MACHINE LEARNING MODELS WORKSHOP I:

METHODS FOR DETECTING & CORRECTING BIAS

AGENDA

- Introduction
- Importance of Fairness
- Definition of Algorithmic Fairness
- Detect/Measure Biasness – Theory and Python Hands on
- Remove/reduce Biasness – Theory and Python Hands on

LEGALLY RECOGNIZED PROTECTED CLASSES (USA)

RACE Civil Risk Act 1964	**Color** Civil Rights Act 1964	**Sex** Equal Pay Act 1963, Civil Rights Act 1964
Religion Civil Risk Act 1964	**National Origin** Civil Rights Act 1964	**Citizenship** Immigration Reform and Control Act
Age Age Discrimination in Employment Act 1967	**Pregnancy** Pregnancy Discrimination Act	**Family Status** Civil Rights Act 1968
Disability Status Rehabilitation Act 1973, Americans with Disabilities Act 1990	**Veterans Status** Vietnam Era Veterans' Readjustment Assistance Act 1974, Uniformed Services Employment and Reemployment Rights Act	**Genetic Information** Genetic Information Nondiscrimination Act

REGULATED DOMAINS - USA

Credit
Equal Credit Opportunity Act

Employment
Civil Rights Act 1964

Education
Civil Rights Act 1964, Education Amendments 1972

Housing
Fair Housing Act

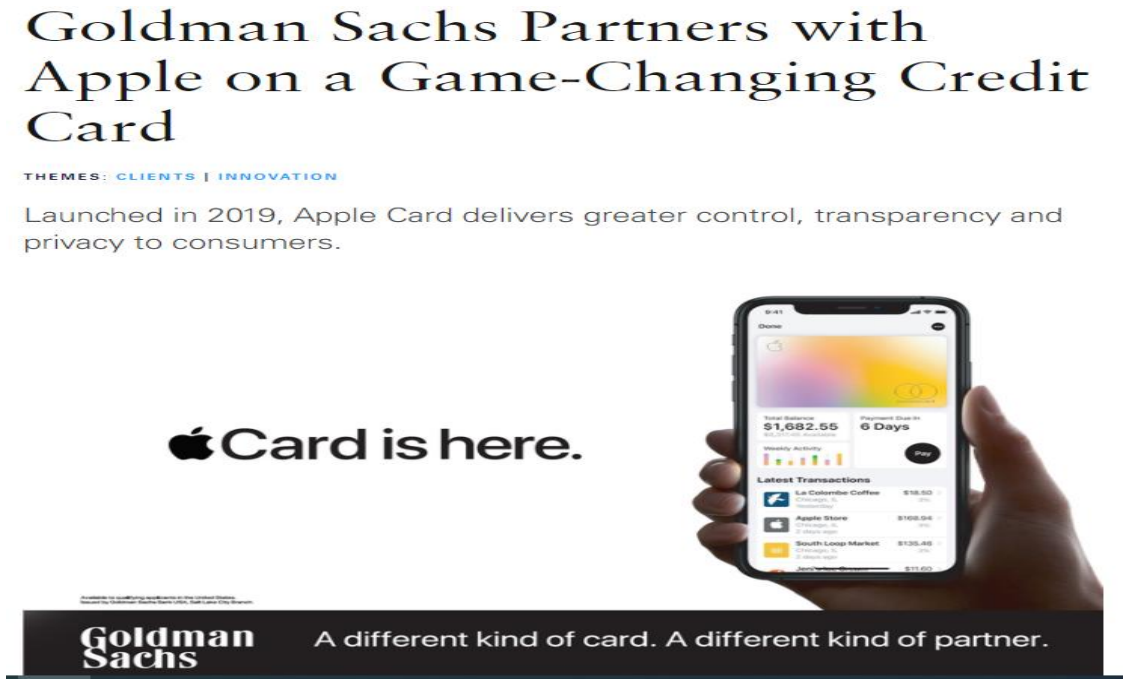
Public Accommodation
Civil Rights Act 1964

Extends to Marketing and Advertising
Not Limited to Final Decision

The above List sets aside complex web of Laws that regulates the government

EXAMPLES OF BIASNESS

Apple, Goldman Face Criticism Over Alleged Sexism in Credit Card Algorithm



Amazon scraps secret AI recruiting tool that showed bias against women

INTRODUCTION

- Bias in this context is unfairness (more or less)
- True definition is a wild good chase. Lack of concrete definition. How to develop model correctly?
- Researchers and practioners have come up with many definition of fairness
- Predictive Accuracy - Model developers/decision makers goal is to maximize accuracy subject to fairness constraint
- Standard statistical bias in machine learning (the bias in the bias vs. variance tradeoff)
- Bias = Expected value of model - true value
- Is statistical bias an adequate fairness criteria?
- Why statistical bias is not enough?
  - Error or distribution of errors
  - Data bias
- Real challenge is how to make algorithms systems support human values? We need to align with this objective

WHY FAIRNESS IS IMPORTANT

Fairness is becoming one of the most popular topics in machine learning in recent years. The main motivation is that it is highly related to our own benefits

- We are at an age where many things have become or are becoming automated by ML systems. e.g.
  - Driverless cars are around the corner and are estimated to be widely used within 5–10 years;
  - Employers use ML system to select job applicants;
  - Courts in United States use COMPAS algorithm for recidivism prediction;
  - Linked-in uses ML to rank job candidates queried;
  - Amazon uses recommender system to recommend items and decide the order of items appearing on a page.
  - Netflix uses recommender system to present customized page for every user
- Machine learning systems have been an inseparable part of our daily lives. They will be even more widely used in the near future as more and more fields begin to integrate AI into their existing practice/products
- AI is good, but it can be used incorrectly. ML, the most widely used AI technique, relies heavily on data
- It is a common misconception that AI is absolutely objective. AI is objective only in the sense of learning what human teaches. The data provided by human can be highly-biased

FAIRNESS TERMS ONE SHOULD KNOW

Protected Attribute

An attribute that partitions a population into groups whose outcomes should have parity (e.g. race, caste, gender, religion)

Group Fairness

Groups defined by protected attributes receiving similar treatments or outcomes

Fairness Metric

Measure of unwanted bias in training data or models

Privileged Protected Attribute

A protected attribute value indicating a group that has historically been at systematics advantage

Individual Fairness

Similar Individuals receiving similar treatments or outcomes

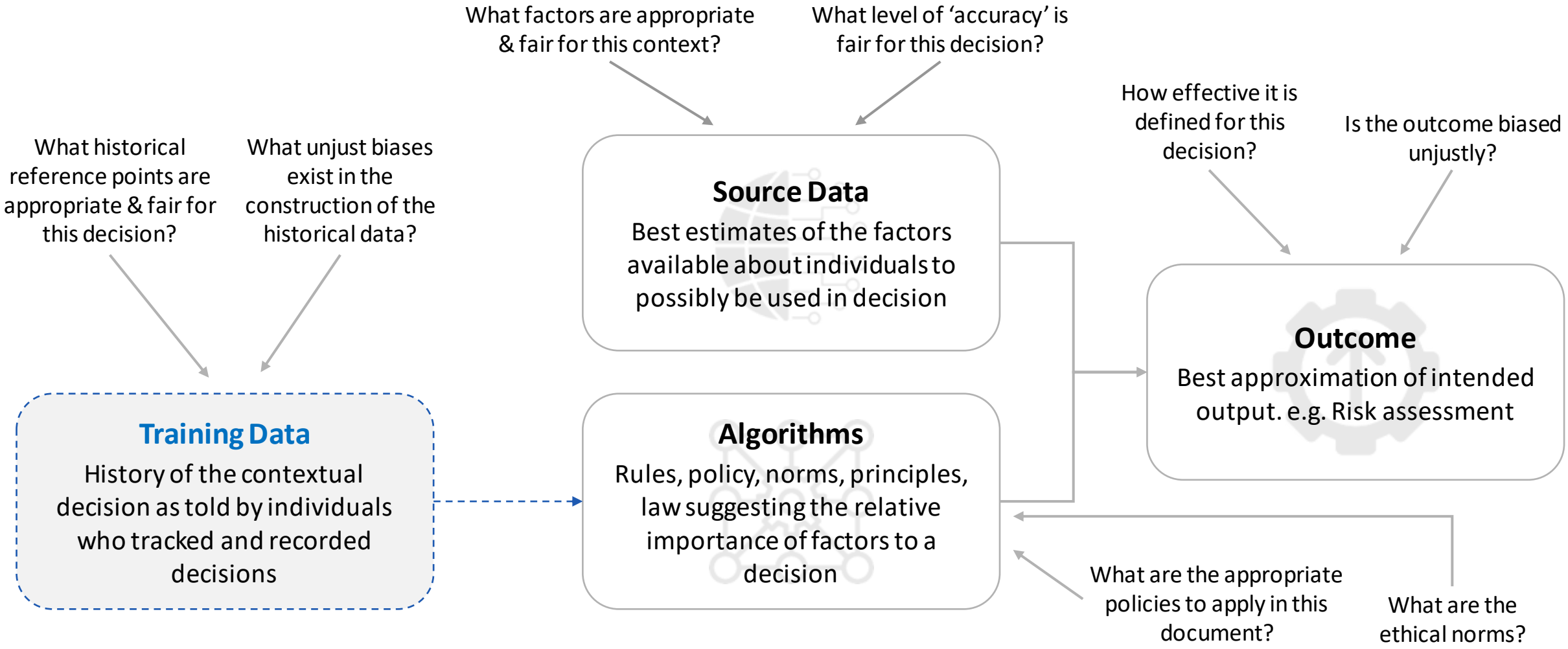
Favourable Label

A label whose value corresponds to an outcome that provides an advantage to recipient

FAIRNESS IN MACHINE LEARNING – SOME STANDARD TERMINOLOGY

- A **favorable label** is a label whose value corresponds to an outcome that provides an advantage to the recipient. Examples are receiving a loan, being hired for a job, and not being arrested.
- A **protected attribute** is an attribute that partitions a population into groups that have parity in terms of benefit received. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific.
- A **privileged value** of a protected attribute indicates a group that has historically been at a systematic advantage
- **Group fairness** is the goal of groups defined by protected attributes receiving similar treatments or outcomes
- **Individual fairness** is the goal of similar individuals receiving similar treatments or outcomes
- **Bias** is a systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage.
- A **fairness metric** is a quantification of unwanted bias in training data or models
- A bias mitigation algorithm is a procedure for reducing unwanted bias in training data or models

ML MODEL DEVELOPMENT



ML/AI MODELS: FTC, FRCA AND ECOA EXPECTATIONS

Be Transparent

- Don't deceive consumers about how you use your ML models
- Be transparent when collecting sensitive data
- If you make automated decisions based on information from a third-party vendor, you may be required to provide the consumer with an "adverse action" notice.

Explainability

- Explain if you deny consumers something based on algorithmic decision-making
- For algorithm-based risk scores to consumers, disclose the key factors that affected the score, rank ordered for importance
- Tell consumers if terms of a deal changes based on automated tools

Sound Data and Models

- Written policies and procedures to ensure that the data they furnish is accurate and has integrity
- Ensuring that the data is accurate and up to date
- Make sure that ML/AI models are validated and revalidated to ensure that they work as intended, and do not illegally discriminate

Ensure Fairness

- Don't discriminate based on protected classes
- Focus on inputs, but also on outcomes
- Give consumers access & an opportunity to correct information used to make decisions about them

Accountability

- Ask questions:

a) How representative is your data set?
b) Does your data model account for biases?
c) How accurate are your predictions based on big data?
d) Does your reliance on big data raise ethical or fairness concerns?

- Consider your accountability mechanism
- Protect your algorithm from unauthorized use.

MACHINE LEARNING: BIAS IN – BIAS OUT

Bias in Data

Data is often heterogeneous, generated by subgroups with their own characteristics and behaviors.

- Historical bias
- Representation bias
- Measurement bias
- Population bias
- Collection bias
- External source bias

Data Cleansing and Variable Selection

Algorithmic bias is added purely by the algorithm

- Sampling bias
- Missing values & outliers bias
- Omitted variable bias
- Cause and effect bias
- Inappropriate qualitative analysis bias
- Transformation, cleansing and normalization bias

Algorithmic Bias

Algorithmic bias is added purely by the algorithm

- Choice of methodology
- Estimation methodology bias
- Aggregation bias
- Non-interpretable and non-explainable models
- Model automation bias

Evaluation Bias

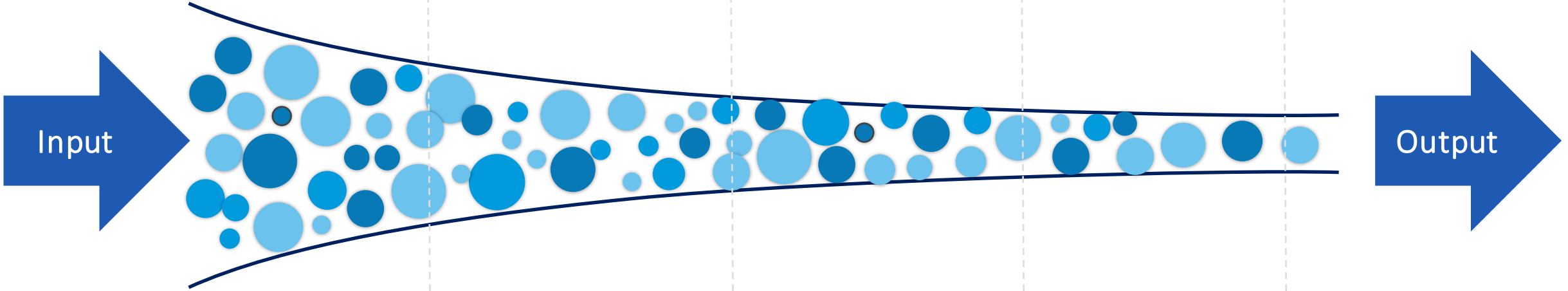
Evaluation bias happens during model evaluation

- Ranking bias
- Equal Metrics bias
- Inappropriate benchmark
- Performance metrics bias
- Replicability and reproducibility bias

Interaction Bias

Triggered from two sources - the user interface and through the user

- Presentation bias
- Linking bias
- Content Production bias
- Emergent bias



IMPACT OF BIAS IN FINANCIAL SERVICES ML MODELS

Consumer Financial Protection Problem

- Data and ML bias can lead to problems: product steering, discriminatory pricing, unfair credit rationing, exclusionary filtering, accountability and digital redlining
- The stakes become higher when developers does not care of fairness and ethical issues from the start
- Innovative tools, oversight policy and frameworks would be useful for consumer financial protection regulatory regimes such as FTC, FCRA, ECOA, BSA and AML

Discrimination

- If the data and ML models are biased, they could disparate impact the way millions of consumers borrow, save and manage their money.
- Differentiate customers w.r.t to age, ethnicity, color, race, religion, gender, locality, profession, legal etc.,
- Discrimination on the basis of any protected class is illegal under the US Equal Credit Opportunity Act of 1974.

