Responsible AI

Lecture 7

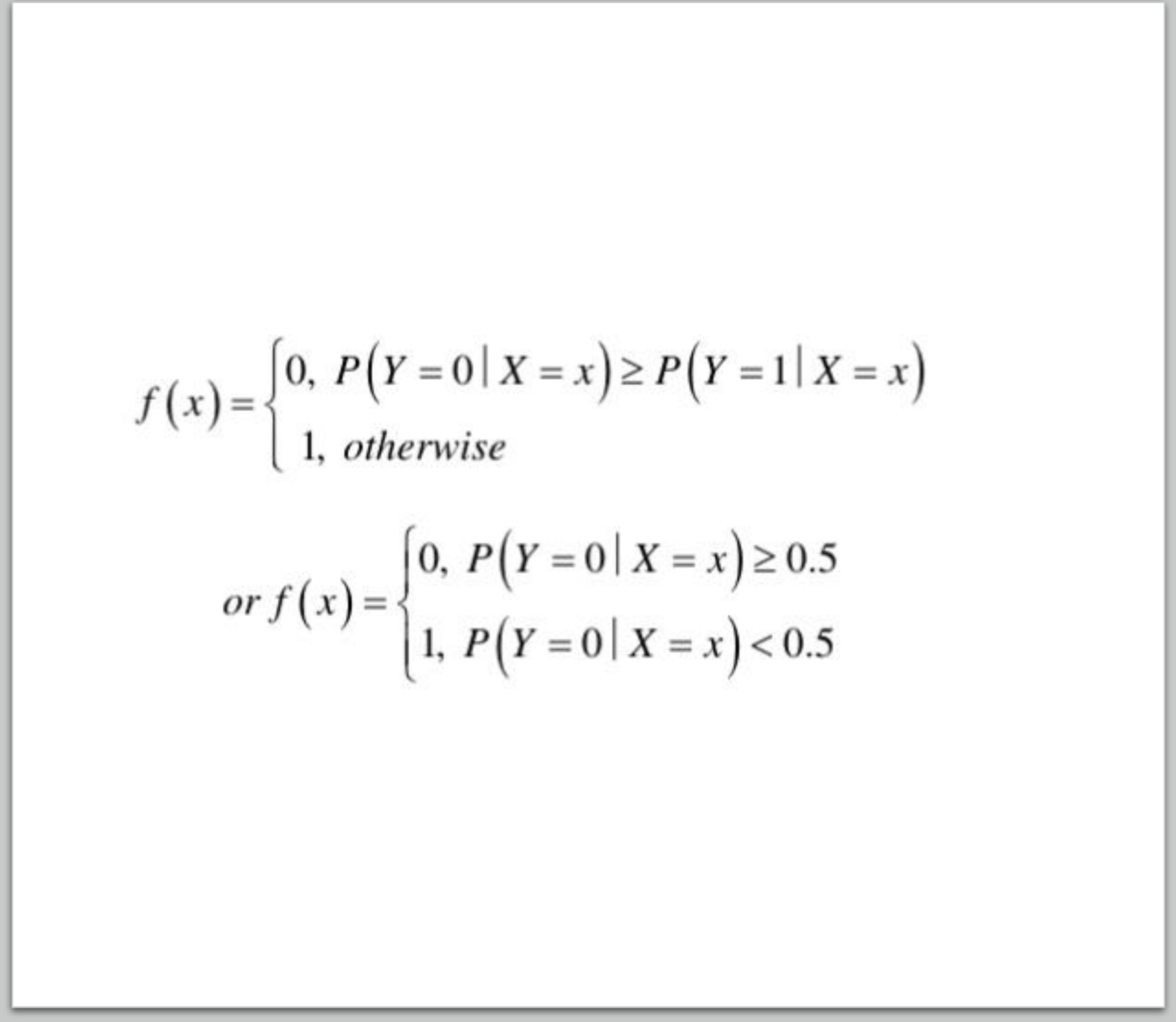
Today

1. Reject option classifier
2. Optimising the ROC
3. Handling multiple features in ROC

What is decision boundary in classification

Decision Boundary

ROC applies to two class classification (binary) problems. A binary classification algorithm would yield a result where a classifier f(X) returns an output (ˆY) between 0 and 1 (predicted probabilities). Conventionally, the decision boundary for the decision is at P(Y = 0 |X = x) = 0.5.



