Holistic AI × The Alan Turing Institute

Bias in Regression Tasks – Part III

Content by: Sachin Beepath, Giulio Filippi, Nigel Kingsman, Cristian Munoz, Roseline Polle, Sara Zannone

Speaker: Sara Zannone

Contents

- Part I – Introduction to Regression
- Part II – Fairness in Regression
- **Part III – Measuring Bias in Regression**
- Part IV – Mitigating Bias in Regression

Measuring Bias in Regression Tasks

In our previous lecture, we have seen multiple definitions of Fairness for Regression tasks.

We will now show how there definitions can be operationalized into metrics.

Bias metrics allow us to estimate the bias of an AI system or dataset.

Equality of Opportunity

Equality of Opportunity:

The performance of an AI system should be the same across all groups

For example, a facial recognition algorithm is trained to predict age from photos. We want to make sure that it's equally accurate for black and white people.

Equality of Opportunity

Equality of Opportunity:
The performance of an AI system should be the same across all groups

For example, a facial recognition algorithm is trained to predict age from photos. We want to make sure that it's equally accurate for black and white people.

Equality of Opportunity

Equality of Opportunity:
The performance of an AI system should be the same across all groups

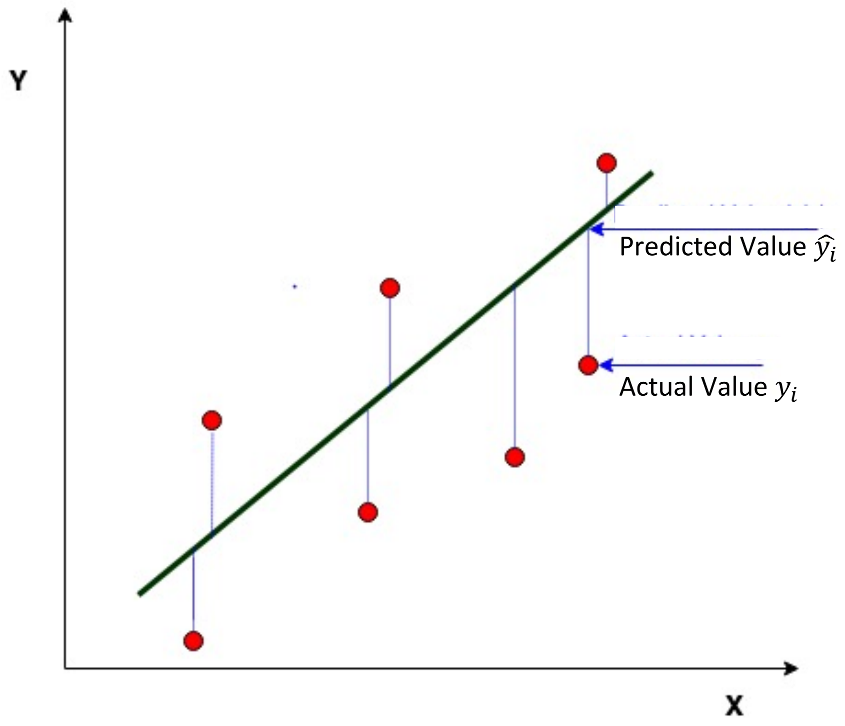
For example, a facial recognition algorithm is trained to predict age from photos. We want to make sure that it’s equally accurate for black and white people.

- We can measure this by <u>comparing the error</u> the model makes across groups

RMSE ratio

We want to compare the error of an AI system for group A and group B (e.g. white / black people).

