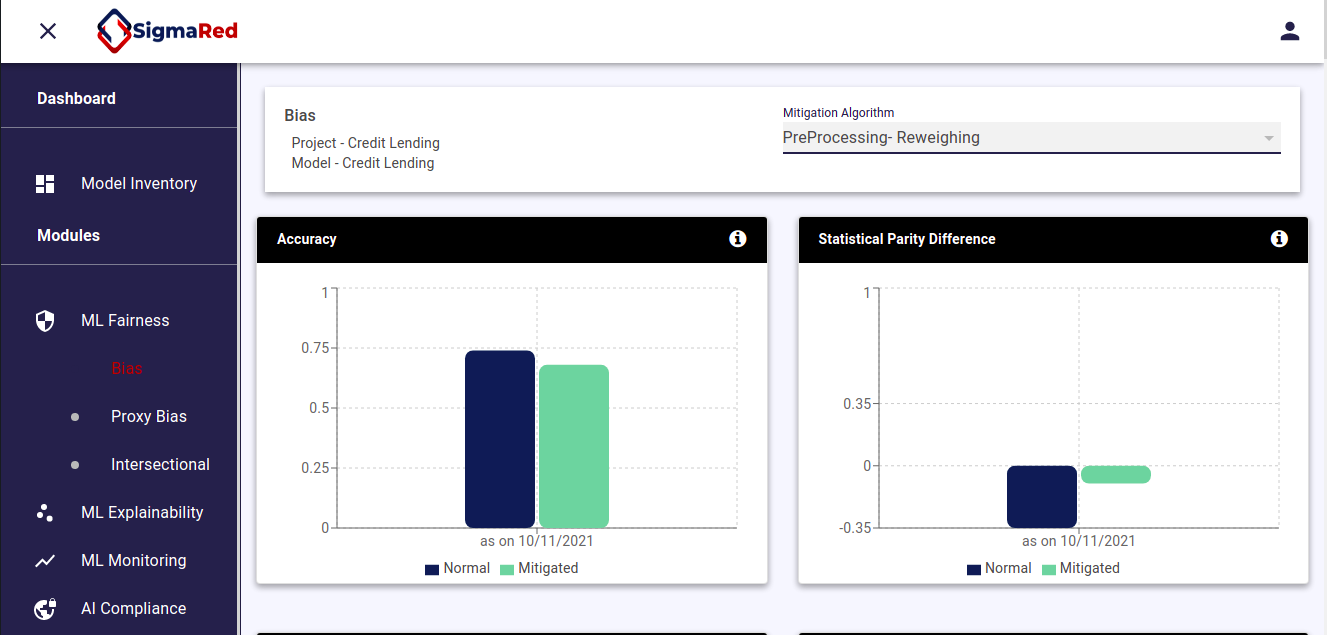
**Manual for Fariness Dashboard**

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| SL No. | Section | Sub-Section | Page No. |
| 1. | Algorithm Overview | PreProcessing- Reweighing | 3 |
|  | | PreProcessing- Disparate Impact Remover | 4 |
| InProcessing- Prejudice Remover | 4 |
| InProcessing- GerryFair Classifier | 5 |
| InProcessing- Adversarial Debiasing | 5 |
| PostProcessing- Reject Option Classifier | 5 |
| PostProcessing- Callibrated EquiOdds | 6 |
| PostProcessing- Adjusted Reject Option Classifier | 6 |
| InProcessing- Custom Adversarial Debiasing | 7 |
| 2 | Proxy Bias | Adjusted Mutual Info Score | 8 |
|  | | Normalised Mutual Info Score | 8 |
| Correlation plot details | 9 |
| Proxy bias mitigation overview | 10 |
| 3. | Deep Evaluation | Overall Evaluation | 11 |
|  | | Intersectional Evaluation | 11 |
| 4 | Metrics Details | Terminologies | 12 |
|  | | AI Bias Page Metrics Details | 12-17 |
| Additional Metrics details | 17-23 |

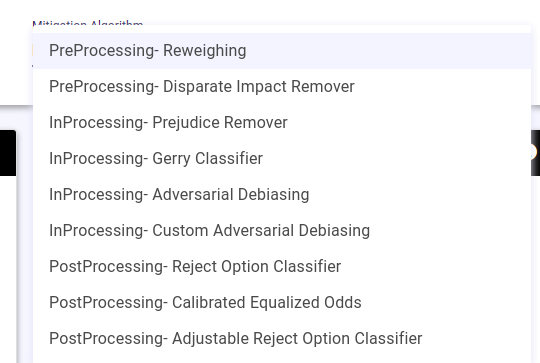
The AI Mitigation Dashboard can be found under ML Fairness section. The Overview of the mitigation section of the dashboard is as shown below. In order to perform mitigation, you can select any algorithm from the dropdown menu in “Mitigation Algorithm”.

**Bias Mitigation Dashboard:**



**Dropdown menu (selection of algorithms):**

Below is the enlarged picture of the dropdown menu for selection of algorithms. Based on the necessity, the respective algorithm can be chosen.



The below explains briefly each algorithm and the bias metrics before and after mitigation

1. **Reweighing:-**

1. Reweighing is the pre-processing technique (only working on dataset and model)

2. In this algorithm we need protected feature (gender/race) as a input variable to calculate the sample weights for model.

