# Cross-attribute evaluation of fairness mechanisms

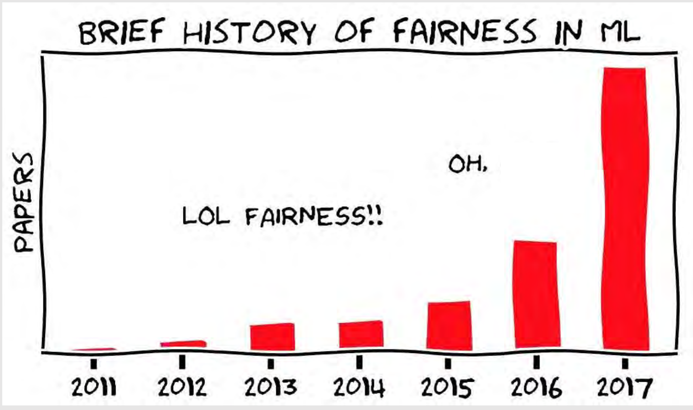
Student: Razvan-Andrei Morariu
Supervisors: Univ.-Prof. Dr. Benjamin Roth, Pedro Henrique Luz de Araujo, Anastasiia Sedova

universität wien

## Introduction

Fairness in ML ensures that algorithms produce unbiased outcomes, treating all individuals equally regardless of characteristics like race, gender, or age.

ML models used on certain datasets yield biased and unfair outcomes. As a result, mechanisms ensuring fairness in ML have emerged.

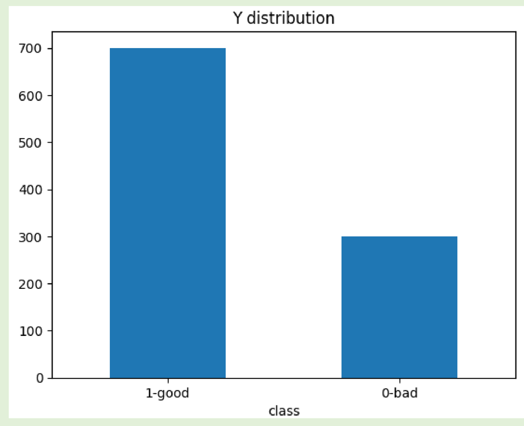



## German Credit Dataset

The dataset has:

- 1000 entries
- 20 features/descriptors (such as age, gender, credit amount, purpose)
- 2 sensitive features: **age and gender**

**Binary classification** : If the credit of a person is good or bad in terms of risk



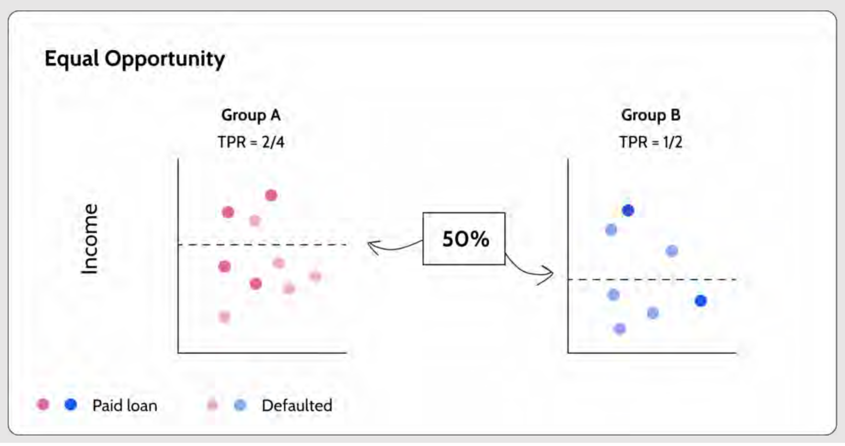

## Fairness

Biased Dataset + ML Model → Unfair Predictions

Biased Dataset + ML Model + Fairness Mechanism → Fair Predictions

Fairness mechanisms: pre-process, in-process, **post-process**
**Post-process** : Do the fairness optimization on the predictions of ML model

Measure: **Equal opportunity difference** :
$|P(\hat{Y}=1|Y=1,A=1) - P(\hat{Y}=1|Y=1,A=0)| \in [0,1]$



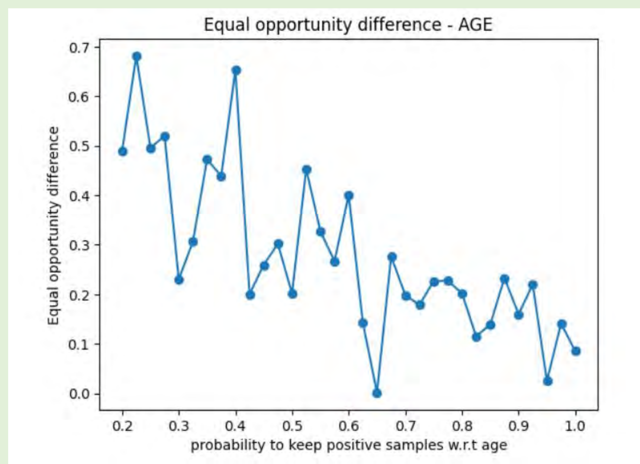

## Bias

One very common source of bias is under-representation bias.
There are bias injection methods to simulate this type of bias in order to understand its consequences.

