Fairness in Artificial Intelligence

Sheikh Rabiul Islam

shislam@hartford.edu

What is Fairness in AI

- AI algorithms are increasingly being used in high-stake decision making applications that affect individual lives.

- However, AI might exhibit algorithmic discrimination behaviors with respect to protected groups, potentially posing negative impacts on individuals and society.

- Fairness in AI (FAI) aims to build fair and unbiased AI/machine learning systems, that ensure the benefits are broadly available across all segments of society.

- Specific topics include but are not limited to: theoretical understanding of algorithmic bias, defining measurements of fairness, detection of adverse biases, and developing mitigation strategies.

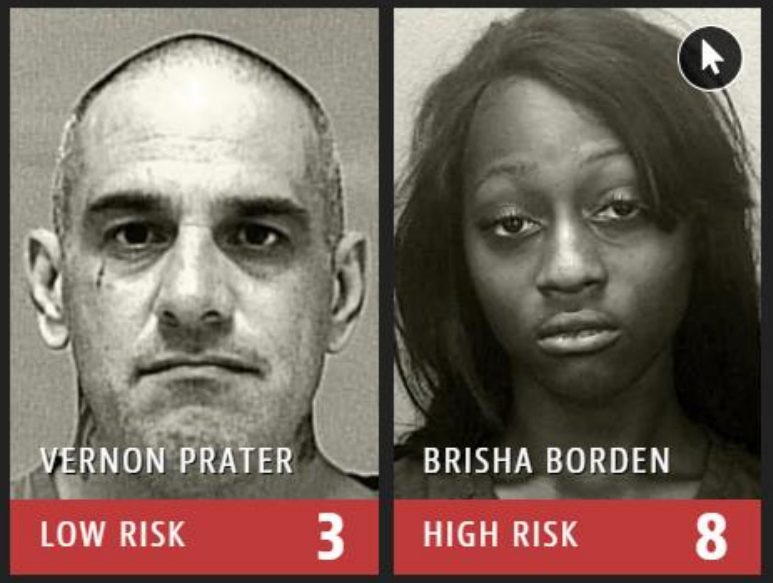
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two DUI Arrests

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

ProPublica is an independent, non-profit newsroom that produces investigative journalism in the public interest

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

Predictive Policing System is Bias

- “According to an NYU Law Review, predictive policing systems to forecast criminal activity and allocate police resources are built on data produced during documented periods of flawed, racially-biased, and sometimes unlawful practices and policies (a.k.a. ”**dirty policing**”). As a result, decisions grounded on systematically biased data (”**dirty data**”) can not escape the legacies of unlawful and biased policing practices”

Source: Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, Forthcoming, 2019.

Racial Bias in Health Algorithms

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

https://science.sciencemag.org/content/366/6464/447

Bias in Census data and Allocated Public Benefits

"A recent report reveals that native Hawaiians and Pacific Islanders are **underrepresented** in the census, due to different **systematic barriers and a reluctance to be counted,** resulting in a disproportionate allocation of resources (e.g., public funds)"

https://abcnews.go.com/Politics/pandemic-shows-native-hawaiians-pacific-islanders-participation-census/story?id=70873566

More on Bias

- Machine Learning (ML), the most common form of AI, is always a form of **statistical discrimination** by nature, and the discrimination becomes objectionable when it places certain privileged groups or individuals at a systematic advantage, and certain under-privileged groups at a systematic disadvantage

More on Bias

- “In the context of decision making, a fair decision is free from favoritism or prejudice towards individuals or groups based on their inherent or acquired characteristics; in contrast, a biased decision is skewed towards a particular person or group. Data bias and algorithmic bias are the primary contributing factors for fairness related risks”
- “A particular group can be disproportionately represented in the data due to natural or systematic bias in the data collection process; and a group could be subject to statistical discrimination that is inherent in AI algorithms”

Different Kinds of Discriminations

- Different kinds of discrimination that might occur include:

(1) **Direct Discrimination**: when the protected attributes (e.g., sex, race) of individuals explicitly result in a non-favorable outcome toward them;

(2) **Indirect discrimination**: when non-protected attributes (e.g., zip code) are used for decision making (e.g., loan approval decision), but the individual can still be discriminated from the implicit effect (e.g., an implicit guess of race from zip code) of the protected attribute (e.g., race);

(3) **Systemic discrimination**: results from flawed policies, custom, or behaviors (i.e., perpetuating discrimination against certain groups) that are part of the culture or structure of an organization;

(4) **Statistical discrimination**: results from the use of group statistics to judge an individual belonging to that group

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635, 2019.

Primary Factors for Fairness Related Risks

Data bias and **algorithmic bias** are the primary contributing factors for fairness related risks in AI-based decision making.

A particular group can be disproportionately represented in the data due to **natural or systematic bias** in the data collection process; and a group could be subject to statistical discrimination that is inherent in AI algorithms.

Tools

- Aequitas: Bias and Fairness Audit Toolkit (http://aequitas.dssg.io)
- Scikit-Fairness (https://scikit-fairness.netlify.app)
- AI Fairness 360 (https://aif360.mybluemix.net/)

Aequitas

- The Aequitas toolkit is a flexible bias-audit utility for algorithmic decision-making models, accessible via Python API, command line interface (CLI), and through their web application: http://aequitas.dssg.io/

- Aequitas can help you:
  - Understand where biases exist in your model(s)
  - Compare the level of bias between groups in your sample population (bias disparity)
  - Visualize absolute bias metrics and their related disparities for rapid comprehension and decision-making
- URL: https://dssg.github.io/aequitas

Preliminary Concepts

Name	Notation	Definition
Score	$S \in [0, 1]$	a real valued score assigned to each entity by the predictor.
Decision	$\widehat{Y} \in \{0, 1\}$	a binary prediction assigned to a given entity (data point).
True Outcome	$Y \in \{0, 1\}$	binary label of a given entity.
Attribute	$A = \{a_i, a_2, ..., a_n\}$	a multi-valued attribute, e.g., gender= {*female, male, other*}
Group	$g(a_i)$	all entities that share the same attribute value, e.g., gender=female.
Reference group	$g(a_r)$	one of the groups of A that is used as reference for calculating bias measures.
Labeled Positive	LP_g	number of entities labeled as positive within a group.
Labeled Negative	LN_g	number of entities labeled as negative within a group.
Predicted Positive	PP_g	number of entities within a group which decision is positive,i.e., $\widehat{Y} = 1$.
Total Pred. Positive	$K = \sum_{A=a_1}^{A=a_n} PP_{g(a_i)}$	total number of entities predicted positive across groups defined by A.
Predicted Negative	PN_g	number of entities within a group which decision is negative,i.e., $\widehat{Y} = 0$.
False Positive	FP_g	number of entities of the group with $\widehat{Y} = 1 \wedge Y = 0$
False Negative	FN_g	number of entities of the group with $\widehat{Y} = 0 \wedge Y = 1$.
True Positive	TP_g	number of entities of the group with $\widehat{Y} = 1 \wedge Y = 1$.
True Negative	TN_g	number of entities of the group with $\widehat{Y} = 0 \wedge Y = 0$.

Reference groups: Fairness is always determined in relation to a reference group. By default, aequitas uses the **majority group** level for a given group as the **reference group**.

