Usability Experiment for Library focused tool

Aequitas

Aequitas is an open-source bias audit toolkit specifically designed for the machine learning workflow. It was developed at the University of Chicago. It helps identify bias in various stages, from data preparation to model evaluation.

Metrics that Aequitas uses to identify biases

1. Equal parity - each group is represented equally among the selected set.
2. Proportional parity - each group represented proportional to their representation in the overall population
3. False positive parity - each group to have equal False Positive Rates
4. False Negative parity - each group to have equal False Negative Rates

Important links related to Aequitas

1. Project website: http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/
2. Tool Website: http://aequitas.dssg.io/
3. Github: https://github.com/dssg/aequitas
4. Documentation: https://dssg.github.io/aequitas/

What I liked:
The tool caters to both coders and non-coders. The tool has an option for users to either use the website or the library.

What I disliked:
It requires a different data format if the user wants to use a website or library. I believe it should be uniform across the tool.

Experiment

I focused on using the website version as in the future my idea is to develop a tool that helps the user visualize their large datasets even without prior experience in coding.

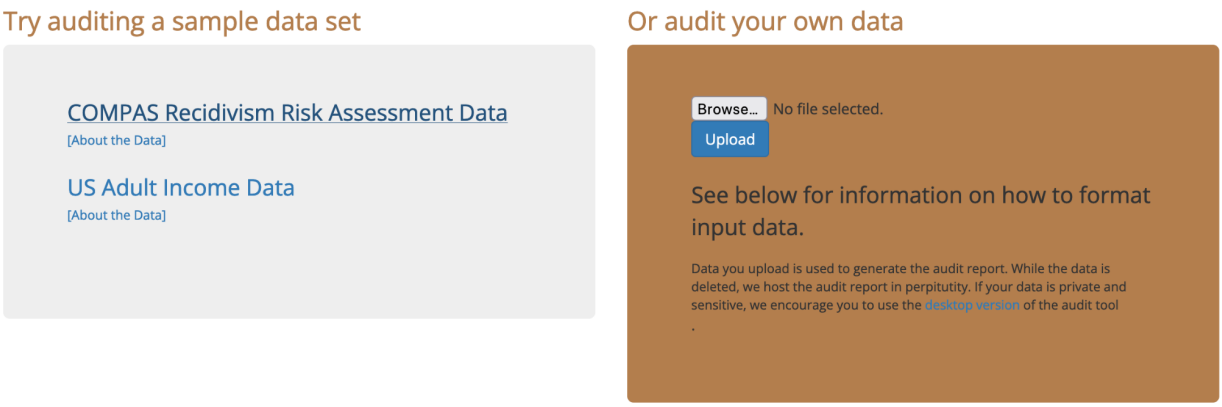
Process:
Using the website version is fairly straightforward. Even users who have no deep knowledge of machine learning can figure out the usage. The instructions are quite clear and it involves 4 simple steps

1. Upload the data
2. Select the groups
3. Select Fairness metrics

4. A bias report is generated

Uploading the data

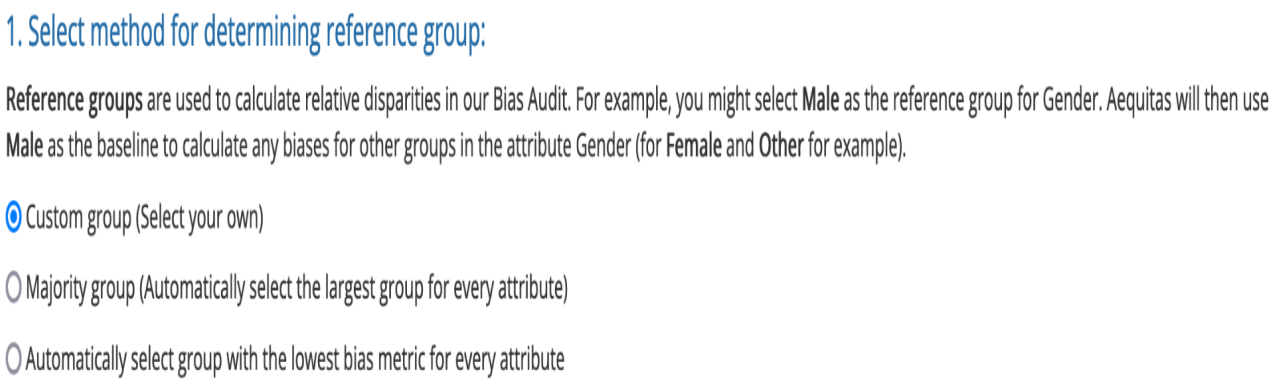
The tool also provides two sample datasets for the users to try out, and an option to upload our own data. Although there are few restrictions on how to format the Input data, it is nothing out of the ordinary that an ML dataset won't have otherwise.



