Measuring Algorithmic Biases and Fairness

Swati Mishra

Applications of Machine Learning (4AL3)

Fall 2024

McMaster University

ENGINEERING

Review

- Biases
- Types of Biases - Cognitive, Social
- Why Biases are so important?
- How do we measure biases?

How do we improve fairness

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome $P(Y' = 1|Y = 1) > \tau$

negative outcome $P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Recall $= 0.60$ Precision $= 0.92$

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Recall $= 0.88$ Precision $= 0.80$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

How do we improve fairness

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome $P(Y' = 1|Y = 1) > \tau$

negative outcome $P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Recall $= 0.61$ Precision $= 1$

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	0	65

Recall $= 0.88$ Precision $= 0.84$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	15	80

Dataset Fairness

- If the number of members in Group A significantly exceeds group B, this **Group Imbalance** (GI) may bias the model.

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	1	1	1	0	0	4583	1508.0	128.0	360.0	1.0	0	0
2	1	1	0	0	1	3000	0.0	66.0	360.0	1.0	2	1
3	1	1	0	1	0	2583	2358.0	120.0	360.0	1.0	2	1
4	1	0	0	0	0	6000	0.0	141.0	360.0	1.0	2	1
5	1	1	2	0	1	5417	4196.0	267.0	360.0	1.0	2	1

$$GI = \frac{n_B}{n_A} \geq 1 - \beta$$

β = bias threshold extent to allowable bias under this metric

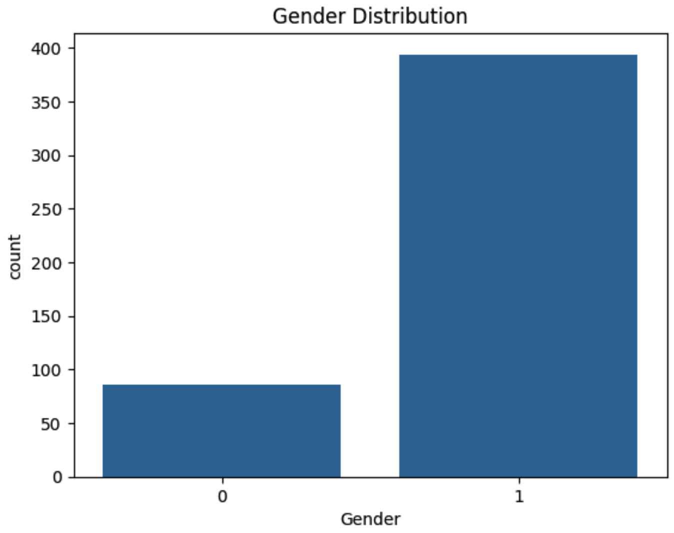
Dataset Fairness

- If the number of members in Group A significantly exceeds group B, this **Group Imbalance** (GI) may bias the model.

```
Gender distribution in the dataset:

Gender
1    394
0     86
Name: count, dtype: int64
```

```
Loan Approval Rate for each Gender:
        Not Approved  Approved
Gender
0           0.372093  0.627907
1           0.294416  0.705584
```




```
Classes in 'Gender': ['Female' 'Male']
Corresponding numeric labels:  [0 1]
```

Dataset Fairness

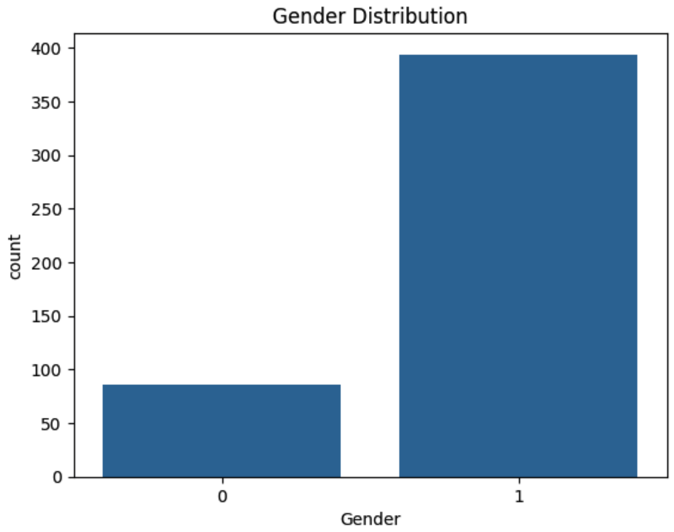
- If the number of members in Group A significantly exceeds group B, this **Group Imbalance** (GI) may bias the model.