Bias Metrics

Name	Notation	Definition
Prevalence	$Prev_g = LP_g / \|g\| = \Pr(Y=1 \mid A=a_i)$	fraction of entities within a group which true outcome was positive.
Predicted Prevalence	$PPrev_g = PP_g / \|g\| = \Pr(\widehat{Y}=1 \mid A=a_i)$	fraction of entities within a group which were predicted as positive.
Predicted Positive Rate	$PPR_g = PP_g / K = \Pr(a=a_i \mid \widehat{Y}=1)$	fraction of the entities predicted as positive that belong to a certain group.
False Discovery Rate	$FDR_g = FP_g / PP_g = \Pr(Y=0 \mid \widehat{Y}=1, A=a_i)$	fraction of false positives of a group within the predicted positive of the group
False Omission Rate	$FOR_g = FN_g / PN_g = \Pr(Y=1 \mid \widehat{Y}=0, A=a_i)$	fraction of false negatives of a group within the predicted negative of the group
False Positive Rate	$FPR_g = FP_g / LN_g = \Pr(\widehat{Y}=1 \mid Y=0, A=a_i)$	fraction of false positives of a group within the labeled negative of the group
False Negative Rate	$FNR_g = FN_g / LP_g = \Pr(\widehat{Y}=0 \mid Y=1, A=a_i)$	fraction of false negatives of a group within the labeled positives of the group

COMPAS Analysis using Aequitas

- In 2016, **ProPublica** reported on racial inequality in automated criminal risk assessment algorithms.
- Northpointe's **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) is one of the most widely utilized risk assessment tools/ algorithms within the **criminal justice system** for guiding decisions such as **how to set bail**.
- The ProPublica dataset represents two years of COMPAS predicitons from Broward County, FL.
- COMPAS produces a risk score that **predicts a person's likelihood of committing a crime in the next two years**.

https://dssg.github.io/aequitas/examples/compas_demo.html

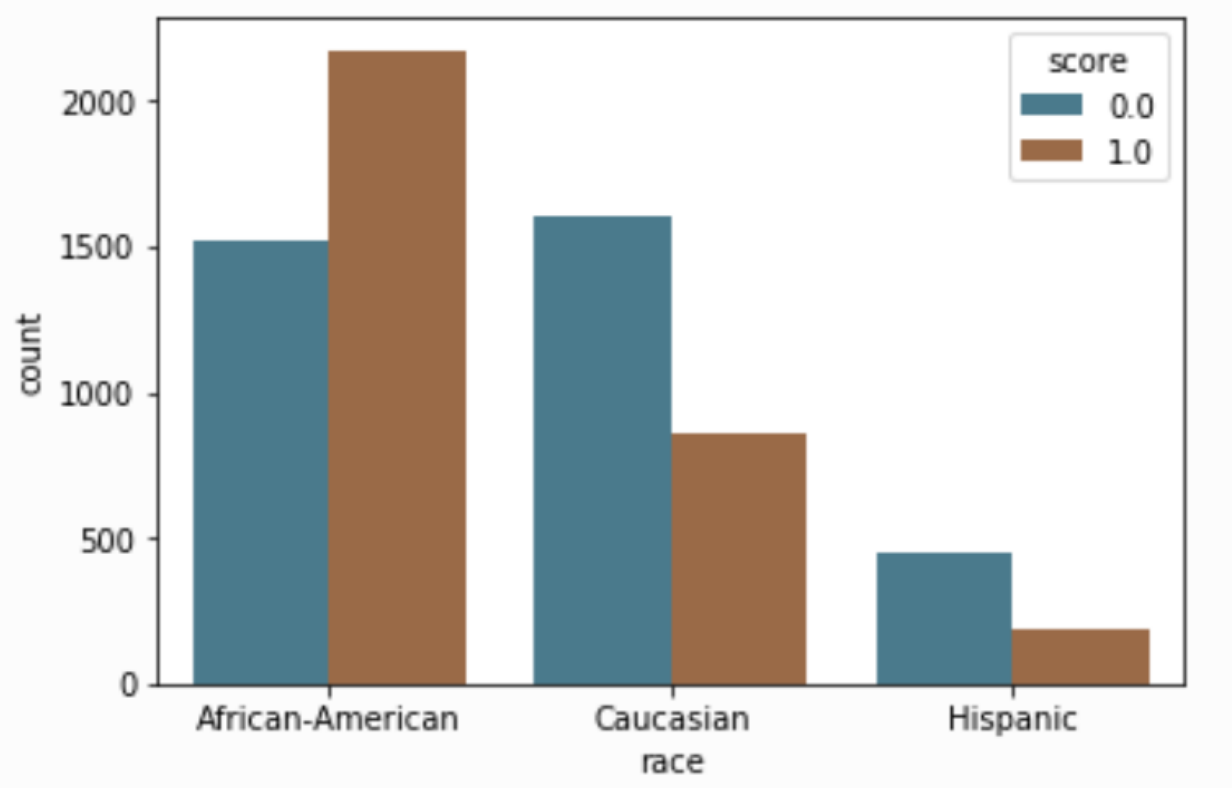
COMPAS Analysis using Aequitas

A score of 0 indicates a prediction (chance of recidivism) of “low” risk according to COMPAS, while a 1 indicates “high” or “medium” risk.

score	label_value	race	sex	age_cat
0.0	0	Other	Male	Greater than 45
0.0	1	African-American	Male	25 - 45
0.0	1	African-American	Male	Less than 25
1.0	0	African-American	Male	Less than 25
0.0	0	Other	Male	25 - 45

https://dssg.github.io/aequitas/examples/compas_demo.html

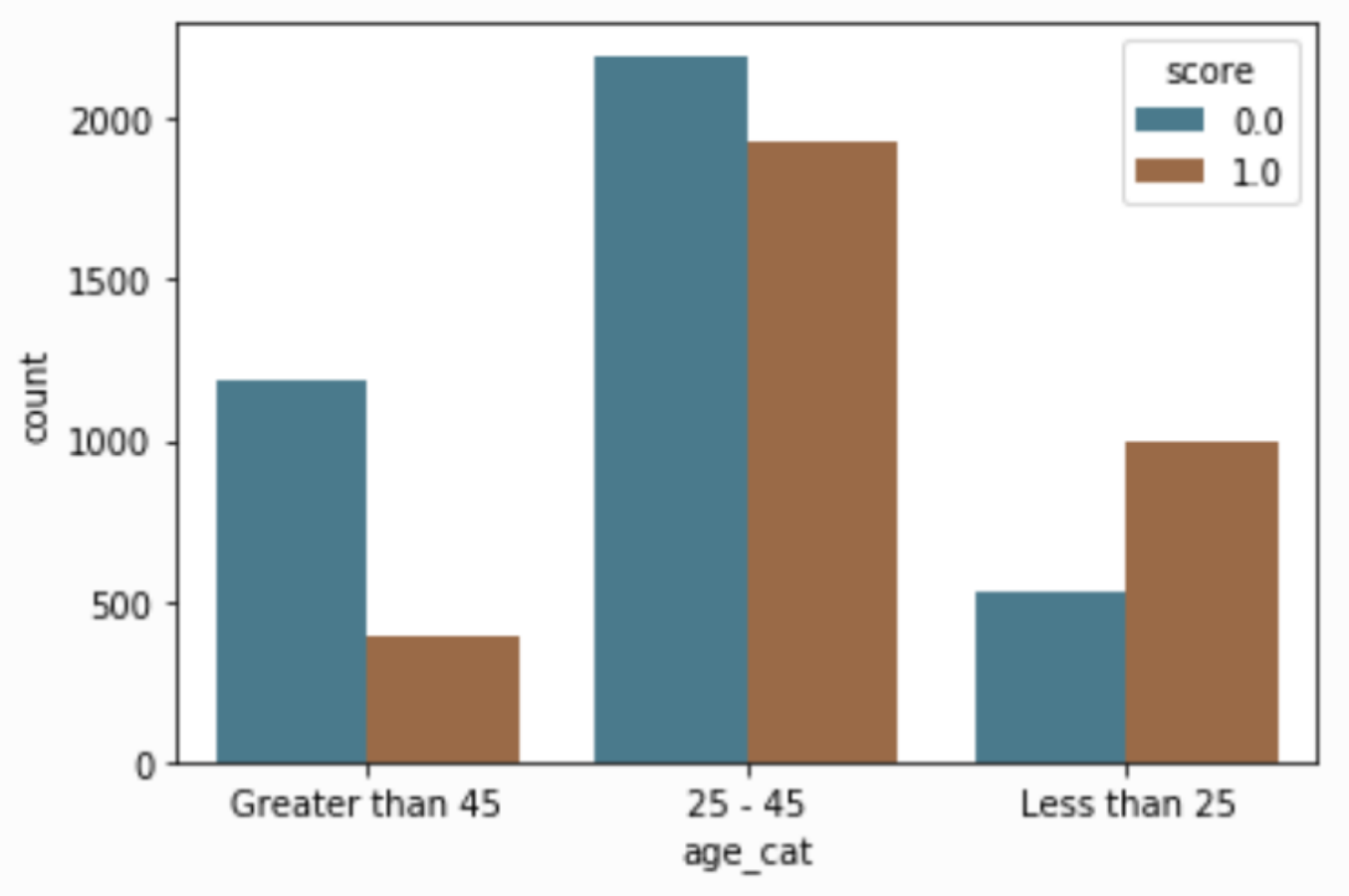
COMPAS Analysis using Aequitas : Predictiction



