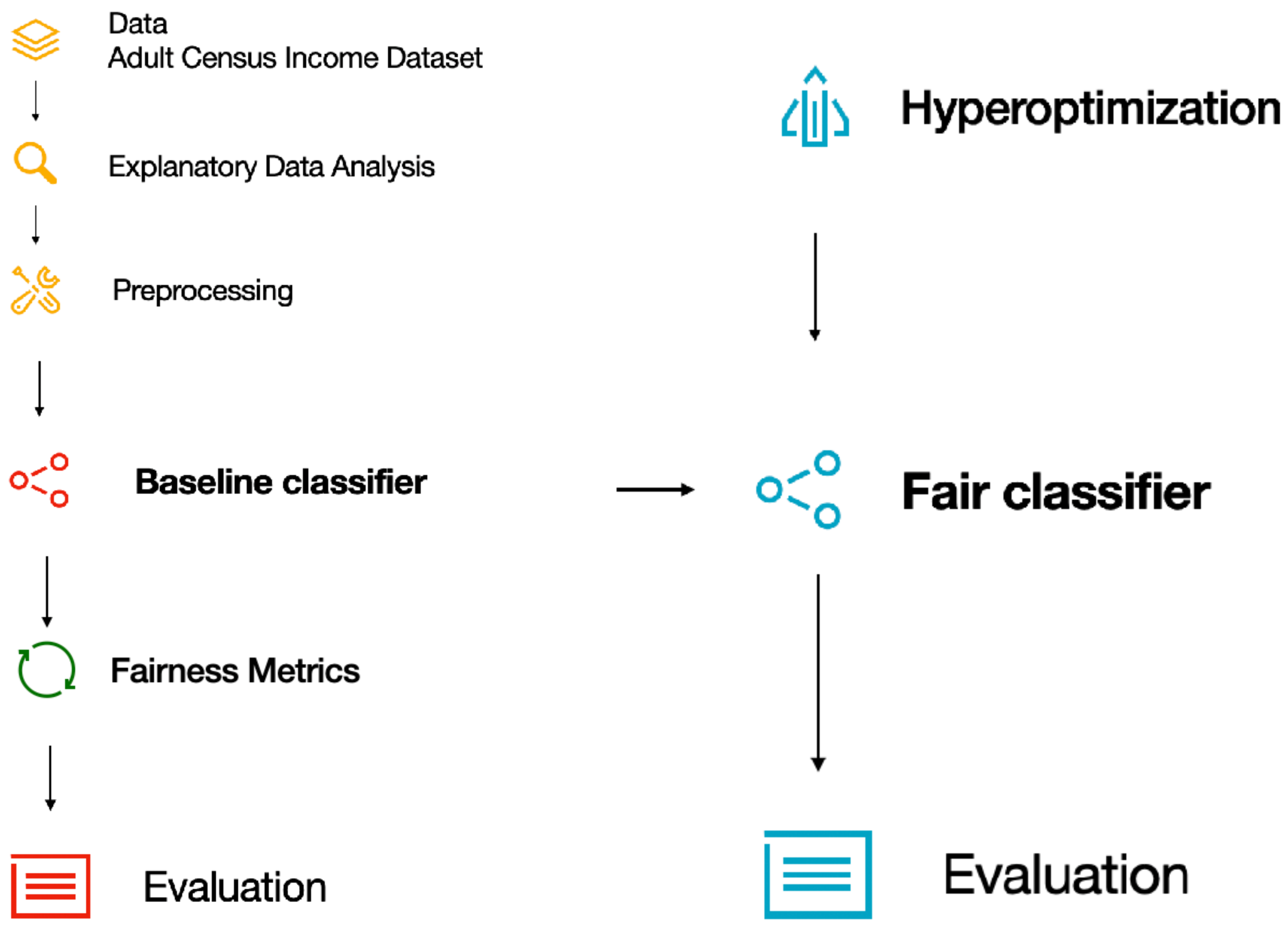
Fairness Metrics

↓

Evaluation

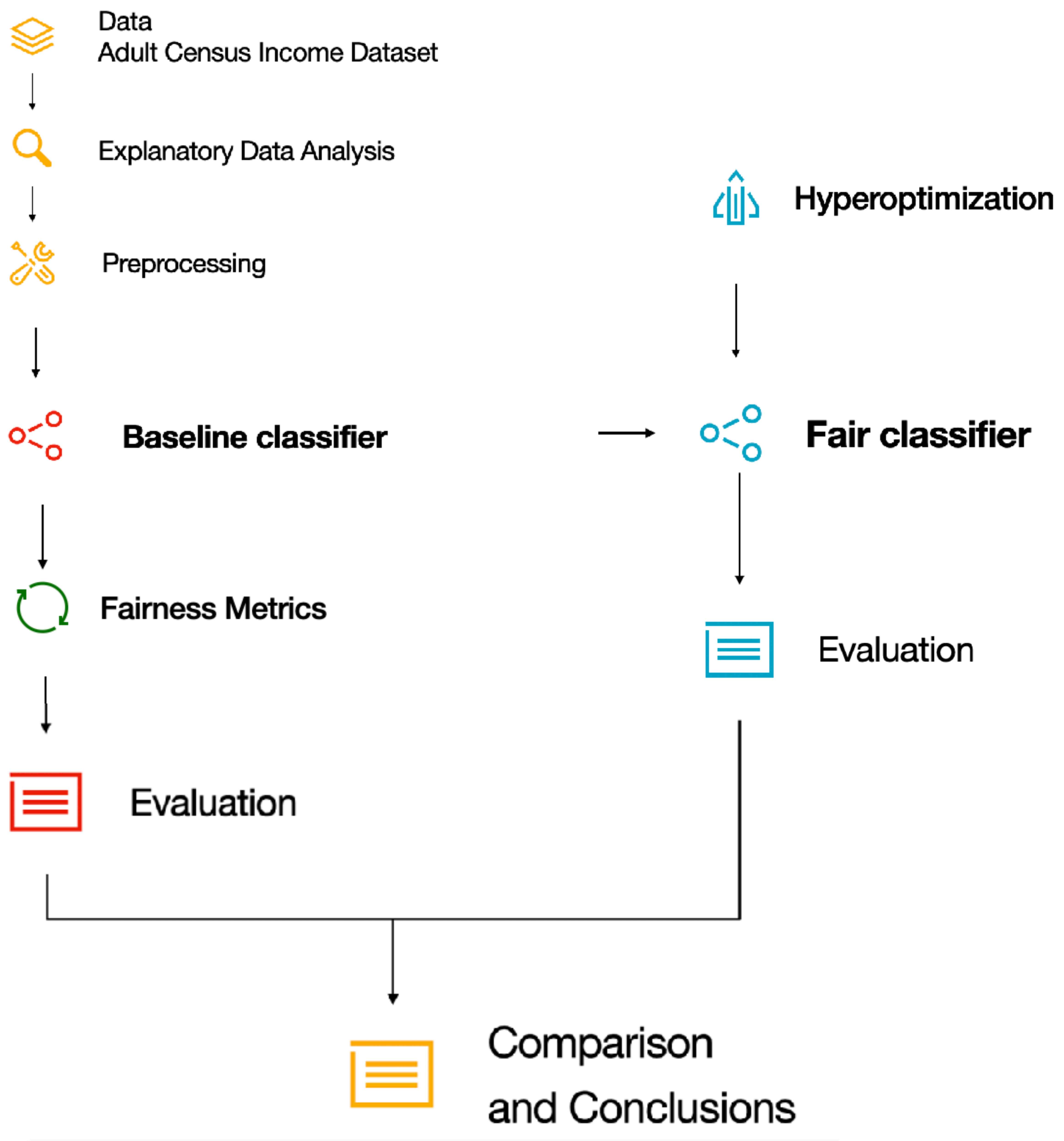
PROJECT OVERVIEW

THE WORKFLOW



PROJECT OVERVIEW

THE WORKFLOW



DATA:
Adult Census Income Dataset

RAW DATA

- 48842 records with 14 attributes regarding social, demographic and financial aspects of U.S. citizens.
- 2 sensitive attributes: Race (Black, White), Sex (Male, Female)
- 1 binary class variable: Income (1 if $Income \geq 50K$, 0 otherwise)

CLEANING

- Records with race different than black or white were discarded from the analysis.
- Few missing data (0.071%) removed

