Measuring Discrimination Under Model Usage Uncertainty

Nicholas Schmidt, BLDS, LLC and SolasAI
Matthew Boutte, BLDS, LLC and SolasAI

Introduction

It has become standard practice for lenders to perform discrimination testing on the algorithms they use for high-stakes decisions, such as those used to make credit offers. For this testing, there are a number of metrics that have gained widespread acceptance by lenders because of their common use in courts and apparent acceptance by regulators. These are rigorous and provide crucial information about the fairness of models. However, these metrics all suffer from a common limitation: they all assume that it is known at the time of model development and testing how a model will be used in production.

This assumption is rarely true. Data scientists may develop a model before the business has decided how the model's predictions will be used. Alternatively, a single model may be used in different ways for different use cases. Or how a model is used may change over time in response to changing economic conditions or new business strategies. Each of these common scenarios decreases the usefulness of existing disparity testing metrics.

In a series of papers, Discover Financial Services (Discover) proposed a new metric to measure disparity that overcomes these limitations of existing metrics. The authors of this white paper have reviewed Discover's papers, developed software to compute and analyze the metric, and conducted an independent analysis of how the metric performs across various use cases. We find that the metric is a promising innovation that overcomes the limitation of unknown model usage while continuing to adhere to existing legal and regulatory frameworks and best practices.

The Wasserstein Distance

The metric proposed by Discover is based on the Wasserstein Distance – a measure of the distance between two probability distributions. The metric is colloquially known as the "earth mover's distance" because of a commonly used illustration to gain an intuitive understanding of the Wasserstein Distance. Imagine two piles of soil (representing two probability distributions). The Wasserstein Distance is a measure of the minimal amount of effort (or, more precisely, work) required to move every point of the first pile of soil so that it is identical in shape to the second pile.

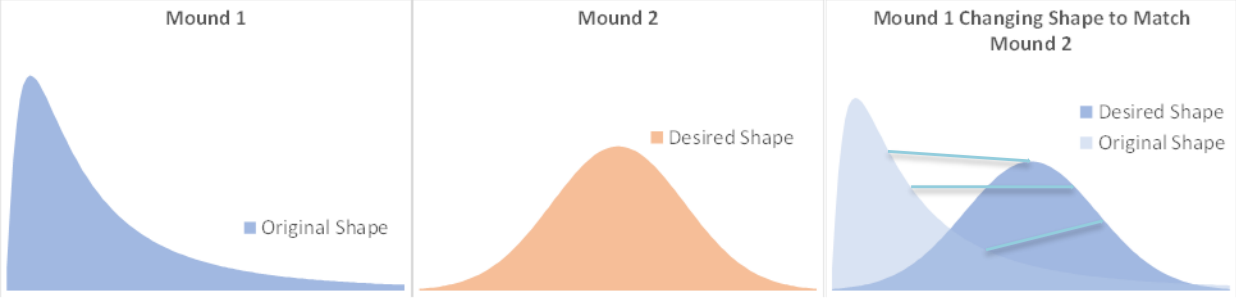

Figure 1: The Earth-Mover Distance

Here, we show that there are two piles of earth, and we ask what it will take to change the pile shown in blue on the far left to look like the shape of the tan pile in the middle. The challenge is to do so in an optimal way – that is, minimizing the amount of work performed. An optimal strategy – which corresponds to the Wasserstein Distance – accomplishes this by taking a point from Mound 1, finding its percentile, and then moving it to the point on Mound 2 that corresponds to the same percentile. For example, the point at the 25th percentile of the blue distribution should move the point at the 25th percentile of the tan distribution; the median point should move to the median point, and so on. The total amount of work required to carry out this transformation is the Wasserstein metric.

The Discover papers cleverly identified that this concept can be directly applied to assess the fairness of models: the authors use the Wasserstein Distance to assess the effort required to move the distribution of model scores of one group into the distribution of scores of another group. For example, given the distribution of model scores of men and the distribution of scores of women, the Wasserstein metric asks how much women's scores would need to be changed for their distribution of scores to be identical to the distribution of scores of men. Below, this is shown graphically. As is clear, it is identical to the graph above, except that the names of the axes have changed.

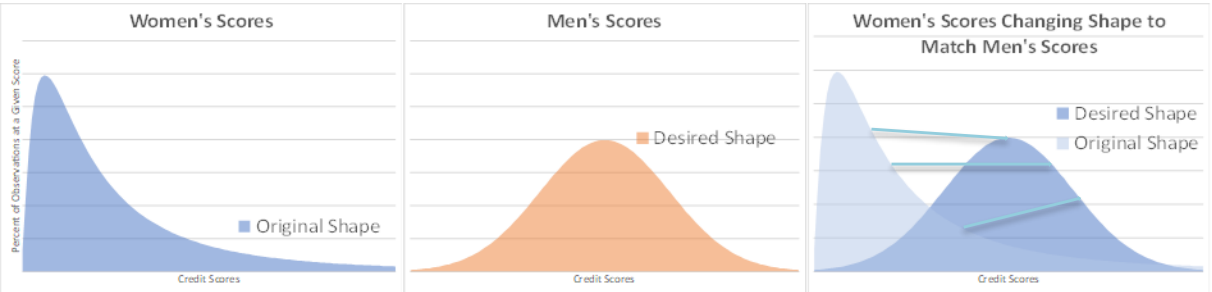

Figure 2: Differences in Score Distributions

One can easily see why this may be a useful metric for assessing discrimination: one may assume that if the distributions of scores for two groups are similar, i.e., there is mostly overlap in the rightmost picture above, then the two groups likely receive similar treatment. When this occurs, the Wasserstein Distance will be small. On the other hand, if, like in the picture shown above, there is relatively little

