Audit

Dataset

Choose a dataset that you want to audit.

○ **COMPAS**
The COMPAS dataset was collected by ProPublica for their article "Machine Bias." We preprocessed the dataset to make it usable for this demo. The predicted scores are the original (decimal) scores from COMPAS.

- Y=0: Was arrested within two years
- Y=1: Was not arrested within two years
- D=0: Predicted to be rearrested
- D=1: Predicted not to be rearrested
- Group A: Black
- Group B: white

You can find the notebook here to see how we prepared the data.

○ **Credit lending (UCI German Credit)**
The German Credit dataset is available in the UCI repository. It is a small dataset of German credit loans from the 1970s. The scores have been predicted with a vanilla logistic regression.

- Y=0: Defaulted on the loan
- Y=1: Repaid the loan
- D=0: Predicted to default
- D=1: Predicted to repay
- Group A: female
- Group B: male

You can find the notebook here to see how we prepared the data.

○ **ACSEmployment (California)**
The ACSEmployment dataset is derived from US Census data and is available through the Folktables GitHub repository. It is a large dataset of US adults from California. The task is to predict whether an individual is employed. The scores have been predicted with a vanilla logistic regression.

- Y=0: Is not employed
- Y=1: Is employed
- D=0: Predicted to be unemployed
- D=1: Predicted to be employed
- Group A: Black
- Group B: white

You can find the notebook here to see how we prepared the data.

◉ **Choose your own dataset:** Datei auswählen ACSData_full.csv
If you want to upload your own dataset as a CSV file, please make sure that it has

- a column named 'Y' (only 0 and 1 allowed)
- a column named 'sensitive-attribute' (only 0 and 1 allowed)
- a column named 'scores' (values have to be between 0 and 1) and/or a column named 'D' (only 0 and 1 allowed)

You can also upload a JSON file with an array of objects that contain the previously mentioned attributes

Terminology

Y: The actual outcome, also known as the "ground truth"; not known at prediction time.

Label the two possible outcomes:
Y=1 an employed individual (Y=1)
Y=0 an unemployed individual (Y=0)

D: The decision in question; is trying to predict Y.

Decision maker: The people or organization designing the algorithm, deciding on its design and thereby ultimately taking the decisions in question.
Decision subjects: The people subjected to the decisions of the algorithm. They may or may not be aware that this algorithm is being deployed and used to make decisions about them.

Configuration

Decision maker's utility

How much utility does the decision maker derive from the decisions?

Currency of the decision maker

In what unit do you want to measure the utility of the decision maker (e.g., USD, well-being)? 00k USD

Quantification of the decision maker's utility

How much utility does the decision-maker derive from an employed individual (Y=1) that is getting predicted to be employed (D=1)?
-2 00k USD
How much utility does the decision-maker derive from an unemployed individual (Y=0) that is getting predicted to be employed (D=1)?
-5 00k USD
How much utility does the decision-maker derive from an employed individual (Y=1) that is getting predicted to be unemployed (D=0)?
-3 00k USD
How much utility does the decision-maker derive from an unemployed individual (Y=0) that is getting predicted to be unemployed (D=0)?
-3 00k USD

Fairness score

How should the utility of the decision subjects be distributed?

Sensitive attribute

What are the two groups that you want to compare and that are defined by the 'sensitive-attribute' column?
Group A (sensitive-attribute=0) caucasian
Group B (sensitive-attribute=1) african-american

Claim differentiator

Do the socio-demographic groups have the same moral claims to utility or is it only a subgroup of them? For example, one could argue that the subgroup of people with Y=1 is deserves a higher (or lower) utility than people with Y=0.

Define the subgroup in which people are deserving of the same amount of utility:
- (x) Everyone deserves the same utility
- ( ) People with Y=0 deserve the same utility
- ( ) People with Y=1 deserve the same utility
- ( ) People with D=0 deserve the same utility
- ( ) People with D=1 deserve the same utility

Decision subjects' utility

How much utility do the decision subjects derive from the decisions?

