Identifying Drivers of Discrimination in Machine Learning Models

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

Introduction

In a series of papers, Discover Financial Services (Discover) explored the use of the Wasserstein Distance – a measurement of the distance between two probability distributions – for detecting, understanding, and mitigating discrimination in predictive algorithms. BLDS, LLC is writing three companion pieces intended for a broader audience on the use of the Wasserstein Distance for these three use cases.

The first paper, Measuring Discrimination Under Model Usage Uncertainty, explained the mechanics of what the Wasserstein Distance is, how it is used to measure disparity, and the advantages it has over existing metrics. This paper explores how the Wasserstein Distance can be used to determine what is driving disparity in a model. The final paper will show how the Wasserstein Distance can be used to mitigate disparity that may be found in a model.

Measuring Discrimination Under Model Usage Uncertainty

As a metric for assessing whether there is evidence of discrimination, the fundamental insight of Discover's work shows that the Wasserstein Distance is a useful measure of expected model disparity when a model's specific use in production has not been determined. This is a particularly powerful innovation because, more often than not, how a model will be used is not known at the time of development and testing. This can be for a number of reasons, such as having multiple use cases or changing use cases due to evolving economic conditions or business strategies. This fundamental uncertainty undermines the efficacy of existing disparity metrics, which all require prior knowledge of how the model will be used.

Discover solves this problem by developing a methodology that aggregates any disparity that occurs across all possible use cases of a model. This makes it possible to assess the disparity of a model across the entire domain of use cases, and to consider the possible tradeoffs without specifying specific thresholds as existing disparity metrics require. More details are available in the Measuring Discrimination Under Model Usage Uncertainty paper. For the purposes of this paper, the following concepts are important.[1]

Protected Group is a group that has historically been subject to discrimination along a prohibited basis, such as sex, ethnicity, or age. Examples include women, racial minorities, and older people.

Reference Group is the group that a protected group is compared to when testing for discrimination on a prohibited basis. For example, women are compared with men, minorities with the non-minority group, and those who are older versus those who are younger.

Total Bias (Wasserstein Distance) is a measure of the most efficient way to change the protected group's scores to match the corresponding reference group's scores. It is a measure of total bias, regardless of whether the bias is helping or hurting the protected group.

Positive Bias is the portion of the total bias (or Wasserstein Distance) where the protected group is harmed. Thus, it is a measure of how much the protected group's scores must be moved in the favorable direction to match the reference group's scores.

Negative Bias is the portion of the total bias (or Wasserstein Distance) where the protected group is favored. Thus, it is a measure of how much the protected group's scores could be moved in the non-favorable direction to match the reference group's scores.

Net Bias is the positive bias minus the negative bias. It is particularly useful when it is compared to the total bias since the amount of offsetting due to negative bias can be inferred. This is the metric Discover proposes to evaluate the disparity of a model. In practice, Net Bias frequently coincides with Total Bias since Negative Bias is often zero or negligible.

These definitions and the associated methodology for disparity testing are extended here to gain insight into which features in a model are driving disparity.

[1] The papers by Discover Financial Services use the term "bias" to mean a result that has an unfavorable impact on a protected class. In regulatory and legal settings, this would more commonly be referred to as an "adverse impact" or "disparate impact." Further, in legal and regulatory settings, bias is typically differentiated from adverse/disparate impact in that, when one defines a model as being biased against a protected class, it means that it not only causes an adverse/disparate impact, but it does so at a rate that is higher than the underlying data suggests should occur.

For example, suppose the true default rate for a model is 10% for men and 15% for women. Suppose further that the model predicts that only 8% of men will default, but 17% of women will default. This 2% overprediction for women and 2% underprediction for men would be called "model bias" that disfavors women.

While it is essential to understand that these are two distinct concepts, in this paper, we adopt Discover's use of the term "bias" to mean adverse/disparate impact.