$$f(x)=\begin{cases}0, & P\left(Y=0\mid X=x\right)\geq P\left(Y=1\mid X=x\right)\\ 1, & \textit{otherwise}\end{cases}$$

$$\textit{or}\ f(x)=\begin{cases}0, & P\left(Y=0\mid X=x\right)\geq 0.5\\ 1, & P\left(Y=0\mid X=x\right)<0.5\end{cases}$$

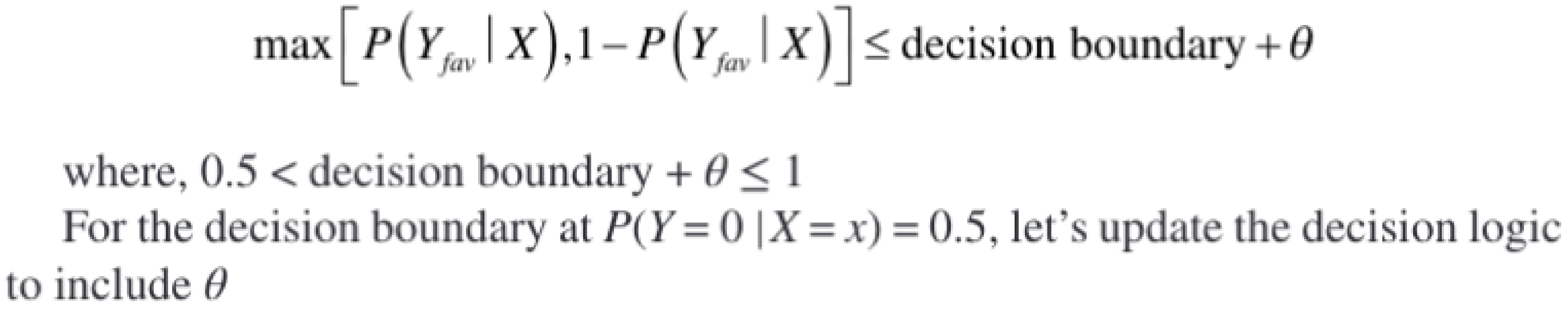
I don’t Know

The decisions can be placed into two broad categories: hard and soft. The hard decisions are the ones where the prediction probability is sufficiently far away from the decision boundary for us to make a confident position, e.g. $P(Y = 0 X = x) = 0.1$ or $P(Y = 0 X = x) = 0.8$, while the soft decision is the one where the prediction probability is too close to make a confident decision, e.g. $P(Y = 0 X = x) = 0.45$ or $P(Y = 0 X = x) = 0.51$.

Reject Option Classifier for Fairness

- It is important to note that the decision boundary to define the hard and soft pre-dictions would depend on the problem you are going to solve.
- It is also an important lever available to you as you utilize the ROC to reduce the bias from the predictions your model has made.
- The critical region centres around the decision boundary and covers the range of the soft decision.
- The distance of the edge of the critical region, or the range of the reject option, from the decision boundary will be denoted by θ.
- The critical region plays an important role in reducing the discrimination and is another important lever available to you.

Formulate ROC



$$\max\left[P\left(Y_{fav} \mid X\right), 1 - P\left(Y_{fav} \mid X\right)\right] \leq \text{decision boundary} + \theta$$

where, $0.5 < \text{decision boundary} + \theta \leq 1$

For the decision boundary at $P(Y = 0 \mid X = x) = 0.5$, let's update the decision logic to include θ