Entire dataset + chose a category to be under-represented → Biased dataset

**Probability p**: keep every positive sample from the selected category
**Probability 1-p**: discard it

**Pipeline: Inject bias at different levels and check the effects on fairness optimization algorithms**




## Goal of the Project

Fairness can be evaluated w.r.t. different features (age, gender, race).

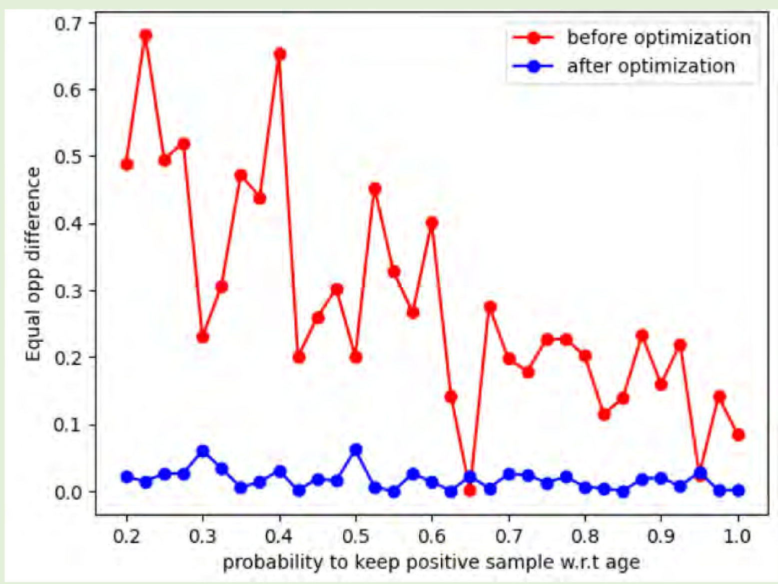
How does fairness optimization for one feature influence the others?

Fairness for age ↔ Fairness for gender

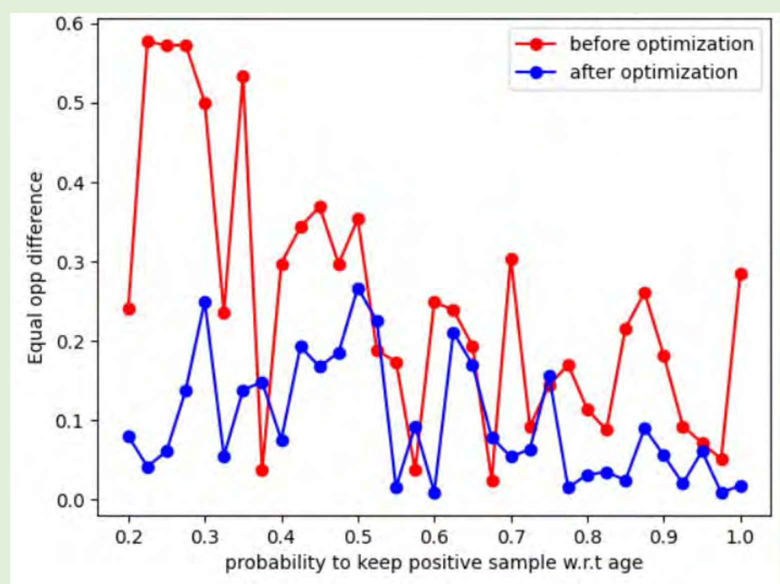
Accuracy

## Results

Fairness for age - train set




Fairness for age - test set




Similar results were obtained for gender.
There is no clear correlation between age fairness improvement and gender fairness improvement.
There might be a negative correlation between age fairness improvement and accuracy.

## Selected References

Avrim Blum, Kevin Stangly: Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?
Moritz Hardt, Eric Price, Nathan Srebro: Equality of Opportunity in Supervised Learning
Yuji Roh, Kangwook Lee, Steven Euijong Whang, Changho Suh: Sample Selection for Fair and Robust Training
XiaoqianWang, Heng Huang: Approaching Machine Learning Fairness through Adversarial Network