DATASCI207-005/007 Applied Machine Learning

Vilena Livinsky, PhD(c)

School of Information, UC Berkeley

Week 12: 04/02/2025 & 04/03/2025

Today’s Agenda

- Baseline Presentations
- Fairness
- Walkthroughs:
  - Fairness Examples

Fairness in Machine Learning

- Protected features → Sensitive attributes
  - Features that are not to be used to make decisions (could lead to discrimination)
  - Protected attributes
    - Legal mandates
    - Organizational values (ethics)
    - Examples: race, religion, gender (sex), marital status, age, etc.

Fairness in Machine Learning: How to identify & define/measure

Exploratory data analysis in terms of fairness

- Unbalanced samples
- Prevalence
- Proxy variables

Definitions of fairness

- Equal opportunity
- Equalized odds
- Disparate impact

Source: Conor O'Sullivan, A Data Odyssey

Unbalanced Datasets

Issue

- Model parameters can be skewed towards the majority
  - Ex.: female vs. male trends (relationships between features and the target variable)
- A model will try to maximize accuracy across the whole population
  - Might favour trends in, for ex., the male population ("privileged")
    - Result: lower accuracy on the female population

Fairness analysis

- Define *protected* features: use *sensitive* attributes and create binary variables

Source: Conor O'Sullivan, A Data Odyssey

Prevalence

Prevalence

- The proportion of individuals who belong to the positive class
  - Ex.: individuals who earn more than $50K (y > 50K)
- Overall Prevalence = Positive cases / All cases

Prevalence (Fairness)

- Helps us understand the baseline distribution of positive cases across different groups
- This provides context for understanding if the model is biased in favour of certain groups

Source: Conor O'Sullivan, A Data Odyssey

Proxy Variable (Example: Fair Lending)

The Home Mortgage Disclosure Act (HMDA)

- requires certain financial institutions to collect, report, and disclose information about their mortgage lending activity
  - Originally enacted by the Congress in 1975
- HMDA was enacted given public concern over credit shortages in certain neighborhoods
  - Congress believed that some financial institutions had contributed to the decline of various geographic areas through their failure to provide adequate home financing to qualified applicants on reasonable terms and conditions
- Thus, one statutory purpose of HMDA:
  - Provide the public with information that will help show whether financial institutions are serving the housing credit needs of the communities and neighborhoods in which they are located

Proxy Variable

- "A variable used instead of the variable of interest when that variable of interest cannot be measured directly." (Source: Oxford Reference)
- Features that are correlated with protected features
  - How to measure association?
    - replace target with protected feature/s

Source: FDIC Consumer Compliance Examination Manual — July 2021

Examples of Proxy Variables

Intended Variable

- True body fat percentage
- Quality of life
- Cognitive ability
- ?
- ?
- Example 1?
  - DOJ vs. Associates National Bank
- Example 2?
  - Ex. domains: *economic, environmental, social well-being, public safety indicators, etc.*

Proxy Variable

- Body Mass Index (BMI)
- Per-capita GDP
- Years of education
- Occupation: Nurse
- Shopping at Whole Foods
- Example 1?
  - DOJ vs. Associates National Bank
- Example 2?
  - Your example

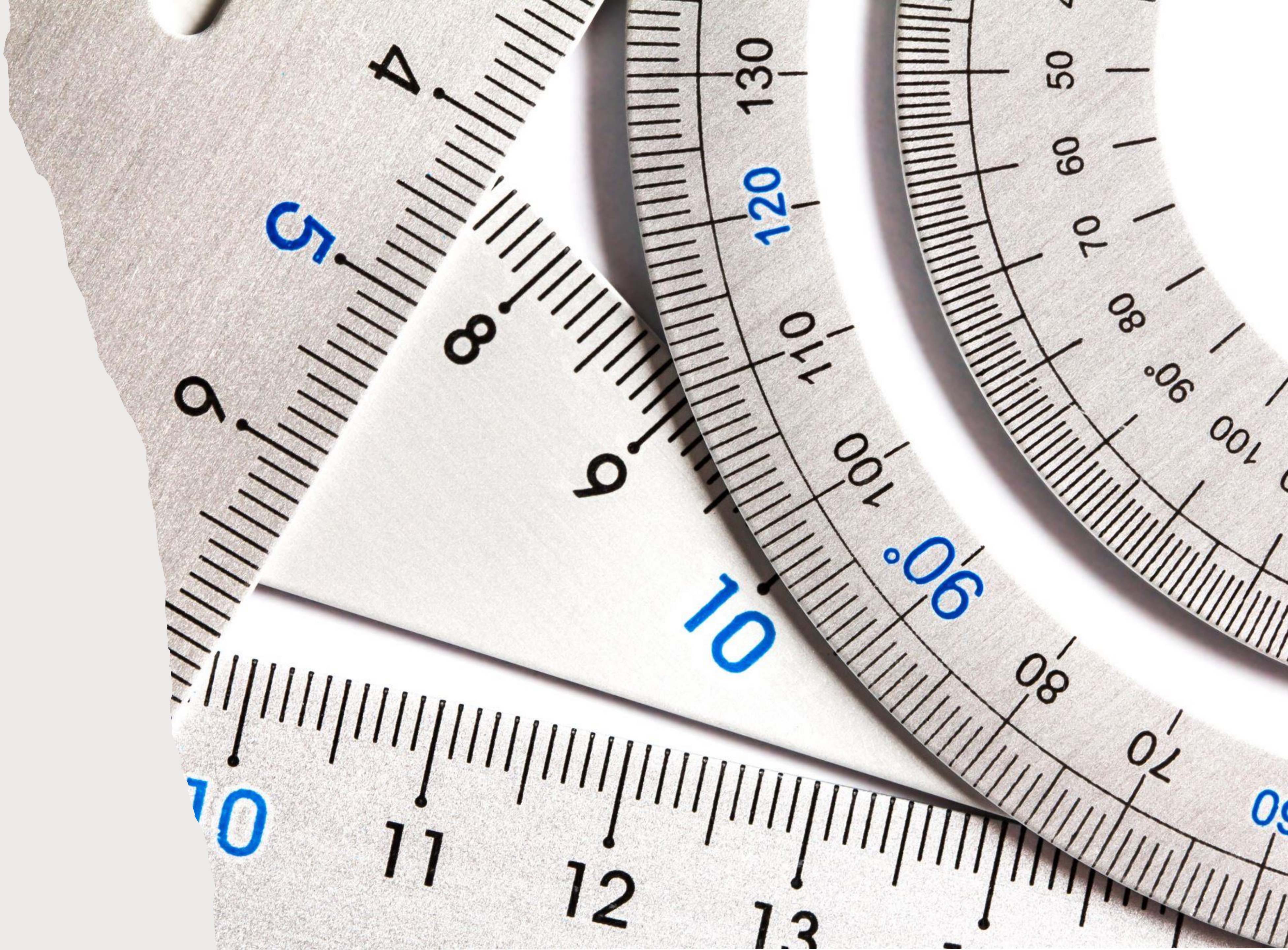
United States v. Associates National Bank

