Algorithmic fairness is complicated!

- There are lots of ways to measure algorithmic fairness
- It’s impossible to satisfy all fairness measures simultaneously

But… it is *extremely* important

- Ethically: we don’t want to inadvertently cause harm to individuals because of biases in data or models
- Legally: we could get in a lot of trouble if out models are prejudiced against legally-protected groups

- There are lots of ways to measure algorithmic fairness
  - What are different fairness metrics?
  - How do I calculate them?
  - What do they measure?
  - What are the differences between them?

Comprehend differences between fairness metrics

- It's impossible to satisfy all fairness measures simultaneously
  - Which metric should I use?

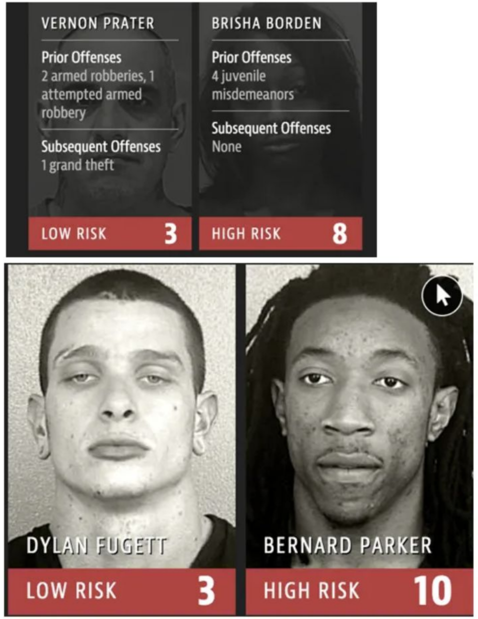
Choose appropriate fairness metrics

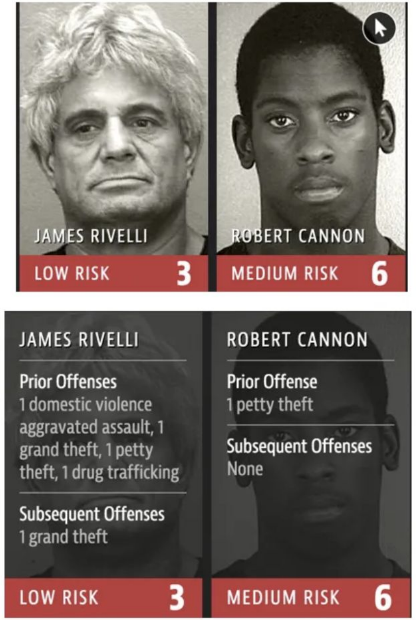
Outline

- Look at Recidivism example
- Interactive exercise to develop intuition to understand…
    - Differences between fairness metrics
    - How to choose a fairness metric
- Review exercise
- Discuss additional resources

Recidivism

"The tendency of a convicted criminal to reoffend"





https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The COMPAS software

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was used to predict the likelihood that a convicted criminal would reoffend ("recidivism risk")
- Has been used in New York, Wisconsin, California, and Florida
- Informed court decisions such as whether to grant parole

→ The COMPAS algorithm was shown to be fair according to some metrics and unfair according to others

→ Received major backlash because many people believed they chose the wrong fairness metric to optimize for given the context of recidivism

https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

COMPAS Data from Bowen County, FL (2013-2014)

Legally protected sensitive attributes used to predict recidivism risk

Predicts risk level which can be categorized as “High” or “Low”

```
RangeIndex: 60843 entries, 0 to 60842
Data columns (total 28 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person_ID                60843 non-null  int64
 1   AssessmentID             60843 non-null  int64
 2   Case_ID                  60843 non-null  int64
 3   Agency_Text              60843 non-null  object
 4   LastName                 60843 non-null  object
 5   FirstName                60843 non-null  object
 6   MiddleName               15624 non-null  object
 7   Sex_Code_Text            60843 non-null  object
 8   Ethnic_Code_Text         60843 non-null  object
 9   DateOfBirth              60843 non-null  object
 10  ScaleSet_ID              60843 non-null  int64
 11  ScaleSet                 60843 non-null  object
 12  AssessmentReason         60843 non-null  object
 13  Language                 60843 non-null  object
 14  LegalStatus              60843 non-null  object
 15  CustodyStatus            60843 non-null  object
 16  MaritalStatus            60843 non-null  object
 17  Screening_Date           60843 non-null  object
 18  RecSupervisionLevel      60843 non-null  int64
 19  RecSupervisionLevelText  60843 non-null  object
 20  Scale_ID                 60843 non-null  int64
 21  DisplayText              60843 non-null  object
 22  RawScore                 60843 non-null  float64
 23  DecileScore              60843 non-null  int64
 24  ScoreText                60798 non-null  object
 25  AssessmentType           60843 non-null  object
 26  IsCompleted              60843 non-null  int64
 27  IsDeleted                60843 non-null  int64
dtypes: float64(1), int64(9), object(18)
```