RMSE: $\sqrt{\sum_{i=1}^{n} \frac{(y_i - \widehat{y_i})^2}{n}}$

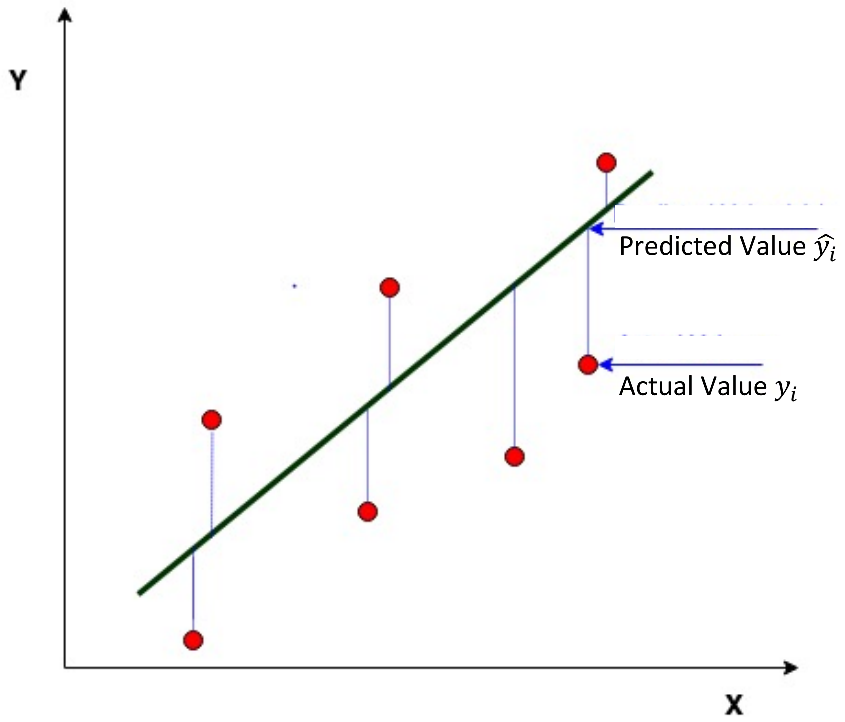


RMSE ratio

We want to compare the error of an AI system for group A and group B (e.g. white / black people).

RMSE: $\sqrt{\sum_{i=1}^{n} \frac{(y_i - \widehat{y_i})^2}{n}}$

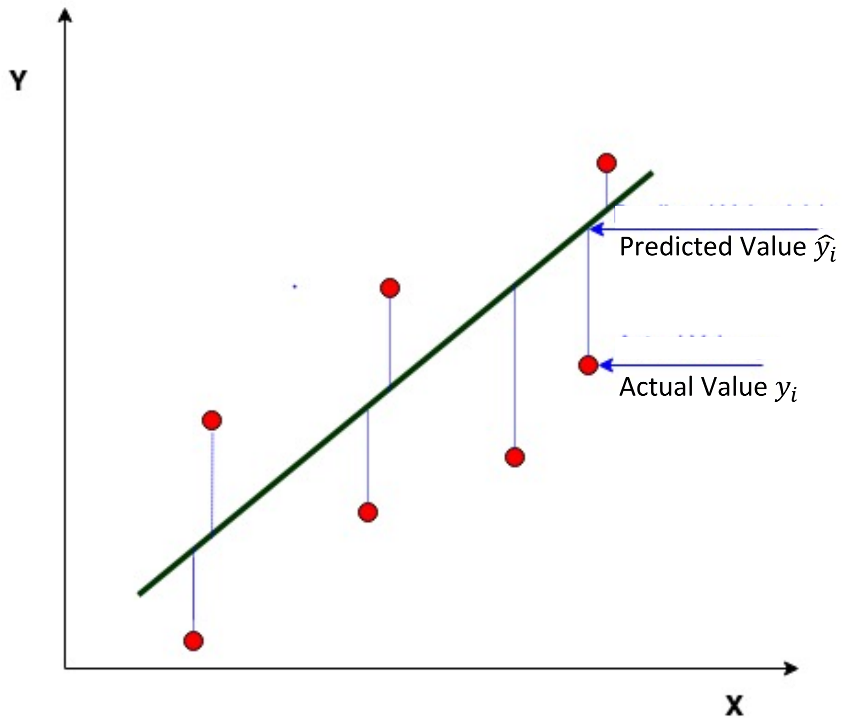
RMSE ratio: $\frac{RMSE_a}{RMSE_b}$



MAE ratio

We want to compare the error of an AI system for group A and group B (e.g. white / black people).