3. After getting all the sample weights for each record we will use this sample weight while training the model.

4. Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.72 | More the better |
| statistical parity difference | -0.35 | -0.1 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | -0.02 | -0.10 to + 0.15 |
| average odds difference | -0.28 | -0.02 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 0.80 | 0.85 to 1.15 |
|  |  |  |  |

**2. Disparate Impact Remover:-**

1.Disparate impact remover is the pre-processing technique (only working in dataset)

2. In this algorithm we can pre-process the dataset based on statistical analysis for each feature without taking the protected attribute in consideration and changing the distribution of each feature in a such a way that model will not give debiased results after introducing the protected attribute

3. In this algorithm how much change in features is in our hands.

Eg- if we only want to change the features by 20% then we can only change the feature values by 20% and pass it to the model

4. Results: :

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.72 | More the better |
| statistical parity difference | -0.35 | -0.1 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | -0.03 | -0.10 to + 0.15 |
| average odds difference | -0.28 | -0.02 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 0.81 | 0.85 to 1.15 |

**3. Prejudice Remover:-**

1. Prejudice Remover is the In-Processing algorithm. The input of this method is dataset and prejudice remover will give the prediction values

2. Prejudice remover is the regularizer technique which will be helpful to reduce the bias from the model

3. Results

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.66 | More the better |
| statistical parity difference | -0.35 | -0.21 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | -0.45 | -0.10 to + 0.15 |
| average odds difference | -0.28 | -0.28 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 0.1 | 0.85 to 1.15 |

**4. GerryFair Classifier :-**

1. Gerryfair classifier is the In-Processing algorithm. The input of this method is dataset and GerryFair will give the prediction values. GerryFair is the model itself

2. For this classifier we need to select the metrics (like false positives, true positives) to optimize the GerryFair classifier optimisation function for getting debiased model .

3. Results

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.47 | More the better |
| statistical parity difference | -0.35 | 0.0 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | 0.0 | -0.10 to + 0.15 |
| average odds difference | -0.28 | 0 | -0.85 to + 0.15 |

**5. Adversarial Debiasing:-**

1. Adversarial debiasing is the In-processing technique. This technique is based on deep learning.

2. In this technique we are following the GAN’s approach for building the debiased model

3. In this method we are making sure that features are not maintaining any correlation with protected attribute while training.

4. Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.71 | More the better |
| statistical parity difference | -0.35 | -0.27 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | -0.6 | -0.10 to + 0.15 |
| average odds difference | -0.28 | -0.37 | -0.10 to + 0.15 |

**6. Reject Option Classifier :-**

1. Reject Option Classifier is Post Processing technique (Working on model prediction values (predicted probability values))

2. In this approach we have regions where we can test each prediction for protected attributes. Based on the results ROC fine tuning the prediction values and give the debiased prediction values

3. Result :-

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.71 | More the better |
| statistical parity difference | -0.35 | -0.03 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | 0.02 | -0.10 to + 0.15 |
| average odds difference | -0.28 | -0.04 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 0.93 | 0.85 to 1.15 |

**7. Calibrated EquiOdds :-**

1. Calibrated EquiOdds is the postprocessing technique (working in model prediction)

2. It calibrates the prediction probability based on equal false positive rate and false negative rate constraint to equalize between privilege and unprivileged groups

3. Result :-

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.5 | More the better |
| statistical parity difference | -0.35 | 0.0 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | 0.0 | -0.10 to + 0.15 |
| average odds difference | -0.28 | 0.0 | -0.10 to + 0.15 |

**8. Adjusted Reject Option classifier :-**

1. Adjusted reject option classifier is the post-processing method. Here we are working on the model prediction probability values.

2. In this method we proposed two critical regions for prediction probability based on unprivileged condition.

3. Results

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.79 | More the better |
| statistical parity difference | -0.35 | 0.16 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | 0.06 | -0.10 to + 0.15 |
| average odds difference | -0.28 | 0.0 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 0.68 | 0.85 to 1.15 |

**9. Custom Adversarial Debiasing:-**

1. This algorithm is in-processing algorithm. Here we are building the deep learning model based on Gan’s approach

2. In this Adversarial debiasing approach we proposed method to customisable loss and sample weight which will work as a lever of adversarial debiasing model architecture. Using customised loss weights and sample weights we are able to generate the debiased model

3. Results :-

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.75 | 0.68 | More the better |
| statistical parity difference | -0.35 | 0.02 | -0.10 to + 0.10 |
| equal opportunity difference | -0.31 | 0.2 | -0.10 to + 0.15 |
| average odds difference | -0.28 | 0.19 | -0.10 to + 0.15 |
| disparate impact ratio | 0.28 | 1.15 | 0.85 to 1.15 |

**10. Proxy feature generation via perturbation of features (mitigation of proxy bias)**

1. This technique is used for mitigating the proxy bias.

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Disparate Impact Ratio | 0.42 | 0.51 | 0.85 to 1.15 |
| statistical parity difference | -0.08 | -0.08 | -0.10 to + 0.10 |
| False positive rate ratio | 0.35 | 0.55 | 0.85 to 1.15 |
| average odds difference | -0.07 | -0.07 | -0.10 to + 0.15 |

**11. Custom Equiodds IP**

1. This technique is post processing used for mitigating the bias.

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics Name | Score before mitigation | Score after mitigation | Acceptable Range |
| Balanced Accuracy | 0.74 | 0.66 | More the better |
| statistical parity difference | -0.22 | -0.09 | -0.10 to + 0.10 |
| equal opportunity difference | -0.15 | -0.17 | -0.10 to + 0.15 |
| average odds difference | -0.16 | -0.11 | -0.10 to + 0.15 |
| disparate impact ratio | 0.46 | 0.37 | 0.85 to 1.15 |

**15. Bias Evaluation And Feature Interaction Detection Using Model Interpretation**

Details: Our inventions are two metrics, one is to detect and quantify the bias in the model, and the other is to find the features and their contribution to the bias.

