AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 8: AI: Supervised ML and Trust Mitigation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

14TH SEP, 2021

Carolinian Creed: "I will practice personal and academic integrity."

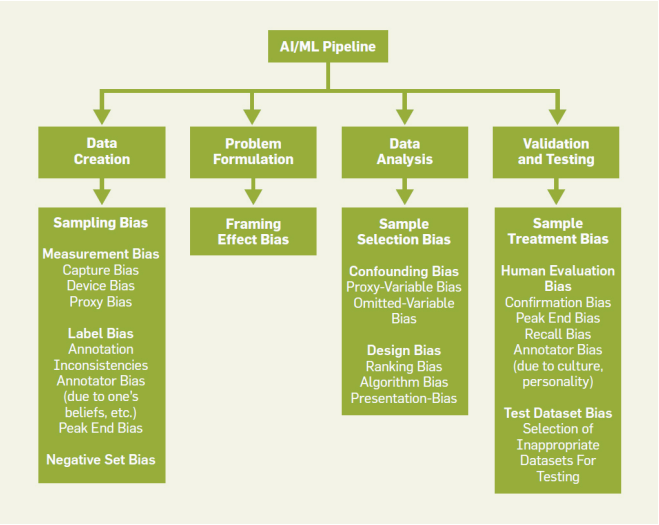
Organization of Lecture 8

- Introduction Segment
  - Recap from Lecture 7
  - Review of Quiz 1
  - Project discussion
- Main Segment
  - Paper: Fairness Definitions Explained
  - Focus on what it means for German-credit data
- Concluding Segment
  - About next lecture – Lecture 8
  - Ask me anything

Introductory Segment

Recap of Lecture 7

Looked at trust issues



- Trust Issues – Reading Material – Bias in AI
  - Illustrative problem: German credit data
  - Issues possible during data creation, problem formulation, data analysis, validation

- Trust Issues – Reading Material - Turing Institute report on AI for Covid in UK
  - Illustrative problem : vaccinate distribution
  - Issues in robust and timely data, and communication

Quiz 1 Recap and Discussion

Look at water data from Florida for WaterAtlas project and do four tasks

- WaterAtlas information
  - Website: https://orange.wateratlas.usf.edu/
  - About data collection and group: https://orange.wateratlas.usf.edu/about/
  - APIs to get data - Data download: https://dev.chnep.wateratlas.usf.edu/data-download/beta/
- Local cache with data
  - https://github.com/biplav-s/course-tai/tree/main/sample-code/common-data/water
  - Data for one lake: https://github.com/biplav-s/course-tai/blob/main/sample-code/common-data/water/WaterAtlas-OneLake.csv

Task 1 – Load data programmatically (10 points), summarize its statistics (10 points) and report on missing data (10 points). Note that a number of parameters are reported for the same date/time in successive rows.

Task 2 - Create plots for all the parameters with X-axis showing time and Y-axis showing the parameter value.

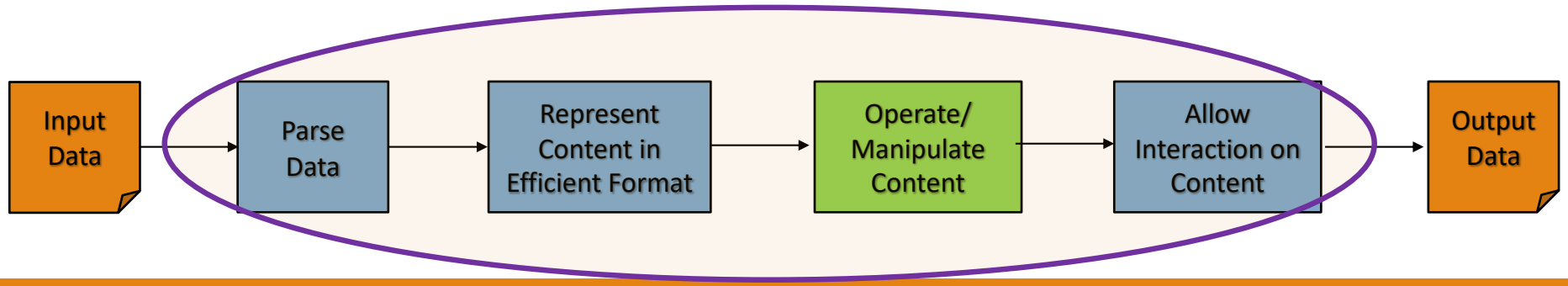
Task 3 – List at least 3 feature pairs with strong correlations (> 0.5 or <-0.5) among them? Show heatmap of correlation, if possible. What does this indicate?