Interpretability

- Data bias and bias in ML model development stages can lead to ubiquity, unfair, opaque and non-interpretable models
- Transparency around ML models are regulators' greatest obstacle
- Need of robust oversight to ensure ML applications remain accountable to society, the people and government and circumvent discriminatory bias

Vendor Management

- The black box and complex ML algorithms offered by third-parties and opens sources may not perform as intended and may lack data protection, interpretability and quality documentation.
- Such gaps calls for conducting third-party risk management (SR 13-19), data protection standards, SR 11-7 compliant documentation, and technology management guidelines geared toward AI applications.

Supervision

- As data set, ML algorithms and model complexity increases, so the models will become even more inscrutable
- The shift toward relatively more sophisticated, non-linear (often metaphorically described as "black box") models necessitates new governance processes
- Call for continuously monitoring performance and timely maintenance of developed models.

Coffee Break

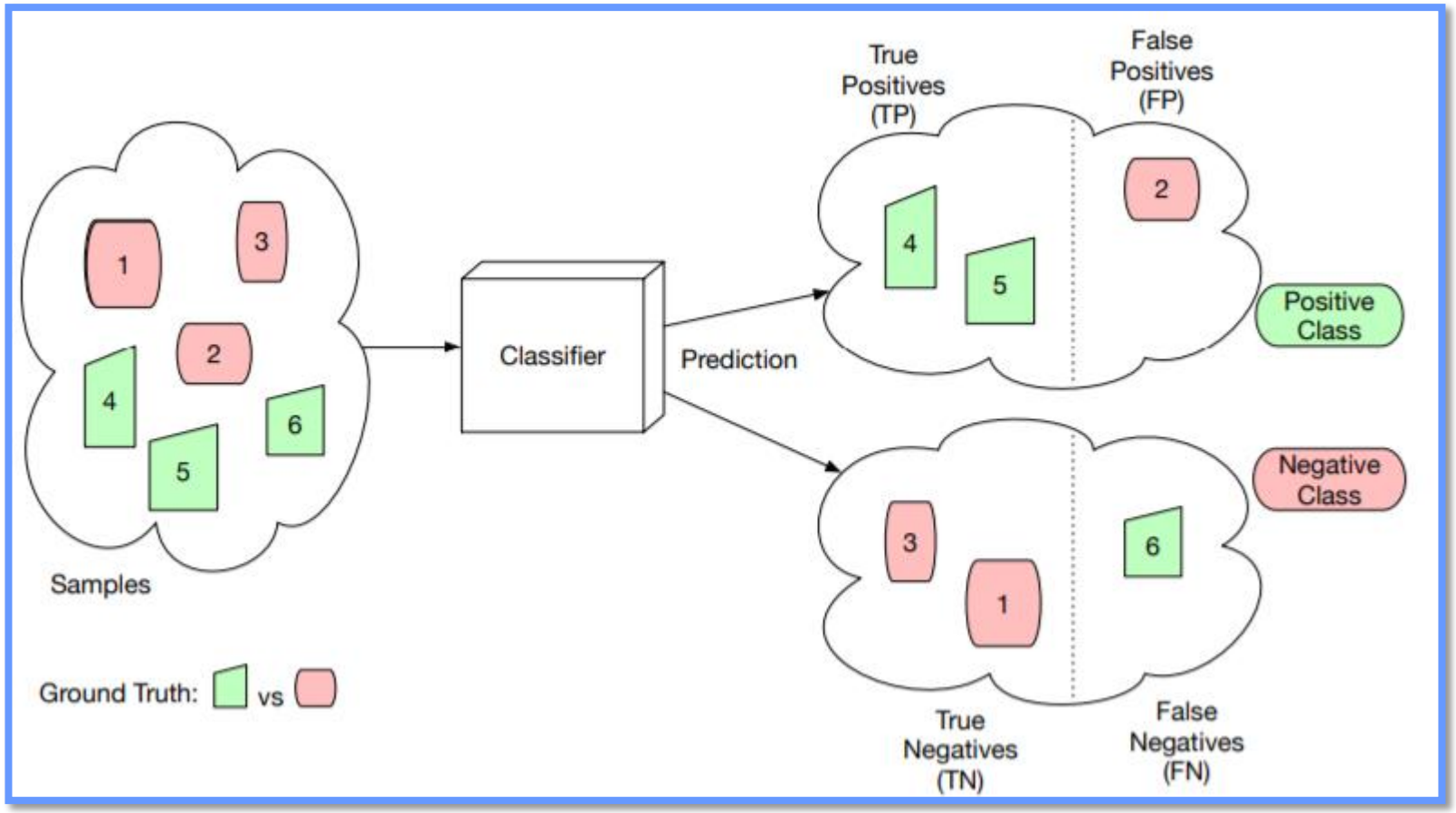
Lets Meet in 15 Minutes

FAIRNESS METRICS EVALUATION SETTING

- $X \in \mathbb{R}^d$ denote a non-sensitive feature vector describing an individual.
- $g \in \{0,1\}$ denote sensitive binary feature describing the gender of the individual. g is assumed to be a binary variable representing the sex (male or female) of the credit seeker.
- $Y \in \{0,1\}$ represent the actual or true labels. Y is either an approved or rejected decision (or good vs. bad credit). As such, it must be denoted as a binary variable in the dataset.
- $S \in [0,1]$ is the predicted probability. For a logistic regression model, it is defined as:

- $S = Pr(Y = 0, 1 \mid X, g) = \exp[\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta g g] / (1 + \exp[\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta g g])$
  where α, β_i, and βg are the estimated regression coefficients.
- $\hat{y} \in \{0,1\}$ is the prediction decision of the algorithm.
- It depends on the predicted probability (S) of an applicant having good or bad credit. For a threshold value of S^*, $\hat{y} = 1$ when $S > S^*$, and $\hat{y} = 0$ when $S < S^*$.

OVERVIEW OF CLASSIFICATION



DISCRIMINATION IN MACHINE LEARNING

Disparate Treatment	▪ Situation where a policy or decision is made that explicitly discriminates against a demographic group of people based on race, age, gender etc. ▪ Due to the fairness blind optimization functions often used in practice, a machine learning algorithm may end up making decisions based on a protected attribute like gender
Disparate Impact	▪ Disparate Impact involves policies or decisions that implicitly discriminate against a certain group of people. ▪ A decision making system suffers from disparate impact if it provides outputs that benefits (hurt) a group of people sharing a value of a sensitive feature more frequently than other groups of people . ▪ With the massive amounts of data we feed our machine learning algorithms, they are likely to find features that correlate with sensitive attributes such as race. ▪ Therefore even if an algorithm is not explicitly trained to make unequal decisions for different demographic groups, it may do just so.

GROUP LEVEL FAIRNESS

Group-level notions of fairness require just (equal) distributions of benefit or harm across groups. Metrics belonging to this category are based on observational criteria as they depend on one of the following joint distributions

Subcategory 1

- Distribution between predicted outcome $\hat{y}$ and sensitive attribute g (marginalizing features X and the actual outcome Y)
- There are two metrics that belong to this category commonly known as **demographic parity** and **conditional demographic parity**

Subcategory 2

- Distribution between predicted outcome $\hat{y}$, sensitive attribute g, and the actual outcome Y (marginalizing features X)
- Metrics belonging to this subcategory look beyond the absolute distribution of predictions $\hat{y}$ by conditioning fairness on accuracy or imperfections of the model

Subcategory 3

- The fairness notions in this subcategory are based on some distribution between predicted probability S, sensitive attribute g, and the actual outcome Y

Source: Verma and Rubin (2018), Fairness Definition Explained, 2018 ACM/IEEE International Workshop on Software Fairness

SUBCATEGORY1 FAIRNESS METRICS

Fairness Metrics

- **Demographic parity** requires that the positive prediction between the two groups of males and females be equal
- **Conditional demographic parity** extends the definition of demographic parity by conditioning the outcome to be based on certain legitimate factors (L)
- While there is no unified agreement on what legitimate factors are, it is useful to mention that conventionally, legitimacy o f loan approval decisions is based on evaluating key factors, commonly known as credit history, character, collateral, and capacity

	Metric	Mathematical notation	Requires
Independence	Demographic parity	$Pr(\hat{y} = 1 \mid g = 0) = Pr(\hat{y} = 1 \mid g = 1)$	Equal probability of positive predictions for both males and females.
	Conditional demographic parity	$Pr(\hat{y} = 1 \mid g = 0, L) = Pr(\hat{y} = 1 \mid g = 1, L)$	Equal probability of positive predictions for both males and females, but condition on a subset of legitimate input features.

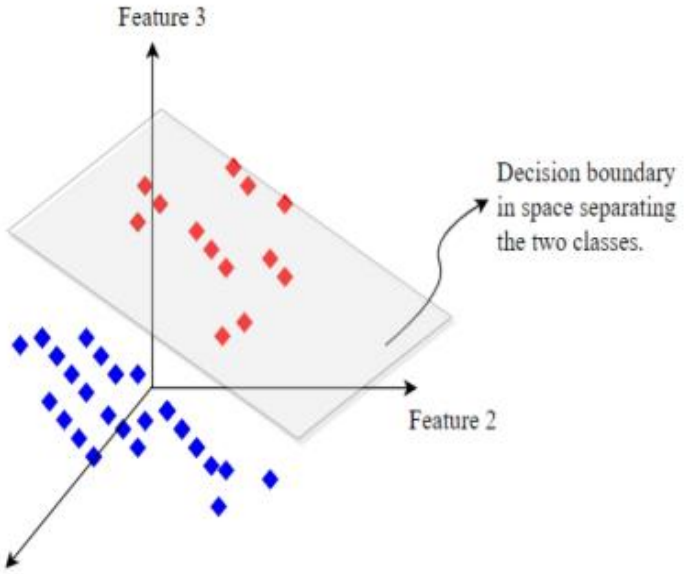
Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

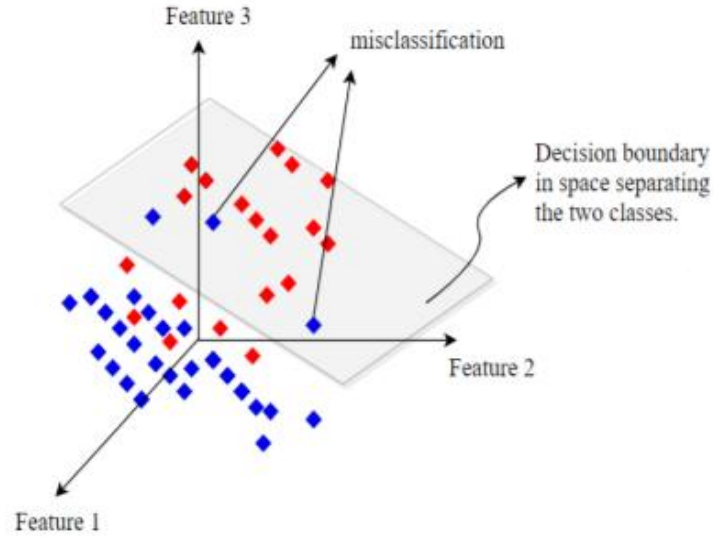
- The two metrics require that the decision outcomes $\hat{y}$ be uncorrelated with sensitive features g: $\hat{y} \perp g$.
- Parity measures can be problematic?

