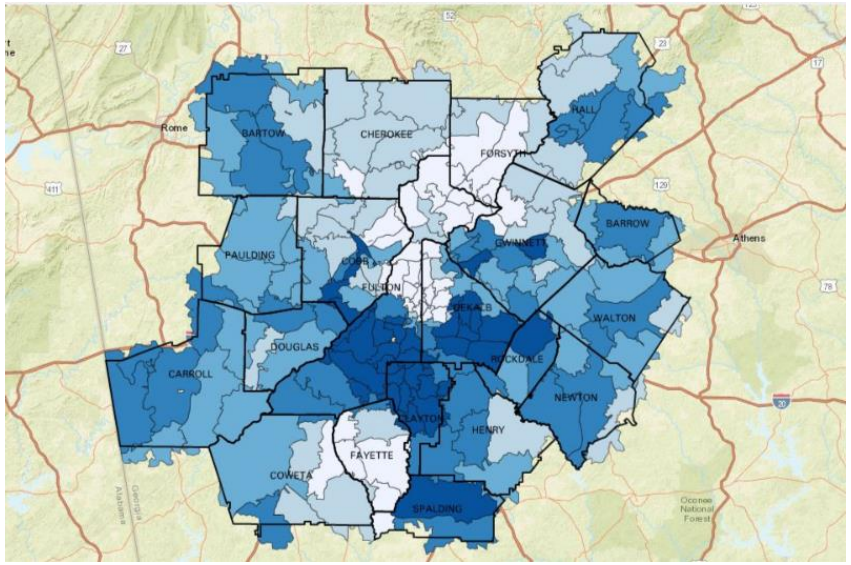
# Responsible AI

Lecture 3

# Today

1. Fairness and Proxies
2. Fairness metrics
3. Proxy features
4. Methods to detect proxy features
    1. Variance Inflation Factor (VIF)
    2. Linear association method using variance

# Proxies


Rome
BARTOW
CHEROKEE
FORSYTH
HALL
BARROW
PAULDING
GWINNETT
Athens
FULTON
WALTON
DOUGLAS
ROCKDALE
CARROLL
NEWTON
HENRY
FAYETTE
COWETA
SPALDING


Jobs & Occupations
waiter
paramedic
dentist
train conductor
nurse
electrician
doctor
businessman
American football player
student
surgeon
doorman


Occupations
boxer
cameraman
detective
journalist
housewife
diver
pope
priest
salesman
librarian
pirate
singer


-65,45
-250,00
-27,90
-42,74
+1.662,00
-29,20
-27,68
-39,34
-337,89
-44,98
TAX

# How to detect

Correlation

**Feature Combinations**

**Feature Redundancy**

# Common methods

$$SE = \sqrt{1 - R^2_{adj}} \times \sigma_y$$

$$\text{where } R^2_{adj} = 1 - \frac{\left(1 - R^2\right)\left(n - 1\right)}{n - k - 1}$$

Linear Regression

$$VIF = \frac{1}{1 - R^2}$$

Variance Inflation factor

$$Assoc = \frac{cov\left(X_1, X_2\right)^2}{Var\left(X_1\right) Var\left(X_2\right)}$$

Association

# PII data

- Race (Civil Rights Act of 1964)
- Colour
- Sex including gender, pregnancy, sexual orientation, and gender identity (Equal Pay Act of 1963)
- Religion or creed
- National origin or ancestry
- Citizenship (Immigration Reform and Control Act)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familial status
- Physical or mental disability status (Rehabilitation Act of 1973)
- Veteran status
- Genetic information

# Remember…

For any problem we are trying to solve, or any model we are working to train, the various possible outputs of the model are the possible outcomes – these out-comes can be favourable or unfavourable

Protected features can introduce biases in our models. We have looked at some examples of protected features recognized by law, but in order to remove the biases efficiently for our use case, we need to identify all protected features for our dataset

The possible user classes that comprise a given protected feature will have a privileged and unprivileged groups

Privileged groups are the ones that are more likely to have a favourable outcome than the unprivileged groups.


# Confusion Matrix

Individually define

- TP, FP, FN, TN

Examples on:

- FPR
- TPR
- FNR
- TNR
- Precision
- Recall

# Fairness concept

Independence: $\hat{Y}$ independent of S, if the predicted features and protected features are independent of each other. In other words, probability of being in a favourable (or non-favourable) class has nothing to do with the group of S.