Task 4 – If you are a resident living near this location and looking at this water data. You want to know answers for questions like if it safe to go to swim in the water, use water to irrigate your garden or event drink from it? Can this data answer any such questions? Discuss.

Project Discussion – Review Projects

- Information to be shared by students
  - Go to Google sheet: https://docs.google.com/spreadsheets/d/1VAX8ntb5zBQ-vOdsMHMhvEdwoaCZtuBaO4kJdkSA4eQ/edit?usp=sharing
  - Create a Google drive called "CSCE 590-1 Trusted AI (<YourName>)" and share with instructor: firstname.lastname@gmail.com
    - Put shared url in Column E
    - Put project title in column G
    - Create a folder in shared directory call project. Under it, have a Google doc called "Project Description". In it, have the following as bullets with associated details: Problem, User, AI Method, Data, Reliability: Testing, Holding Human Values, Human-AI interaction. See next slide for framework and guidance on what to put.
  - Put Github location for your code in F
    - Create one repository
    - For each quiz, project, etc, create a sub-folder

Main Segment



Discussion on Reading Material - 1

- Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (*FairWare '18*). Association for Computing Machinery, New York, NY, USA, 1–7. DOI: https://doi.org/10.1145/3194770.3194776
  - Pdf available on Arxiv, also stored in Blackboard

- 21 definitions of bias and their politics, Prof. Arvind Narayanan, Princeton, https://www.youtube.com/embed/jIXIuYdnyyk

German Credit Data

https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

- Dataset that classifies people's credit risk based on their individual attributes such as Age, Income, Gender, etc.
  - 1000 rows of data, each with 20 attributes to check bias against
- Each entry represents an individual who takes credit from a bank
- Each entry is classified as *Good* or *Bad* credit risk based on their profile
  - It is **worse** to class a **customer as good when they are bad**, than it is to class a **customer as bad when they are good**.

1. Credit amount (numerical);
2. Credit duration (numerical);
3. Credit purpose (categorical);
4. Status of existing checking account(categorical);
5. Status of savings accounts and bonds (categorical);
6. Number of existing credits (numerical);
7. Credit history(categorical);
8. Installment plans (categorical);
9. Installment rate(numerical);
10. Property (categorical);
11. Residence (categorical);
12. Period of present residency (numerical);
13. Telephone (binary);
14. Employment (categorical);
15. Employment length (categorical);
16. Personal status and gender (categorical); 1
17. Age (numerical);
18. Foreign worker (binary);
19. Dependents (numerical);
20. Other debtors (categorical);
21. Credit score (binary)

Review detailed data exploration at:

- https://www.kaggle.com/sanyalush/predicting-credit-risk

Example record: Alice is requesting a loan amount of 1567 DM for a duration of 12 months for the purpose of purchasing a television, with a positive checking account balance that is smaller than 200 DM, having less than 100 DM in savings account, and having one existing credit at this bank. She duly paid existing credits at the bank till now and has no other installment plan. She possesses a car and owns a house, has been living at the present residence for one year and has a registered telephone. She is a skilled employee, working in the present employment for past four years. She is a 22-year-old married female and is a German citizen. She has one dependent and no guarantors. The recorded outcome for Alice (attribute #21) is a good credit score.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science