MAE: $\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{n}$

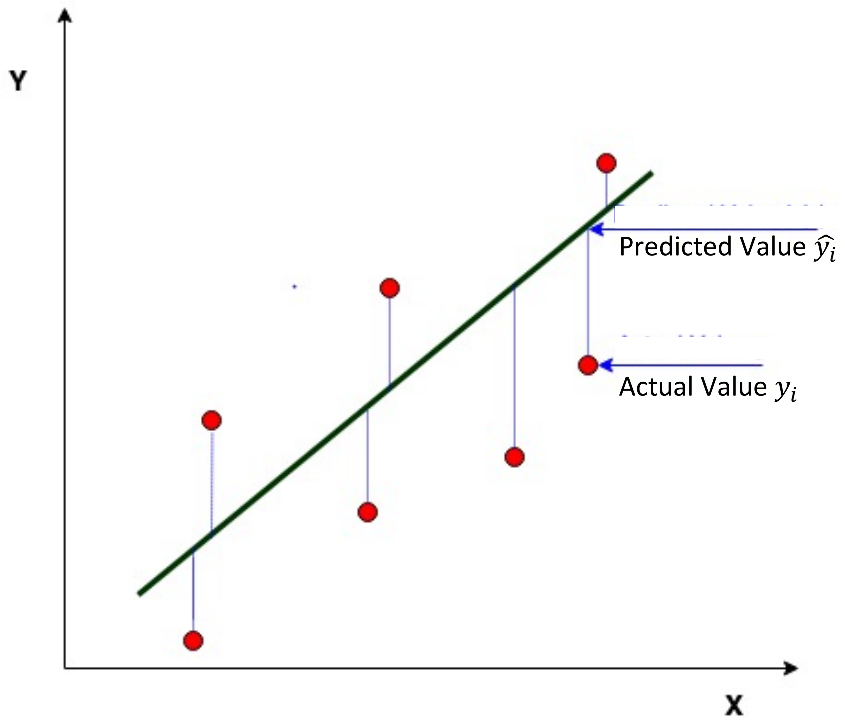


MAE ratio

We want to compare the error of an AI system for group A and group B (e.g. white / black people).

MAE: $\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{n}$

MAE ratio: $\frac{MAE_a}{MAE_b}$



Equality of Outcome

Equality of Outcome:
The model's output should be similar across groups

- In practice, we need to discretize the outputs to compare their distributions

For example, an algorithm scores candidates CVs. We want to make sure that the distribution of scores is similar for both men and women.

Equality of Outcome

Equality of Outcome:
The model's output should be similar across groups

- In practice, we need to discretize the outputs to compare their distributions

For example, an algorithm scores candidates CVs. We want to make sure that the distribution of scores is similar for both men and women.

Equality of Outcome

Equality of Outcome:
The model's output should be similar across groups

- In practice, we need to discretize the outputs to compare their distributions

For example, an algorithm scores candidates CVs. We want to make sure that the distribution of scores is similar for both men and women.

Binary Classification Bias Metrics

In binary classification, the outcome $\hat{y} \in \{0,1\}$.

- For equality of outcome metrics, we can compare the success rate SR for different groups

$$SR = \frac{\#\, Successful\ outcomes}{\#\, Total\ outcomes}$$

For example, an algorithm decides which candidates should be hired. We want to make sure that the success rate is similar for both men and women.

Binary Classification Bias Metrics

In binary classification, the outcome $\hat{y} \in \{0,1\}$.

- For equality of outcome metrics, we can compare the success rate SR for different groups

$$SR = \frac{\#\, Successful\ outcomes}{\#\, Total\ outcomes}$$

For example, an algorithm decides which candidates should be hired. We want to make sure that the distribution of the scores is similar for both men and women.

Binary Classification Bias Metrics

In binary classification, the outcome $\hat{y} \in \{0,1\}$.

- For equality of outcome metrics, we can compare the success rate SR for different groups

$$SR = \frac{\#\,Successful\ outcomes}{\#\,Total\ outcomes}$$

$\text{Disparate Impact} = \frac{SR_a}{SR_b}$

Ideal value = 1

$\text{Statistical Parity} = SR_a - SR_b$

Ideal value = 0

Binary Classification Bias Metrics

In binary classification, the outcome $\hat{y} \in \{0,1\}$.