- Built a logistic regression model to predict low vs. high risk of recidivism
  - 0: low risk
  - 1: high risk
- 200 people in my test set

Validation Dataset Testing Results:

Accuracy: 0.715

We do pretty well!

[[75 32]

[25 68]]

	precision	recall	f1-score	support
0	0.75	0.70	0.72	100
1	0.68	0.73	0.70	100
accuracy			0.72	200
macro avg	0.72	0.72	0.71	200
weighted avg	0.72	0.72	0.71	200

- Built a logistic regression model to predict low vs. high risk of recidivism
  - 0: low risk
  - 1: high risk
- 200 people in my test set

However, 100 are White and 100 are not

Is the model fair according to race?

Validation Dataset Testing Results:

Accuracy: 0.715

[[75 32]

[25 68]]

	precision	recall	f1-score	support
0	0.75	0.70	0.72	100
1	0.68	0.73	0.70	100
accuracy			0.72	200
macro avg	0.72	0.72	0.71	200
weighted avg	0.72	0.72	0.71	200

Parity-based Fairness Metrics

Confusion Matrix:

		Predicted class	
		Positive	**Negative**
Actual class	**Positive**	TP	FN
	Negative	FP	TN

Accuracy: $\frac{TP+TN}{TP+FN+FP+TN}$

Precision: $\frac{TP}{TP+FP}$

Recall: $\frac{TP}{TP+FN}$

Parity-based fairness metrics:

Calculate some ratio of TP/FN/FP/TN for each subgroup and find the distance

E.g., Recall Parity: $\left|\frac{TP_A}{TP_A+FN_A} - \frac{TP_B}{TP_B+FN_B}\right|$