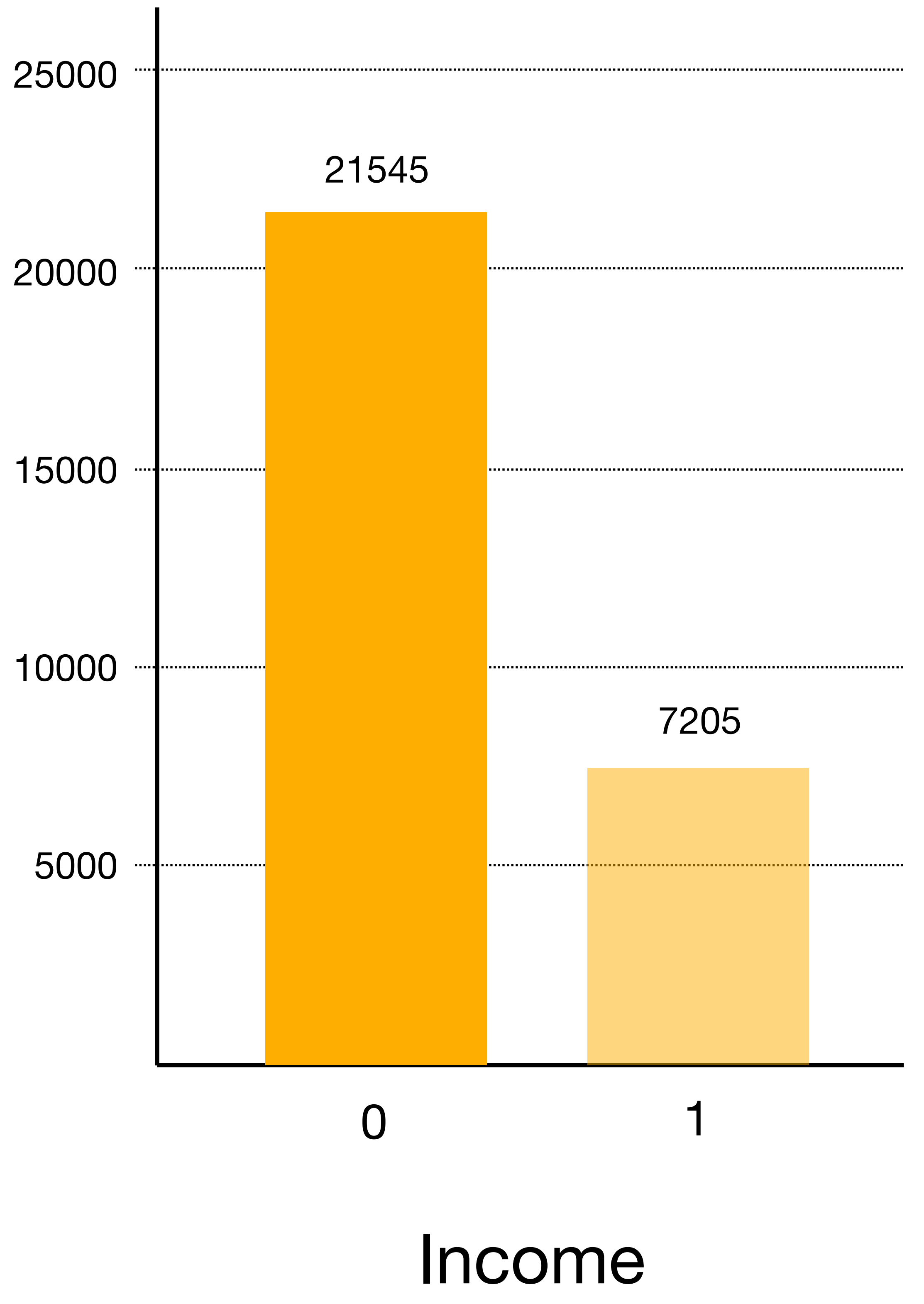
FINAL DATASETS

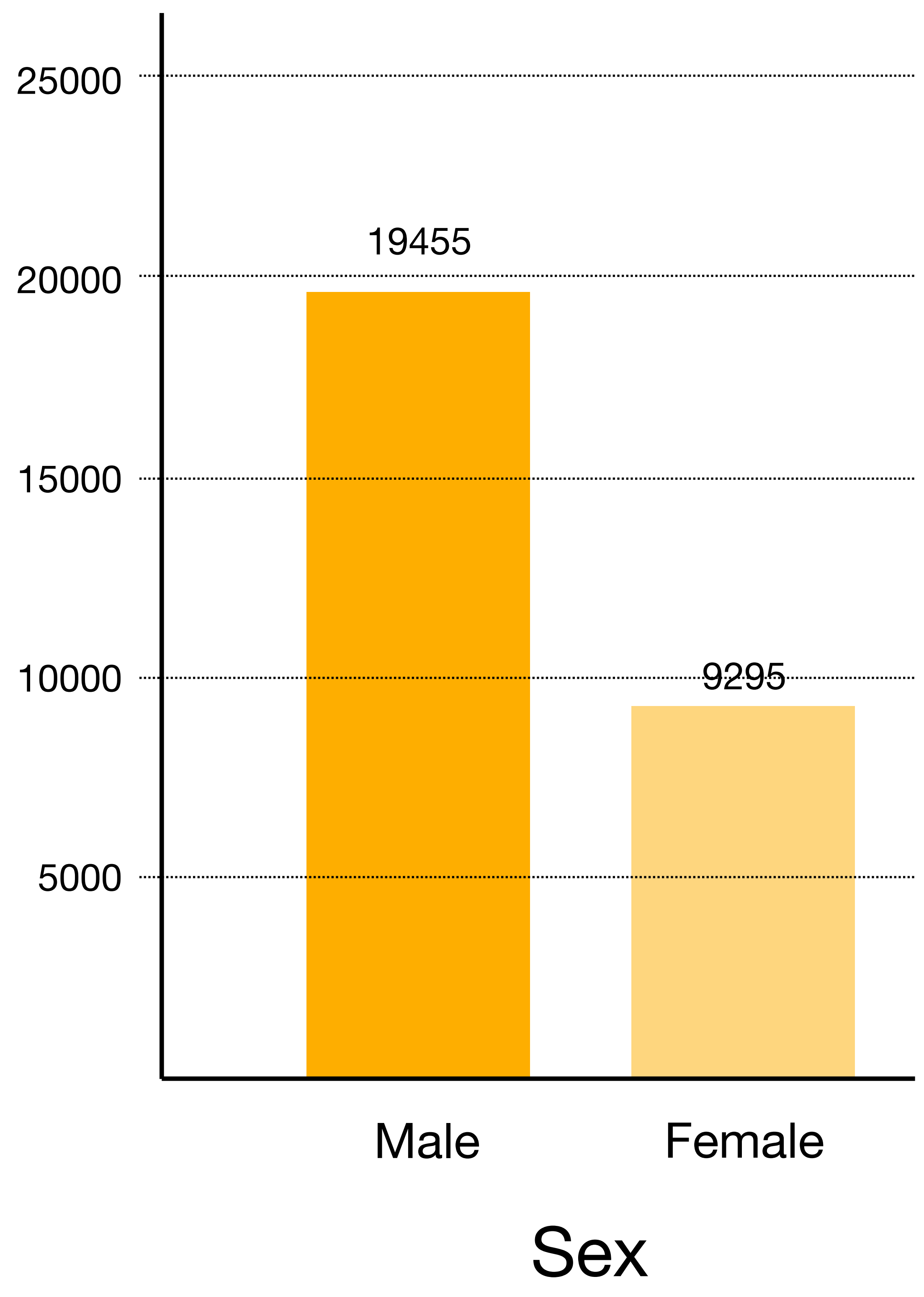
- 28750 records