- For equality of outcome metrics, we can compare the success rate SR for different groups

$$SR = \frac{\#\, Successful\ outcomes}{\#\, Total\ outcomes}$$

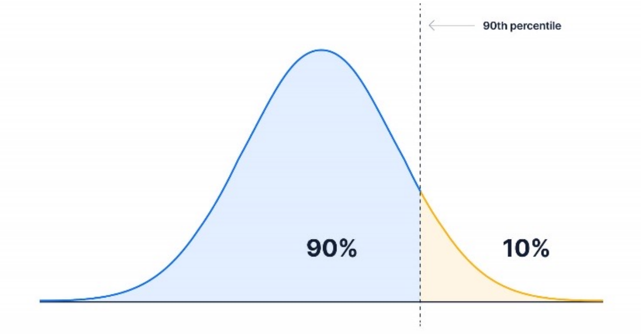
$$\text{Disparate Impact} = \frac{SR_a}{SR_b} \qquad\qquad \text{Statistical Parity} = SR_a - SR_b$$

How can these metrics be extended to regression?

Fixed binarization

A simple solution consists of fixing a threshold to binarize the regression data.

- Let's go back to the previous example: we have an algorithm that scores CVs. We can decide to hire only the top 10% of the applicants (0.9 quantile/ 90th percentile), our data will thus become binary.



More generally, we can define:

$$SR_g = \frac{\#\, Outcomes\ in\ group\ g\ that\ fall\ in\ the\ top\ percentile}{\#\, Total\ outcomes\ in\ group\ g}$$

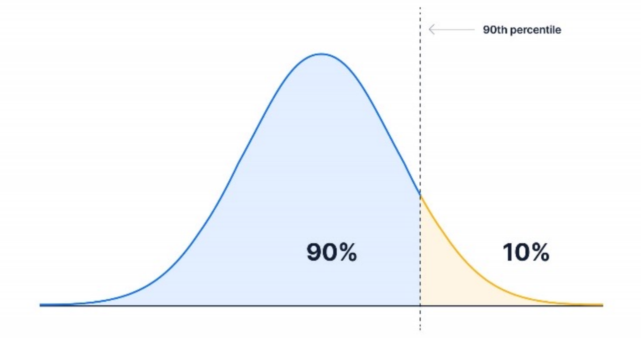
$$\text{Disparate Impact} = \frac{SR_a}{SR_b} \qquad \text{Statistical Parity} = SR_a - SR_b$$

Fixed binarization

A simple solution consists of fixing a threshold to binarize the regression data.

- Let's go back to the previous example: we have an algorithm that scores CVs. We can decide to hire only the top 10% of the applicants (0.9 quantile/ 90th percentile), our data will thus become binary.

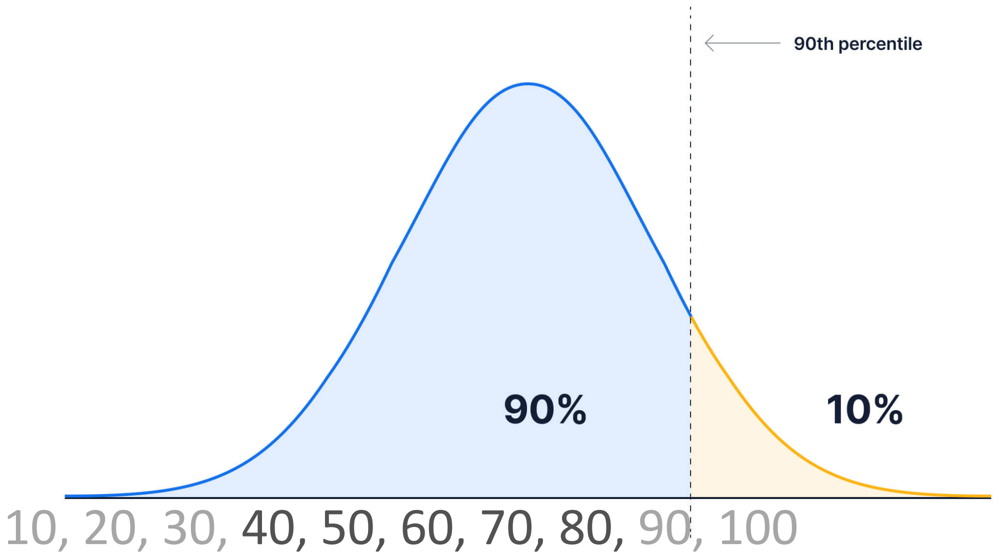
More generally, we can define:



$$SR_g = \frac{\#\, Outcomes\ in\ group\ g\ that\ fall\ in\ the\ top\ percentile}{\#\, Total\ outcomes\ in\ group\ g}$$

$$\text{Disparate Impact} = \frac{SR_a}{SR_b} \qquad \text{Statistical Parity} = SR_a - SR_b$$

Example



M = [10, 20, 30, 90, 100]

F = [40, 50, 60, 70, 80]

If we consider the 90th percentile, then only one of the male candidates will fall in it.

