Evaluating the Fairness Impact of Differentially Private Synthetic Data

Blake Bullwinkel[1], Kristen Grabarz[1], Lily Ke[1], Scarlett Gong[1], Chris Tanner[1], Joshua Allen[2]
[1]IACS, Harvard University, [2]Microsoft

ICML International Conference On Machine Learning

The Problem

As data containing sensitive information on individuals are being collected in an increasing number of domains, differentially private (DP) synthetic data has emerged as a promising approach to maximizing its utility. However, the fairness implications of training ML models on DP synthetic data are not well understood.

Objectives

Due to the suppression of underrepresented classes that is often required to achieve privacy, it may be in conflict with fairness. With this knowledge, we have two main research objectives:

1. Understand the fairness impact of differential privacy in the context of DP synthetic data by evaluating various synthesizers at a range of privacy budgets.
2. Propose a pre-processing method to mitigate bias in DP synthetic data while retaining predictive accuracy.

Evaluation Metrics

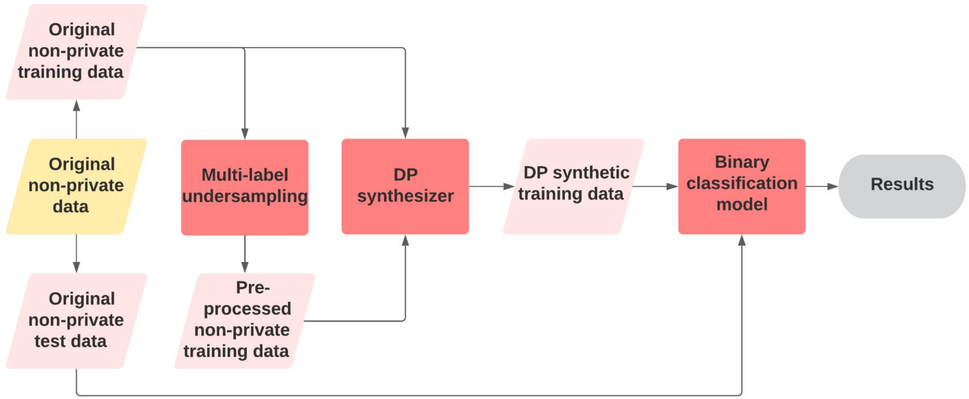
Fairness	Equalized Odds Distance	$\delta_y = Pr[\hat{Y} = 1 \mid A = 0, Y = y] - Pr[\hat{Y} = 1 \mid A = 1, Y = y], \quad y \in [0, 1]$
Accuracy	F1 score	$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$

Datasets

	Adult	ACS Income	COMPAS
Protected att.	Gender	Gender	Race
Example features	Age, work class, education, occupation, race, hours worked, etc	Age, work class, education, occupation, race, hours worked, etc.	Age prior counts, sex, juvenile felonies / misdemeanor, charges
	Income >=$50k	Income >=$50k	Recidivism

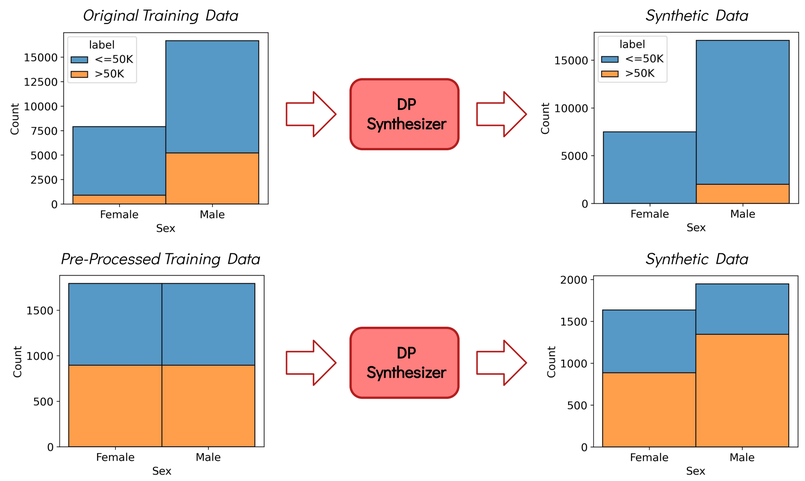
Data Pipeline

We evaluated the fairness outcomes of binary classifiers trained on DP synthetic data. Synthesizers were trained on the original or pre-processed data, while classifiers were trained on synthetic data and tested on the original, non-private data. The flow chart below illustrates our pipeline, which was automated to run across datasets.



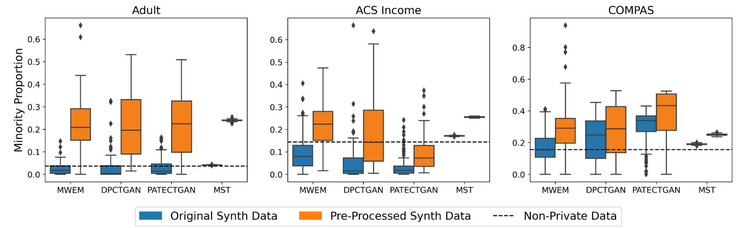
Pre-Processing

We observed that DP synthesizers tended to exacerbate pre-existing class imbalances in the original data, leading to less fair predictions by binary classifiers. To mitigate this issue, we pre-processed the data synthesizers were trained on using a multi-label undersampling technique, illustrated below.



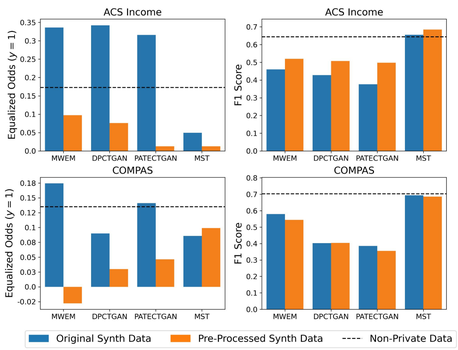
Results

We first analyzed the proportion of the minority group present in the synthetic data sets. QUAIL-MWEM, QUAIL-DPCTGAN, and QUAIL-PATECTGAN frequently decreased this proportion relative to the non-private data, as shown below. Interestingly, MST did not appear to suffer from this issue.



We observed a strong association between the minority group proportions visualized above and the downstream fairness outcomes reflected below. Synthetic data sets with lower minority proportions than non-private data are associated with less fair outcomes, while synthetic data sets that do not decrease this proportion are less likely to degrade fairness.

Conversely, our pre-processing method is also associated with lower equalized odds distances (more fair outcomes). Similarly, MST did not significantly alter the minority group proportions in comparison to non-private data and therefore did not degrade fairness.



Conclusions

- QUAIL-MWEM, QUAIL-DPCTGAN, and QUAIL-PATECTGAN frequently degraded fairness outcomes on binary classification tasks.
- We found an association between less fair outcomes and decreased proportions of minority groups in the synthetic data.
- Our pre-processing technique mitigated unfair outcomes without sacrificing predictive accuracy.
- The MST synthesizer achieved fairness and accuracy metrics that were close to those obtained on non-private data and may be a preferable option for real-world applications involving DP synthetic data.

Theory and Practice of Differential Privacy (TPDP), ICML, 2022