A Comparative Analysis of AI fairness Toolkits

Rutuja Padgilwar, Swetha Srihari

November 2023

1 Introduction

In the rapidly evolving landscape of artificial intelligence (AI), the pursuit of fairness has emerged as a critical concern. As AI systems increasingly influence various aspects of our lives, from hiring processes to judicial decision-making, the need for tools that ensure fairness and mitigate bias becomes paramount. In this project, we have tried to do a comprehensive exploration of two of the AI fairness tool-kits currently available, aiming to provide a comparative analysis that sheds light on their functionalities, strengths, and limitations. By delving into the intricacies of these tool-kits, we endeavor to deepen our understanding of how they contribute to the broader goal of fostering equitable and unbiased AI systems.

2 Bias

Bias can creep into algorithms in several ways. AI systems learn to make decisions based on the training data, which can include biased human decisions or reflect historical or societal inequalities. This makes certain elements of a dataset to be over-represented. These biased datasets don't accurately represent the machine learning model's use case, which leads to skewed outcomes. Often, the results generated using biased dataset discriminates against a specific group or groups of people. For example, data bias reflects prejudice against age, race, culture, or sexual orientation. In today's world where AI systems are increasingly used everywhere, the danger of bias lies in amplifying discrimination. It takes a lot of training data for machine learning models to produce viable results. If you want to perform advanced operations (such as text, image, or video recognition), you need millions of data points. Biased data will result in inaccurate predictions because the quality of the outputs is determined by the quality of the inputs.

3 What are AI fairness tools?

AI fairness tools are software and frameworks designed to assess and mitigate biases and unfairness in artificial intelligence (AI) systems. These tools aim to

ensure that AI models and algorithms treat all individuals and groups fairly and avoid perpetuating or exacerbating existing social, economic, or demographic biases. IBM fairness 360 and Aequitas are the two fairness tool-kits that we have explored for this project.

4 IBM Fairness 360

IBM 360 degree toolkit contains a comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets at the pre-processing and model training stages.

4.1 Algorithms

IBM fairness 360 uses ten state-of-the-art bias mitigation algorithms that can address bias throughout AI systems.

Optimized Pre-processing Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing Used to mitigate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing Used to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification Used to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover Used to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations Used to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover Used to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing Used to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing Used to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier Used to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

4.2 Metrics

The following are some of the metrics that measure individual and group fairness.

Statistical Parity Difference The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

Equal Opportunity Difference The difference of true positive rates between the unprivileged and the privileged groups.

Average Odds Difference The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.

Disparate Impact The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

Theil Index Measures the inequality in benefit allocation for individuals.

Euclidean Distance The average Euclidean distance between the samples from the two datasets.

Mahalanobis Distance The average Mahalanobis distance between the samples from the two datasets.

Manhattan Distance The average Manhattan distance between the samples from the two datasets.

4.3 AI Fairness 360 - Demo

In the IBM fairness 360 website, there is a demo integrated with a few datasets. We picked the adult census income data set. It had many sensitive data like race and sex. From the two figures 1 and 2, we can see that the dataset is biased for 2 out of 5 metrices for Race and 4 out of 5 metrices are biased for Sex.

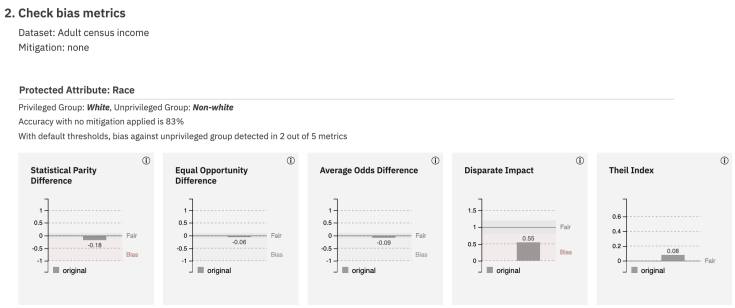


Figure 1: Checking for bias against Race attribute

To mitigate bias, we chose the reweighing algorithm. This algorithm is Used to mitigate bias in training data. Modifies the weights of different training examples. Figures 3 and 4 show the metrics of Race and Sex attributes after mitigation. We can see that one of the two biased metrics have been reduced to acceptable levels for the Race attribute and 2 of the 4 previously biased metrics have been reduced to acceptable levels after mitigation for the Sex attribute .