For my experiment, I selected US Adult Income Data

Selecting the groups

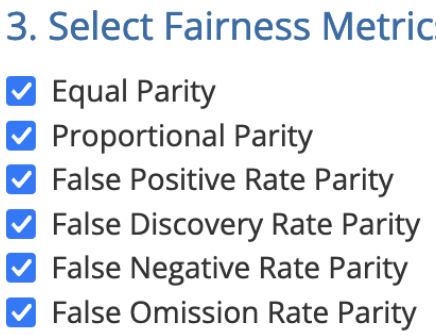
One of the best features of the tool is that it has automatic selection of majority groups so it makes it easy to calculate biases against other groups. Although custom options might sound like the best way to select a group of your interest it does introduce a chance of human error.

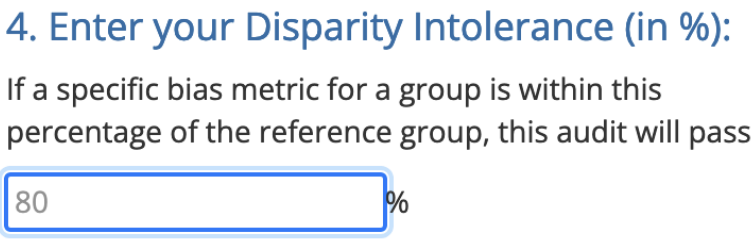


For my experiment, I selected "Automatically select group with the lowest bias metric for every attribute"

Selecting Fairness metrics

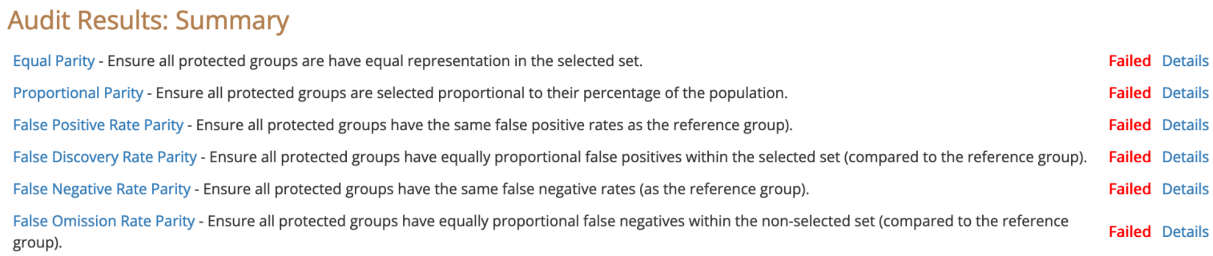
The tool does a very good job of providing various options for Fairness calculation and custom parity Intolerance.





For my experiment, I have selected all the metrics with 70% as Disparity Intolerance

Result presentation



Although it provides a good summary of all the metrics at a glance and also a detailed summary of each metric.

The below image shows a detailed breakdown of one of the metrics

Which groups failed the audit:

For education (with reference group as **Preschool**)
- Assoc-voc with **18.88X** Disparity
- 9th with **9.25X** Disparity
- 10th with **17.94X** Disparity
- 7th-8th with **11.25X** Disparity
- 1st-4th with **2.69X** Disparity
- Doctorate with **1.81X** Disparity
- Masters with **12.69X** Disparity
- 12th with **7.00X** Disparity
- 11th with **20.62X** Disparity
- HS-grad with **174.00X** Disparity
- Prof-school with **2.75X** Disparity
- Some-college with **113.69X** Disparity
- Assoc-acdm with **14.00X** Disparity
- 5th-6th with **4.88X** Disparity
- Bachelors with **56.00X** Disparity

For gender (with reference group as **Female**)
- Male with **1.60X** Disparity

For race (with reference group as **Other**)
- White with **93.40X** Disparity
- Asian-Pac-Islander with **3.34X** Disparity
- Black with **12.73X** Disparity

Conclusion: Very easy-to-use bias detection tool

IBM AI Fairness 360

The AI Fairness 360 toolkit is a flexible open-source library by IBM that incorporates methodologies created by the research community. It is designed to identify and address bias in machine learning models at all stages(Pre-process, in-process, and post-process) of the AI application journey. The AI Fairness 360 package is accessible in both Python and R. It is still a work in progress.

Metrics used to calculate biases

1. Statistical Parity Difference
2. Equal Opportunity Difference
3. Average Odds Difference
4. Disparate Impact
5. Theil Index

Important links

1. Github: https://github.com/Trusted-AI/AIF360#python
2. Website: https://aif360.res.ibm.com/
3. Demo: https://aif360.res.ibm.com/data

What I liked: The bias can be detected during pre-process, in-process, and post-process which means the data can be checked for bias and changed by Reweighing, Optimized pre-processing, Adversarial Debiasing, and Reject Option Based Classification.