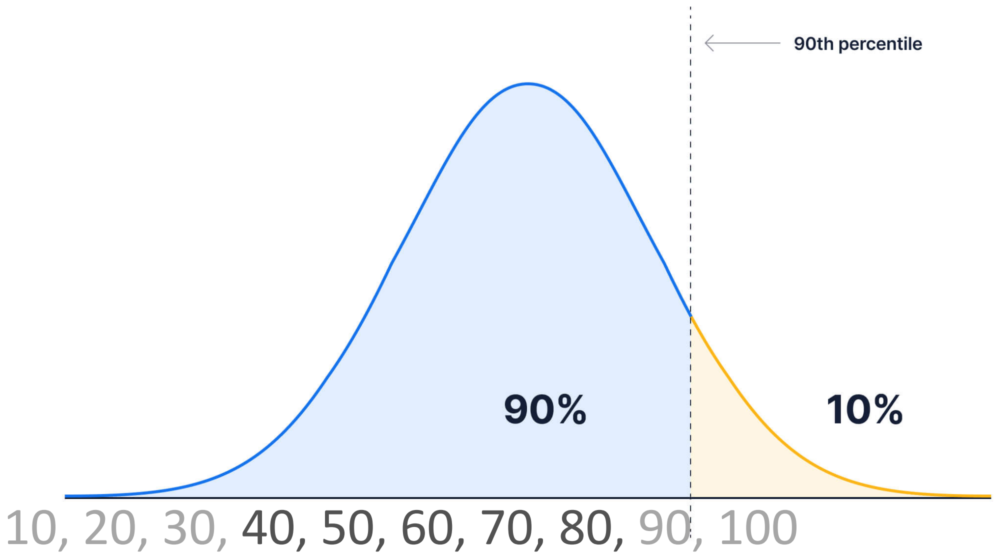
$$SR_M = 0.2$$

$$SR_F = 0$$

Disparate Impact $= 0$ Statistical Parity $= -0.2$

- With this binarization, the data is unfair towards women.

Example



M = [10, 20, 30, 90, 100]

F = [40, 50, 60, 70, 80]

If we consider the 90th percentile, then only one of the male candidates will fall in it.

$$SR_M = 0.2$$

$$SR_F = 0$$

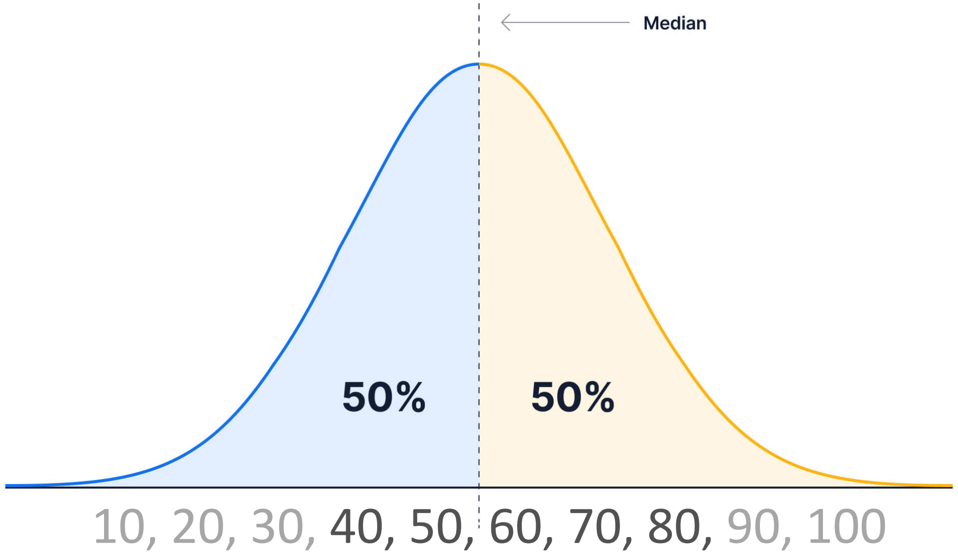
Disparate Impact $= 0$ Statistical Parity $= -0.2$

- With this binarization, the data is unfair towards women.

