PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

April 23, 2024

Benjamin Eva (2022). *Algorithmic Fairness and Base Rate Tracking*. Philosophy & Public Affairs, 50(2), pp. 239 - 266.

Fairness Criterion

Some fairness criterion involve studying the internal workings of the algorithm. E.g., the algorithm cannot be based on certain features.

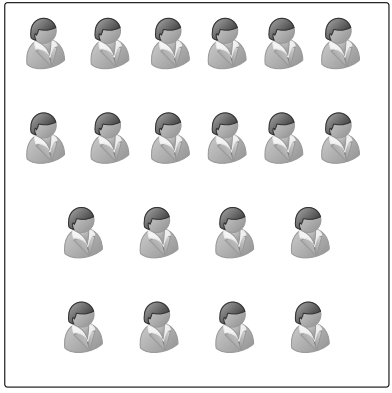
Fairness Criterion

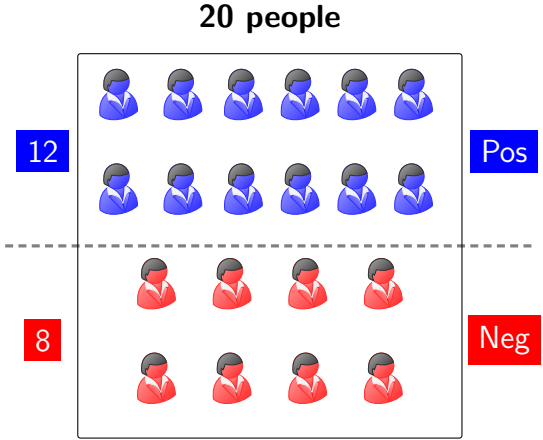
Some fairness criterion involve studying the internal workings of the algorithm. E.g., the algorithm cannot be based on certain features.

Statistical Criteria of Fairness: Criteria that require that certain relations between predictions and actuality be the same for each of the groups in question.

The criteria can be evaluated without actually looking at the inner workings of the algorithm, which may be proprietary or otherwise opaque. Instead, we just have look at the results—what the algorithm Predicted and what actually happened.

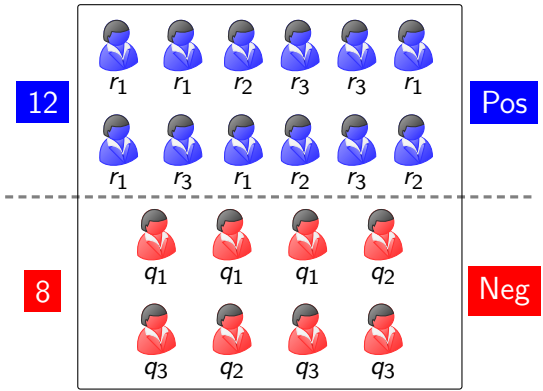
20 people





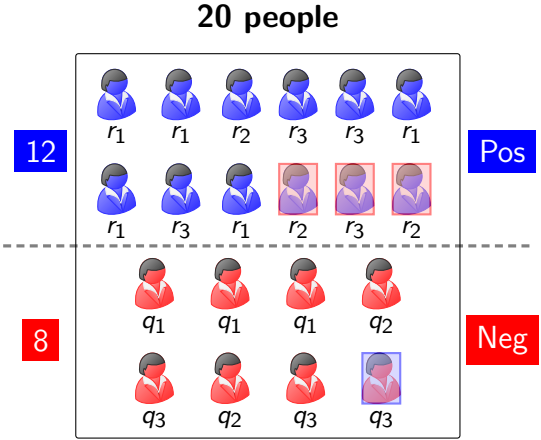
Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

20 people



Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

Predict Risk Scores: $0 \leq q_1, q_2, q_3, r_1, r_2, r_3 \leq 1$



Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

Predict Risk Scores: $0 \leq q_1, q_2, q_3, r_1, r_2, r_3 \leq 1$

Actuality: 3 classified as Pos are misclassified, 1 classified as Neg is misclassified

Fairness (1)

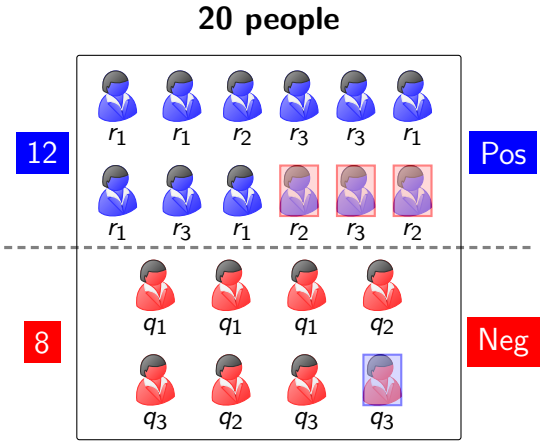
Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Fairness (1)

Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

The idea is that fairness requires a given risk score to "mean the same thing" for each relevant group. We want the assignment of a given risk score to have the same evidential value, regardless of the group to which the individual belongs.