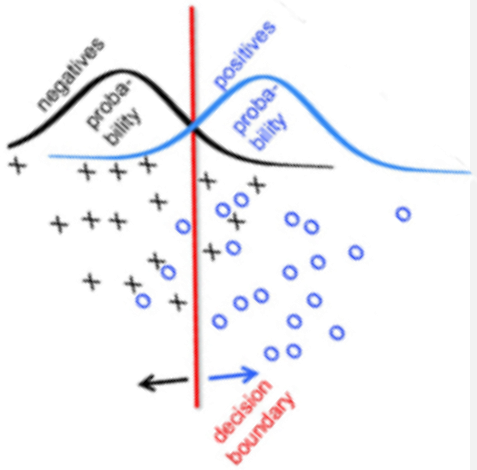
SUBCATEGORY2 FAIRNESS METRICS (1/6)

Fairness Metrics

- Metrics belonging to this **subcategory 2** look beyond the absolute distribution of predictions ŷ by conditioning fairness on accuracy or imperfections of the model.
- In an ideal case, a logistic regression model or other ML model may find a decision boundary such that the two binary classes are linearly separable and that output decisions are 100% accurate.







- If the items in the training data are not perfectly, linearly separable, the hyper plane/linear boundary estimated will misclassify some of the data points.

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

SUBCATEGORY2 FAIRNESS METRICS (2/6)

Do outcomes systematically different between demographic group or other population group?

Predicted	Actual			
	Positive Class		Negative Class	
Positive Class	**True Positive**	Hit	**False Positive**	Type I Error (False Alarm)
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Bad** transaction predicted	**Prediction**	**Bad** Transaction predicted
	Results	Good	**Results**	Unnecessary work and cost
Negative Class	**False Negative**	Type II Error (Miss)	**True Negative**	Correct reject
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Good** transaction predicted	**Prediction**	**Good** transaction predicted
	Results	Not a good sign	**Results**	Good

- Applicable to many context – e.g. Loan (default versus non default), Hiring (Succeed at Job vs. did not)
- Yields useful insights
- What do different stakeholders want of the binary classifier (Logistic regression, ANN, GBM, SVM etc)?
- E.g. Decision maker/model developers and Defendants (wrongly classified as future criminals)

SUBCATEGORY2 FAIRNESS METRICS (3/6)

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

If the model satisfies predictive parity but the prevalence differs between groups, the model can achieve equal false positive rates and equal false negative rates across the groups.

SUBCATEGORY2 FAIRNESS METRICS (4/6)

✓	**True positive (TP)**	A case when the predicted and actual outcomes are both in the positive class
✓	**False positive (FP)**	A case predicted to be in the positive class when the actual outcome belongs to the negative class
✓	**False negative (FN)**	A case predicted to be in the negative class when the actual outcome belongs to the positive class
✓	**True negative (TN)**	A case when the predicted and actual outcomes are both in the negative class
✓	**Positive predictive value (PPV)**	The fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, TP/(TP+FP). PPV is often referred to as precision, and represents the probability of a subject with a positive predictive value to truly belong to the positive class, P(Y = 1|d = 1). In our example, it is the probability of an applicant with a good predicted credit score to actually have a good credit score
✓	**False discovery rate (FDR)**	The fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases, FP/(TP+FP) . FDR represents the probability of false acceptance, P(Y = 0|d = 1), e.g., the probability of an applicant with a good predicted credit score to actually have a bad credit score
✓	**False omission rate (FOR)**	The fraction of positive cases incorrectly predicted to be in the negative class out of all predicted negative cases, FN/(TN +FN) . FOR represents the probability of a positive case to be incorrectly rejected, (P(Y = 1|d = 0)), e.g, the probability of an applicant with a bad predicted credit score to actually have a good score

SUBCATEGORY2 FAIRNESS METRICS (5/6)

✓	**Negative predictive value (NPV)**	The fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, TN/(TN +FN) . NPV represents the probability of a subject with a negative prediction to truly belong to the negative class, P(Y = 0|d = 0), e.g., the probability of an applicant with a bad predicted credit score to actually have such score
✓	**True positive rate (TPR)**	The fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, TP/(TP+FN) . TPR is often referred to as sensitivity or recall; it represents the probability of the truly positive subject to be identified as such, P(d = 1|Y = 1). In our example, it is the probability of an applicant with a good credit score to be correctly assigned with such score
✓	**False positive rate (FPR)**	The fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases, FP/(FP+TN) . FPR represents the probability of false alarms – falsely accepting a negative case, P(d = 1|Y = 0), e.g., the probability of an applicant with a actual bad credit score to be incorrectly assigned with a good credit score
✓	**False negative rate (FNR)**	The fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases, FN/(TP+FN). FNR represents the probability of a negative result given an actually positive subject, P(d = 0|Y = 1), e.g., the probability of an applicant with a good credit score to be incorrectly assigned with a bad credit score
✓	**True negative rate (TNR)**	The fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, TN/(FP+TN) . TNR represents the probability of a subject from the negative class to be assigned to the negative class, P(d = 0|Y = 0), e.g., the probability of an applicant with a bad credit score to be correctly assigned with such score

SUBCATEGORY2 FAIRNESS METRICS (6/6)

	Metric	Mathematical notation/Alias	Requires
Accuracy/Sufficiency	Predictive parity	$Pr(Y=1 \mid \hat{y}=1, g=0) = Pr(Y=1 \mid \hat{y}=1, g=1)$ *False discovery rate:* $Pr(Y=0 \mid \hat{y}=1, g=0) = Pr(Y=0 \mid \hat{y}=1, g=1)$	Equal fraction of correct positive predictions for both male and females.
	Overall accuracy equality	$Pr(\hat{y}=Y, g=0) = Pr(\hat{y}=Y, g=1)$ *Overall misclasification rate:* $Pr(Y \neq \hat{y} \mid g=0) = Pr(Y \neq \hat{y} \mid g=1)$	Equality of positive and negative predictive values for both males and females.
	Conditional accuracy equality	$Pr(Y=1 \mid \hat{y}=1, g=0) = Pr(Y=1 \mid \hat{y}=1, g=1) \cap$ $Pr(Y=0 \mid \hat{y}=0, g=0) = Pr(Y=0 \mid \hat{y}=0, g=1)$	Equal correct positive predictions and equal correct negative predictions for both males and females.
Misrepresentation/Separation	False positive error rate	$Pr(\hat{y}=1 \mid Y=0, g=0) = Pr(\hat{y}=1 \mid Y=0, g=1)$ *True negative rate:* $Pr(\hat{y}=0 \mid Y=0, g=0) = Pr(\hat{y}=0 \mid Y=0, g=1)$	Equal incorrect positive predictions for both male and females.
	False negative error rate	$Pr(\hat{y}=0 \mid Y=1, g=0) = Pr(\hat{y}=0 \mid Y=1, g=1)$ *True negative rate:* $Pr(\hat{y}=0 \mid Y=0, g=0) = Pr(\hat{y}=0 \mid Y=0, g=1)$	Equal incorrect negative predictions for both male and females.
	Equalized odds	$Pr(\hat{y}=1 \mid Y=0, g=0) = Pr(\hat{y}=1 \mid Y=0, g=1) \cap$ $Pr(\hat{y}=0 \mid Y=1, g=0) = Pr(\hat{y}=0 \mid Y=1, g=1)$	Equality of false positive and false negative error rates for both males and females.
	Treatment equality	(FNR/FPR) for males = (FNR/FPR) for females	Equal ratio of false positive and false negative error rates for both males and females.

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

Two statistical fairness criteria, separation and sufficiency are introduced.

Sufficiency requires the true label Y to be statistically independent of group-membership conditional on prediction outcome ŷ (or decision):
$Y \perp g \mid \hat{y}$.

Individuals about whom the same decision is made have the same statistical prospects of being either true label, regardless of their irrelevant feature.

Separation is satisfied when the prediction/decision ŷ is statistically independent of group-membership g conditional on the true label
Y: $\hat{y} \perp g \mid Y$.

Individuals with the same true label have the same statistical prospects of either decision, regardless of their irrelevant feature

Coffee Break

Lets Meet in 15 Minutes

EVALUATE FOR FAIRNESS

Female Applicants Results	
True Positive (TP) = 10	False Positive (FP)=1
False Negative (FN)=1	True Negative (TN)=488

Male Applicants Results	
True Positive (TP) = 6	False Positive (FP)=3
False Negative (FN)=5	True Negative (TN)=48

Precision =	TP/(TP+FP) =	10/(10+1) =	0.9090

Predictive parity

Recall=	TP/(TP+FN) =	10/(10+1) =	0.9090

Equality of Opportunity

Precision =	TP/(TP+FP) =	6/(6+3) =	0.6666

Recall=	TP/(TP+FN) =	6/(6+5) =	0.5454

SUBCATEGORY3 FAIRNESS METRICS

Fairness Metrics

- The fairness notions in this subcategory are based on some distribution between predicted probability S, sensitive attribute g, and the actual outcome Y.
- The fairness criteria imposed on these metrics are in correspondence to their similar metric in Subcategory 2

Metric	Mathematical Notation	Requires	Similar to
Calibration	$Pr(Y=1 \mid S=s, g=0) = Pr(Y=1 \mid S=s, g=1)$	Equal probability to belong to true positive label.	Predictive parity (except it considers fraction of positive predictions)
Well-calibration	$Pr(Y=1 \mid S=s, g=0) = Pr(Y=1 \mid S=s, g=1) = s$	Equal probability to belong to true positive label and the probability to equal S.	Predictive parity (except it considers fraction of positive predictions)
Balance for positive class	$E(S \mid Y=1, g=0) = E(S \mid Y=1, g=1)$	Equal expected value of probability S for individuals with positive true label.	• False negative error rate • True positive rate.
Balance for negative class	$E(S \mid Y=0, g=0) = E(S \mid Y=0, g=1)$	Equal expected value of probability S for individuals with negative true label.	• False positive error rate balance • true negative rate

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

Ideally, one would desire to achieve an algorithmic decision-making model that satisfies all the group level notions

But this is not achievable unless for some improbable circumstances.

Consequently, careful choices have to be made between one subcategory and another

EVALUATE FOR FAIRNESS

Female Applicants Results	
True Positive (TP) = 10	False Positive (FP)=1
False Negative (FN)=1	True Negative (TN)=488

Male Applicants Results	
True Positive (TP) = 6	False Positive (FP)=3
False Negative (FN)=5	True Negative (TN)=48

Precision =	TP/(TP+FP) =	10/(10+1) =	0.9090

Predictive parity

Recall=	TP/(TP+FN) =	10/(10+1) =	0.9090

Equality of Opportunity

Precision =	TP/(TP+FP) =	6/(6+3) =	0.6666

Recall=	TP/(TP+FN) =	6/(6+5) =	0.5454

INDIVIDUAL FAIRNESS

"Treat similar examples similarly" i.e. similar individuals should be treated similarly

Flaws

- It is hard to determine what is an appropriate metric function to measure the similarity of two inputs
- In our case, it is hard to quantify the difference between two job candidates. Imagine three job applicants, A, B and C
  - A has a bachelor degree and 1-year related work experience
  - B has a master degree and 1-year related work experience
  - C has a master degree but no related work experience
- Is A closer to B than C? If so, by how much? Such question is hard to answer
- It becomes even worse when the sensitive attribute(s) comes into the play

SIMILARITY –BASED MEASURES (INDIVIDUAL LEVEL)

Metric	Mathematical notation	Requires
Causal discrimination	$(X_i = X_j) \cap (g_i \mathrel{!}= g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification.
Fairness through awareness	$(X_i = X_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification, and no sensitive feature is explicitly used in the model.
Fairness through unawareness	$(X_i \sim X_j) \cap (g_i \mathrel{!}= g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Similar individuals receive similar classification

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

- Causal discrimination requires the model to be trained on the sensitive feature g
- Fairness through unawareness excludes sensitive attribute g from the feature set
- Ensuring that identical or similar individuals that only differ by gender receive similar classification outcomes and failing to do so is an indication of direct discrimination or indirect discrimination

AGENDA – SESSION II

- Estimating Fairness Metrics Using Python
- Pre-Processing
- In-Processing
- Post processing
- Exercise of Some Popular Techniques - Python

MACHINE LEARNING MODELS WORKSHOP II:

METHODS FOR DETECTING & CORRECTING BIAS

CONFUSION MATRIX

Do outcomes systematically different between demographic group or other population group?

Predicted	Actual			
	Positive Class		Negative Class	
Positive Class	**True Positive**	Hit	**False Positive**	Type I Error (False Alarm)
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Bad** transaction predicted	**Prediction**	**Bad** Transaction predicted
	Results	Good	**Results**	Unnecessary work and cost
Negative Class	**False Negative**	Type II Error (Miss)	**True Negative**	Correct reject
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Good** transaction predicted	**Prediction**	**Good** transaction predicted
	Results	Not a good sign	**Results**	Good

- Applicable to many context – e.g. Loan (default versus non default), Hiring (Succeed at Job vs. did not)
- Yields useful insights
- What do different stakeholders want of the binary classifier (Logistic regression, ANN, GBM, SVM etc)?

E.g. Decision maker/model developers and Defendants (wrongly classified as future criminals)

FAIRNESS METRICS

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

GROUP FAIRNESS

Metrics	Equalized Under
Equal Selection of probability between different groups	Demographic Parity
Equal Positive Predictive value between different groups	Predictive Parity
Equal Negative Predicted value between groups	Predictive Parity
Equal False Positive Rates between different groups	Error Rate Balance
Equal False Negative Rates between different groups	Error Rate Balance
Equal Accuracy rates between different groups	Accuracy Equity

Different metrics matter to different stakeholders. There is no right definition of fairness.

