Identifying and Mitigating Bias in Machine Learning Models

Jana Gunn

Sr. Principal Data Scientist

OPTUM Labs®

Objectives

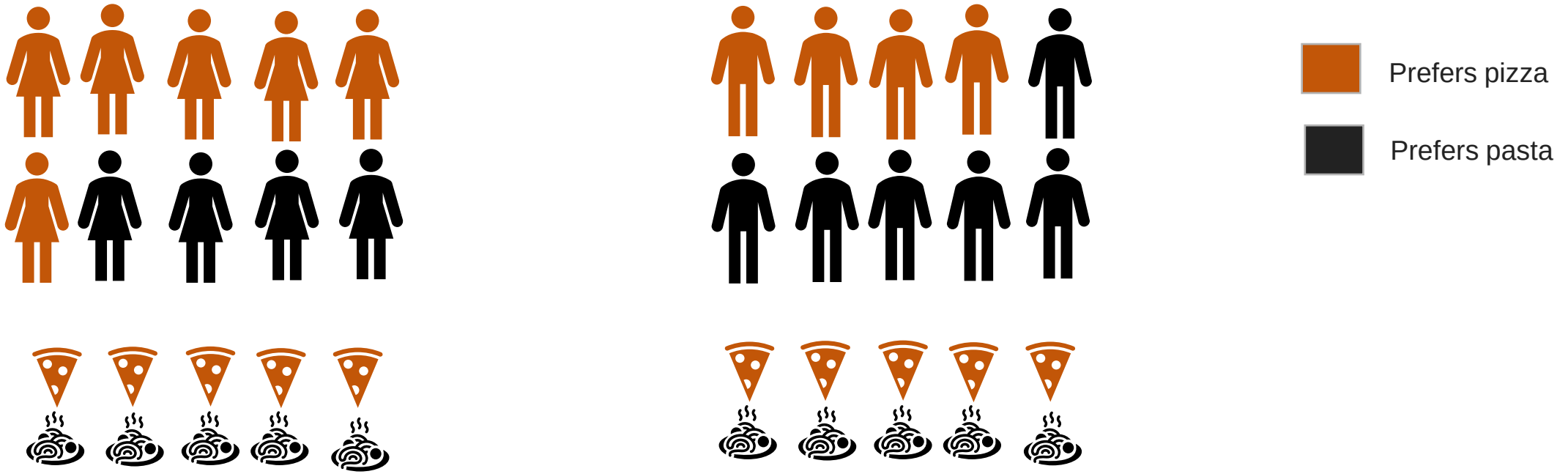
1) Identify and discuss different definitions of fairness
2) Understand disparate treatment in the healthcare system
3) Quantify harms found in machine learning models
4) Review strategies for mitigating bias in machine learning models

What is Fairness?

Important to understand that there is no single definition of fairness.

Parity or Preference?