How Data Can Cause Algorithmic Discrimination

Before describing how discrimination is measured, it is important to recall how a model uses information in order to generate predictions, so that the meaning of discrimination in this context is properly understood. Fundamentally, a model is a tool that uses the relationships between variation in features and variations in the true outcome in a dataset to make predictions when the outcome of interest is not known. In traditional regression models, this relationship takes the form of a regression coefficient which captures the change in the prediction as a result of differences in one observation's feature values relative to another observation's feature values. In more sophisticated machine learning model architectures, the process of incorporating information about features to predict an outcome is more complex, but the purpose is the same.

However, when a feature correlates strongly and unfavorably with protected class status, the inclusion of this feature in a model often has the effect of causing unfavorable disparities in a model's predictions for that protected class. The challenge of identifying and mitigating bias is therefore understanding whether a model's predictions are causing disparities, and then rooting out the causes of those disparities. In traditional regression models, this is a relatively straightforward process. However, in machine learning models that can use features in complex and sometimes hidden ways, determining how a feature's relationship to protected class ultimately feeds through to the predictions can be extremely difficult and requires more complex analyses to achieve a good understanding of these relationships. As described below, Discover's work with Shapley Values, Owen Values, and the Wasserstein metric provides effective tools for identifying which features cause disparities so that, ultimately, discrimination caused by the model can be mitigated.

Shapley and Owen Values

Shapley values are a powerful and widely used explainable AI (xAI) tool for understanding how a model uses data to create predictions.[2] Shapley Values originated in the Nobel Prize-winning work of Lloyd Shapley, who aimed to find the fairest and best way to allocate the profits of a business to its stakeholders.

To illustrate how the approach works, suppose that three people, A, B, and C, start a company. At the end of the year, they have $1 million in profit to distribute. The Shapley Value approach

[2] Shapley values were originally developed out of game theory but have been applied in various contexts. Because the application of Shapley values to machine learning is an adaptation from their original purpose, various techniques for applying them have been developed. Research into these various techniques has led to an important distinction. One approach known as the marginal explanation (sometimes interventional explanation) is focused on the particulars of how a specific model generates predictions from data. For this reason, it is also known as the *true to the model* approach. A second approach known as the observational explanation or conditional explanation is focused on the relationships between all features in a dataset and the label that is being predicted. For this reason, it is also known as the *true to the data* approach. Although these distinctions are highly technical, they are important for Discover's mitigation methodology, which will be reviewed in a subsequent paper.

suggests that the fairest way to determine how much of the $1 million each person should get is equal to what the profits would be assuming that the person joined the company, less the profits that the company would have earned if that person had not joined the company. While there are multiple ways to do this, the Shapley approach suggests estimating what the profits would have been under all possible combinations of founders. Here, this includes the combination (or "coalition") of all three founders: A+B+C (reality); coalitions of just two founders: A+B, A+C, and B+C; and then just one founder each: A alone, B alone, or C alone (the profit with no founders is assumed to be zero). The distribution of profits can then be calculated for each person. For person A, this is equal to a certain weighted average of profits of the combinations including person A, which are A, A+B, A+C, A+B+C, minus the weighted average of profits of combinations excluding A, which are B+C, B, and C alone.

How does this relate to machine learning predictions? Conceptually, we replace profits with the observation's prediction, and we replace each person with each variable used in the model. In this case, the Shapley values allow one to ask, how much did each feature contribute to the difference between observation's prediction and the dataset's average prediction?

