Introduction

Concerns surrounding the integrity and fairness of ML model use in important decision-making systems with serious consequences (e.g. justice system, hiring practices, and so on) have rightfully been extensively raised and discussed. The growth and development of the field has revealed that models often reflect the real-life biases of the humans developing them. Biases in these algorithms can lead to real problems that affect real people, namely systemic discrimination against people based on sensitive attributes, further perpetuating real, existing social inequality. Ensuring (or at least, striving for) fairness in AI models is critical in preventing needless real-world harm and maintaining the integrity of such automated systems. As someone who understands the effects of such systemic discrimination on a personal level, this issue is important to me.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool is a perfect example as justification for these concerns. It is commonly used in American courts to predict defendants' likelihood of recidivism, yet studies have highlighted significant disparities in predictions based on race. A ProPublica analysis (Julia Angwin 2016) states that Black defendants are twice as likely as white defendants to be incorrectly predicted to reoffend. Further, white defendants are increasingly mislabelled as low risk to reoffend, despite going on to commit crimes regardless of that prediction.

This report focuses on enhancing fairness testing methodologies for AI models beyond a given baseline (random search), where sensitive attributes are flipped randomly across test samples and predictions are assessed for bias. The focus will be on the race attribute. This baseline method's reliance on chance overlooks systematic biases and is insufficient to reliably assess model fairness; hence, the necessity for improvement. To be clear, the goal here is to test bias, not mitigate it.

Related Works

Several methodologies and approaches have been developed to tackle this issue, but below are options I researched to improve my grasp on subject matter before settling on my approach.

- **Permutation tests** (Cyrus DiCiccio 2020)**:** assessing independence of model predictions from sensitive attributes by using exhaustive permutations of these attributes to evaluate fairness.
    - Pros:
        - rigorous, dependable approach that will reveal bias if it exists
        - Robust across fairness metrics and contexts
    - Cons:
        - Very intensive computationally, especially as datasets increase in size (exponential effect on overhead)
        - Careful interpretation of results needed, as specific metrics other than manual interpretation not specified
- **Local Group Bias Detection by Clustering (LOGAN)** (Jieyu Zhao 2020)**:** uses clustering to identify local regions in data where behaviour is different between groups. The paper argues that bias evaluation at corpus level is not enough uncover/understand embedded bias in model.
    - Pros:
        - Specific biases affecting even subgroups within data can be revealed.
        - More granular analysis of bias that would not be apparent from corpus level testing.
    - Cons:
        - Very deliberate/careful optimisation of hyperparameters needed to net meaningful results.
        - Potential sensitivity to noise/outliers.
- **Bias Estimation Using Deep Neural Network (BENN)** (Amit Giloni 2020)**:** uses pre-trained unsupervised DNN to provide bias estimation for given ML model and dataset, for every feature based on model predictions. The motivation is the difficulty of generalising fairness testing methods across contexts, and the need for domain experts.
    - Pros:
        - generic approach that can be applied to any model without need for an expert.
    - Cons:
        - Effectiveness depends on quality of pretrained model
        - Only forms of bias present in the training set of that pretrained model may be uncovered in further tests of models and datasets.

Solution

I propose the following to improve beyond the given baseline:

- a greedy heuristic that actively selects those attribute modifications that maximise changes in predictions, in an effort to expose more subtle biases.
- the use of non-parametric Wilcoxon signed-rank hypothesis test to statistically determine whether flipping a sensitive attribute causes consistent directional shifts in prediction scores to ensure that subtle systemic bias is revealed (in case IDI ratios remain largely the same compared to baseline).

The starting point is to apply this to the COMPAS dataset, given its relevance and prominence in the model fairness testing space, then extending to another dataset (KDD) to check robustness across contexts.

Baseline Design: Random Search with Decision Comparison

The baseline implementation follows a random search approach for detecting bias:

> **Data handling:**
> The pre-trained DNN is loaded, along with the corresponding dataset. The target column ('Recidivism' for the COMPAS set) is dropped so code can keep focus on only input features. Features are already in numerical form, so no encoding or further preprocessing is necessary.