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

FAIRNESS IN MATHEMATICAL TERMS

Demographic Parity: Demographic or statistical parity suggests that a predictor is unbiased if the prediction $\hat{Y}$ is independent of the protected attribute p so that $\Pr(\hat{Y}|p) = \Pr(\hat{Y})$

Here, the same proportion of each population are classified as positive. However, this may result in different false positive and true positive rates if the true outcome y does actually vary with the protected attribute p

Deviations from statistical parity are sometimes measured by the statistical parity difference or the disparate impact which replaces the difference in this equation with a ratio.

Equality of Odds: Equality of odds is satisfied if the prediction $\hat{Y}$ is conditionally independent to the protected attribute p, given the true value Y: $\Pr(\hat{Y}|Y,p) = \Pr(\hat{Y}|Y)$

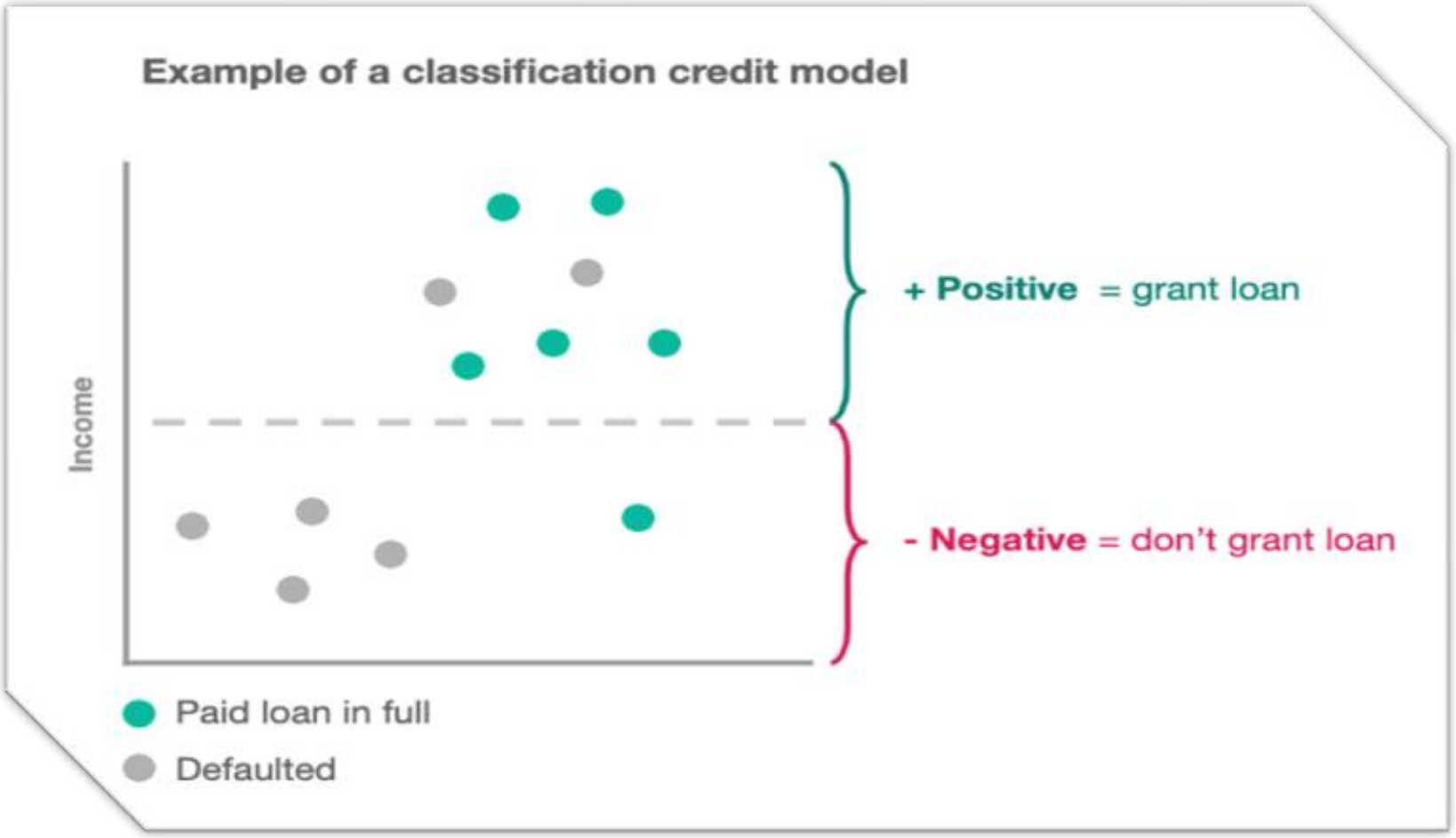
This means that the true positive rate and false positive rate will be the same for each population; each error type is matched between each group.

Equality of Opportunity: It has the same mathematical formulation as equality of odds, but is focused on one particular label Y=1 of the true value so that:$\Pr(\hat{Y}|Y = 1,p) = \Pr(\hat{Y}|Y = 1)$

In this case, we want the true positive rate to be the same for each population with no regard for the errors when Y=0. In effect it means that the same proportion of each population receive the "good" outcome Y=1

Deviation from equality of opportunity is measured by the equal opportunity difference:

FAIRNESS IN MACHINE LEARNING



- Bank A is building a ML based credit model purely on "income" variable.
- ML model will learn the income pattern of applicants who can pay their loan in full and differentiate them from others who may default.
- ML based need to set an income threshold using the in-sample data/ training set to decide who gets a loan.
- Applicants above the threshold, will get the loan (positive predictions).
- Applicants below the threshold who will not get a loan (negative predictions).

Image Credit: Valeria Cortez, How to define fairness to detect and prevent discriminatory outcomes in Machine Learning, Sept 24, 2019

DEMOGRAPHIC/STATISTICAL PARITY

Demographic Parity (Mitigating Bias in ML)

- One of the simplest and intuitive definitions of fairness and widely known mitigation strategies for bias in ML.
- Have a collection of pre-defined groups and then use parity of some statistic of the outcome across these groups.
- Assume there exist two groups A and B. **75%** of all the applicants are in group A and **25%** of them are in group B
- Apart from their income, there is no difference between group A and B applicants. All the other features are same.
- In case, the credit decisioning model using machine learning algorithm satisfies demographic parity, we expect about 75% of group A and 25% of group B applicants within the "Acceptable" range.
- We expect the fractions of group A and B applicants in the global population to be reflected in the "favorable" group of "Acceptable" applicants.
- An unfair algorithm, that "favors" group A and discriminates against group B, would put more than 75% of group A in the "Acceptable" bin.
- Here, demographic parity is a perfectly fine measure.

https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-three-framework/fairness-criteria/

DEMOGRAPHIC/STATISTICAL PARITY

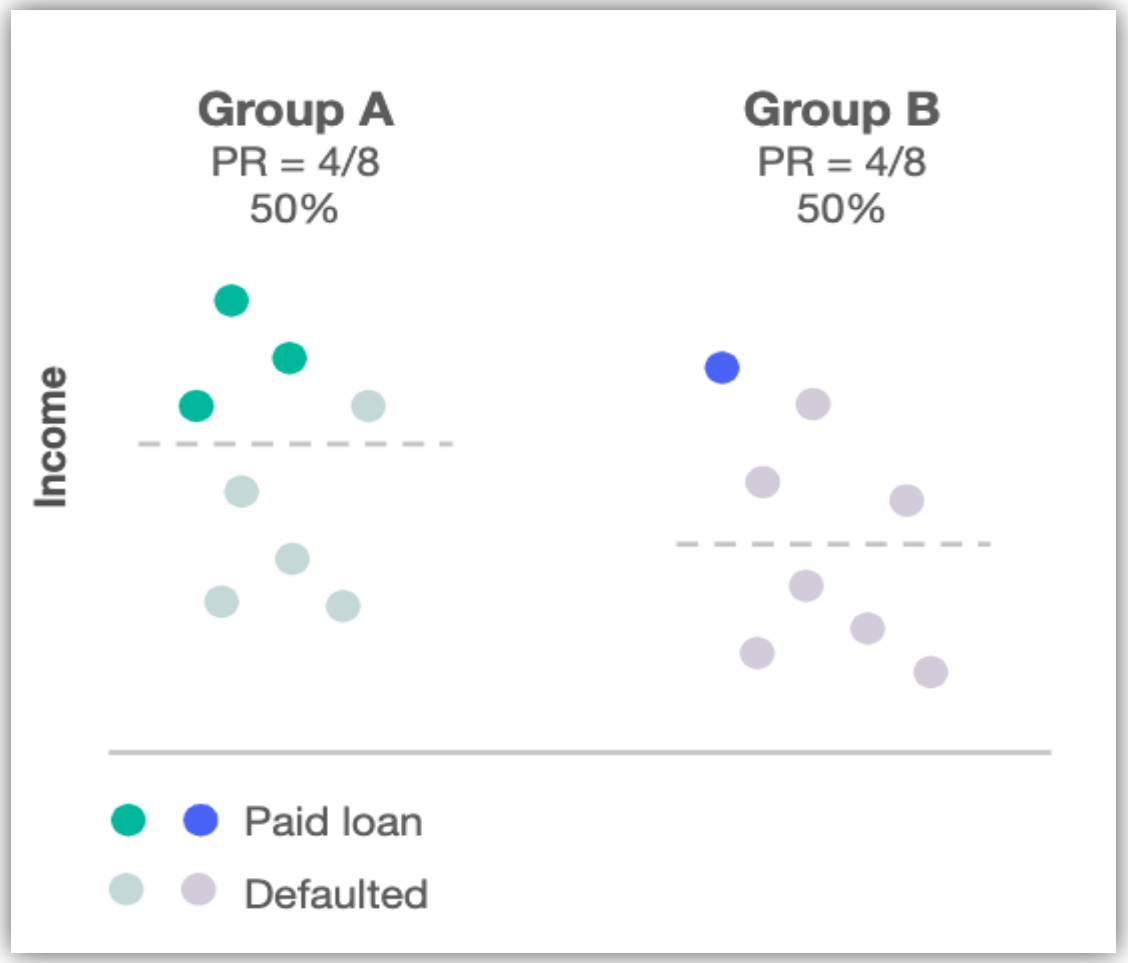


Image Credit: Valeria Cortez

Positive outcome is the preferred decision – e.g. getting selected for interview, getting a credit card, getting student loan etc.,

Demographic Parity requires that a decision of accepting or denying a loan application be independent of the protected attribute.

Demographic Parity: The proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates.

Membership in a **protected class should have no correlation with the decision**.

WHAT COULD GO WRONG WITH DEMOGRAPHIC PARITY

Demographic Parity (Mitigating Bias in ML)

- Imposing demographic parity can lead to two problems.
  - Introduces some unfairness - to achieve demographic parity, the algorithm needs to put Group B applicants into the "Acceptable" bin while, at the same time, prevent some of the Group A applicants from going in there.
  - Trade-off between accuracy and fairness - If the credit decisioning model was what is called a perfect classifier (in practice a perfect classifier does not exist but for the sake of the argument let's consider it does), it would not make any mistakes and place all applicants in the correct side of decision boundary. As such, this algorithm is fair as it treats every applicants the way they "deserves".
- Enforcing demographic parity on this perfect classifier would actually detune it –there could be a misalignment between optimizing a classifier and satisfying demographic parity.
- Therefore, demographic parity usually leads to larger costs in accuracy and, therefore, costs an organization more money than other fairness measures.

CONCLUSION

Demographic Parity (Mitigating Bias in ML)

- Demographic parity is both too strong and too weak.
- If you find a different condition that's strictly stronger, it'll still be flawed.
- If you come up with something strictly weaker, it'll still be flawed.
- **Demographic parity**, asks that the prediction must be uncorrelated with the sensitive attribute. This might sound intuitively desirable, but the outcome itself is often correlated with the sensitive attribute.
- You also won't rescue or salvage demographic parity by developing more sophisticated ML model such as ANN - backpropping through a 1200 layer neural net or Extreme Gradient Boosting algorithms
- Government Policies/University admissions/Public Distribution System tend to work towards achieving Demographic Parity.