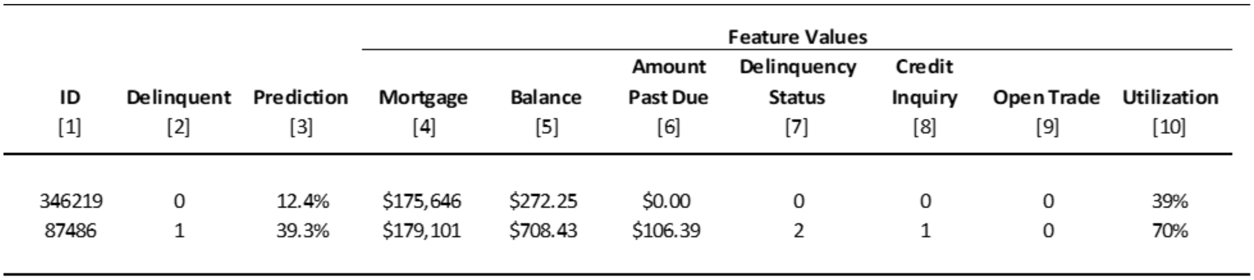
To illustrate, suppose a person has a default prediction of 10% based on a model with two variables (or "features" to use machine learning parlance), Time on File and Number of Delinquencies. Further suppose that the average probability of default in the dataset is 2%. We would like to know how much each of those features contributed to driving the person's prediction from the 2% average up to 10% (i.e., an increase of 8%). While the mechanism by which this occurs is complex, the Shapley Value for Time on File would be equal to the average model score for models that include Time on File (i.e., Time on File alone and Time on File + Number of Delinquencies) minus the average model score for models that exclude Time on File. If we perform this calculation and find that the Shapley Value for Time on File is 5%, then the Shapley value method of feature attribution suggests we can attribute 5% of the 8% increase to Time on File. Correspondingly, we would find that the remaining 3% of the increase was due to Number of Delinquencies.

In summary, Shapley Values reveal, for each observation in the data, how much each feature contributes to the model's prediction for that observation. The contributions for all features for an observation plus a fixed constant term sum to the model's prediction for that observation. They provide useful insight into which features have a large contribution versus a small contribution to a prediction and which features increase versus decrease a prediction.[3]

[3] An alternative, but closely related, approach to Shapley values is Owen values. At a high level, Owen values account for the fact that features may occur in "groups" of interrelated features. All of the ensuing discussion in this paper will be in terms of Shapley values; however, all of the discussion is equally valid for Owen values. Elements of Discover's methodology are predicated on the use of Owen values, but these represent an important technical improvement that is outside of the scope of this paper.

In order to understand how Shapley values work in practice, we created a hypothetical credit default model using synthetic data designed to closely mimic real credit data.[4] The model included seven features, listed below in Table 1. The model was built using an XGBoost machine learning model. XGBoost is likely the most commonly used machine learning model architecture for credit default modeling at this time.

In Table 1 below, we see the input records for two observations. Column [2] shows that the first person in the data did not become delinquent, but that the second person did. Columns [4] through [10] show the individuals' values for the features that went into the model. Noteworthy is that the second person did have prior delinquencies (Delinquency Status = 2 and Amount Past Due = $106.39). The second person also has a much higher trade line utilization percentage than the first person (70% versus 39%). One expects that these factors will lead to a higher prediction of delinquency for the second person. Column [3] shows this is true: the model predicted the first person as having a 12.4% chance of becoming delinquent, but the second person has a 39.3% chance of becoming delinquent.



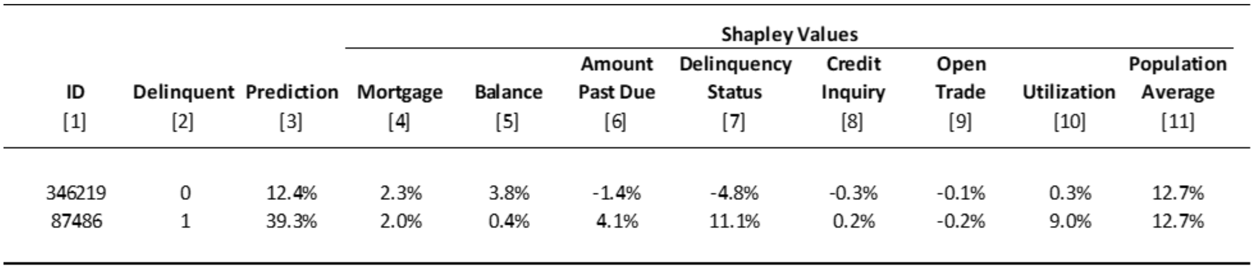
			Feature Values						
ID [1]	Delinquent [2]	Prediction [3]	Mortgage [4]	Balance [5]	Amount Past Due [6]	Delinquency Status [7]	Credit Inquiry [8]	Open Trade [9]	Utilization [10]
346219	0	12.4%	$175,646	$272.25	$0.00	0	0	0	39%
87486	1	39.3%	$179,101	$708.43	$106.39	2	1	0	70%