What I disliked: It took a lot of time, It took around 2 hours to check bias during the in-process phase. It is not one model that fits all. The demo has only selected datasets and no tuning parameters

Experiment

I tried running the demo version of the tool as I am focusing on the tool's UI

Process:

It has 4 steps

Choosing the data: There are only 3 options and can not upload your own data

1. Choose sample data set

Bias occurs in data used to train a model. We have provided three sample datasets that you can use to explore bias checking and mitigation. Each dataset contains attributes that should be protected to avoid bias.

◉ **Compas (ProPublica recidivism)**

Predict a criminal defendant's likelihood of reoffending.
Protected Attributes:
- **Sex**, privileged: ***Female***, unprivileged: ***Male***
- **Race**, privileged: ***Caucasian***, unprivileged: ***Not Caucasian***

Learn more

◯ **German credit scoring**

Predict an individual's credit risk.
Protected Attributes:
- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

◯ **Adult census income**

Predict whether income exceeds $50K/yr based on census data.
Protected Attributes:
- **Race**, privileged: ***White***, unprivileged: ***Non-white***
- **Sex**, privileged: ***Male***, unprivileged: ***Female***

Learn more

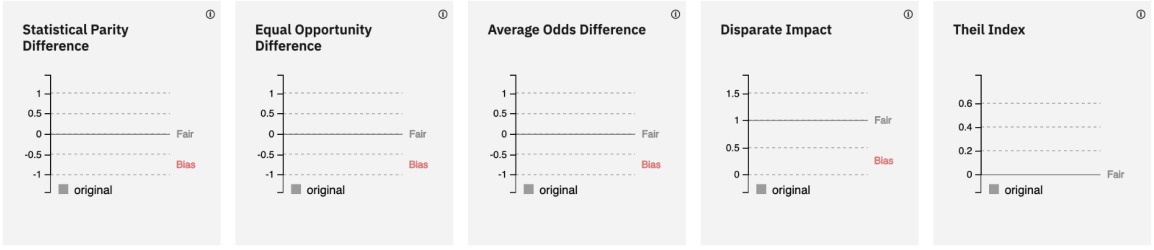
I chose the "Compas" datasets

Check Bias Metrics

The demo tool does not allow users to set bias parameters, it is rather fixed and the information is just displayed.

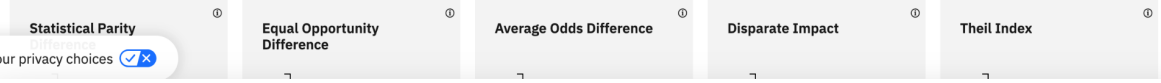
Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***
Accuracy with no mitigation applied is 66%
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