Calibration



risk score	proportion Pos
r_1	1.0
r_2	1/3
r_3	3/4
q_1	0
q_2	0
q_3	1/3

Fairness (2)

Equal Positive Predicative Value: The (expected) percentage of individuals Predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predicative Value: The (expected) percentage of individuals Predicted to be negative who are actually negative is the same for each relevant group.

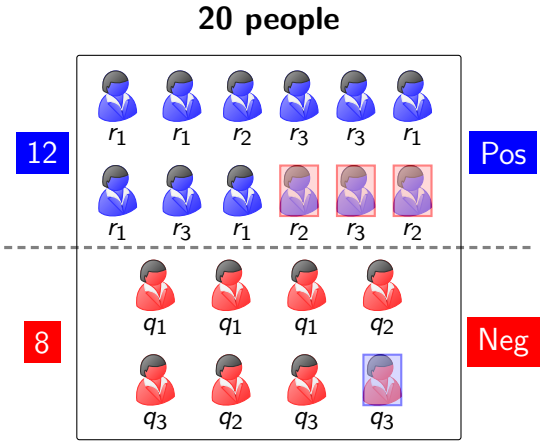
Fairness (2)

Equal Positive Predicative Value: The (expected) percentage of individuals Predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predicative Value: The (expected) percentage of individuals Predicted to be negative who are actually negative is the same for each relevant group.

The idea is that fairness requires a prediction of positive to mean the same thing, or to have the same evidential value, regardless of the group to which the individual belongs (similarly for a prediction of negative).

Pos/Neg Predictive Value



Pos Predicative Value: 9/12

Neg Predicative Value: 7/8

Fairness (3)

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

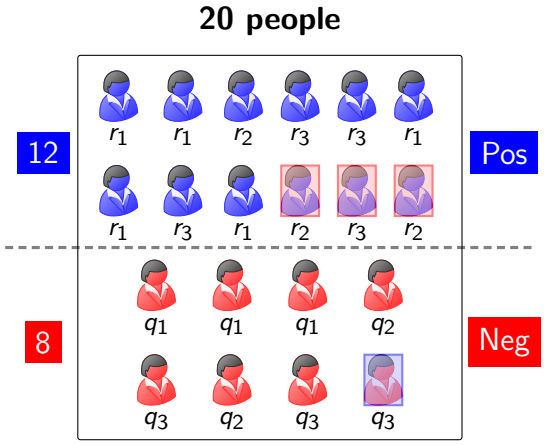
Fairness (3)

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

The idea is that fairness requires individuals from different groups who exhibit the same behavior to, on balance, be treated the same by the algorithm in terms of whether they are Predicted to be positive or negative. It would be unfair, for instance, if individuals from one group who are actually negative tended to be Predicted to be positive at higher rates than actually negative members of the other group.

False Pos/Neg Rate



False Pos Rate: 3/10

False Neg Rate: 1/10

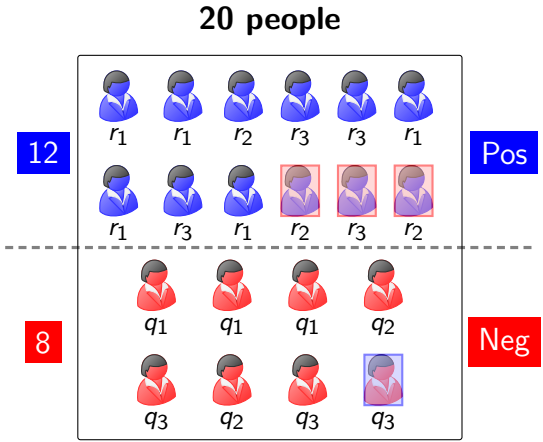
Fairness (4)

Balance for the Positive Class: The (expected) average risk score assigned to those individuals who are actually positive is the same for each relevant group.

Balance for the Negative Class: The (expected) average risk score assigned to those individuals who are actually negative is the same for each relevant group.