Table 1: Model Feature Values and Predictions for Two Applicants

But how do each of these variables contribute to the prediction? In order to answer this, we can calculate the Shapley values for each applicant.

The Shapley values for the two observations in Table 1 are given below in Table 2. There are a few important elements Shapley values introduce that help to explain the way that the features contribute to the predictions. First, Shapley values provide a value that, when combined with the average (constant) value itself (Column [11], 12.7%), equals the value of the prediction. This additive aspect allows model developers to decompose their model's predictions into their constituent parts. In this example, what this reveals is the decisive role of the Delinquency Status (Column [7], 11.1%) and Utilization (Column [10], 9.0%) variables in increasing the predicted probability of delinquency for the second applicant. Taken together, these two features more than double the predicted probability of default for this applicant as compared to the baseline average of 12.7%. In this way, Shapley values provide a formula for 'explaining' feature importance for each variable, for each observation.

[4] The original version of the source data was created by Wells Fargo, for use in the Python Interpretable Machine Learning (PiML) package (see https://github.com/SelfExplainML/PiML-Toolbox). The version of the dataset used here was slightly modified and is available in the SolasAI GitHub repository (see https://github.com/SolasAI/solas-ai-disparity).



			Shapley Values							
ID	Delinquent	Prediction	Mortgage	Balance	Amount Past Due	Delinquency Status	Credit Inquiry	Open Trade	Utilization	Population Average
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
346219	0	12.4%	2.3%	3.8%	-1.4%	-4.8%	-0.3%	-0.1%	0.3%	12.7%
87486	1	39.3%	2.0%	0.4%	4.1%	11.1%	0.2%	-0.2%	9.0%	12.7%

Table 2: Explaining Individuals' Predictions Using Shapley Values

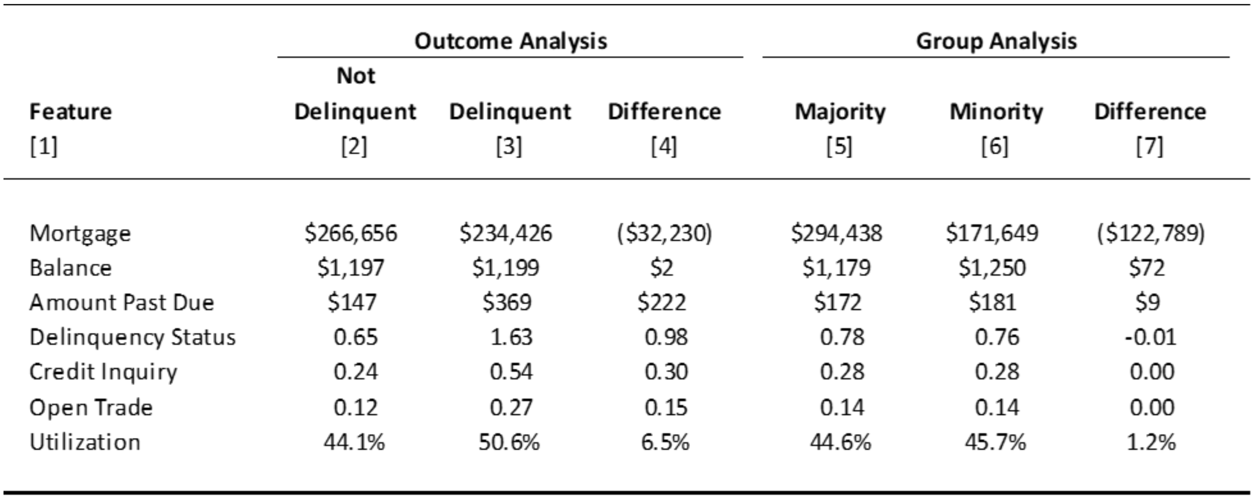
By providing insight into the prediction process for a model, Shapley values help to inform the relationship between the inputs and outputs of a model. What remains to be done is to link this important tool to Discover's methodology around disparity testing. Additionally, making this link robust to cases where a model's exact usage is indeterminate provides a vital anchor for developers to quantify the contributions of individual features to aggregate disparities that a model may generate. In the next section, we will see how Discover connects a thread between Shapley values and the Wasserstein Distance to solve these challenges.

