Responsible AI

Lecture 4

Today

1. Introduction to fairness
2. Statistical parity difference
3. Disparate impact
4. Binary features with continuous output
5. Continuous features with binary output

Data Fairness

Stat Parity Difference

Disparate Impact

Terminology

- Y - actual
- Yhat – predicted
- X – Independent features
- S – Sensitive features
- S_a – Sensitive feature advantageous group
- S_d – Sensitive feature disadvantageous group
- Y^+ - Y with favourable outcome
- Y^- - Y with unfavourable outcome


What do we want?
EQUALITY!
When will we get it?
ANOTHER 100 YEARS!
imgflip.com

Statistical Parity Difference

$$SPD = P\left(Y = 1 \mid S = S_a\right) - P\left(Y = 1 \mid S = S_d\right) = 0$$

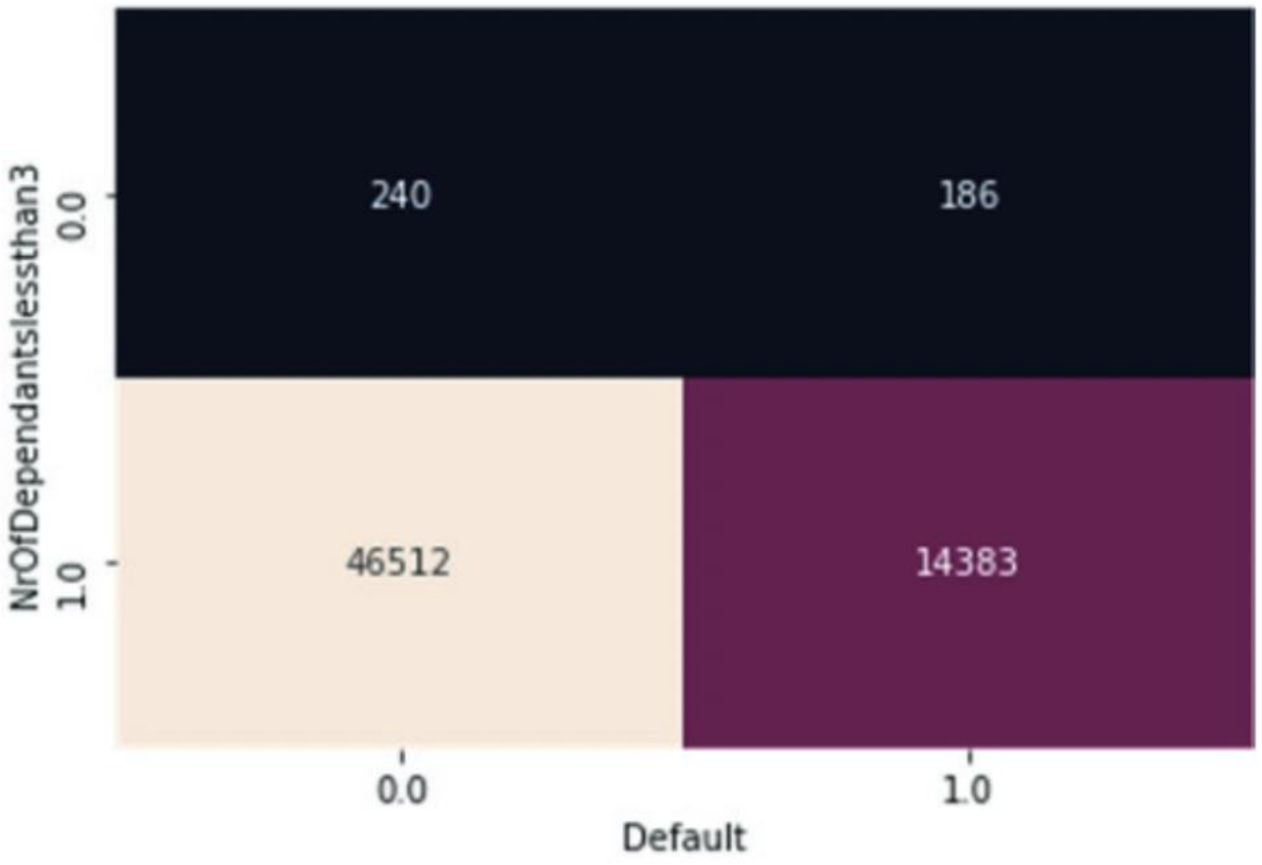
the acceptable values of SPD will range between 0 and 0.1