overlap between the two groups, then the value of the Wasserstein Distance will be large. This may signal evidence of discrimination; it almost certainly indicates that the difference is worth reviewing to understand its cause.

A direct application of the Wasserstein Distance to model fairness is straightforward. A model will produce a distribution of predictions for a protected group and a second distribution of predictions for the corresponding reference group. The more similar these distributions are to each other, the lower the level of disparity and the smaller the Wasserstein Distance; the less similar the distributions are to each other, the higher the level of disparity and the greater the Wasserstein Distance. If the distributions are identical, then the Wasserstein Distance would be zero (no "soil" would have to be moved in one distribution to make it match the second distribution) and there would be no disparity.

Measuring Disparate Impact with the Wasserstein Distance

In addition to being a novel application for the metric, the Wasserstein metric also dovetails with existing measures of disparity. Of these, the most intuitive measure of discrimination is likely the metric known as "Marginal Effects." Marginal effects are equal to the difference in average outcomes between a protected group and its associated reference group.[1] For example, suppose 75% of Black or African American applicants (the protected group) receive loan offers, but 81% of White applicants (the reference group) receive offers. The Marginal Effect is 81% - 75% = 6%. In other words, White applicants are an additional 6 percentage points more likely to receive loans than Black or African American applicants. Because of its simplicity, using marginal effects to measure discrimination is appealing. However, in credit models, this becomes difficult when multiple cutoffs may be used. Fortunately, the Wasserstein Distance can be used to aggregate marginal effects measures into a single measure of potential discrimination.

To illustrate with a simplified example, consider the table below, Table 1, where we assume a credit default model is to be used to give offers assuming two possible cutoff points. Which of the two cutoffs will be used will be determined based on macroeconomic conditions after the model is put into production (i.e., macroeconomic conditions will not be known until the model is being used). In the more restrictive first cutoff, "Cutoff 1: Recession," we see that 50% of women receive offers, but 65% of men do. In the less restrictive offer, "Cutoff 2: Expansion," 65% of women receive offers and 75% of men receive offers. In the first case, men are 15% more likely to receive loans (i.e., the marginal effect equals 15%). In the second case, the men are 10 percentage points more likely to receive loans. Thus, while there are disparities in both situations, they are worse under the Recession scenario.

[1] Marginal effects, when used in regression analysis, can have a slightly more technical interpretation. That usage is outside of the scope of this paper.

Cutoff	Women	Men	Marginal Effects	Notes
Cutoff 1: Recession	50%	65%	15.0%	Disfavors women
Cutoff 2: Expansion	65%	75%	10.0%	Disfavors women
Wasserstein Distance (Average Marginal Effects)			**12.5%**	

Table 1: Calculating the Wasserstein Distance Across Two Cutoffs

Since no one knows what the macroeconomic conditions will be when the model is put into production, it is important to be able to measure the overall disparity of the model assuming either of the two use cases. Being able to do this using some aggregation of marginal effects would be especially helpful, given that this metric has an intuitive meaning and is already commonly used in fair lending analyses. What the Discover researchers realized is that, under certain conditions, the Wasserstein Distance does aggregate marginal effects by being equal to the average of the marginal effects over the two cutoffs. In this example, the Wasserstein Distance is equal to 12.5% as that is the average of the 15% difference and the 10% difference. It is crucial to point out that this identity is reliant on the marginal effects favoring the reference group for both cutoffs. This will be discussed in more detail below.

There is an intuitive interpretation of using the average marginal effects across cutoffs (i.e., using the Wasserstein Distance) that further makes it an appealing measure of disparity. This is that, if all we know is a person's protected class status and that the model will be implemented using one of these two cutoffs, then the Wasserstein Distance is the best measure of the expected percentage shortfall as a result of using the model.[2] In the case above, since the macroeconomic conditions cannot be known during model development, our best estimate of the shortfall of women's offers is the Wasserstein Distance value of 12.5%.

As mentioned above, the Wasserstein Distance is not always equal to the average marginal effects: they are only equal when the protected class is negatively impacted across all cutoffs. In more general settings, the Wasserstein Distance is equal to the average of the absolute marginal effects, i.e., without taking into account whether the protected class or the reference group is negatively affected. To illustrate, if the protected class is disfavored by 10% over one cutoff, but favored by 5% over the second cutoff, then the Wasserstein Distance is equal to the average of those two values, 7.5%. In other words, the Wasserstein Distance is the deviation from parity across all thresholds. While one might initially find this unappealing, it proves to be a helpful tool in understanding potential discriminatory or beneficial effects of variables that go into a model.