Example

M = [10, 20, 30, 90, 100]

F = [40, 50, 60, 70, 80]



If we consider the median ($m = 55$), then we will have 3 women and 2 men who are successful.

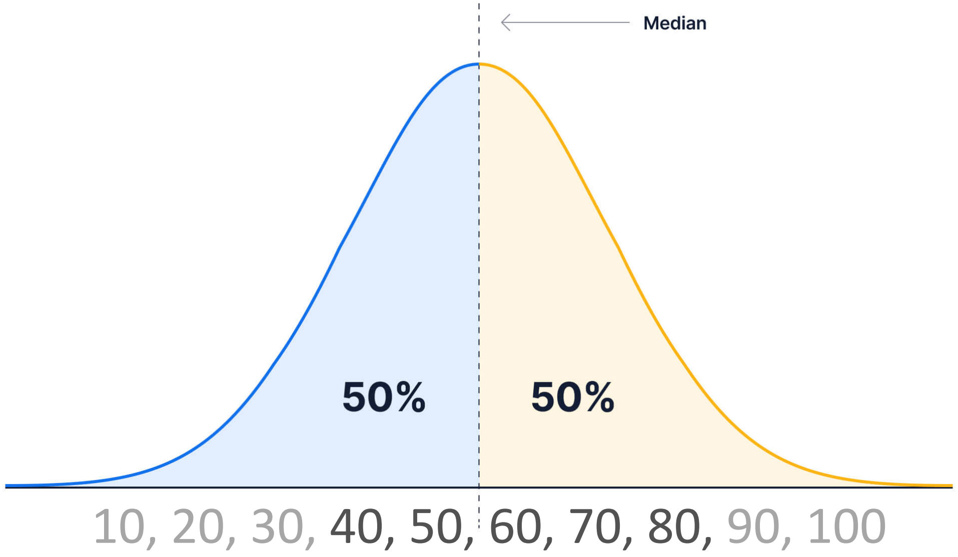
$$SR_M = 0.4$$

$$SR_F = 0.6$$

$\text{Disparate Impact} = 1.5$ $\quad\quad$ $\text{Statistical Parity} = 0.2$

- Changing the binarization, the data becomes unfair towards men.

Example



M = [10, 20, 30, 90, 100]

F = [40, 50, 60, 70, 80]

If we consider the median ($m = 55$), then we will have 3 women and 2 men who are successful.

$$SR_M = 0.4$$

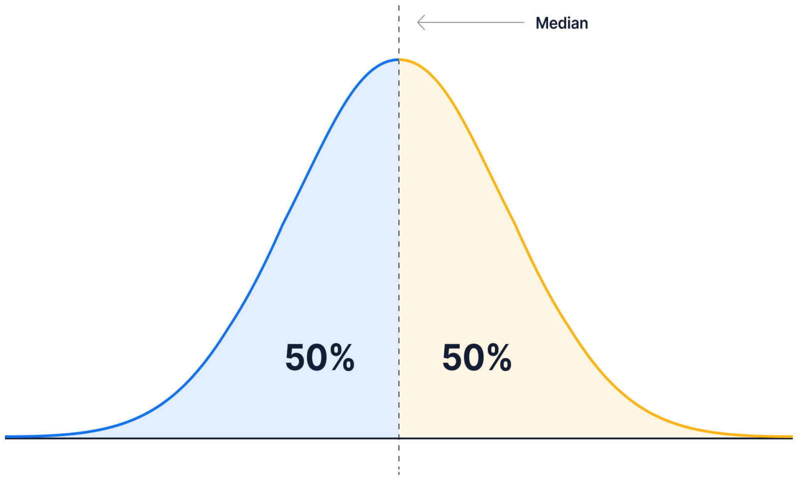
$$SR_F = 0.6$$

$\text{Disparate Impact} = 1.5$ $\text{Statistical Parity} = 0.2$

➢ Changing the binarization, the data becomes unfair towards men.