Disparate Impact

$$\frac{P\left(Y=1 \mid S=S_d\right)}{P\left(Y=1 \mid S=S_a\right)} \geq 0.8$$

$$\frac{P\left(Y=1 \mid S=S_a\right)}{P\left(Y=1 \mid S=S_d\right)} \leq 1.25$$

Heatmaps for investigation



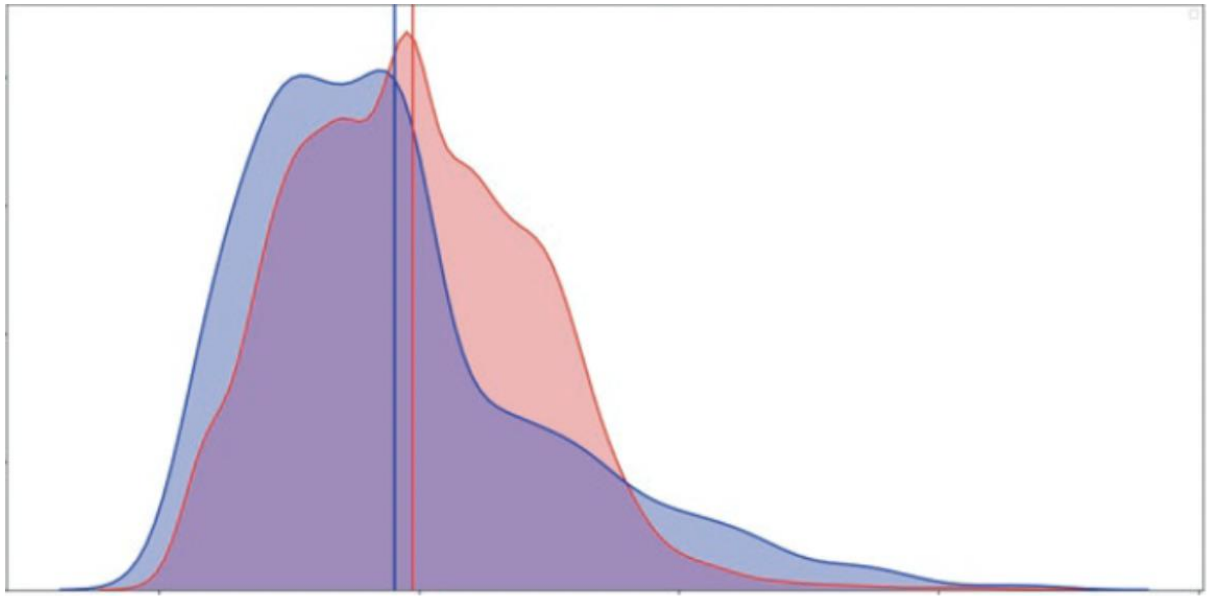
Calculate:

- SPD
- DI

When the Y Is Continuous and S Is Binary

Difference between des stats of two groups of a protected feature

- Mean
- Skewness
- Kurtosis
- Density plots



When the Y Is Binary and S Is Continuous

- Create iterative bins to find the bin with the maximum SPD and DI
- Create multiple bins

WHY?

$$\frac{\bar{m} - \overline{nm}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Self read: two-sample t test


WHEN I REBALANCE

Reweighting the Data

Reweighting assigns weights to the data based on a protected feature. These weights are then used along with the input data for loss function optimization. A big benefit of this approach is that the data is not altered to achieve any reduction in the bias

Why Rw

- It can handle one protected feature at a time.
- Only for classification-based algorithm.
- You can create composite features to handle multiple features together – composite features can be an incredibly powerful way to handle multiple features together, but the features you can combine will depend on the features and the problem at hand.
- And finally, no dip in accuracy (or hardly any!).

Steps

1. Choose a protected feature to generate weights
2. Generate weights and add them to the dataset
3. Choose an algorithm to train a model using both unweighted and weighted data
4. Calculate statistical parity difference for both models to compare the discrimination
5. Generate the accuracy and fairness metrics to compare the difference between the two group and between two models (with and without weights)
6. Repeat the above steps for all the protected features you want to analyse and compare to decide.