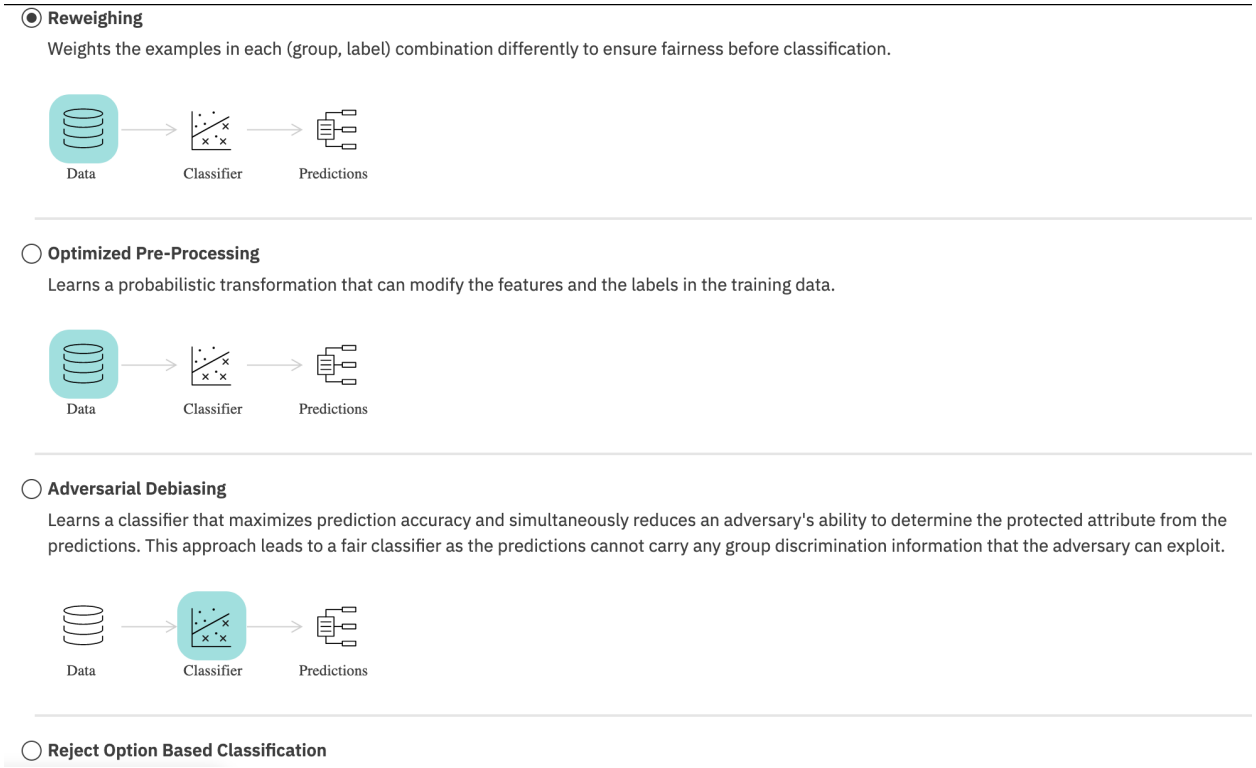
Protected Attribute: Race

Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***
Accuracy with no mitigation applied is 66%
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



Choosing a bias mitigation algorithm

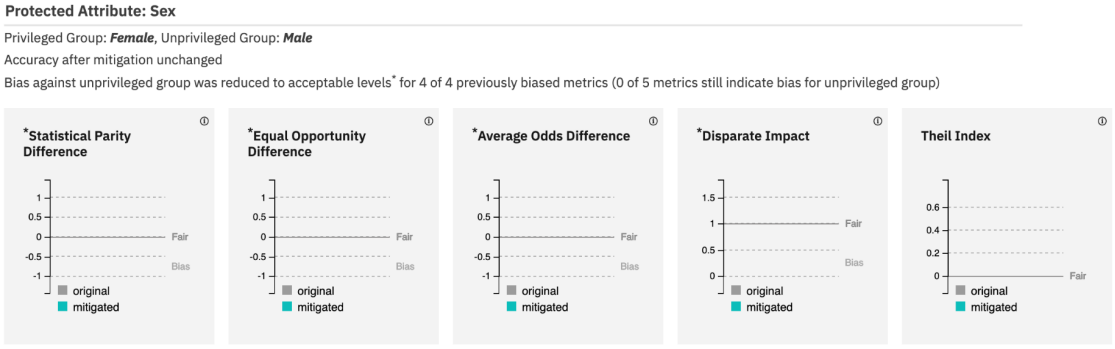
Here you can choose if you want to check for bias in the different processes of the model lifecycle



I selected to option to Reweighing the data in the initial stage

Compare original vs. mitigated results

It displays results in each of the bias metrics specified for calculations



Conclusion: The demo tool does not give a clear picture of how the tool works. It is not interactive and does not provide the option to tune parameters.

Google’s What-If Tool

Description

Google's What-If Tool is a platform for identifying biases in machine learning models. It provides a user-friendly interface to explore model behavior, compare outcomes across different demographic groups, and detect potential biases. This tool enables data scientists and developers to gain a deeper understanding of model predictions and assess fairness, making it an essential resource for enhancing transparency and equity in AI systems.

Metrics calculations

1. Demographic parity
2. Equal Opportunity
3. Equal Accuracy
4. Group Threshold

Important Links

Site: https://pair-code.github.io/what-if-tool/

Demo: https://pair-code.github.io/what-if-tool/demos/uci.html

What I like: It provides multiple options such as integrating into your notebook or Colab platform(Cloud option). For visualization, it has the option to use TensorBoard(But it is not a web-based tool)