- X: 12 features
- Z: 2 sensible attributes
- y: 1 class variable

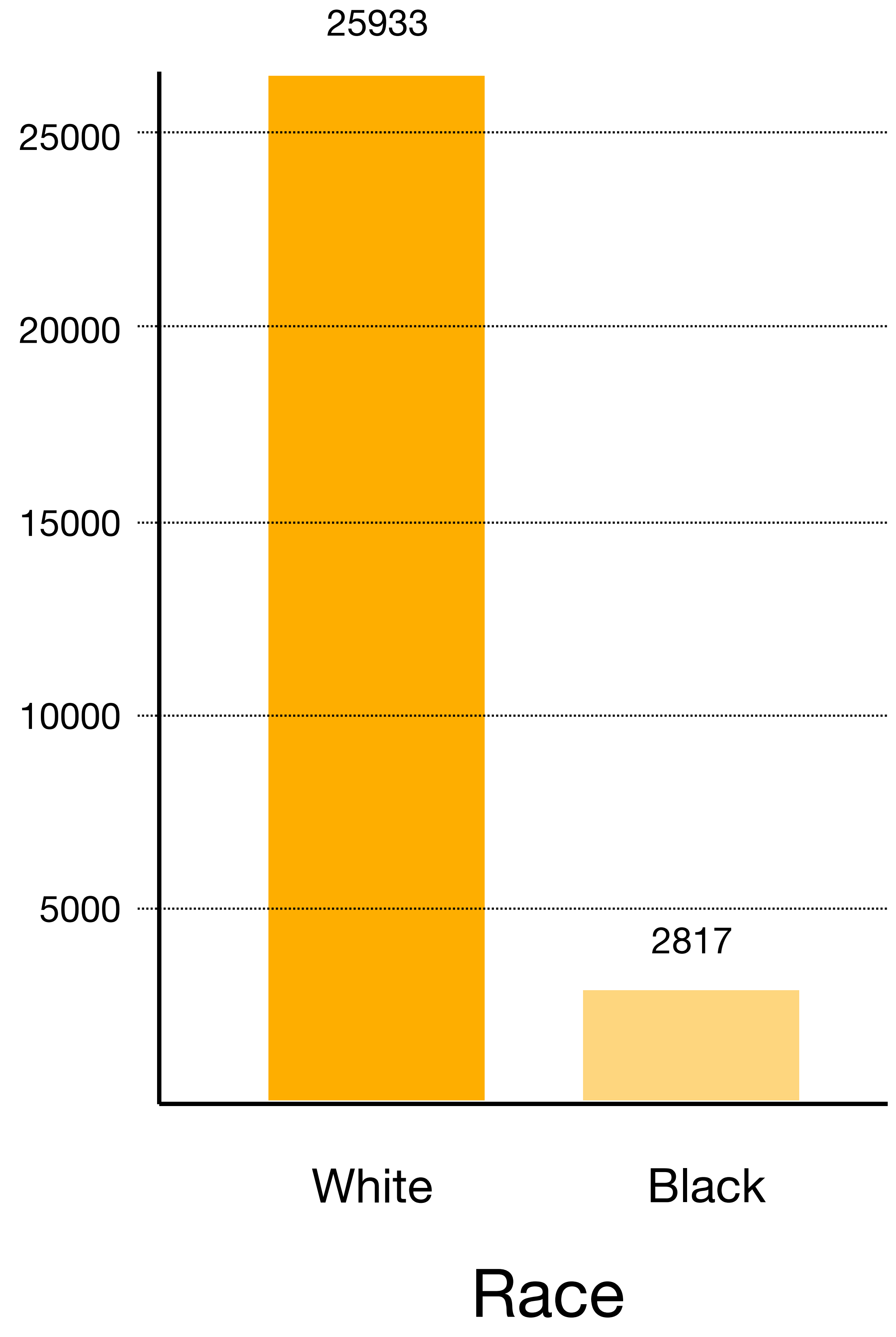
DATA:

Adult Census Income Dataset

Explanatory Data Analysis



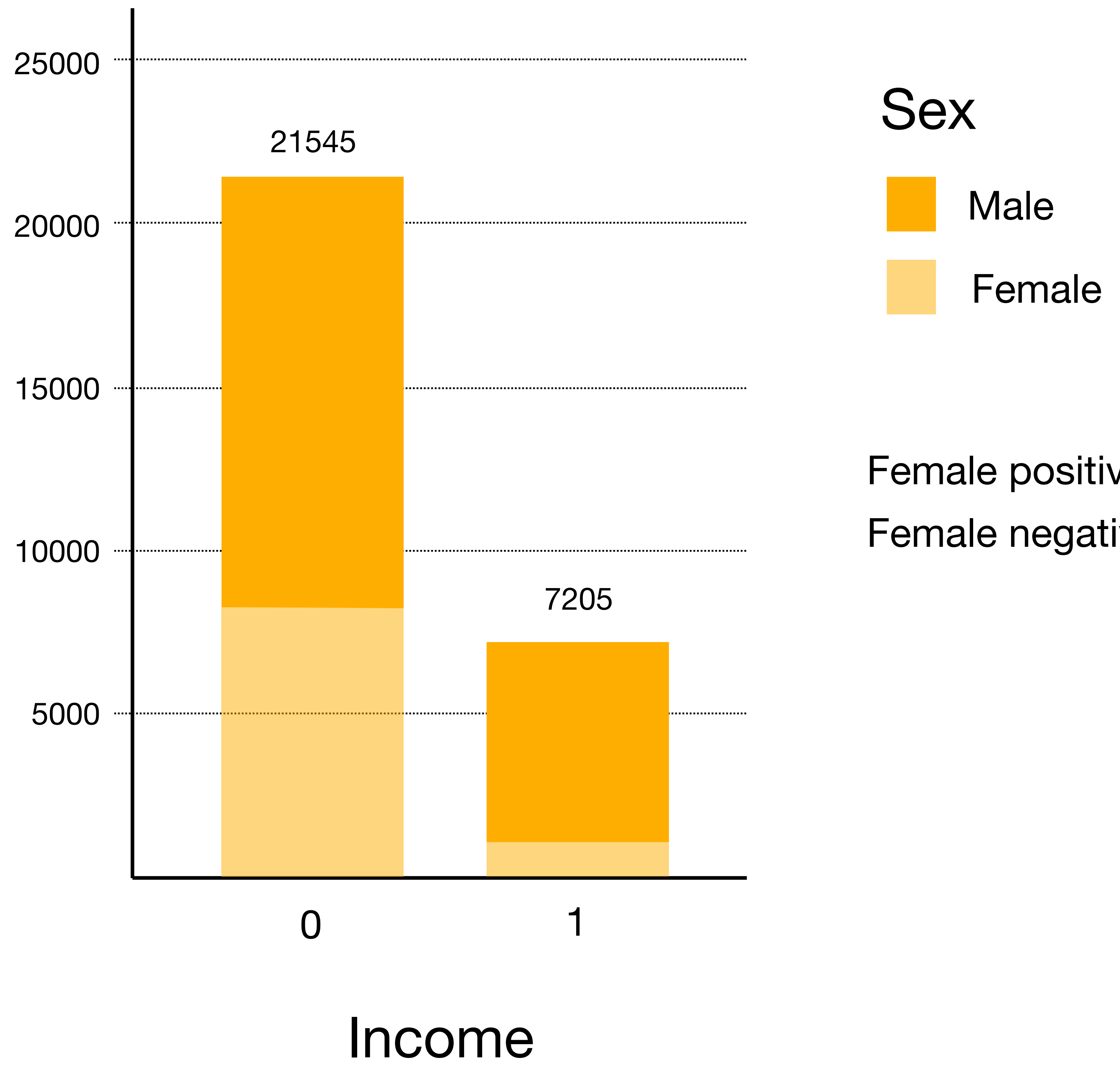




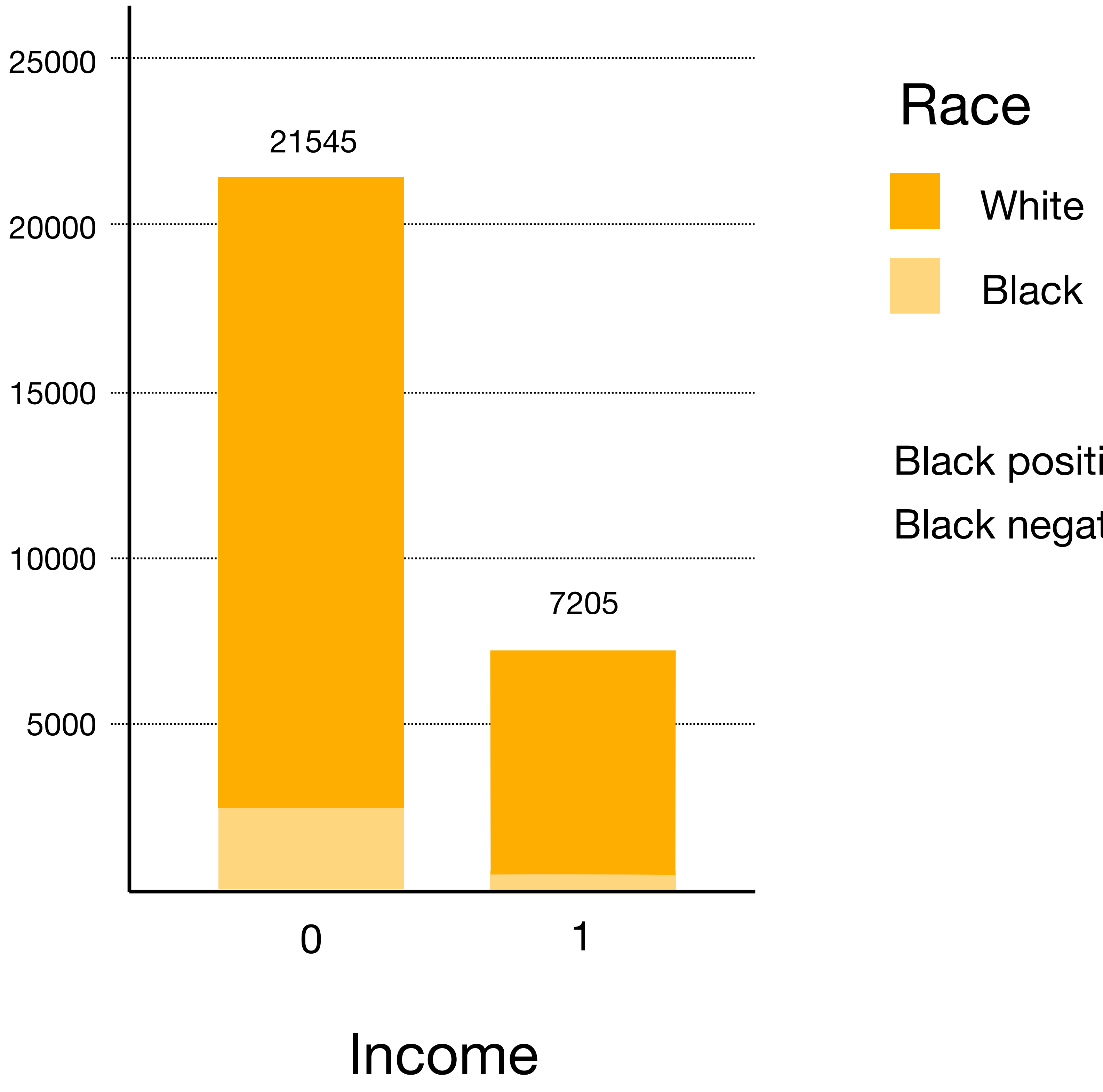
DATA:

Adult Census Income Dataset

Explanatory Data Analysis



Female positive: 14.66%
Female negative: 38.24%



Black positive: 5.08%
Black negative: 11.38%

DATA:

Adult Census Income Dataset