Explaining Bias

Although Shapley values are generated on an individual basis to explain how the features for a single observation lead to the prediction for that observation, they can be aggregated within groups to explain differences in outcomes between groups. We will consider the Mortgage feature as an example. Due to the monotonic constraints[5] imposed on the model, we know that the mortgage balance and the likelihood of default will be negatively correlated. That is, all other features being held constant, applicants with higher mortgage balances will receive lower predictions of the likelihood of delinquency (and thus will receive more favorable outcomes). This can also be seen in the Outcome Analysis section of Table 3. The average mortgage balance for those who are not delinquent is $32,230 greater than the average for those who are delinquent.

As can be seen in Table 3, the average amount of outstanding mortgage loans for the Majority group is $294,438, while the average for the Minority group is $171,649. Thus, it is reasonable to conclude that this feature will contribute to more favorable outcomes for the Majority group relative to the Minority group (and thus increase disparities in outcomes). However, we cannot quantify the size of this contribution from these summary statistics alone. That is, we cannot conclude whether the Mortgage feature is a significant contributor to disparities or not.

[5] Monotonic constraints are a technique used in machine learning by which a modeler can enforce either a positive or a negative relationship between a variable and what is being predicted. In this example, a negative relationship between the mortgage balance and the likelihood of default is enforced. Thus, if the mortgage balance increases (and all other variables remain the same), the predicted probability of default can only decrease or remain the same.



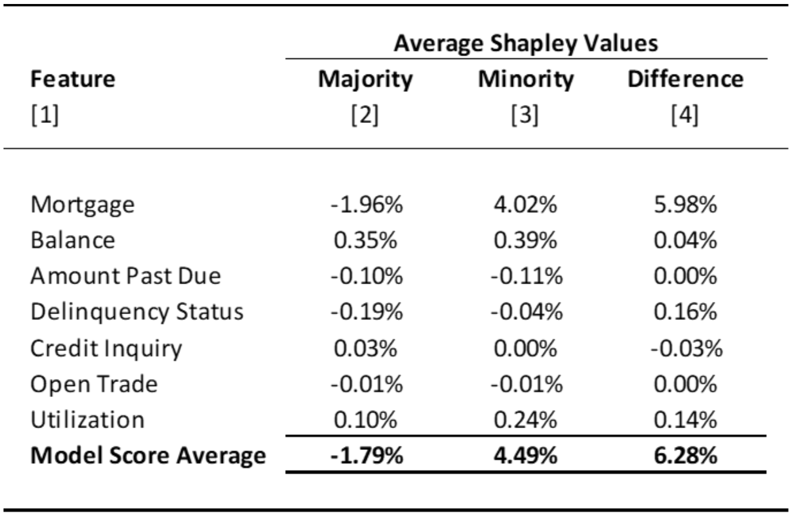
	Outcome Analysis			Group Analysis		
Feature [1]	Not Delinquent [2]	Delinquent [3]	Difference [4]	Majority [5]	Minority [6]	Difference [7]
Mortgage	$266,656	$234,426	($32,230)	$294,438	$171,649	($122,789)
Balance	$1,197	$1,199	$2	$1,179	$1,250	$72
Amount Past Due	$147	$369	$222	$172	$181	$9
Delinquency Status	0.65	1.63	0.98	0.78	0.76	-0.01
Credit Inquiry	0.24	0.54	0.30	0.28	0.28	0.00
Open Trade	0.12	0.27	0.15	0.14	0.14	0.00
Utilization	44.1%	50.6%	6.5%	44.6%	45.7%	1.2%