For $\beta = 0.2$ what is GI?

```
Classes in 'Gender': ['Female' 'Male']
Corresponding numeric labels:  [0 1]
```

```
Gender distribution in the dataset:

Gender
1    394
0     86
Name: count, dtype: int64
```



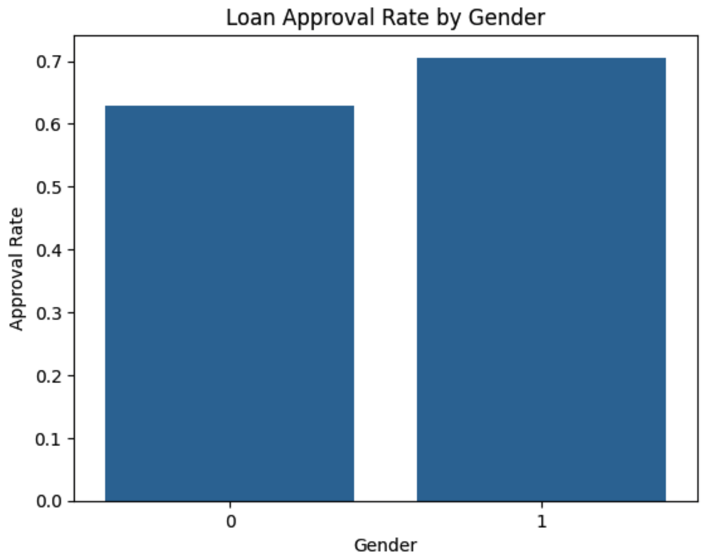

```
Loan Approval Rate for each Gender:
        Not Approved  Approved
Gender
0           0.372093  0.627907
1           0.294416  0.705584
```

Dataset Fairness

- If the ratio of approved to declined applications differs significantly across groups, it is called **Class Imbalance** (CI) and leads to biased results.

```
Loan Approval Rate for each Gender:
        Not Approved  Approved
Gender
0           0.372093  0.627907
1           0.294416  0.705584
```

$$CI = \frac{\#(Y=1|A)}{\#(Y=0|A)} - \frac{\#(Y=1|B)}{\#(Y=0|B)} \leq \beta$$



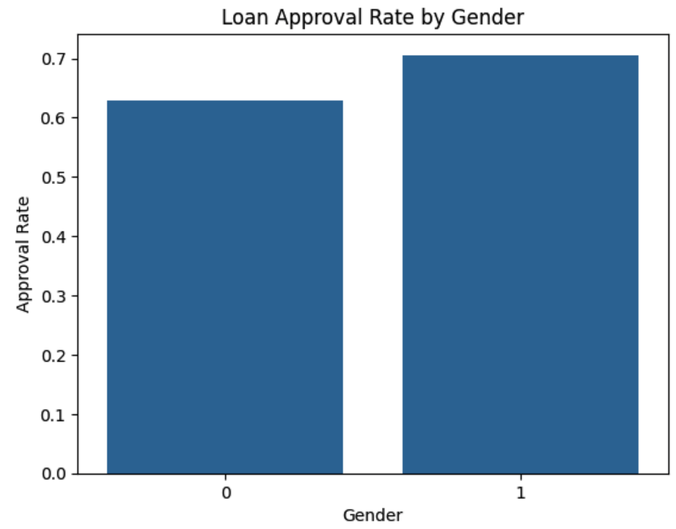

```
Classes in 'Gender': ['Female' 'Male']
Corresponding numeric labels:  [0 1]
```

Dataset Fairness

- If the ratio of approved to declined applications differs significantly across groups, it is called **Class Imbalance** (CI) and leads to biased results.

```
Loan Approval Rate for each Gender:
        Not Approved  Approved
Gender
0           0.372093  0.627907
1           0.294416  0.705584
```

$$CI = \frac{\#(Y = 1|A)}{\#(Y = 0|A)} - \frac{\#(Y = 1|B)}{\#(Y = 0|B)} \leq \beta$$



For $\beta = 0.2$ what is CI?

```
Classes in 'Gender': ['Female' 'Male']
Corresponding numeric labels:  [0 1]
```

Dataset Fairness

- Class imbalance can be generalized to multiple classes and then it is called **Distributional Imbalance** (CI)

$$DI = \sum_{y} f_A(y)\, log_2\left(\frac{f_A(y)}{f_B(y)}\right) \leq \beta$$

where, $f_A(y) = \frac{\#(Y=y|g)}{n_g}$

g = group,
y = category in binary class

Classifier Fairness

- **Disparate Impact** (Dimp): When decisions made about positive or negative outcomes lead to unintentional discrimination.

$$DI = \frac{Pr(Y' = 1|B)}{Pr('Y = 1|A)} \geq 1 \ - \beta$$

Classifier Fairness

- **Disparate Impact** (Dimp): When decisions made about positive or negative outcomes lead to unintentional discrimination.

$$DI = \frac{Pr(Y' = 1|B)}{Pr('Y = 1|A)} \geq 1 - \beta$$

- Disparate impact is presented as a ratio. It can also be computed as a difference, which is commonly known as “demographic parity”,

Classifier Fairness

- **Equal Opportunity** (EOpp): A metric to assess whether a model is predicting outcomes equally well for all groups with respect to both the positive and negative class—not just one class or the other exclusively.