Explanatory Data Analysis

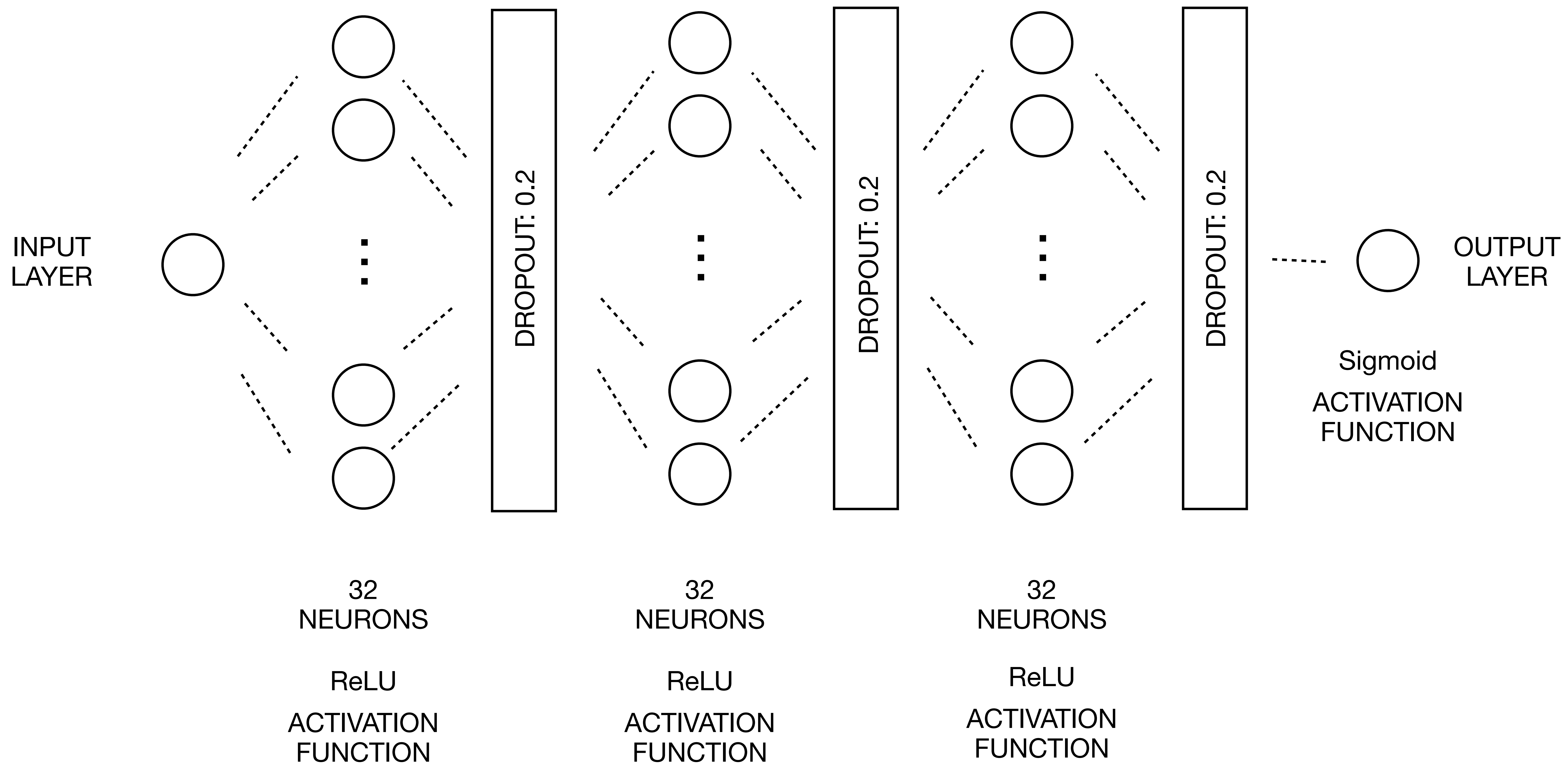
Preprocessing

- *“education”* attribute removed because redundant
- Less represented countries grouped into *“others”*
- Standardization
- Training-Vaidation-Test split:
    - Training set: 16100 samples
    - Validation set: 4025 samples
    - Test set: 8625 samples