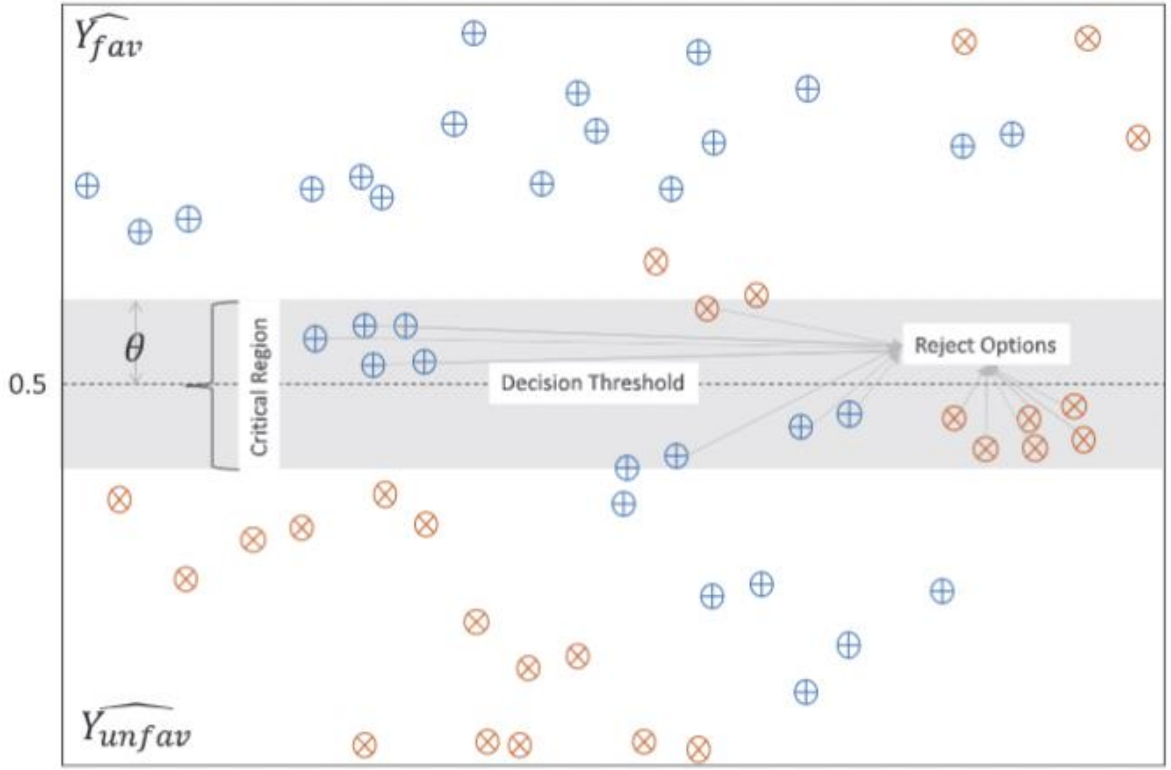


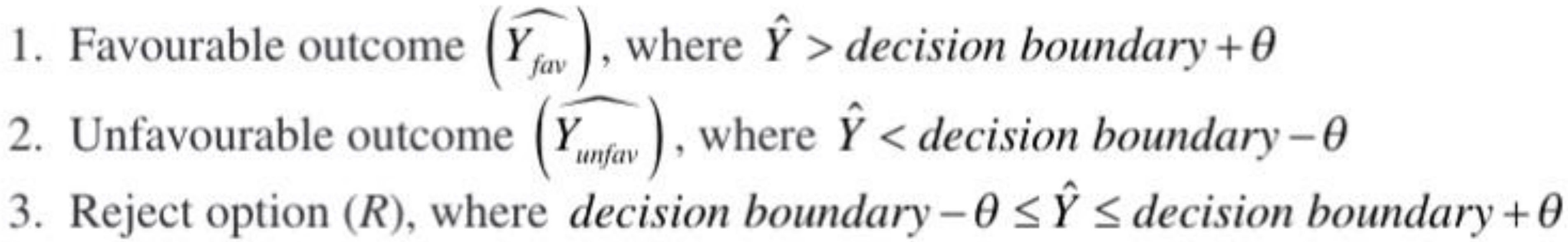
Fig. 6.1 Image depicting parameters of ROC

$$f(x)=\begin{cases}0, & P\left(Y=0 \mid X=x\right)>0.5+\theta \\ 1, & P\left(Y=0 \mid X=x\right)<0.5-\theta \\ R, & 0.5-\theta \leq P\left(Y=0 \mid X=x\right)\leq 0.5+\ \theta\end{cases}$$

In order to remove discrimination, we would like to minimize the following, within the constraints of the business objectives for the problem.

$$\left| P\left(\hat{Y} = \widehat{Y_{fav}} \mid S = S_a\right) - P\left(\hat{Y} = \widehat{Y_{fav}} \mid S = S_d\right) \right|$$