EQUAL OPPORTUNITY

Equal Opportunity (Mitigating Bias in ML)

- The core **problem** of **demographic parity** is that it **does not care** whether or not applicants are "Acceptable", it just requires the fractions of group A and group B applicants in the global population being represented in the "Acceptable" bin.
- At its most basic, **Equality of Opportunity** requires that all human beings are equal in the sphere of opportunity.
- In order for opportunities to be equal within a group, each member of that group must face the same relevant obstacles, none insurmountable, with respect to achieving the same desirable goal.
- Individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome.
- The rate of 'low risk' predictions among people who actually pay back their loan should not depend on a sensitive attribute like race or gender. This is the principle *equality of opportunity* in machine learning.
- Applicants who pay back their loan, have an equal opportunity of getting the loan in the first place.
- Applicants that will end up defaulting, do *not* require equal opportunity of getting a loan in the first place.

https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-three-framework/fairness-criteria/

EQUAL OPPORTUNITY

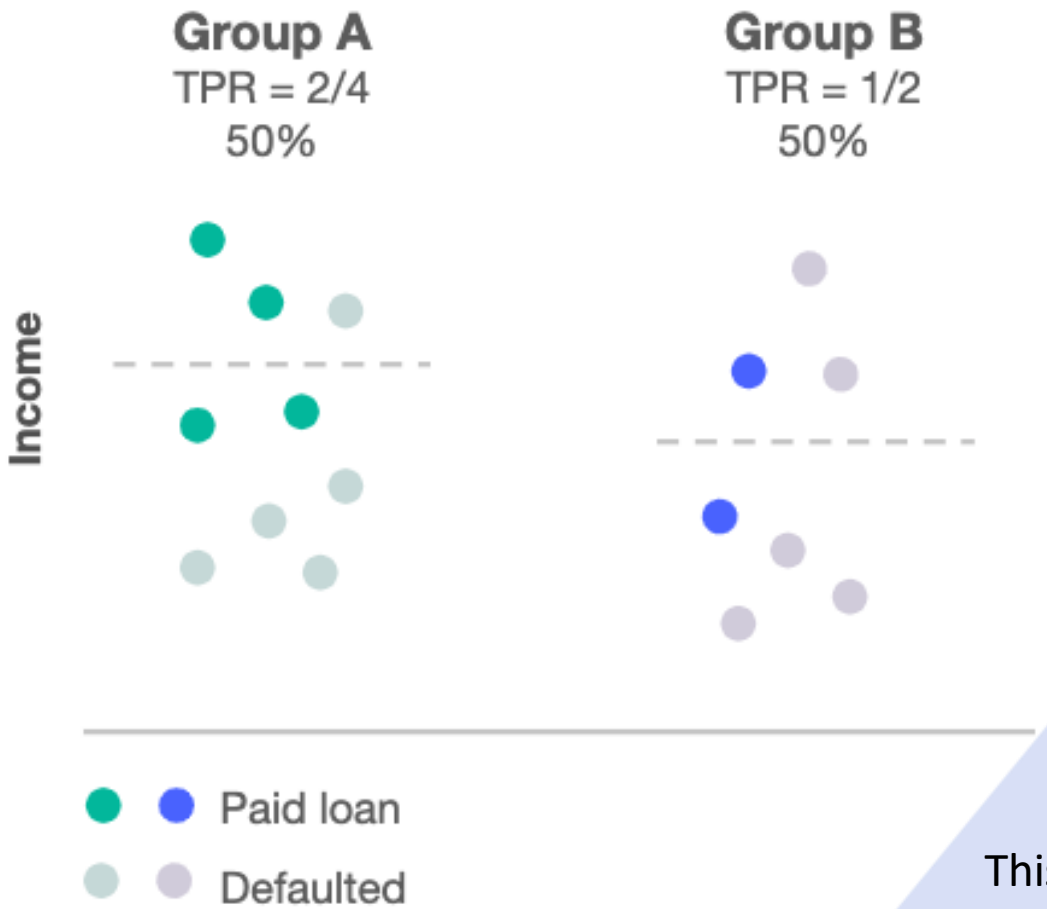


Image Credit: Valeria Cortez

Equal Opportunity requires the positive outcome to be independent of the protected class A, conditional on Y being an actual positive

Based on the confusion matrix, the True Positive Rate (TPR) to be the same for each segment of the protected class.

To achieve equal opportunity, ML should pick per-group thresholds such that the fraction of *non-defaulting* group members that qualify for loans is the same across all groups.

In figure, the percentage of positives that were accurately predicted is 50% for both groups.

This is like saying: *"You're allowed to be part of our New York City Football club, assuming you can play Football".*

WHAT COULD GO WRONG WITH EQUAL OPPORTUNITY (1/2)

Equal Opportunity

Mitigating Bias in ML

- Assume group B applicants have higher unemployment rate or lower score (FICO).
- The fraction of group A applicants in the "Acceptable" bin should be larger than 75% as more of the group B applicants need to be rejected. Thus, Equal Opportunity forces equality even when the data set is fundamentally unequal.
- Or True Positive Rate (TRP) may be the same for both groups as required by Equal Opportunity. Group B, might have higher False Positives.
- False Positives in this case are those who get a loan when in fact they're likely to default.

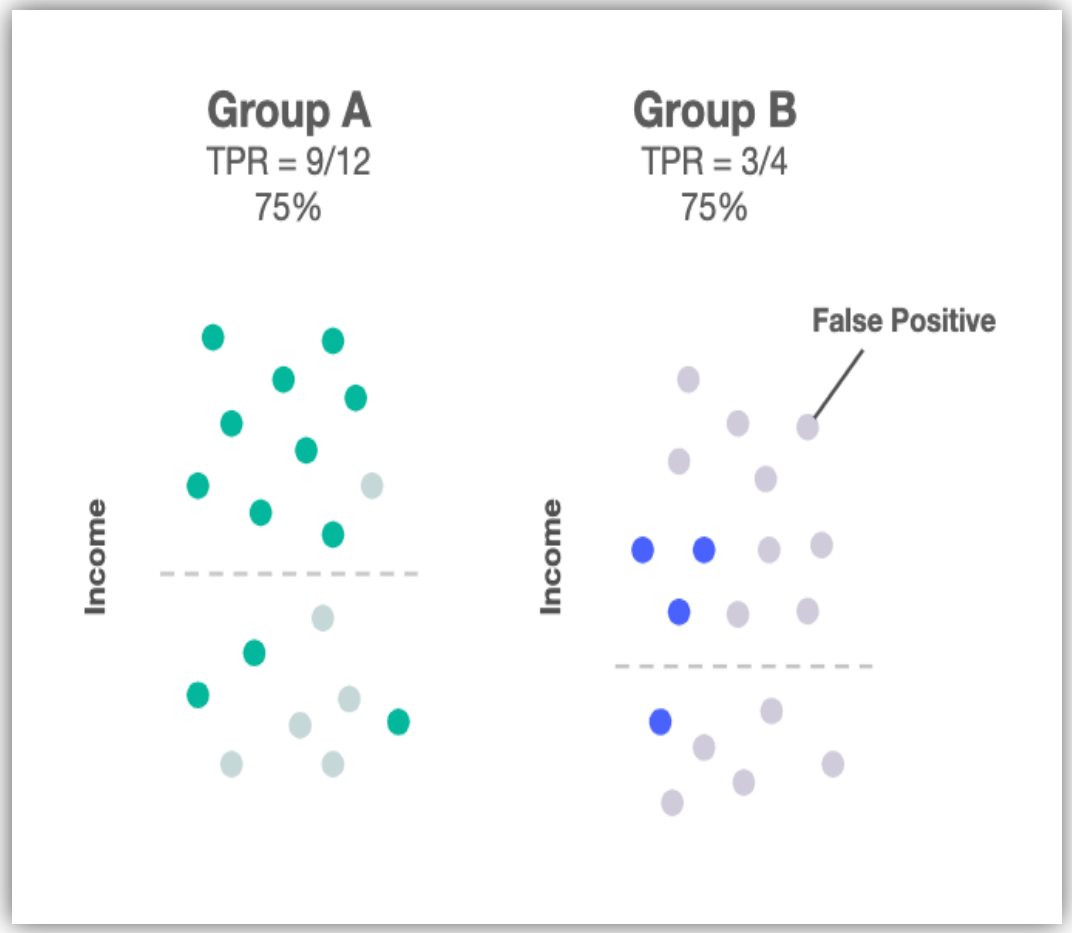


Image Credit: Valeria Cortez

WHAT COULD GO WRONG WITH EQUAL OPPORTUNITY (2/2)

Equal Opportunity
Mitigating Bias in ML

- We would be damaging the credit score at a larger scale in Group B, which results in disparate impact.
- ML classifier satisfies the *equal opportunity but does not satisfy the group fairness definitions.*
- There must be an adequate data set enabling the true positive rates to be estimated and subsequently equalized.
- Enforcing a fairness criterion does come at the cost of accuracy, which is to be expected as a fairness criterion is an additional constraint

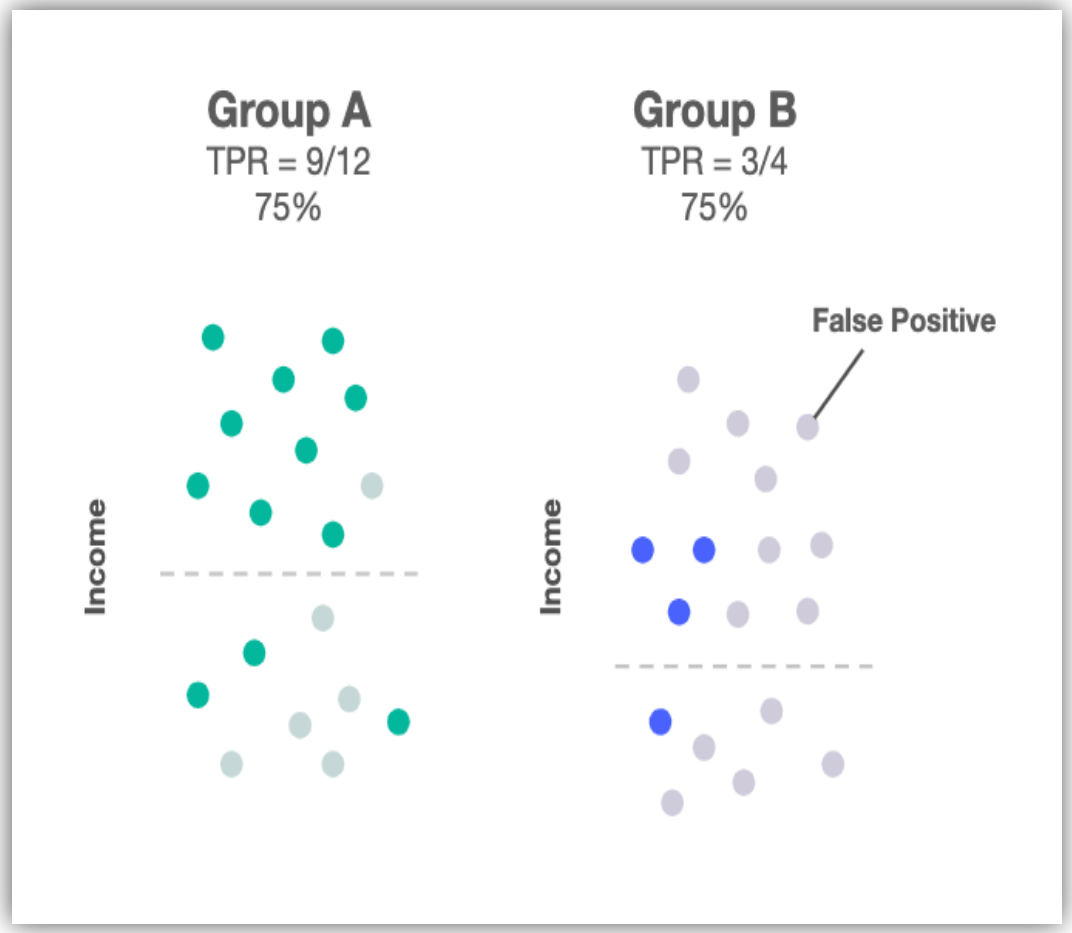


Image Credit: Valeria Cortez

CONCLUSION

Equal Opportunity (Mitigating Bias in ML)

- Equalized opportunity means matching the true positive rates for different values of the protected attribute.
- This is a less interventionist approach of equalizing the odds and may be more achievable.
- Equal Opportunity may be a suitable case for fraud detection/laundered transactions, given that mistakes in form of 'False Positives' do not generate a costly negative impact on customers or the company handling the situation.
- Also, labeling a transaction as suspicious/fraudulent should be a clear decision, not subject to major bias.

EQUALIZED ODD

Equalized ODD (Mitigating Bias in ML)

- Previous fairness measures does take into account the ground truth by, for example, balancing or equalizing the errors the ML algorithm makes for both sorts of groups.
- For example - **Equality of false-negatives** measure that enforces constant false-negative rates across groups. In our fresh applicants – irrespective of their group – have the same probability of falsely ending up in the "rejected" bin.
- Equalized odds (aka predictive value parity) is an approach to fairness in ML that is similar to equalized opportunity but places an additional constraint on the algorithm.
- In equalized odds, both the true positive rates and the false positive rates are equalized between the protected groups.
- Equalized odds most often drives the ML algorithm to sacrifice accuracy in order to satisfy additional criteria of fairness.

https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-three-framework/fairness-criteria/