How to calculate weights

Remember this?

$$P\left(S=S_a \mid Y=Y^+\right)-P\left(S=S_d \mid Y=Y^+\right)$$

With the weights, the discrimination can be calculated using

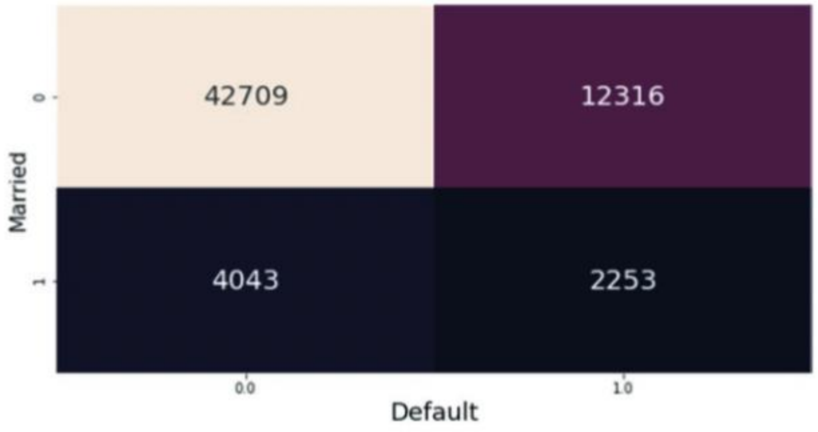
$$\left\{P\left(S=S_a \mid Y=Y^+\right)\times W_{S_a \wedge Y_{fav}}\right\}-\left\{P\left(S=S_d \mid Y=Y^+\right)\times W_{S_d \wedge Y_{fav}}\right\}$$

After Rw

There will always be 4 combinations

- $S_a \land Y_{fav}$: S = advantageous (S_a), Y = positive (Y^+ or Y_{fav})
- $S_a \land Y_{unfav}$: S = advantageous (S_a), Y = negative (Y^- or Y_{unfav})
- $S_d \land Y_{fav}$: S = disadvantageous (S_d), Y = positive (Y^+ or Y_{fav})
- $S_d \land Y_{unfav}$: S = disadvantageous (S_d), Y = negative (Y^- or Y_{unfav})

Lets calculate some weights



Total number of observations (n) = 61,321
Total number of observations including unprivileged class (S_d) = 6296
Total number of favourable observations (Y_{fav}) = 46,752
Total number of observations where unprivileged class had a favourable outcome ($S_d \wedge Y_{fav}$) = 4043

$$P\left(\text{Observed}_{S_d \wedge Y_{fav}}\right) = \frac{\left(S_d \wedge Y_{fav}\right)}{n}$$

$$P\left(\text{Observed}_{S_a \wedge Y_{fav}}\right) =$$

$$\text{Weight}_{S_a \wedge Y_{fav}} = \frac{P\left(\text{Expected}_{S_a \wedge Y_{fav}}\right)}{P\left(\text{Observed}_{S_a \wedge Y_{fav}}\right)}$$

$$P\left(\text{Expected}_{S_d \wedge Y_{fav}}\right) = \frac{Y_{fav}}{n} \times \frac{S_d}{n}$$

$$P\left(\text{Expected}_{S_a \wedge Y_{fav}}\right) =$$

$$\text{Weight}_{S_a \wedge Y_{fav}} = \frac{Y_{fav} \times S_a}{\left(S_a \wedge Y_{fav}\right) \times n}$$

Let's divide the class in 4 groups and calculate weights for all four combinations

Which combination got highest weight? Why

Can you also calculate discrimination before and after

New loss function in LR

How weights impacts loss function?
What is this doing actually??

$$J(\theta) = -\sum_{i=1}^{n} w_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]$$

Where w_i is the weight for the record X_i.

Use the metrics we discussed last time

EoD

EO

DP

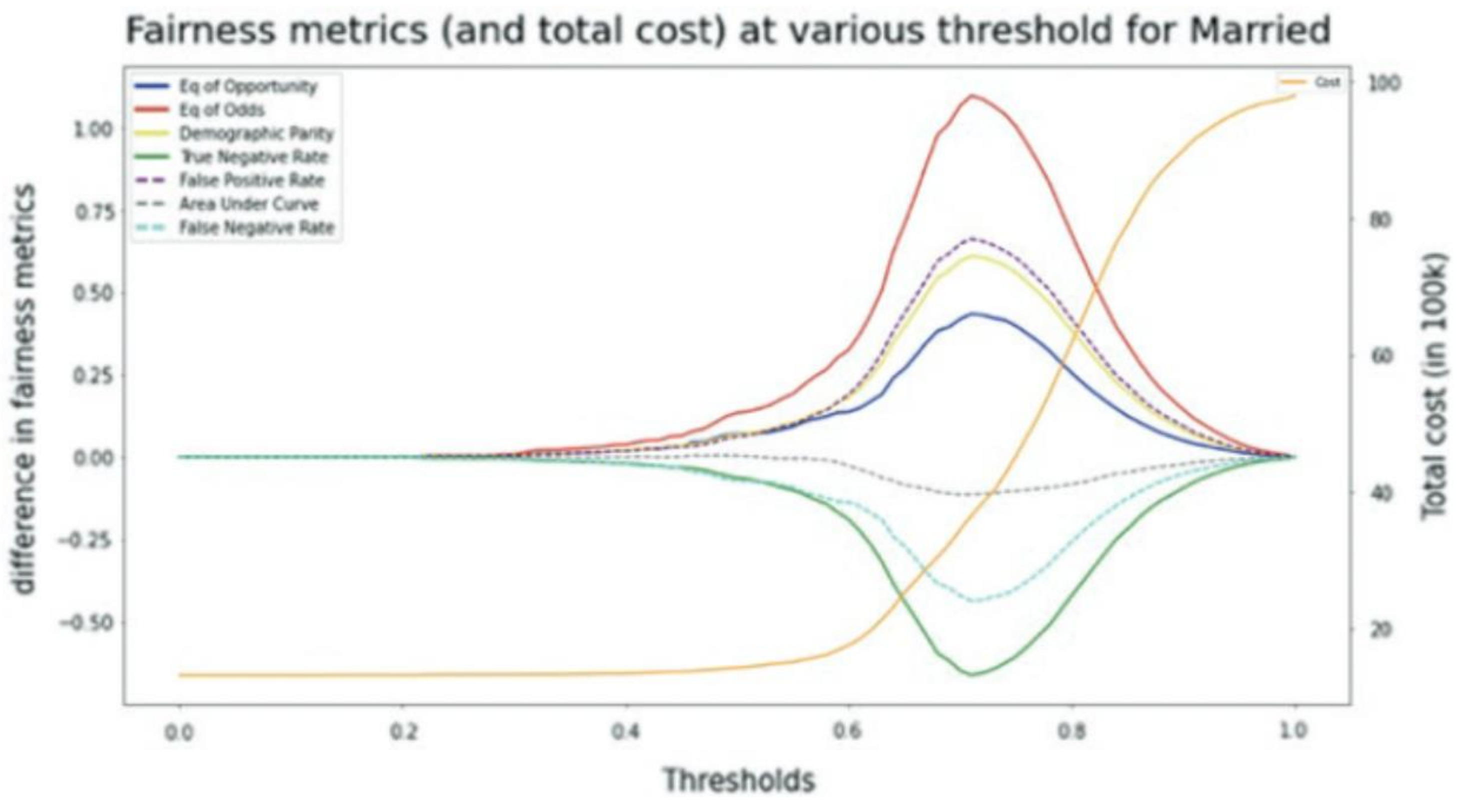
TPR

TNR

FPR

FNR

Calibrating Decision Boundary



What is cost?

Which is the right threshold?