CORRELATION MATRIX
between the sensitive attributes and training features

BASELINE CLASSIFIER

ARCHITECTURE



HYPERPARAMETERS

- LOSS FUNCTION: *Binary Crossentropy*
- OPTIMIZER: *Adam* (LR, B_1, B_2: default)
- WEIGHTS: balanced
- EPOCHS: 50
- EARLY STOPPING: Validation loss,
  Patience:10

FAIRNESS METRICS

P-RULE

$$min\left(\frac{P(\hat{y}=1|z=1)}{P(\hat{y}=1|z=0)}, \frac{P(\hat{y}=1|z=0)}{P(\hat{y}=1|z=1)}\right) \geq \frac{p}{100}$$

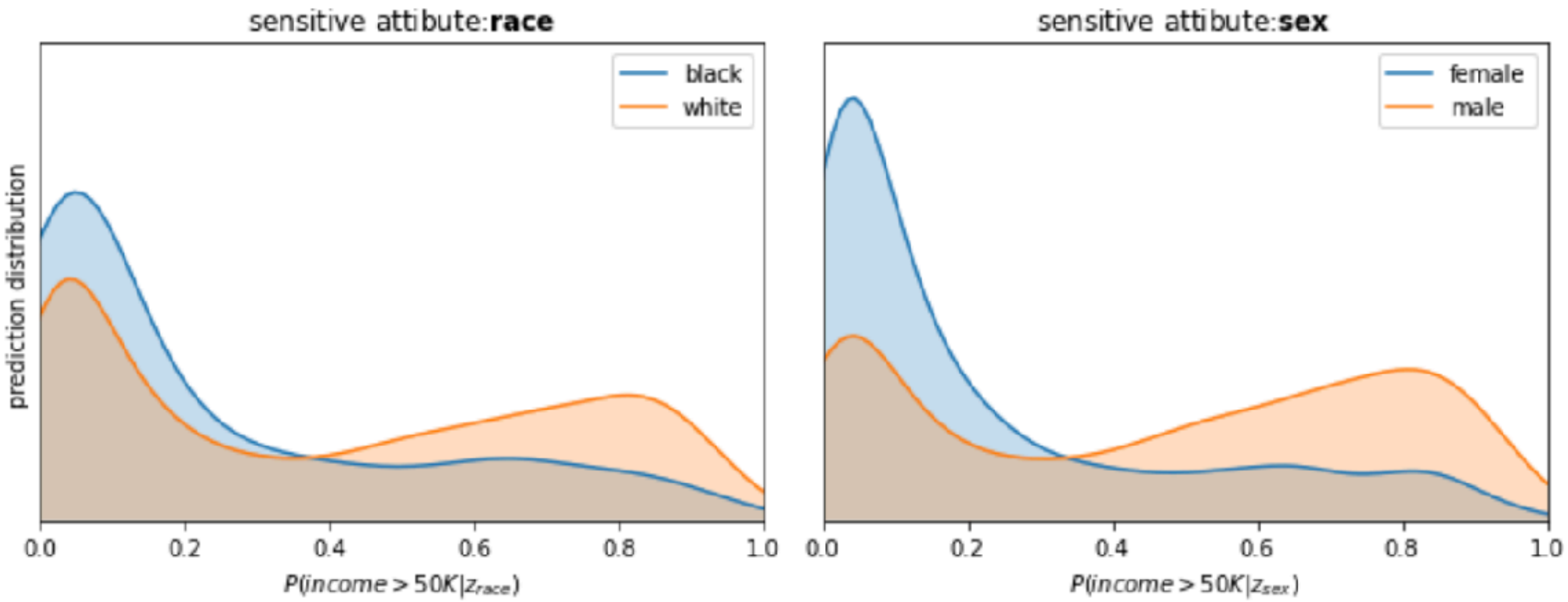
Pivotal Property

Good proxy for **disparate impact**

Threshold for fairness: $p \geq 80$

EVALUATIONS

BASELINE CLASSIFIER



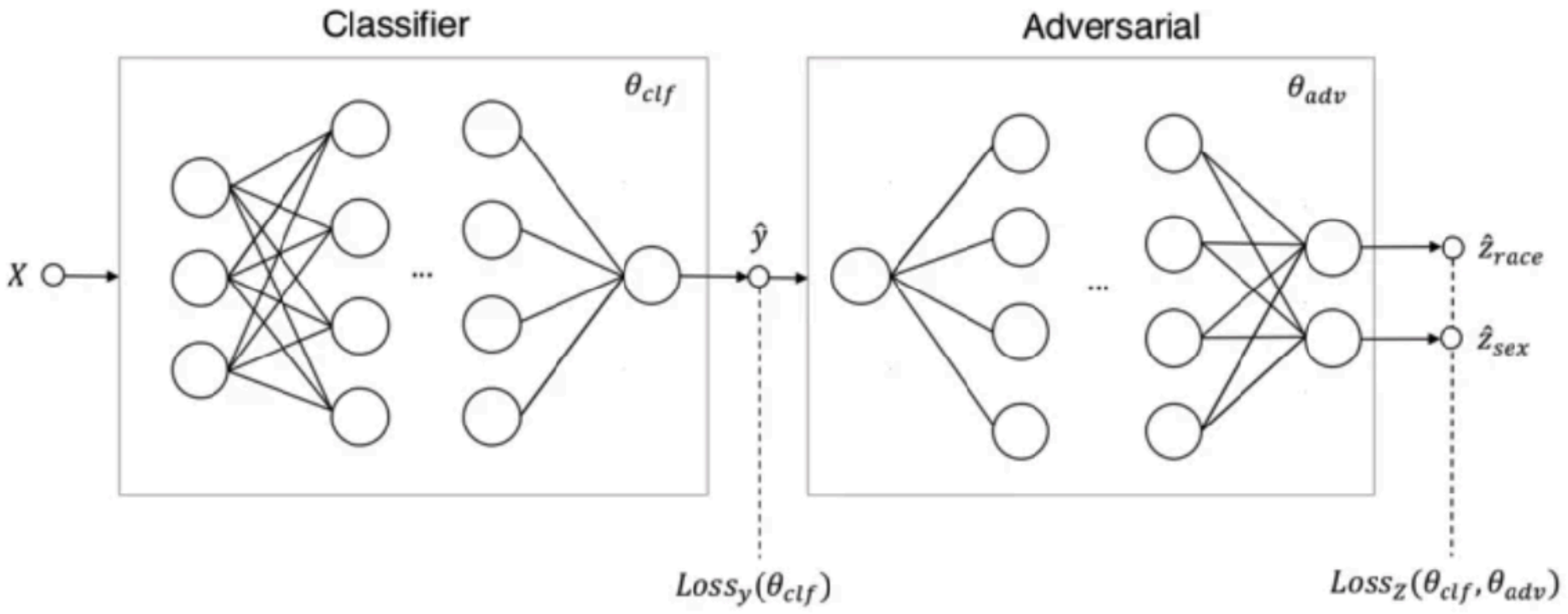
Training iteration #0

- ROC AUC: 0.89
- Accuracy: 0.7743

- race: 48%-rule