[2] The calculations here assume that expansions are equally as common as recessions. If this is not accurate, then the calculation can easily be adjusted by weighting by how likely the economy will expand or contract. For example, if the developer expects that, over the life of the model, 75% of the time will be during economic expansions, and 25% of the time will be during contractions, then the Wasserstein Distance is calculated as (75% * 10%) + (25% * 15%) = 11.25%.

Below, we show an example of a full decomposition of bias metrics across five thresholds, three where the protected group, women, are disfavored (Cutoffs 1-3), and two where women are advantaged (Cutoffs 4-5). Descriptions of the calculations and their usefulness are then provided.

Cutoff	Women	Men	Marginal Effects	Notes
Cutoff 1	50%	65%	15%	Favorable to Men (Positive Bias)
Cutoff 2	65%	75%	10%	Favorable to Men (Positive Bias)
Cutoff 3	83%	85%	2%	Favorable to Men (Positive Bias)
Cutoff 4	88%	87%	-1%	Favorable to Women (Negative Bias)
Cutoff 5	91%	89%	-2%	Favorable to Women (Negative Bias)
Average Unfavorable Marginal Effects			9.0%	When Women are Negatively Impacted
		Positive Bias	5.4%	When Women are Negatively Impacted
		Negative Bias	0.6%	When Women are Favorably Impacted
		Net Bias	4.8%	Net Disparity to Women
Total Bias (Wasserstein Distance)			6.0%	Total Absolute Disparities to Men and Women

Table 2: Calculating the Wasserstein Distance Across Five Cutoffs

Average Unfavorable Marginal Effects are equal to the average of the three cutoffs where women are disfavored. (15% + 10% + 2%) / 3 = 9%. This is a measure of how extensive the disparity is across cutoffs where women are disfavored. It is also a standard measure of disparity used by courts and regulators to understand whether members of protected groups are disadvantaged by a model. It does assume that no offsetting is allowed in measuring disparity. In other words, when one cutoff gives a favorable result and another gives an unfavorable result, this measure only considers the unfavorable amount. When offsetting is allowed, the correct measure of bias is Net Bias, as defined below.

Positive Bias is calculated as the sum of the marginal effects where women are disfavored divided by the total number of cutoffs. (15% + 10% + 2%) / 5 = 5.4%. Importantly, when women are favored, the calculation does not offset the unfavorable amounts by the favorable ones. As such, this can also be calculated as (15% + 10% + 2% + 0% + 0%) / 5 = 5.4%. In other words, the -1% and -2% favorable marginal effects are set to zero.

This is an aggregate measure of the level of disparity harming women. Returning to the definition of the Wasserstein Distance, positive bias is technically the effort required to move predictions for the protected group in the favorable direction.

Negative Bias is calculated as the sum of the marginal effects where women are favored divided by the total number of cutoffs. (1% + 2%) / 5 = 0.6%. This is an aggregate measure of the level of disparity favoring women. In a manner like positive bias, cutoffs where women are disfavored are set to zero. As such, the formula can also be written as (0% + 0% + 0% + 1% + 2%) / 5 = 0.6%.

In practice, negative bias acts to offset positive bias to make a model fairer on average across all possible cutoffs. Returning to the definition of the Wasserstein Distance, negative bias is technically the effort required to move predictions for the protected group in the unfavorable direction.

Net Bias is the average of the marginal effects across all thresholds. (15% +10% + 2% - 1% - 2%) / 5 = 4.8%. It is also equal to the positive bias minus the negative bias. If this is positive, then the model on average disfavors women across all cutoffs; if it is negative, then women are, on average, favored across the cutoffs. It is further equal to the Average Unfavorable Marginal Effects when protected classes are disfavored across all cutoffs. Finally, when a legal standard allows offsetting benefits and harms from one score usage to another, then this would be the appropriate metric of disparity instead of the Average Unfavorable Marginal Effects.

Net bias is particularly useful when compared to the total bias since the amount of offsetting due to negative bias can be inferred. This is the metric that Discover uses to evaluate the disparity in a model.

Total Bias (Wasserstein Distance) is equal to the positive bias plus the negative bias. 5.4% + 0.6% = 6%. It is also equal to the average of the absolute value of the marginal effects across all cutoffs. This is a measure of total bias in both directions – favoring and disfavoring women.

Conclusion

The Wasserstein Distance methodology developed by Discover is an innovation that overcomes a real-world shortcoming inherent to existing metrics while continuing to adhere to the principles of disparate impact theory. It enables more thorough discrimination testing on a model at the time of development and, as such, represents a tool that will enable lenders to minimize the chances that their models will cause discrimination before they are put into production. Of course, being able to minimize the chance of discrimination before it occurs is far better and more effective than trying to remedy it after it occurs. By being an effective tool based on traditional approaches to discrimination testing, the Wasserstein Distance methodology represents a conservative, incremental development that has the potential to increase the algorithmic fairness of models deployed in the real world – even when measured by already accepted metrics.