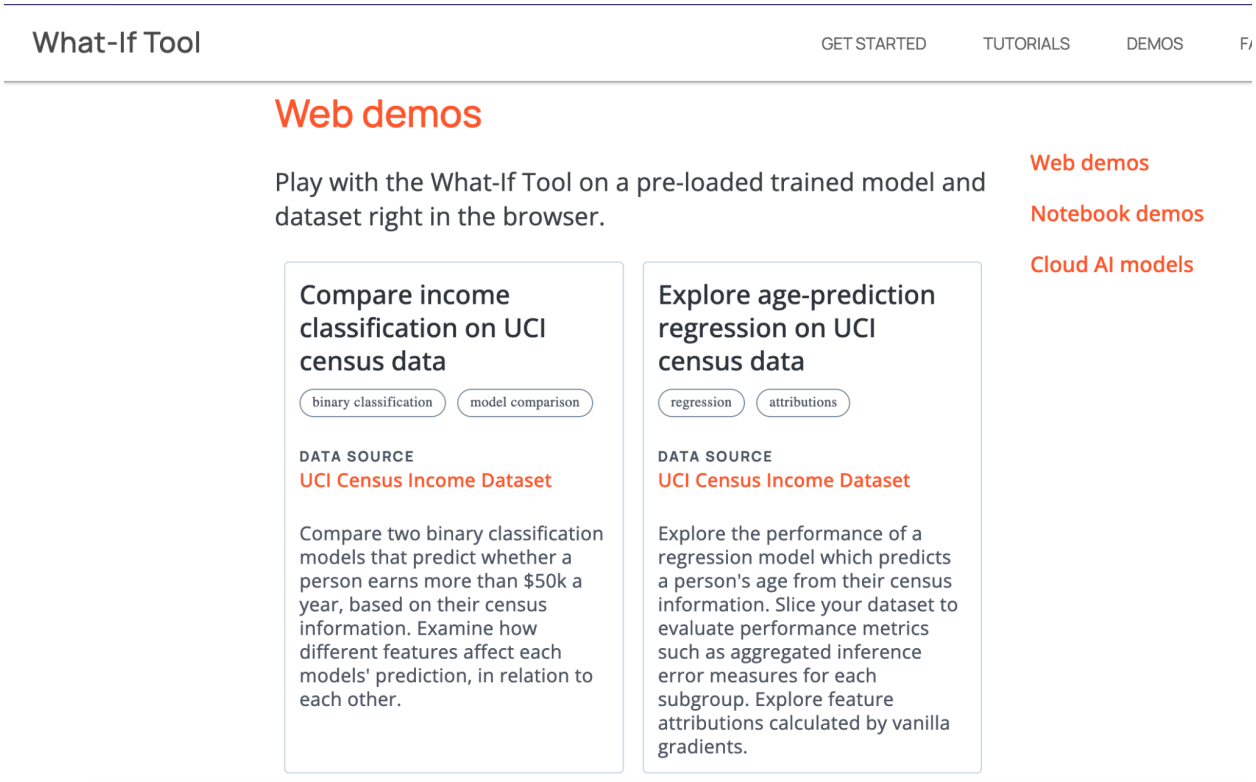
What I dislike: It has a huge learning curve. Unlike other tools that I could start using without any documentation, I had to go through the documentation thoroughly to get started. Identifying biases is only after the processing of the data. Fairness can only be calculated in Binary classification models.

Experiment

I have covered web demos in this experiment

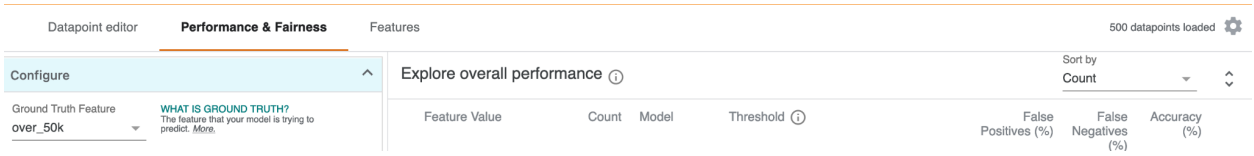
Process:

Selecting a demo: Here I chose Compare income classification on UCI census data



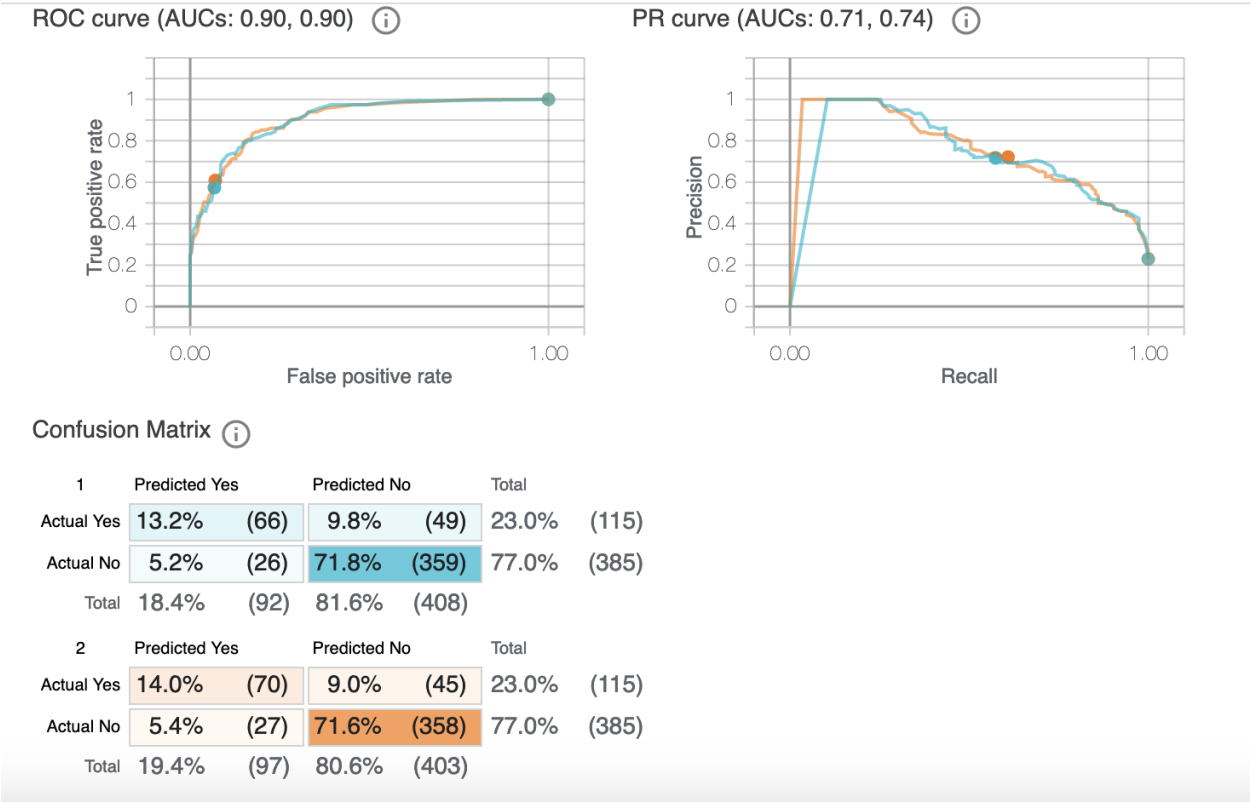
The tool provides 3 options in terms of exploring