We see a large difference in how these scores are distributed by race, with a **majority of white and Hispanic people predicted as low risk** (score = 0) and a **majority of black people predicted high and medium risk** (score = 1).

https://dssg.github.io/aequitas/examples/compas_demo.html

COMPAS Analysis using Aequitas: Prediction

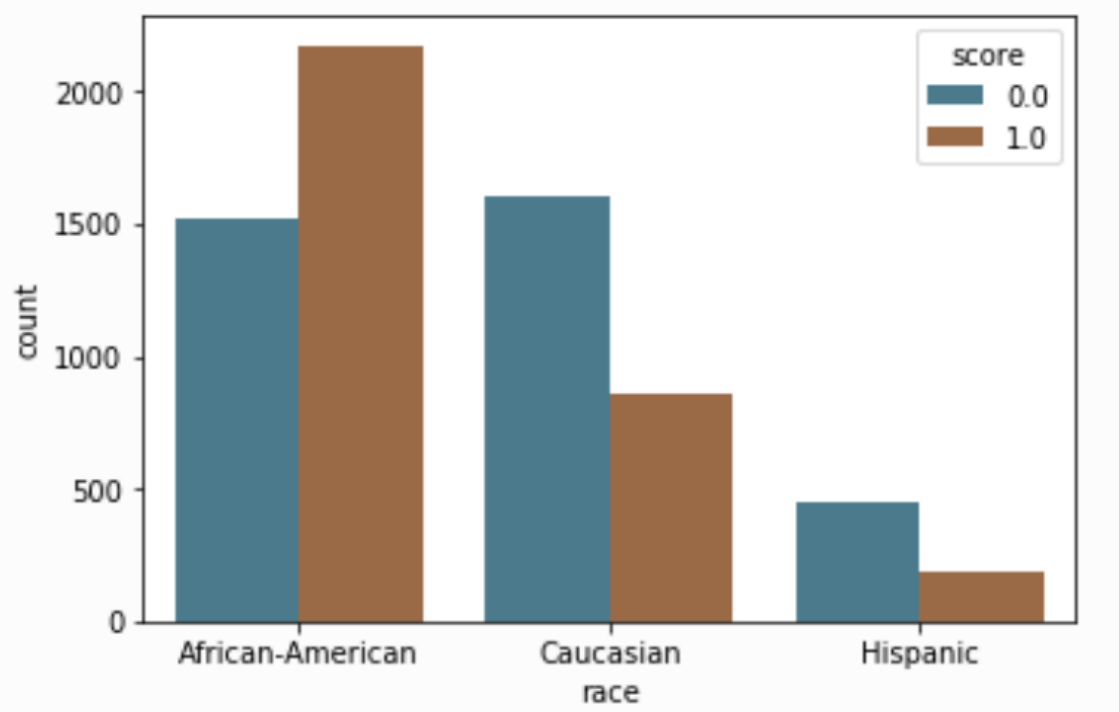


We also see that while the majority of people in age categories over 25 are predicted as low risk (score = 0), the majority of people below 25 are predicted as high and medium risk (score = 1)

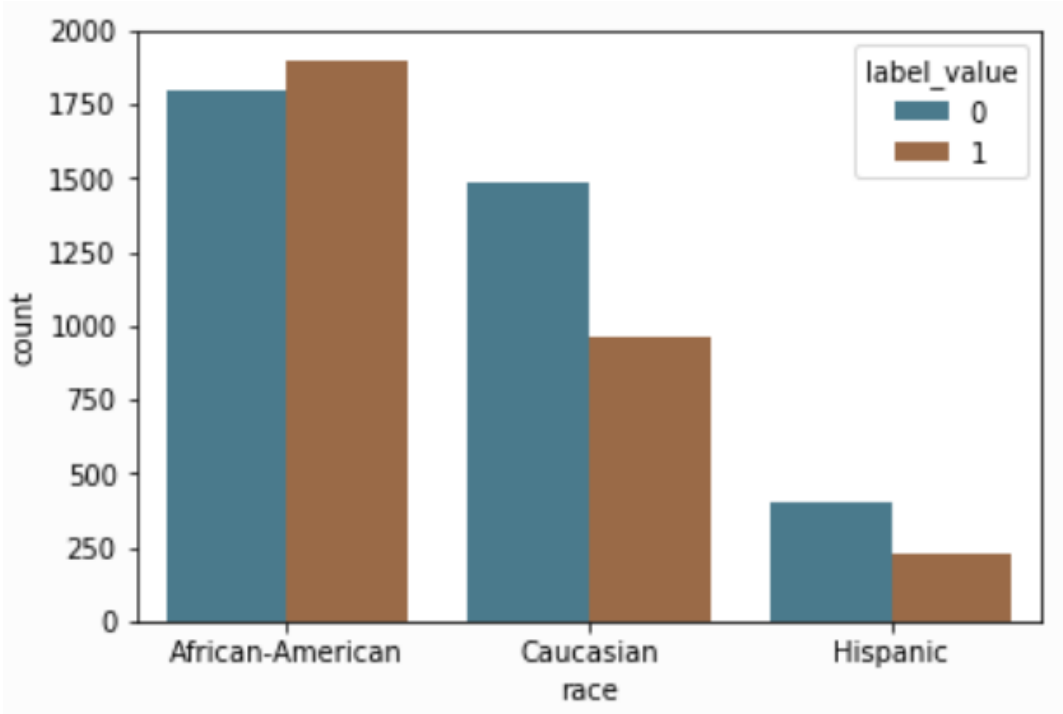
https://dssg.github.io/aequitas/examples/compas_demo.html