TPR difference : $| Pr(Y' = 1|B, Y = 1) - Pr(Y' = 1|A, Y = 1) | \leq \beta$

Classifier Fairness

- **Equal Opportunity** (EOpp): A metric to assess whether a model is predicting outcomes equally well for all groups with respect to both the positive and negative class—not just one class or the other exclusively.

TPR difference : $| \, Pr(Y' = 1 | B, Y = 1) - Pr(Y' = 1 | A, Y = 1) \, | \leq \beta$

It accounts for both true labels and predicted labels when measuring disparate impact

```
# Equal Opportunity Difference: True positive rate difference between groups
def equal_opportunity_diff(y_true, y_pred, sensitive_attribute):
    tpr_male = sum((sensitive_attribute == 1) & (y_pred == 1) & (y_true == 1)) / sum((sensitive_attribute == 1) & (y_true == 1))
    tpr_female = sum((sensitive_attribute == 0) & (y_pred == 1) & (y_true == 1)) / sum((sensitive_attribute == 0) & (y_true == 1))
    return abs(tpr_male - tpr_female)
```

Classifier Fairness

- **Equal Opportunity** (EOpp): A metric to assess whether a model is predicting outcomes equally well for all groups with respect to both the positive and negative class—not just one class or the other exclusively.

TPR difference : $| \, Pr(Y' = 1 | B, Y = 1) - Pr(Y' = 1 | A, Y = 1) \, | \leq \beta$

True Positive Rate

```
# Equal Opportunity Difference: True positive rate difference between groups
def equal_opportunity_diff(y_true, y_pred, sensitive_attribute):
    tpr_male = sum((sensitive_attribute == 1) & (y_pred == 1) & (y_true == 1)) / sum((sensitive_attribute == 1) & (y_true == 1))
    tpr_female = sum((sensitive_attribute == 0) & (y_pred == 1) & (y_true == 1)) / sum((sensitive_attribute == 0) & (y_true == 1))
    return abs(tpr_male - tpr_female)
```

Classifier Fairness

- **Equalized Odds** (EOdds): A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal true positive rate and equal false positive rate

$$FPR\ difference = \mid Pr(Y' = 1|B, Y = 0) - Pr(Y' = 1|A, Y = 0) \mid \leq \beta$$

True Positive Rate and False Positive Rate

Equalized odds is related to equality of opportunity, which only focuses on error rates for a single class (positive or negative).

Classifier Fairness

- **Accuracy Difference** (AD): Difference in accuracies of individual groups.

$$Pr(Y' = Y|A) - Pr(Y' = Y|B) \leq \beta$$

$Y' = 1$ if the loan is approved

positive outcome $P(Y' = 1|Y = 1) > \tau$
negative outcome $P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Accuracy= 0.743

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Accuracy= 0.85

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

Classifier Fairness

- **Treatment Equality** (TE): This metric assesses whether the kinds of error made by the algorithm are similar across groups, i.e., are the ratios of false negatives to false positives the same.

$$\frac{\#(Y' = 0|A, Y = 1)}{\#(Y' = 1|A, Y = 0)} - \frac{\#(Y' = 0|B, Y = 1)}{\#(Y' = 1|B, Y = 0)} \leq \beta$$

Classifier Fairness

- **Treatment Equality** (TE): This metric assesses whether the kinds of error made by the algorithm are similar across groups, i.e., are the ratios of false negatives to false positives the same.

$$\frac{\#(Y' = 0|A, Y = 1)}{\#(Y' = 1|A, Y = 0)} - \frac{\#(Y' = 0|B, Y = 1)}{\#(Y' = 1|B, Y = 0)} \leq \beta$$

positive outcome $P(Y' = 1|Y = 1) > \tau$

$Y' = 1$ if the loan is approved

negative outcome $P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

Are errors more harmful to one group than another?

Classifier Fairness

- **Predictive Parity** (PPP, NPP): It is the difference in the ratio of false negatives to false positives.