These are generalizations of the previous two conditions from the case of binary predictions to the case of risk scores, and are motivated in the same way.

Average Risk Scores



Average Pos Risk Score:
$(5 * r_1 + r_2 + 3 * r_3 + q_3)/10$

False Neg Rate:
$(3 * q_1 + 2 * q_2 + 2 * q_3 + 2 * r_2 + r_3)/10$

Brian Hedden has recently presented a counterexample which seems to simultaneously refute 10 of the 11 most influential criteria from the literature on algorithmic fairness.

As a result, it is far from clear exactly which criteria you should employ when evaluating the fairness of the suspect lending algorithm.

In this article, I will present, motivate and defend a novel statistical criterion of algorithmic fairness, that is, both resistant to Hedden's counterexample, and well equipped to accurately diagnose unfairness when one does not have access to the internal workings of the algorithm.

Calibration Within Groups (Strong): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Calibration Within Groups (Strong): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Calibration Within Groups (Weak): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group.

Calibration Within Groups (Strong): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Calibration Within Groups (Weak): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group.

Like the strong formulation, the weak formulation requires that every possible risk score should have the same evidential import for all relevant groups in order for the algorithm to count as fair.

Example

If 9 percent of the white drivers who are assigned a risk score of 10 percent actually get involved in accidents, then the strong formulation will deem the algorithm to be unfair, even if it is also the case 9 percent of those drivers from all other relevant groups who are assigned a risk score of 10 percent get involved in accidents.

Weak vs. Strong Calibration

Suppose that there are two rooms, A and B, containing 10 people each.

Weak vs. Strong Calibration

Suppose that there are two rooms, A and B, containing 10 people each.

- All people are assigned 2 coins, the first of which is a fair coin with a known bias of $1/2$. The second coins have unknown biases that are not available to the algorithm.

Weak vs. Strong Calibration

Suppose that there are two rooms, A and B, containing 10 people each.

- All people are assigned 2 coins, the first of which is a fair coin with a known bias of $1/2$. The second coins have unknown biases that are not available to the algorithm.
- As it turns out, the biases of the second coins are all $3/5$.

Weak vs. Strong Calibration

Suppose that there are two rooms, A and B, containing 10 people each.

- All people are assigned 2 coins, the first of which is a fair coin with a known bias of $1/2$. The second coins have unknown biases that are not available to the algorithm.
- As it turns out, the biases of the second coins are all $3/5$.
- The algorithm aims to predict whether both of a subject's two coins will land heads when flipped. Since the biases of the second coins are not available to the algorithm, it operates by assuming that all the second coins have a uniform bias of $1/2$ and then assigns each subject a risk score equal to the products of the biases of their two coins, i.e., $1/2 * 1/2 = 1/4$.

Weak vs. Strong Calibration

I think this is obviously fair. The algorithm assigns everyone from both groups the same risk score on the basis of the same evidence. And indeed, the algorithm trivially satisfies the weak formulation of calibration within groups.

Weak vs. Strong Calibration

I think this is obviously fair. The algorithm assigns everyone from both groups the same risk score on the basis of the same evidence. And indeed, the algorithm trivially satisfies the weak formulation of calibration within groups.

- The only risk score assigned by the algorithm is $1/4$ and the proportion of Room A people assigned this score whose coins both land heads is $3/10$, which is equal to the proportion of Room B people assigned the score whose coins both land heads.

Weak vs. Strong Calibration

I think this is obviously fair. The algorithm assigns everyone from both groups the same risk score on the basis of the same evidence. And indeed, the algorithm trivially satisfies the weak formulation of calibration within groups.

- The only risk score assigned by the algorithm is $1/4$ and the proportion of Room A people assigned this score whose coins both land heads is $3/10$, which is equal to the proportion of Room B people assigned the score whose coins both land heads.

- However, the algorithm also violates the strong formulation of calibration within groups, since the expected proportion of people from either room assigned the risk score $1/4$ who actually tossed two heads ($3/10$) is not equal to that risk score.

Weak vs. Strong Calibration

This example shows that only the weaker formulation of calibration within groups is plausibly a necessary condition for algorithmic fairness:

Weak vs. Strong Calibration

This example shows that only the weaker formulation of calibration within groups is plausibly a necessary condition for algorithmic fairness:

- While I agree that the above algorithm is non-ideal in the sense that it systematically underestimates the risk of agents tossing two heads, I also think it is clear that this shortcoming is not helpfully described as "unfairness."

Weak vs. Strong Calibration

This example shows that only the weaker formulation of calibration within groups is plausibly a necessary condition for algorithmic fairness:

- While I agree that the above algorithm is non-ideal in the sense that it systematically underestimates the risk of agents tossing two heads, I also think it is clear that this shortcoming is not helpfully described as "unfairness."

- If one insists on calling this kind of shortcoming "unfair," then it is clear that we need to distinguish between two conceptions of algorithmic unfairness: one that applies to uniform failings of accuracy that do not track divisions between groups, and one that manifests itself in inequitable differences in the way that different groups are treated by the algorithm.

Calibration is not Necessary

Age	Credit score	Base rate	Risk score
Young	Good	$\frac{3}{80}$	$\frac{1}{20}$
Young	Bad	$\frac{3}{80}$	$\frac{1}{10}$
Old	Good	$\frac{1}{40}$	$\frac{1}{20}$
Old	Bad	$\frac{1}{20}$	$\frac{1}{10}$

- On average, 3/80 young drivers are involved in accidents, regardless of their credit scores, while 1/20th of older drivers with bad credit scores are involved in accidents, compared to only 1/40th of those with good credit scores.

Calibration is not Necessary

Age	Credit score	Base rate	Risk score
Young	Good	$\frac{3}{80}$	$\frac{1}{20}$
Young	Bad	$\frac{3}{80}$	$\frac{1}{10}$
Old	Good	$\frac{1}{40}$	$\frac{1}{20}$
Old	Bad	$\frac{1}{20}$	$\frac{1}{10}$

- The algorithm simply assigns risk scores of 1/20 to all drivers with good credit scores, and 1/10 to drivers with bad credit scores.

Calibration is not Necessary

Age	Credit score	Base rate	Risk score
Young	Good	$\frac{3}{80}$	$\frac{1}{20}$
Young	Bad	$\frac{3}{80}$	$\frac{1}{10}$
Old	Good	$\frac{1}{40}$	$\frac{1}{20}$
Old	Bad	$\frac{1}{20}$	$\frac{1}{10}$

- For simplicity, assume that the algorithm is applied to an equal number of drivers from each of the four profiles, which implies that young drivers and old drivers both have an overall base rate of $3/80$.

Calibration is not Necessary

Age	Credit score	Base rate	Risk score
Young	Good	$\frac{3}{80}$	$\frac{1}{20}$
Young	Bad	$\frac{3}{80}$	$\frac{1}{10}$
Old	Good	$\frac{1}{40}$	$\frac{1}{20}$
Old	Bad	$\frac{1}{20}$	$\frac{1}{10}$

- The algorithm violates calibration within groups, since the base rate for young drivers with a risk score of $1/20$ is $3/80$ while the base rate for old drivers with the same risk score is $1/40$, which means that the risk score $1/20$ has different evidential implications for young drivers than it does for older drivers.

- However, it seems wrong to say that the algorithm treats older drivers unfairly in comparison to young drivers.

  The algorithm does not systematically treat younger drivers more favourably than older drivers or vice versa. On balance, it gives them equal treatment, evinced by the fact that the average risk score for both groups is 3/40, equal to half the overall base rates for both groups.

- However, it seems wrong to say that the algorithm treats older drivers unfairly in comparison to young drivers.

  The algorithm does not systematically treat younger drivers more favourably than older drivers or vice versa. On balance, it gives them equal treatment, evinced by the fact that the average risk score for both groups is 3/40, equal to half the overall base rates for both groups.

  Calibration within groups says that the algorithm treats old drivers unfairly in comparison to young drivers, but that is clearly not correct in this case. Neither group is systematically preferred to the other.

It is important to draw a distinction between two distinct possible interpretations of the calibration within groups criterion.

1. One can interpret the criterion as a diagnostic tool for identifying whether an algorithm treats some specific groups unfairly in comparison to some others.

It is important to draw a distinction between two distinct possible interpretations of the calibration within groups criterion.

1. One can interpret the criterion as a diagnostic tool for identifying whether an algorithm treats some specific groups unfairly in comparison to some others. On this interpretation the criterion can be used to check whether the pricing algorithm above treats young drivers unfairly in comparison to old drivers, for instance. And as we've just seen, the criterion gives an intuitively incorrect verdict here, since it identifies age bias where there does not seem to be any.

It is important to draw a distinction between two distinct possible interpretations of the calibration within groups criterion.

1. One can interpret the criterion as a diagnostic tool for identifying whether an algorithm treats some specific groups unfairly in comparison to some others. On this interpretation the criterion can be used to check whether the pricing algorithm above treats young drivers unfairly in comparison to old drivers, for instance. And as we've just seen, the criterion gives an intuitively incorrect verdict here, since it identifies age bias where there does not seem to be any.

2. One can interpret the criterion as a more coarse grained diagnostic tool that simply helps to identify whether the algorithm is unfair overall. On this interpretation, the algorithm is unfair just in case it is possible to identify any groups with respect to which the calibration criterion is violated.

I am skeptical of the idea that we should treat all violations of calibration as conclusive evidence of injustice.

I am skeptical of the idea that we should treat all violations of calibration as conclusive evidence of injustice.

For instance, one can imagine an algorithm, that is, calibrated with respect to age, gender, race, education, income, nationality, zip code, sexual orientation and political and religious beliefs, but that is not calibrated with respect to whether someone lives in an odd or even numbered house.

I am skeptical of the idea that we should treat all violations of calibration as conclusive evidence of injustice.

For instance, one can imagine an algorithm, that is, calibrated with respect to age, gender, race, education, income, nationality, zip code, sexual orientation and political and religious beliefs, but that is not calibrated with respect to whether someone lives in an odd or even numbered house.

In this case, it might be right to say that the algorithm treats even dwellers unfairly in comparison to odd dwellers, but that does not seem like a good reason to simply dismiss the algorithm as "unfair."

Clearly, we are more interested in evaluating statistical markers of "significant" group distinctions (e.g., race, gender, age, etc.) that track group distinctions with important social, political, economic and historical origins and ramifications.

Indeed, it seems unrealistic to expect our algorithms to be even roughly calibrated with respect to every possible group distinction, which suggests that the most we can reasonably demand is that they be calibrated with respect to all "significant" group distinctions.

But then we run straight back into the counterexample outlined above.

One could certainly make a case for the claim that the group distinction young/old is a significant one, while the distinction young & good credit/ young & bad credit/old & good credit/old & bad credit is not (if one is not convinced by this case, replace credit score with something more trivial).

But then we run straight back into the counterexample outlined above.

One could certainly make a case for the claim that the group distinction young/old is a significant one, while the distinction young & good credit/ young & bad credit/old & good credit/old & bad credit is not (if one is not convinced by this case, replace credit score with something more trivial).

This then suggests that the algorithm is actually fair after all, since it seems to be fair with respect to age, which is the only significant group

This gives two choices:

This gives two choices:

1. Argue that all group distinctions are equally relevant to an algorithm's fairness, in which case they avoid the counterexample (because the more fine grained distinction is treated as relevant to the algorithm's fairness, which implies that the algorithm is unfair), at the cost of placing unrealistic and unreasonable demands on predictive algorithms, or

This gives two choices:

1. Argue that all group distinctions are equally relevant to an algorithm's fairness, in which case they avoid the counterexample (because the more fine grained distinction is treated as relevant to the algorithm's fairness, which implies that the algorithm is unfair), at the cost of placing unrealistic and unreasonable demands on predictive algorithms, or

2. Argue that only "significant" group distinctions really matter when it comes to an algorithm's fairness, in which case the counterexample still stands (because the more fine grained partition is not treated as relevant to the algorithm's fairness, which means that the algorithm is fair even though calibration is violated).

Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same area.

Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same area.

The bank can achieve its discriminatory agenda by assigning risk scores to applicants based purely on their zip code, and ignoring other relevant factors like income, credit history, and so on.

This is an idealized illustration of a real historical phenomena called "redlining," which lenders used to avoid giving mortgages to minority applicants in the 1930s.

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- There are two zip codes, TR10 and TR11.
- Blacks are a minority in TR10 but are a majority in TR11.

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- On average, applicants in TR10 have a lower default rate than those in TR11.
- The discriminatory algorithm assigns all applicants in TR10 a risk score of 1/4 and applicants in TR11 a risk score of 3/4.

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- For both zip codes, the proportion of black and white applicants with good credit scores is the same (3/4 for TR10 and 1/2 for TR11), as is the default rate (1/8 for TR10 and 3/20 for TR11).

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- An applicant's credit score is a perfect indicator of their true default risk, in the sense that, regardless of their race and zip code, 20 percent of applicants with bad credit scores go on to default, and 10 percent of applicants with good credit scores do so.

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- By ignoring credit score and basing risk scores purely on applicants' zip codes, the algorithm seems to treat black applicants unfairly in comparison to white applicants.

However, it is easy to see that the algorithm satisfies the weak formulation of the calibration within groups criterion.