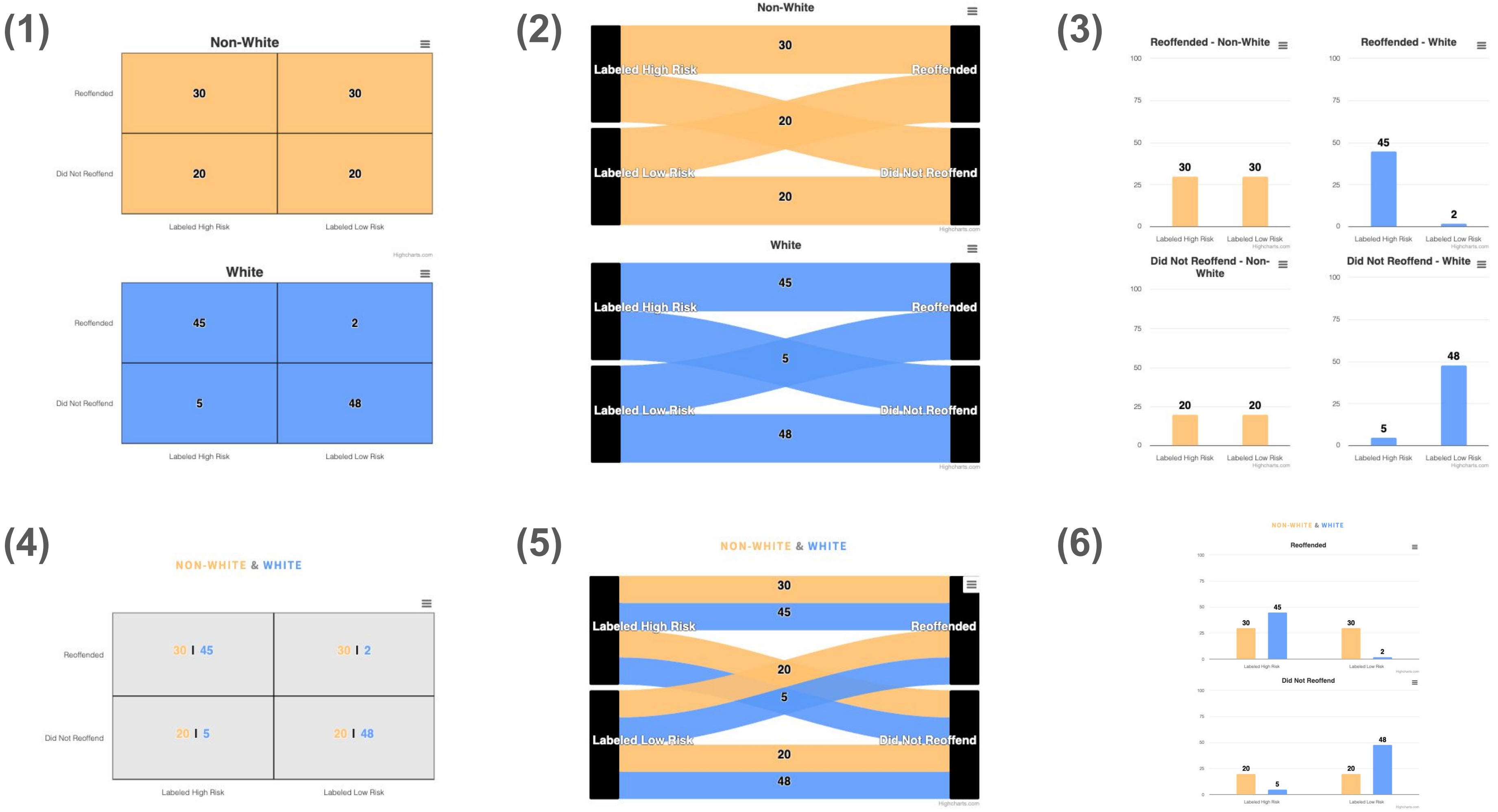
Exercise - 20 minutes

Survey: https://qualtricsxmr7dy5rld6.qualtrics.com/jfe/form/SV_1KS40JEIVYKsIbY

FAIR-EDU tool: https://fair-edu.github.io/FAIR-EDU/index.html

Based on **last name**, use the following visualization:

- A-D: Confusion Matrix
- E-H: Sankey Diagrams
- I-L: Bar Plots
- M-P: Confusion Matrix - Superimposed
- Q-U: Sankey Diagrams - Superimposed
- V-Z: Bar Plots - Superimposed



(1)
Non-White
Reoffended
Did Not Reoffend
30
30
20
20
Labeled High Risk
Labeled Low Risk
Highcharts.com
White
Reoffended
Did Not Reoffend
45
2
5
48
Labeled High Risk
Labeled Low Risk
(2)
Non-White
Labeled High Risk
Labeled Low Risk
Reoffended
Did Not Reoffend
30
20
20
Highcharts.com
White
Labeled High Risk
Labeled Low Risk
Reoffended
Did Not Reoffend
45
5
48
Highcharts.com
(3)
Reoffended - Non-White
30
30
Reoffended - White
45
2
Did Not Reoffend - Non-White
20
20
Did Not Reoffend - White
5
48
Labeled High Risk
Labeled Low Risk
Highcharts.com
(4)
NON-WHITE & WHITE
Reoffended
Did Not Reoffend
30 | 45
30 | 2
20 | 5
20 | 48
Labeled High Risk
Labeled Low Risk
(5)
NON-WHITE & WHITE
Labeled High Risk
Labeled Low Risk
Reoffended
Did Not Reoffend
30
45
20
5
20
48
Highcharts.com
(6)
NON-WHITE & WHITE
Reoffended
30
45
30
2
Did Not Reoffend
20
5
20
48

PREDICTIVE PARITY

FALSE DISCOVERY RATE PARITY

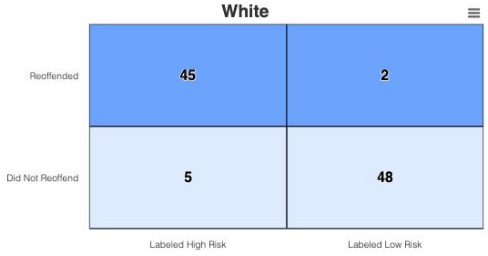
RECALL PARITY

This is a fairness metric that compares the proportion of true positives among all the actual positives for each sensitive group. The parity condition is satisfied when the recalls are equal across all groups.