Currency of decision subjects

In what unit do you want to measure the utility of the decision subject (e.g., USD, well-being)? happy

Quantification of the decision subjects' utility

For the group: caucasian

How much utility does an employed individual (Y=1) derive from getting predicted to be unemployed (D=0)?

-1 happy

How much utility does an unemployed individual (Y=0) derive from getting predicted to be unemployed (D=0)?

-1 happy

For the group: african-american

How much utility does an employed individual (Y=1) derive from getting predicted to be employed (D=1)?

1 happy

How much utility does an unemployed individual (Y=0) derive from getting predicted to be employed (D=1)?

0 happy

How much utility does an employed individual (Y=1) derive from getting predicted to be unemployed (D=0)?

-1 happy

How much utility does an unemployed individual (Y=0) derive from getting predicted to be unemployed (D=0)?

-1 happy

Pattern of Justice

How should the utility be distributed between the two groups (defined by the sensitive attribute)?

Egalitarianism: Fairness is if individuals in both groups are expected to derive the same utility from the decision rule. Equality in itself is valued.
→ Measured as: *How close are the average utilities to being equal?*
Maximin: Fairness is if the average utility of the worst-off group is maximized by the decision rule. Inequalities are okay if they benefit the worst-off group.
→ Measured as: *What's the lowest average utility?*
Prioritarianism: Fairness is if the aggregated utility of the groups is maximized by the decision rule, with the utility of the worst-off group being weighted higher than the other groups' utilities.
→ Measured as: *What's the aggregated utility with the worst-off group having a higher weight?*
Sufficientarianism: Fairness is if all groups' have an average utility that is above the defined threshold. Inequalities are okay if every group is above the defined threshold.
→ Measured as: *How many groups are above the defined threshold?*

Choose a pattern: egalitarianism
If you're unsure what to choose here, we recommend egalitarianism for your first evaluation.

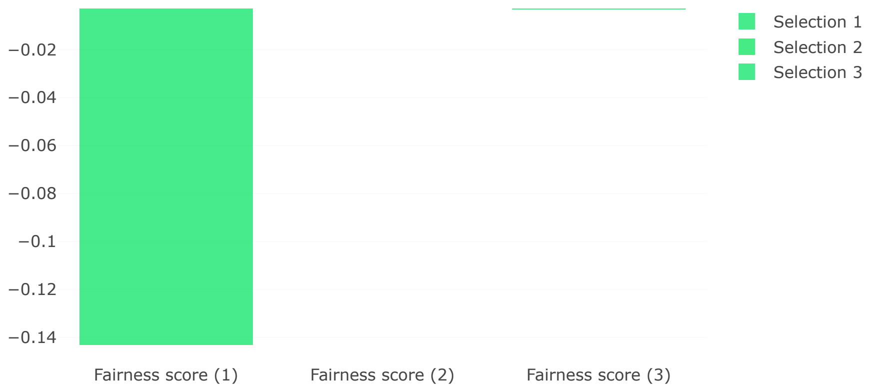
Audit

Resulting fairness metric

In the audit, we will use the fairness metric that you defined with your inputs above. Specifically, we will look at the following fairness metric:
Negative absolute difference in average utility of caucasian and african-american (so 0 is perfect equality)