EQUALIZED ODD

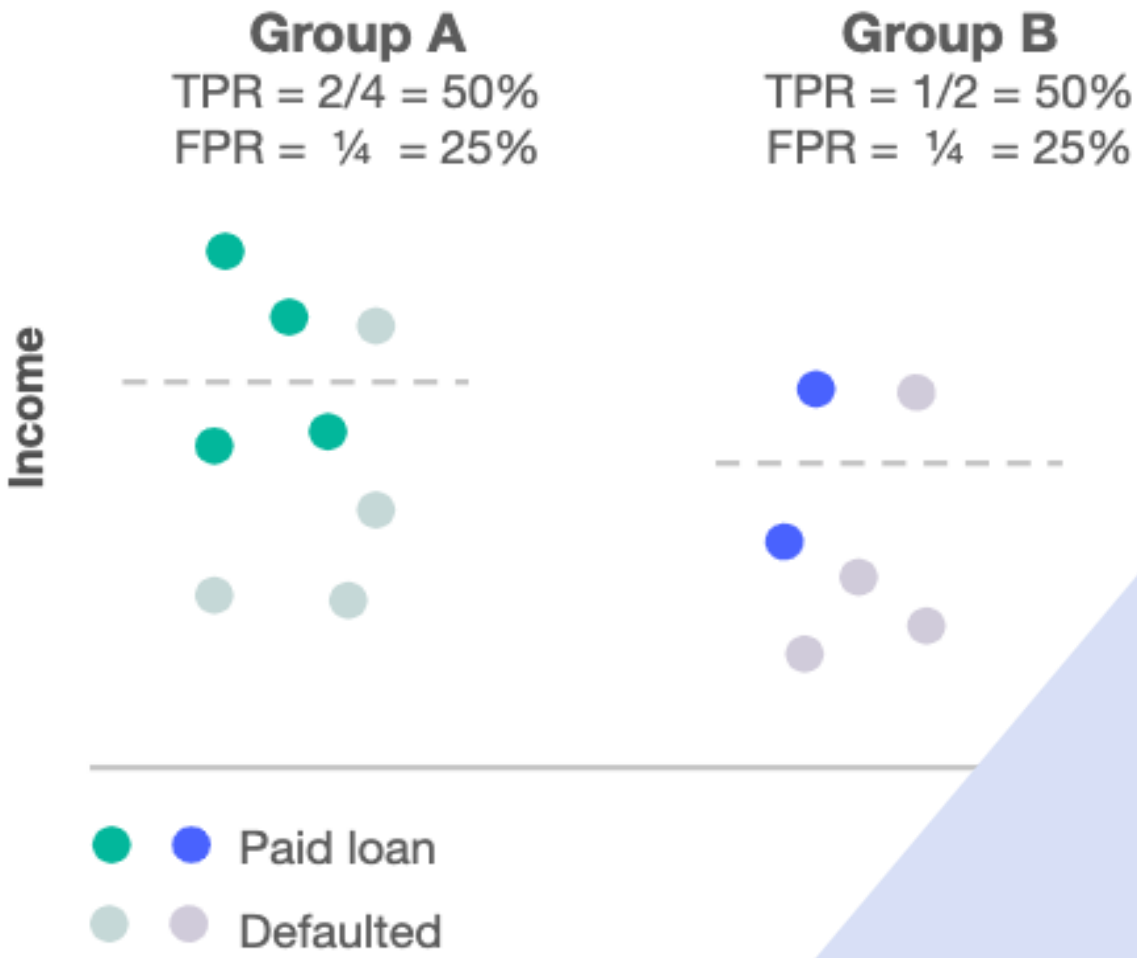


Image Credit: Valeria Cortez

Equalized Odds is the most restrictive concept

It correctly identify the positive outcome at equal rates across groups *(same as in Equal Opportunity).*

Miss-classify the positive outcome at equal rates across groups *(creating the same proportion of False Positives across groups)*

Identifying the positive outcome at a rate of 50% for both groups, but also having the same proportion of False Positives at 25% for both groups

WHAT COULD GO WRONG WITH EQUALIZED ODD

Equalized Odd

Mitigating Bias in ML

- It is the most restrictive of the definitions, trying to achieve equal TPR and FPR for each group may lead to a drop in accuracy/profits.
- This is because model performance could be compromised by not being able to optimize accuracy on the majority group.
- A credit decisioning model being used to identify who should get a loan and analyzed definitions of fairness for people over 35 and under 35.
- Accuracy/profits may be penalized during Equalized Odds compared to other definitions of fairness.

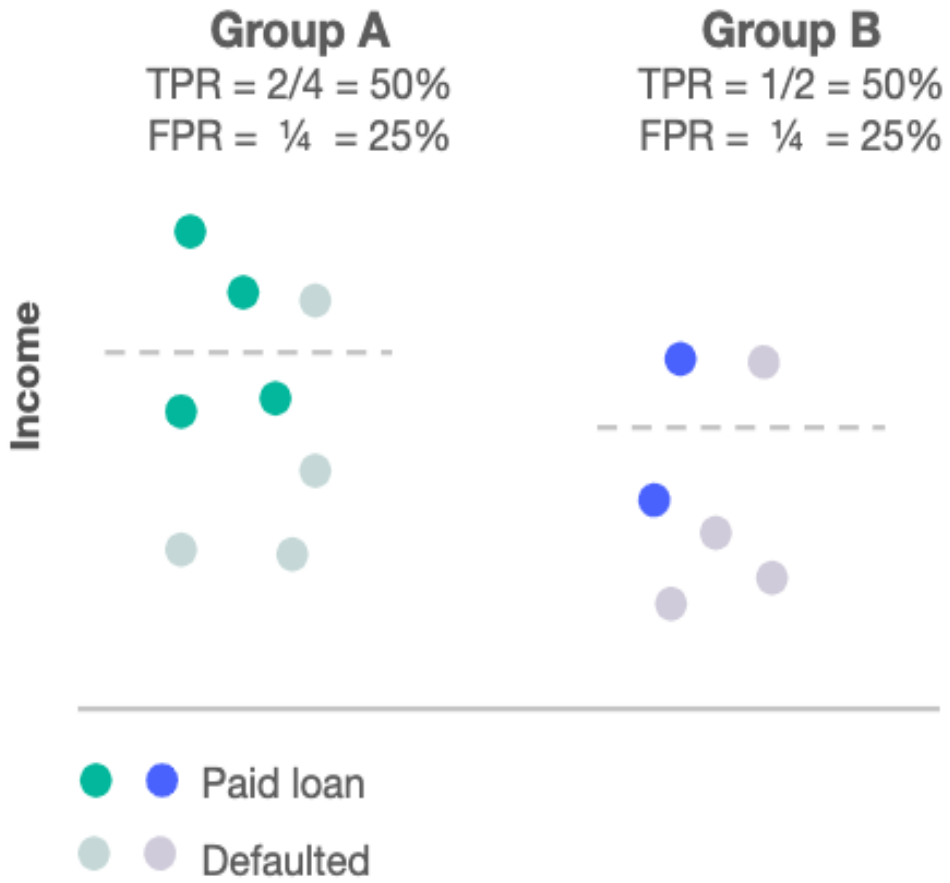


Image Credit: Valeria Cortez

CONCLUSION

Equal ODD (Mitigating Bias in ML)

- Based on the confusion matrix, Equalized odd requires the True Positive Rate (TPR) and False Positive Rate (FPR) to be the same for each segment of the protected class.
- This cannot be achieved with a single threshold for each group.
- There are many ways to do it - pick *two* thresholds for each group, so above both thresholds people always qualify, and between the thresholds people qualify with some probability.
- Equalized odds enforces that the accuracy is equally high in all groups, punishing models that perform well only on the majority.
- Credit models are an excellent example to use Equalised Odds as a definition of fairness.

ADVANTAGES AND DISADVANTAGES

Criterion	Advantages	Disadvantages
Demographic parity	• Conceptually simple • Can have legal standing (disparate treatment)	• Doesn't address individual-level fairness • May unacceptably compromise accuracy
Equalized Opportunity	• Appears to a reasonable interpretation of fairness • Can be good option if the TPR is most consequential factor	• Disparate FNR may remain between two population • Requires lots of labeled historical data
Equalized Odds	• Appeals to be a reasonable interpretation of fairness	• Can be inconsistent with the levels of accuracy and profits

Source: MiT Dlab Jan 2020
https://d-lab.mit.edu/sites/default/files/inline-files/Exploring_fairness_in_machine_learning_for_international_development_2802 2020_pages.pdf

WHEN TO USE WHAT

When to use Demographic Parity?

- We want to **change the state of our current world to improve it** *(e.g.: we want to see more minority groups getting to the top)*
- We are aware of **historical biases may have affected the quality of our data** *(e.g.: ML solution trained to hire software engineers, where nearly no women was hired before)*
- We have a **plan in place to support the unprivileged group** and **to prevent the reinforcement of historical biases** *(e.g.: setting policies that penalize non inclusive behavior at a board)*

When to use Equal Opportunity?

- There is a strong emphasis on **predicting the positive outcome correctly** *(e.g.: we need to be very good at detecting a fraudulent transaction)*
- Introducing **False Positives are not costly to the user nor the company** *(e.g.: wrongly notifying a customer about fraudulent activity will not be necessarily expensive to the customer nor the bank sending the alert)*
- The target variable is **not considered subjective** *(e.g.: labelling who is a 'good' employee is prompt to bias and hence very subjective)*

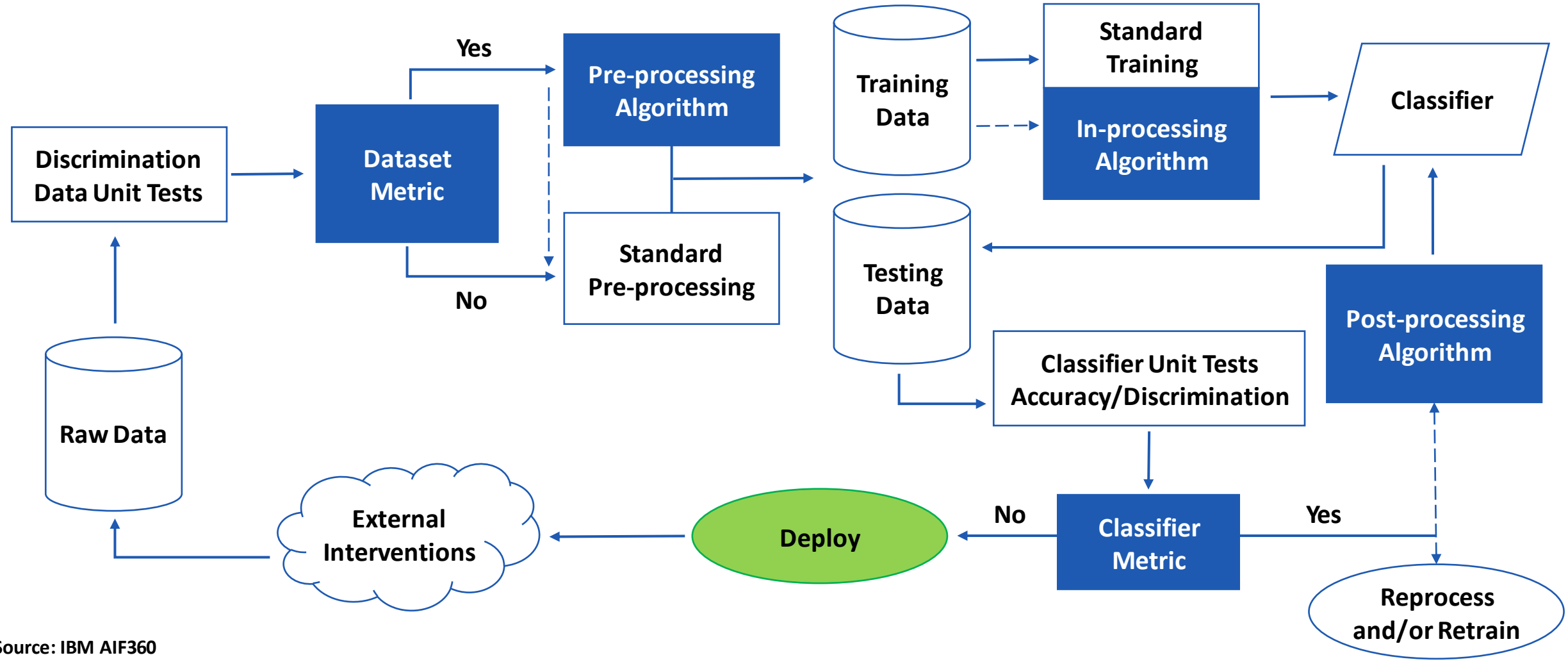
When to use Equalized Odds?

- There is a strong emphasis on **predicting the positive outcome correctly** *(e.g.: correctly identifying who should get a loan drives profits), and*
- We strongly care about **minimizing costly False Positives** *(e.g.: reducing the grant of loans to people who would not be able to pay back)*
- The target variable is **not considered subjective** *(e.g.: labeling who is a 'good' employee is prompt to bias and hence very subjective)*
- The **reward function of the model is not heavily compromised** *(e.g.: revenue or profit function for the business remains high)*

Coffee Break

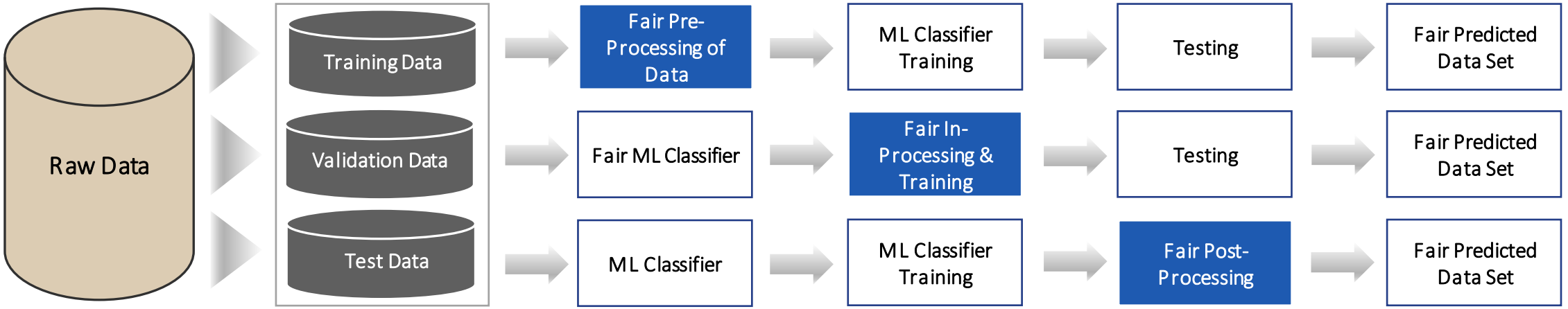
Lets Meet in 15 Minutes

BIAS WORKFLOW IN MACHINE LEARNING PROCESS



Source: IBM AIF360

BIAS IN MACHINE LEARNING PROCESS



Fair Pre-processing

- If training data can be pre-processed using fair exploratory analysis and processing, then pre-processing can be used
- Pre-processing algorithms are applied to training data to mitigate bias
- Algorithm examples: Reweighing, Disparate Impact Remover, Optimized Pre-processing, and Learning Fair Representation

