Algorithmic Fairness in Lending: How Should We Measure Fairness?

Written by Andrew Green

OPEN BANKER
Open Banker
September 23, 2025

Andrew Green is a mission-driven data scientist with a passion for using data to improve how government designs policies, delivers services, and protects consumers. He previously worked on algorithmic fairness issues at the CFPB.

Open Banker curates and shares policy perspectives in the evolving landscape of financial services for free.

Fairness is something we care deeply – maybe even instinctively – about. So as an increasing number of decisions are made by automated systems, we're increasingly concerned that the algorithms powering those decisions are fair. But one of the most basic questions in algorithmic fairness is also one of the thorniest: what do we mean by *fair*? Is a model fair if it produces similar outcomes across groups? Or should a model have similar predictive performance across groups to be considered fair? Which metrics best capture these conceptions of fairness? And how should we handle metric conflicts, where one metric indicates a model produces unfavorable disparities for group A and favorable disparities for group B, while another metric shows the exact opposite?

These are difficult questions. The answers can't be solved for in an optimization algorithm or found in the various titles of the United States Code. But that's also what makes them interesting. Because they are fundamentally *normative* questions, it's not just data scientists or lawyers that get to weigh in; the answers can and should be shaped by our collective sense of what fairness means to us.

Fair Lending

In the lending context, discussions of algorithmic fairness often fall under the realm of disparate impact, one of the two primary types of discrimination that the Equal Credit Opportunity Act (ECOA) protects against.[1] Disparate impact discrimination is when a lender uses a policy — an underwriting model, for example — that is facially neutral, but produces disparities on the basis of a protected class like race, sex, or age.

Disparate impact analysis usually focuses on metrics related to a key outcome, such as being approved or denied for credit, or the model scores that are used to make those decisions. A common metric for binary outcomes, like the approve/decline decision, is the adverse impact ratio (AIR), which is calculated as the approval rate of a given protected class group, divided by the approval rate of a "control" protected class group. For example, the AIR for Black applicants may be calculated by dividing the approval rate for Black applicants by the approval rate for White applicants. While there aren't any hard and fast rules on what level of disparity would raise disparate impact concerns, many lenders use AIR thresholds of 80% or 90% to flag potential risk.

While outcomes-based measures like AIR — which focus on disparities in outcomes without regard to differences in underlying characteristics across groups — have traditionally been the focus of disparate impact testing, some have argued that lenders and regulators should instead focus on measures of differential validity, or how accurate a model is for different groups. This could involve analyzing whether disparities persist when controlling for common model inputs like FICO scores, or it could involve looking at metrics that compare



A Question of Values

This seemingly dry question of whether we should use outcomes-based measures or differential validity-based measures of fairness is first and foremost a question of values.

It's well-established that there are large disparities in income and wealth by race/ethnicity. These disparities are driven by historical and current discrimination in the form of redlining, employment discrimination, reduced access to traditional credit, and biases in the criminal justice system, among other sources. These forms of discrimination make it harder to earn income, accumulate wealth, and get access to credit on favorable terms; those factors then make it more likely that someone will become delinquent on their credit, which in turn will hurt measures of their creditworthiness, such as FICO scores.

These mechanisms are generally well-accepted and usually not the subject of debate; what *is* the subject of *fierce* debate is what should be done about them. Do lenders have a responsibility to address the effects of past discrimination on their models and lending decisions? Or should credit models simply reflect the world as it is?

Critics of outcomes-based metrics argue that disparities on these measures are simply downstream of broader socioeconomic disparities, and it's not the role of credit modeling to correct for inequities in society. They may prefer to look at measures of disparities that control for things like FICO scores and argue that a model is fair if it tends to deny Black applicants and White applicants with similar FICO scores at around the same rate, for example.[3]

On the other hand, critics of differential validity-based metrics argue that measuring disparities in this way is self-defeating, given that the inputs we're controlling for have discrimination baked into them. In doing so, we may be masking the exact discrimination that we're trying to address.