1. Datapoint editor: With this users can edit their datapoint features and also see the model prediction after the edit
2. Performance & Fairness: This was my point of focus. Here I was able to access the fairness of the model through the confusion matrix
3. Features: Where the histogram of each feature is displayed



Results:

The way to identify bias is through a Confusion matrix. The web does provide a few opinions for the users to select Such as custom threshold or single threshold(Optimize a single threshold for all datapoints based on the specified cost ratio.)



TensorBoard can be used for visualization purposes. It is not a web-based tool and is also to be installed like a library in the notebook.

Usability Experiment for Visualization focused tool

FairSight

Description

FairSight was developed by FairDM to assess fairness in the ML model. It is a visualization platform built with React and Django. It uses D3.js for Visualization. The framework is agnostic to the model being used and aims to offer a fairness pipeline that facilitates the assessment of fairness at every stage, spanning from input to output, within the workflow.

Metrics calculations:

The tool helps to identify bias during Pre-processing, in-processing, and post-processing.

a. Understand how every step in the machine learning process could potentially lead to biased/unfair decision-making
b. Measure the existing or potential bias

c. Identify the possible sources of bias
d. Mitigate the bias by taking diagnostic actions. We provide the rationale for each action in the following.

Important Links

Github : https://github.com/ayong8/FairSight

Paper: https://arxiv.org/abs/1908.00176

What I like the most: The paper is very detailed to understand how the system identifies biases

What I dislike the most: It is very difficult to understand or even find the link to use it

Experiment:

Could not conduct experiment as no link to the system was found

FairVis

Description

FairVis is a visual analytics platform enabling users to conduct audits on their **classification** models, specifically for intersectional bias*. Users have the capability to create data subgroups and explore whether the model exhibits disparities in performance across different demographic groups

Intersectional bias: recognizes that people can experience unique and compounded forms of discrimination when multiple aspects of their identity interact.

Metrics calculations: Focused on sub-groups, uses Confusion Matrix to identify biases based on subgroup

Important Links

Paper: https://arxiv.org/abs/1904.05419

Github: https://github.com/poloclub/FairVis

Demo: https://poloclub.github.io/FairVis/

What I like the most: Upon loading a dataset, a histogram of the features is presented which makes it easier to see data distribution.

What I dislike the most: The demo site kept crashing and I had to restart the entire process from the beginning. There's a learning curve and not very easy to figure out. Response time is more than 5 seconds