- **Statistical Parity**: Representative division regardless of preference

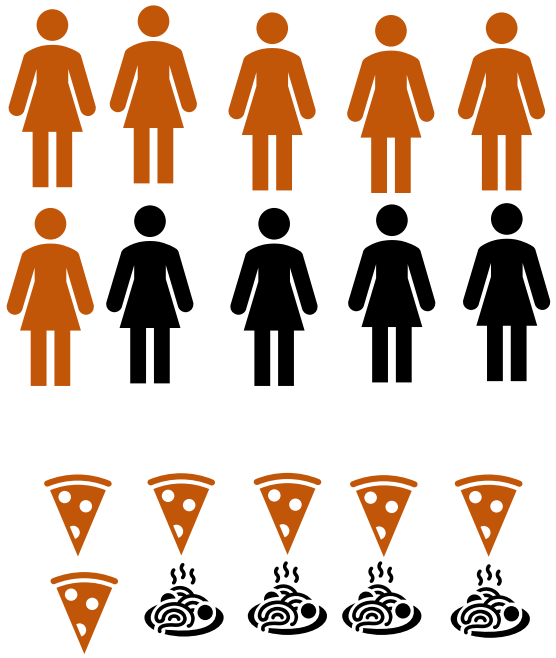


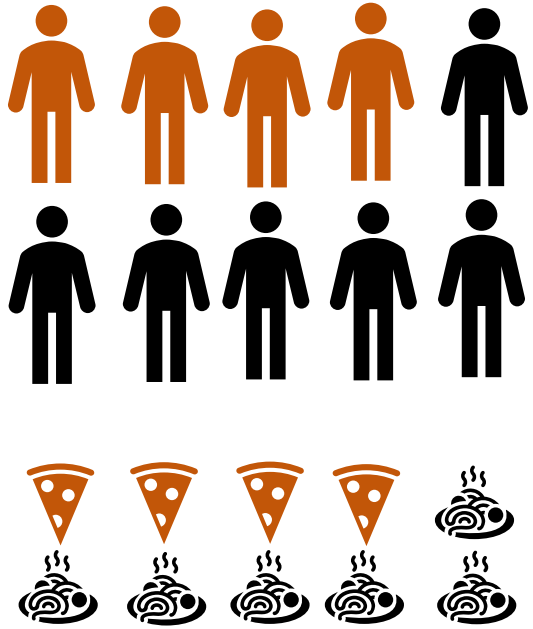
What is Fairness?

Important to understand that there is no single definition of fairness.

Parity or Preference?

- **Preference Preservation**: Strive to optimize for the preference of users





What is Fairness?

Avoid Disparate Treatment or Disparate Outcomes?

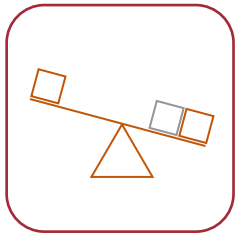
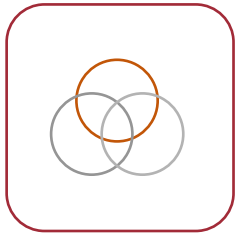
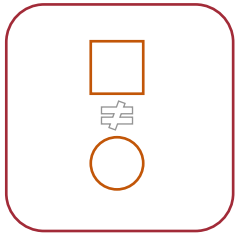
- Procedural Fairness: All individuals with similar characteristics should receive similar decisions from the system
  - Equal Opportunity
- Distributive Justice: All groups of individuals (as defined by specific identifiable characteristics such as race or sex) should receive similar outcomes from the system
  - Equal Outcome

OPTUM Labs®

Bias and Fairness

Any decision-making system can exhibit bias towards certain factors and thus, needs to be evaluated for fairness.

Sources of bias in machine learning

Data collection	Representation	Fit			Post processing
					
Historical	Representation	Measurement	Aggregation	Evaluation	Deployment
What may continue even with perfect sampling and feature selection	When population is under-represented and fails to generalize for sub-groups	When features make models blind and/or add factors that skew predictions	When subgroups are inappropriately combined	When there is unequal results representation or there are faulty performance metrics	When there is misuse or misinterpretation of predictive models
(E.g., racism)	(E.g., CEO, hiring predictions)	(E.g., spurious proxy)	(E.g., sex and race)	(E.g., A1C range variations by subclass; working history as indication of credit worthiness)	(E.g., use of diabetes risk score to determine donor eligibility; using a recidivism prediction model to determine length of sentence)

OPTUM Labs®

Approaches to Fairness

Fairness through Unawareness

- Protected attributes are not used in prediction process

  PROS:

  - Guaranteed not to make an explicit judgment based on a sensitive attribute

  CONS:

  - Sometimes the attribute is important to the decision-making process (e.g. different symptoms based on the sex of a patient)
  - Unfairness (such as racism) can and does happen in "color-blind settings", and unawareness can mask and hide this.