This argument is fundamentally a question of values. Focusing on disparities in outcomes puts us on a path toward addressing historical discrimination, by reducing the extent to which it affects future lending decisions. Alternatively, focusing on disparity measures that control for creditworthiness characteristics aims to ensure a model doesn't introduce *any new discrimination* beyond what may exist in these factors as they're measured today, but at the risk of "entrenching" historic discrimination.

> There aren't any easy answers to these questions. Where we ultimately land should reflect what we collectively believe to be fair, and the obligations we think lenders should have to create a fairer system.



Given the thorny normative questions raised here, lenders may seek a balanced approach that considers both outcomes-based metrics and differential validity-based metrics in some capacity.

However, a significant complicating factor is that these metrics may lead to very different conclusions and may even directly contradict one another. For example, an outcomes-based metric may show that Black and Hispanic applicants have unfavorable disparities relative to White applicants, while Asian applicants have a favorable disparity relative to White applicants; meanwhile a differential validity-based metric may show that Black and Hispanic applicants have a favorable disparity relative to White applicants, while Asian applicants have an unfavorable disparity relative to White applicants.

How do we resolve this type of conflict? While it's important to monitor both types of metrics, lenders – and their regulators – should have a clear sense for how they will "break the tie" if and when the metrics conflict. There are two dimensions lenders and regulators should consider for these decisions.

First, they should consider the context/use case of the decision or model in question, and how that may impact their choice of fairness metric. For example, when evaluating disparate impact in the decision of whether an applicant is approved or denied credit, lenders may choose to look at both types of metrics but ultimately set risk thresholds and act primarily based on an outcomes-based metric like AIR. Given the centrality of access to credit in modern financial well-being, there's a clear normative case for combating large disparities in credit access across groups.[4] [5]

On the other hand, other credit modeling contexts have a less clear normative case for focusing on an outcomes-based metric. For example, consider automated valuation models (AVMs), which estimate property values. AVMs are used in various mortgage financing contexts, and a property being over- or under-valued may have different impacts on the buyer and seller, depending on the use case. As a result, disparities in model accuracy are likely the primary fairness concern, so disparate impact analysis of AVMs may involve focusing primarily on a measure of differential validity.

Second, lenders and regulators should consider the magnitudes of the disparities across fairness metrics. For example, let's say a model generates large unfavorable outcomes-based disparities for groups A and B, and relatively small unfavorable differential validity-based disparities for groups C and D. It may be prudent to focus on reducing the outcomes-based disparities for groups A and B, even if doing so slightly exacerbates the differential validity-based disparities for groups C and D. This sounds like an obvious point — we should care about large disparities more than small ones — but these discussions often only focus on the conceptual merits of the different types of metrics, neglecting to also consider practical concerns like magnitudes.

An Illustrative Example



Let's say a lender is designing their framework for disparate impact analysis related to loan application decisions and the underwriting model used for those decisions.

The lender starts out by selecting one outcomes-based metric and one differential validity-based metric of interest. They select AIR for outcomes-based disparities and FPR for differential validity-based disparities. They observe large unfavorable disparities on AIR for group A, and small unfavorable disparities on FPR for group B.[6]

Given the strong normative case for combating large disparities in credit access, the lender decides to prioritize monitoring and improving AIR disparities, with FPR disparities as a secondary focus. They decide to take action based on the following tiered decision-making logic:

- If AIR is less than 80% for one or more groups, the lender will conduct a search for less discriminatory alternative (LDA) models focused on reducing that disparity.
    - FPR will be monitored, but not necessarily acted upon; the focus is on combating the large disparities in access to credit.
- If AIR is between 80% and 90% for one or more groups, the lender will conduct a LDA search focused on reducing that disparity, subject to constraints on how FPR is affected.
    - The lender will define a ratio of AIR improvement for group A to FPR degradation for group B and only consider LDA models that meet or exceed this ratio.[7]
- If AIR is greater than 90% for all groups, the lender may decide to either a) take no action, or b) conduct a LDA search focused on reducing FPR disparities, subject to the constraint that AIR must remain above 90% for all groups.

Closing Thoughts

The question of how we should measure fairness in lending decisions is thorny. Measures that focus only on disparities in outcomes and metrics that attempt to measure disparities in predictive performance each draw valid criticisms that shouldn't be dismissed out of hand.