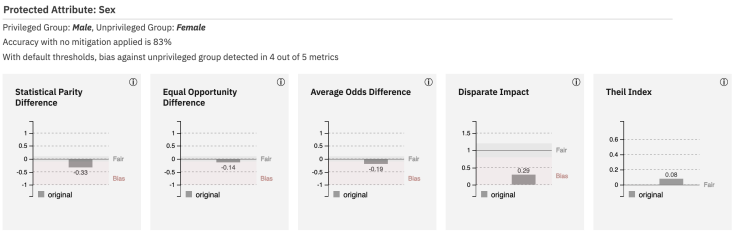


Figure 2: Checking for bias against Sex attribute

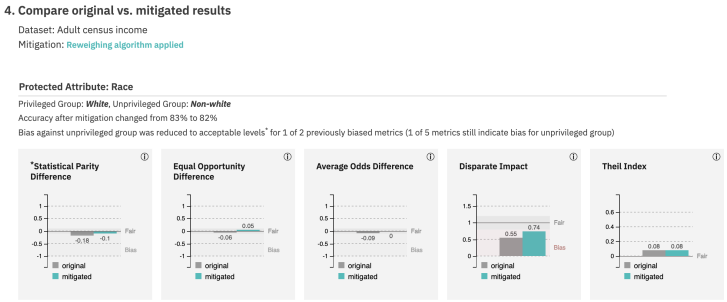


Figure 3: Race attribute after bias mitigation

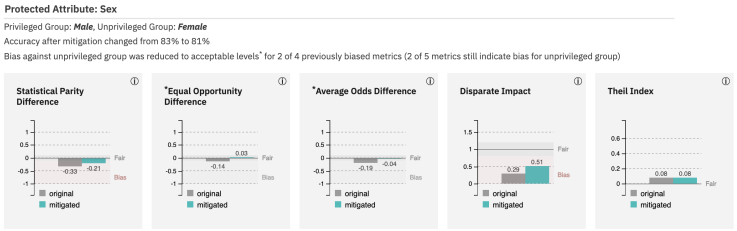
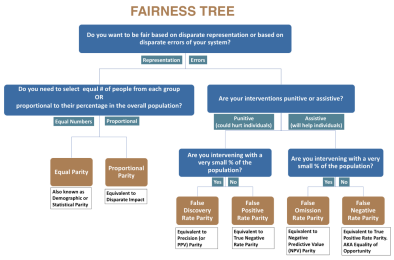


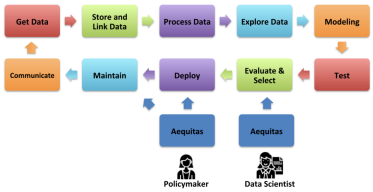
Figure 4: Sex attribute after bias mitigation

5 Aequitas

Aequitas, an open-source bias audit toolkit developed by the Center for Data Science and Public Policy at the University of Chicago, can be used to audit the predictions of machine learning-based risk assessment tools to understand different types of biases and make informed decisions about developing and deploying such systems.

Aequitas, provides the Bias Report, Detailed Fairness and Bias Statistics, and Interactive Bias Dashboard. Users can use the Aequitas in three ways Python Library, Command Line Tool, and Web Audit Tool. In our project, we explored the Web Audit Tool.





5.1 Audit Paradigm

Aequitas employs a comprehensive audit paradigm to assess bias in machine learning models at various stages. Data scientists use it during model building to compare bias measures across different models, while policymakers utilize it before operationalization to understand and mitigate biases in AI systems. By standardizing the audit paradigm, Aequitas ensures that both internal (data scientists) and external (policymakers) stakeholders consistently consider bias and fairness throughout the decision-making process, from model selection to deployment and beyond. This approach aligns with the machine learning workflow, making bias assessments an integral part of the AI development lifecycle

5.2 Metrics

The following are some of the metrics that uses to generate the audit report.

Equal Parity The difference between true positive rates and False positive rates across different groups.

Proportional Parity Calculate and compare the Predicted Positive Rates across different groups.

False Positive Rate Parity Evaluates the equality of false positive rates across various subgroups defined by a sensitive attribute. Fraction of false positives within the labeled negatives of the group.

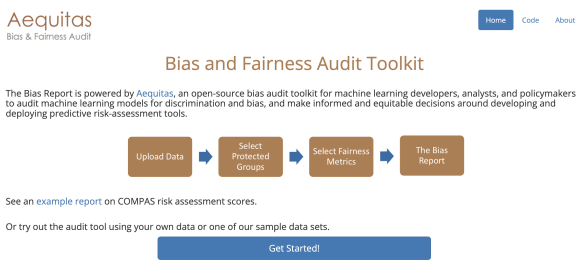
False Discovery Rate Parity Compares the difference in false discovery rates between unprivileged and privileged groups. Fraction of false positives within the predicted positives of the group.

False Negative Rate Parity Compares the difference in false negative rates between unprivileged and privileged groups. Fraction of false negatives within the labeled positives of the group.