Approaches to Fairness

Individual Fairness

- Similar outputs for similar individuals

PROS:

- Consistent between individuals

CONS:

- "Similarity" can be difficult to define, especially when multiple overlapping metrics are involved
- There is no way to check whether group fairness is also satisfied under these definitions

Approaches to Fairness

Group Fairness

- Predicts a particular outcome for individuals across groups with similar probability (e.g. a healthcare algorithm that assigns 5% of white patients as eligible for dialysis also assigns 5% of Black patients as eligible for dialysis)

PROS:

- Satisfies notions of avoiding penalizing or harming a specific group
- Aligns with concerns about group equity (e.g. similar dialysis spending is granted to both Black and white patients)

CONS:

- No requirement to pick the "most qualified" within each group
- Can be less accurate and potentially inappropriate if base rates of a label differ

Approaches to Fairness

Equality of Opportunity

- Probability of an outcome is the same across different classes (e.g. if a man has a 40% chance of being hired for a job, so does a woman with similar experience)

PROS:

- True positive rate is the same for all groups

CONS:

- If base rates of the labels are different, there would be different false positive rates (e.g. if a higher proportion of women are qualified for the job, more unqualified men may be hired).

Approaches to Fairness

Counterfactual Fairness

- Decisions for a person who is a member of group X are the same as they would have been were that person a member of group Y (e.g. the algorithm makes the same decision for a Black woman as it would have if she were a white woman)

PROS:

- Aligns with an intuitive and aspirational sense of fairness without being colorblind

CONS:

- Different factors are interrelated, and the world is too complex to build models that truly estimate the counterfactual

- Defining the "similarly situated" member of the non-minority group can be difficult

-Intersectional identities further the complexity

OPTUM Labs®

Unfairness and Proxy Variables

It is entirely possible for an algorithm that has zero knowledge of the protected characteristic (e.g. sex) to be unfair and discriminatory.

- Proxy variables are closely related to sensitive feature
- Examples: hair length and gender, race and zip code

An algorithm trained on a dataset that embeds systemic bias will learn and perpetuate its patterns even if blind to the protected class.

- Using healthcare costs as a proxy for disease severity

OPTUM Labs®

Disparate Treatment in the Healthcare System

- ***Disparities in Treatment:***
  - African Americans, both children and adults are less likely to receive appropriate pain treatment when in acute distress, even when controlling for age, sex, and time of treatment.
  - Up to half of medical students in one study hold the physiologically false belief that Black and white patients experience pain differently

- ***Disparities in Diagnosis:***
  - Women presenting to the emergency department with coronary syndromes are less likely than men with the same symptoms to be admitted to acute care.
  - Black patients experiencing symptoms of major depression are more likely to receive a potentially inappropriate diagnosis of schizophrenia
  - Medical students often only learn to diagnose important and common dermatologic findings on light colored skin, with darker skin being treated as an afterthought.

- ***Structural Disparities:***
  - 54% of rural counties do not have a hospital with obstetrics services
  - Elderly LGBT individuals experience isolation, and a lack of access to culturally competent healthcare and social service providers

Mitigation Strategies

Mitigation Strategies

Within the model training stage, mitigation may occur at different steps relative to model training:

Preprocessing: A mitigation algorithm is applied to transform the input data to the training algorithm; for example, some strategies seek to remove the dependence between the input features and sensitive features.

At training time: The model is trained by an optimization algorithm that seeks to satisfy fairness constraints.

Postprocessing: The output of a trained model is transformed to mitigate fairness issues; for example, the predicted probability of readmission is thresholded according to a group-specific threshold.

Fairlearn Mitigation Algorithms

Algorithm	Description	Binary Classification	Regression	Supported Fairness Definitions	Sensitive Features
Threshold Optimizer	Postprocessing algorithm based on the paper *Equality of Opportunity in Supervised Learning*[3]. This technique takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the classifier's prediction to enforce the specified parity constraints.	✓	✗	DP, EO, TPRP, FPRP	Categorical
ExponentiatedGradient	A wrapper (reduction) approach to fair classification described in *A Reductions Approach to Fair Classification*[1].	✓	✓	DP, EO, TPRP, FPRP, ERP, BGL	Categorical
GridSearch	A wrapper (reduction) approach described in Section 3.4 of *A Reductions Approach to Fair Classification*[1]. For regression it acts as a grid-search variant of the algorithm described in Section 5 of *Fair Regression: Quantitative Definitions and Reduction-based Algorithms*[2].	✓	✓	DP, EO, TPRP, FPRP, ERP, BGL	Binary
CorrelationRemover	Preprocessing algorithm that removes correlation between sensitive features and non-sensitive features through linear transformations.	✓	✓	✗	Binary

DP refers to *demographic parity*, EO to *equalized odds*, TPRP to *true positive rate parity*, FPRP to *false positive rate parity*, ERP to *error rate parity*, and BGL to *bounded group loss*.

Use Case – Increasing downloads of a MSK mobile application

Objective: Develop a classification model that identifies which members are most likely to download a mobile application for musculoskeletal (MSK) therapy. Model will be used to send marketing materials to individuals that qualify, encouraging them to download the application.

Sample: Members that qualified for the mobile app between 1/1/2021 and 7/15/2021.

Data:

- Amerilink Consumer Database
- Medical and Rx Claims
- County Health Rankings
- Food Environment Atlas

Target Variable: Members that have downloaded the mobile application.

Fairness Considerations

Data

Determining what data to use to train the model is another critical component of the machine learning process. Several factors should be considered including:

- Sample sizes of different groups (i.e. race, ethnicities)
- Representativeness of population
- Appropriateness of choice of label
- Label imbalance
- Adequate features for prediction
- Privacy considerations

Types of fairness-related harms

Allocation harms: Opportunities are unintentionally withheld from certain members of society

Quality of service harms: System works well for some groups but not others

Representation harms: System produces results that exacerbate stereotypes or under-represent certain populations

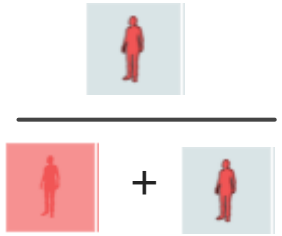
These are not mutually exclusive, meaning AI systems can exhibit more than one type of harm.

Quantify Harms

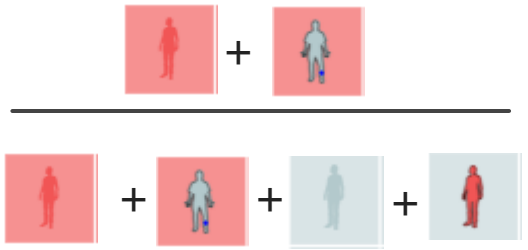
Members could experience allocation harm if they would benefit from the mobile app but are not notified about it. These individuals would be considered **false negatives** in our classification model.

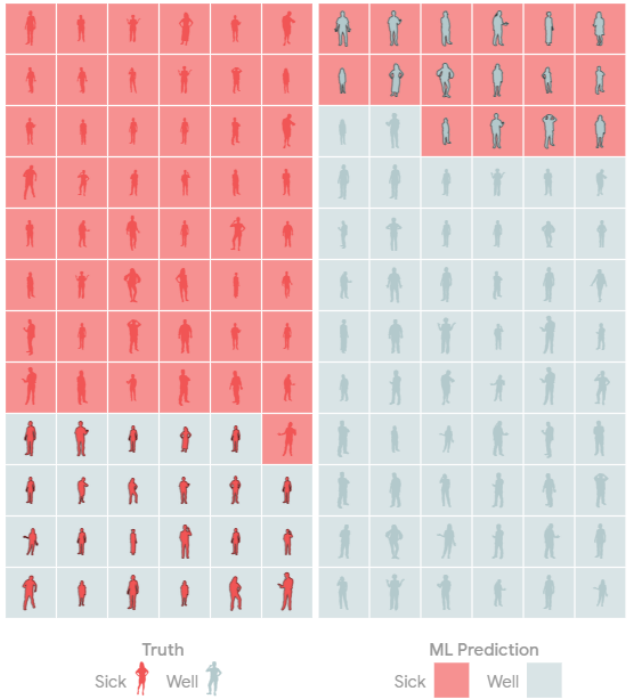
We will use two metrics to measure harm:

- **False Negative Rate**: fraction of members that would download the app, but are NOT marketed to
  - False Negative / (True Positive + False Negative)



- **Selection Rate**: overall fraction of members that are sent marketing materials (regardless of whether they would download it)
  - (True Positive + False Positive) / (True Positive + False Positive + True Negative + False Negative)

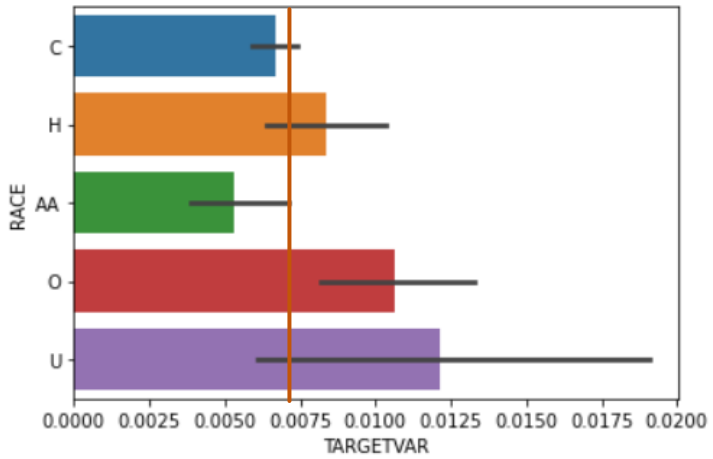




Train the model

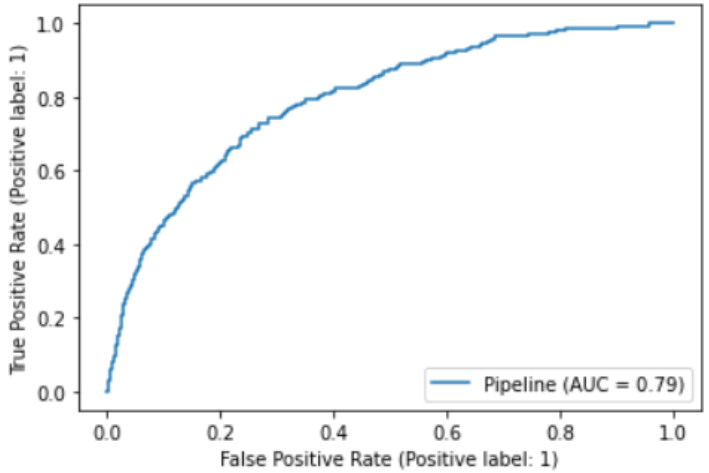
1) Split data into training and testing datasets
   - Balance training data – there should be an equal amount of positive and negative labels
   - Testing data will remain unbalanced
2) Run the logistic regression model
3) Examine performance metrics and coefficients