$$\text{Recall} = \left| \frac{TP}{TP + FN} - \frac{TP}{TP + FN} \right|$$

$$\text{Recall} = \left| \frac{30}{30 + 30} - \frac{45}{45 + 2} \right|$$

$$\text{Recall} = 0.46$$



FALSE NEGATIVE RATE PARITY

NEGATIVE PREDICTIVE VALUE PARITY

FALSE OMISSION RATE PARITY

SPECIFICITY PARITY

FALSE POSITIVE RATE PARITY

OVERALL ACCURACY EQUALITY

Each of the metrics have two components that relate to what they measure:

1. Classification outcome (numerator)
2. Conditioning factor (denominator)

E.g., for Recall Parity, we look at the difference in true positives (labeled high risk and reoffended) conditioned on a positive class label (actually reoffended)

PREDICTIVE PARITY

FALSE DISCOVERY RATE PARITY

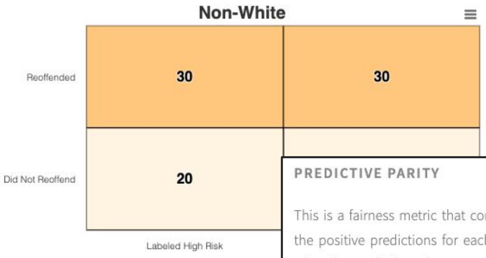
RECALL PARITY

This is a fairness metric that compares the proportion of true positives among all the actual positives for each sensitive group. The parity condition is satisfied when the recalls are equal across all groups.

$$\text{Recall} = \left|\frac{TP}{TP+FN} - \frac{TP}{TP+FN}\right|$$

$$\text{Recall} = \left|\frac{30}{30+30} - \frac{45}{45+2}\right|$$

$$\text{Recall} = 0.46$$



FALSE NEGATIVE RATE PARITY

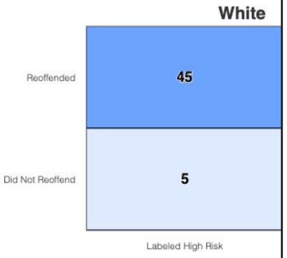
NEGATIVE PREDICTIVE VALUE PARITY

FALSE OMISSION RATE PARITY

SPECIFICITY PARITY

FALSE POSITIVE RATE PARITY

OVERALL ACCURACY EQUALITY



Each of the metrics have two components that relate to what they measure

1. Classification outcome (numerator)
2. Conditioning factor (denominator)

PREDICTIVE PARITY

This is a fairness metric that compares the proportion of true positives among all the positive predictions for each sensitive group. The parity condition is satisfied when the predictive values are equal across all groups.

$$\text{Predictive Parity} = \left|\frac{TP}{TP+FP} - \frac{TP}{TP+FP}\right|$$

$$\text{Predictive Parity} = \left|\frac{30}{30+20} - \frac{45}{45+5}\right|$$

$$\text{Predictive Parity} = 0.30$$

FALSE DISCOVERY RATE PARITY

RECALL PARITY

FALSE NEGATIVE RATE PARITY

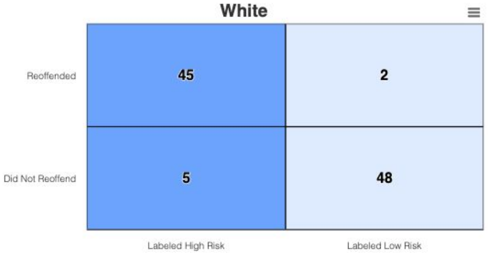
NEGATIVE PREDICTIVE VALUE PARITY

FALSE OMISSION RATE PARITY

SPECIFICITY PARITY

FALSE POSITIVE RATE PARITY

OVERALL ACCURACY EQUALITY



E.g., for Predictive Parity, we look at the difference in true positives (labeled high risk and reoffended) conditioned on a positive model prediction (labeled high risk)

When choosing fairness metrics,

1. Decide on the most critical classification outcome
   a. False positive? Critical for recidivism because someone could be denied parole
   b. False negative? Critical in a medical setting because we could fail to detect a disease
2. Decide on the conditioning factor
   a. What actually happened?
   b. What the model predicted to happen?