Fair In-processing

- If ML training algorithms can be modified, then in-processing can be used
- A bias mitigation algorithm applied to a ML model during its training
- Algorithm examples: Adversarial De-biasing, Variational Fair Autoencoders, Discrimination Aware Ensemble, Model Regularization

Fair Post-processing

- If training data or ML training algorithm can't be modified, then the developed ML model is treated in the post-processing
- Post-processing methods try to modify the model's predictions or decision boundary in order to ensure fairness
- Algorithm examples: Reject Option Classification, Calibrated Equalized Odds Post Processing, Equalized Odds Post-processing

MACHINE LEARNING PROCESS – BIAS MITIGATION ALGORITHMS

Source: IBM AIF360

Pre-Processing Algorithms
Mitigating Bias in Training Data

Reweighing
Modifies the weight of training examples

Disparate Impact Remover
Edits feature values to improve group fairness

Optimized Pre-processing
Modifies training data features and labels

Learning Fair Representations
Learns fair representations by obfuscating information about protected attributes

In-Processing Algorithms
Mitigating Bias in Classifiers

Adversarial Debiasing
Uses adversarial techniques to maximize accuracy & reduce evidence of protected attributes in predictions

Prejudice Remover
Adds a discrimination-aware regularization term to the learning objective

Meta Fair Classifier
Takes the fairness metric as part of the input & returns a classifier optimized for the metric

Post-Processing Algorithms
Mitigating Bias in Predictions

Calibrated Equalized Odds Post-Processing
Optimizes over calibrated classifier score outputs that lead to fair output labels

Reject Option Classification
Changes predictions from a classifier to make them fairer

Equalized Odds Post-Processing
Modifies the predicted label using an optimization scheme to make predictions fairer

PREPROCESSING - REWEIGHT

ID	Weight	Name	Gender	Age	Label
e_1	1.0	John	M	20	1
e_2	1.0	Joe	M	20	0
e_3	1.0	Joseph	M	20	0
e_4	1.0	Sally	F	30	1
e_5	1.0	Sally	F	40	0
e_6	1.0	Sally	F	300	1

- An initial set of training examples with features for predicting whether a person will have high income.
- The data is not clean (e2 and e3 are duplicates), which may introduce bias that affects model fairness. In addition, e6 has an anomalous age.

- A simplified reweighing technique to increase the weights of positively labeled examples in sensitive groups whose ratio of weighted positive labels is lower than other groups.
- The sensitive groups Gender = M and Gender = F have ratios of 1/(1+1+1)=0.33 and (1+1)/1+1+1) = 0.67
- So we can increase the weight of e1 from to 4so that the ratio is: 4/(4+1+1)=0.67
- Suppose e2 and e3 are merged in e23 which would have total weight of 1 and label 0.
- Then the ratio of positive examples in gender = M increase to 1/(1+1)=0.5 so e1 weights needs to be changed from 1.0 to 2.
- 2/(2+1)=0.67

PREPROCESSING – REMOVING SENSITIVE VARIABLE

Redlining Effect

- Variables are usually correlated with each other:
  - Postal code with race
  - Working hours with gender
- Suppose **Salary** is decided (in decision maker's head) as : Salary = f(education, ethnicity)
- ML Modeling team assumes: Salary = f(education)
- Thus, Salary would reward for education and the model punishes ethnical minorities
- The development team need to assessed modification of the effect by redlining and the role of redlining

PRE-PROCESSING – DISPARATE IMPACT REMOVER

Disparate Impact Remover

- Disparate Impact is a metric to evaluate fairness.
- It compares the proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group.

$$\frac{Pr(Y=1|D=\text{unprivileged})}{Pr(Y=1|D=\text{privileged})}$$

- The industry standard is a four-fifths rule: if the unprivileged group receives a positive outcome less than 80% of their proportion of the privileged group, this is a disparate impact violation.
- One approach for mitigating bias is simply to remove the feature that should be protected. For example, you have gender available in your data set, remove it from the features passed to the machine learning algorithm.
- Disparate Impact Remover edits values, which will be used as features, to increase fairness between the groups.
- A feature can give a good indication as to which group a data point may belong to. Disparate Impact Remover aims to remove this ability to distinguish between group membership.

PREPROCESSING – DISPARATE IMPACT REMOVER

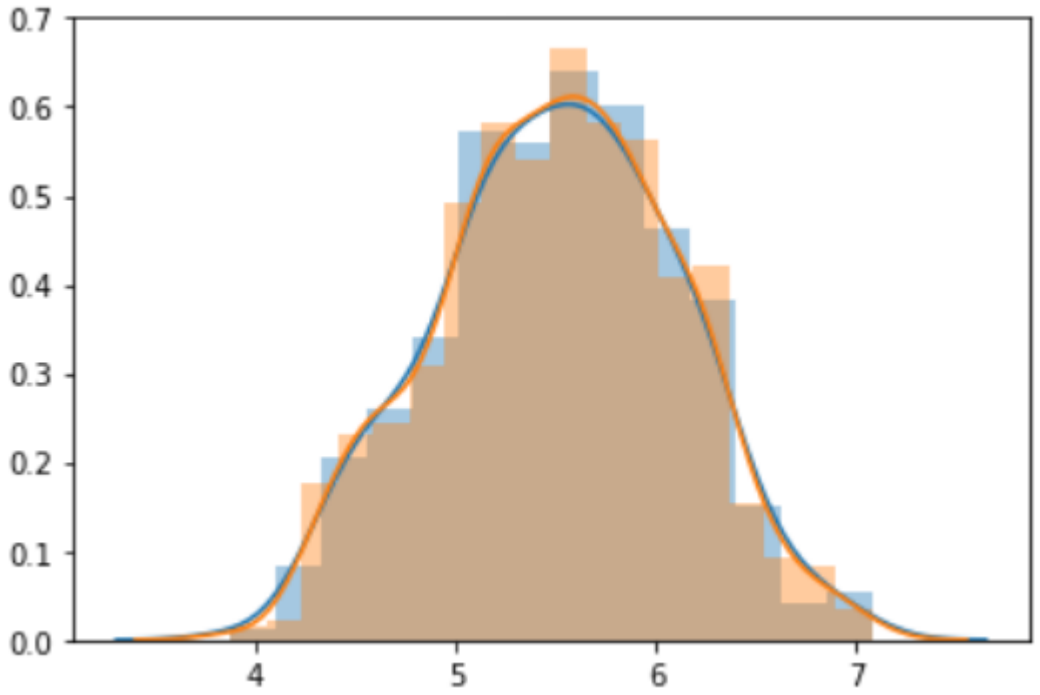
Disparate Impact Remover

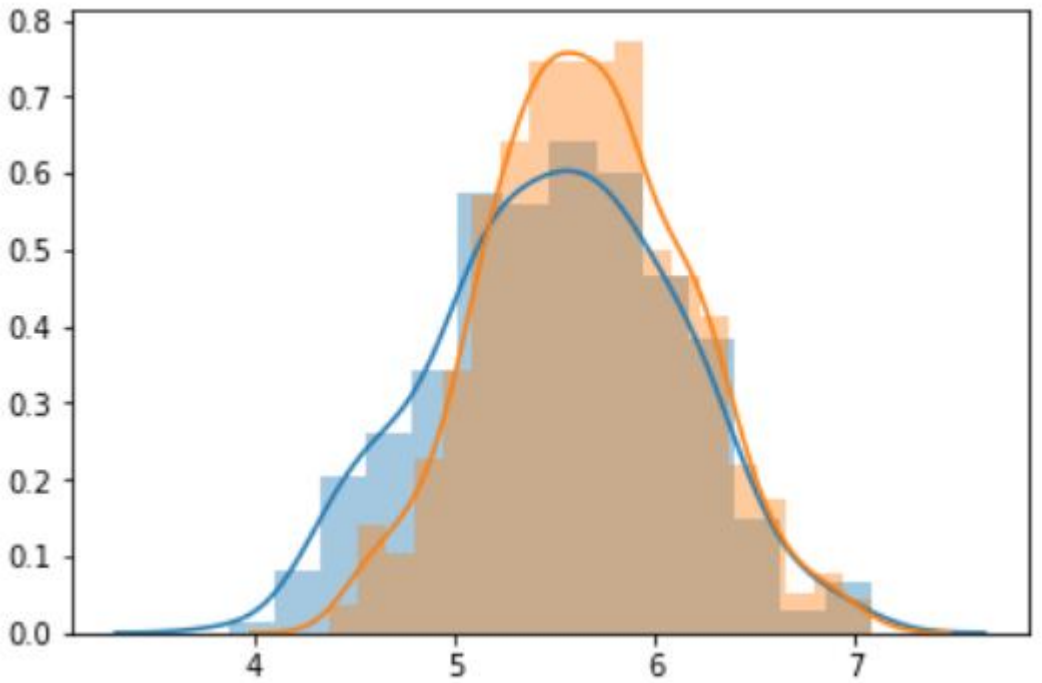
- The algorithm requires the user to specify a **REPAIR LEVEL.** This indicates how much you wish for the distributions of the groups to overlap.
- Repair level ranges between 0 and 1.
- 1 indicates complete overlap between the two groups.
- Developers are no longer able to select a point and infer which group it belongs to. This would ensure no group bias is discovered by a machine learning model.
- A value 0.80 indicates, the distributions of the two groups do not entirely overlap but developers would still struggle to distinguish between membership, making it more difficult for a model to do so.
- Disparate Impact Remover preserves rank-ordering within groups; if an individual has the highest score for group Blue, it will still have the highest score among Blues after repair.

Source: "Certifying and removing disparate impact" by M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, https://arxiv.org/abs/1412.3756

PREPROCESSING – DISPARATE IMPACT REMOVER

Disparate Impact Remover

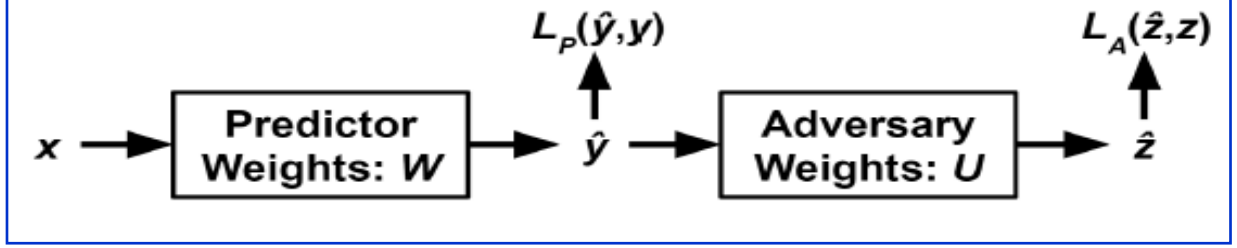




Source: "Certifying and removing disparate impact" by M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, https://arxiv.org/abs/1412.3756

IN-PROCESSING – ADVERSARIAL DEBIASING

Adversarial Debiasing



We have a base predictor model and an adversary model. The base predictor takes in input features x (including some protected variable z) and outputs a prediction.

The objective is to satisfy a given fairness definition; we add an adversary to the original predictor which penalizes the original predictor if its prediction is biased against the protected variable z.

The inputs for the adversary depend on the fairness metric. we are trying to optimize:

Demographic Parity Model: A takes as input the prediction and learns to predict protected variable z. If our prediction is biased against z, we will be able to predict z and Adversary will have high predictive accuracy.

Equality of Odds Model: A takes as input, the prediction and true label y and learns to predict protected variable z. If output of predictor and z are not conditionally independent given y, A will have high predictive accuracy.

IN-PROCESSING – PREJUDICE REMOVER REGULARISER

Prejudice Remover (Kamishima et al, 2012)

Prejudice: Statistical dependences of an objective variable or non-sensitive features on a sensitive feature

Direct Prejudice:
- Clearly unfair state where a ML model uses sensitive feature as inputs
- Imply the conditional dependence between Y and S given X
- To remove this type of prejudice, we have simply eliminate the sensitive variable from the prediction

Indirect Prejudice:
- Statistical dependence of an objective variable on a sensitive feature
- Even if a prediction model lacks a direct prejudice, the model can have an indirect prejudice and can make an unfair determination.
- To remove this indirect prejudice, we must use a prediction model that satisfies the condition of Y independent of S

Latent Prejudice:
- Statistical dependence of non-sensitive features on a sensitive feature
- This dependence doesn't cause a sensitive information to influence the final determination, but it would be exploited for training learners; thus, this might violate some regulations or laws.
- Removal of potential prejudice is achieved by making X and Y independent from S simultaneously

IN-PROCESSING – PREJUDICE REMOVER REGULARISER

Prejudice Remover (Kamishima et al, 2012)

Quantifying Prejudice:

- To quantify the degree of indirect prejudice, which is straightforwardly defined as the mutual information between Y and S.
- This index is referred as a (indirect) prejudice index (PI for short).
- Similar to a PI, the degree of a latent prejudice can be quantified by the mutual information between X and S.