Results:

Biased logistic regression model: 0.1

Unbiased logistic regression model: -0.0012

Biased random forest model: 0.319

Unbiased random forest model: -0.00121

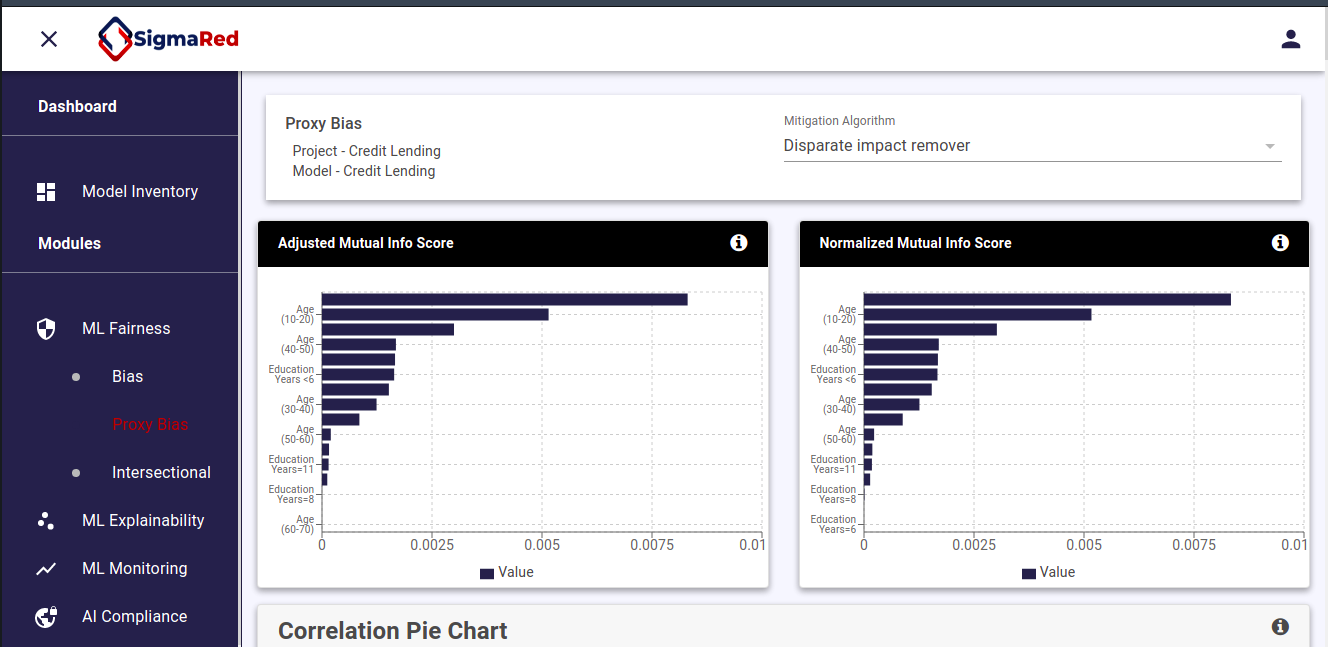
Acceptable range: -0.01 to 0.01.

**AI Proxy Bias Dashboard.**

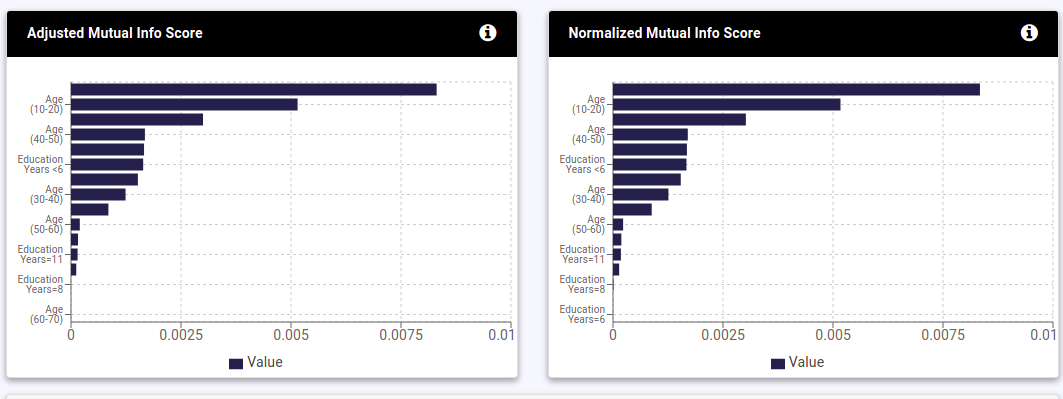
The AI Proxy Mitigation Dashboard can be found under ML Fairness section The dashboard includes bar plots to indicate the potential proxyness of features, using adjusted mutual information and normalized mutual information score, the correlation pie charts based on the sensitive attributes and metrics and standard bias metrics.

Upon finding the proxy features, we can mitigate them by selecting the mitigation algorithm from the

drop down menu.