Average download rate: 0.7%



AUC: 0.79

Balanced Accuracy: 0.73



OPTUM Labs®

Calculate Fairness Metrics

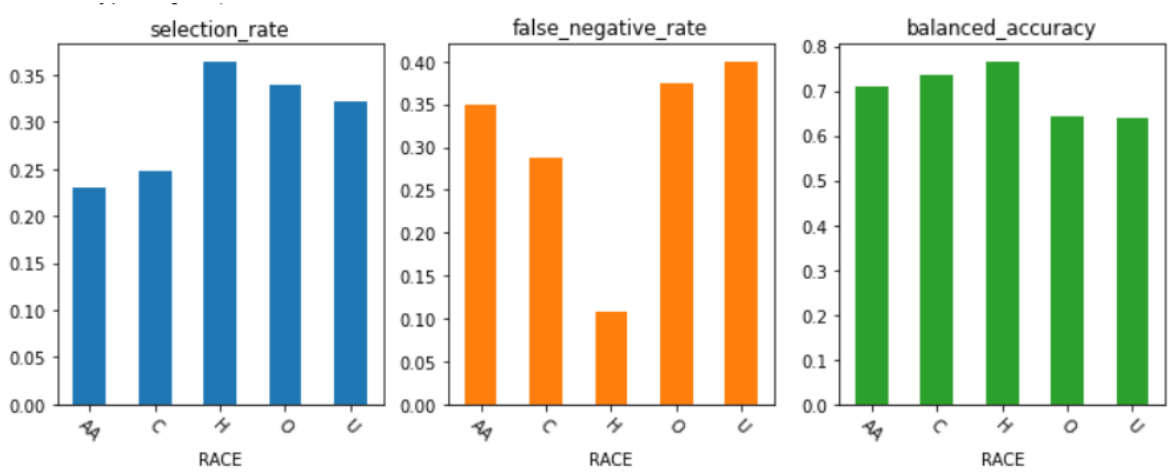
Identifying harms

Individuals with “Other” or “Unknown” race have the **higher false negative rates** and **lower accuracy**.

- A larger fraction of group members that are likely to benefit from the mobile App are not selected (false negative rate).
- Individuals are not correctly classified as often as other groups (accuracy).

RACE	selection_rate	false_negative_rate	balanced_accuracy
AA	0.229666	0.35	0.71116
C	0.247614	0.287356	0.734156
H	0.364608	0.107143	0.766116
O	0.338668	0.375	0.644678
U	0.321053	0.4	0.640708

	selection_rate	false_negative_rate	balanced_accuracy
difference	0.134942	0.292857	0.125408
ratio	0.629898	0.267857	0.836307
group_min	0.229666	0.107143	0.640708
group_max	0.364608	0.4	0.766116



OPTUM Labs®

Postprocessing with Fairlearn ThresholdOptimizer

ThresholdOptimizer is a **meta-algorithm** in that it acts as a wrapper around any standard (fairness-unaware) machine learning algorithm.

ThresholdOptimizer is a **postprocessing** technique.

- It takes an already trained model and a dataset as an input and seeks to fit a transformation function to the model's outputs to satisfy some (group) fairness constraint(s).

One limitation of ThresholdOptimizer is it needs access to sensitive features during training and at deployment time.

OPTUM Labs®

Use Case – ThresholdOptimizer Results

	Unmitigated			ThresholdOptimizer_Training			ThresholdOptimizer_Testing		
Race	Selection Rate	False Negative Rate	Balanced Accuracy	Selection Rate	False Negative Rate	Balanced Accuracy	Selection Rate	False Negative Rate	Balanced Accuracy
African American	23%	35%	71%	46%	16%	84%	32%	30%	69%
Caucasian	25%	29%	73%	45%	22%	82%	22%	34%	72%
Hispanic	36%	11%	77%	55%	17%	81%	31%	14%	77%
Other	34%	38%	64%	57%	21%	77%	31%	38%	66%
Unknown	32%	40%	64%	54%	22%	89%	26%	60%	57%
Difference	13%	29%	13%	13%	6%	11%	10%	46%	21%

Threshold Optimizer minimizes the difference in the training data, however, does not hold up in the testing data suggesting the need for a larger sample size.

OPTUM Labs®

Reductions approach with Fairlearn ExponentiatedGradient

ExponentiatedGradient is a **reductions approach** of Agarwal et. al (2018) to obtain a model that satisfies the fairness constraints, but does not need access to sensitive features at deployment time.

ExponentiatedGradient creates a sequence of reweighted datasets and retrains the wrapped model on each of them. The retraining process is guaranteed to find a model that satisfies the fairness constraints while optimizing the performance metric.

