Fairness Impact of Privacy

Milestone #1 Presentation

Blake Bullwinkel, Scarlett Gong, Kristen Grabarz, Lily Ke
October 4, 2021

Problem statement

Original Data	
Age	State
23	NY
47	NE
35	NY
29	CT
...	...
52	CT

Average Age: 44
State: 0.8% NE

+ Noise

Synthesis

Private Data	
Age	State
24	NY
45	NY
33	NY
31	CT
...	...
51	CT

Average Age: 45
State: 0% NE

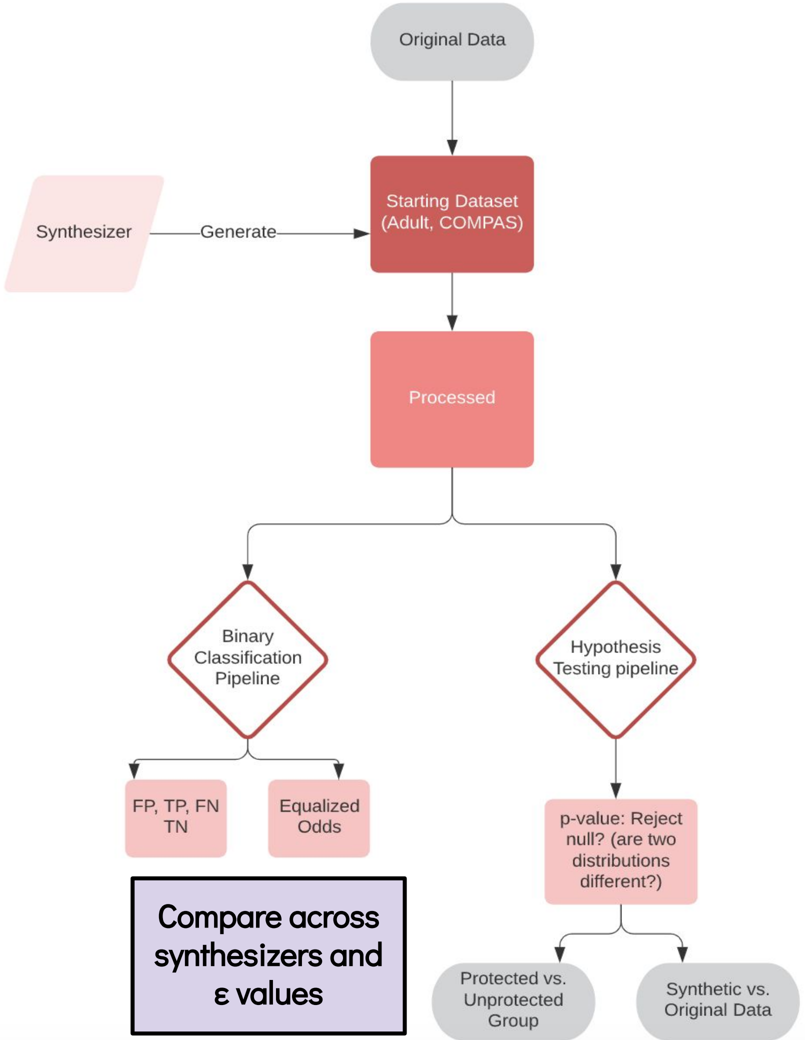
Differential privacy protects sensitive information by adding noise to data. However, it can have a disparate impact on model accuracy.

Our goal is to understand how changing ε (privacy loss) across various differentially private synthesizers affects our ability to achieve “fair” outcomes.

Scope of work

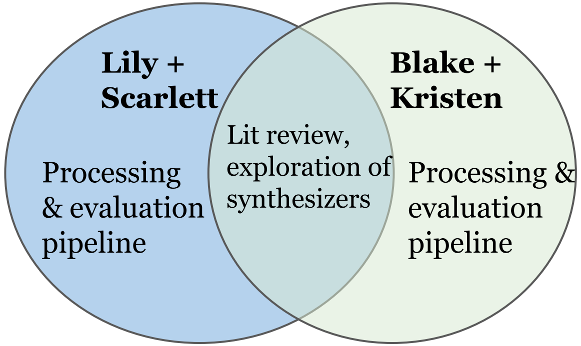
Using 3 initial datasets popular in the fairness literature, we will:

- **Generate** synthetic datasets using 2+ synthesizers and 8 ε values.
- **Perform** tasks including binary classification and hypothesis testing.
- **Measure and compare** fairness outcomes across these variants to understand the tradeoff between privacy and fairness.