VERMA, S., AND RUBIN, J. 2018. FAIRNESS DEFINITIONS EXPLAINED. IN PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS, FAIRWARE '18, 1–7. NEW YORK, NY, USA: ASSOCIATION FOR COMPUTING MACHINERY, HTTPS://WWW.ECE.UBC.CA/~MJULIA/PUBLICATIONS/FAIRNESS_DEFINITIONS_EXPLAINED_2018.PDF

Over 20 Definitions of Bias

Divides them along 5 categories:

1. C1: predicted outcome
2. C2: predicted and actual outcome
3. C3: predicted probabilities and actual outcome
4. C4: similarity based
5. C5: causal reasoning

Definitions from literature
as of January 2018!
Many more since.

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

Paper's Experiment

- Created a logistic regression model
  - Attributes: Used binary and numeric attributes as-is; created categorical attribute to a set of binary features; total 48 attributes
  - 90% training, 10% test
- Model favors single males when deciding on the credit score and disadvantage divorced males; behaves similarly for females and married males similarly

Attribute	Coefficient
Personal status and gender: single male	0.16
Personal status and gender: married male	-0.04
Personal status and gender: married/divorced female	-0.08
Personal status and gender: divorced male	-0.14

Metrics for Model Performance

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Positive predictive value (PPV, **Precision**)
False discovery rate (FDR)
False omission rate (FOR)
Negative predictive value (NPV)

True positive rate (TPR, **Recall**)
False positive rate (FPR)
False negative rate (FNR)
True negative rate (TNR)

Definitions - 1

- **[C1] Group fairness ((a.k.a. statistical parity, equal acceptance rate, benchmarking):** if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class
  - **P(d = 1 | G = m) = P(d = 1 | G = f)**
  - the probability to have a good predicted credit score for married / divorced male and female applicants is 0.81 and 0.75, respectively. Hence, fails this metric.

- **[C1] Conditional statistical parity**: if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L. L considered are: requested credit amount, applicant's credit history, employment, and age
  - P(d = 1 | L = l,G = m) = P(d = 1 | L = l,G = f).
  - Controlling for L, the probability for married / divorced male and female applicants to have good predicted credit score is 0.46 and 0.49, respectively. Considers metric **satisfied**.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C1: Definitions Based on Predicted Outcome

Definitions - 2

- **[C2] Predictive parity ((a.k.a. outcome test).**: if both protected and unprotected groups have equal PPV
  - for both male and female applicants, the probability of an applicant with a good predicted credit score to actually have a good credit score should be the same. Represented as: $P(Y = 1 | d = 1, G = m) = P(Y = 1 | d = 1, G = f)$.
  - A classifier with equal PPVs will also have equal FDRs: $P(Y = 0 | d = 1, G = m) = P(Y = 0 | d = 1, G = f)$.
  - PPV for married / divorced male and female applicants is 0.73 and 0.74, respectively. Inversely, FDR for male and female applicants is 0.27 and 0.26, respectively. Hence, metric **satisfied**.

- **[C2] False positive error rate balance (a.k.a. predictive equality)**: if both protected and unprotected groups have equal FPR
  - the probability of a subject in the negative class to have a positive predictive value: $P(d = 1 | Y = 0, G = m) = P(d = 1 | Y = 0, G = f)$
  - A classifier with equal FPRs will also have equal TNRs: $P(d = 0 | Y = 0, G = m) = P(d = 0 | Y = 0, G = f)$.
  - FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively. Inversely, TNR is 0.30 and 0.45. Hence, fails this metric.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 3

- **[C2] False negative error rate balance (aka, equal opportunity)**: if both protected and unprotected groups have equal FNR
  - the probability of a subject in a positive class to have a negative predictive value: P(d = 0|Y = 1,G = m) = P(d = 0|Y = 1,G = f).
  - A classifier with equal FNRs will also have equal TPR: P(d = 1|Y = 1,G = m) = P(d = 1|Y = 1,G = f).
  - FPRs for married / divorced male and female applicants are the same – 0.14. Inversely, TPR is 0.86. Hence, metric **satisfied**.