****

**Proxy Indicator section:**



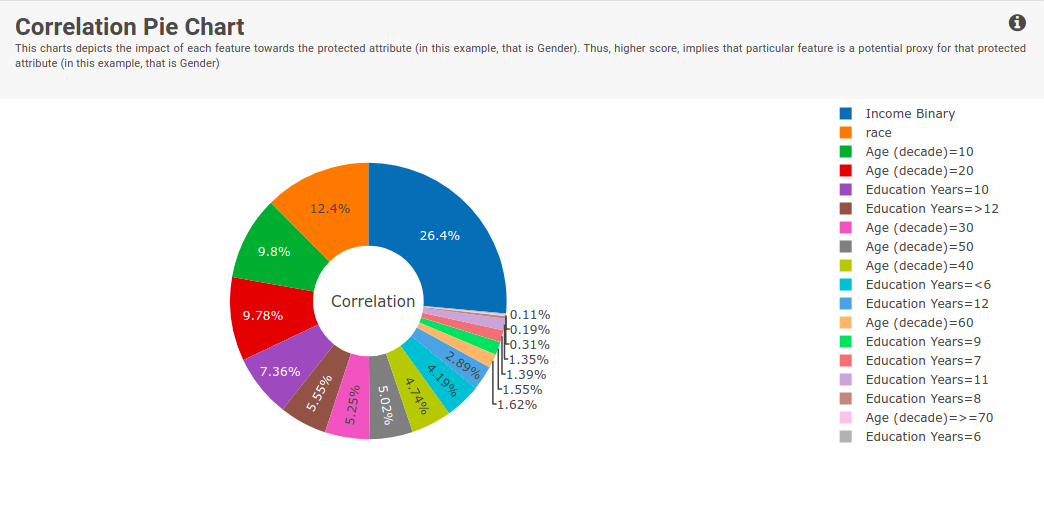
1. Adjusted Mutual Information score.

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters. The strength of proxy is given by this score. We wish to measure how strongly a random feature X is a proxy for protected feature Z

2. Normalised mutual information score

"Normalized Mutual Information (NMI) is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). The strength of proxy is given by this score. We wish to measure how strongly a random feature X is a proxy for protected feature Z.

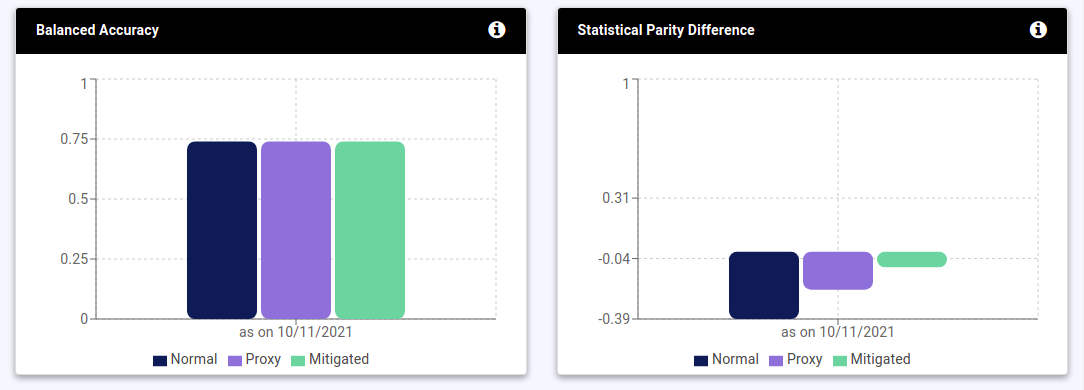
**Proxy correlation pie chart:**



This charts depicts the impact of each feature towards the protected attribute . Thus, higher score, implies that particular