COMPAS Analysis using Aequitas

Prediction



Ground Truth

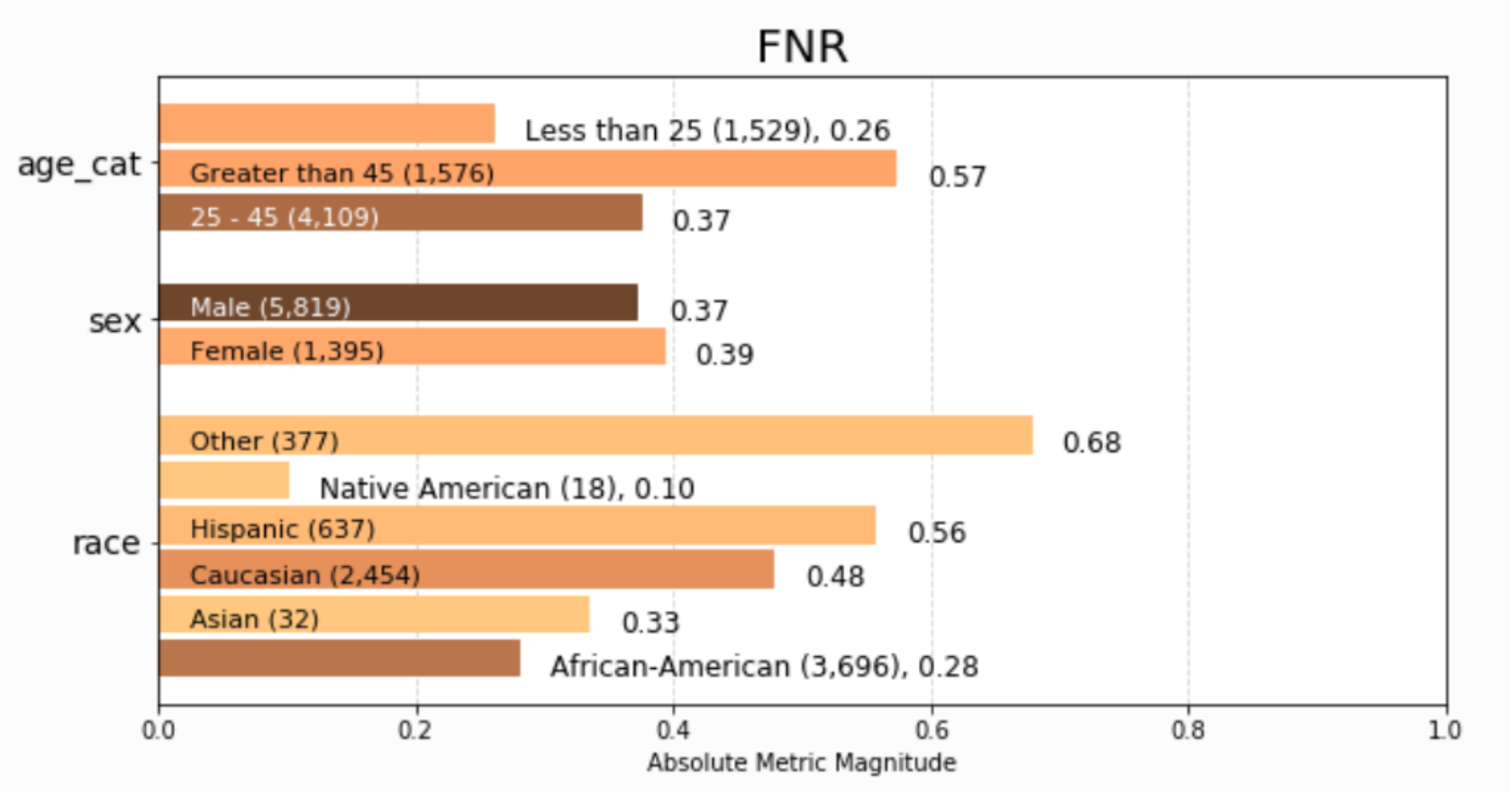


- A **recidivism score** is meant to predict a new misdemeanor or felony offense within two years of the COMPAS administration date

- The graphs above (**ground truth**) show the base rates for recidivism are higher for black defendants compared to white defendants (.51 vs .39), though the **predictions** do not match the base rates.

https://dssg.github.io/aequitas/examples/compas_demo.html

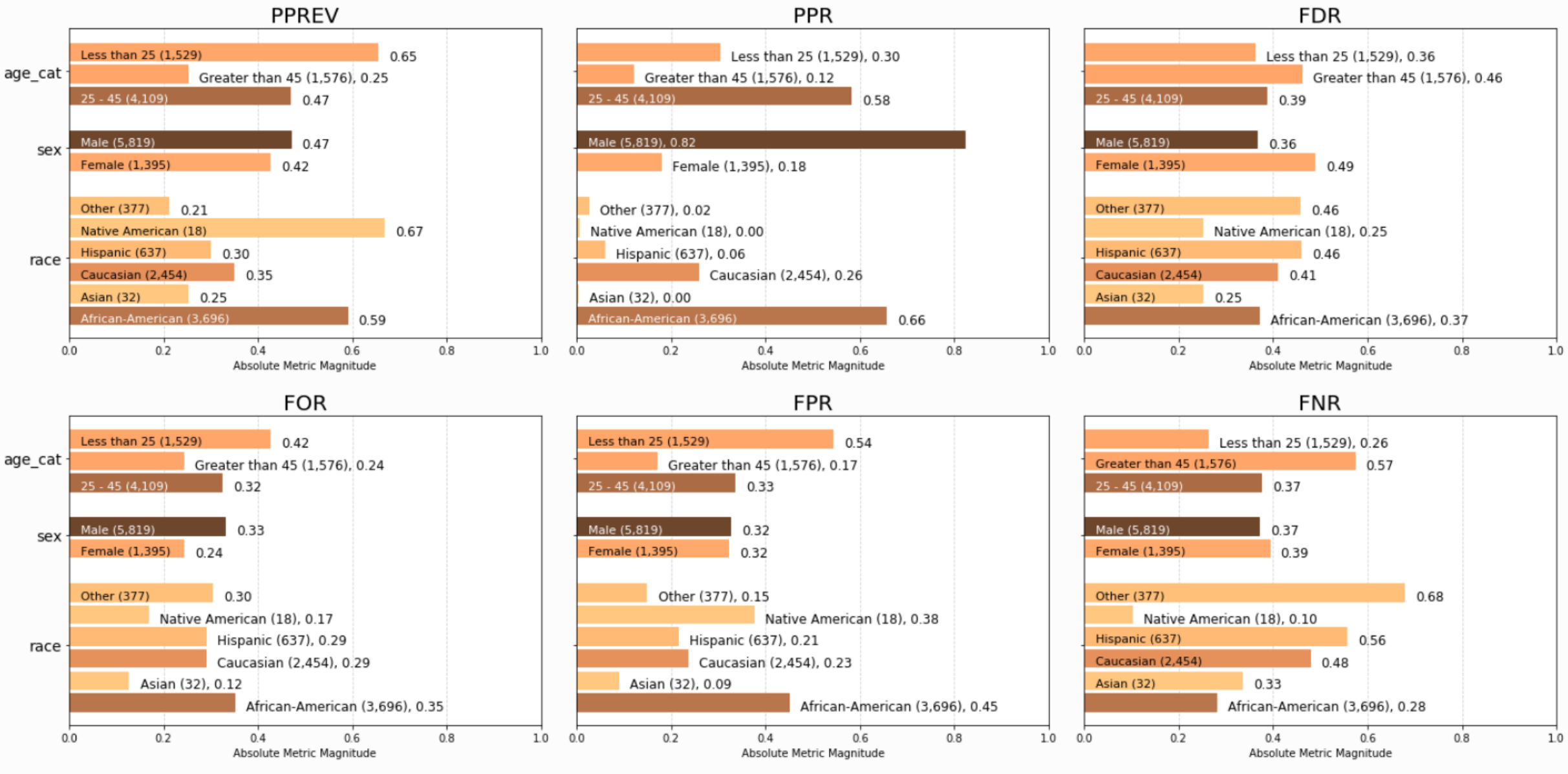
COMPAS Analysis using Aequitas



- False Negative Rate (FNR) calculated across each attribute, colored based on number of samples in the attribute group.
- FNR of African American is less compared to most of other groups.