Team & collaboration infrastructure

- To streamline our progress, we are operating on sub-workstreams as pairs.
    - Built processing and evaluation pipelines for respective datasets
- Tooling:
    - **Deepnote:** Initial exploratory data analysis and pipeline development.
    - **SmartNoise synthesizers:** Require dependencies that are better suited to install and run locally.
    - **Github:** Code sharing for synthesis and downstream evaluation.



Lit review: Differential privacy & fairness

2006

***Differential Privacy* by Cynthia Dwork**

Cynthia Dwork formalizes differential privacy, which bounds privacy loss and ensures that the output of a computation does not leak information about any individual

2012

MWEM synthesizer released

The first popular differentially private data synthesizer is released

2015

Differentially private models and tradeoffs

A great deal of research is focused on developing differentially private implementations of machine learning models and balancing privacy with predictive accuracy

2019

The impact of privacy on accuracy is not borne equally

Bagdasaryan et al. show that differentially private ML models may be less fair

2020

SmartNoise releases multiple synthesizers

SmartNoise, a collaboration between OpenDP (Harvard) and Microsoft, releases implementations of MWEM and multiple GAN-based synthesizers

2021

Harvard IACS

An ambitious group of IACS students sets out to characterize the interplay between differentially private synthetic data and fairness

Lit review: Fairness data sets

Protected Attributes:

Based on the literature, we identify the following protected attributes in our data:

a. **Adult:**
   i. Gender (male: privileged; female: unprivileged);
   ii. Race (white: privileged; nonwhite: unprivileged)
b. **COMPAS:**
   i. Race (white: privileged; nonwhite: unprivileged)
c. **German Credit:**
   i. Gender (male: privileged; female: unprivileged);
   ii. Age (> 25: privileged; < 25 : unprivileged)
   iii. Nationality (non-foreigners: privileged; foreigners: unprivileged)

Lit review: Fairness metrics

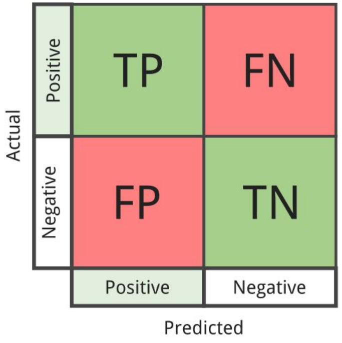
Core Fairness Metrics:

a. **Binary Classification:**
   i. Equalized Odds Distance:

$$\delta_y = \Pr(\hat{y}=1|A=0,Y=y) - \Pr(\hat{y}=1|A=1,Y=y),\ y \in \{0,1\}$$

   ii. Demographic/statistical parity

b. **Hypothesis Testing:**
   i. Method: Difference in proportions hypothesis testing
   ii. Target outcomes across protected / unprotected groups
   iii. Target outcomes across original versus synthetic data for protected and unprotected groups



DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., & Agarwal, D. (2020, August). Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1467-1477).

Lit review: Differentially private synthesizers

1. MWEM (2012): simple but effective with shorter runtime
2. PrivBayes (2014): developed by dataResponsibily
3. GAN-based (2018-2019): based on GAN architecture, privatized by DPSGD
    a. PATE-GAN
    b. DPGAN
    c. DP-CTPGAN
4. FFPDG (2021): “native fair” synthesizer developed by Amazon