Separation: $\hat{Y}$ independent of S given Y, if the predicted feature given the target value Y is independent of protected feature. The predicted probability of being in any class, given their actual class, has nothing to do with their protected group membership.

Sufficiency: Y independent of S given $\hat{Y}$. Here we expected protected class to be independent of actual value given the predicted value. The probability of actually being in each of the groups has nothing to do with membership of protected feature. In simple words, the prediction should not depend on the protected group.

# Equal Opportunity

This talks about equal FNR

E.g., the probability of an actual non-defaulting applicant to be incorrectly predicted as a defaulter should be the same for both groups of a protected feature of the protected feature applicants; no group will have an advantage of a reduced miss rate.

**Why? What about TPR?**

$$P\left(\hat{Y}=0\,|\,Y=1,S=S_a\right)=P\left(\hat{Y}=0\,|\,Y=1,S=S_d\right)$$

# Predictive Equality

Both advantageous and disadvantageous groups have equal FPR.

Probability that of an actual defaulter to be incorrectly predicted as a non-defaulter should be the same for both subsets of the protected class applicants

$$P\left(\hat{Y}=1 \mid Y=0, S=S_a\right)=P\left(\hat{Y}=1 \mid Y=0, S=S_d\right)$$

# Equalized Odds

Also known as disparate mistreatment, equalized odds requires that both advantageous and disadvantageous groups have equal TPR and FPR.

The main idea behind the definition is that loan applicants with a good actual credit score and loan applicants with a bad actual credit score should have a similar classification, regardless of their membership to a protected class.

$$P\left(\hat{Y}=1|Y=i,S=S_a\right)=P\left(\hat{Y}=1|Y=i,S=S_d\right),i\subset 0,1$$

# Predictive Parity

Predictive parity is also known as the outcome test; this requires both advantageous and disadvantageous groups to have equal PPV/precision

The percentage of correct positive predictions should be the same for both the groups of a protected class. It implies that the errors are spread homogeneously among all the groups of a protected class. Furthermore, the odds for receiving a favourable outcome would be same irrespective of their membership to the group. In our example, this implies that, for both the advantageous and disadvantageous subsets of a protected group, the probability that a non-defaulting loan applicant is predicted to be a non-defaulter should be the same

# What if PPV is 1?

$$P\left(Y=1 \mid \hat{Y}=1, S=S_a\right)=P\left(Y=1 \mid \hat{Y}=1, S=S_d\right)$$

# Demographic Parity

Membership in a protected class should have no correlation with being predicted a favourable outcome

$$P\left(\hat{Y}=1, S=S_a\right)=P\left(\hat{Y}=1, S=S_d\right)$$

- For the privileged group, we reduce false positives and increase true negatives. This is the ideal case. Here not only the cost would decrease as FP is more expensive for the business but also ensuring that unfavourable outcomes are kept at a bay to prevent any gain by privileged group. This is akin to reducing the privilege of the privileged class.

- For the unprivileged group, we reduce false negatives and increase true positives. Here we will observe that the probability of unprivileged class to have a favourable outcome (increase in TP) will see a boost and will also ensure that unprivileged class does not see a decline in the favourable outcome (reduced FN) when they deserve a favourable outcome. This will also increase the number of candidates to whom service should be offered and is similar to positive discrimination.

# Average Odds Difference

Average of difference in FPR and TPR for advantageous and disadvantageous groups

This metric encompasses both the predictive equality difference (FPR to TNR diff) and equal opportunity difference. A lower difference (or a zero difference) would mean equal benefit (positive or favourable outcome) for both the group.

$$\frac{1}{2}\left[\left(FPR_{s_d} - FPR_{s_a}\right) + \left(TPR_{s_d} - TPR_{s_a}\right)\right]$$

# Prioritizing Fairness Metrics

Demographic Parity vs Equal Opportunity

# Explore

- FACET
- Github (https://github.com/srayagarwal/JIO_RAI/blob/main/Ch%208%20Data%20and%20Model%20Privacy.ipynb)

# Assessment

In Groups use any data of your choice having PII data:

Implement one among Cosine similarity, Distance method, Mutual Information to detect proxies