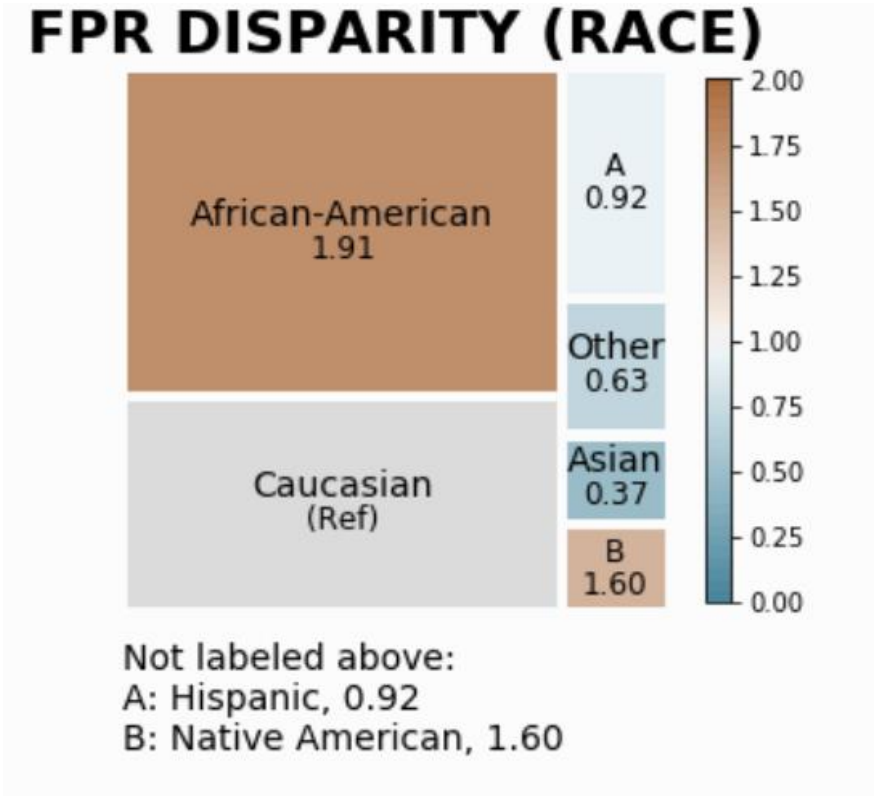
https://dssg.github.io/aequitas/examples/compas_demo.html

COMPAS Analysis using Aequitas

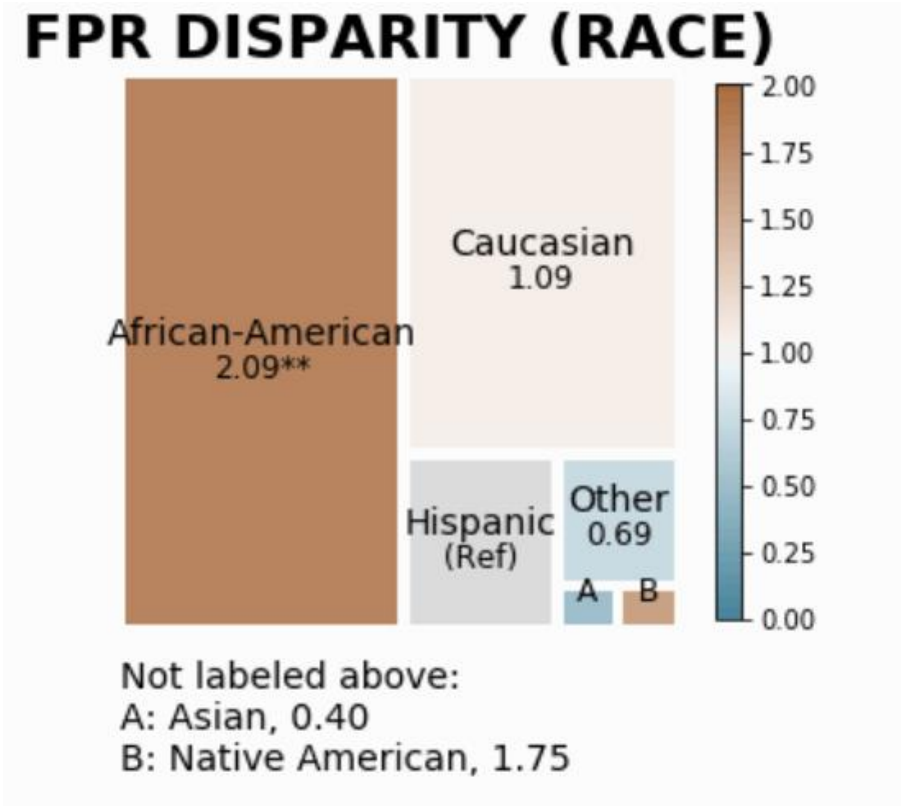


https://dssg.github.io/aequitas/examples/compas_demo.html

COMPAS Analysis using Aequitas



Reference group: Caucasian



Reference group: Hispanic

https://dssg.github.io/aequitas/examples/compas_demo.html

COMPAS Analysis using Aequitas

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://dssg.github.io/aequitas/examples/compas_demo.html
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS Analysis using Aequitas : Web Demo

Aequitas
Bias & Fairness Audit

Home Code About

Bias and Fairness Audit Toolkit

The Bias Report is powered by Aequitas, an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

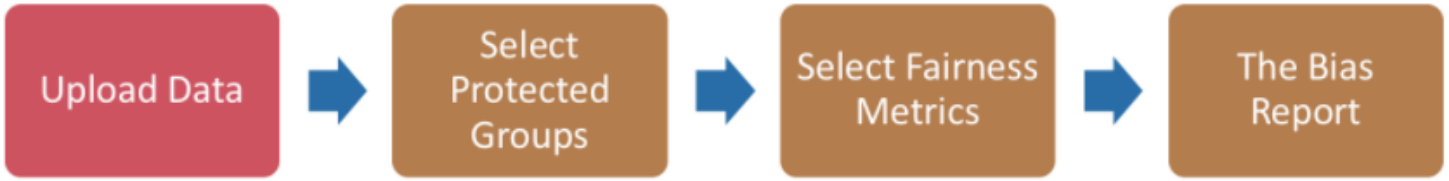
See an example report on COMPAS risk assessment scores.

Or try out the audit tool using your own data or one of our sample data sets.

Get Started!

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo



Try auditing a sample data set

COMPAS Recidivism Risk Assessment Data
[About the Data]

US Adult Income Data
[About the Data]

Or audit your own data

Choose File No file chosen

Upload

See below for information on how to format input data.