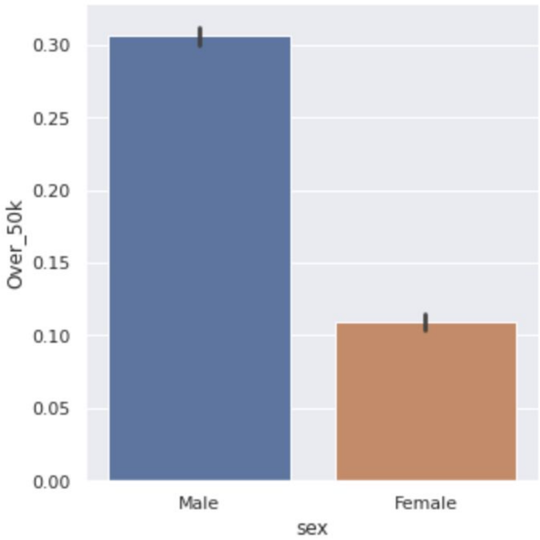
Project ideas

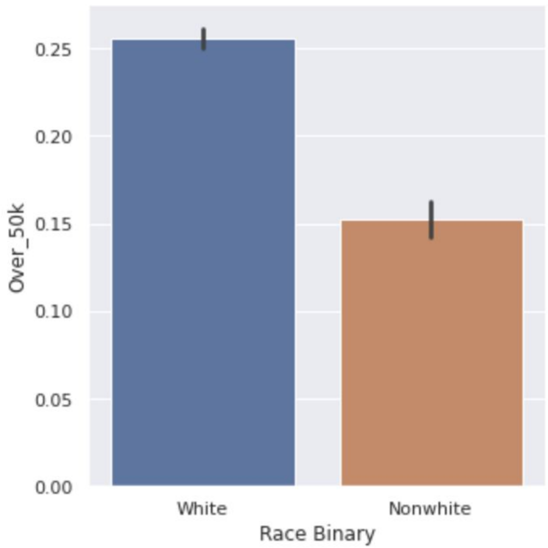
Milestone 1	● Literature review (fairness definition, fairness metrics) ● EDA on three datasets - Adult, COMPAS, and German Credit ● Prepare fairness evaluation pipelines ● Explore SmartNoise synthesizers
Milestone 2	● Use various synthesizers to generate synthetic data and apply them through the established pipelines ● Compare and analyze results and understand the conditions that lead to good or bad results
Milestone 3	● Recommend pre/post-processing steps that mitigates the bias we observe ● Gain a deeper understanding on the tradeoff between privacy and fairness
Final deliverable	● Wrap up and write paper ● Prepare the final presentation

Learning Goals

Milestone 1	Literature review ● fairness definitions & fairness metrics ● Adult, COMPAS, German Credit datasets
Milestone 2	SmartNoise synthesizers and DP algorithms ● GAN algorithm ● privacy loss ε ● tradeoff between fairness and privacy
Milestone 3	Existing pre/post-processing bias mitigation algorithm ● reweighting ● optimized pre-processing ● adversarial debiasing ● reject option based classification
Final deliverable	Improve our academic research skills

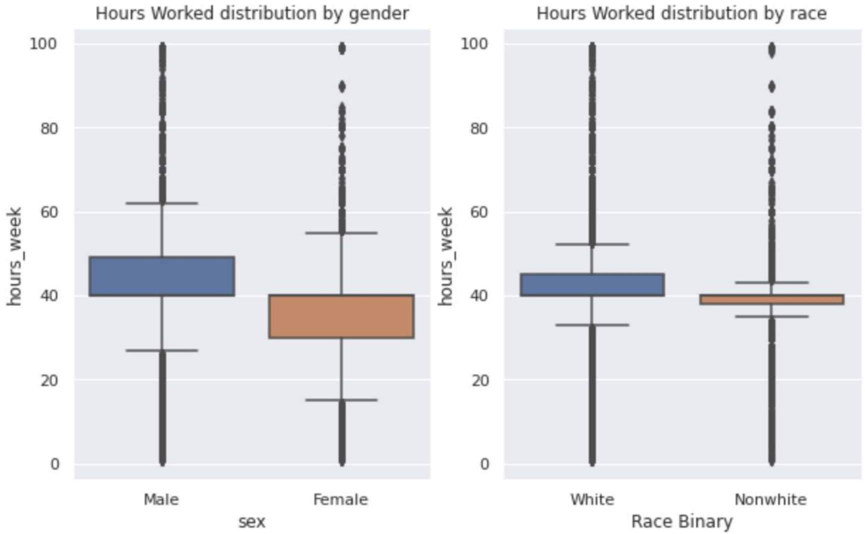
Adult data set: EDA

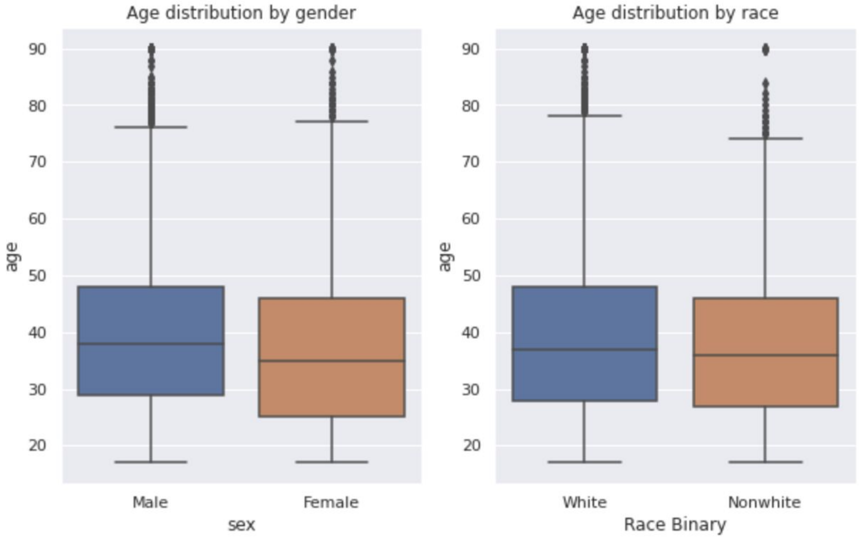




Both unprotected groups (Women and Nonwhite individuals) are less likely to make an income of at least $50k

Adult data set: EDA

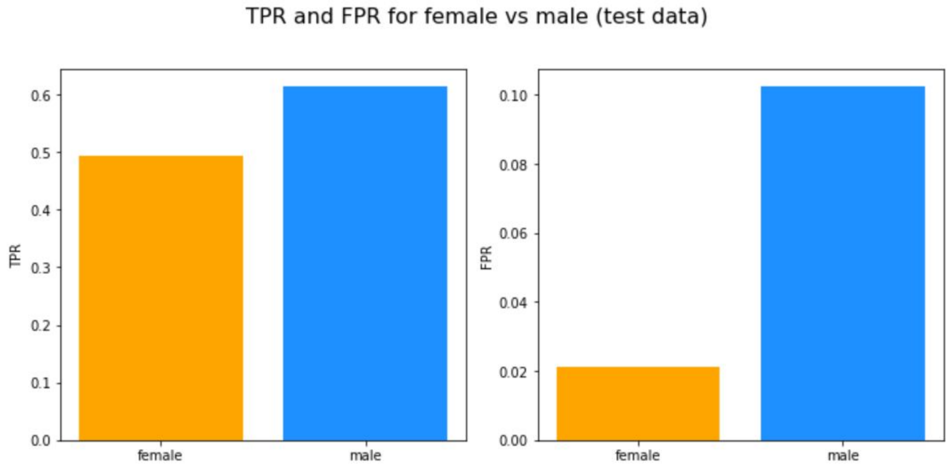




Women and Nonwhite people tend to work fewer hours per week and, especially for women, they appear to be slightly younger.

Adult: Binary Classification Pipeline & Results

We confirmed that men is the privileged class and has higher TPR and FPR than women, both of which are associated with favorable outcomes in the Adult data set.

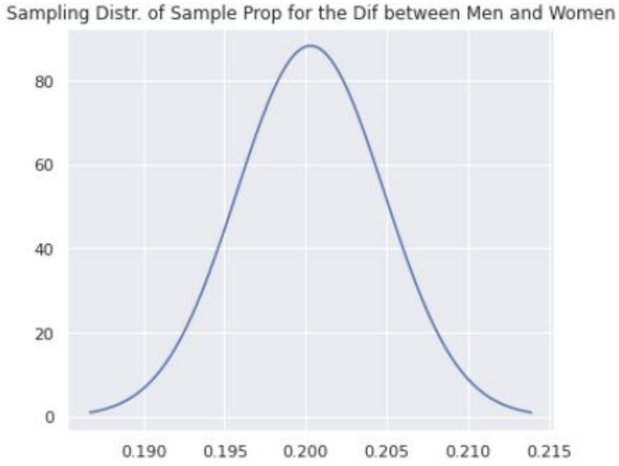


Test Set Fairness Metrics	**Gender**
Equalized Opportunity (TP rate difference)	0.120
Equalized Odds (FP rate difference)	0.081
Demographic Parity (FP+TP rate difference)	0.201

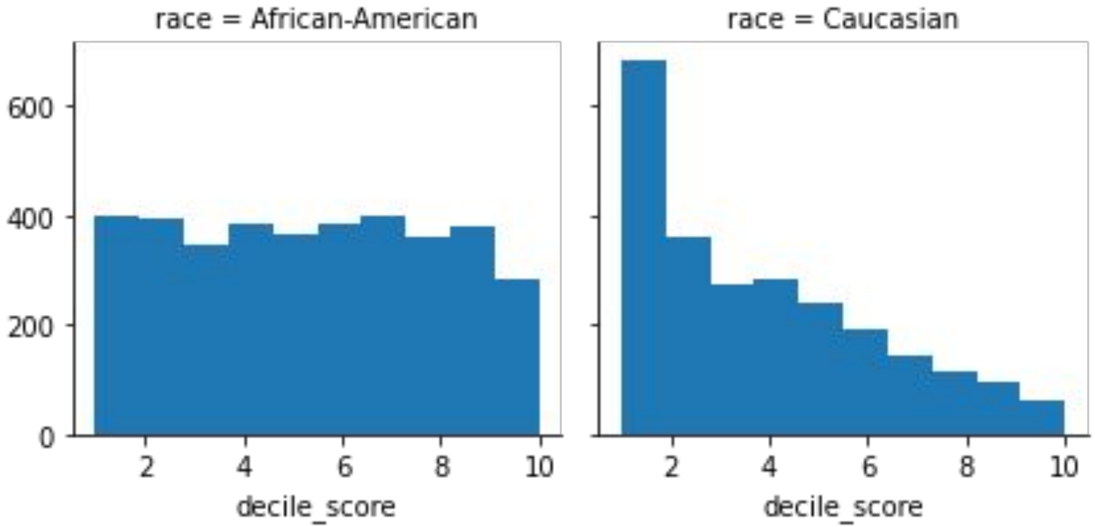
Adult: Hypothesis Test Pipeline & Results

- Compared rate of favorable outcome (>$50k) across protected versus unprotected group: Men versus Women

- **Reject the null hypothesis of no difference**:
    - Men significantly more likely than women to yield a positive outcome
    - Plan to expand comparison to original versus synthetic data



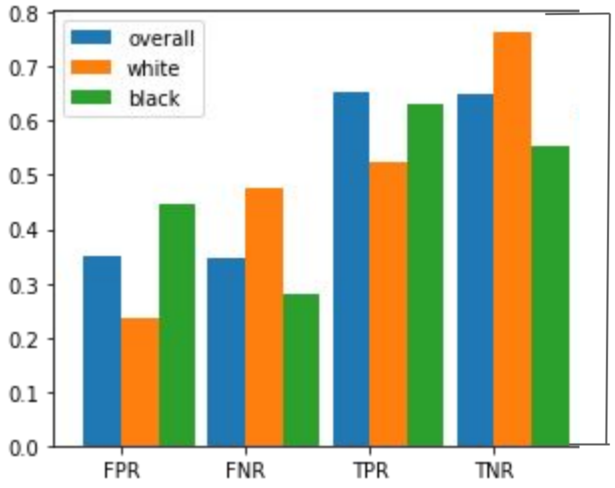
COMPAS dataset: EDA



Histogram of decile_score provided by COMPAS tool

Plotting the decile scores produced by COMPAS tool as a prediction score, the distribution for white individuals is right-skewed

COMPAS dataset: EDA



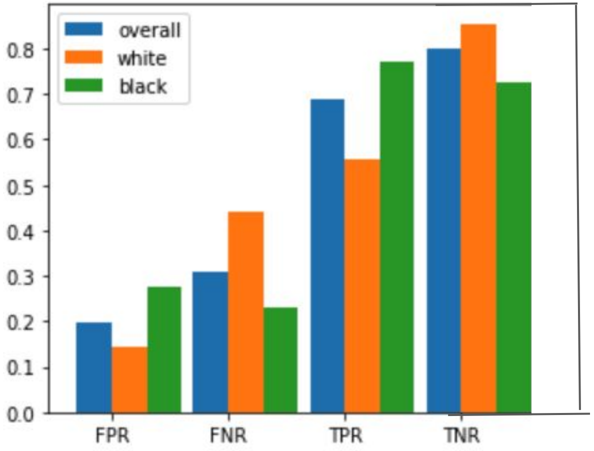
decile_score provided by COMPAS tool

- False Positive Rate:
  - White: 23.5%
  - Black: 44.9%
- False Negative Rate:
  - White: 47.7%
  - Black: 28.0%

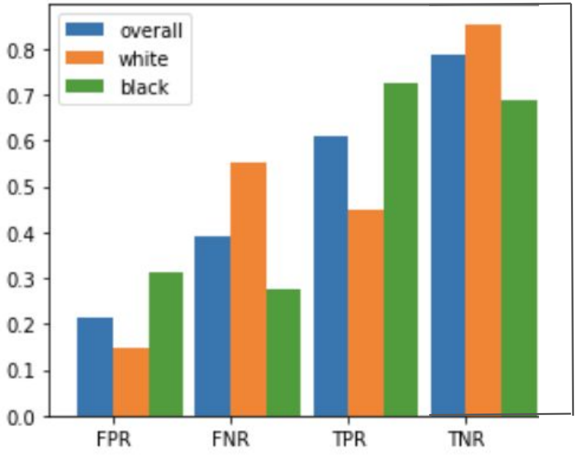
Black individuals have a higher FP rate and lower FN rate than white people.

COMPAS: Binary Classification Pipeline & Results

This shows the classifications appeared to favor white defendants over black defendants by underpredicting recidivism for white and over predicting recidivism for black defendants.



Logistic Regression



Decision Tree

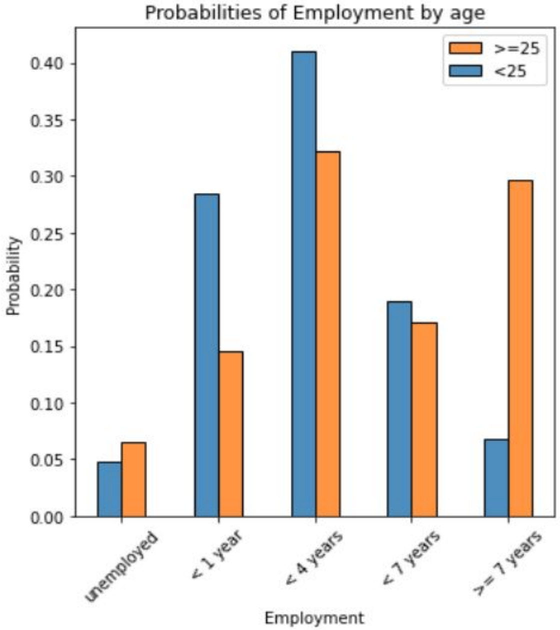
COMPAS: Hypothesis Test Pipeline & Results

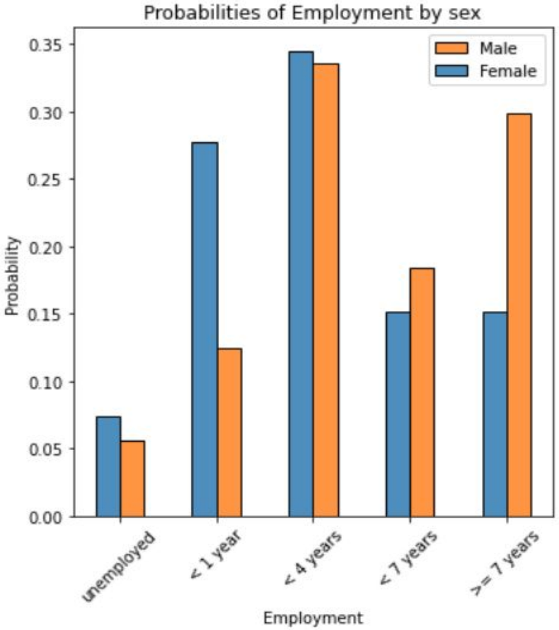
Compared recidivism rate across protected versus unprotected group: African American versus Caucasian individuals.

Reject the null hypothesis of no difference:

- Mean of the African American predicted recidivism rate > the mean of the Caucasian predicted recidivism rate
- Mean of the African American predicted recidivism rate > the mean of the African American real recidivism rate

German Credit dataset: EDA





People who are older and Male tend to have longer employment histories that can influence their likelihood of obtaining credit.

German Credit: Binary Classification Pipeline & Results

- Equalized Opportunity

  the TP rates for different values of the protected attribute should match

- Equalized Odds

  the TP and FP rates for different values of the protected attribute should match

	Gender	Age	Nationality
Equalized Opportunity (TP rate difference)	0.035	0.084	0.108
Equalized Odds (FP rate difference)	0.239	0.011	0.102

German Credit: Hypothesis Test Pipeline & Results

- Compared rate of credit risk across protected versus unprotected group:
  Men versus Women, Young and Older population(age >25)

- **Reject the null hypothesis of no difference**:
  - Men significantly more likely than women to have a good credit risk
  - Younger population is more likely to have a bad credit risk

Thank you!