- Intended variable: ?
- Proxy variable: ?

On March 29, 1999, the United States filed a lawsuit against Associates National Bank of Delaware [ANB], a leading issuer of Visa and MasterCard bank cards, claiming that the bank violated the Equal Credit Opportunity Act [ECOA] by discriminating on the basis of national origin, specifically, against persons of Hispanic origin. Our complaint asserted that individuals applying for an ANB/UNOCAL MasterCard through the bank's Spanish-language application were processed through a separate approval system, which utilized a credit scoring system that required higher scores than those required for English-language applicants. As a consequence, some Spanish-language applicants were denied credit on a discriminatory basis. The United States also claimed that approved Spanish-language UNOCAL applicants were given lower credit line assignments than applicants processed through the English-language decision system.

Source: https://www.justice.gov/crt/housing-cases-summary-page

Measuring Fairness



Algorithmic Fairness: Basic Steps

- Reframe **target** variable
  - Where positive prediction = incurs some benefit
  - Ex.: predicting loan award
    - `[1 if y == '>50K'else 0 for y in df['y']]`

```
ŷ = { 1 → loan
      0 → no loan
```

- Reframe **protected** features:
  - 1 = privileged group
  - 0 = unprivileged group
  - Ex.: Race (1= white, 0=others), Sex (1=male, 0=female)

- Goal: Look at model performance by splitting the population into groups
  - Use fairness metrics (based on confusion matrix)

Accuracy

Predicition

Actual		0	1
	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Accuracy is the percentage of correct predictions:

$$\text{Accuracy} = \frac{TN + TP}{N}$$

$$\text{where } N = TN + FP + FN + TP$$

Accuracy of a model by protected features:

	1	0	Ratio
Race	83.9%	89.3%	1.07
Sex	81.0%	92.1%	1.14

Source: Conor O'Sullivan, A Data Odyssey

Equal Opportunity

- Assume: positive prediction will lead to some benefit

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

num of correctly **predicted positives**

number of **actual positives**

% of **actual** positives that were **correctly** predicted as positive

---

% who **rightfully benefitted** from the model

Under **equal opportunity we consider a model to be fair if the TPRs of the privileged and unprivileged groups are equal

	1	0	Ratio
Race	61.1%	53.3%	0.87
Sex	63.2%	44.3%	0.70

Source: Conor O'Sullivan, A Data Odyssey

Equal Opportunity

- Assume: positive prediction will lead to some benefit

Equal opportunity

$$TPR_0 = TPR_1 \quad (1)$$

$$TPR_1 - TPR_0 < \text{Cutoff} \quad (2)$$

$$\frac{TPR_0}{TPR_1} > \text{Cutoff} \quad (3)$$

Under **equal opportunity we consider a model to be fair if the TPRs of the privileged and unprivileged groups are equal

	1	0	Ratio
Race	61.1%	53.3%	0.87
Sex	63.2%	44.3%	0.70

Source: Conor O'Sullivan, A Data Odyssey

Equalized Odds

- FPR is the percentage of actual negatives incorrectly predicted as positive

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

incorrect positive predictions

num of **actual negatives**

% of people who have **wrongfully benefited** from the model

Equalized Odds

TPR_0 = TPR_1

FPR_0 = FPR_1

**Equalized odds: require that the FPRs are equal; overall benefit should be equal (right or wrong)

FPR = the num of low-income earners predicted as having a high income:

	1	0	Ratio
Race	8.1%	3.9%	0.48
Sex	10.9%	1.7%	0.16

Source: Conor O'Sullivan, A Data Odyssey

Disparate Impact

- % of people who will benefit from the model
  - % of people who have either been correctly (TP) or incorrectly (FP) predicted as positive (income >50k)

$$\text{\% predicted as positive (PPP)} = \frac{TP + FP}{N}$$

****DI:** a model is fair if we have equal PPP rates

	1	0	Ratio
Race	22.0%	11.7%	0.53
Sex	27.3%	6.6%	0.24

Disparate Impact

$$PPP_0 = PPP_1 \quad (1)$$

$$\frac{PPP_0}{PPP_1} > \text{Cutoff} \quad (3)$$

In the U.S. there is a **legal precedent**!
- Cutoff: **0.8**
- the unprivileged group's PPP must not be less than **80%** of that of the privileged group

Source: Conor O'Sullivan, A Data Odyssey

Walkthrough

Fairness analysis/metrics (measuring bias)