feature is a potential proxy for that protected attribute

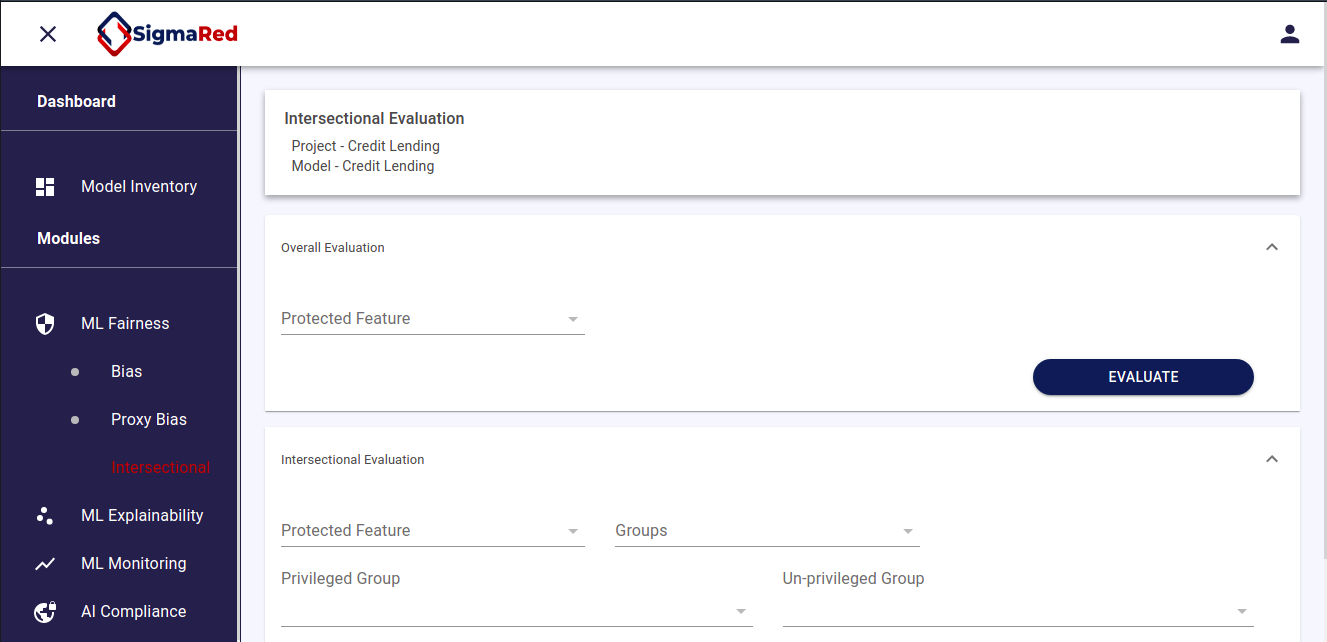
**Overview of some bias metrics:**



**Custom Bias evaluation**

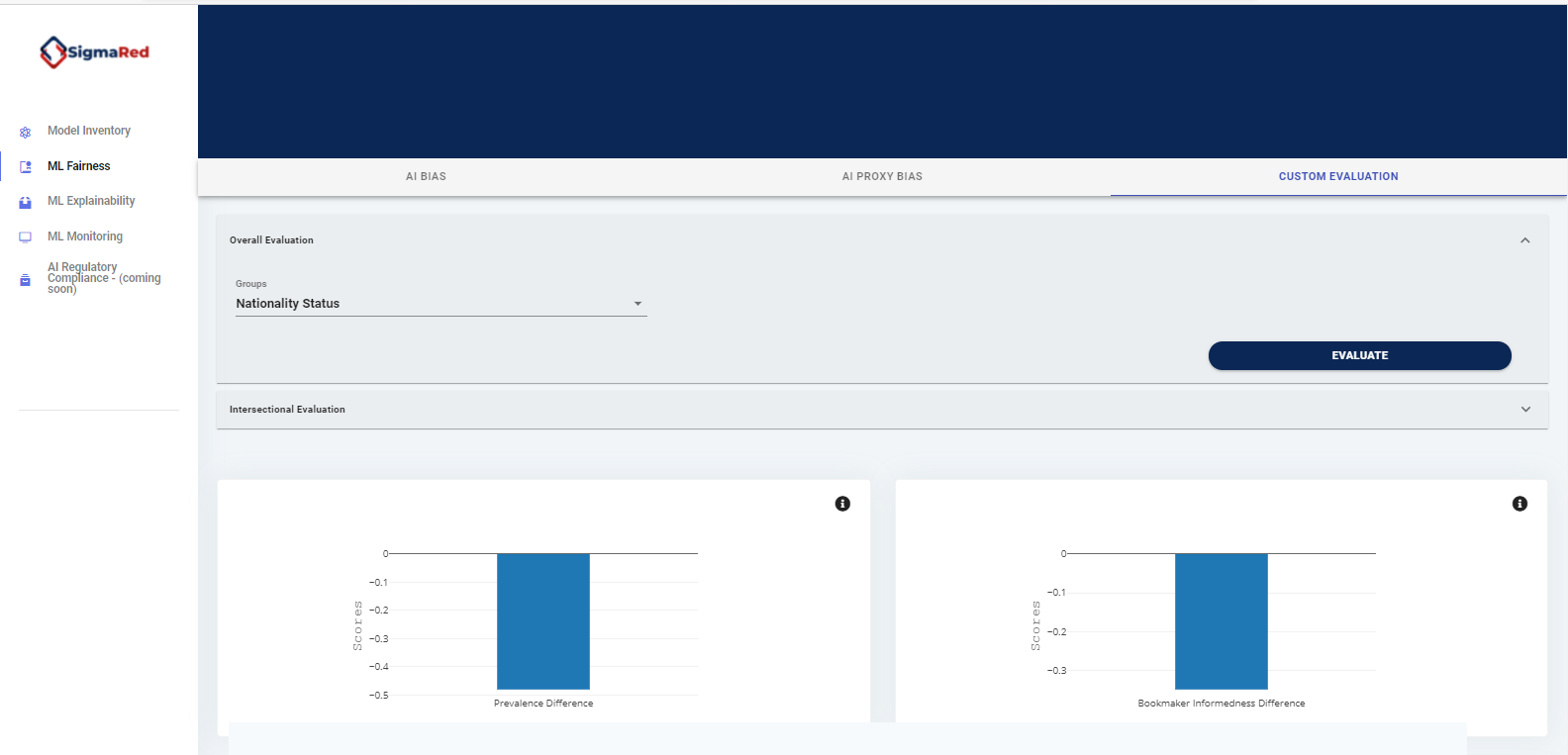
The custom bias evaluation dashboard can be found under the ML fairness section. This contains an overall evaluation and intersectional evaluation. The overall evaluation can be done on a single feature such as nationality, gender etc .. whereas the intersectional evaluation allows in depth analysis on nature of bias for combination of features, eg: Hispanic female in her 20’s