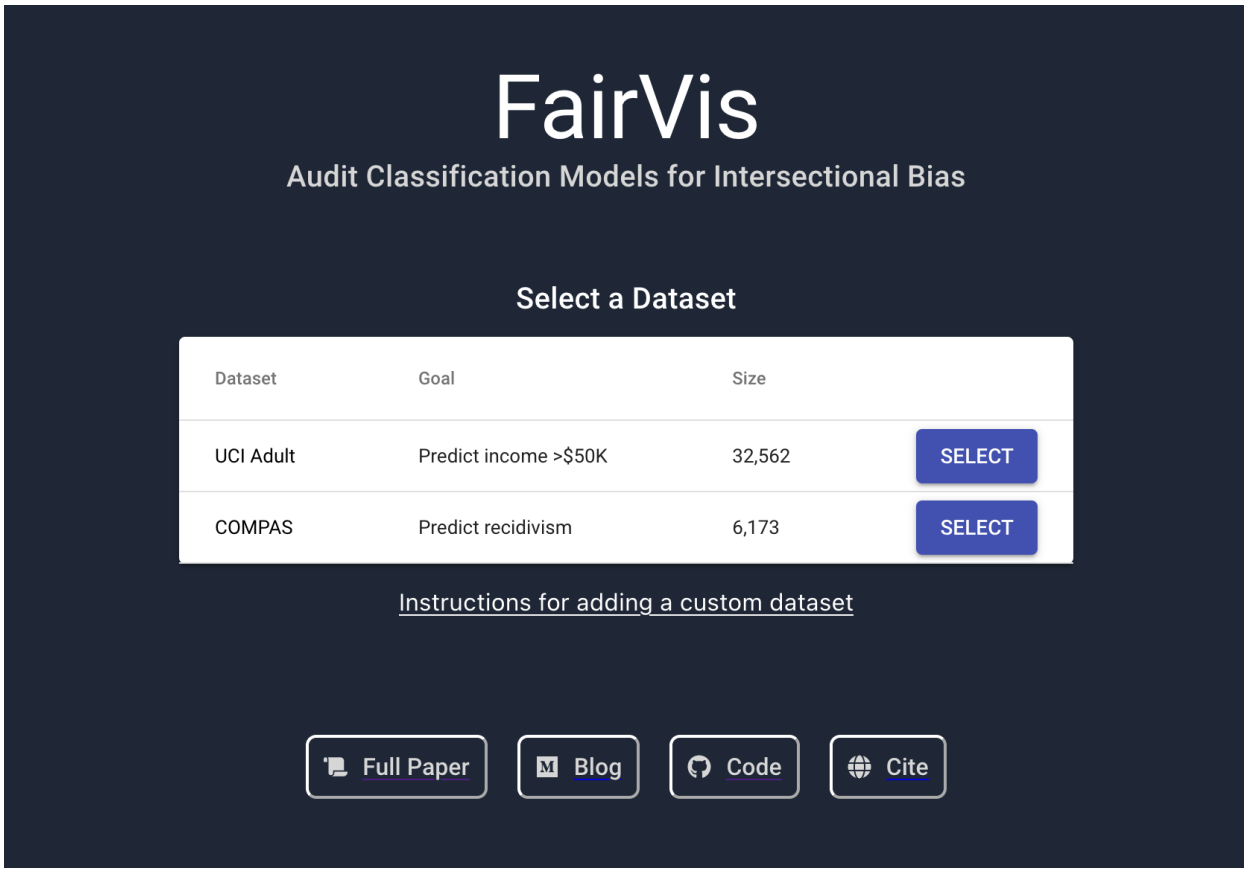
Experiment

When trying the live demo; the tool had two option datasets to choose from. Since it is a library, users need to clone the repo and install npm to start the application.

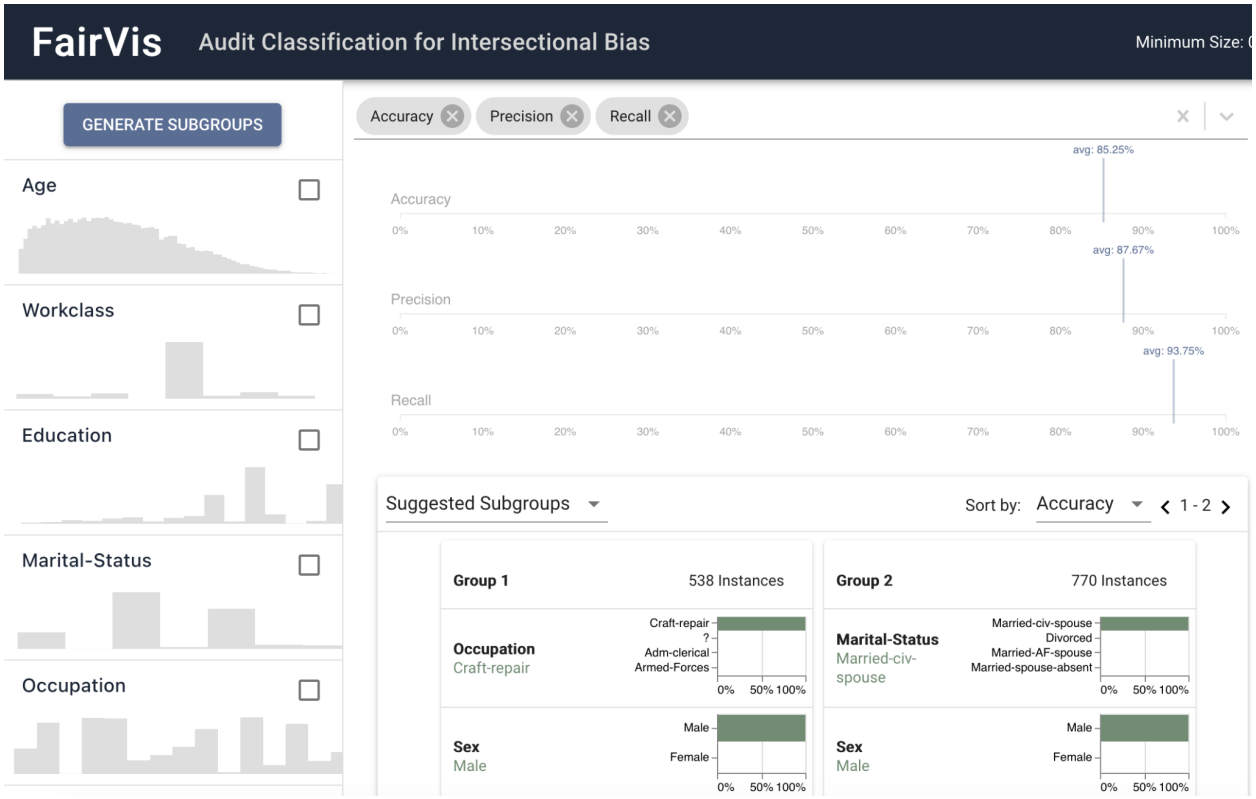
The response/feedback time of the system is delayed.