Negative Legacy:

- It is unfair sampling or labeling in the training data.
- If a bank has been rejecting/refusing credit to certain groups who should have been approved/without assessing them, labels and the records of that group are less sampled in a training data set.
- A sample selection bias is caused by such biased sampling depending on the features of samples.
- It is known that the problem of a sample selection bias can be avoided by adopting specific types of classification algorithms

General Framework

- Y , X, and S are random variables corresponding to a class, non-sensitive features, and a sensitive feature, respectively
- A training data set consists of the instances of these random variables, i.e., D = {(y, x, s)}
- The conditional probability of a class given non-sensitive and sensitive features is modeled by $M[Y|X, S; \Theta]$, where Θ is the set of model parameters
- These parameters are estimated based on the maximum likelihood principle

IN-PROCESSING – PREJUDICE REMOVER REGULARISER

Prejudice Remover (Kamishima et al, 2012)

General Framework

- Define a prejudice index as the mutual information between the protected attribute and the prediction.
- Formulate a regularization loss based on the expectation of the prejudice index over the data set.

$$\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) = \sum_{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}} \ln \mathcal{M}[y_i|\mathbf{x}_i,s_i;\boldsymbol{\Theta}]$$

Regulariser

- Two types of regularizers to be used. The first regularizer is L2 regulariser is a standard one to avoid over-fitting.
- The second regularizer, R(D, Θ), is introduced to enforce fair classification.

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model Fairness Regularizer L2 Regularizer

where λ and η are positive regularization parameters.

POST-PROCESSING – REJECT OPTION CLASSIFICATION

ROC (Chow, 1970)

General Framework

- In this approach, the assumption is that most discrimination occurs when a model is least certain of the prediction i.e. around the decision boundary (classification threshold).
- Exploit the low confidence region of a classifier for discrimination reduction and reject its predictions.
- We can reduce the bias in model predictions.

Prediction Score:

- For an ML model with a classification threshold of 0.5, if the model prediction is 0.81 or 0.1, we would consider the model certain of its prediction
- but for 0.51 or 0.49 (for example), the model is not certain about the chosen category.
- For model predictions with the highest uncertainty around the decision boundary, when the favorable outcome is given to the privi leged group or the unfavorable outcome is given to the unprivileged, we modify them.

POST-PROCESSING – REJECT OPTION CLASSIFICATION

ROC (Chow 1970)

Advantages

- One can directly intervene at the last stage of the modeling workflow.
- Valuable for situations where at the prediction time (or in the deployment environment), the protected or sensitive attributes are available.
- ROC based Post-processing techniques provide the option to mitigate without modifying the learning stage and so are not restricted by any specific learning algorithm.
- Additionally, this approach is applicable to different fairness definitions as well.

POST-PROCESSING – EQUALIZED ODDS POST-PROCESSING

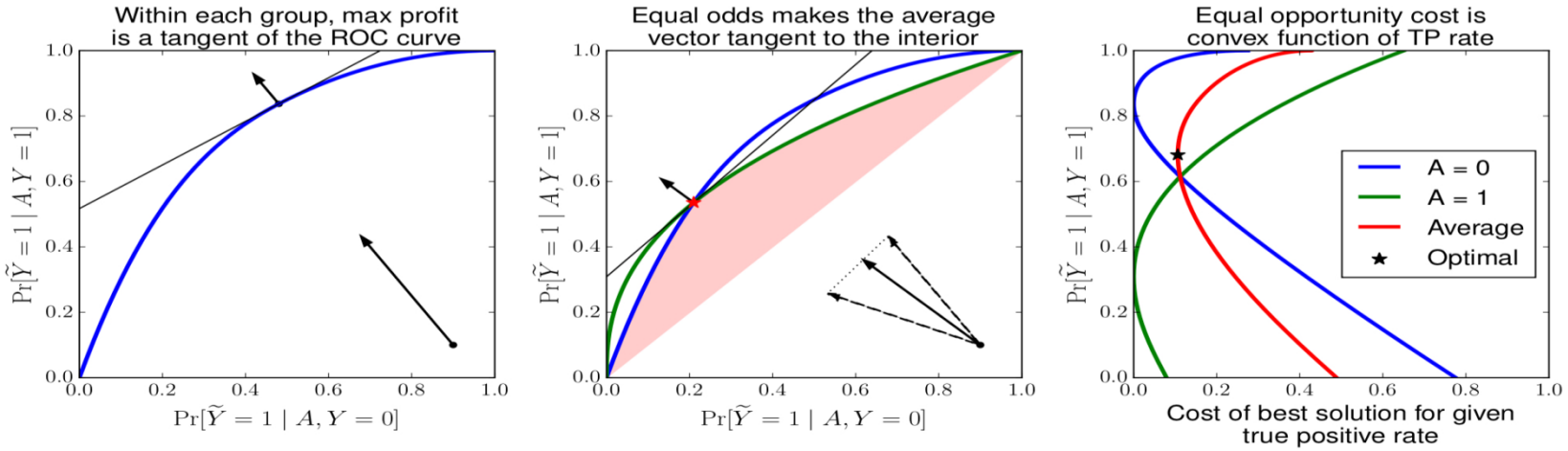
ROC (Hard et al. 2016)

General Framework

- In this approach, draw the corresponding ROC curves for both groups.
- Next, find threshold based on the ROC curves
- Equalized odds(both true positive rates and false positive rates are equal) is only satisfied when the ROC curves of the two groups intersect
- Equality of opportunity, as a weaker notion, can be satisfied by taking threshold such that the true positive rates of the two groups are equal
- To satisfy equalized odds, these must be at the same point in the false/true positive plane
- When the ROC curves do not agree, we might choose different thresholds for the different protected groups.
- So ML model can learn a possibly discriminator learned binary predictor (or score R), and then *derive* an equalized odds or equal opportunity predictor from it.
- An existing training pipeline untouched, and add an anti-discriminatory step on the back-end of it.

POST-PROCESSING – EQUALIZED ODDS POST-PROCESSING

ROC (Hard et al. 2016)



- The feasible set of false/true positive rates of possible equalized odds predictors is thus the intersection of the areas under the A-conditional ROC curves, and above the main diagonal
- Since for any loss function the optimal false/true-positive rate will always be on the upper-left boundary of this feasible set, this is effectively the ROC curve of the equalized odds predictors.
- This ROC curve is the point-wise minimum of all A-conditional ROC curves.
- The performance of an equalized odds predictor is thus determined by the minimum performance among all protected groups

POST-PROCESSING – EQUALIZED ODDS POST-PROCESSING

ROC (Hard et al. 2016)

Advantages

Can be applied after any classifiers

Relatively good performance especially fairness measures

No need to modify classifier

Disadvantages

Require test-time access to the protected attribute

Lack the flexibility of picking any accuracy–fairness tradeoff

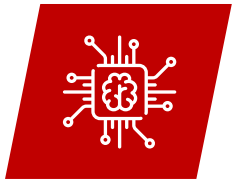
FAIRNESS TOOLS IN PYTHON

FairML

- Adebayo (M. S Thesis) developed FairML, an end-to-end toolbox for quantifying the relative significance of the feature dimensions of a given model.
- FairML is written in Python and available on GitHub

Themis ML

- Themis-ML is unrelated to the previously presented Themis testing tool
- Themis-ML giving access to a number fairness metrics and algorithms from literature.
- It is written in Python and available on GitHub.1



Aequitas

- The Center for Data Science and Public Policy of the University of Chicago published Aequitas
- It is usable as a Python library or as a standalone command line utility and available on GitHub

Themis

- Themis was developed by the Laboratory for Advanced Software Engineering Research at the University of Massachusetts Amherst
- It is a tool for testing software for discrimination.
- Themis is written in Python and available on GitHub

AIF360

- IBM Research released AI Fairness 360 (AIF360), an extensible Python toolkit comprising multiple fairness-aware algorithms known from literature
- It is available on GitHub and provides a website with interactive tutorials and a gentle introduction to its concepts and capabilities.

Fat Forensics

- For inspecting fairness, accountability, and transparency of all aspects of a machine learning system
- Started as an academic collaboration between the University of Bristol and the Thales Group, and is available on GitHub

MITIGATING BIAS IN MACHINE LEARNING – OUR APPROACH

Review of Data Completeness

- Mapping of risk and gap analysis
- Build harmonised standards for data labelling
- Develop standards to track the provenance, development and use of training data through their life cycle

Data review for Bias

- Map out risks
- Review, audit, data quality and data validation
- Use reliable statistical methods to ensure fairness in input data and ensure data representativeness
- Cleansing and preprocessing and EDA without discrimination

Sampling & Discriminatory Feature

- Sampling and data segmentation
- Training, validation and test sample creation
- Feature engineering based on application of ML model in particular domain (industry, geography, population, etc.)
- Algorithms to validate features

Model Development

- Use a reguliser to ensure fairness
- Wrap SHAP, LIME, ICE etc. around base ML models to ensure model interpretability and explainability
- Calibrate threshold to maintain fair outcomes for all groups
- Development of challenger model
- Use of FairML models

Model Testing

- Evaluating the algorithm performance by changing the training data, target variable, sampling method, hyper-parameters, cost functions, or other elements of the ML application)
- Cross validation sampling
- Error analysis, under and overfitting, evaluation of model stability and goodness-of-fit
- Benchmarking & alternate Fair ML models

Documentation

- Drafting details of development approach
- Include alternate models and reasons of rejections
- Documentation as per SR11-7 and OCC

Monitoring & Governance

- Evaluate existing governance policies and frameworks
- Redesigning monitoring under the lens of ML
- Develop standards to evaluate fairness, inclusion, and accountability
- Monitoring with new age metrics
- Thresholds, escalation & resolution

BEST PRACTICES TO CHECK FOR BIAS IN MACHINE LEARNING

Posing the following questions can help you check bias in ML Models:

- Are any particular groups suffering from systematic data error or ignorance?
- Have you intentionally or unintentionally ignored any group?
- Are all groups represented proportionally, e.g. when it comes to the protected feature of age, ethnicity, color, race, religion, gender, locality, profession, legal etc., are all protected features being identified or merely one or two?
- Do you have enough features to explain minority groups?
- Are you sure you aren't using or creating features that are tainted?
- Have you considered stereotyping features?
- Are you models apt for the underlined use case?
- Is your model accuracy similar for all groups?
- Are you sure that your predictions are not skewed towards certain groups?
- Are you optimizing all required metrics and not just those that suit the business?

BEST PRACTICES FOR MITIGATING BIAS IN MACHINE LEARNING

Area	Challenges		Best Practices	
Data Collection	• Biased data • Incomplete data	• The curse of dimensionality • Sparsity	• Take time to understand the business problem and its context • Enrich the data	• Dimension-reduction techniques • Change data representation
Outliers and Missing Values	• Out-of-range numeric values and unknown categorical values • Undue influence on squared loss	functions • Information loss • Bias	• Robust methods (e.g. Huber loss function) • Discretization (binning)	• Winsorizing • Weight of evidence • Leave-one-out event rate
Imbalance/Sparse Target Variable	• Low primary event occurrence rate	• Overwhelming preponderance of zero or missing values in target	• Over-sampling/Under sampling/SMOTE	• Inverse prior probabilities • Mixture models
High Cardinality Variables	• Overfitting	• Unknown categorical values in holdout data	• Discretization (binning) • Weight of evidence	• Leave-one-out event rate
Multicollinearity	• Unstable parameter estimates		• Regularization • Dimension reduction (PCA, ICA)	• Information Value/Mutual Entropy/VIF
Overfitting	• High-variance and low-bias models that fail to generalize		• Regularization • Noise injection	• Partitioning or K-fold-cross validation
Hyperparameter tuning	• Large number of hyper-parameters in conventional	algorithms (e.g. deep neural networks, Super Learners)	• Grid search, random search • Local search optimization,	including genetic algorithms/ simulated annealing/ swarm
Model Explainability/ Interpretability	• Large number of parameters, rules, or other complexity	obscures model explainability/interpretation	• Variable selection by regularization • Surrogate models • Partial dependency	plots/Variable importance measures • LIME/SHAP/ICE
Model Deployment	• Trained model must be migrated from a development ecosystem to an operational computing	ecosystem to assist in decision making processes	• Deployment strategies like Shadow Model and Canary • Platform as a Service (PaaS)	• Infrastructure as a Service (IaaS)

OVERCOMING ML BIAS KEY CHALLENGES

Gap Between Theoretical Process & Practice

- Many issues like complex system interactions and ethical questions often fall beyond the scope of those theoretical steps
- It has problematic implications for metrics, testing, and process improvements

Unforeseen Issues

- Many algorithmic fairness issues can't easily be detected before implementing
- Issues emerge only after the model is in production or over a period
- Leverage a couple of useful mechanisms in addressing fairness

No One Size Fit All Solution

- ML algorithmic fairness is a challenge for which there is unlikely to be a single right question or answer
- Checklists, automated testing and similar tools may be inadequate because of contextual variation

Actual Distribution of Problems May Be Different

- Likely a significant selection bias in the problems
- Small systemic biases are harder to detect
- Need to perform various rigorous testing

Managing Expertise is a Challenge

- Need to recruit, train, and retain experts
- Capture knowledge so it can be shared and ultimately systematized
- Integrate experts into the product life cycle and empower them to negotiate with development teams

Not just Explainability and Interpretability

- Explainability and interpretability have potential upside, however, they appear to be a means to an end rather than an end goal
- Focus should also be on contestability (the ability to challenge ML models) or recourse (the ability to change the decision via changes to input).

Thank you