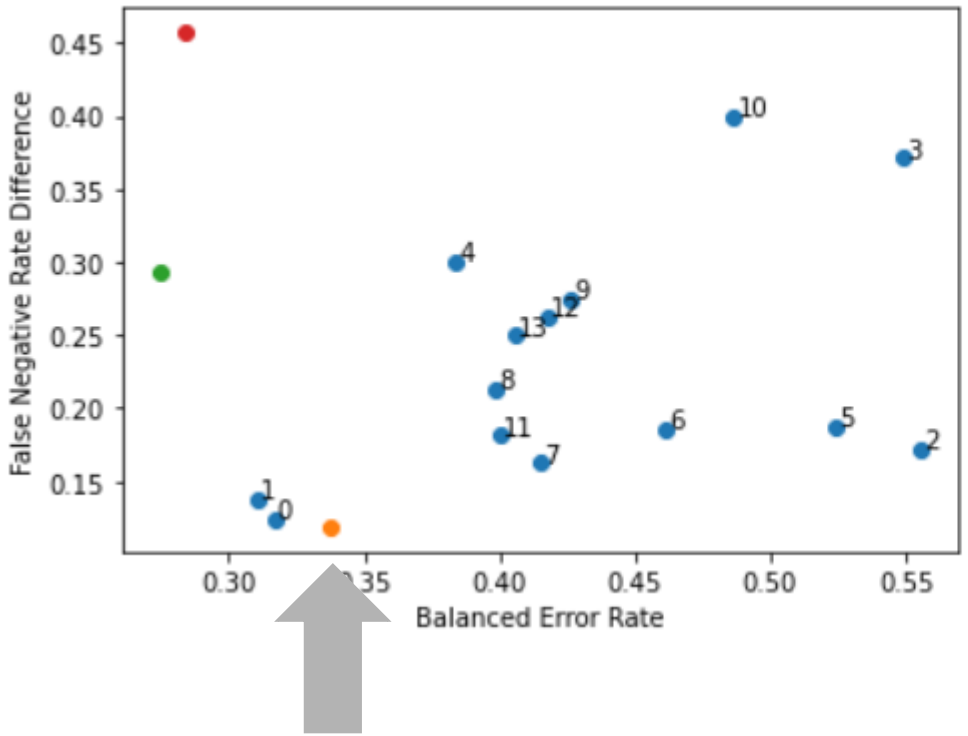
Difference

	selection_rate	false_negative_rate	balanced_accuracy
Unmitigated	0.134942	0.292857	0.125408
Postprocessing	0.103943	0.457143	0.205838
Reductions	0.087041	0.11875	0.080159

The reductions approach reduces the difference in all three metrics.

OPTUM Labs®

Use Case – Mitigation Comparison



	selection_rate	false_negative_rate	balanced_accuracy
Unmitigated	0.266321	0.285714	0.725585
Postprocessing	0.246928	0.324324	0.715908
Reductions	0.298497	0.378378	0.662718

The mitigated model using ExponeniatedGradient has a higher balanced error rate but a much smaller false negative rate difference.

OPTUM Labs®

Pre-Processing Techniques

Imbalanced Data: Some sensitive group are under-represented causing the model to be unable to perform similarly across all groups

- Sampling and Reweighing: Adjusts balance of different groups presenting more learning opportunities for under-represented groups

Data with Historical Bias: Causes model to learn discriminatory behavior from existing bias in the historical data

- Relabeling: Biased labels cause significant differences in false positives and false negatives between sensitive groups
  - Change labels for records that have same features but different outcomes
- Data Transformations: Models can indirectly infer association of certain records with sensitive groups based on certain features (i.e. zip code, sports)

  - Reduce or eliminate the correlation between sensitive attributes and other features as well as the target

Due to the blindness of these techniques to models' inference of the data, some level of bias can still creep into the model predictions.

Reweighing Technique

Disparate Impact (DI)= Probability of download given the person is not Caucasian divided by the probability of download given the person is Caucasian.

- For our fairness benchmark, we require that 1 - min(DI, 1/DI) < 0.2

Race	Download (1 = Yes, 0 = No)	Percent Downloaded
Not Caucasian	0	99.2%
Not Caucasian	1	0.8%
Caucasian	0	99.3%
Caucasian	1	0.7%

Disparate Impact (DI)	1/DI	1 - Min(DI, 1-DI)
1.21	0.83	0.17

Reweighing Technique

Reweighing is a preprocessing technique that weights the examples in each (group, label) combination differently to ensure fairness before classification.

Calculate weight for each group and label combination:

Group - Label	Numerator	Denominator	Weight
Caucasian - Download	Total # Caucasian * Total # Download	Total # * # Caucasian Download	1.065883
Caucasian - No Download	Total # Caucasian * Total # No Download	Total # * # Caucasian No Download	0.999558
Non-Caucasian - Download	Total # Non-Caucasian * Total # Download	Total # * # Non-Caucasian Download	0.883384
Non-Caucasian - Not Download	Total # Non-Caucasian * Total # No Download	Total # * # Non-Caucasian No Download	1.000945

Use Case -Takeaways

1) There is tradeoff between overall accuracy of the model and minimizing the difference in false negative rates between the groups.

2) The mitigated model using ExponeniatedGradient has a higher balanced error rate but a much smaller difference in false negative rate.

3) The disparate impact between Caucasian and Non-Caucasians is within the threshold of being fair.

4) More data will help identify if a mitigation algorithm is required.

Summary

1) Machine learning models should demonstrate parity across identified sensitive groups.
2) It is important to not only *identify* disparities, but also *mitigate* them.
3) Mitigation can occur during pre-processing, at training time, or during post-processing.
4) Determining which algorithm to use is a tradeoff between performance and level of parity.

Resources

Example code to run Fairlearn's MetricFrame, ThresholdOptimizer, and Exponeniated Gradient:

SciPy 2021 Tutorial.ipynb - Colaboratory (google.com)

Example code to run Fairlearn's GridSearch:

https://github.com/fairlearn/fairlearn/blob/main/notebooks/Grid%20Search%20for%20Binary%20Classification.ipynb

Reweighing using AIF360:

https://nbviewer.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb

Model remediation with TensorFlow:

https://www.tensorflow.org/responsible_ai/model_remediation

Technique for dealing with bias when no sensitive features are available:

https://arxiv.org/pdf/1806.08010.pdf

Summary of multiple bias mitigation techniques:

https://towardsdatascience.com/algorithmic-solutions-to-algorithmic-bias-aef59eaf6565

References

[1] Agarwal, Beygelzimer, Dudik, Langford, Wallach “A Reductions Approach to Fair Classification”, ICML, 2018.

[2] Agarwal, Dudik, Wu “Fair Regression: Quantitative Definitions and Reduction-based Algorithms”, ICML, 2019.

[3] Hardt, Price, Srebro “Equality of Opportunity in Supervised Learning”, NeurIPS, 2016.

Accuracy vs Balanced Accuracy

Balanced accuracy should be used as the target metric on imbalanced datasets.

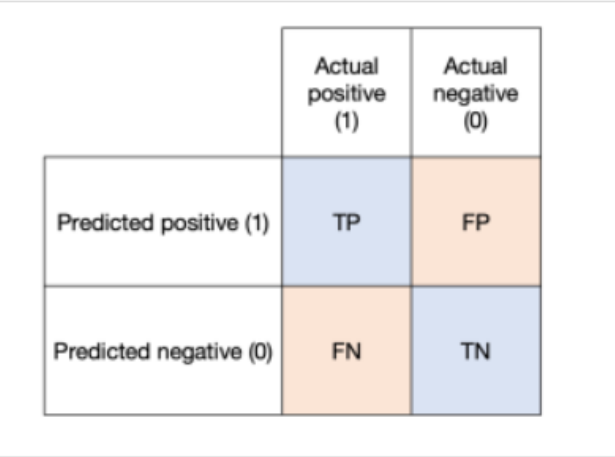
Accuracy: (TP + TN) / (TP + FN + FP + TN)

Sensitivity: TP / (TP + FN)

Specificity: TN/ (TN + FP)

Balanced Accuracy: (Sensitivity + Specificity) / 2

Example:



Accuracy: (5 + 10000) / (5 + 10 + 50 + 10000) = 99.4%

Sensitivity: 5 / (5 + 10) = 33.3%

Specificity: 10000 / (10000 + 50) = 99.5%

Balanced Accuracy = (33.3% + 99.5%) / 2 = 66.4%

Model is not doing much better than a random guess (i.e. 50%)