**Below is an overview of the Custom Bias evaluation dashboard:**



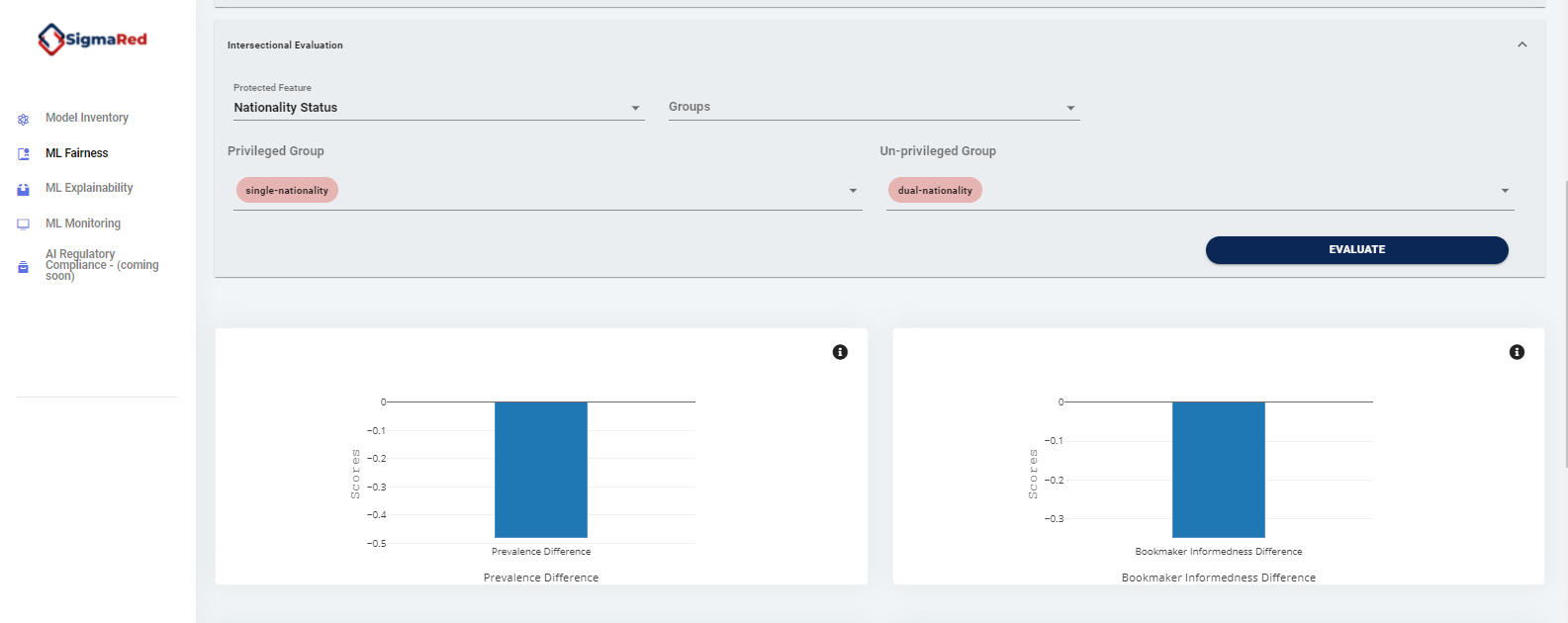
**Overall evaluation section:**

Here we have selected “Nationality status”, and the evaluation is performed on the same.



**Custom intersectional evaluation:**

Here we select Nationality status and Gender(by default) and pick the privileged and unprivileged groups for evaluation



**Metrics Documentation**

**Terminologies:-**

1. Bias: A systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage.
2. Fairness metric: A quantification of unwanted bias in training data or models.
3. Favourable label: A label whose value corresponds to an outcome that provides an advantage to the recipient. The opposite is an unfavourable label.
4. Favourable Classes: - Positive Class
5. Un-Favourable Class: - Negative Class
6. Condition = [{'Gender': 1, 'age': 1}, {'Gender': 0}]:-  (Gender == 1 AND age == 1) OR (Gender == 0)
7. Protected attribute: An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific.
8. Privileged protected attribute: A value of a protected attribute indicating a group that has historically been at systematic advantage.

* Privileged Groups: - Groups which are highly privileged or high importance in protected features. Ex: Race / Ethnicity column, then ‘White’
* Unprivileged Groups: - Groups which are less important than privileged groups.

Ex: Race column, then ‘Black’

\*\*Note: - Selected groups must fall under the protected attributes.

**Descriptions for Metrics**

1. True Positive (TP):- No of data points which are actually positive and predicted as positive

2. False Positive (FP):- No of data points which are actually Negative but predicted as positive

3. True Negative (TN):- No of data points which are actually negative and also predicted as negative

4. False Negative (FN):- No of data points which are actually Positive but predicted as Negative

5. True Positive rate (TPR) (TP /P):- Out of all positives how many predicted as positive.

6. False Positive rate (FPR) (FP/N):- Out of all negative points how many points are falsely predicted as positive

7. False Negative Rate (FNR) (FN /P):- Out of all positive points how many are falsely predicted as negative (Ratio of FN over all the positive points in dataset)

8. True Negative Rate (TNR) (TN/N):- Out of all negative points how many are truly predicted as negative

9. Positive Predicted Values (Precision) (PPV) (TP / TP+FP):- Out of all predicted positive points how many of them are actually predicted correctly

More the scores better the model is.

10. False Discovery Rate (1 - Precision) (FDR) (FP / TP+FP):- Out of all predicted positive points how many of them are actually predicted as false (Falsely predicted data points)

Lower the score better the model

11. False Omission Rate (FOR) (FN /TN+FN):- Out of all predicted Negative points how many of them are actually predicted as false (Falsely predicted data points)

Near to zero is better

12. Negative Predicted Value (NPV) (TN /TN+FN):- Out of all predicted Negative points how many of them are actually predicted correctly

More the scores better the model is.

13. Accuracy (TP+TN / P+N):- More the scores better the model is.

14. Error rate (FP +FN) / (P+N):- [1- accuracy]. Lower the scores better the model

15. True Positive Rate difference (TPR (unprivileged) – TPR (privileged Groups) :-

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group. 

Fairness for this metric is between - 0.1 and 0.1

16. False Positive rate difference (FPR (unprivileged) – FPR (privileged Groups)):-

This metric is computed as the difference of false positive rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group. 

Fairness for this metric is between -0.1 and 0.1

17. False negative rate difference (FNR (unprivileged) – FNR (privileged Groups)):-

This metric is computed as the difference of false negative rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group. 

Fairness for this metric is between -0.1 and 0.1

18. False omission rate difference (FOR (unprivileged) – FOR (privileged Groups)):-

This metric is computed as the difference of false omission rates between the unprivileged and the privileged groups.

FOR :- (FN /TN+FN)

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group. 

Fairness for this metric is between -0.1 and 0.1

19. False Discovery rate difference (FDR (unprivileged) – FDR (privileged Groups) ) :-  FDR  :- (FP / TP+FP )

This metric is computed as the difference of false discovery rates between the unprivileged and the privileged groups.

The ideal value is 0. A value of < 0 implies higher benefit for the Unprivileged group and a value > 0 implies higher benefit for the Privileged group. 

Fairness for this metric is between -0.1 and 0.1

20. False positive rate ratio: - FPR Unprivileged / FPR Privileged

Ratio of the FPR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false positive rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group. 