**Submission on 4th lecture**

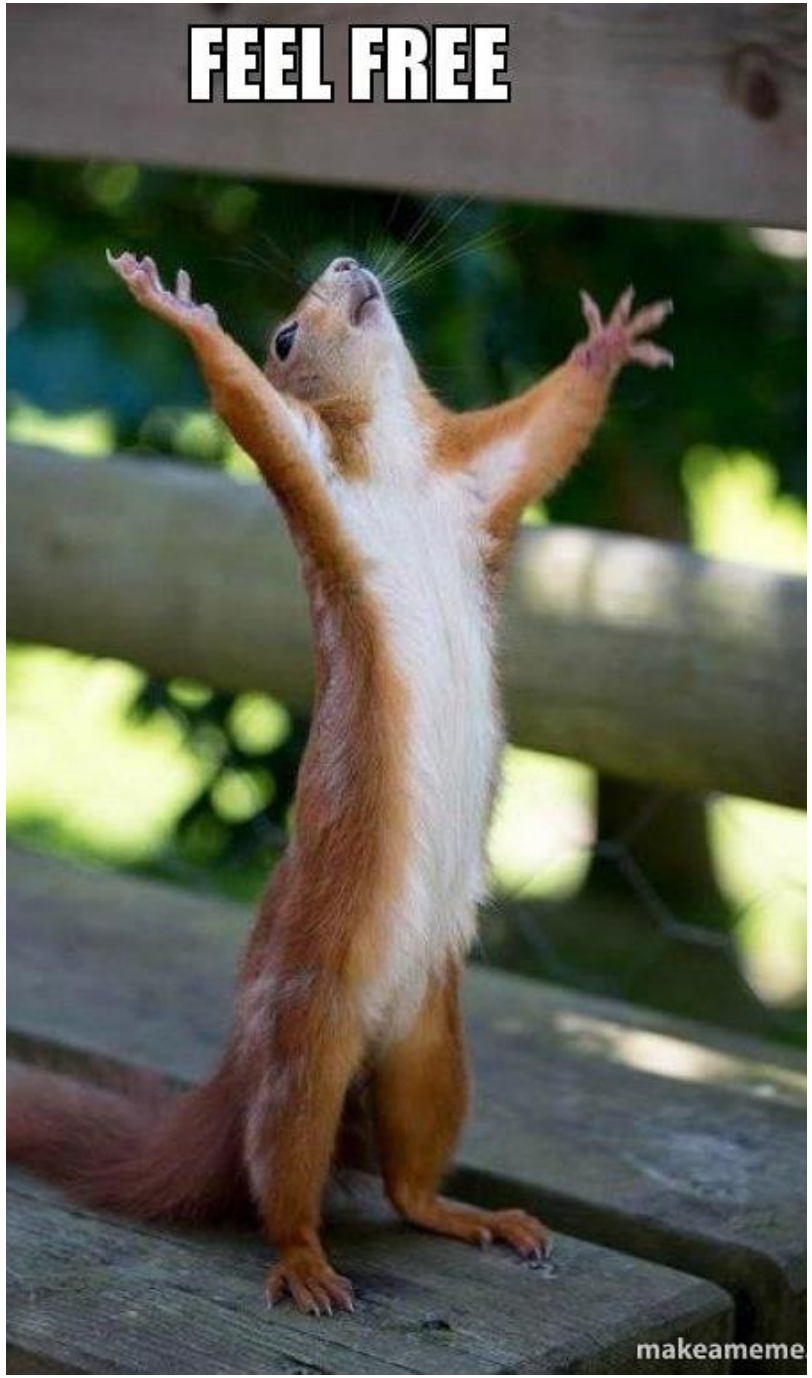
# Explore

Github (https://github.com/srayagarwal/JIO_RAI/blob/main/Ch%208%20Data%20and%20Model%20Privacy.ipynb)

# Next

1. Statistical parity difference
2. Disparate impact
3. Binary features with continuous output
4. Continuous features with binary output

**Chapters from book, research paper on fairness metrics**
**Revise: Confusion Matrix**


FEEL FREE
makeameme


YOU CAN CALL ME


DID YOU REALLY JUST SEND THAT EMAIL!


KNOCK, KNOCK!