

AUROC difference in PD classification after counterfactually removing protected attributes