- **[C2] Equalized odds ((a.k.a. conditional procedure accuracy equality and disparate mistreatment):** if protected and unprotected groups have equal TPR and FPR
  - the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and the probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should both be same for male and female applicants:
  - **P(d = 1|Y = i,G = m) = P(d = 1|Y = i,G = f), i ∈ 0, 1.**
  - FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively and TPR is 0.86 for both males and females. Hence, fails this metric

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 4

- **[C2] Conditional use accuracy equality**: if equal PPV and NPV
  - the probability of subjects with positive predictive value to truly belong to the positive class and the probability of subjects with negative predictive value to truly belong to the negative class: (P(Y = 1 | d = 1,G = m) = P(Y = 1 | d = 1,G = f)) ∧ (P(Y = 0 | d = 0,G = m) = P(Y = 0 | d = 0,G = f)).
  - The calculated value for male and female applicants is 0.73 and 0.74, respectively. NPVs for male and female applicants is 0.49 and 0.6
  - More likely for a male than female applicant with a bad predicted score to actually have a good credit score. Fails this metric.

- **[C2] Overall accuracy equality:** if both protected and unprotected groups have equal prediction accuracy
  - the probability of a subject from either positive or negative class to be assigned to its respective class
  - **P(d = Y,G = m) = P(d = Y,G = f)**
  - **T**he overall accuracy rate is 0.68 and 0.71 for male and female applicants, respectively. Paper considers metric **satisfied**.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 5

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C2] Treatment equality:** satisfies this definition if both protected and unprotected groups have an equal ratio of false negatives and false positives.
  - ratio of FP to FN is same for male and female applicants: [FN/ FP] m = [FN / FP] f
  - Calculated ratios are 0.56 and 0.62 for male and female applicants. Fails this metric.

- **[C3] Test fairness (a.k.a. calibration, matching conditional frequencies):** if for any predicted probability score S, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class
  - for any given predicted probability score s in [0, 1], the probability of having actually a good credit score should be equal for both male and female applicants: P(Y = 1 | S = s,G = m) = P(Y = 1 | S = s,G = f).
  - It is more likely for a male applicant with a bad predicted credit score (low values of S) to actually have a good score (definition 3.2.5), but applicants with a good predicted credit score (high values of S) have an equivalent chance to indeed have a good credit score, regardless of their gender Paper considers metric **satisfied**.

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1\|S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1\|S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4: Calibration scores for different values of s

C3: Definitions Based on Predicted Probabilities and Actual Outcome

Definitions - 6

- **[C3] Well-calibration:** for any predicted probability score S, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to S.
  - for any given predicted probability score s in [0, 1], the probability of having actually a good credit score should be equal for both male and female applicants: $P(Y = 1 | S = s, G = m) = P(Y = 1 | S = s, G = f) = s$.
  - Paper considers metric (partially) **satisfied**.

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1 \mid S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1 \mid S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4: Calibration scores for different values of s

- **[C3] Balance for positive class:** if subjects constituting positive class from both protected and unprotected groups have equal average predicted probability score S.
  - expected value of probability assigned by the classifier to male and female applicant with good actual credit score should be same: $E(S | Y = 1, G = m) = E(S | Y = 1, G = f)$
  - The calculated expected value of predicted probability score is 0.72 for both males and females. Paper considers metric **satisfied**.
  - Paper sees this along with equal opportunity/ false negative error rate balance which was also satistied.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C3: Definitions Based on Predicted Probabilities and Actual Outcome

Definitions - 7

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C3] Balance for negative class:** if subjects constituting negative class from both protected and unprotected groups have equal average predicted probability score S.
  - expected value of probability assigned by the classifier to male and female applicant with good actual credit score should be same: E(S |Y = 0,G = m) = E(S |Y =0,G = f)
  - The the expected value of having bad predicted credit score is 0.61 and 0.52 for males and females. Paper sees failing of this metric.
  - Paper sees this along with predictive equality/ false positive error rate balance which was also not satisfied