Table 3: Model Feature Values by Delinquency Status and Protected Group Status

One appealing way to quantify a feature's contribution to the bias is, for example, to consider the difference in means of the Shapley values for the Mortgage feature between the Majority group and the Minority group, which informs us of how much the Mortgage feature contributes to the average prediction for each group. As can be seen below in Table 4, the mean Shapley value for the Majority group is -1.96% while it is 4.02% for the Minority group. Using this information, we can move beyond the information conveyed in the summary statistics and directly quantify the size of the effect that the Mortgage feature is having on bias in predictions. Specifically, the Mortgage feature contributes an average difference of 5.98 percentage points to the predicted likelihood of delinquency between the Majority and Minority groups.

The sum of average Shapley values for the Majority group is -1.79% while it is 4.49% for the Minority group (yielding an average difference of 6.28%). Thus, we know that the average prediction for the Minority group will be 6.28 percentage points higher than the average for the Majority group. We further know that virtually all of this bias comes from the Mortgage feature.

The above method is simple, as it only uses the average difference between the Shapley values distributions for the Majority and Minority groups. A more informative way, as we will see below, is to quantify the bias by assessing the distributional distance between the Shapley values of the two groups.



Feature	Average Shapley Values		
	Majority	Minority	Difference
[1]	[2]	[3]	[4]
Mortgage	-1.96%	4.02%	5.98%
Balance	0.35%	0.39%	0.04%
Amount Past Due	-0.10%	-0.11%	0.00%
Delinquency Status	-0.19%	-0.04%	0.16%
Credit Inquiry	0.03%	0.00%	-0.03%
Open Trade	-0.01%	-0.01%	0.00%
Utilization	0.10%	0.24%	0.14%
Model Score Average	**-1.79%**	**4.49%**	**6.28%**

Table 4: Average Shapley Values for Majority and Minority Applicants

Using Model Explanations and the Wasserstein Distance to Understand Bias

The example above merely dealt with average Shapley values for a single feature in order to develop an intuition for how Shapley values can be used to quantify the bias caused by features. A much richer analysis is possible because (1) the Shapley values for each feature plus the population’s average default rate sum to the model’s predictions, and (2) the Shapley values for each group can be considered as a distribution. These considerations allow Discover to pinpoint how much positive, negative, and net bias each feature causes within a model. This leads to three classes of features: those that mainly favor the Majority group (when only the positive bias explanation is significant), those that favor the Minority group (when only the negative bias explanation is significant), and those of mixed type (when both positive and negative bias explanations are significant, which means that for different thresholds the feature impacts the model bias differently).

Because a model prediction can be decomposed into the sum of Shapley values for all features plus the population’s average default rate, one can calculate the Wasserstein Distance between the Shapley values for a protected group and those of a reference group in order to assess the sources of bias in the model. Thus, all the analysis of total, net, positive, and negative bias measurements discussed in the first paper (and above) can be applied to the Shapley values for features as well. This key insight by Discover allows one to move beyond simply measuring or detecting discrimination under model usage uncertainty to understanding the sources of that discrimination.

