Algorithm Fairness in Lending

3/4/21 6:18 PM

Speaker: Adam Lieberman (Finastra)
Date: 03/04/2021

AI in Lending

- Automated and more accurate decisioning
- Automatically extract data from various forms to prepopulate front end systems for the end user
- Know your customer
- Credit invisible

As ML engineers, the thought about bias and fairness is often not taken into consideration

What is fairness

In ML, an algorithm is fair if its result are independent of given variables, especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome

Individual Fairness

- Seeking to ensure that statistical measures of outcomes are equal for similar individuals
- Similar individuals are treated similarly

Group Fairness

- Partitioning a population into groups predefined by protected attributes and seeking to ensure that statistical measures of outcomes are equal across groups
- Making model's predictions/outcomes equitable across groups

Two Worldviews

- We're all equal
- What you see is what you get

Why Fairness is Important

- Customers are aware - when customer trust is lost, there's no guarantee we can get it back
- Feedback loops - Decisions made by unfair algorithms are propagated as training data

Causes of Unfairness

Inherent Data Bias

- Data can be inherently biased without intention
- If we train on biased data we should expect biased results

Distorted Representation

Data not representative of the target population

- Missing data
- Sample bias
- Exclusion bias
  - Deletion of valuable data though to be unimportant
- Measurement bias
  - Faulty measurements cause collected data to not represent our target population
- Label bias
  - Similar data points are labeled inconsistently

Algorithmic Objectives

Objective Function Modification

- When modeling we want to minimize errors and maximize performance
- Most predictive analysis is optimized around maximizing an objective function tuned for accuracy of the outcome
- Bias can stem from minimizing overall aggregated prediction errors by benefiting a majority group over a minority group unknowingly
- Need to define appropriate measures in terms of data and algorithmic objectives to ensure our performance makes sense to our minority groups
- Instead of just accuracy we can optimize for particular fairness metrics

Proxies

Links to Sensitive Data

- Non inclusive data
- Related fields
  - Related fields can serve as proxies to sensitive fields
  - There can be a correlation to or be predictors of sensitive fields
  - If related fields are used we can still introduce bias to our models

Measuring Fairness

Generic Metrics	Group Fairness Metrics	Individual Fairness Metrics
Specificity Score	Demographic Parity Difference	Generalized Entropy Index
Sensitivity Score	Mean Difference	Generalized Entropy Error
Base Rate	Disparate Impact Ratio	Theil Index
Selection Rate	Equal Opportunity Difference	Coefficient of Variation
Generalized FPR	Average Odds Difference	Consistency Score
Generalized FNR	Average Odds Error	+ More
	Between Group Generalized Entropy Error	

Demographic Parity

- $P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$
  - The proportion of each segment of a protected class should receive the positive outcome of equal rates
  - The difference in the groups should be close to 0
  - Each group of people should have the same percentage chance of receiving the loan
  - We might need different thresholds for the groups so the percentage of people in each group have an equal chance of getting a loan
  - When to use
    - We are aware historical biases may have affected our data quality
    - Or we decide to support unprivileged group