21. False negative rate ratio: - FNR Unprivileged / FNR Privileged

Ratio of the FNR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false negative rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

22. False omission rate ratio: - FOR Unprivileged / FOR Privileged

Ratio of the FOR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false omission rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

23. False discovery rate ratio: - FDR Unprivileged / FDR Privileged

Ratio of the FDR metric of Privileged and unprivileged group.

This metric is computed as the ratio of false discovery rates between the unprivileged and the privileged groups.

The ideal value is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the Privileged group.

24. Average Odds Difference:-

* Computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.
* ((FPR unprivileged – FPR privileged) + (TPR unprivileged – TPR privileged)) \* 0.5.
* The ideal value of this metric is 0.
* A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.
* Fairness for this metric is between -0.1 and 0.1

25. Average Absolute Odds Difference: -

(abs (FPR unprivileged – FPR Privileged ) + abs(TPR unprivileged – TPR Privileged ) ) \* 0.5

26. Error Rate Difference:- Error rate (Unprivileged) – Error Rate (Privileged)

The ideal value of this metric is 0. A value of < 0 implies higher benefit for the unprivileged group and a value > 0 implies higher benefit for the privileged group.

27. Error rate ratio:-Error rate (Unprivileged) / Error Rate (Privileged)

The ideal value of this metric is 1. A value of <1 implies higher benefit for the Unprivileged group and a value >1 implies higher benefit for the privileged group.

28. Selection rate (Fp+TP)/ Total\_no\_of\_points:-

Out of all points, how many of them are predicted as positive

29. Disparate Impact:-

* Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.
* No\_Predicted\_positive\_Unprivileged\_Group /  No\_Predicted\_positive\_Privileged\_Group.
* The ideal value of this metric is 1.0 A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.
* Fairness for this metric is between 0.8 and 1.25.

30. Statistical Parity Difference:-

* Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.
* No of Predicted positive Unprivileged Group - No of Predicted positive Privileged Group.
* The ideal value of this metric is 0.
* Fairness for this metric is between -0.1 and 0.1.
* A value < 0 implies higher benefit for the privileged group and a value >0 implies a higher benefit for the unprivileged group.

31. Theil Index:-

* Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1 or 0 most number of times.
* It measures the inequality in benefit allocation for individuals.
* A value of 0 implies perfect fairness.
* Fairness is indicated by lower scores, higher scores are problematic.

1.  np.mean(np.log((b / np.mean(b))\*\*b) / np.mean(b))

[Equation]

where b = 1 + newypred – newytrue

newytrue = dataframe[actual][dataframe[actual] == 1]

    newypred = dataframe[predicted][dataframe[actual] == 1]

2.    elif alpha == 0:

      - np.mean(np.log(b / np.mean(b)) / np.mean(b))

[Equation]

 3. else: (alpha other than 0 or 1)

np.mean((b / np.mean(b))\*\*alpha - 1) / (alpha \* (alpha - 1))

[Equation]

32. Equal Opportunity Difference:-

* This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.
* The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.
* Fairness for this metric is between -0.1 and 0.1.

**AdditionalMetrics for Bias**

1. Prevalence:-

* Returns the proportion of actual positives versus total population.
* The high value is 100% and lower value is 0.
* PREV = ((TP+FN) / (TP+TN+FP+FN))
* Ideally must be 50% only when TP are high and FN are low.

1. Matthews Correlation Coefficient:-

* Returns a reliable statistical rate which produces a high score only if the prediction obtained good results in all of the (tp, fn, tn, fp), proportionally both to the size of positive values and the size of negative values.
* Regarded as a balanced measure which can be used even if the classes are of very different sizes.
* MCC = ((TP\*TN)-(FP\*FN)) / np.sqrt((TP+FP)\*(TP+FN)\*(TN+FP)\*(TN+FN))
* Ranges between (-1 and +1)
* +1 indicates best, 0 indicates random, -1 indicates worst.

1. Fowlkes Mallows Index:-

* Returns how similar the actual values and predicted values are.
* Ranges between (0 and +1)
* FMI = np.sqrt(PPV\*TPR)
* +1 indicates best, 0 indicates worst.

1. Bookmaker Informedness:-

* Returns the performance of a dichotomous diagnostic test.
* Dichotomous diagnostic test – dividing into parts with mutually exhaustive points and estimating the probability of informed decision.
* BMI = ((TPR+TNR)-1)
* Ranges between (0 and +1).
* +1 indicates best, 0 worst.
* BMI = ((TPR + TNR) - 1)
* If FP and FN are only results then we switch the labels.

1. Markedness:-

* Returns the measure of how much one variable is marked as a predictor or cause of another.
* MN = ((PPV+NPV)-1)
* Ranges between (0 and +1)
* +1 indicates best, 0 worst.
* If FP and FN are only results then we switch the labels.

1. Cohen Kappa Coefficient:-

* Returns the degree of agreement.
* Ranges between (-1 and +1)
* +1 indicates best, 0 indicates random, -1 indicates worst.

1. Critical Success Index or Jaccard Index:-

* Returns the prediction performance of correct predictions.  (TP / (TP+FP+FN))
* This is not affected by rejection predictions. (TN)
* Ranges between (0 and +1).
* +1 indicates best, 0 worst.

1. Gilbert Skill Score:-

* This is unbiased version of CSI adjusted for chance.
* Returns the prediction performance of correct predictions. (TP - chance / (TP+FP+FN - chance)), chance = ((TP+FP)\*(TP+FN))/(TP+FP+FN+TN)
* This doesn’t require rejection predictions. (TN)
* Ranges between (-1 and +1)
* +1 indicates best, 0 indicates random, -1 indicates worst.