False Omission Rate Parity Compares the difference in false omission rates between unprivileged and privileged groups. Fraction of false negatives within the predicted negatives of the group.

5.3 Demo

Using Aequitas is super easy—just upload your data, choose the protected group, select fairness matrices, and it will generate a report.



`http://aequitas.dssg.io/`

6 Comparison

We utilized the 'Adult Censure Dataset' to assess the dataset's fairness using two distinct tools. Given that both tools employ a common bias metric, referred to as "Equal Opportunity Difference" in AI Fairness 360 and "Equal Parity" in Aequitas, our primary emphasis was on comparing the performance of these tools based on this particular bias metric. We observed how each tool operated in evaluating bias within the dataset.

Both tools yielded results indicating bias in the dataset concerning the protected attributes of "Race" and "Gender." Significantly, even though both tools utilize the same bias metric, they diverge in their calculation methodologies, underscoring the nuanced approaches adopted by AI Fairness 360 and Aequitas in assessing fairness.

7 Results

7.1 AI Fairness 360 - Results

This metric calculates the difference in true positive rates between unprivileged and privileged groups. The true positive rate represents the ratio of correctly identified positives to the total actual positives for a specific group. The optimal value is 0, indicating fairness. A value less than 0 suggests greater benefit for the privileged group, while a value greater than 0 indicates greater benefit for the unprivileged group. Fairness, in this context, is achieved when the metric falls within the range of -0.1 to 0.1.

7.2 Aequitas - Results

Aequitas produces the results below, generating a report that reveals notable disparities:

The predicted positive rate for males is 1.6 times higher than for females, indicating gender-related bias. Among racial attributes, significant disparities

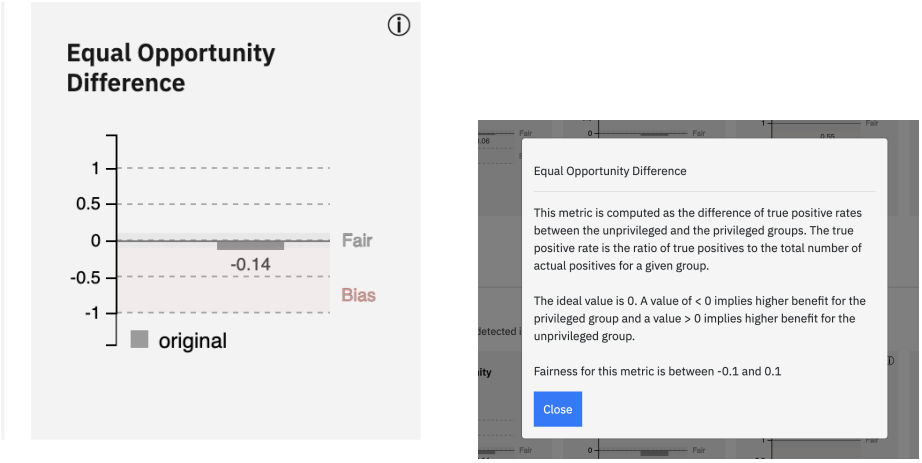


Figure 5: AI Fairness 360 Result

Audit Results: Bias Metrics Values

gender

Attribute Value	Predicted Positive Rate Disparity
Female	1.0
Male	1.6

Go to Previous
Go to Top

race

Attribute Value	Predicted Positive Rate Disparity
Amer-Indian-Eskimo	1.39
Asian-Pac-Islander	3.34
Black	12.73
Other	1.0
White	93.4

Go to Previous
Go to Top

Figure 6: Aequitas Result Report

exist: "Asian-Pac-Islander" has a predicted positive rate disparity of 3.34, while "White" exhibits the highest disparity with a rate of 93.4. These findings emphasize the importance of addressing bias in the model's predictions across different gender and racial groups.

8 Conclusion

When we look at both tools, we notice they have different processes and results. However, the main goal for both is the same: to figure out bias and fairness in our systems, find any differences, and try to make AI systems that are fair and clear. Each tool has its way of doing things, but together they help us understand our models better and fix any biases, moving us toward a future where AI is fair, accountable, and inclusive.

9 Acknowledgements

We would like to express our sincere gratitude to Ameeta Agrawal for their invaluable guidance, mentorship, and support throughout the project.

We are also deeply thankful to Ekata Mitra for their assistance and constructive feedback.

10 References

- http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/
- http://aequitas.dssg.io/
- https://dssg.github.io/aequitas/
- https://arxiv.org/pdf/1811.05577.pdf
- https://github.com/dssg/aequitas
- https://youtu.be/6-ceLhDBwxg?feature=shared
- https://youtu.be/DjSYRb8lWd0?feature=shared
- https://aif360.res.ibm.com/
- https://github.com/Trusted-AI/AIF360
- Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.