PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

April 25, 2024

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

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.

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.

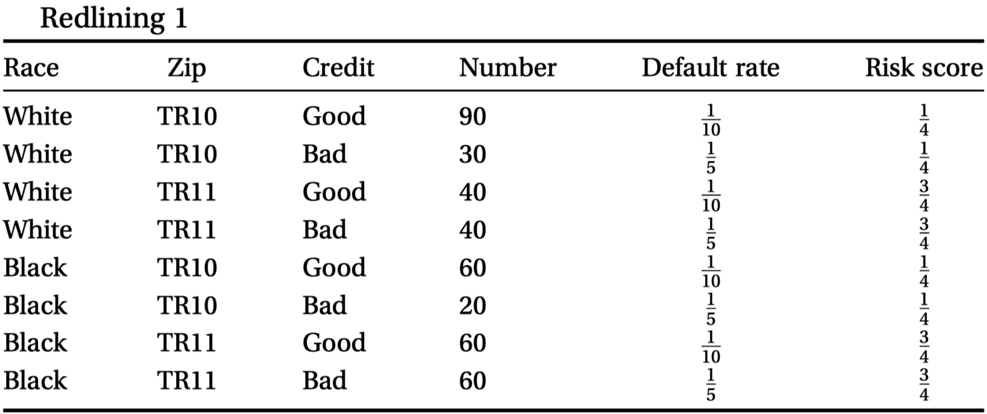
Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same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

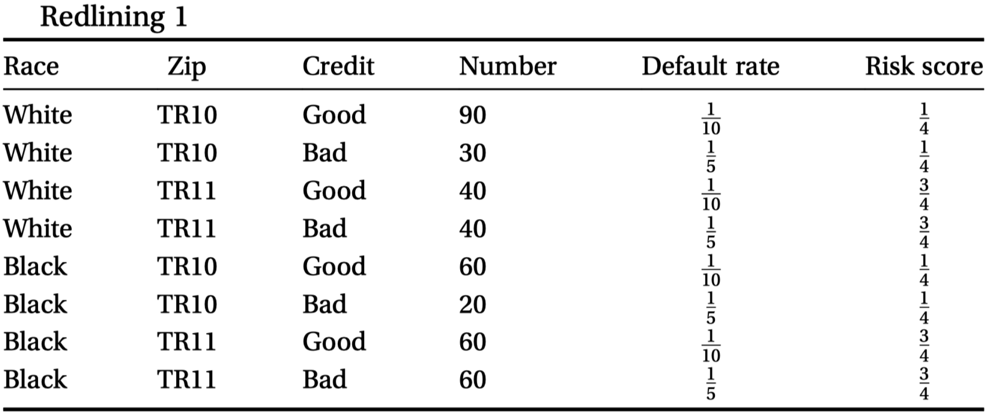
Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20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

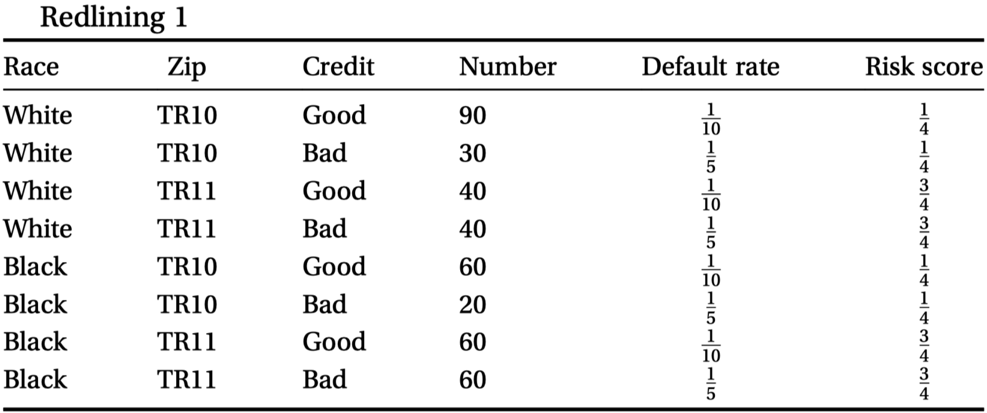
- There are two zip codes, TR10 and TR11.
- Blacks are a minority in TR10 but are a majority in TR11.



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

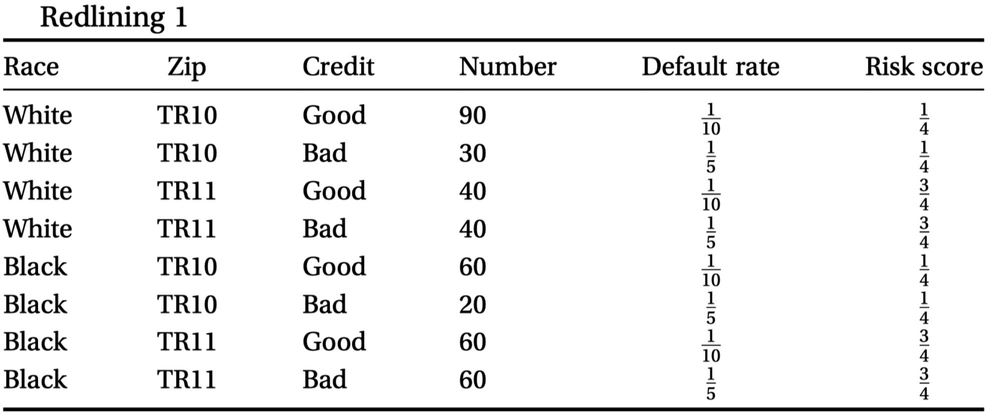
Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20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

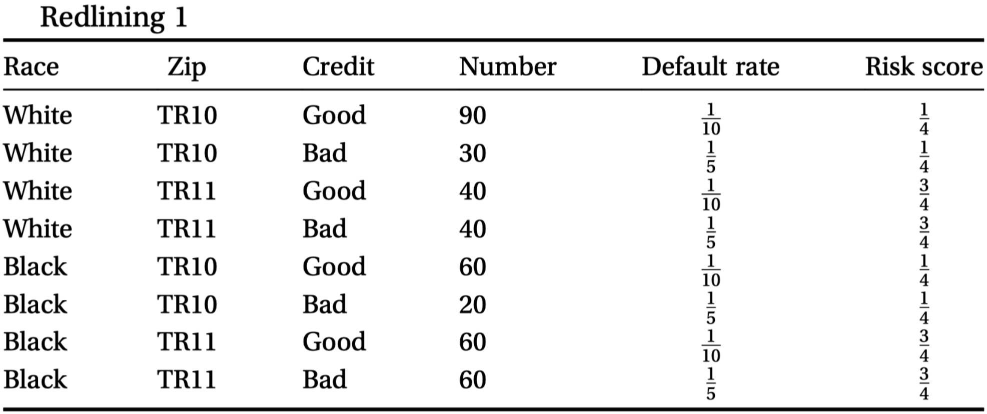


Redlining 1

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.

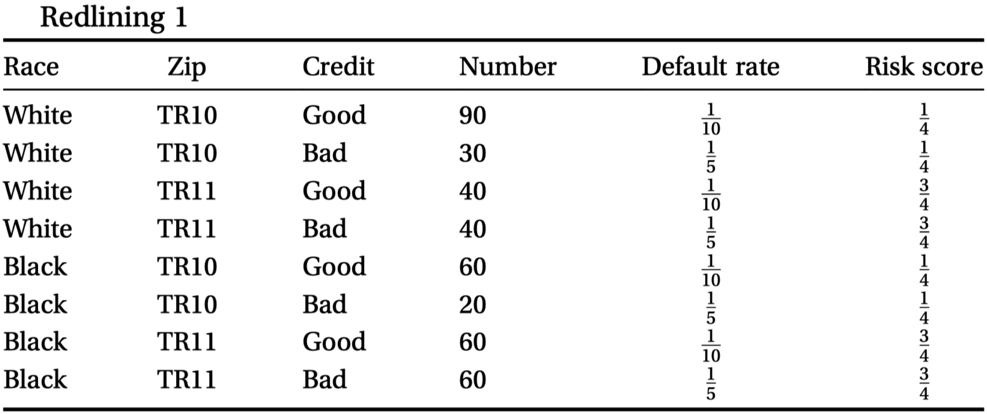
However, it is easy to see that the algorithm satisfies the weak formulation of the calibration within groups criterion.



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	$\frac{1}{5}$	$\frac{3}{4}$

- The proportion of white applicants assigned a risk score of $1/4$ who actually default is $1/8$, which is equal to the proportion of black applicants assigned a risk score of $1/4$ who actually default.



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	$\frac{1}{5}$	$\frac{3}{4}$

- The proportion of white applicants assigned a risk score of 3/4 who actually default is 3/20, which is equal to the proportion of black applicants assigned a risk score of 3/4 who actually default.

This means that both risk scores have the same evidential import for both groups, and hence that the algorithm satisfies the weak formulation of calibration within groups.

This in turn establishes that calibration within groups is not a sufficient condition for algorithmic fairness, and that even if one still thinks that calibration is a necessary condition for algorithmic fairness, one would still need further criteria in order to diagnose unfairness in cases like this.

What aspects of the Redlining example generate the obvious unfairness.

What aspects of the Redlining example generate the obvious unfairness.

- If, as in the actual historical case, the creators of the algorithm crafted it with the intention of disadvantaging black applicants, then it is obvious that the designer's actions in designing and constructing the algorithm themselves constitute a source of injustice and unfairness.

What aspects of the Redlining example generate the obvious unfairness.

- If, as in the actual historical case, the creators of the algorithm crafted it with the intention of disadvantaging black applicants, then it is obvious that the designer's actions in designing and constructing the algorithm themselves constitute a source of injustice and unfairness.
- Even if the designers of the algorithm did not explicitly intend to disadvantage black applicants, one could argue that the correlations between race, zip code and default rates are themselves the product of unjust social economic historical trends, and hence that it is unjust to apply an algorithm that exploits those correlations without recognizing, and in some way compensating for, their unjust historical origin.

In discussions of algorithmic fairness, it is crucial to keep track of distinctions between different kinds of unfairness, since the mechanisms that are best employed to combat or compensate for one kind of unfairness (e.g., the unjust historical origins of the correlations exploited by an algorithm) may not be effective in dealing with another kind of unfairness (e.g., an unfair statistical imbalance in the predictive tendencies of an algorithm).

Is there anything intrinsically unfair about the redlining algorithm or its predictions in and of themselves?

Is there anything intrinsically unfair about the redlining algorithm or its predictions in and of themselves?

Perhaps the most obvious thing to say here is that the algorithm is intrinsically unfair simply in virtue of its using zip codes as a **proxy** for race.

Problem: There is good reason to think that fairness sometimes requires predictive algorithms to explicitly base their predictions on group membership traits like gender and race.

Problem: There is good reason to think that fairness sometimes requires predictive algorithms to explicitly base their predictions on group membership traits like gender and race.

> *[I]t is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics. In the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. Recognizing this problem, some jurisdictions, like Wisconsin, have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women.* *(Corbett-Davies and Goel)*

Sam Corbett-Davies and Sharad Goel (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. https://arxiv.org/abs/1808.00023.

It is difficult to define exactly what it means for a predictive feature to be used as a proxy for a group membership trait.

It is difficult to define exactly what it means for a predictive feature to be used as a proxy for a group membership trait.

On what grounds can one say that zip code counts as a proxy for race in the above case, while other variables that are also correlated with race do not count as proxies?

This problem is further compounded when we recall that the predictive algorithms whose fairness we hope to assess are often proprietary, meaning that we do not actually know exactly which predictive features are being employed by the algorithm.

It is clear that merely citing the use of a proxy variable does not helpfully identify what is intrinsically wrong with the algorithm in the Redlining example.

If the algorithm used some other features rather than zip code to obtain the same predictions, it would still be just as unfair.

It is clear that merely citing the use of a proxy variable does not helpfully identify what is intrinsically wrong with the algorithm in the Redlining example.

If the algorithm used some other features rather than zip code to obtain the same predictions, it would still be just as unfair.

It seems to me that there is something *intrinsically unfair* in the predictions themselves, and that we should not need to refer to the predictive features used by the algorithm in order to diagnose that unfairness.

We should be able to diagnose the intrinsic unfairness of the algorithm's predictions using statistical criteria alone.

But as we have just seen, the most popular statistical criterion of algorithmic fairness from the literature, calibration within groups, is unable to identify any unfairness in this case.

We need a new criterion to help us clearly diagnose the sense in which the predictions of the algorithm in the Redlining example are intrinsically unfair.

What does it mean to say that the Redlining algorithm is “intrinsically unfair”?

- It is clear that the fairness of an algorithm is a function of many factors, including e.g., the intentions of its designers, the social and historical context of its design and application, the historical origins of the correlations that it exploits and the statistical profile of its predictions.

What does it mean to say that the Redlining algorithm is “intrinsically unfair”?

- In order to get a full picture of the (un)fairness of the algorithm, we generally need to have access to all of these features and many more. And since many of these factors concern properties of the social/historical situation inhabited by the algorithm, it seems clear that the algorithm’s overall (un)fairness is not an intrinsic property.

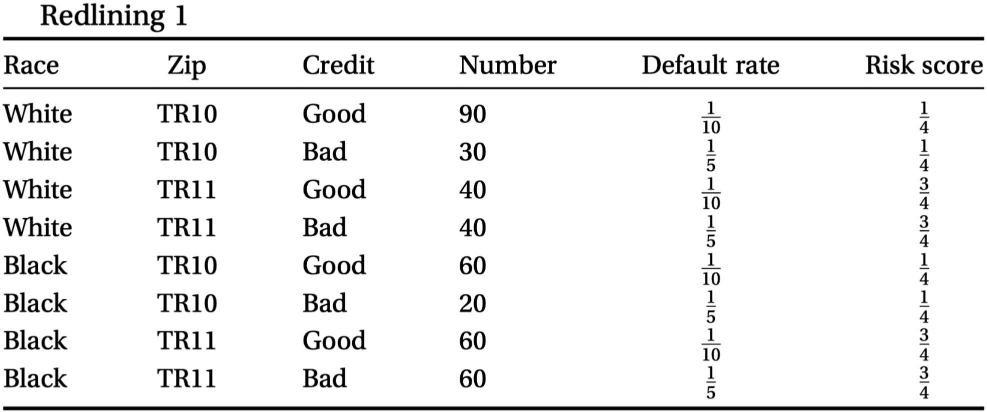
What does it mean to say that the Redlining algorithm is "intrinsically unfair"?

- But it is possible to acknowledge that the (un)fairness of an algorithm is generally far from an intrinsic property while also recognizing that there are some intrinsic properties of algorithms such that any algorithm that has those intrinsic property is bound to be unfair to some degree, regardless of its other non-intrinsic properties.

What does it mean to say that the Redlining algorithm is “intrinsically unfair”?

- When algorithms make predictions that systematically favor one group over another, we can conclude that those algorithms are unfair, at least to some degree, in a way that is independent of their social/historical context (in the sense that any algorithms with the same statistical profiles will be similarly unfair, regardless of their internal workings and social context).

Base Rate Tracking

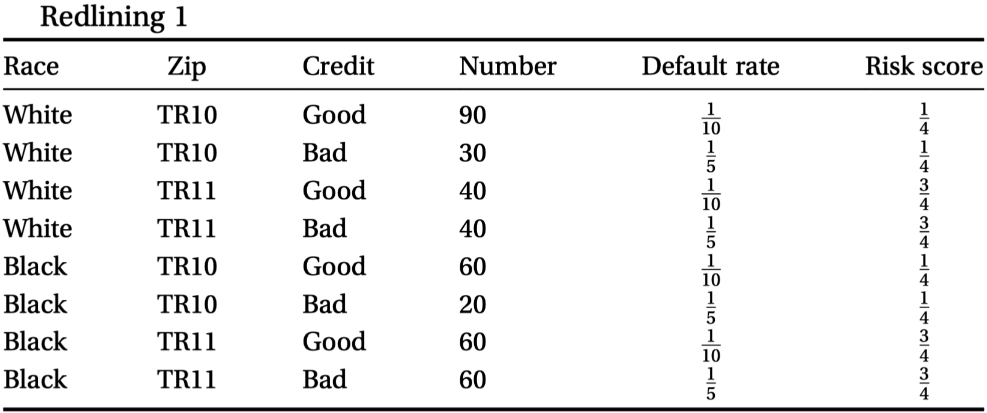


Redlining 1

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	$\frac{1}{5}$	$\frac{3}{4}$

- The overall average risk score for white applicants is $9/20$, while the overall average risk score for black applicants is $11/20$.

Base Rate Tracking

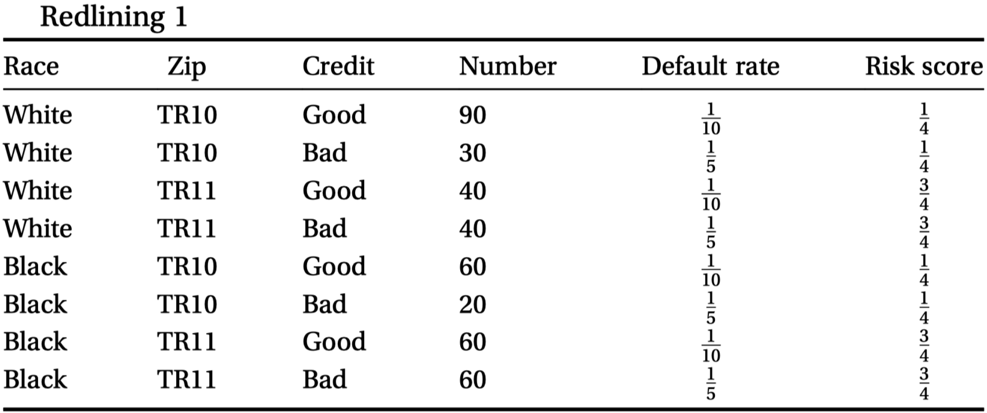


Redlining 1

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	$\frac{1}{5}$	$\frac{3}{4}$

- The overall default rate for white applicants is 27/200, while the overall default rate for black applicants is 28/200.

Base Rate Tracking



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	$\frac{1}{5}$	$\frac{3}{4}$

- The difference between the average risk scores of the two groups is 20 times as great as the difference between their actual default rates.

Base Rate Tracking

If an algorithm assigns one group a higher average risk score than another, that discrepancy has to be justified by a corresponding discrepancy between the base rates of those two groups, and the magnitudes of those discrepancies should be equivalent.

Base Rate Tracking

In slogan form: an algorithm should only treat one groups as much more risky than another if it really is much more risky.

Base Rate Tracking: The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.

Applying Base Rate Tracking to the Redlining algorithm:

Since the difference between the average risk scores assigned to white and black applicants is 20 times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants.

Applying Base Rate Tracking to the Redlining algorithm:

Since the difference between the average risk scores assigned to white and black applicants is 20 times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants.

If we were to rely only on calibration within groups, then we would need to refer to the designers' intentions, or the unjust historical origins of the relevant correlations, or the internal workings of the algorithm, in order to diagnose the unfairness in this case. But base rate tracking allows us to directly identify the algorithm as intrinsically unfair on the basis of its predictions alone.

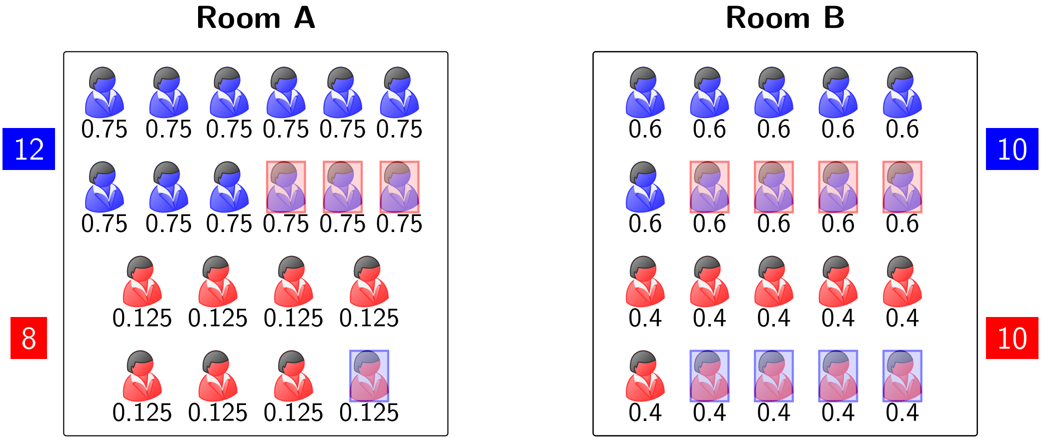
Unlike calibration within groups, base rate tracking really is a statistical criterion of algorithmic fairness, i.e., a necessary condition that any fair algorithm must satisfy.

Base Rate Tracking allows us to directly identify the algorithm as intrinsically unfair on the basis of its predictions alone. Given the lack of information that is generally available regarding the design process and internal architecture of predictive algorithms, this is important, since it shows that base rate tracking allows us to identify algorithmic unfairness in many cases where we would otherwise be unable to do so.

Base Rate Tracking is motivated by a natural philosophical intuition regarding the nature of fairness: that any difference in the way that an algorithm treats two groups needs to be justified by a corresponding difference in the relevant behaviors/properties of the two groups.

It is unfair to treat white loan applicants as if they have a much lower average risk of defaulting compared to black applicants if they do not actually have a much lower default rate.

Base Rate Tracking, unlike the 10 influential criteria of fairness discussed last week, is not undermined by Hedden's counterexample.



Since the base rates for the two rooms are equal to the average risk scores assigned to the people in those rooms, base rate tracking is trivially satisfied by the optimal predictive algorithm.

Base Rate Tracking...

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,
2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,
2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and
3. ...significantly expands the diagnostic scope of calibration within groups in some important cases.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.

And one might plausibly object that this is obviously unfair, since black applicants have the same default rate as white applicants within any given zip code.

This in turn implies that base rate tracking is not a plausible statistical criterion of algorithmic fairness.