Binarization

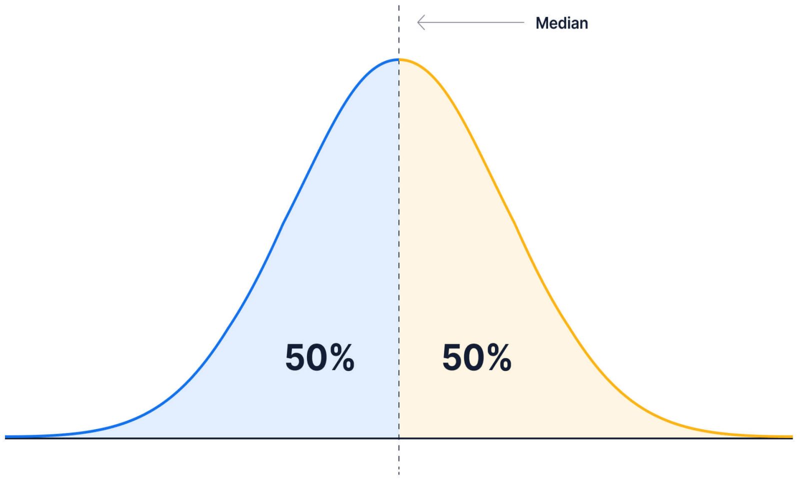
- The NYC bias audit mandate (Local Law 144) requires companies to measure the bias of AI systems used in recruitment
  - For regression tasks, the metric required computes the disparate impact by binarizing at the median (like in our example)
  - Filippi et al. 2023



- However, other solutions are possible:
  - We can decide a fixed threshold NOT based on the distribution (Agarwal et al.2019) e.g. all candidates that score above 50
  - We could consider the ranking of the scores (e.g. Raj and Ekstrand 2022)
  - We can use metrics that take into account the whole distribution, not only at a fixed threshold

Binarization

- The NYC bias audit mandate (Local Law 144) requires companies to measure the bias of AI systems used in recruitment
  - For regression tasks, the metric required computes the disparate impact by binarizing at the median (like in our example)
  - [Filippi et al. 2023]



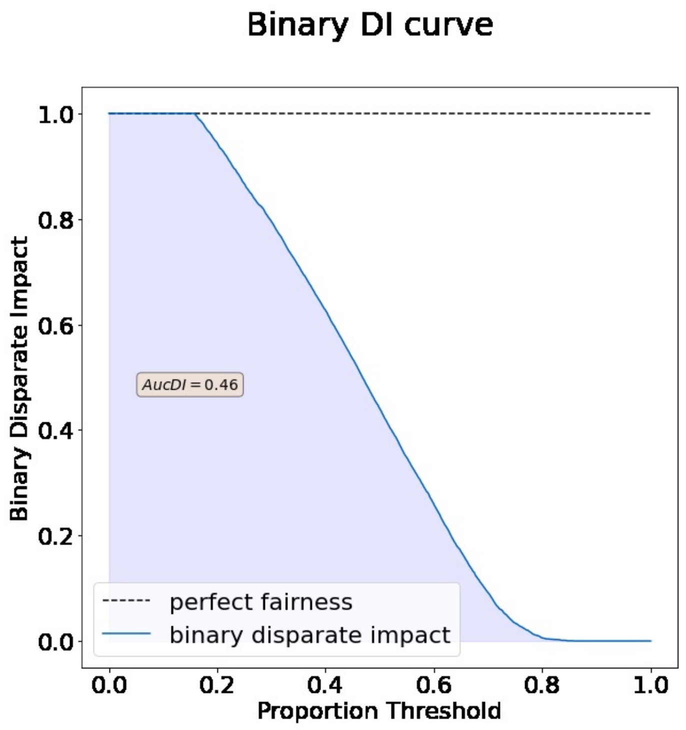
- However, other solutions are possible:
  - We can decide a fixed threshold NOT based on the distribution (Agarwal et al.2019) e.g. all candidates that score above 50
  - We could consider the ranking of the scores [(e.g. Raj and Ekstrand 2022)]
  - We can use metrics that take into account the whole distribution, not only at a fixed threshold

AUC Disparate Impact

- Thresholding the data at a fixed value (like the median) is often not sufficient to describe the whole distribution.