Data you upload is used to generate the audit report. While the data is deleted, we host the audit report in perpitutity. If your data is private and sensitive, we encourage you to use the desktop version of the audit tool
.

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

Configure the Bias Audit

Select attributes to audit and a method for determining the reference groups.

1. Select method for determining reference group:

Reference groups are used to calculate relative disparities in our Bias Audit. For example, you might select **Male** as the reference group for Gender. Aequitas will then use **Male** as the baseline to calculate any biases for other groups in the attribute Gender (for **Female** and **Other** for example).

- (●) Custom group (Select your own)
- ( ) Majority group (Automatically select the largest group for every attribute)
- ( ) Automatically select group with the lowest bias metric for every attribute

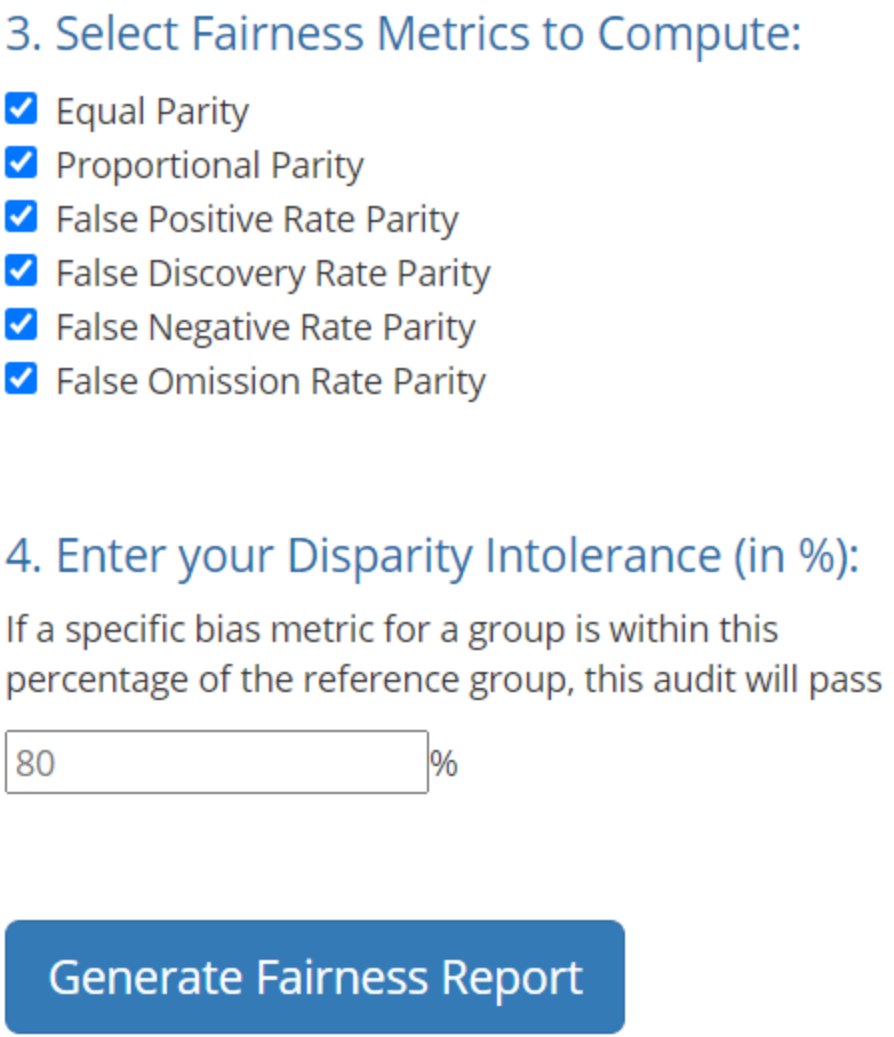
2. Select protected attributes that need to be audited for bias.

Attribute	Reference Group
[x] race	Caucasian
[x] sex	Male
[ ] age_cat	25 - 45

Next!

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo



http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

The Bias Report

Audit Date:	08 Nov 2020
Data Audited:	7214 rows
Attributes Audited:	race, sex
Audit Goal(s):	Equal Parity - Ensure all protected groups are have equal representation in the selected set.
	Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population.
	False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group).
	False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group).
	False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group).
	False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group).
Reference Groups:	Custom group - The reference groups you selected for each attribute will be used to calculate relative disparities in this audit.
Fairness Threshold:	80%. If disparity for a group is within 80% and 125% of the value of the reference group on a group metric (e.g. False Positive Rate), this audit will pass.

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

Audit Results: Summary

Equal Parity - Ensure all protected groups are have equal representation in the selected set.	Failed	Details
Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population.	Failed	Details
False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group).	Failed	Details
False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group).	Failed	Details
False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group).	Failed	Details
False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group).	Failed	Details

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

Audit Results: Summary

Equal Parity - Ensure all protected groups are have equal representation in the selected set.	Failed	Details
Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population.	Failed	Details
False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group).	Failed	Details
False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group).	Failed	Details
False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group).	Failed	Details
False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group).	Failed	Details

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

Audit Results: Details by Fairness Measures

Equal Parity: **Failed**

What is it?

This criteria considers an attribute to have equal parity is every group is equally represented in the selected set. For example, if race (with possible values of white, black, other) has equal parity, it implies that all three races are equally represented (33% each)in the selected/intervention set.

When does it matter?

If your desired outcome is to intervene equally on people from all races, then you care about this criteria.

Which groups failed the audit:

For race (with reference group as **Caucasian**)
- Native American with **0.01X** Disparity
- Other with **0.09X** Disparity
- African-American with **2.55X** Disparity
- Asian with **0.01X** Disparity
- Hispanic with **0.22X** Disparity

For sex (with reference group as **Male**)
- Female with **0.22X** Disparity

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

False Positive Rate Parity: **Failed**

What is it?

This criteria considers an attribute to have False Positive parity if every group has the same False Positive Error Rate. For example, if race has false positive parity, it implies that all three races have the same False Positive Error Rate.

When does it matter?

If your desired outcome is to make false positive errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is punitive and has a risk of adverse outcomes for individuals. Using this criteria allows you to make sure that you are not making false positive mistakes about any single group disproportionately.

Which groups failed the audit:

For race (with reference group as **Caucasian**)

- Other with 0.63X Disparity
- Asian with 0.37X Disparity
- Native American with 1.60X Disparity
- African-American with 1.91X Disparity

http://aequitas.dssg.io

COMPAS Analysis using Aequitas : Web Demo

False Negative Rate Parity: **Failed**

What is it?

This criteria considers an attribute to have False Negative parity if every group has the same False Negative Error Rate. For example, if race has false negative parity, it implies that all three races have the same False Negative Error Rate.

When does it matter?

If your desired outcome is to make false negative errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is assistive (providing helpful social services for example) and missing an individual could lead to adverse outcomes for them. Using this criteria allows you to make sure that you're not missing people from certain groups disproportionately.

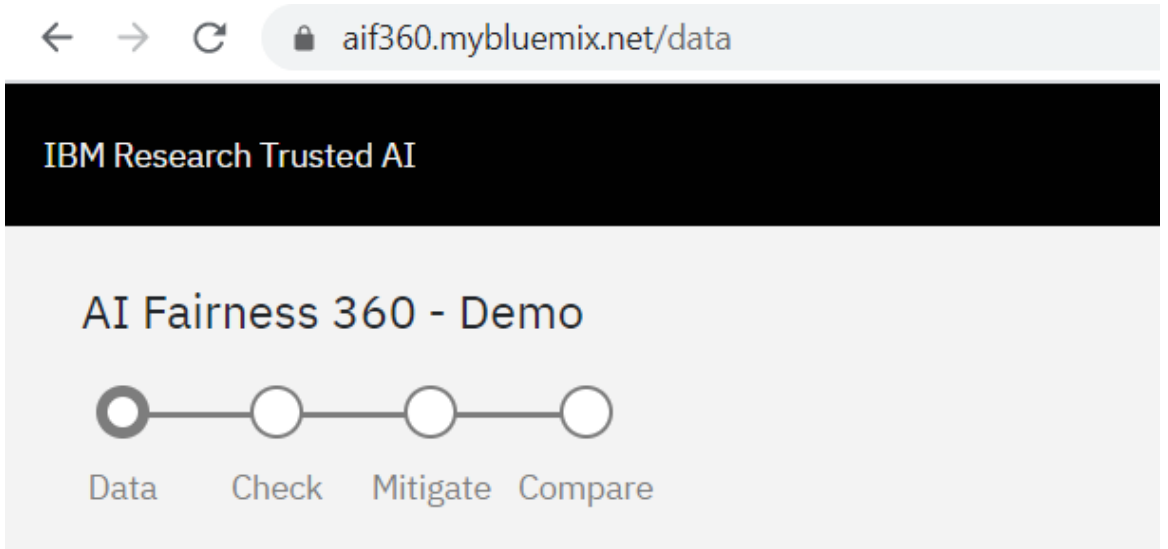
Which groups failed the audit:

For race (with reference group as **Caucasian**)

- Native American with 0.21X Disparity
- African-American with 0.59X Disparity
- Asian with 0.70X Disparity
- Other with 1.42X Disparity

http://aequitas.dssg.io

AI Fairness 360 - Demo



https://aif360.mybluemix.net/data

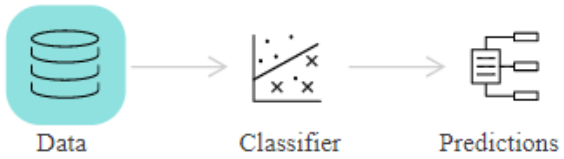
AI Fairness 360 – Demo: Bias Mitigation

3. Choose bias mitigation algorithm

A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (i process), or the predictions (post-process). Learn more about how to choose.

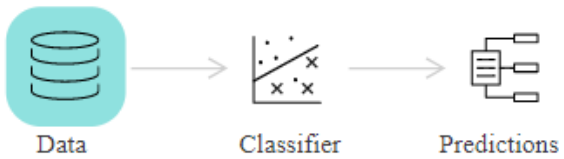
◯ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.



◯ **Optimized Pre-Processing**

Learns a probabilistic transformation that can modify the features and the labels in the training data.



◯ **Adversarial Debiasing**

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the

https://aif360.mybluemix.net/data

Statistical Parity Difference

Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

The ideal value of this metric is 0

Fairness for this metric is between -0.1 and 0.1

Equal Opportunity Difference

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.

The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference

Computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.

The ideal value of this metric is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Fairness for this metric is between -0.1 and 0.1

Disparate Impact

Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

The ideal value of this metric is 1.0 A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Fairness for this metric is between 0.8 and 1.25

Theil Index

Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals.

A value of 0 implies perfect fairness.

4. Compare original vs. mitigated results

Dataset: Compas (ProPublica recidivism)

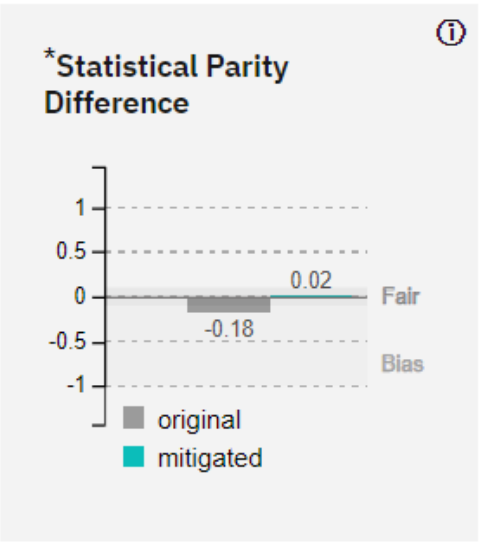
Mitigation: **Reweighing algorithm applied**

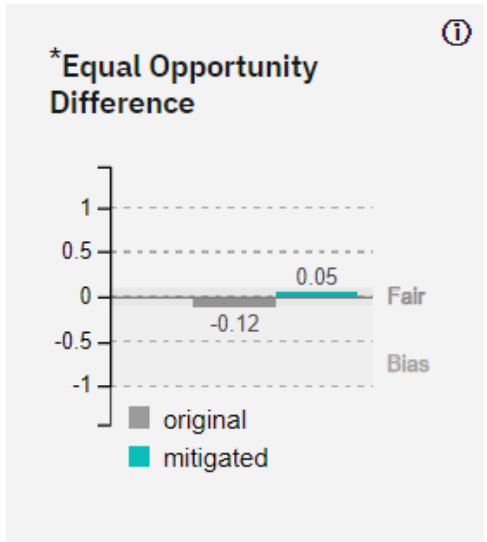
Protected Attribute: Race

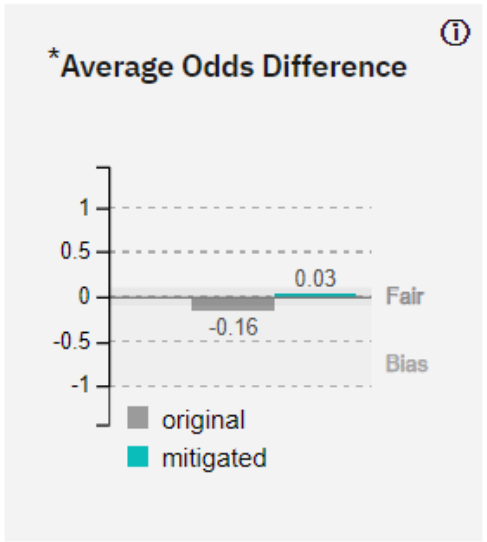
Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***

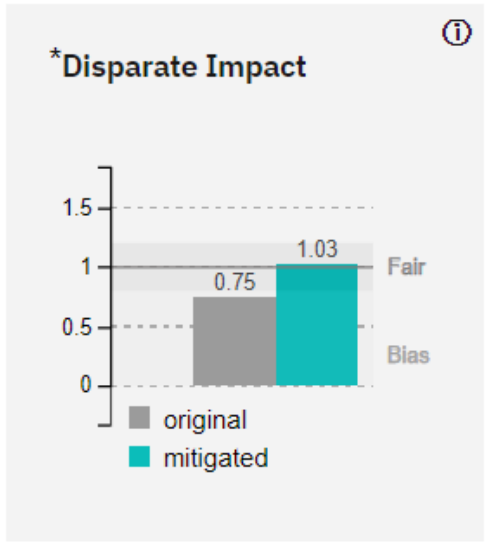
Accuracy after mitigation unchanged

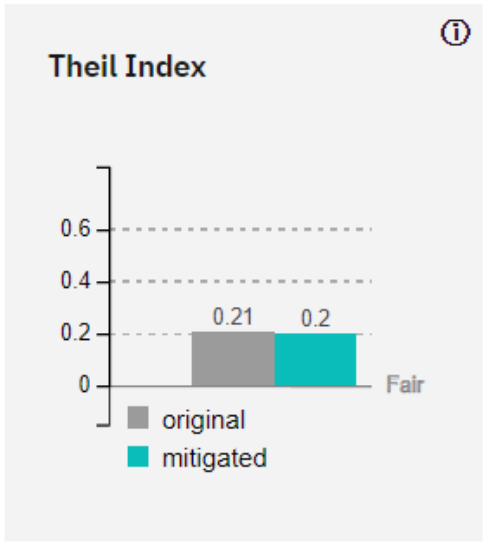
Bias against unprivileged group was reduced to acceptable levels* for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)











https://aif360.mybluemix.net/data

Reweighing

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	–
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Education	–
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Board	+

i.e., the weight of an object will be the expected probability to see an instance with its sensitive attribute value and class given independence, divided by its observed probability.