When the analysis of total, net, positive, and negative bias is applied to feature contributions, it yields insight into which features are harming, benefiting, or both harming and benefiting a protected group. In the contribution analysis in Table 5 below, the Mortgage feature is composed solely of positive bias – thus harming the protected group relative to the reference group across the entire distribution of contributions. All other features (with the exception of Open Trade, which has negligible total bias) have a mix of positive and negative bias. In addition, all the features except Credit Inquiry have a positive net bias – indicating that Credit Inquiry is the only feature that reduces the net bias experienced by the protected group.

It is worth noting that the trained model is positively biased as the negative model bias component is zero. For this reason, the total net bias (6.28%) matches the average difference in Shapley values in Table 4. This is expected since the Wasserstein bias measures in Table 5 are merely a richer decomposition of the average difference in Shapley values from Table 4. The decomposition allows us to see the positive and negative biases, whereas the average differences in Shapley values only allow us to see the net bias.

	Wasserstein Bias Measures			
Feature	**Total**	**Positive**	**Negative**	**Net**
[1]	[2]	[3]	[4]	[5]
Mortgage	5.98%	5.98%	0.00%	5.98%
Balance	0.60%	0.32%	0.28%	0.04%
Amount Past Due	0.49%	0.24%	0.24%	0.00%
Delinquency Status	0.77%	0.46%	0.31%	0.16%
Credit Inquiry	0.07%	0.02%	0.05%	-0.03%
Open Trade	0.01%	0.01%	0.00%	0.00%
Utilization	0.30%	0.22%	0.08%	0.14%
			Model Outcome's Net Bias	**6.28%**

Table 5: Decomposition of Model Bias by Feature

Application

When disparity is detected in a model’s predictions, applying this Wasserstein Distance analysis to feature contributions allows Discover to determine which features drive that disparity. Specifically, features can be ranked in terms of which features contribute the most to positive bias (and thus increasing disparity) and which features contribute the most to negative bias (and thus decreasing disparity). It also allows Discover to determine which features have both positive and negative biases. When this occurs, it means that under different model thresholds the feature could cause either an increase or decrease in the disparity. Understanding how each feature contributes to the overall bias helps guide the bias mitigation strategy.

In particular, this feature-level analysis of bias can be used in the search for less discriminatory alternatives. For example, a feature with a high level of positive bias could potentially be removed or replaced with a feature with a lower level of positive bias. Alternatively, feature engineering, such as binning or winsorizing, could be performed on a feature with a high level of positive bias in order to generate a derivative feature with a lower level of positive bias. The best way to mitigate bias varies with respect to the context, but access to a robust and intuitive methodology for identifying how a model's features contribute to its predictions undergirds each of them.

Relationship to Existing Best Practices

Shapley values (and, by extension, Owen values) are an established explainable AI tool widely used in industry to analyze models and are understood by regulators. They have been used to gain insight into what is driving a model's quality, to generate adverse action notices for individual applicants, and to understand what is driving disparity in a model.

Discover's application of the Wasserstein Distance to Shapley values is an innovation with regard to this last purpose: understanding the sources of model disparity. Their methodology allows them to identify how much each feature exacerbates and mitigates disparity and contributes to overall model disparity. It has the added benefit of being directly related to the Wasserstein Distance based disparity metric that Discover has developed and implemented.

Of course, the ultimate goal of detecting and measuring the sources of bias is to mitigate the disparity. Having discussed the Wasserstein metric in the first paper, and detailed how Discover's methodology can be leveraged for this purpose in this paper, their approach towards mitigating model bias will be covered in a third and final paper. These three papers together encapsulate the process of quantifying group level disparities, explaining feature-level contributions to that disparity, and mitigating the disparity, respectively.

Conclusion

When disparity is found in a predictive model, it is essential to analyze the sources of that disparity. By applying the Wasserstein Distance to distributions of explanation values between protected and non-protected classes, Discover has developed a powerful methodology for determining and quantifying the sources of disparity in a model. This approach fits squarely into existing disparate impact testing and less discriminatory alternatives search best practices.