Note: the conditioning factor depends on the stakeholder's goals

Scenario #1: Recidivism

COMPAS is a machine learning model which predicts whether defendants are of high or low risk of reoffending if released on parole. However, civil rights groups have raised concerns that the model is less accurate for non-white defendants.

With recidivism, the worst-case scenario is when non-white defendants are more likely than white defendants to be denied parole because they are predicted to be high risk when they would not actually reoffend. Which fairness metric is most appropriate for measuring whether the model is fair according to this worst-case scenario?

- False Discovery Rate Parity
- False Positive Rate Parity
- Specificity Parity
- Overall Accuracy Equality

Classification Outcome: FP
Conditioning Factor: did not reoffend (FP + TN)

$$\text{False Positive Rate Parity} = \left| \frac{FP}{TN + FP} - \frac{FP}{TN + FP} \right|$$

Scenario #2: Election Forecasting

USNews has a machine learning model which predicts whether politicians will be elected based on factors such as leadership skills and whether they have a lot of funding. However, certain candidates have raised concerns that the model is biased towards candidates with influential family members who have previously held office.

With election forecasting, the worst-case scenario is that a model propagates biases by correctly predicting that politicians who have influential families are likely to get elected at a higher rate than for those without influential families. Which fairness metric is most appropriate for measuring whether the model is fair according to this worst-case scenario?

- ○ Negative Predictive Value Parity
- ○ Predictive Parity
- ○ Overall Accuracy Equity
- ○ Recall Parity

Classification Outcome: TP
Conditioning Factor: did not reoffend (TP + FP)

$$\text{Predictive Parity} = \left| \frac{TP}{TP + FP} - \frac{TP}{TP + FP} \right|$$

Scenario #3: Lung Cancer Detection

The CDC has a machine learning model which attempts to predict whether a patient has lung cancer from a number of attributes such as age, genetic risk, and whether or not they smoke. However, doctors have raised concerns that the model appears to be less accurate for women than for men.

With lung cancer detection, the worst-case scenario is that women are more likely than men to be predicted to be cancer free when they actually have lung cancer. Which fairness metric is most appropriate for measuring whether the model is fair according to this worst-case scenario?

- False Omission Rate Parity
- Predictive Parity
- False Negative Rate Parity
- Recall Parity

Classification Outcome: FN
Conditioning Factor: did not reoffend (TP + FN)

$$\text{False Negative Rate Parity} = \left| \frac{FN}{TP + FN} - \frac{FN}{TP + FN} \right|$$

Scenario #4: Job Hiring

Facebook has a machine learning model which screens resumes and predicts whether or not an applicant is qualified for the job based on factors such as educational background and prior experience. However, news outlets have been spreading stories that the model is biased against applicants with a low socioeconomic status.

With job hiring, the worst-case scenario is that a model propagates biases by predicting that qualified people with a low socioeconomic status are unqualified at a higher rate than for those with a high socioeconomic status. Which fairness metric is most appropriate for measuring whether the model is fair according to this worst-case scenario?

- False Positive Rate Parity
- False Discovery Rate Parity
- False Negative Rate Parity
- False Omission Rate Parity

Classification Outcome: FN
Conditioning Factor: did not reoffend (TN + FN)

$$\text{False Omission Rate Parity} = \left| \frac{FN}{TN + FN} - \frac{FN}{TN + FN} \right|$$

Additional Resources

- Python packages for measuring and addressing bias in ML models
  - **AIF360:** https://aif360.readthedocs.io/en/stable/
  - **Dalex:** https://dalex.drwhy.ai/python-dalex-fairness.html
  - **Fairlearn:** https://fairlearn.org/v0.8/auto_examples/index.html
  - **Responsibly:** https://docs.responsibly.ai/_modules/responsibly/fairness/metrics/visualization.html
  - **Google What-If Tool:** https://pair-code.github.io/what-if-tool/
    - Jupyter extension for analyzing models
- Useful links
  - **Wikipedia - Fairness:** https://en.wikipedia.org/wiki/Fairness_(machine_learning)
  - **Fairness and Machine Learning (fairmlbook):** https://fairmlbook.org/index.html