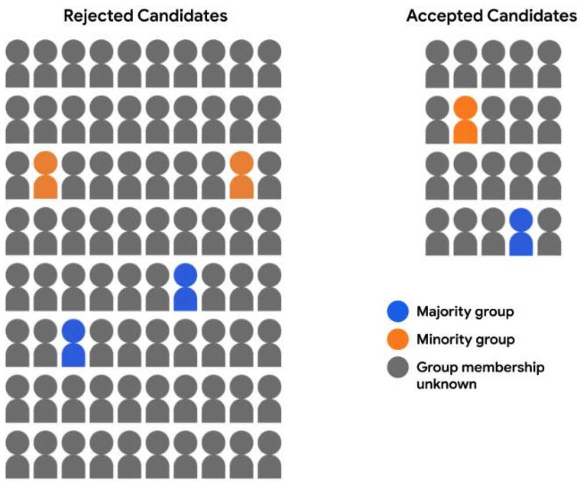
Positive predictive parity $|Pr(Y = 1|A, Y' = 1) - Pr(Y = 1|D, Y' = 1) \leq \beta$

Negative predictive parity $|Pr(Y = 0|A, Y' = 0) - Pr(Y = 0|D, Y' = 0) \leq \beta$

Classifier Fairness

- **Individual Fairness**: This is based on the simple idea that similar individuals should receive similar outcomes.

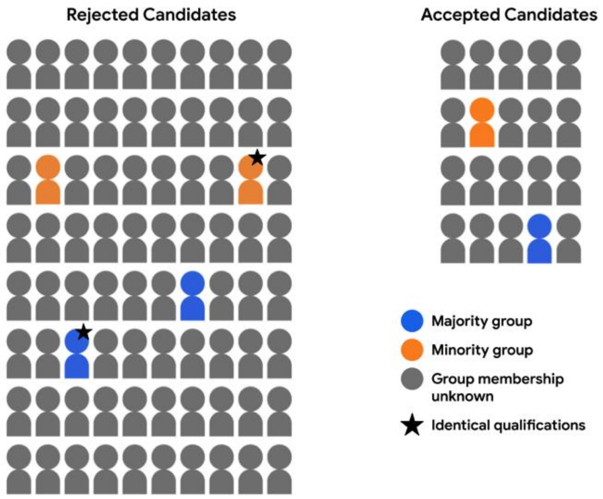
$$|Pr(Y_i' = y|X_i) - Pr\left(Y_j' = y\middle|X_j\right)| \leq \beta \text{ if } \mathrm{d}(X_i, X_j) \leq \epsilon$$

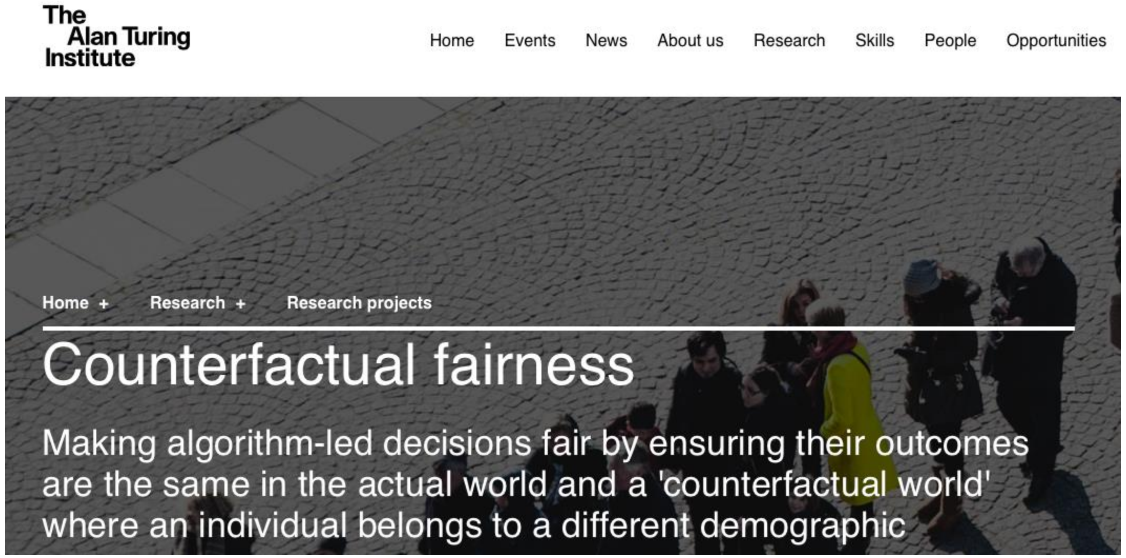


Classifier Fairness

- **Counterfactual Fairness:** It states that two examples that are identical in all respects, except a given sensitive attribute, should result in the same model prediction.

$$|Pr(Y_i' = 1|A, X_i) - Pr\left(Y_j' = 0\middle|B, X_j\right)| \leq \beta \text{ if } \mathrm{d}(X_i, X_j) \leq \epsilon$$





Source: https://developers.google.com/machine-learning/crash-course/fairness/counterfactual-fairness

Readings

Reference Material:

- Algorithmic Fairness (Sanjiv Das, Richard Stanton, and Nancy Wallace)
  Annual Review of Financial Economics

- Source links included in slides

Thank You

McMaster University