- **[C4] Causal discrimination:** if it produces the same classification for any two subjects with the exact same attributes X.
  - A male and female applicants who otherwise have the same attributes X will either both be assigned a good credit score or both assigned a bad credit score:
    (Xf = Xm ∧ Gf != Gm)→ df = dm.
  - For 8.8% married / divorced male and female applicants, the output classification was not same. Paper sees metric not being satisfied.
  - **Closest to legal notion on dissimilar treatment ?**

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C4: Similarity Based Measures

Definitions - 8

- **[C4] Fairness through unawareness:** if no sensitive attributes are explicitly used in the decision-making process
  - the classification outcome should be the same for applicants i and j who have the same attributes X: Xi = Xj →di = dj
  - Trained a model without gender and checked. Found classification identical for both genders. Metric **satisfied**.

- **[C4] Fairness through awareness :** similar individuals should have similar classification
  - For a set of applicants V , a distance metric between applicants k : VxV → R, a mapping from a set of applicants to probability distributions over outcomes M : V → δA, and a distance D metric between distribution of outputs, fairness is achieved iff D(M(x),M(y)) ≤ k(x,y).
  - Dependent on distance metric
  - k: gender-based; similar to causal discrimination (8.8% difference)
    k: age-based; the distance between outcomes (column 3) grew much faster than the distance between ages (column 3). Thus, the percentage of applicants who did not satisfy this definition (column 4) increased.
    Metric unsatisfied

Age difference	k	Avg. D	% violating cases
5	0.09	0.02	0.0
10	0.18	0.05	0.5
15	0.27	0.10	1.8
20	0.36	0.2	4.5
25	0.45	0.3	6.7

Table 5: Fairness through awareness with age-based distance

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C4: Similarity Based Measures

Definitions - 9

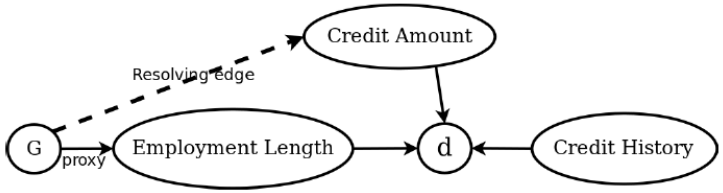


Figure 1: Causal graph example

- **Definitions based on knowledge of the domain:** given a relationship between attributes, check the relationship between decision and protected variables, other variables
  - Proxy attribute: whose value can be used to derive a value of another attribute
  - Resolving attribute: attribute that is influenced by the protected attribute in a non-discriminatory manner

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C5: Causal Reasoning

Group Discussion – 10 mins

- A: Suppose you are **a loan applicant** and your application was
  - Accepted (good credit). Which metric would you use to explain to some other applicant whose application was rejected (bad credit)?
  - Rejected (bad credit). Which metric would you use to ask some other applicant whose application was accepted (bad credit)

- B: Suppose you are the executive at the bank incharge of making loans. Which metric will you use to justify that your bank does not consider gender or age to discriminate in your processing?

- C: Suppose you are the government regulator or a journalist. Which metric will you use to check if the bank discriminates in it loan practices ?

Concluding Segment

Lecture 8: Concluding Comments

- We looked at bias definitions
  - Five categories: C1: predicted outcome, C2: predicted and actual outcome, C3: predicted probabilities and actual outcome , C4: similarity based, C5: causal reasoning
  - Reviewed with respect to German-credit as example
- Metrics should not only be technically sound but practically useful
  - Did role-playing to discuss
- Most are theoretical exercises while law catches up; little technical guidance to developers

About Next Lecture – Lecture 9

Lecture 9: Supervised ML – Explanation Methods

- Explanation method: LIME
- Mitigation methods
  - Book: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence