Ethical & Privacy Considerations

Contents

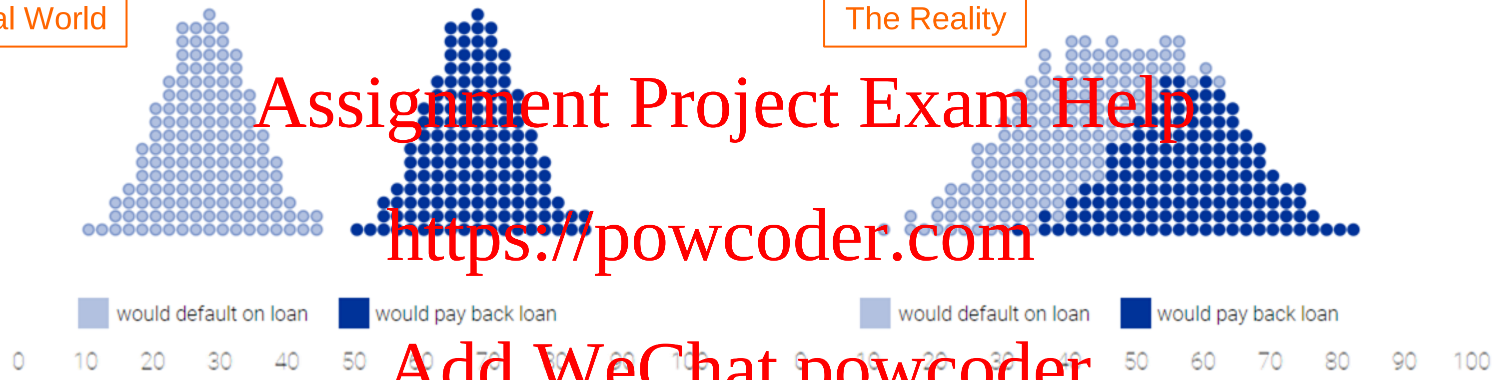
- Algorithmic Fairness
- Source of Bias
- Aequitas Discrimination & Bias Audit Toolkit
- Bias Mitigation
- Ethical Machine Learning
- Deon Data Science Ethics Checklist

Algorithmic Fairness

Ideally, we would use statistics that cleanly separate categories but overlapping categories are the norm

An Ideal World

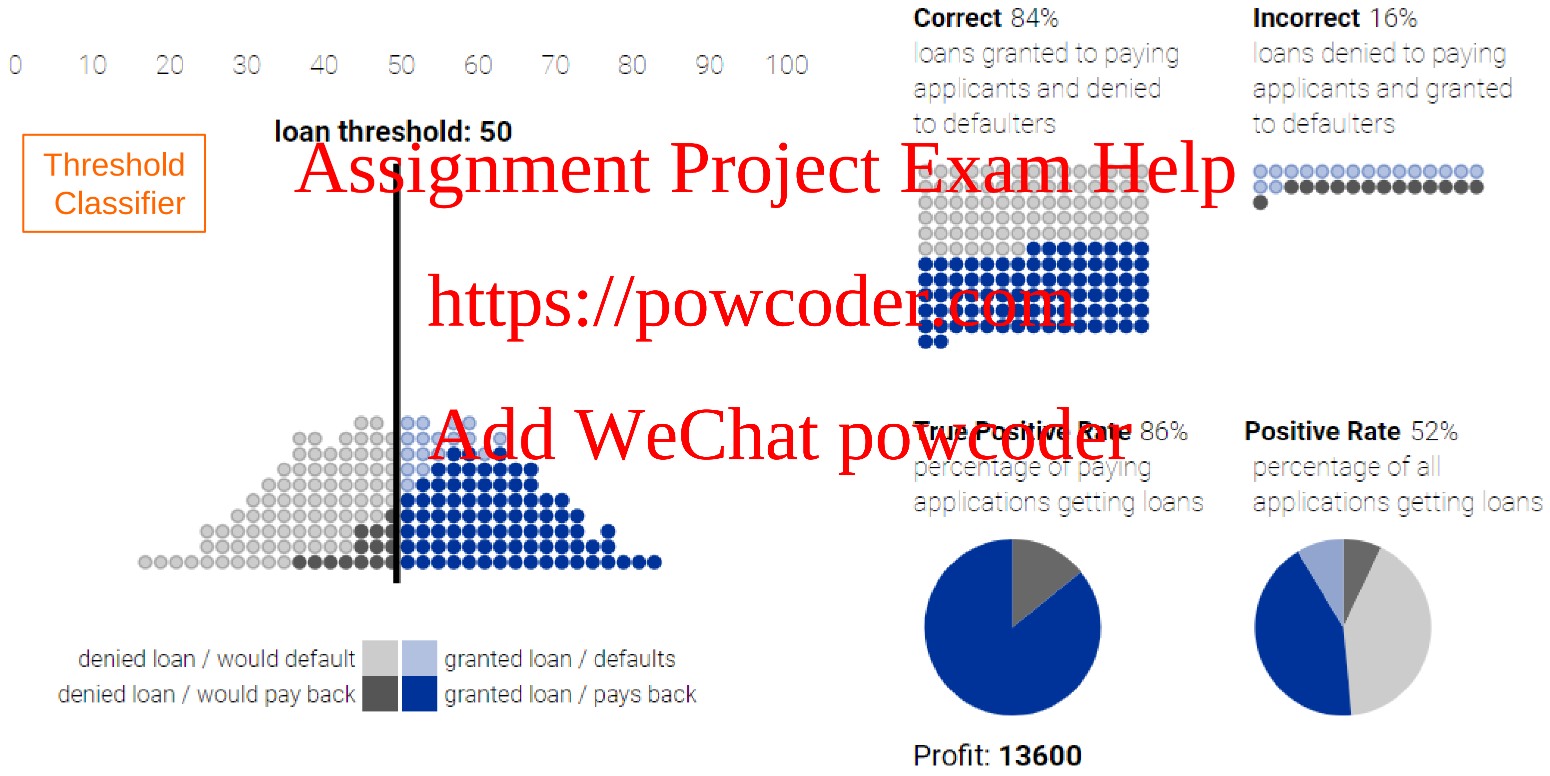
The Reality



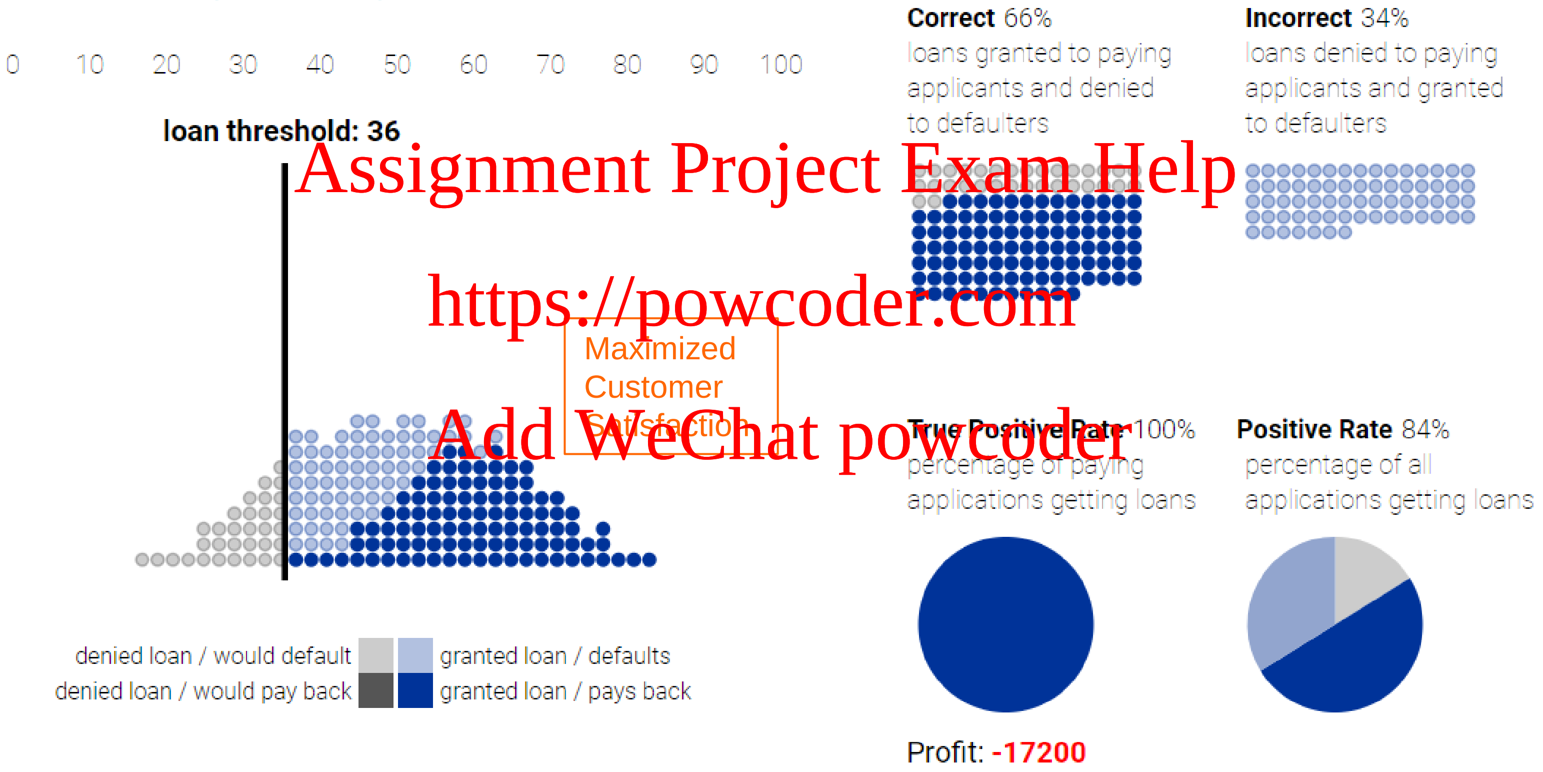
Credit Score: higher scores represent higher likelihood of payback
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

- A single statistic can stand in for many different variables, boiling them down to one number
- In the case of a credit score, which is computed looking at a number of factors, including income, promptness in paying debts, etc., the number might correctly represent the likelihood that a person will pay off a loan, or default
- Or it might not
- The relationship is usually fuzzy – it is rare to find a statistic that correlates perfectly with real-world outcomes

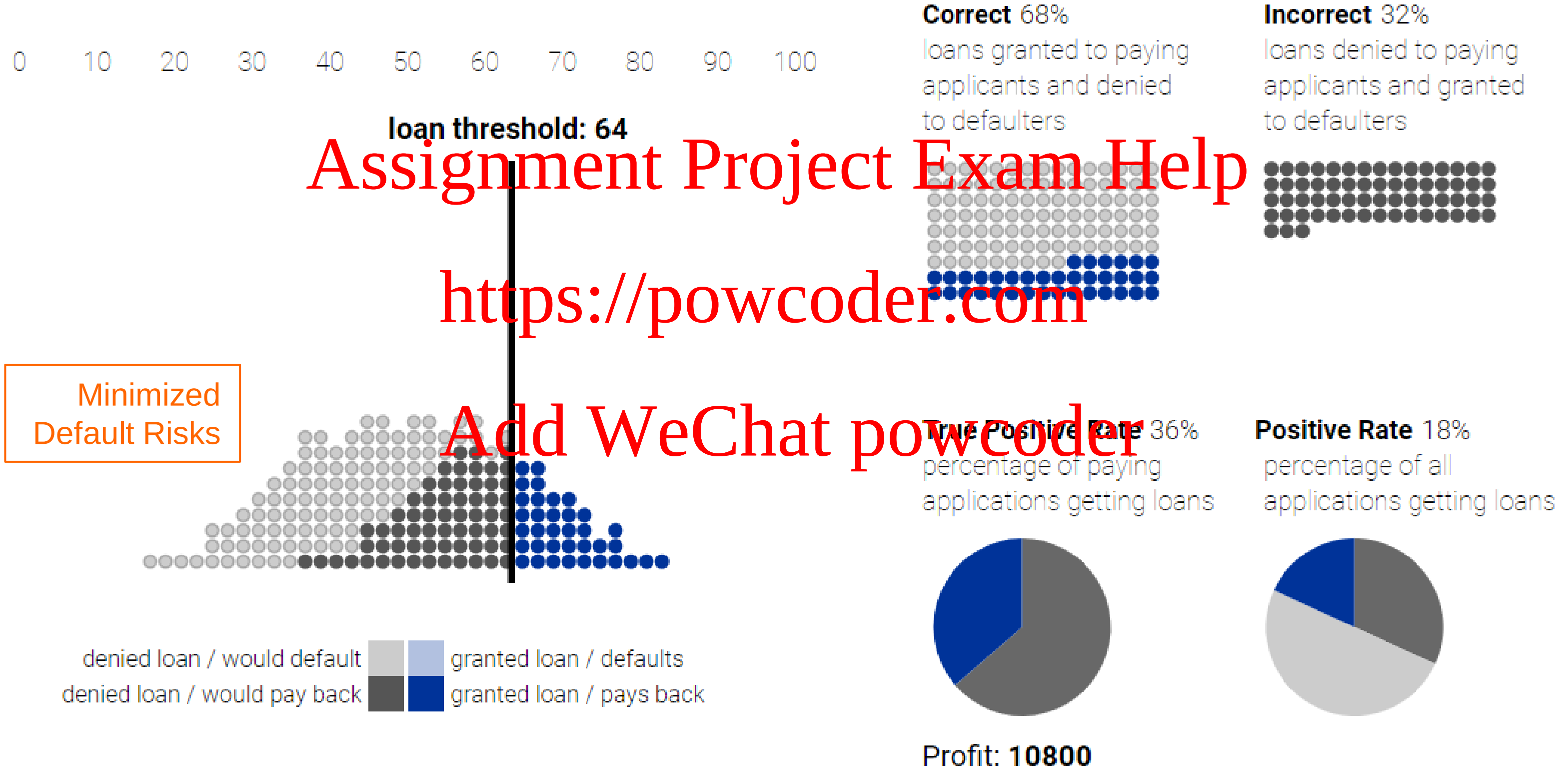
People whose credit scores are below the cut-off / threshold are denied the loan, people above it are granted the loan



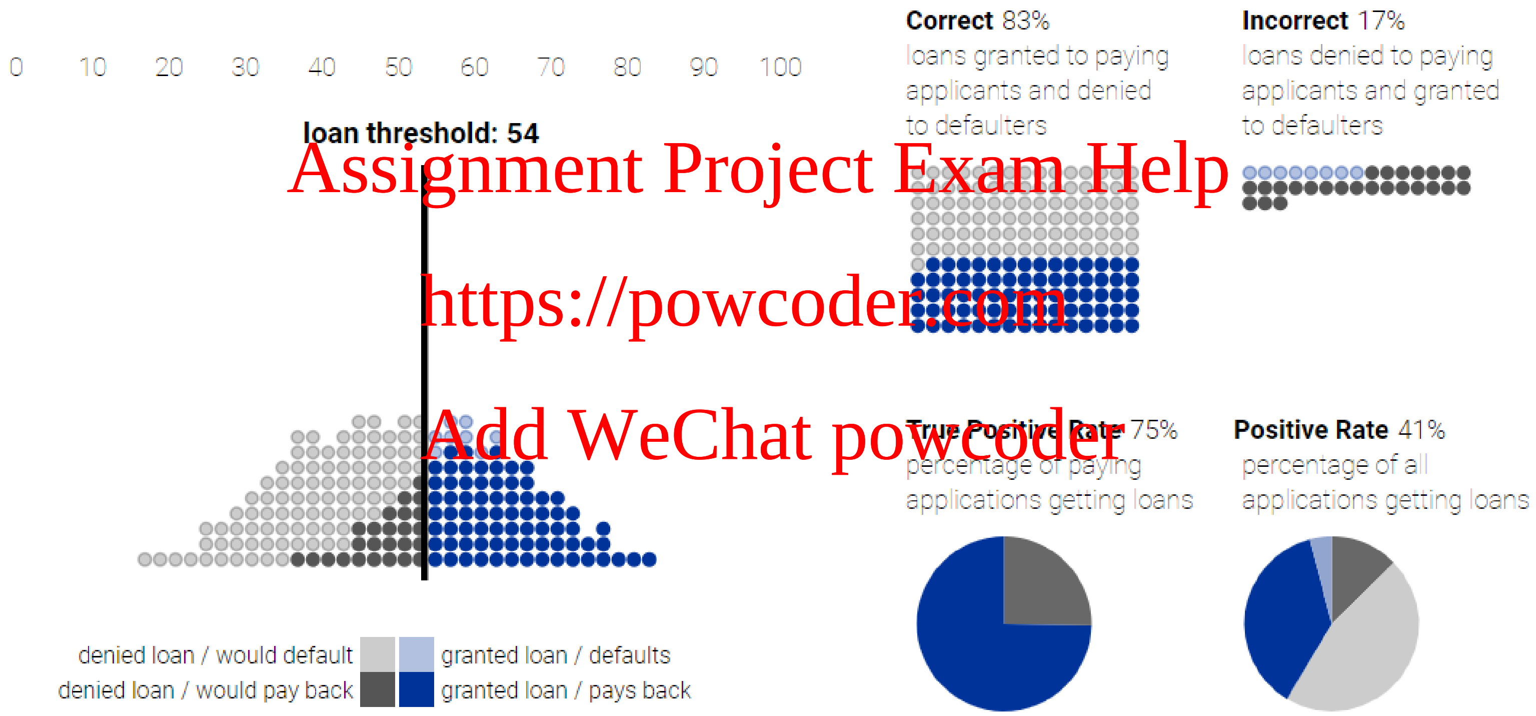
All paying applicants get granted the loan but the number of defaulters getting the loan also increases



All defaulters are denied the loan but a large number of paying applicants are also wrongly denied the loan



Optimal profit is attained using a threshold credit score of 54



Profit could be a driver of the threshold

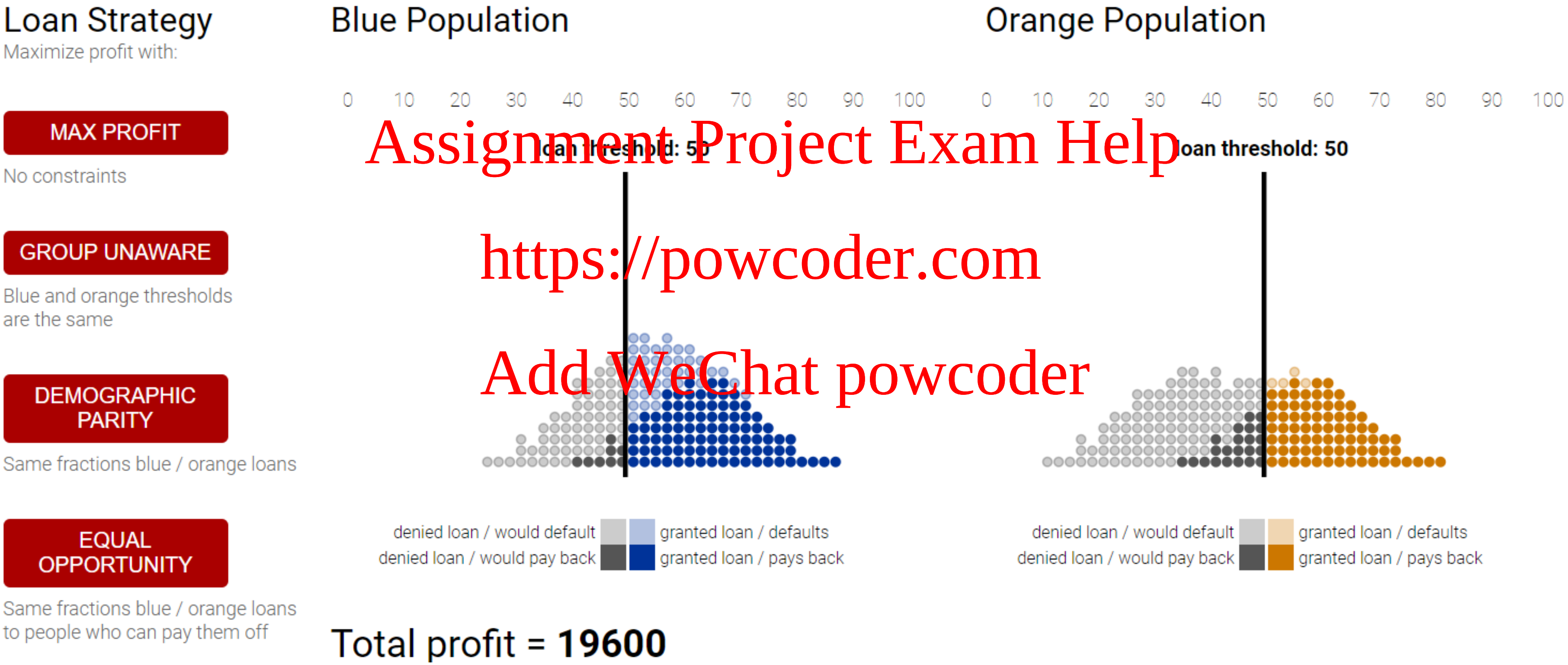
Maximum correctness @ threshold 50

Maximum profit @ threshold 54

The statistic behind a score may distribute differently across various groups

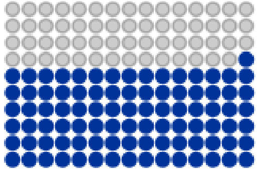
- The issue of how the correct decision is defined and with sensitivities to which factors, becomes particularly thorny when a statistic like a credit score ends up distributed differently between two groups
- Imagine we have two groups of people: blue and orange
- We are interested in making small loans, subject to the following rules
  - A successful loan makes $300
  - An unsuccessful loan costs $700
  - Everyone has a credit score between 0 and 100

The two distributions are slightly different, even though blue and orange people are equally likely to pay off a loan

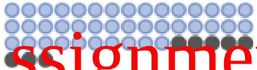


Total profit = **19600**

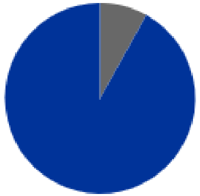
Correct 76%
loans granted to paying applicants and denied to defaulters



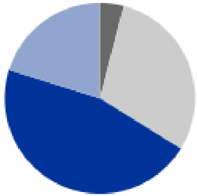
Incorrect 24%
loans denied to paying applicants and granted to defaulters



True Positive Rate 92%
percentage of paying applications getting loans

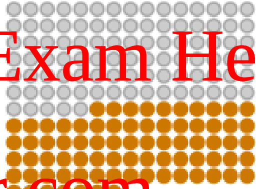


Positive Rate 66%
percentage of all applications getting loans

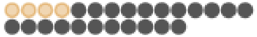


Profit: **-700**

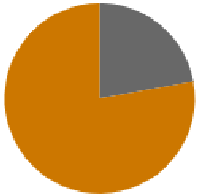
Correct 87%
loans granted to paying applicants and denied to defaulters



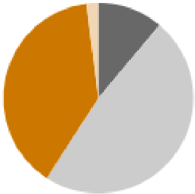
Incorrect 13%
loans denied to paying applicants and granted to defaulters



True Positive Rate 78%
percentage of paying applications getting loans

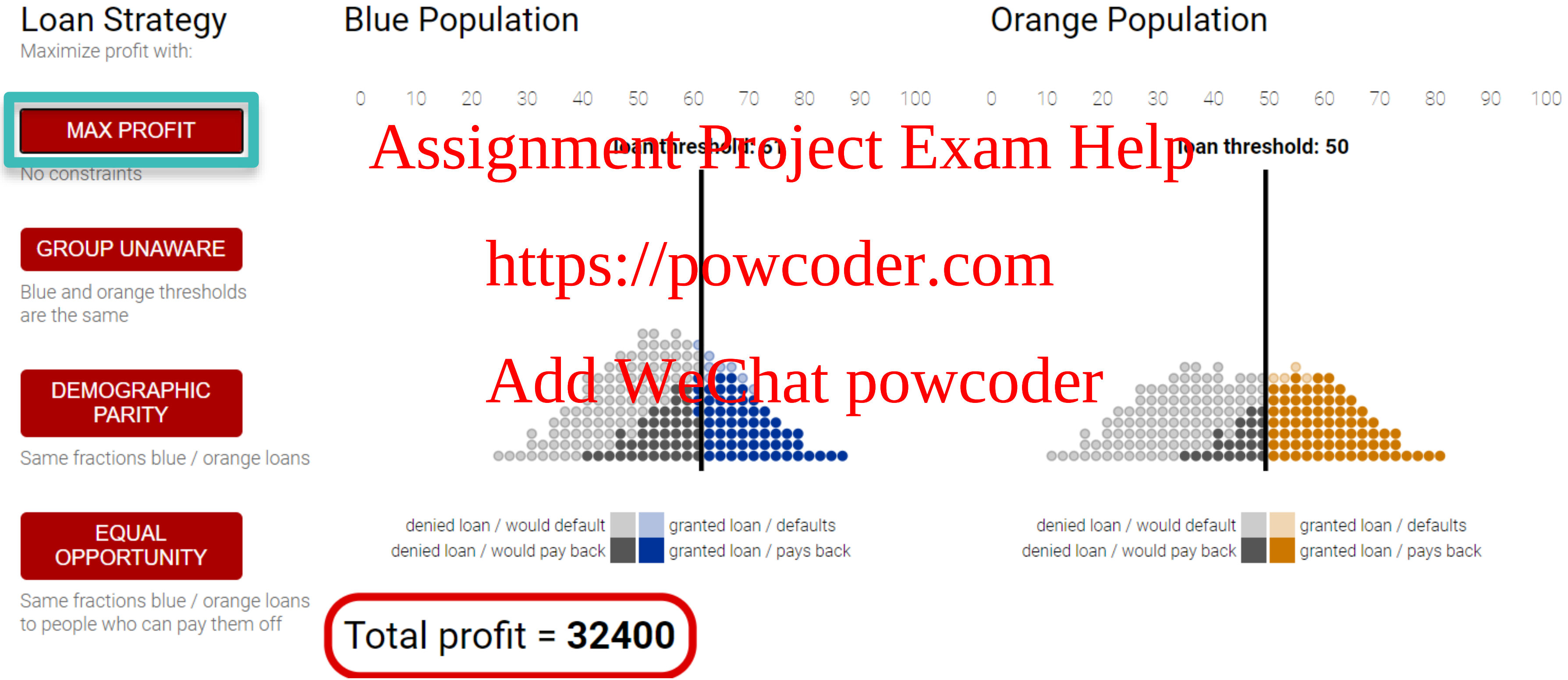


Positive Rate 41%
percentage of all applications getting loans

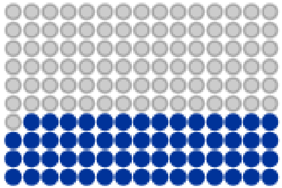


Profit: **20300**

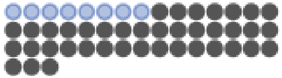
To maximize profit, the two groups have different thresholds, meaning they are held to different standards



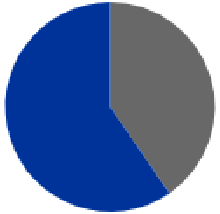
Correct 76%
loans granted to paying applicants and denied to defaulters



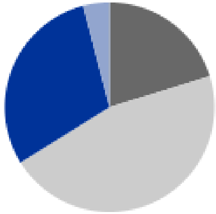
Incorrect 24%
loans denied to paying applicants and granted to defaulters



True Positive Rate 60%
percentage of paying applications getting loans

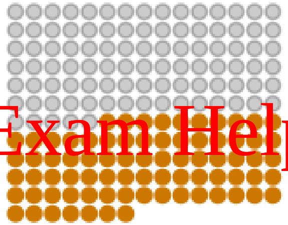


Positive Rate 34%
percentage of all applications getting loans



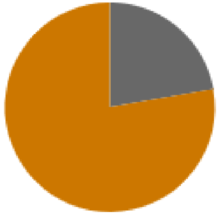
Profit: **12100**

Correct 87%
loans granted to paying applicants and denied to defaulters

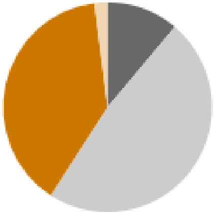


Incorrect 13%
loans denied to paying applicants and granted to defaulters

True Positive Rate 78%
percentage of paying applications getting loans



Positive Rate 41%
percentage of all applications getting loans



Profit: **20300**

Same threshold but orange has fewer loans overall. Among paying applicants, orange is also at a disadvantage.

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

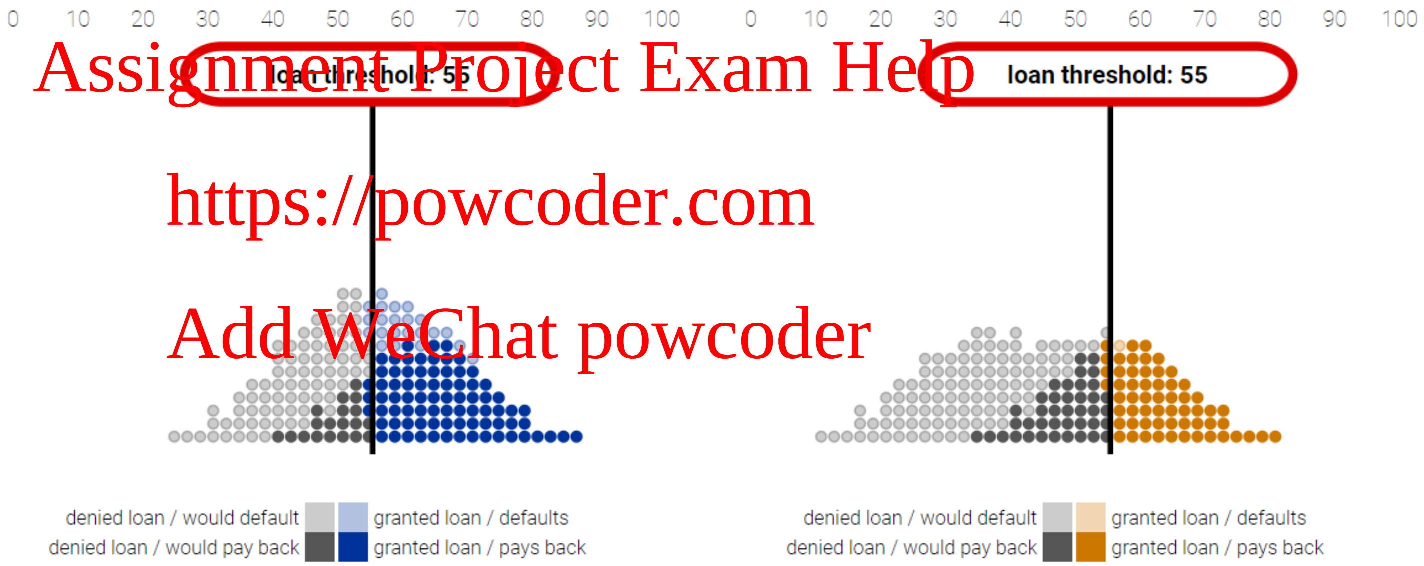
Same fractions blue / orange loans

EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

Blue Population

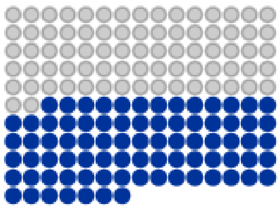
Orange Population

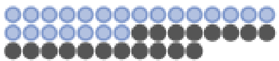


Total profit = **25600**

Correct 79%
loans granted to paying applicants and denied to defaulters

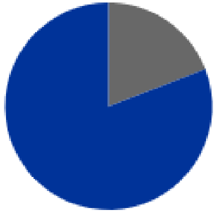
Incorrect 21%
loans denied to paying applicants and granted to defaulters

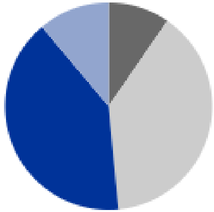




True Positive Rate 81%
percentage of paying applications getting loans

Positive Rate 52%
percentage of all applications getting loans

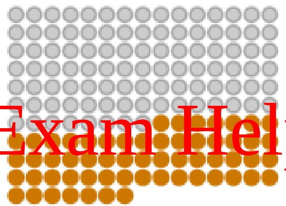


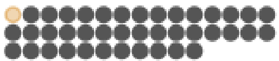


Profit: **8600**

Correct 79%
loans granted to paying applicants and denied to defaulters

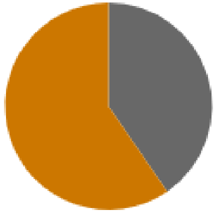
Incorrect 21%
loans denied to paying applicants and granted to defaulters

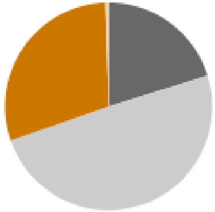




True Positive Rate 60%
percentage of paying applications getting loans

Positive Rate 30%
percentage of all applications getting loans





Profit: **17000**

Same proportion of loans given to each group but among paying applicants, blue is at a disadvantage

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

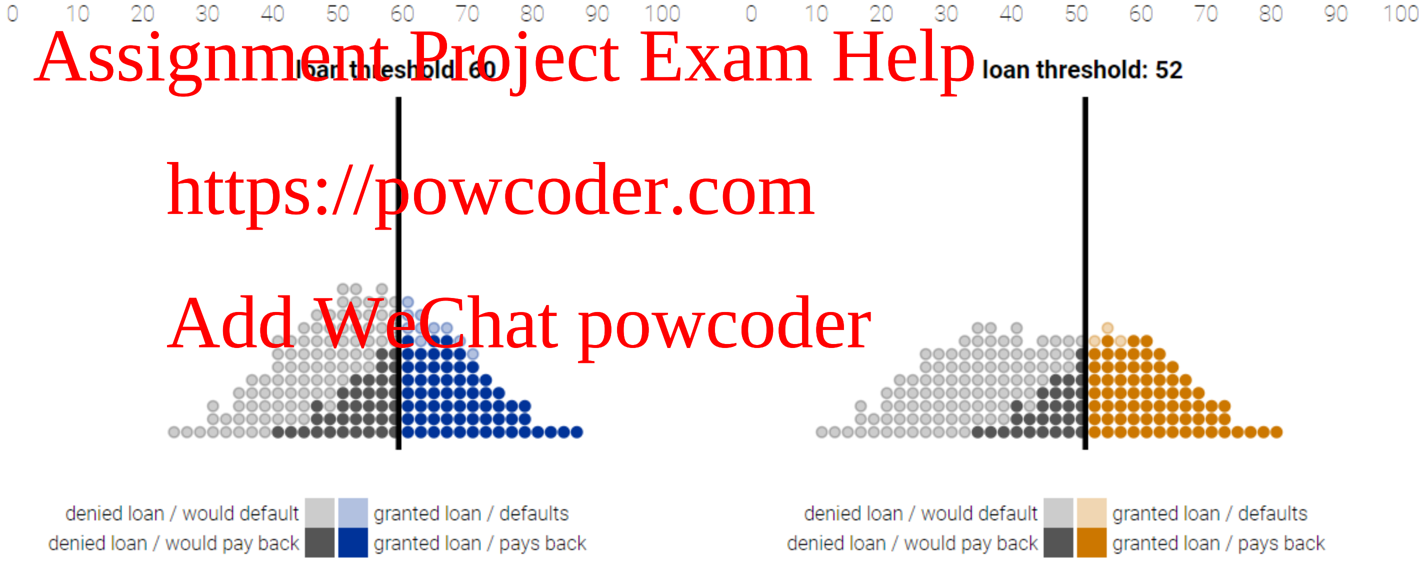
Same fractions blue / orange loans

EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

Blue Population

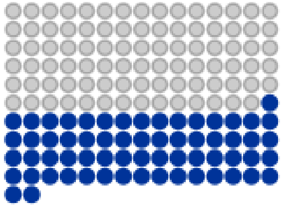
Orange Population

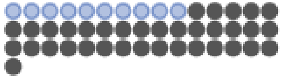


Total profit = **30800**

Correct 77%
loans granted to paying applicants and denied to defaulters

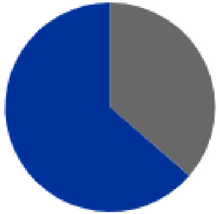
Incorrect 23%
loans denied to paying applicants and granted to defaulters

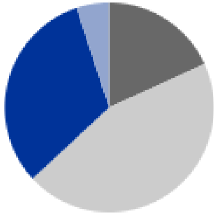




True Positive Rate 64%
percentage of paying applications getting loans

Positive Rate 37%
percentage of all applications getting loans





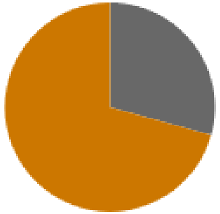
Profit: **11900**

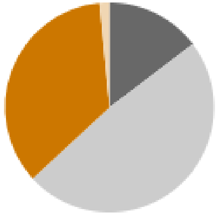
Correct 84%
loans granted to paying applicants and denied to defaulters

Incorrect 16%
loans denied to paying applicants and granted to defaulters

True Positive Rate 71%
percentage of paying applications getting loans

Positive Rate 37%
percentage of all applications getting loans





Profit: **18900**

Same proportion of loans to paying participants for each group, similar profit & grants as demographic parity

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

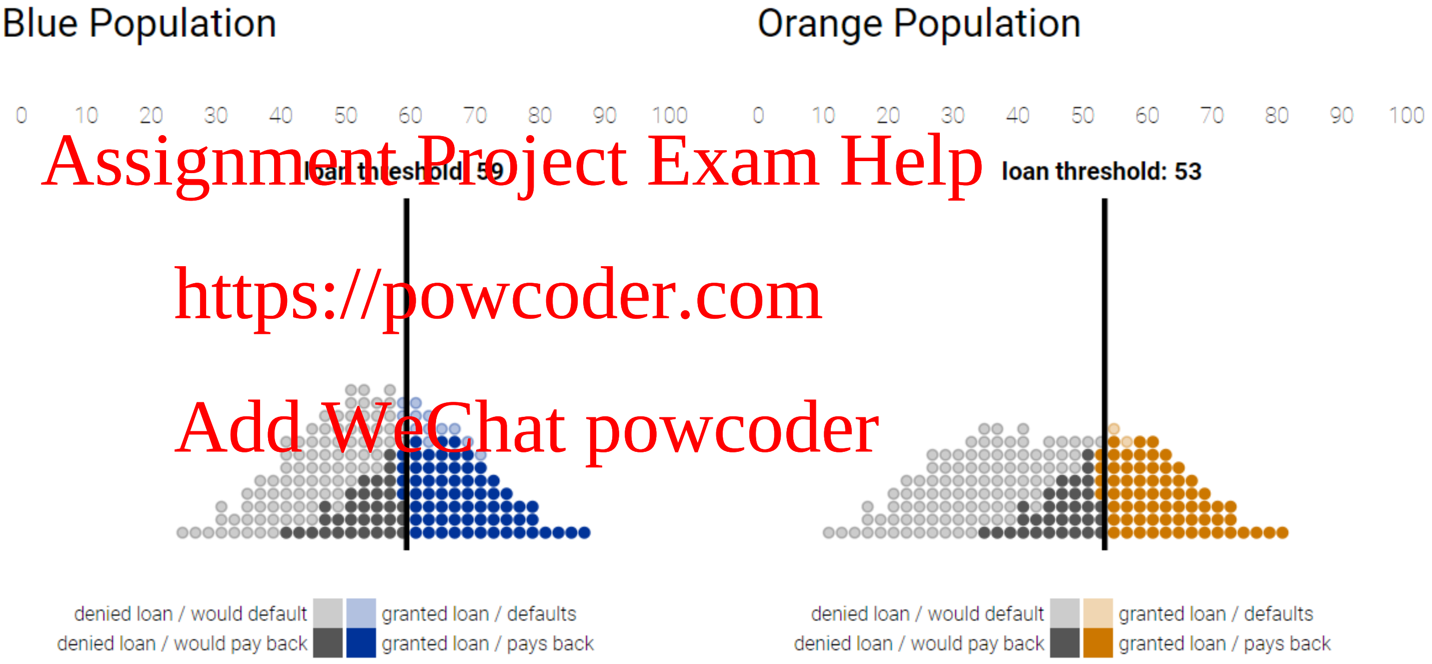
Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

Same fractions blue / orange loans

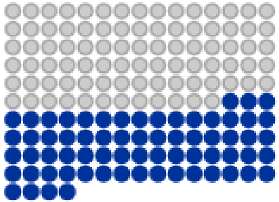
EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

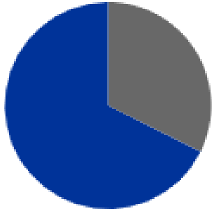


Total profit = **30400**

Correct 78%
loans granted to paying applicants and denied to defaulters

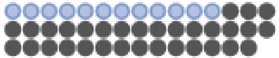


True Positive Rate 68%
percentage of paying applications getting loans

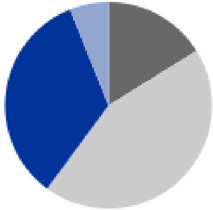


Profit: **11700**

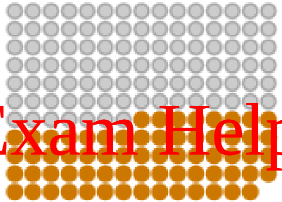
Incorrect 22%
loans denied to paying applicants and granted to defaulters



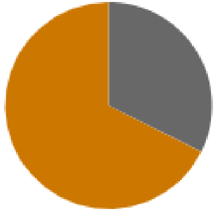
Positive Rate 40%
percentage of all applications getting loans



Correct 83%
loans granted to paying applicants and denied to defaulters



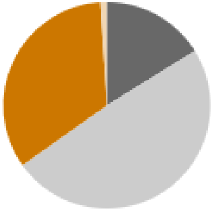
True Positive Rate 68%
percentage of paying applications getting loans



Profit: **18700**

Incorrect 17%
loans denied to paying applicants and granted to defaulters

Positive Rate 35%
percentage of all applications getting loans



Group Unaware (一視同仁)

- Fairness through unawareness
  - The group attribute is not used in the classification
- All groups to the same & one standard
- Ignore real differences between groups
  - Women generally pay less for life insurance than men since they tend to live longer
  - Differences in score distributions causes the orange group gets fewer loans if the most profitable group-unaware threshold is used

Demographic Parity (群体均等)

- Aka statistical parity or group fairness
- Same fraction of each group will receive intervention
  - Same positive rate for each group
  - The bank uses different loan thresholds that yield the same fraction of loans to each group
- Similar individuals (having similar attribute values) but in different groups may be discriminated

Equal Opportunity (機會均等)

- Same chance for the positive ones in each group
  - True Positive Rate (TPR) is identical between groups
  - For people who can pay back a loan, the same fraction in each group should actually be granted a loan

Why does fairness matter?

Regardless of one's definition of fairness, everyone wants to be treated fairly

Ensuring fairness is a moral and ethical imperative

What is fairness anyway?

There are 20+ definitions of fairness

Some of the definitions are contradictory

The way fairness is defined impacts bias

Data + Math ≠ Objectivity

Given essentially any scoring system, it is possible to efficiently find thresholds that meet the criteria earlier

In other words, even if you don't have control over the underlying scoring system (a common case) it is still possible to attack the issue of discrimination

Source of Bias

Most bias come from data used in classification

bias induced by sensitive information

direct bias

indirect bias

protected attribute

non-protected attributes

target

Gender	Spending Power	…	Number of Accidents	ZIP Code	Premium
M	8	…	5	06230	?
F	8	…	3	92100	?
F	10	…	0	75300	?
F	8	…	0	13500	?
M	12	…	2	45000	?
…	…	…		…	…

representation bias

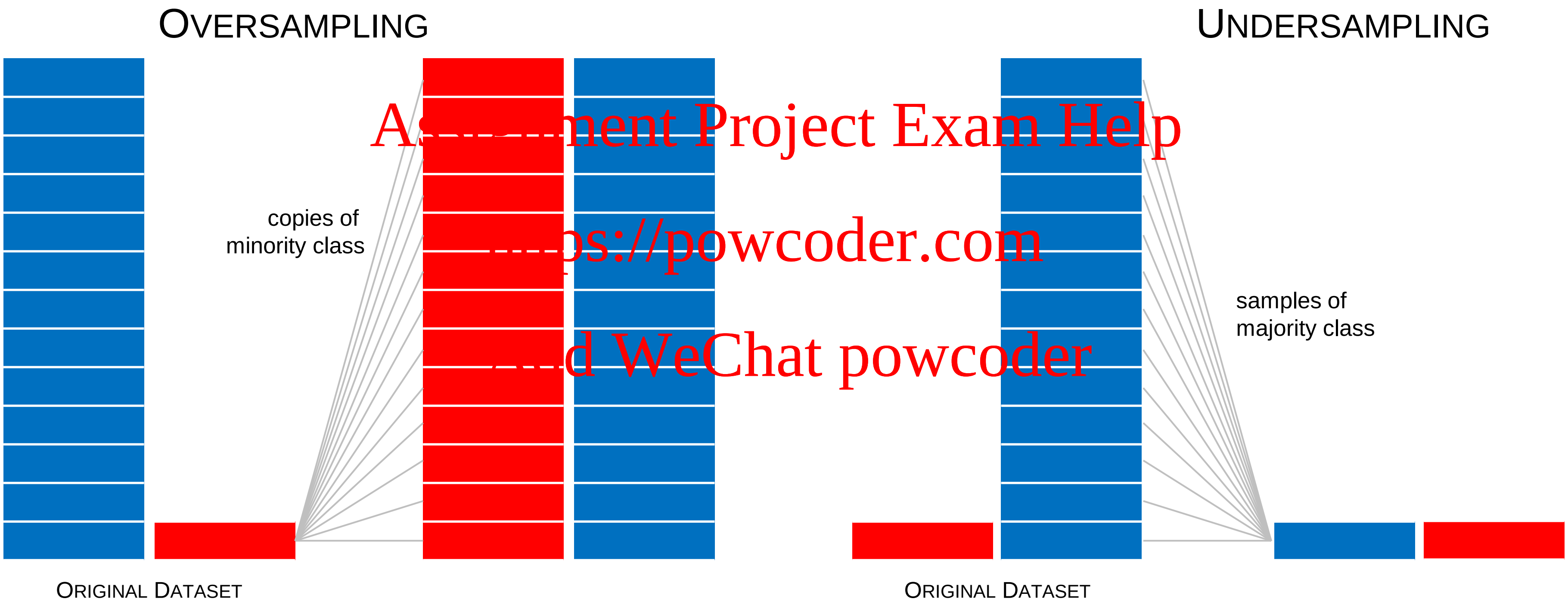
sample bias

label bias

model bias


Human Biases in Data
Reporting bias
Selection bias
Overgeneralization
Out-group homogeneity bias
Stereotypical bias
Historical unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group attribution error
Halo effect
Training data are collected and annotated
Human Biases in Collection and Annotation
Sampling error
Non-sampling error
Insensitivity to sample size
Correspondence bias
In-group bias
Bias blind spot
Confirmation bias
Subjective validation
Experimenter's bias
Choice-supportive bias
Neglect of probability
Anecdotal fallacy
Illusion of validity

Bias can be induced from sample representation



TOP SOCIAL MESSENGERS AROUND THE WORLD

THE MOST POPULAR MESSENGER APP BY COUNTRY / TERRITORY IN DECEMBER 2018

Bias can be induced from sample collection, e.g. using biased sources

SOURCE: BASED ON SIMILARWEB'S ALGORITHM INTEGRATING CURRENT INSTALLS FROM THE GOOGLE PLAY STORE WITH ACTIVE APP USERS (DECEMBER 2018).
NOTE: FIGURES IN PARENTHESES IN THE LEGEND REPRESENT THE NUMBER OF COUNTRIES / TERRITORIES IN WHICH EACH PLATFORM IS THE TOP-RANKED MESSENGER APP.


MIT Study of Top Face Recognition Services
99% accurate
for lighter-skinned males
65% accurate
for darker-skinned
females

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

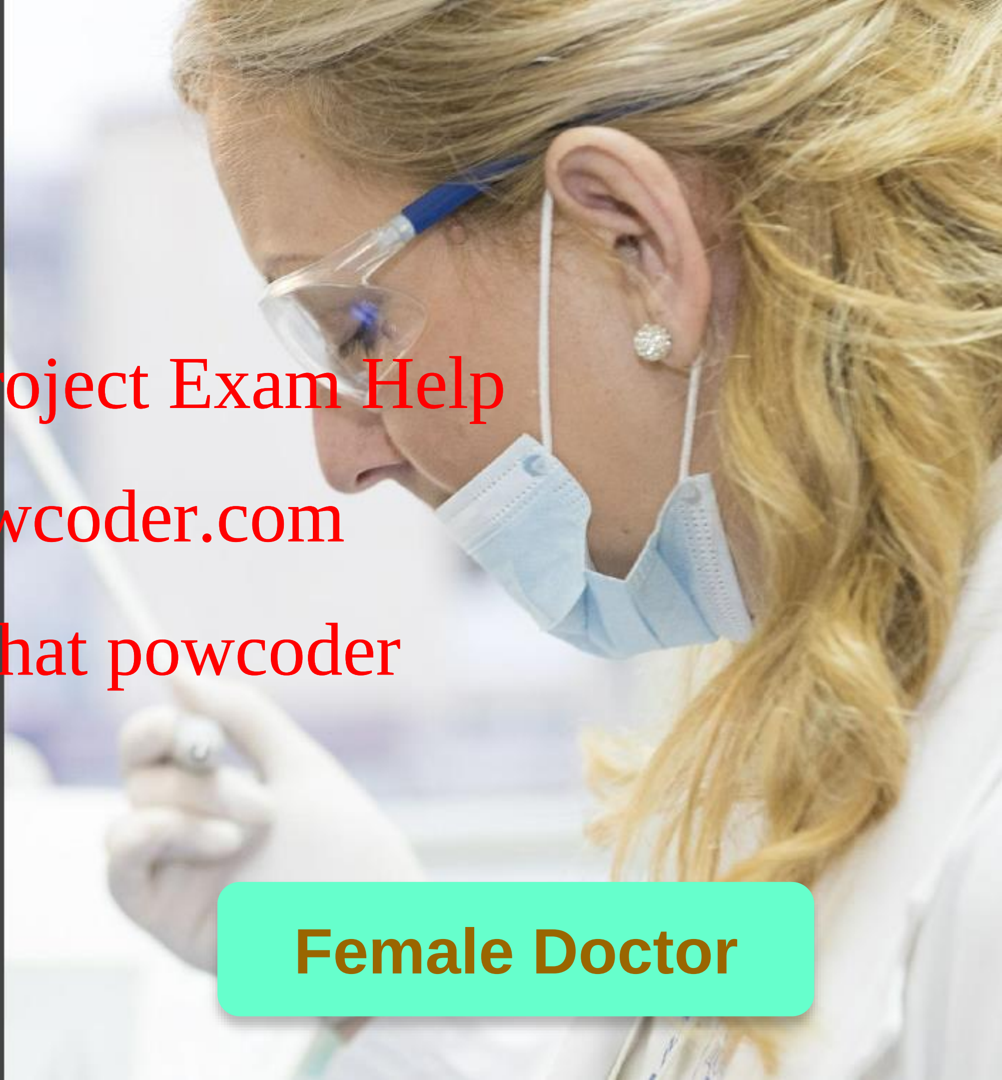
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?



World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

Fairness Terminology

Protected Attributes

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

Privileged Protected Attributes

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

Group Fairness

Groups defined by protected attributes receiving similar treatments or outcomes.

Individual Fairness

Similar individuals receiving similar treatments or outcomes.

Fairness Metric

A measure of unwanted bias in training data or models.

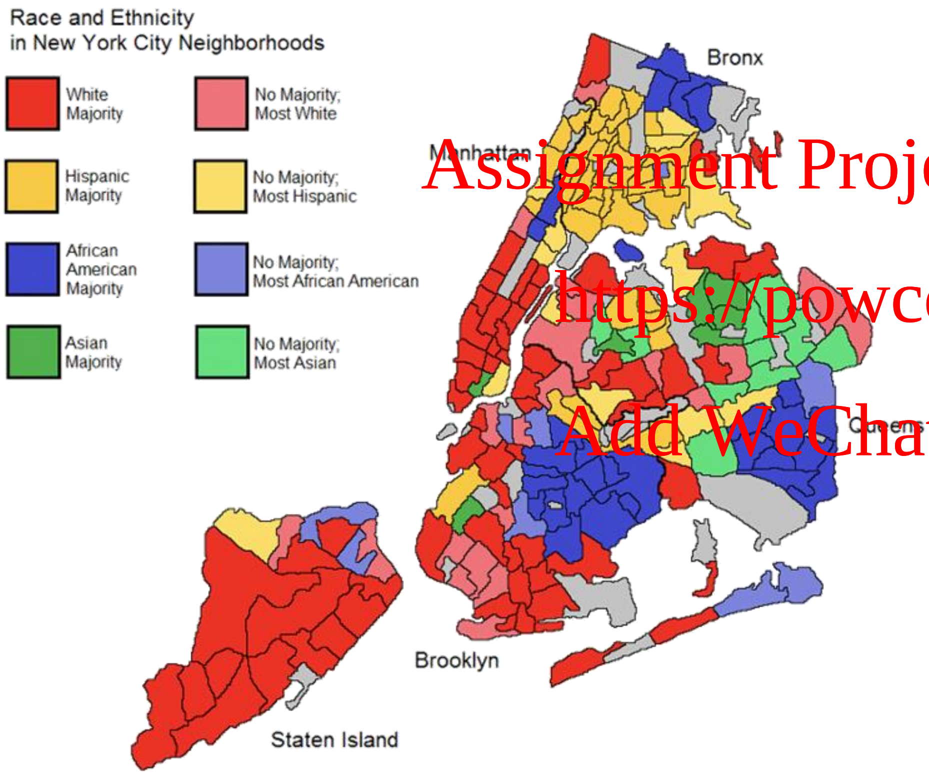
Favorable Label

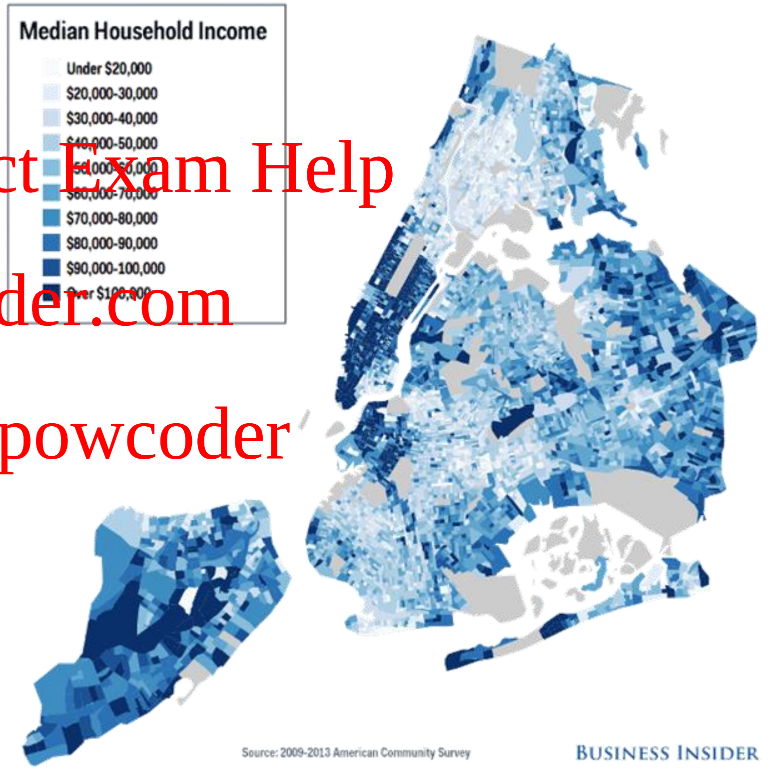
A label whose value correspond to an outcome that provides an advantage to the recipient.

Removing the protected attributes may not be sufficient due to the problem of proxies

- A common concern with AI models is that they may create proxies for protected attributes, where the complexity of the model leads to class membership being used to make decisions in a way that cannot easily be found and improved
- If the attributes used in the model have a strong relationship with the protected attributes, spurious correlation or poor model building could lead to a proxy problem
- Measuring within-class disparities (differences in treatment that only occurs for some members of a class) is much harder

Median household income could be a proxy of race





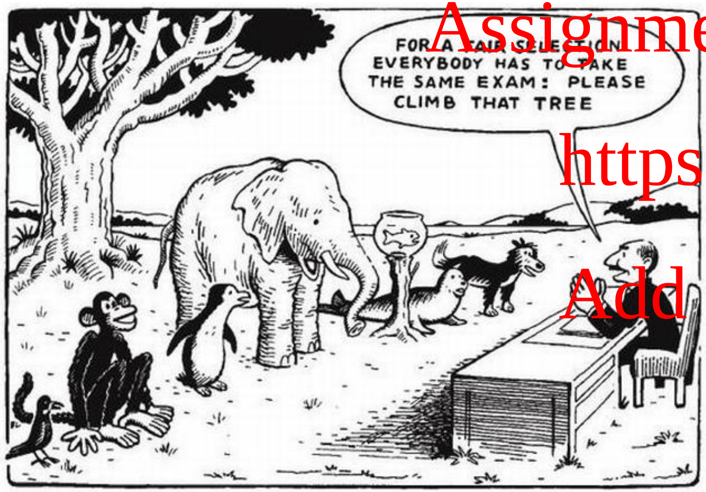
Most of the research on the topic of bias and fairness in AI is about making sure that your system does not have a disproportionate effect on some group of users relative to other groups.

The primary focus of AI ethics is on ***distribution checks*** *and similar analytics.*

Aequitas

Discrimination & Bias Audit Toolkit

Interest in algorithmic fairness and bias has been growing recently



- Machine Learning based predictive tools are being increasingly used in problems that can have a drastic impact on people's lives
  - e.g., criminal justice, education, public health, workforce development, and social services
- Recent work has raised concerns on the risk of unintended bias in these models, affecting individuals from certain groups unfairly
- While a lot of bias metrics and fairness definitions have been proposed, there is no consensus on which definitions and metrics should be used in practice to evaluate and audit these systems

Aequitas audits the predictions of ML-based risk assessment tools to understand different types of biases

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

- Aequitas is used to evaluate model performance across several bias and fairness metrics, and utilize the most relevant metrics in model selection.

- Aequitas will help:

  - Understand where biases exist in the model(s)
  - Compare the level of bias between groups in the samples (bias disparity)
  - Visualize absolute bias metrics and their related disparities for rapid comprehension and decision-making

Aequitas audits the evidence of disparate representation and disparate errors

- Aequitas can audit risk assessment systems for two types of biases
  - **Disparate representation**: biased actions / interventions that are not allocated in a way that is representative of the population
  - **Disparate errors**: biased outcomes through actions or interventions that are result of the system being wrong about certain groups of people
- To assess these biases, the following data are needed
  - Data about the overall population considered for interventions along with the protected attributes that are to be audited (e.g., race, gender, age, income)
  - The set of individuals in the target population that the risk assessment system recommended / selected for intervention or action
    - Unseen data, not the training dataset
  - To audit for biases due to disparate errors of the system, actual outcomes for the individuals who were selected and not selected are also required

Different bias and fairness criteria need to be used for different types of interventions

Equal Parity (均等)

Also known as Demographic Parity (人口平價 / 人口平价) or Statistical Parity (統計平價 / 统计平价)

Each group represented equally among the selected set

Proportional Parity

Also known as Impact Parity or Minimizing Disparate (完全不同) Impact

Each group represented proportional to their representation in the overall population

False Positive Parity

Desirable when interventions are punitive (懲罰性的 / 惩罚性的)

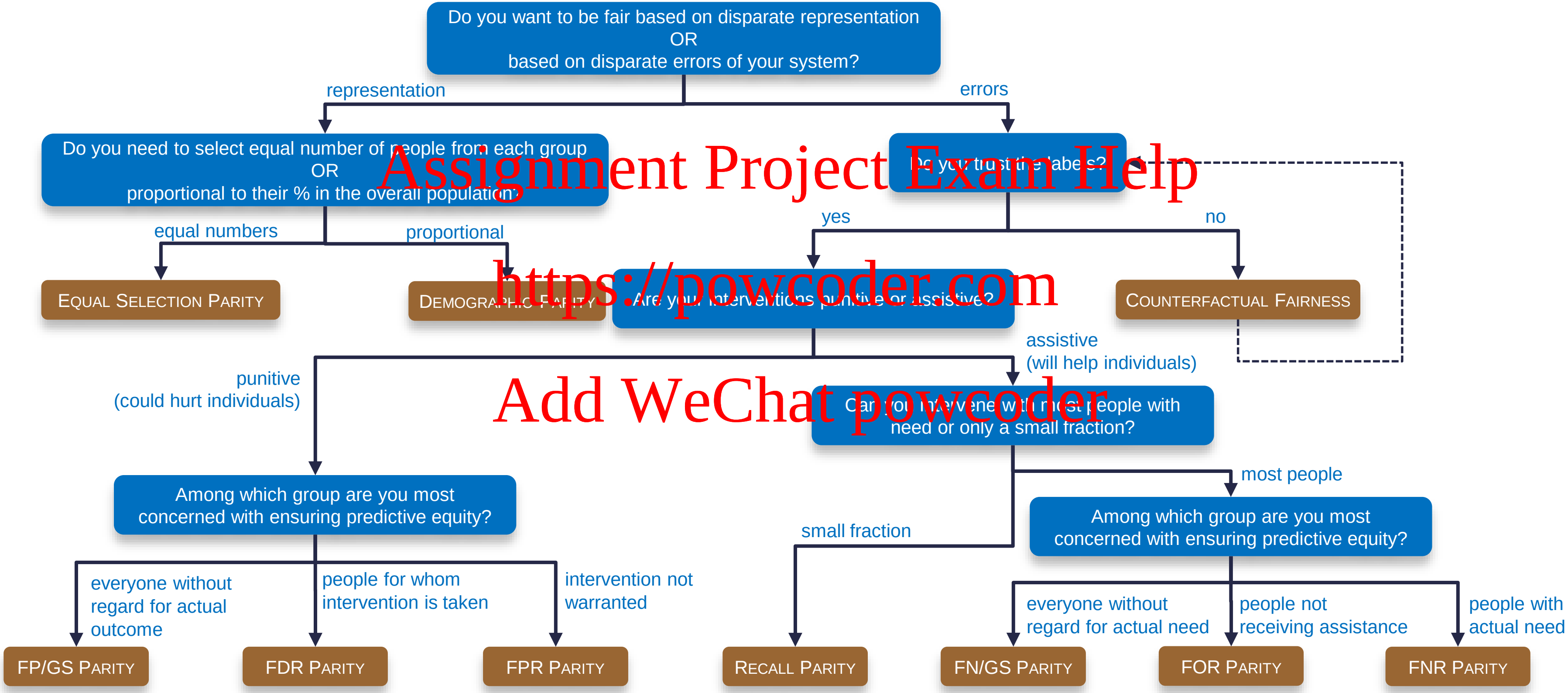
Each group to have equal False Positive Rates

False Negative Parity

Desirable when interventions are assistive / preventative

Each group to have equal False Negative Rates

The fairness tree describes the interpretation of the metrics



Preliminary Concepts

Name	Notation	Definition
Score	$S \in [0,1]$	A real-valued score assigned to each entity by the predictor.
Decision	$\hat{Y} \in \{0,1\}$	A binary-valued prediction assigned to an entity.
True Outcome	$Y \in \{0,1\}$	A binary-valued label (ground truth) of an entity.
Attribute	$A = \{a_1, a_2, \cdots, a_n\}$	A multi-valued attribute with multiple possible values, e.g., $gender = \{female, male, other\}$.
Group	$g(a_i)$	A group formed by all entities having the same attribute value, e.g., $race(Asian)$.
Reference Group	$g(a_r)$	A reference group formed by all entities having the reference attribute values, e.g., $gender(male)$.
Labelled Positive	LP_g	Number of entities within $g(a_i)$ with positive label, i.e., $Y = 1$.
Labelled Negative	LN_g	Number of entities within $g(a_i)$ with negative label, i.e., $Y = 0$.
Predicted Positive	PP_g	Number of entities within $g(a_i)$ with positive prediction, i.e., $\hat{Y} = 1$.
Total Predicted Positive	$K = \sum_{A=a_1}^{A=a_n} PP_{g(a_i)}$	Total number of entities with positive prediction across all groups $g(a_i)$ formed by all possible attribute values of A.
Predicted Negative	PN_g	Number of entities within $g(a_i)$ with negative prediction, i.e., $\hat{Y} = 0$.
False Positive	FP_g	Number of entities within $g(a_i)$ with false positive prediction, i.e., $\hat{Y} = 1 \wedge Y = 0$.
False Negative	FN_g	Number of entities within $g(a_i)$ with false negative prediction, i.e., $\hat{Y} = 0 \wedge Y = 1$.
True Positive	TP_g	Number of entities within $g(a_i)$ with true positive prediction, i.e., $\hat{Y} = 1 \wedge Y = 1$.
True Negative	TN_g	Number of entities within $g(a_i)$ with true negative prediction, i.e., $\hat{Y} = 0 \wedge Y = 0$.

Basic Metrics

Name	Notation	Definition & Example
Prevalence (Prev)	$Prev_g = \frac{LP_g}{\lvert g \rvert} = P(Y = 1 \mid A = a_i)$	Fraction of entities within $g(a_i)$ with positive label. *Given your race, what is your chance of being denied bail?*
Predicted Prevalence (PPrev)	$PPrev_g = \frac{PP_g}{\lvert g \rvert} = P(\hat{Y} = 1 \mid A = a_i)$	Fraction of entities within $g(a_i)$ with positive prediction. *Given your race, what is your predicted chance of being denied bail?*
Predicted Positive Rate (PPR)	$PPR_g = \frac{PP_g}{K} = P(A = a_i \mid \hat{Y} = 1)$	Ratio of number of entities within $g(a_i)$ with positive prediction to the Total Predicted Positive over all $g(a_i)$ formed by all possible attribute values of A. *Given the predicted denials of bail over all races, what is the chance of your race being denied bail?*
Recall / True Positive Rate (TPR)	$TPR_g = \frac{TP_g}{LP_g} = P(\hat{Y} = 1 \mid Y = 1, A = a_i)$	Fraction of entities within $g(a_i)$ with positive label that are also with positive prediction. *Among people with need, what is your chance of receiving assistance given your gender?*

Basic Metrics

Name	Notation	Definition & Example
False Negative Rate (FNR)	$FNR_g = \frac{FN_g}{LP_g} = P(\hat{Y} = 0 \mid Y = 1, A = a_i)$	Fraction of entities within $g(a_i)$ with positive label but have negative prediction. *Among people with need, what is your chance of not receiving any assistance given your gender?*
False Positive Rate (FPR)	$FPR_g = \frac{FP_g}{LN_g} = P(\hat{Y} = 1 \mid Y = 0, A = a_i)$	Fraction of entities within $g(a_i)$ with negative label but have positive prediction. *Among people who should be granted bail, what is your chance of being denied bail given your race?*
False Discovery Rate (FDR)	$FDR_g = \frac{FP_g}{PP_g} = P(Y = 0 \mid \hat{Y} = 1, A = a_i)$	Fraction of entities within $g(a_i)$ with positive prediction that are false. *Among people being denied bail, what is your chance of being innocent given your race?*
False Omission Rate (FOR)	$FOR_g = \frac{FN_g}{PN_g} = P(Y = 1 \mid \hat{Y} = 0, A = a_i)$	Fraction of entities within $g(a_i)$ with negative prediction that are false. *Among people who do not receive assistance, what is your chance of requiring assistance given your gender?*

Basic Metrics

Name	Notation	Definition & Example
False Positive over Group Size (FP/GS)	$FP/GS_g = \frac{FP_g}{\lvert g \rvert} = P(\hat{Y} = 1, Y = 0 \mid A = a_i)$	Ratio of number of entities within $g(a_i)$ with positive prediction that is wrong to the number of entities within $g(a_i)$ *What is your chance of being wrongly denied bail given your race?*
False Negative over Group Size (FN/GS)	$FN/GS_g = \frac{FN_g}{\lvert g \rvert} = P(\hat{Y} = 0, Y = 1 \mid A = a_i)$	Ratio of number of entities within $g(a_i)$ with negative prediction that is wrong to the number of entities within $g(a_i)$ *What is your chances of being wrongly left out of assistance given your race?*

Unfair Disparities in COMPAS

COMPAS was reported to have unfair disparities, Northpointe pushed back, who is right and who is wrong

- In 2016, Propublica reported on racial inequality in COMPAS, a risk assessment tool
- The algorithm was shown to lead to unfair disparities in False Negative Rates and False Positive Rates

- In the case of recidivation (累犯), it was shown that black defendants faced disproportionately high risk scores, while white defendants received disproportionately low risk scores

- Northpointe, the company responsible for the algorithm, responded by arguing they calibrated the algorithm to be fair in terms of False Discovery Rate, also known as calibration
- The Bias Report provides metrics on each type of disparity, which add clarity to the bias auditing process

The COMPAS Dataset

COMPAS Recidivism Risk Assessment Dataset

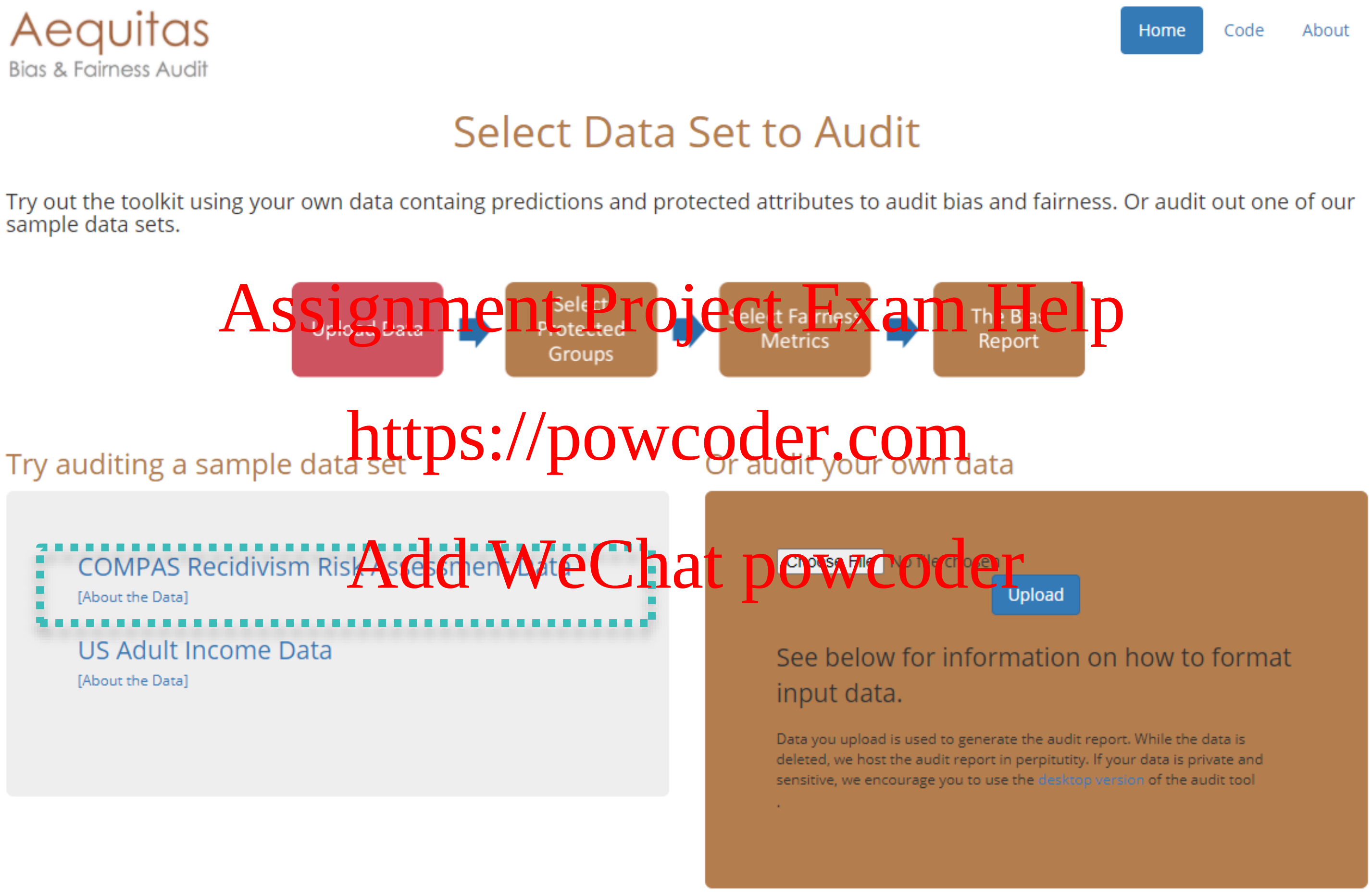
***score** is a binary assessment made by the predictive model and 1 denotes an individual selected for the intervention*

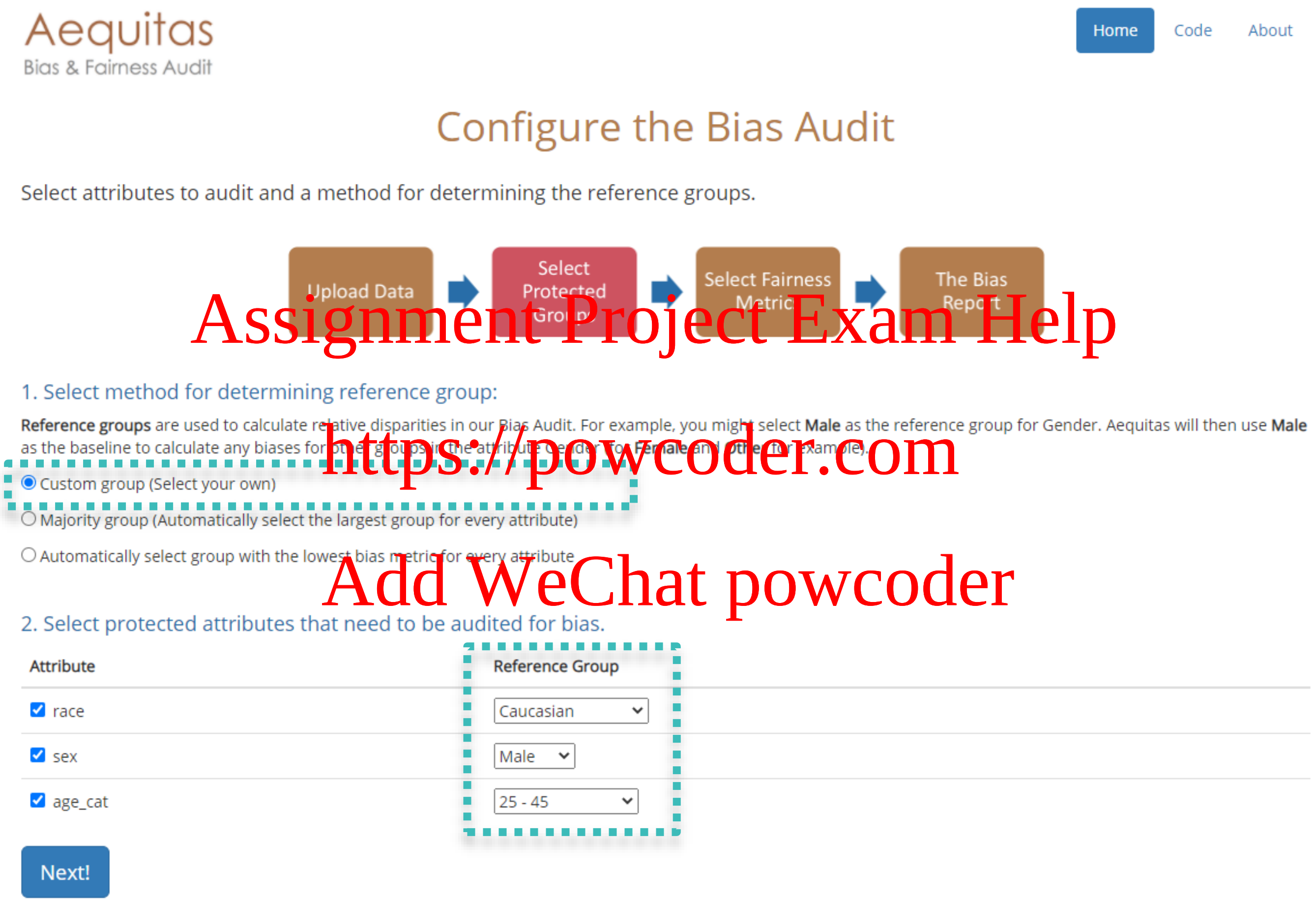
label_value is the binary valued ground truth and 1 denotes a biased case based on disparate errors

	entity_id	score	label_value	race	sex	age_cat
0	1	0.0	0	Other	Male	Greater than 45
1	3	0.0	1	African-American	Male	25 - 45
2	4	0.0	1	African-American	Male	Less than 25
3	5	1.0	0	African-American	Male	Less than 25
4	6	0.0	0	Other	Male	25 - 45

The Audit Process

http://aequitas.dssg.io/upload.html




Aequitas
Bias & Fairness Audit
Home
Code
About
Configure the Bias Audit
Select attributes to audit and a method for determining the reference groups.
Upload Data
Select Protected Groups
Select Fairness Metric
The Bias Report
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 (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
race
Caucasian
sex
Male
age_cat
25 - 45
Next!

Aequitas
Bias & Fairness Audit

Home Code About

Configure the Bias Audit

Configure the bias and fairness audit by selecting the fairness measures to audit and the fairness threshold to determine when the audit passes or fails.

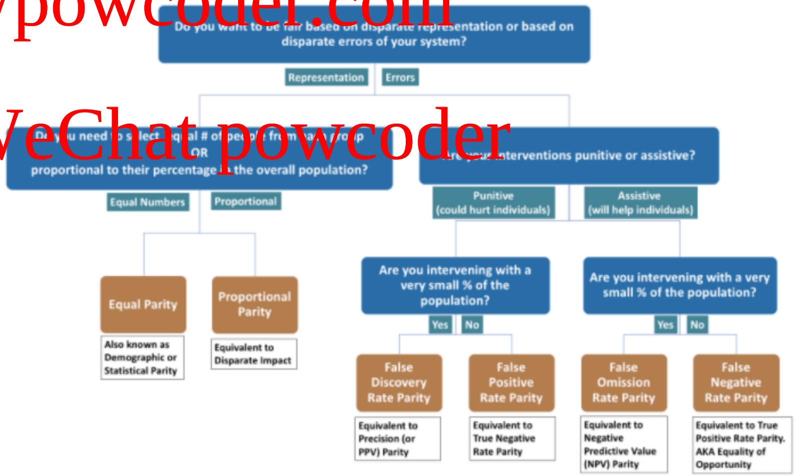
Upload Data → Select Protected Groups → Select Fairness Metrics → The Bias Report

$$\text{threshold} \le \text{Fair Parity} \le \frac{1}{\text{threshold}}$$

3. Select Fairness Metrics to Compute:

- [x] Equal Parity
- [x] Proportional Parity
- [x] False Positive Rate Parity
- [x] False Discovery Rate Parity
- [x] False Negative Rate Parity
- [x] False Omission Rate Parity

FAIRNESS TREE



Fair: between 80~125% of the reference group metric value

Unfair: outside the 80~125% range of the reference group metric value

4. Enter your Disparity Intolerance (in %):

If a specific bias metric for a group is within this percentage of the reference group, this audit will pass

80 %

Generate Fairness Report

The Bias Report

The Bias Report

The Bias Report

Audit Date:	08 Mar 2021
Data Audited:	7214 rows
Attributes Audited:	race, sex, age_cat
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.

All groups in all attributes show disparity outside the 80~125% range hence failing the Equal Parity

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

For age_cat (with reference group as **25 - 45**)
- Greater than 45 with 0.20X Disparity
- Less than 25 with 0.52X Disparity

All groups in all attributes show disparity outside the 80~125% range hence failing the Proportional Parity

Proportional Parity: Failed

What is it?

This criteria considers an attribute to have proportional parity if every group is represented proportionally to their share of the population. For example, if race with possible values of white, black, other being 50%, 30%, 20% of the population respectively) has proportional parity, it implies that all three races are represented in the same proportions (50%, 30%, 20%) in the selected set.

When does it matter?

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

Which groups failed the audit:

For race (with reference group as **Caucasian**)
- Other with 0.60X Disparity
- African-American with 1.69X Disparity
- Native American with 1.92X Disparity
- Asian with 0.72X Disparity

For age_cat (with reference group as **25 - 45**)
- Greater than 45 with 0.53X Disparity
- Less than 25 with 1.40X Disparity

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

For age_cat (with reference group as **25 - 45**)
- Greater than 45 with 0.50X Disparity
- Less than 25 with 1.62X Disparity

False Discovery Rate Parity: Failed

What is it?

This criteria considers an attribute to have False Discovery Rate parity if every group has the same False Discovery Error Rate. For example, if race has false discovery parity, it implies that all three races have the same False Discvery 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 can hurt individuals and where you are selecting a very small group for interventions.

Which groups failed the audit:

For race (with reference group as **Caucasian**)
- Asian with 0.61X Disparity
- Native American with 0.61X Disparity

For sex (with reference group as **Male**)
- Female with 1.34X Disparity

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

For age_cat (with reference group as **25 - 45**)
- Greater than 45 with 1.53X Disparity
- Less than 25 with 0.70X Disparity

False Omission Rate Parity: Failed

What is it?

This criteria considers an attribute to have False Omission Rate parity if every group has the same False Omission Error Rate. For example, if race has false omission parity, it implies that all three races have the same False Omission 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 help social services for example) and missing an individual could lead to adverse outcomes for them , and where you are selecting a very small group for interventions. 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**)
- Asian with 0.43X Disparity
- Native American with 0.58X Disparity

For sex (with reference group as **Male**)
- Female with 0.73X Disparity

For age_cat (with reference group as **25 - 45**)
- Greater than 45 with 0.75X Disparity
- Less than 25 with 1.31X Disparity

The dataset failed

all fairness assessments

Metric values for each group is provided

race

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
African-American	0.51	0.66	0.59	0.37	0.45	0.35	0.28
Asian	0	0.0	0.25	0.25	0.09	0.12	0.33
Caucasian	0.34	0.26	0.35	0.41	0.23	0.29	0.48
Hispanic	0.09	0.06	0.3	0.46	0.21	0.29	0.56
Native American	0	0.0	0.67	0.25	0.38	0.17	0.1
Other	0.05	0.02	0.21	0.46	0.15	0.3	0.68

Only a few bias metric values fall within the 80~125% range

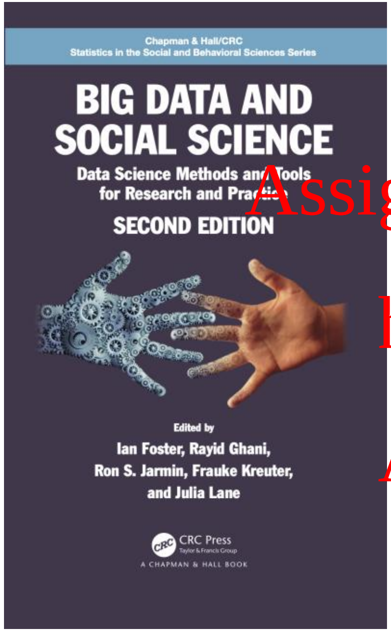
$$\text{FDR Disparity}_{Asian} = \frac{\text{FDR}_{Asian}}{\text{FDR}_{Caucasian}} = \frac{0.25}{0.41} = 0.61$$

race

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
African-American	2.55	1.69	0.91	1.91	1.21	0.59
Asian	0.01	0.72	0.61	0.37	0.43	0.7
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.22	0.86	1.12	0.92	1.0	1.17
Native American	0.01	1.92	0.61	1.6	0.58	0.21
Other	0.09	0.6	1.12	0.63	1.05	1.42

References

References



"Big Data and Social Science – Data Science Methods and Tools for Research and Practice", 2nd edition, Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane, Chapman and Hall/CRC, November 2020 (https://textbook.coleridgeinitiative.org/)

- Aequitas project website (http://www.datasciencepublicpolicy.org/projects/aequitas/)
- Aequitas GitHub page (https://dssg.github.io/aequitas/30_seconds_aequitas.html)
- "Dealing with Bias and Fairness in Data Science Systems", Pedro Saleiro et al, 2020 (https://www.youtube.com/watch?v=N67pE1AF5cM)
- Tutorial: Fairness in Decision-Making with AI: a Practical Guide & Hands-On Tutorial using Aequitas , YouTube, October 2019 (https://www.youtube.com/watch?v=yOR71zBm3Uc)
- "Chapter 11 Bias and Fairness" in "Big Data and Social Science", (https://textbook.coleridgeinitiative.org/)
- "Aequitas: a Bias and Fairness Audit Toolkit", Pedro Saleiro et al, 2019 (https://arxiv.org/pdf/1811.05577.pdf)

Bias Mitigation

What is modifiable determines what mitigation algorithms can be used

Pro-Processing Algorithm

Bias mitigation algorithms applied to training data

In-Processing Algorithm

Bias mitigation algorithms applied to a model during its training

In-Processing Algorithm

Bias mitigation algorithms applied to predicted labels

Bias mitigation can be applied at different phases of the machine learning pipeline

Pre-processing Algorithms
Mitigates Bias in **Training Data**

Reweighing
Modifies the weights of different training examples

Disparate Impact Remover
Edits feature values to improve group fairness

Optimized Preprocessing
Modifies training data features & labels

Learning Fair Representation
Learns fair representation by obfuscating information about protected attributes

In-processing Algorithms
Mitigates Bias in **Classifiers**

Adversarial Debiasing
Uses adversarial techniques to maximise accuracy & reduce evidence of protected attributes in prediction

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
Mitigates Bias in **Prediction**

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

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

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

Ethical Machine Learning

A fully autonomous car is transporting a human being (A) to its desired destination. Suddenly, in a twist of fate, some living being (B) appears on the road. The AI (i.e., the computer) that controls the vehicle (i.e., the machine) must come to a decision within a fraction of a second: take evasive action or continue straight ahead. If it does try to dodge B, the vehicle skids and hits a tree, A dies, and B survives. If not, A survives, but B dies. For simplification purposes, we shall assume that collateral damage is negligible or identical in both cases.

Deon

Data Science Ethics Checklist

deon

- To facilitate data scientists to practice data ethics
- To evaluate considerations related to advanced analytics and machine learning applications from data collection through deployment
- To ensure that risks inherent to AI-empowered technology do not escalate into threats to an organization's constituents, reputation, or society more broadly
- To provide concrete, actionable reminders to the developers that have influence over how data science gets done
- A lightweight, open-source command line tool that facilitates integration into ML workflows

Bias mitigation can be applied at different phases of the machine learning pipeline

Data Collection	Data Storage	Analysis	Modeling	Deployment
Informed Consent	Data Security	Missing Perspective	Proxy Discrimination	Redress
Collection Bias	Right to be Forgotten	Dataset Bias	Fairness Across Groups	Roll back
Limiting PII Exposure	Data Retention Plan	Honest Representation	Metric Selection	Concept Drift
Downstream Bias Mitigation		Privacy in Analysis	Explainability	Unintended Use
		Auditability	Communicating Bias	

Data Collection Checklist

A. Data Collection

A.1	**INFORMED CONSENT**	If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
A.2	**COLLECTION BIAS**	Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
A.3	**LIMIT PII EXPOSURE**	Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?
A.4	**DOWNSTREAM BIAS MITIGATION**	Have we considered ways to enable testing downstream results for biased outcomes (e.g., collecting data on protected group status like race or gender)?

Facebook uses phone numbers provided for two-factor authentication to target users with ads

https://techcrunch.com/2018/09/27/yes-facebook-is-using-your-2fa-phone-number-to-target-you-with-ads/

- FB confirmed it in fact used phone numbers that users had provided it for security purposes to also target them with ads
- Specially a phone number handed over for two factor authentication (2FA)
  - SFA is a security technique that adds a second layer of authentication to help keep accounts secure

Low smartphone penetration areas contribute less to big data and consequently become digitally invisible

https://hbr.org/2013/04/the-hidden-biases-in-big-data

- Data fundamentalism (數據原教旨主義) is the notion that correlation always indicates causation and that massive data sets & predictive analytics always reflect objective truth
- Datasets are not objective, they are creations of human design
- We give numbers their voice, draw inferences from them, and define their meaning through our interpretation
- Hidden biases in both the collection and analysis stages present considerable risks
- Biases are as important to the big-data equation as the numbers themselves

Low smartphone penetration areas contribute less to big data and consequently become digitally invisible

https://hbr.org/2013/04/the-hidden-biases-in-big-data

- The greatest number of tweets (20 millions) were generated from Manhattan around the strike of Hurricane Sandy, creating the illusion that Manhattan was the hub of the disaster
- Very few messages originated from more severely affected locations, such as Breezy Point, Coney Island, and Rockaway
- As extended power blackouts drained batteries and limited cellular access, even fewer tweets came from the worst hit areas
  - In fact, there was much more going on outside the privileged, urban experience of Sandy that Twitter data failed to convey, especially in aggregate
- Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities

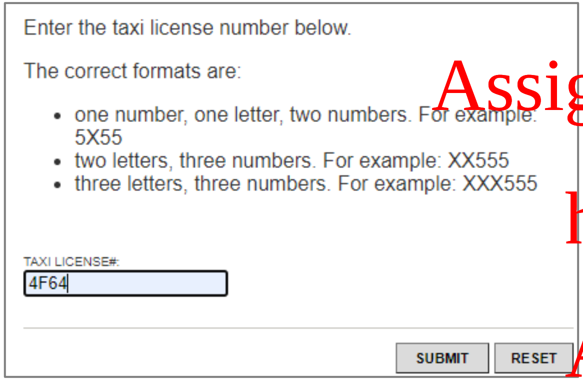
Personal information on taxi drivers can be accessed in poorly anonymized taxi trips dataset of New York City

https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

- New York City released data of 173M individual taxi trips including information like time & location of the pickup & drop off as well as an anonymised licence number and medallion number
- Even though no PII (Personal Identifiable Information) was included in the data, de-anonymising data to reveal personal identity was then found to be trivial
- The anonymised licence number and medallion number had not been anonymised properly and trivial to undo with other publicly available data

Personal information on taxi drivers can be accessed in poorly anonymized taxi trips dataset of New York City

https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

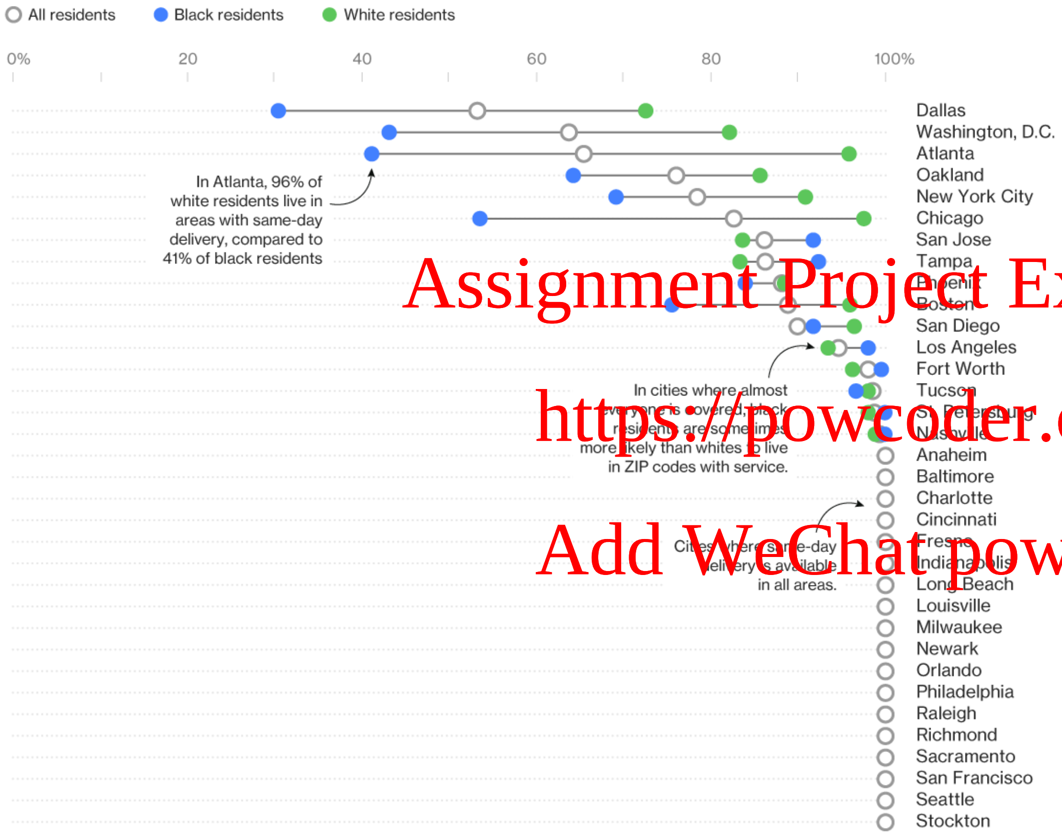


- 3 medallion number formats giving 22M possibilities
  - 9X99, XX999, XXX999 (some numbers are skipped)
- 2 licence number formats giving 2M possibilities
  - 5XXXXX, 5XXXXXX
- Both numbers were anonymised by hashing using MD5
- With the number of possibilities down to 24M, it was a matter of only minutes to determine which number was associated with a hash
- The entire dataset was de-anonymised within one hour

Personal information on taxi drivers can be accessed in poorly anonymized taxi trips dataset of New York City

https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

- There was ton of resources on NYC Taxi and Limousine Commission including a mapping from licence number to driver name and a way to look up owners of medallions
- This anonymisation was so poor that anyone could, with less than two hours work, figure out which driver drove every single trip in this entire dataset, or calculate drivers' gross income or infer where they live
- NYC could have simply assigned random numbers to each licence plate making it much more difficult to work backwards



Source: Bloomberg analyis of data from Amazon.com and the American Community Survey

In six major cities, Amazon's same day delivery service excludes many predominantly black neighborhoods

https://www.bloomberg.com/graphics/2016-amazon-same-day/

Data Storage Checklist

B. Storage Collection

B.1	**DATA SECURITY**	Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
B.2	**RIGHT TO BE FORGOTTEN**	Do we have a mechanism through which an individual can request their personal information be removed?
B.3	**DATA RETENTION PLAN**	Is there a schedule or plan to delete the data after it is no longer needed?

Equifax revealed the exact scope of the massive breach that exposed sensitive data about millions of Americans

https://www.nbcnews.com/news/us-news/equifax-breaks-down-just-how-bad-last-year-s-data-n872496

Data Element Stolen	Data Analyzed	Approximate Number of Impacted US Consumers
Name	First Name, Last Name, Middle Name, Suffix, Full Name	146.6 million
Data of Birth	D.O.B.	146.6 million
Social Security Number	SSN	145.5 million
Address Information	Address, Address2, City, State, Zip	99 million
Gender	Gender	27.3 million
Phone Number	Phone, Phone2	20.3 million
Driver Licence Number	DL	17.6 million
Email Address	Email Address	1.8 million
Payment Card Number and Expiration Date	CC Number, Exp Date	209,000
Tax ID	TaxID	97,500
Driver's License State	DL License State	27,000

- Equifax is one of US's biggest credit reporting agencies.
- In a filing with the SEC in 2018, Equiifax acknowledged that 145.5M Social Security Numbers were compromised, more than 200,000 credit card numbers and expiration dates were also collected, as well as government-issued identification documents (e.g., driver's licenses, taxpayer ID cards, passports and military IDs) that about 182,000 consumers uploaded when they disputed credit reports with Equifax

There is no express right to be forgotten under Hong Kong's PDPO (Personal Data Privacy Ordinance)

https://www.linklaters.com/en/insights/data-protected/data-protected---hong-kong

- The PDPO only includes a general obligation on a data user to take all practicable steps to erase personal data held by it where the data is no longer required for the purpose for which the data was used (unless such erasure is prohibited under any law or it is in the public interest, including historical interest, for the data not to be erased)
- In the banking context, however: (i) the Privacy Commissioner has published a specific code of practice on consumer credit data such that a credit provider must inform data subjects that they have the right to instruct the credit provider to make a request to a credit reference agency to delete account data relating to a terminated account and (ii) the Code of Banking Practice published by the Hong Kong Association of Banks requires institution to have in place appropriate control and protection mechanism that acknowledge the rights of customers to obtain prompt correction and/or deletion of inaccurate, or unlawfully collected or processed data
- The Privacy Commissioner has the power, by way of issuing an enforcement notice, to request a data user to remove personal data if the use of the personal data contravenes the PDPO
- This power has been exercised by the Privacy Commissioner in the past and was upheld on a legal challenge against the Privacy Commissioner's decision

Unsecured server exposed thousands of FedEx customer records

https://www.zdnet.com/article/unsecured-server-exposes-fedex-customer-records/

- FedEx was reported in 2018 to have exposed customer private information after a legacy server was left open without a password

- The server contained more than 112,000 unencrypted files
  - Completed US Postal Service forms (including names, home addresses, phone numbers, and handwritten signatures) used to authorize the handling of mail, drivers' licenses, national ID cards, and work ID cards, voting cards, utility bills, resumes, vehicle registration forms, medical insurance cards, firearms licenses, a few US military identification cards, and even a handful of credit cards that customers used to verify their identity with the FedEx division

- Despite the division's shutdown in 2015, some documents remained valid and the breach put customers at risk of identity theft

Analysis Checklist

C. Analysis

C.1	**MISSING PERSPECTIVES**	Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
C.2	**DATASET BIAS**	Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?
C.3	**HONEST REPRESENTATION**	Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?
C.4	**PRIVACY IN ANALYSIS**	Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
C.5	**AUDITABILITY**	Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

Modeling Checklist

D. Modelling

D.1	**Proxy Discrimination**	Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
D.2	**Fairness Across Groups**	Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
D.3	**Metric Selection**	Have we considered the effects of optimizing for our defined metrics and considered additional metrics?
D.4	**Explainability**	Can we explain in understandable terms a decision the model made in cases where a justification is needed?
D.5	**Communicate Bias**	Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?

Deployment Checklist

E. Deployment

E.1	**REDRESS**	Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?
E.2	**ROLL BACK**	Is there a way to turn off or roll back the model in production if necessary?
E.3	**CONCEPT DRIFT**	Do we test and monitor for concept drift to ensure the model remains fair over time?
E.4	**UNINTENDED USE**	Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed?

References

References

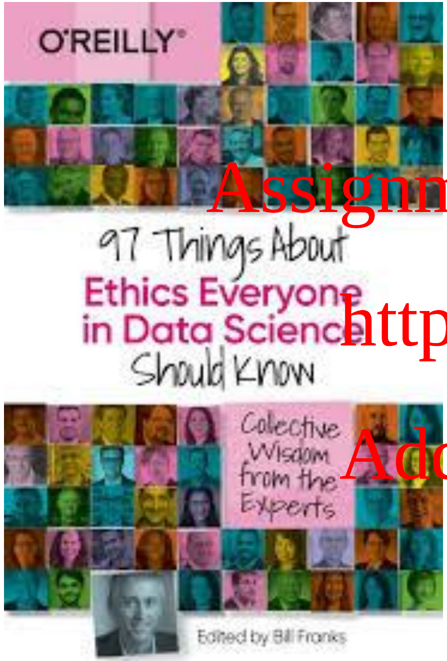
- An Ethics Checklist for Data Scientists (https://deon.drivendata.org/)
- Deon Wiki pages (https://github.com/drivendataorg/deon/wiki)
- Case Study (https://deon.drivendata.org/examples/)

References

References

- Introducing Deon, a Tool for Data Scientists to Add an Ethics Checklist, Natasha Mathur, September 2018 (https://hub.packtpub.com/introducing-deon-a-tool-for-data-scientists-to-add-an-ethics-checklist/)
- Actionable Ethics for Data Scientists (https://github.com/drivendataorg/odsc-actionable-ethics)
- Data Science Ethics Checklist, Marta Ghiglioni, January 2021 (https://www.martaghiglioni.com/2021/01/19/data-science-ethics-checklist/)
- An Ethics Checklist for Data Science Projects, Kelvin Washington, May 2020 (https://kelvinwellington.com/An-Ethics-Checklist-for-Data-Science-Projects/)
- Deon – Data Ethics Checklist for Data Science Projects, David Curry, January 2019 (https://www.youtube.com/watch?v=TEverjcYNkM&t=9s)

References



97 Things About Ethics Everyone in Data Science Should Know
Bill Franks, O'Reilly Media, Inc., August 2020
ISBN 978-1-492-07266-9

Privacy and Big Data, Mary E. Ludloff & Terence Craig
O'Reilly Media, Inc., September 2011
ISBN 9781449305000

Online Training

- "Artificial Intelligence Algorithms Models and Limitations", Brent Summers, Coursera (https://www.coursera.org/learn/ai-algorithm-limitations)
- "Artificial Intelligence Data Fairness and Bias", Brent Summers, Coursera (https://www.coursera.org/learn/ai-data-bias)
- "Artificial Intelligence Privacy and Convenience", Brent Summers, Coursera (https://www.coursera.org/learn/ai-privacy-and-convenience)
- "Artificial Intelligence Ethics in Action", Brent Summers, Coursera (https://www.coursera.org/learn/ai-ethics-analysis)
- "Fairness", Google (https://developers.google.com/machine-learning/crash-course/fairness/video-lecture)
- "Machine Learning Fairness: Lessons Learned", YouTube (https://www.youtube.com/watch?v=6CwzDoE8J4M)
- "Machine Learning, Ethics and Fairness", YouTube (https://www.youtube.com/watch?v=YQpnd7z0HO8)
- "Fairness and Bias in Artificial Intelligence", YouTube (https://www.youtube.com/watch?v=JCGUYFe6P2k)
- "Fairness in Machine Learning", You Tube (https://www.youtube.com/watch?v=_-OMyAn7LWM)
- "Writing the Playbook for Fair & Ethical Artificial Intelligence & Machine Learning, YouTube (https://www.youtube.com/watch?v=5pMQGT3O4CI)

References

- "Trusted AI and AI Fairness 360 Tutorial", YouTube(https://www.youtube.com/watch?v=IXbG2u4lOYI&feature=youtu.be)
- "Removing Unfair Bias in Machine Learning", IBM (https://community.ibm.com/community/user/datascience/viewdocument/removing-unfair-bias-in-machine-lea?_ga=2.14117502.876294236.1606307412-1095965628.1606307412)
- "The Emerging Theory of Algorithm", YouTube (https://www.youtube.com/watch?v=g-z84_nRQhw)
- "Algorithmic Fairness, Privacy & Ethics", YouTube (https://www.youtube.com/watch?v=AzdxbzHtjgs)
- "Making AI Fair", TED, YouTube (https://www.youtube.com/watch?v=4l_LZ5NcIBI)
- "How to Keep Human Bias Out of AI", TED, YouTube (https://www.youtube.com/watch?v=BRRNeBKwvNM)
- "Fair is Not the Default: the Myth of Neutral AI", TED, YouTube (https://www.youtube.com/watch?v=NF98WCdvR6U)
- "Can We Protect AI from Our Biases?", TED, YouTube (https://www.youtube.com/watch?v=eV_tx4ngVT0)
- "Biases Are Being Baked Into Artificial Intelligence", YouTube (https://www.youtube.com/watch?v=NaWJhlDb6sE)
- The Ethical Algorithm

Tools

- AI Fairness 360, IBM (https://aif360.mybluemix.net/)
- Fairlearn, Microsoft (https://fairlearn.github.io/)
- "How to Build Ethics into AI – Part I", Kathy Baxter, 2018 (https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-i-bf35494cce9)
- "How to Build Ethics into AI – Part II", Kathy Baxter, 2018 (https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-ii-a563f3372447)
- "How to Recognise Exclusion in AI", Joyce Chou et al, 2017 (https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850)

THANK YOU