Self learn: Composite Feature

Assessment

In groups use any data of your choice having PII data:

Implement RW in Algo (in one protected feature) of your choice and report before and after values (on original data)

Implement RW in Algo (in one protected feature) of your choice and report before and after values (on DP data)

Repeat the above (both) on composite PII data

Submission on 6th lecture

Explore

Github (https://github.com/srayagarwal/JIO_RAI/blob/main/Ch%203%20Bias%20in%20Data.ipynb

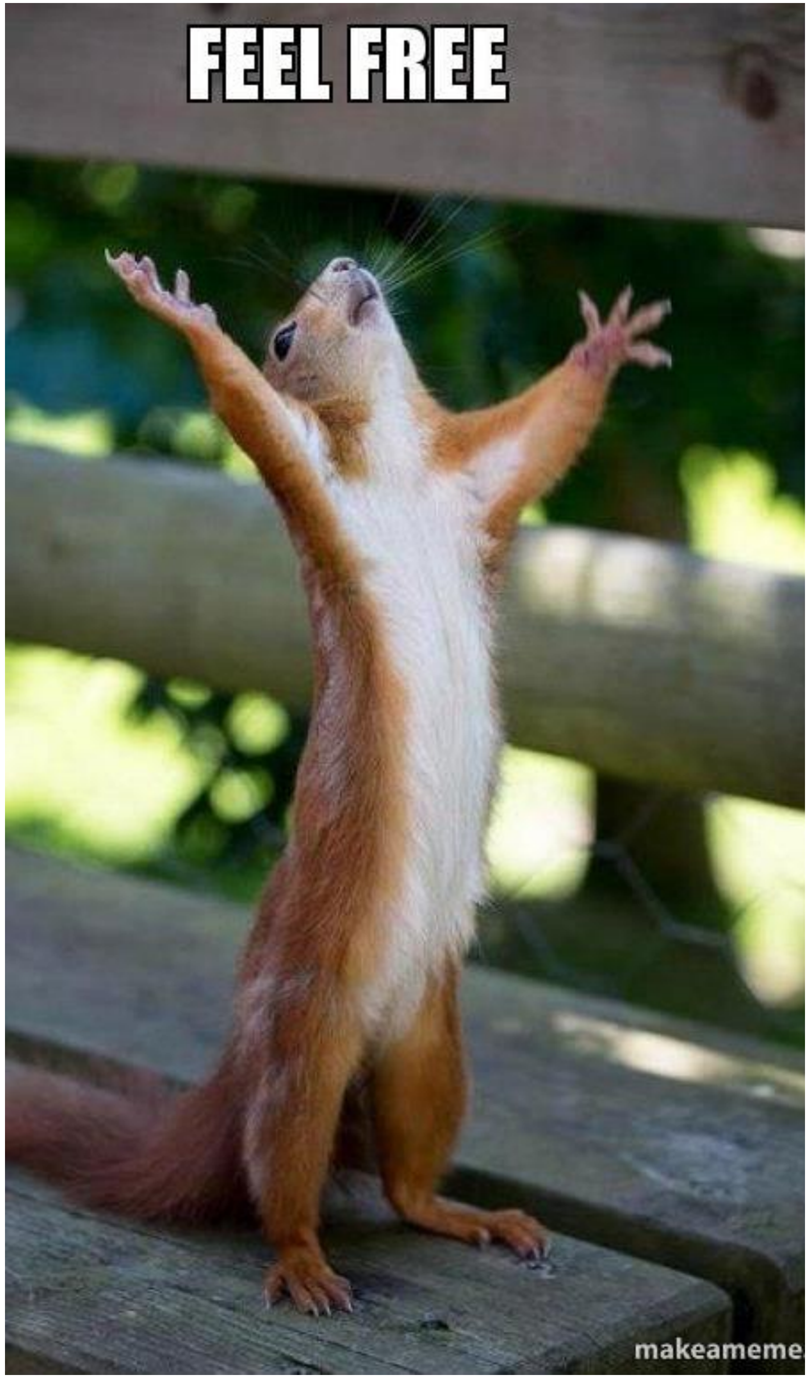
https://github.com/srayagarwal/JIO_RAI/blob/main/Ch%205%20Remove%20Bias%20from%20ML%20Model%20I.ipynb)

Next

1. Explainable AI
2. Introduction to XAI
3. Feature explanation
4. Information value plots
5. Model explanation - split and compare quantiles
6. Explainable models - Generalized Additive Models (GAM)
7. Counterfactual explanation

Revise:

- **Chapters from book**
- **SHAP**
- **LIME**
- **PDP**
- **Information Value**


FEEL FREE
makeameme


YOU CAN CALL ME


DID YOU REALLY JUST SEND THAT EMAIL!


KNOCK, KNOCK!