In this way we assign a weight to every tuple according to its S and *Class*-values. We will call the dataset D with the added weights, D_W. It is easy to see that D_W is unbiased; i.e., if we multiply the frequency of every object by its weight, the discrimination would be 0. On this balanced dataset the discrimination-free classifier is learned.

Example 3 Consider again the dataset in Table 1. The weight for each data object is computed according to its S- and *Class*-value. We calculate the weight of a data object with $X(S) = f$ and $X(Class) +$ as follows. We know that 50% objects have $X(S) = f$ and 60% objects have *Class*-value +, so the expected probability of the object should be:

$$P_{exp}(Sex = f \ \wedge \ X(Class) \ = \ +) = 0.5 \times 0.6 = 30\%$$

but its actually observed probability is 20%. So the weight $W(X)$ will be:

$$W(X) = \frac{0.5 \times 0.6}{0.2} = 1.5 \ .$$

Similarly, the weights of all other combinations are as follows:

$$W(X) := \begin{cases} 1.5 & \text{if } X(Sex) = f \text{ and } X(Class) \ = \ + \\ 0.67 & \text{if } X(Sex) = f \text{ and } X(Class) \ = \ - \\ 0.75 & \text{if } X(Sex) = m \text{ and } X(Class) \ = \ + \\ 2 & \text{if } X(Sex) = m \text{ and } X(Class) \ = \ - \end{cases}$$

The weight of each data object of the Table 1 is given in Table 4.

(1) F. Kamiran and T. Calders, “Data Preprocessing Techniques for Classification without Discrimination,” Knowledge and Information Systems, 2012.
(2) https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.preprocessing.Reweighing.html

Adversarial Debiasing

Adversarial debiasing is an in-processing technique that learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

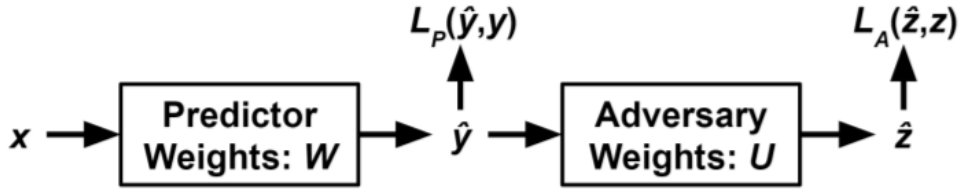


Figure 1: The architecture of the adversarial network.

We begin with a model, which we call the *predictor*, trained to accomplish the task of predicting Y given X. As in Figure 1, we assume that the model is trained by attempting to modify weights W to minimize some loss $L_P(\hat{y}, y)$, using a gradient-based method such as stochastic gradient descent.

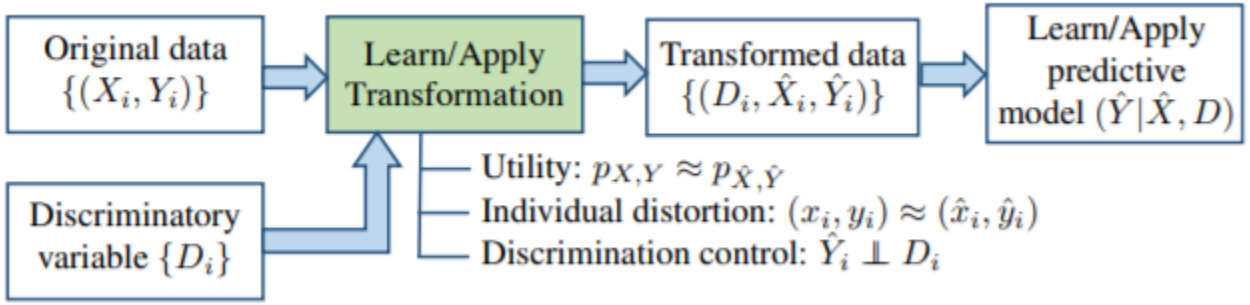
The output layer of the predictor is then used as an input to another network called the *adversary* which attempts to predict Z. This is part of the network corresponds to the *discriminator* in a typical GAN (Goodfellow et al. 2014). We will suppose the adversary has loss term $L_A(\hat{z}, z)$ and weights U. Depending on the definition of fairness being achieved, the adversary may have other inputs.

- For DEMOGRAPHIC PARITY, the adversary gets the predicted label $\hat{Y}$. Intuitively, this allows the adversary to try to predict the protected variable using nothing but the predicted label. The goal of the predictor is to prevent the adversary from doing this.
- For EQUALITY OF ODDS, the adversary gets $\hat{Y}$ and the true label Y.
- For EQUALITY OF OPPORTUNITY on a given class y, we can restrict the training set of the adversary to training examples where $Y = y$.[3]

(1) B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2018.
(2)https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.inprocessing.AdversarialDebiasing.html

Optimized Preprocessing

Optimized preprocessing is a preprocessing technique that learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives



(1) F. P. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. “Optimized Pre-Processing for Discrimination Prevention.” Conference on Neural Information Processing Systems, 2017.
(2)https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.preprocessing.OptimPreproc.html

Reject Option Classification

Reject option classification is a postprocessing technique that gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

- The approach utilize decision theory to make standard probabilistic classifiers (ROC) and classifier ensembles (DAE) discrimination-aware.
- Both ROC and DAE ensure discrimination-aware classifications at run-time without data modification or algorithm tweaking.
- ROC can also be interpreted as a cost-based classification method in which the cost of misclassifying a deprived group instance as negative is much higher than that of misclassifying a favored group instance as negative.

(1) F. Kamiran, A. Karim, and X. Zhang, “Decision Theory for Discrimination-Aware Classification,” IEEE International Conference on Data Mining, 2012.
(2) https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.postprocessing.RejectOptionClassification.html

About Your Term Project

Reminding you about the special requirements for your term project :

You need to assess your term project for possible bias and discriminations. If there is bias, you need to write a section on bias detection and mitigation. If your project doesn't have any fairness related risk then you need to provide justification of that too. You can use Aequitas and IBM AI Fairness 360 or other software for bias detection and mitigation.

https://aif360.mybluemix.net/data
http://aequitas.dssg.io

Other Resources

- NIPS 2017 Video Tutorial on Fairness in Machine LearningAequitas: Bias and Fairness Audit Toolkit https://fairmlbook.org/tutorial1.html
- 21 fairness definitions and their politics (https://www.youtube.com/embed/jIXIuYdnyyk)

- Book: https://fairmlbook.org

Thank you!