1. Treatment Equality:-

* Returns the proportion of false positives and negatives.
* TE = FP/FN
* Ranges between (0 and +1).
* +1 indicates best, 0 indicates worst.

1. Positive Likelihood Ratio:-

* Returns the likelihood ratio for positives.
* Likelihood is all about how strong the probabilities are.
* Ranges between (0 and 100) (high TPR is 1 and low FPR can be 0.01).
* 0 indicates no likelihood, 100 indicates very high likelihood.

1. Negative Likelihood Ratio:-

* Returns the likelihood ratio for negatives.
* Ranges between (0 and 100) (high FNR is 1 and low TNR can be 0.01).
* 100 indicates no likelihood, 0 indicates very high likelihood.

1. Diagnostic Odds Ratio:-

* Returns the likelihood ratio for positives and negatives.
* DOR = PLR/NLR
* Ranges between (0 and inf).
* Higher the values, better the results.

1. F-Beta Score:-

* Returns the harmonic mean of precision and recall.
* Beta is mostly 1, called as F1 score. Sometimes can be 2 etc based upon scenario.
* Ranges between (0 and +1).
* +1 indicates best, 0 indicates worst.

1. Brier Score:-

* The Brier Score is a strictly proper scoring rule that measures the accuracy of probabilistic predictions.
* The Brier score is applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes or classes.
* Ranges between (0 and +1).
* 0 indicates best, +1 indicates worst.
* The lower score, better the predictions are calibrated.

1. Jaccard Similarity:-

* Returns how similar are the sets/values are.
* This score may be a poor metric if there are no positives for some samples or classes and is undefined if there are no true or predicted labels.
* Ranges between (0 and +1)
* +1 indicates best, 0 indicates worst.

1. Jensen Shannon Divergence:-

* This is a method of measuring the similarity between the actual and predicted probability distributions.
* It is more useful as a measure as it provides a smoothed and normalized version of KL divergence, with scores between 0 (identical) and 1 (maximally different)
* Positive values mean the label distributions diverge, the more positive the larger the divergence.

1. Theil T Index:-

* Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1.
* Theil's T and is more sensitive to differences at the top of the distribution.
* It measures the inequality in benefit allocation for individuals.
* A value of 0 implies perfect fairness and higher implies unfairness.
* Fairness is indicated by lower scores, higher scores are problematic.

1. Theil L Index:-

* Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 0.
* Theil's L and is more sensitive to differences at the lower end of the distribution.
* It measures the inequality in benefit allocation for individuals.
* A value of 0 implies perfect fairness and higher implies unfairness.
* Fairness is indicated by lower scores, higher scores are problematic.

1. Equal Opportunity Difference:-

* This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.
* The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.
* Fairness for this metric is between -0.1 and 0.1.

1. P-Rule:-

* The rule states that the ratio between the probability of a positive outcome given the protected attribute being true and the same probability given the protected attribute being false is no less than 100%.
* So, when a classifier is completely fair it will satisfy a 100%-rule. In contrast, when it is completely unfair it satisfies a 0%-rule.
* Acceptable range is more than 80%.

1. Adjusted Mutual Info Score:-

* The strength of proxy is given by this score. We wish to measure how strongly a random feature X is a proxy for protected feature Z.
* Adjusted for chance mean higher chance of falling in proper cluster.
* Measured with epsilon parameter typically range between 0.01 and 0.1
* The measure being 1 implies perfect proxies, and 0 implies statistical independence.

1. Customized influence Score:-

* The influence of the decomposition of program p1 with random feature X and program p2 is given by the output of p1 on p2.
* Measured with delta parameter typically range between 0.01 and 0.1
* The influence measure being 1 implies perfect influence, and 0 implies statistical independence.

1. Rand Score:-

* Computes the how much similar between two features random feature X is a proxy for protected feature Z.
* Adjusted for chance mean higher chance of falling in proper cluster.
* The measure being 1 implies perfect similar, and 0 implies dissimilar.

1. Prevalence Difference:-

* Returns the proportion of actual positives versus total population between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Matthews Correlation Coefficient Difference:-

* Returns  a reliable statistical rate which produces a high score only if the prediction obtained good results in all of the (tp, fn, tn, fp), proportionally both to the size of positive values and the size of values between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Fowlkes Mallows Index Difference:-

* Returns how similar the actual values and predicted values between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Bookmaker Informedness Difference:-

* Returns the performance of dichotomous diagnostic test (dividing into parts with mutually exhaustive points and estimating the probability of informed decision) between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Markedness Difference:-

* Returns the measure of how much one variable is marked as a predictor or the cause of another between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Cohen kappa Coefficient Difference:-

* Returns the degree of agreement between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Critical Success Index Difference:-

* Returns the prediction performance equal to the total number of correct predictions between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Gilbert Skill Score Difference:-

* Is similar to Critical Success Index Difference but is adjusted for chance. Returns the prediction performance equal to the total number of correct predictions between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Treatment Equality Difference:-

* Returns the proportion of false positives and false negatives between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Diagnostic Odds Ratio Difference:-

* Returns the likelihood ratio for positives and negatives between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. F Beta Score Difference:-

* Returns harmonic mean of precision and recall between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Brier Score Difference:-

* Is a strictly proper scoring rule that measures the accuracy of probabilistic predictions between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Jaccard Similarity Difference:-

* Returns how similar are the sets/values between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Jensen Shannon Divergence Difference:-

* Is a method of measuring the similarity between the actual and predicted probability distributions between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.

1. Theil T Index Difference:-

* Measures the inequality in benefit allocation for individuals between unprivileged and privileged groups.
* 0 indicates best.
* Greater than 0 inclined towards unprivileged groups and less than 0 inclined towards privileged groups.