Fairness score

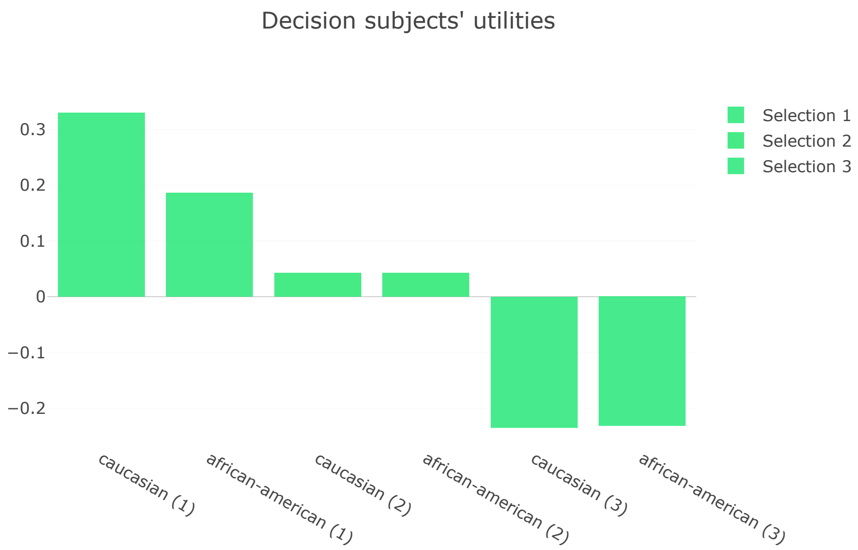
Here, you can see a direct comparison of the fairness scores (for the points selected in the Pareto plot below). The higher the score, the better the decision rule aligns with the configured fairness metric. The lower the score, the worse its alignment with the fairness



metric is.

Decision subjects' utilities

Here you can see a direct comparison of the decision subjects' average utilities (for the points selected in the Pareto plot below).