Process:

Step 1: Select a dataset that is provided for the demo. I selected the UCI Adult dataset for uniformity



Step 2: Upon loading the datasets, the tool shows a histogram of all the features and groups the value providing options to select.



Conclusion: It is a good library-based tool

slice teller

Description

SliceTeller is introduced as an innovative tool designed to facilitate the debugging, comparison, and enhancement of machine learning models, particularly those reliant on crucial data slices. SliceTeller automates the identification of problematic data slices, offering insights into the reasons behind model failures. Significantly, it introduces an efficient algorithm, SliceBoosting, which aids in the estimation of trade-offs when prioritizing slice optimization. Moreover, the system empowers model developers to conduct comprehensive comparisons and analyses of different model iterations, enabling them to select the most suitable version for their specific applications. The system's efficacy is demonstrated through the evaluation of three use cases, including two real-world product development scenarios, underscoring SliceTeller's ability to enhance the debugging and refinement of product-quality machine learning models.

Metrics calculations: Not mentioned

Important Links

Paper: https://ieeexplore.ieee.org/abstract/document/9906903

Youtube: https://www.youtube.com/watch?v=QJZ9iksgDTI

What I like the most: The use cases are very well demonstrated

What I dislike the most: There are no links to the repo or tool that I could find to conduct the experiment

Experiment

Could not conduct the experiment as no links were found

Usability Experiment for Web+vis focused tool

Facets

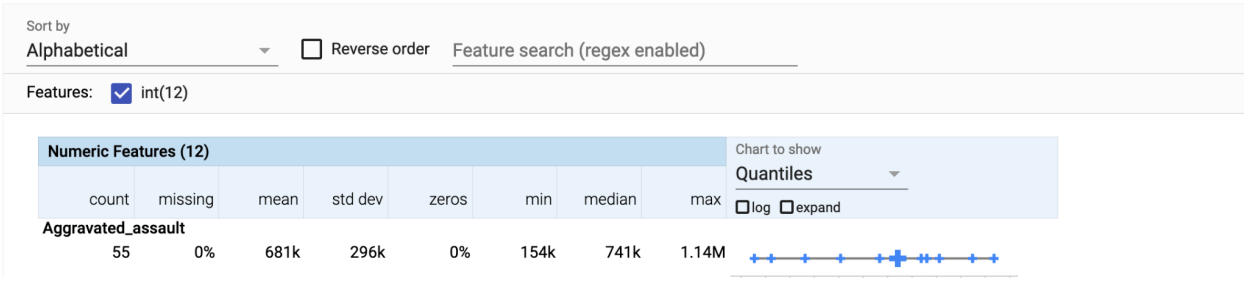
Facets is a web-based data exploration tool by Google. It offers two powerful visualizations that assist in comprehending and examining machine learning datasets. Gain insights into the characteristics of each dataset feature through Facets Overview, or delve into specific observations with Facets Dive.

Important link:
Website: https://pair-code.github.io/facets/

What I like: The best feature of this tool is that it has options for both an overview and a detailed look at the dataset

Experiment:

Facet Overview: This lets you see the overview of the dataset.



It is interactive and provides options for the user to select the sorting order of the features. I was able to analyze the data by uploading my own dataset

Facets Dive: As the name suggests, it allows you to dive deep into the dataset. Initially, it shows all the data points by selected channels and has options for binning and scattering. It is even easier to upload own data to explore

Result: Highly efficient interactive data visualization tool.

SandDance

SandDance is a web-based interactive tool that helps craft data-backed narratives, construct evidence-based arguments, test hypotheses, delve deeper into initial explanations, inform purchasing decisions, and connect data to broader real-world contexts. SandDance employs unit visualizations, ensuring a direct mapping between database rows and on-screen data points, with animated transitions between views assisting in preserving context during data exploration.

Important links:
Web link: https://microsoft.github.io/SandDance/app/

Experiment

Although this tool gives an overview of the data, it does not explicitly claim to detect biases in the datasets. This is the same as Tableau but with advanced features for data exploration.