- sex: 36%-rule

FAIR CLASSIFIER

ARCHITECTURE



OBJECTIVE

$$min_{\theta_{clf},\theta_{adv}}[Loss_y(\theta_{clf} - \lambda Loss_Z(\theta_{clf}, \theta_{adv})]$$

HYPERPARAMETERS

- LOSS FUNCTION: *Binary Crossentropy*
- OPTIMIZER: *Adam* (LR, B_1, B_2: default)
- WEIGHTS: balanced
- WEIGHTS Z: balanced
- EPOCHS: 300
- EARLY STOPPING: P-rule Sex: 80%
  P-rule Race: 80%

HYPEROPTIMIZATION

HYPERPARAMETERS

- λ_{race}
- λ_{sex}

SCENARIO

- Deterministic
- Run objective: quality
- Initial Random configurations: 5
- Further configurations: 45
- Acquisition Function: EI

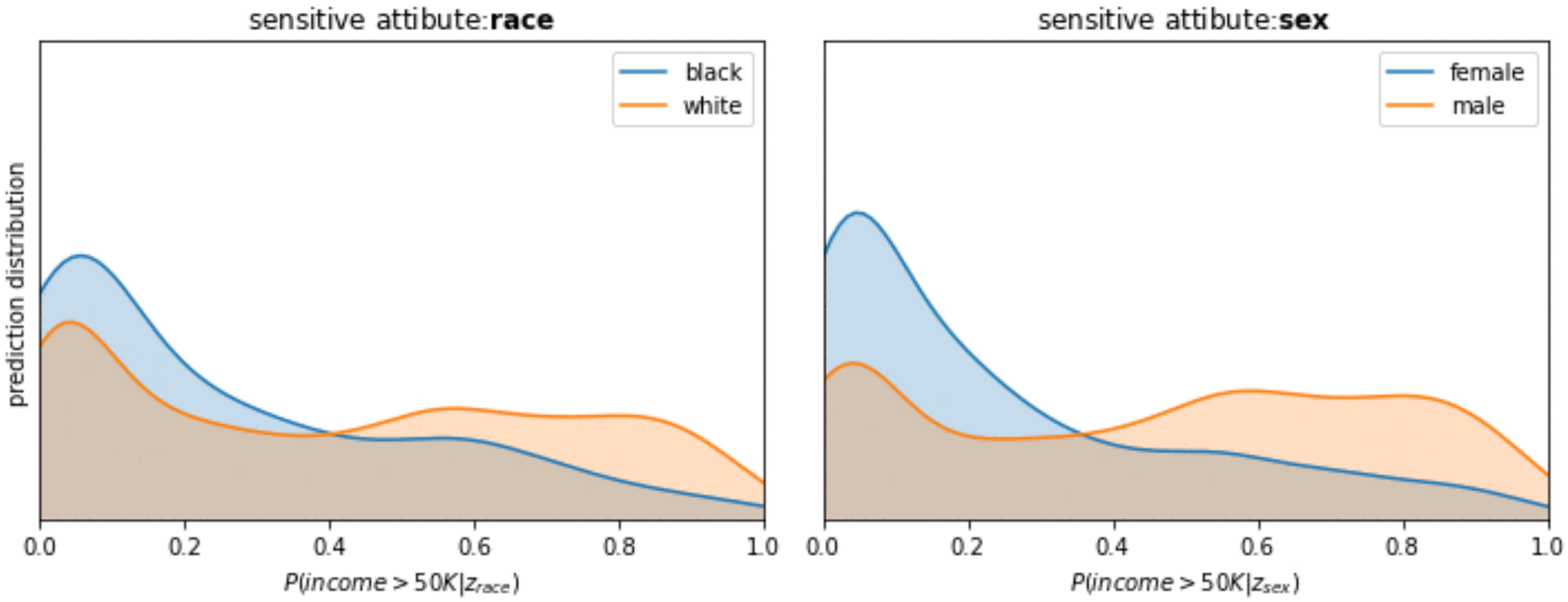
SURROGATE FUNCTION

$$\left(\frac{\Delta p - rule_0/|\Delta acc|}{100} + \frac{\Delta p - rule_1/|\Delta acc|}{100}\right)$$