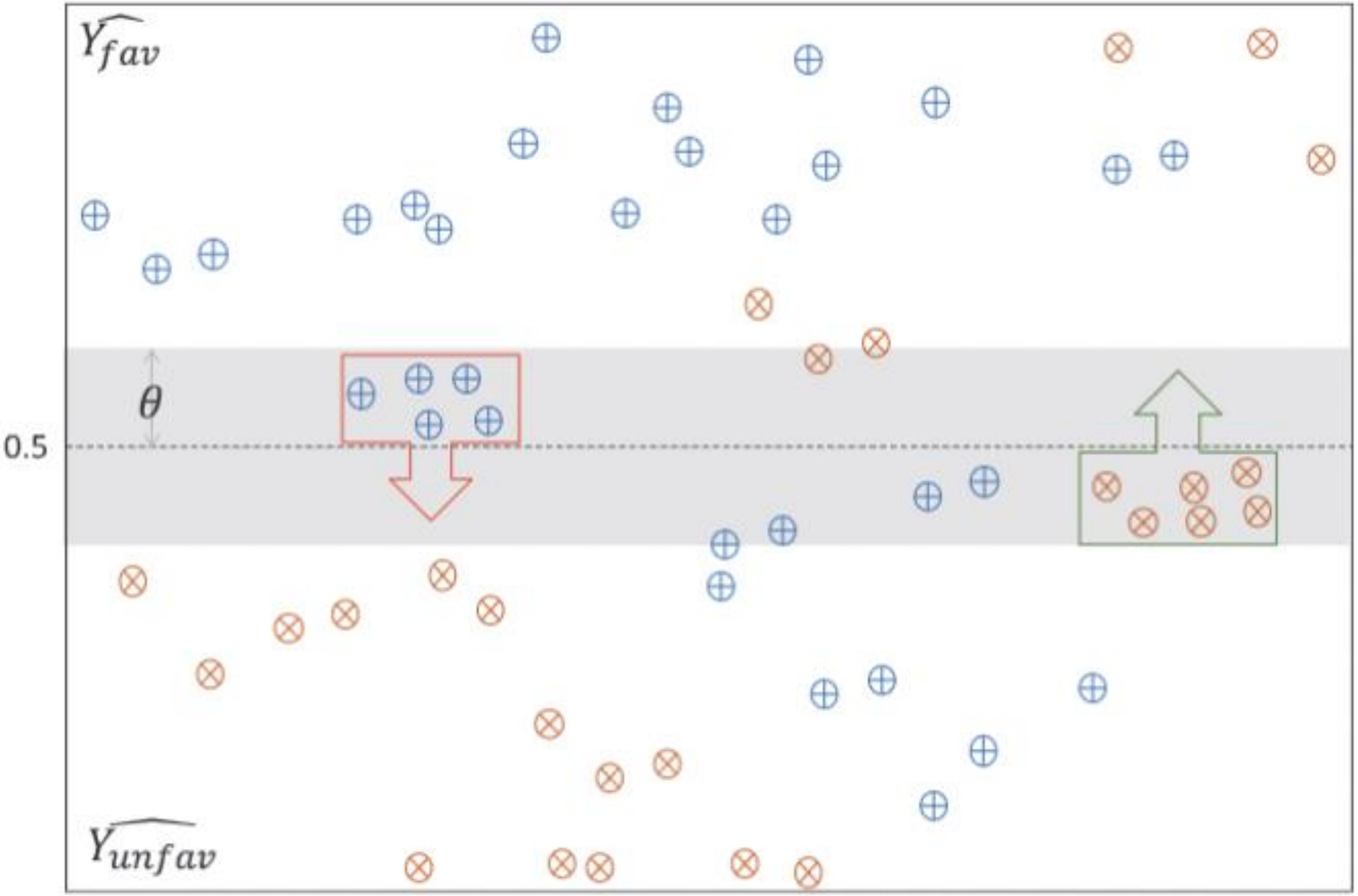
- If $P(Y_{fav}| X)$ is higher (either close to 1 or 0), then the label assigned would be certain and hard (or vice versa – in case of probability near 0.5).
- This means the classifier is very certain about the prediction when its falls into the hard boundary region. But the same can’t be concluded for the observations falling into soft boundary region.



1. Favourable outcome $\left(\widehat{Y_{fav}}\right)$, where $\hat{Y} > \mathit{decision\ boundary} + \theta$
2. Unfavourable outcome $\left(\widehat{Y_{unfav}}\right)$, where $\hat{Y} < \mathit{decision\ boundary} - \theta$
3. Reject option (R), where $\mathit{decision\ boundary} - \theta \leq \hat{Y} \leq \mathit{decision\ boundary} + \theta$

AGREE?

- For all predictions within R, ROC takes the action to promote or penalize based on the group that the record belongs to.
- Prediction that are below the decision boundary and belong to the unprivileged group are promoted above the decision boundary, whereas the predictions that are above the decision boundary and belong to the privileged group are penalized and moved below the decision boundary.
- This results in “flipping” the decision for the two groups.


$\widehat{Y_{fav}}$
θ
0.5
$\widehat{Y_{unfav}}$

Steps

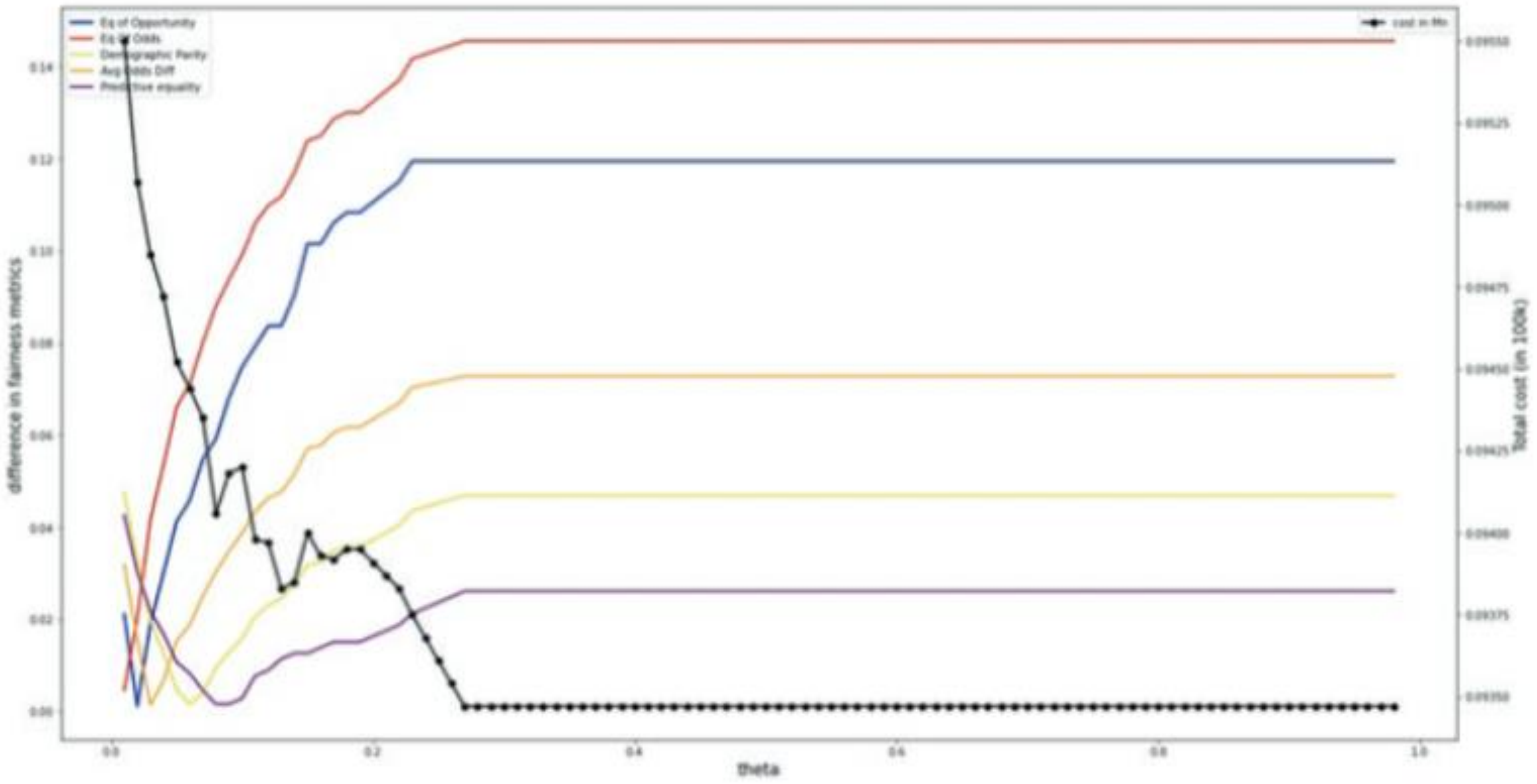
- Get actual labels, predicted labels and predicted probabilities
- Find out the relevant protected features that need treatment
- Create composite protected features, if necessary
- Create copy of predictive probabilities
- Filter the data falling in critical region
- Filter the data for each of four protected and predicted class combination
- Swap the probabilities
- Check the model accuracy after ROC treatment
- Check the fairness accuracy
- Optimize the critical region to reduce accuracy-fairness trade-off.

What are few advantage of ROC?

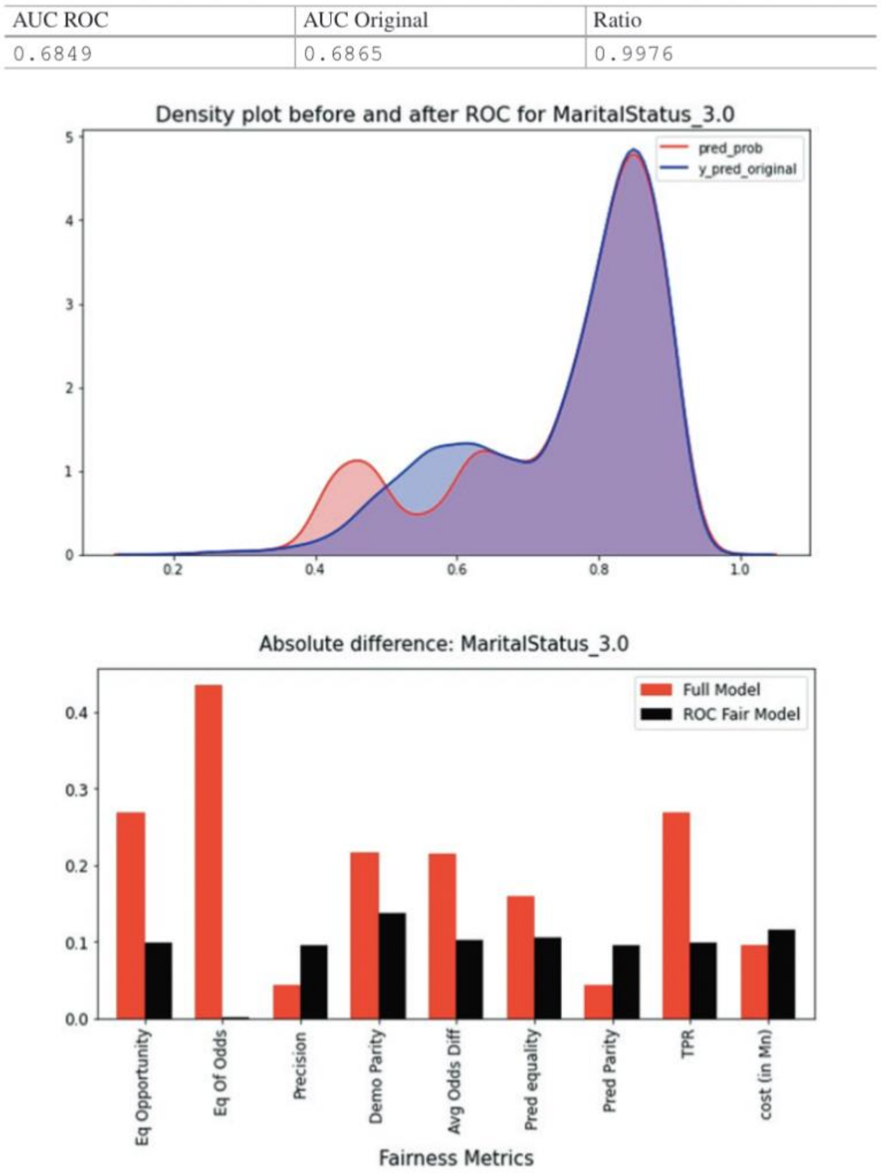
Tuning

- Decision boundary
- Theta
- Non-uniform theta
- Promote or penalize or both

ROC optimization for various value of theta MaritalStatus_1.0



Observations



AUC ROC	AUC Original	Ratio
0.6849	0.6865	0.9976

Handling Multiple Features in ROC

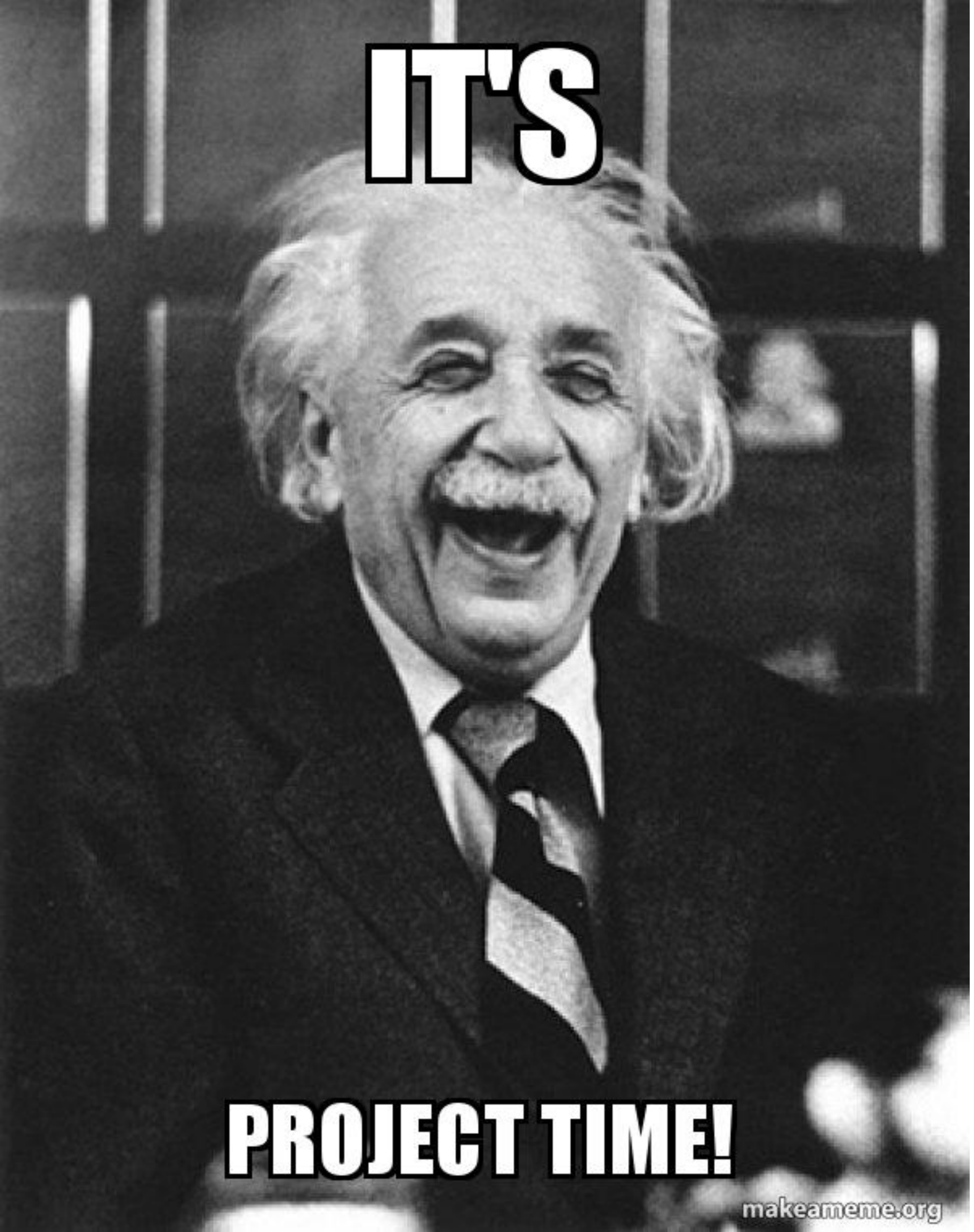
Sequentially treat multiple features on the ROC treatment of previous protected feature

Go for composite feature.

Can we keep the distance intact?

If you have two records with probabilities 0.41 and 0.49, they become 0.59 and 0.51. While the outcome for both of them is now favourable and the same, the one that had a higher score before ROC was applied now has lower score. This is an unintended consequence of the process we have applied

we can take to flip the records to the other side of the decision boundary while keeping their relative positions intact is by adding or subtracting θ from the predicted probability depending on whether we are promoting or penalizing the record. Thus, when promoting a record, we will change the probability to (ˆY−θ), whereas when penalizing it will change to ˆY+ θ

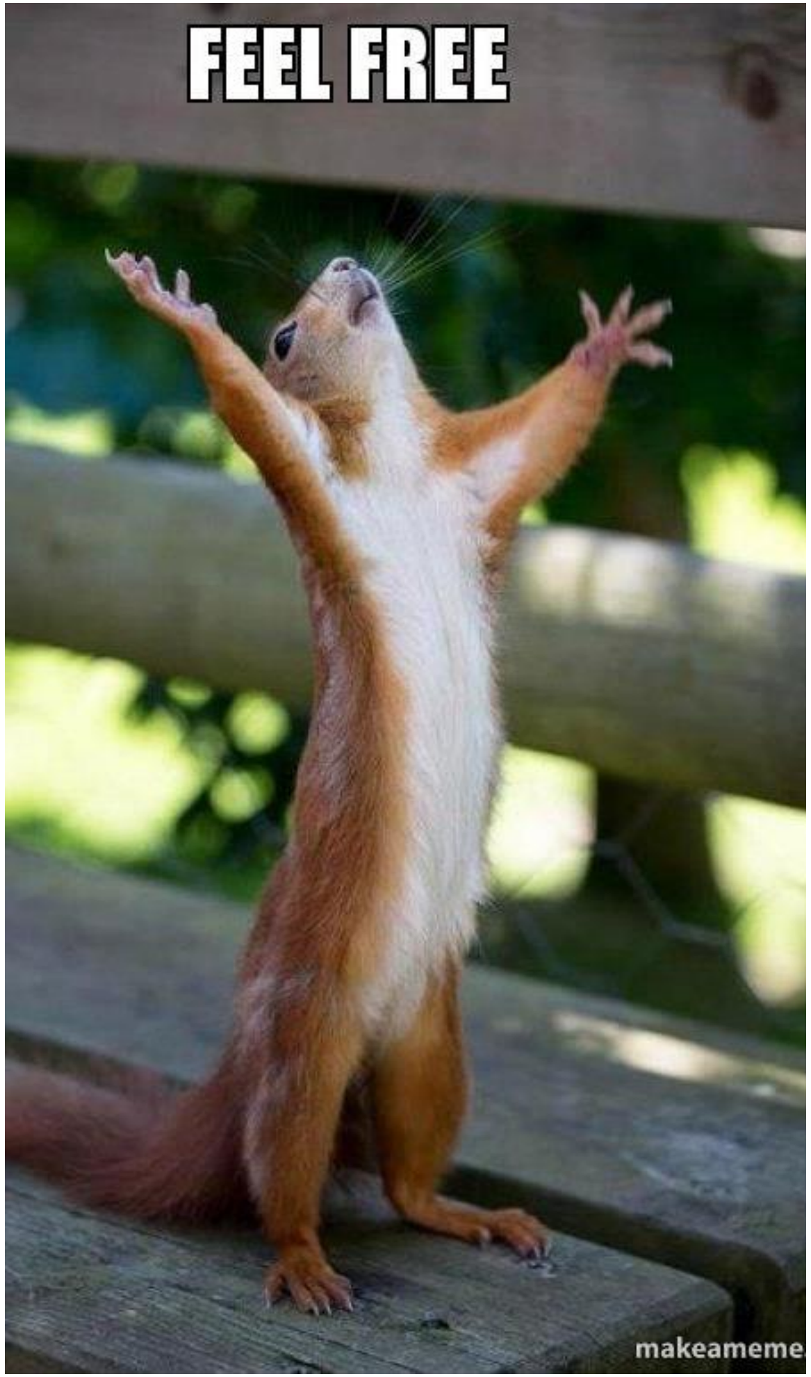


3rd of 3 parts of final project

Implement Reject Option on the output of:

1. Original model
2. Model post reweighting
3. Model post ACF
4. All of the above on DP data and models
5. Report difference and impact

NO PREP FOR YOU!


FEEL FREE
makeameme

YOU CAN CALL ME

DID YOU REALLY JUST SEND THAT EMAIL!

KNOCK, KNOCK!