- An alternative metric that we proposed in our work ([Filippi et al. 2023]) is to consider the evolution of the Disparate Impact while varying the quantile threshold value.

- We can then compute a metric by calculating the area under the curve. The larger the area, the closer it will be to fairness.

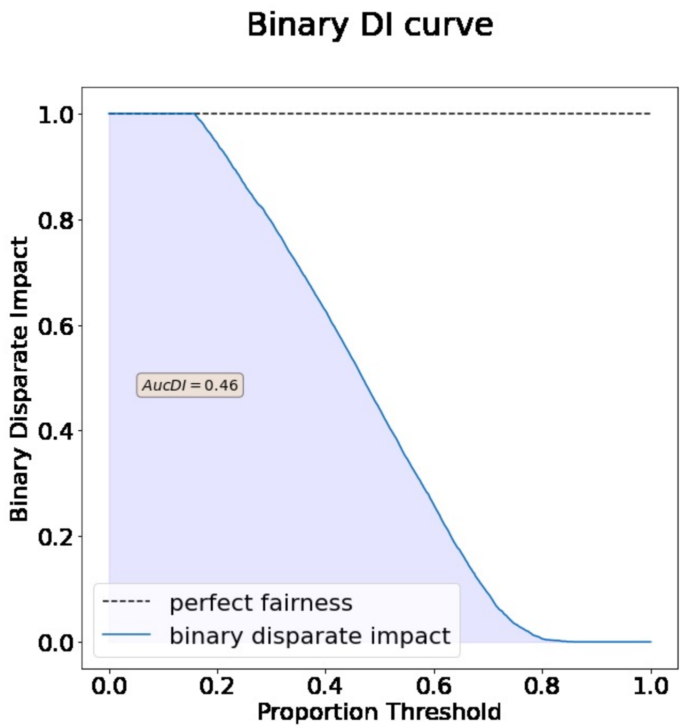


AUC Disparate Impact

- Thresholding the data at a fixed value (like the median) is often not sufficient to describe the whole distribution.

- An alternative metric that we proposed in our work ([Filippi et al. 2023]) is to consider the evolution of the Disparate Impact while varying the quantile threshold value.

- We can then compute a metric by calculating the area under the curve. The larger the area, the closer it will be to fairness.



Computing Bias Metrics with holisticai library

- The first step is installing the library

```
# install the holisticai library
!pip install holisticai
```

- We can now import the bias metrics we would like to use

```
# import regression bias metrics
from holisticai.bias.metrics import statistical_parity_regression
from holisticai.bias.metrics import disparate_impact_regression
from holisticai.bias.metrics import mae_ratio
from holisticai.bias.metrics import rmse_ratio
```

Computing Bias Metrics with holisticai library

- We can then define two binary group membership vectors

```
group_a = np.array(X['sex']=='Male')
group_b = np.array(X['sex']=='Female')
```

- Finally, we can compute the metrics

```
# evaluate fairness metrics for gender
print ('Statistical Parity Q80   : ' + str(statistical_parity_regression(group_a, group_b, y_pred, q=0.8)))
print ('Disparate Impact Q80     : ' + str(disparate_impact_regression(group_a, group_b, y_pred, q=0.8)))
print ('MAE Ratio Q80            : ' + str(mae_ratio(group_a, group_b, y_pred, y_true,q=0.8)))
print ('RMSE Ratio Q80           : ' + str(rmse_ratio(group_a, group_b, y_pred, y_true,q=0.8)))
```

```
Statistical Parity Q80   : 0.10488505747126436
Disparate Impact Q80     : 1.839080459770115
MAE Ratio Q80            : 0.7557387626353143
RMSE Ratio Q80           : 0.8178214225397291
```

Exercise Notebooks

- We have created exercise Notebooks for measuring Bias.

Contents

References

[1] Ekstrand et al, 2023, Overview of the TREC 2022 Fair Ranking Track (https://arxiv.org/abs/2302.05558)

[2] Filippi et al, 2023, Local Law 144: A Critical Analysis of Regression Metrics (https://arxiv.org/abs/2302.04119)