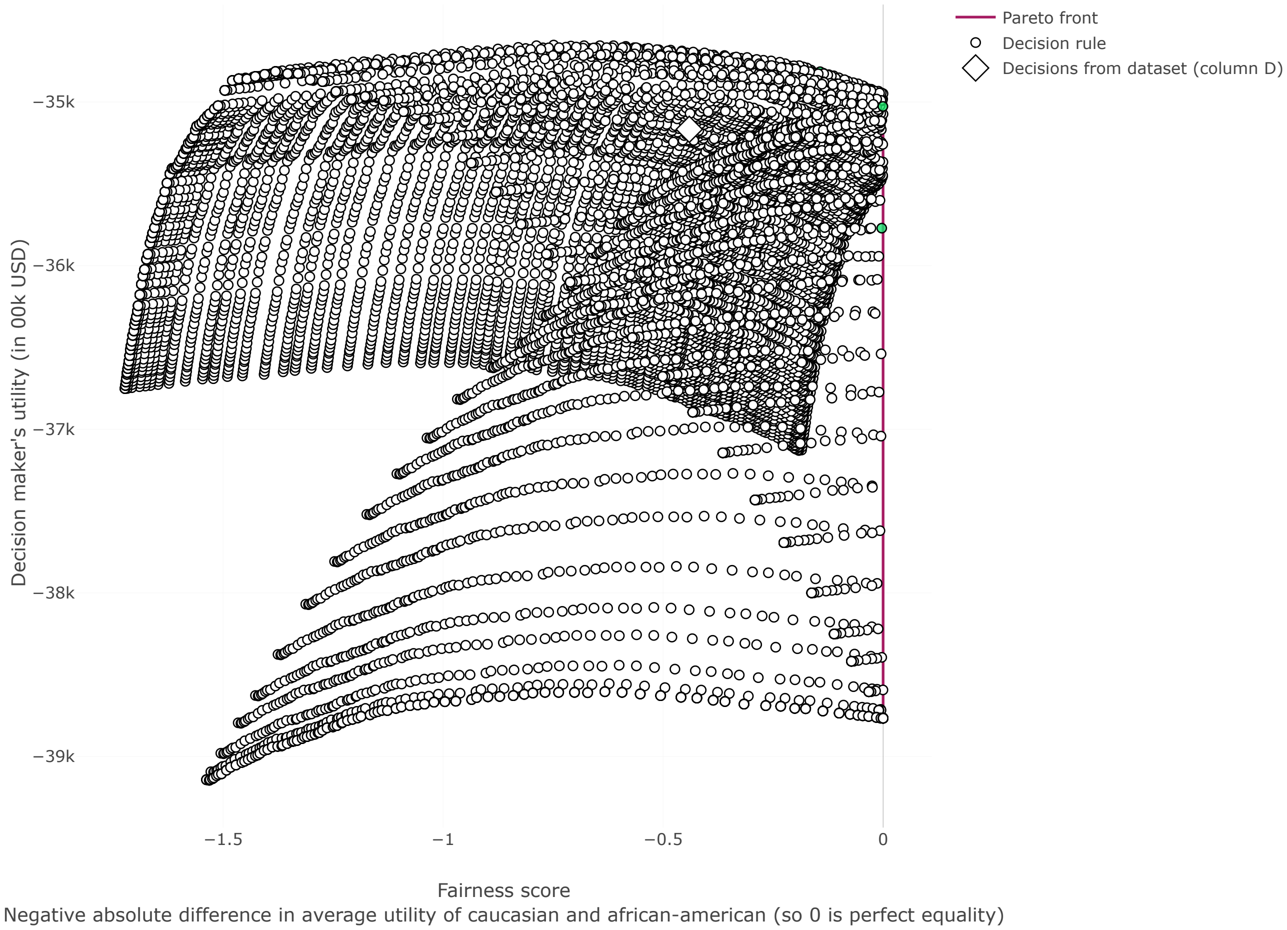
Pareto plot

With the decision maker utility and a fairness metric specified, we can take a simple approach to show the trade-offs between these metrics: We go through different decision rules and calculate the metrics associated with each of them, i.e., the decision maker's utility and the fairness score. For each decision rule, we then plot the associated decision maker’s utility and fairness score in a 2D plot. We use group-specific thresholds as decision rules. Select threshold rules that you want to compare by clicking on the points in the plot.

Decision maker's utility: Higher is better (total utility for the 12922 individuals in the dataset)
Fairness score: Higher is better

Number of thresholds: How many thresholds do you want to test for each group? (min: 2, max: 101) 101

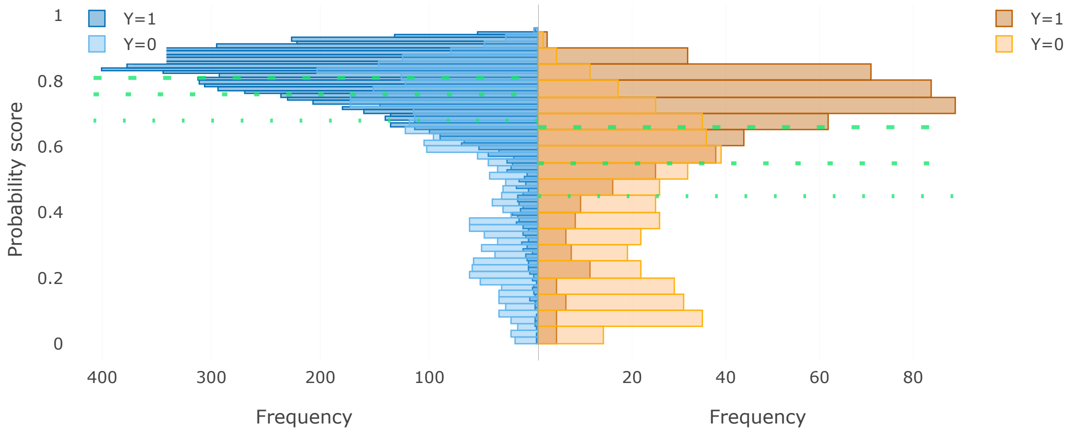
Deselect all points



Selected Decision Rules

Selection	Thresholds	Decision maker's utility	Fairness score
1	caucasian: 0.68; african-american: 0.45	-34815 00k USD	-0.1432
2	caucasian: 0.76; african-american: 0.55	-35027 00k USD	-0.0004
3	caucasian: 0.81; african-american: 0.66	-35771 00k USD	-0.0032

Score distribution


Y=1
Y=0
Y=1
Y=0
Probability score
1
0.8
0.6
0.4
0.2
0
400
300
200
100
20
40
60
80
Frequency
Frequency