SMAC

EVALUATIONS

FAIR CLASSIFIER



Training iteration #1

- ROC AUC: 0.91
- Accuracy: 0.7933

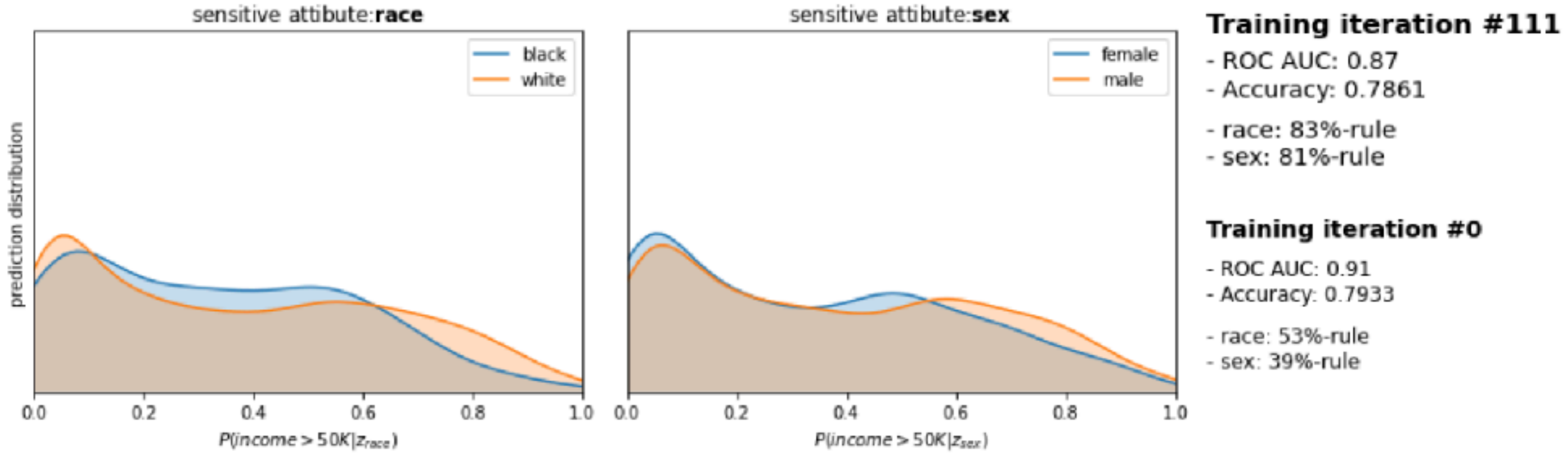
- race: 53%-rule
- sex: 39%-rule

Training iteration #0

- ROC AUC: 0.91
- Accuracy: 0.7933

- race: 53%-rule
- sex: 39%-rule

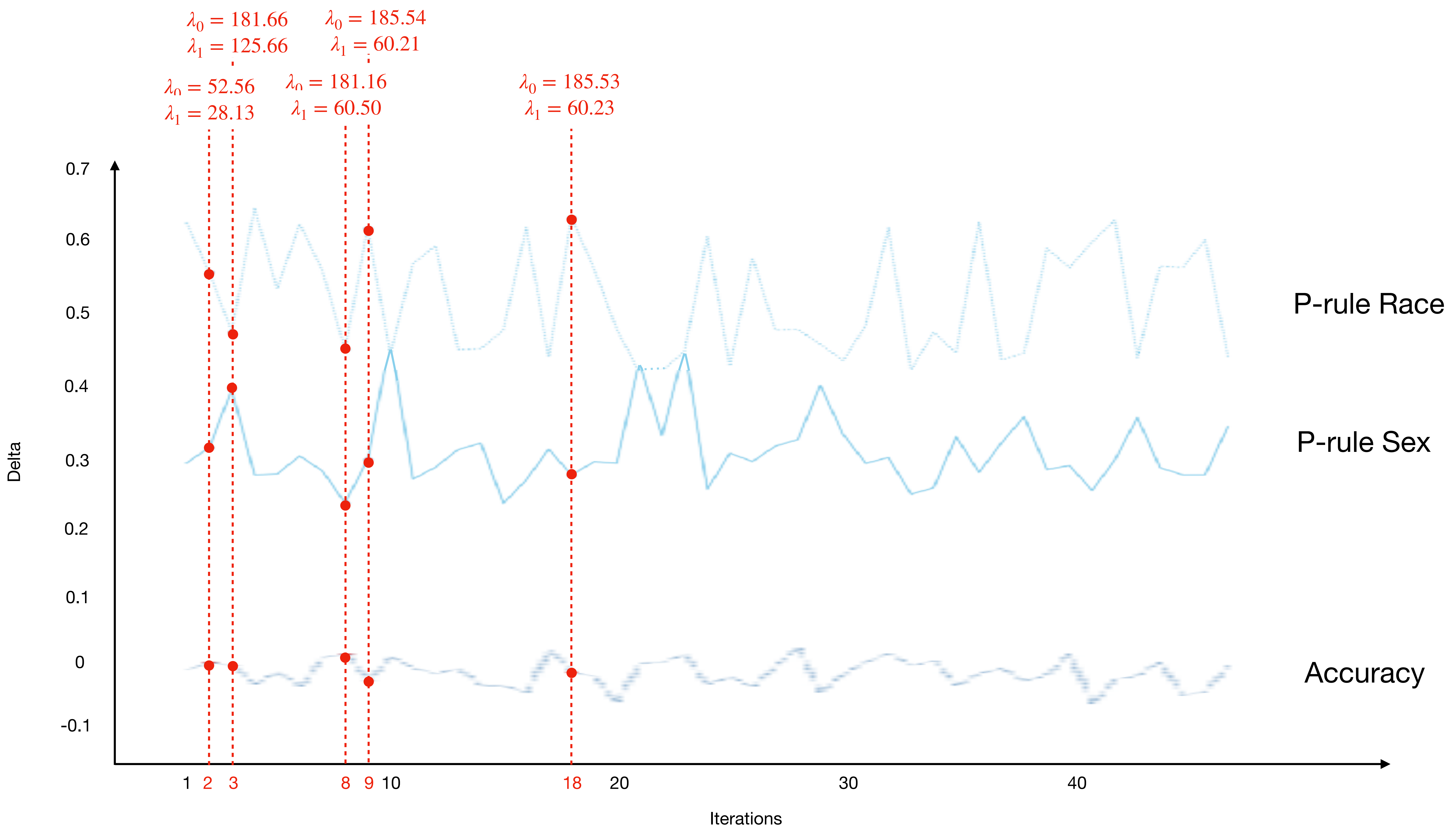
COMPARISON



		BASELINE	FAIR
Classification Metrics	Accuracy	0,793	0,77
	AUC	0,91	0,87
	F-1	0,67	0,64
Validation Metrics	P-rule Sex	53%	83%
	P-rule Race	39%	81%

EVALUATIONS

HYPEROPTIMIZATION



CONCLUSIONS

BASELINE CLASSIFIER

- Good classification performances
- Low fairness performances
- Indirect discrimination

FAiR CLASSIFIER

- Good fairness performances
- Low accuracy waste
- Disparate treatment needed to avoid disparate impact

HYPEROPTIMIZATION

- Task-specific
- Unable to optimized different objectives simultaneously
- Good performances

NEXT STEPS

ADVERSARIAL NETWORK

- Optimization of other hyperparameters
- Preprocessing fair representation of Data (VAE)

HYPEROPTIMIZATION

- Multi-objective optimization
- Use of constraints

“It’s easier to make an algorithm fair than a man!”

Thanks

Raffaele Anselmo - 846842
Lorenzo Pastore - 847212