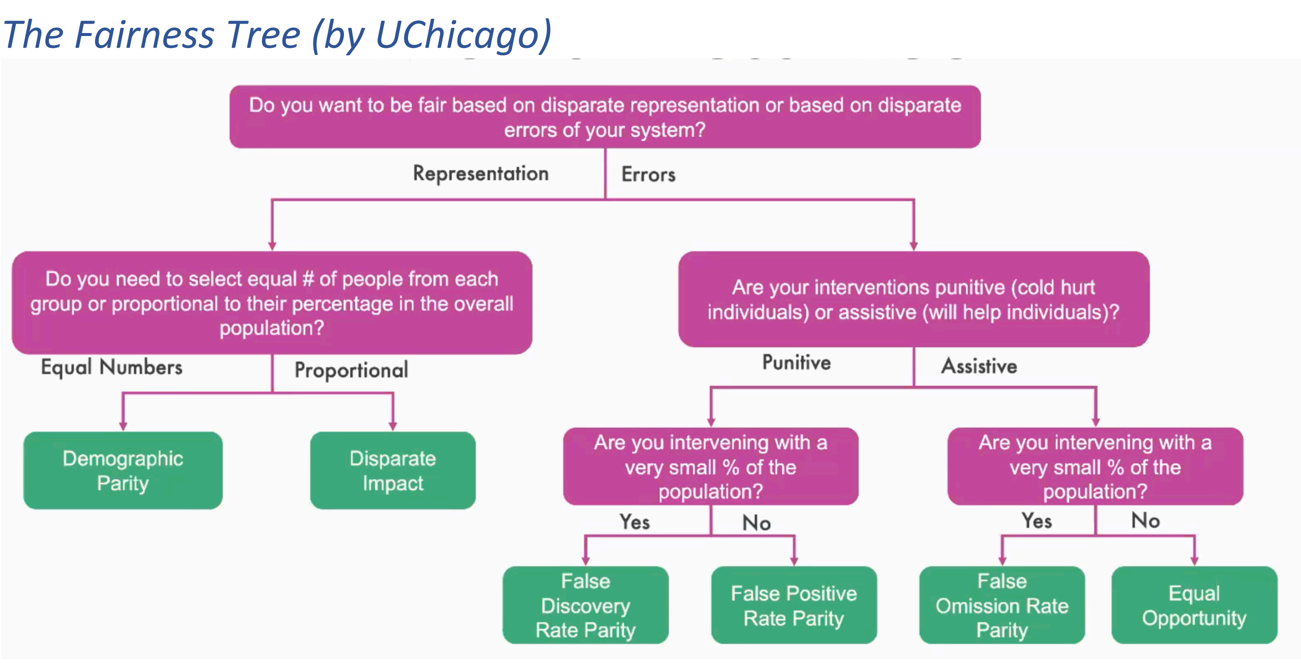
Equal Opportunity

- $P(\hat{Y}=1 \mid A=0,\ Y=1) = P(\hat{Y}=1 \mid A=1,\ Y=1)$
  - Each group should get the positive outcome at equal rates, assuming that people in their group qualify for it
  - Based on the ground truths, we look at our predictions for the positive binary outcomes and measure the probability across the groups
  - When to use
    - There is a strong emphasis on predicting the positive outcome correctly
      - E.g. we need to be very good at detecting a fraudulent loan application
    - Introducing False Positive are not costly to the user nor the company
    - The target variable is not considered subjective
      - E.g. labeling who is a "good" employee is prompt to bias and hence very subjective

Equalized Odds

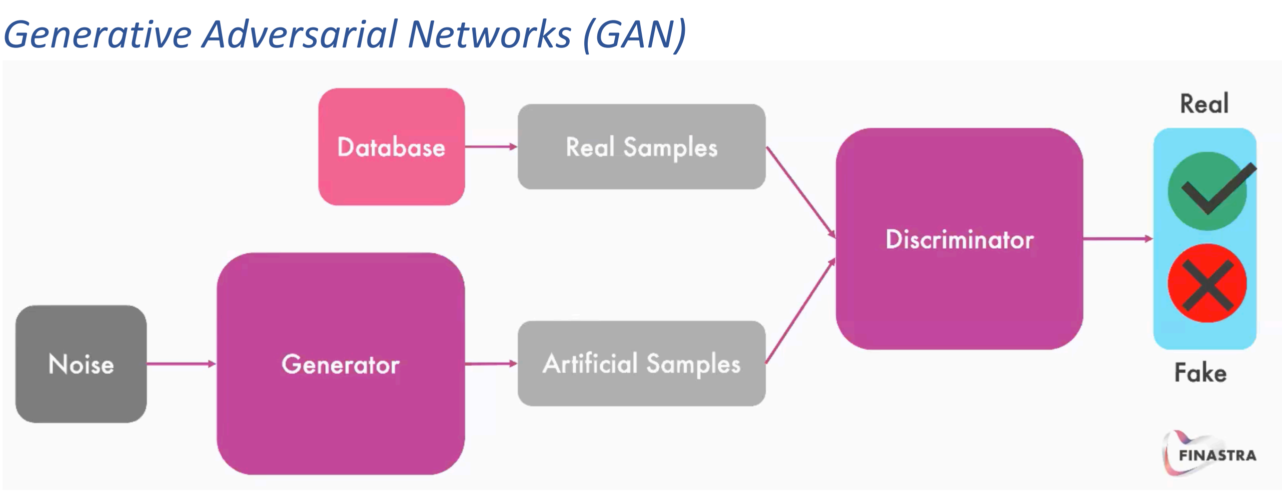
- $P(\hat{Y}=1 \mid A=0,\ Y=y) = P(\hat{Y}=1 \mid A=1,\ Y=y),\ y \in \{0,1\}$
  - A combination of statistical parity for true positives and false positives simultaneously
  - Equalized odds requires the positive outcome to be independent of the protected class A, conditional on the actual Y
  - Based on the confusion matrix, it requires the True Positive Rate (TPR) and False Positive Rate (FPR) to be the same for each segment of the protected class
  - When to use
    - There is a strong emphasis on predicting the positive outcome correctly
      - E.g. correctly identifying who should get a loan drives profits
    - We strongly care about minimizing costly False Positives
    - The reward function of the model is not heavily compromised
    - The target variable is not subjective

The Fairness Tree (by UChicago)



Adversarial Learning

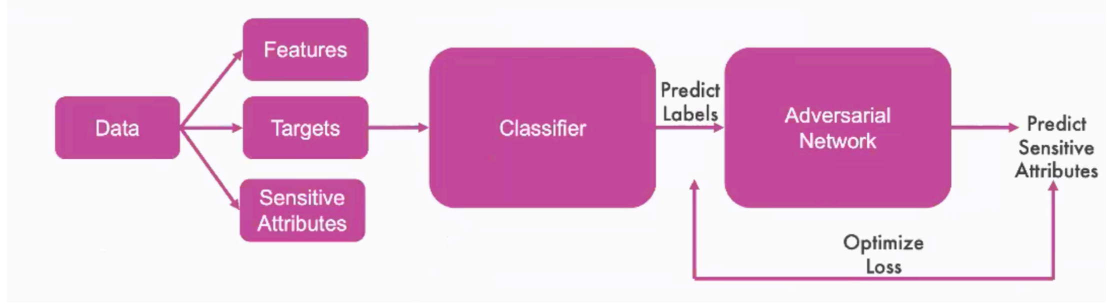
Generative Adversarial Networks (GAN)



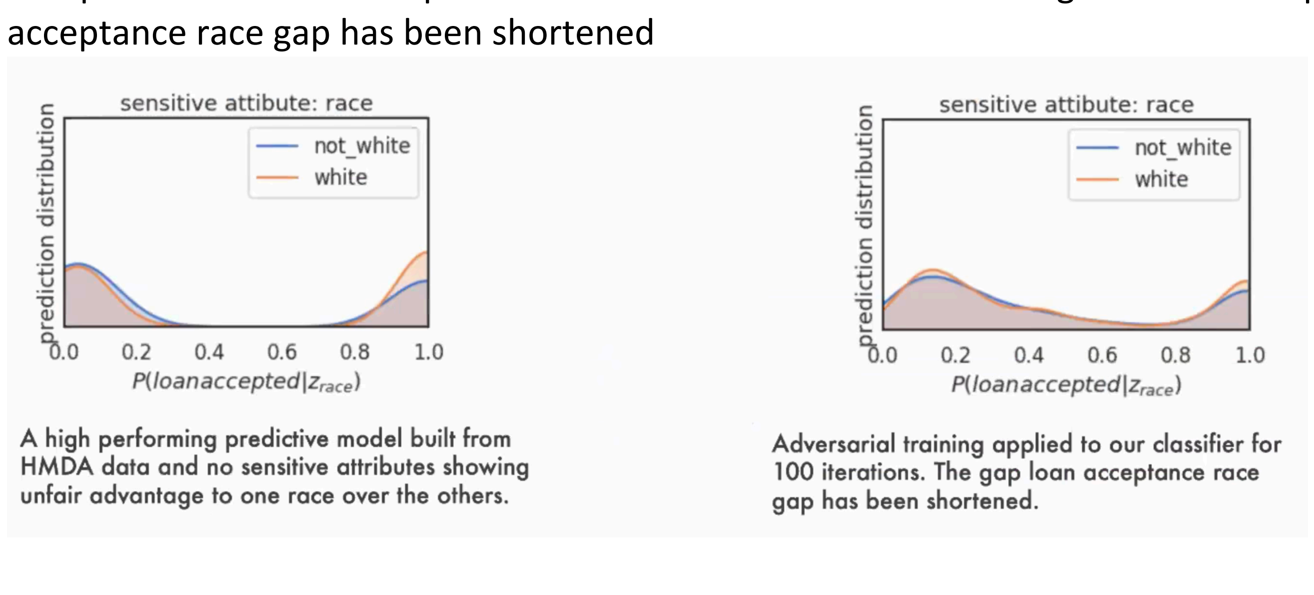
Adversarial Learning

The process:

- Pre-train the classifier on the features/targets
- Pre-train the adversarial on the predictions of the pre-trained classifier
- During T iterations simultaneously train the classifier and the adversarial
  - First train the adversarial for a single epoch and keep the classifier fixed
  - Train the classifier on a single sampled mini batch while keeping the adversarial fixed



- Comparison of traditional predictive model and adversarial training model. The tap loan acceptance race gap has been shortened



A high performing predictive model built from HMDA data and no sensitive attributes showing unfair advantage to one race over the others.

Adversarial training applied to our classifier for 100 iterations. The gap loan acceptance race gap has been shortened.