A sensible approach to disparate impact testing involves considering both types of measures in some capacity, but metric conflicts make this tricky and pose a risk that lenders will be paralyzed with indecision. Lenders and regulators should think deeply about how to structure and prioritize disparate impact testing based on the context/use case of the decision in question and the magnitudes of the



(prospective) customers, and regulators allow for flexibility in how disparate impact testing is conducted.

There aren't any easy answers or one-size-fits-all solutions to the question of how to measure fairness in lending decisions, but talking honestly and rigorously about the factors and tradeoffs involved gives us the best shot at finding the right path forward.

The opinions shared in this article are the author's own and do not reflect the views of any organization they are affiliated with.

---

[1] The other primary type of discrimination is disparate treatment, which is when a lender treats applicants differently based on a protected class. Including a variable for race in a credit model would be a clear-cut case of disparate treatment, while including a variable that acts as a proxy for race could potentially be a disparate treatment issue.

[2] False positive rate (FPR) in this context is defined as the percentage of non-defaulters/goods that are improperly classified as defaulters/bads. A given protected class group would be considered to have an unfavorable disparity if the FPR for that group is higher than the FPR of the control group.

[3] The approach described here uses a measure of disparities that controls for model inputs; an alternative approach could use a measure of disparities that controls for observed outcomes. This type of approach seems conceptually harder to argue with — even the most fierce advocates of combating historic discrimination would likely concede that a lender's model should be able to accurately identify people who have defaulted on their loans from that same lender. However, the challenge is that lenders can't directly observe outcomes for any applicants they denied. Instead, they must rely on reject inference for these applicants, which will inevitably introduce error into the outcome variable. These concerns only multiply when one group has been denied at a higher rate than other groups in the past. Ignoring inferred outcomes altogether is also problematic, because it means focusing on a very biased sample of all applicants (i.e., only those that were approved).

[4] This doesn't mean more access to credit is always good — giving loans to people who clearly have a high risk of not paying them back is bad for both the lender (losing money) and the borrower (damage to credit score and thus future ability to access credit) — but we are generally talking about approving more people around the margin of a lender's decision threshold, who are generally still unlikely to default.

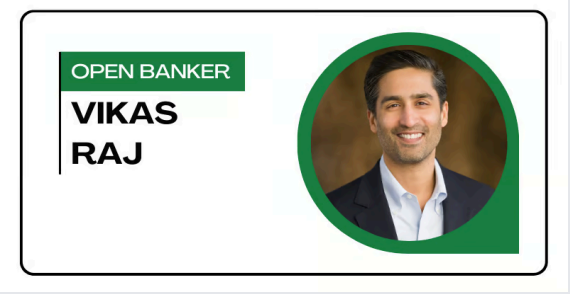


integrity if the underlying socioeconomic disparities are large enough. Achieving modest/incremental improvements to outcomes-based fairness while maintaining comparable predictive performance would likely be a desirable outcome, even if disparities remain.

[6] In reality, these two metrics may not conflict with each other. If that were the case, the logic becomes simpler.

[7] In their paper titled "Modernizing Fair Lending", Caro, Gillis, and Nelson provide a framework for explicitly balancing tradeoffs between outcomes-based metrics and differential validity-based metrics. Under their framework, a lender may choose to pursue a LDA model that offers substantial improvements to a differential validity-based metric at the cost of a small deterioration in an outcomes-based metric, for example. The lender and/or regulator would choose how to set the relative weights of the metrics, expressed as the "slope" of the tradeoff between them. What I'm describing here is an extension of their framework, which would allow the tradeoffs to be made in either direction.

Keep reading



Fintech Policy Needs a Resilience Focus

Written by Vikas Raj

Open Banker /

Modernizing Supervisory Information Sharing: A Former Regulator's...

Written by Kayce Seifert

Open Banker /

Congress Must Reject the GENIUS Act and Remove the Dangers Posed by...

Written by Art Wilmarth

Open Banker /

View more

Open Banker

Financial policy without the paywall.

Home

Posts

E.. Subscribe

Privacy policy Terms of use