> **Sensitive attribute flip:**
> A sensitive attribute is chosen (race). Create a duplicate of each test case and change the sensitive attribute. An alternative race value is selected at random from the set of available discrete options.

> **Random perturbation of non-sensitive features:**
> To ensure the system is not overly reliant on unrealistic ideal inputs, (scalable) random noise is applied to non-sensitive features. The lab baseline requires evaluation of model fairness under minor, realistic input variability.

> **Bias detection (IDI):**
> Predictions for both original and modified inputs are obtained, and if the final prediction after rounding changes, the instance is counted as an Individual Discriminatory Instance (IDI). The IDI ratio is then computed as those such biased cases over the total number of test cases.

> **Statistical evaluation – Wilcoxon signed-rank hypothesis test:**
> In addition to counting final decision changes, the algorithm keeps track of the signed differences between original and modified prediction scores. The Wilcoxon signed-rank test is then used to assess whether these differences are systematically biased in one direction. A low p-value would suggest that the model decisions are indeed significantly altered by flipping a sensitive attribute.

In summary: This is the naïve approach that I am aiming to improve upon. It is generic across datasets and models, and picks out cases where a sensitive attribute is the only factor in a changed outcome. A true naïve approach would not utilise hypothesis testing, but I have implemented that in the baseline code for the purposes of comparison with improved code results. The reasoning for the choice of this test is discussed in the Experiments section.

Because attribute changes are random, any obtained results overly rely on chance, lack reproducibility, and can easily outright miss existing *consistent* bias due to its probabilistic foundation.

Improved Design: Greedy Heuristic + Statistical Validation

To improve upon the baseline, the modified solution incorporates two main enhancements:

1. Greedy Heuristic for Attribute Flipping:

The improved system uses a greedy heuristic that chooses the race flip that causes the maximum prediction shift possible, rather than picking one randomly. Such a method of ensuring the worst-case scenario for bias makes the algorithm meaningfully more sensitive to discriminatory decisions, shifting the focus of the program from chance to a more targeted search. The heuristic ensures that the modified test case duplicate is as meaningfully different to the original test case as possible, rather than being arbitrarily different.

2. Threshold-Based IDI + Wilcoxon test:

Since bias is unlikely to be so pronounced that final decisions are completely altered from a single attribute flip, the improved solution introduces a threshold for classifying test cases as biased. If the absolute difference between the prediction output of the original and modified test cases exceeds said threshold, the case is counted as a biased instance. This is a far more nuanced understanding of bias, recognising that bias does not only exist if prediction scores are altered significantly enough to consistently affect final rounded outcomes. Using this method we can instead focus on more subtle biases in the model.

Like in the baseline code, the Wilcoxon signed rank test for statistical significance is implemented here. As the use of the greedy heuristic reliably maximises prediction score changes, the signed differences should be more prominent, which in theory should result in stronger statistical evidence of bias (lower p values).

In summary: The improvement upon the baseline was designed to address the limitations of random search and the reliance on bias being reflected as changes in final rounded output. It systematically maximises contrast within test case pairs, ensuring that every pair fishes for bias as aggressively as possible, and the threshold mechanism makes the IDI ratio more sensitive to subtle bias. The result is that more bias that would otherwise remain hidden is captured.

Setup

The solution was developed and tested in Python 3.10. TensorFlow was used to handle the model locally, and standard imported libraries (NumPy, Pandas, SciPy and Matplotlib) were also used.

Data:

The main dataset and model used to develop the solution is the COMPAS recidivism dataset, which holds features related to adult and juvenile criminal history and age/gender/race demographics. The target column “Recidivism” is dropped from the testing set before evaluation to focus on the input terms. The sensitive attribute being tested is race (categorical field with 5 discrete numerical options representing races).

Parameters and values:

- **Number of test cases: 250** randomly selected per run. Run seeds can be used for reproducibility purposes, but not used for this project
    - Values smaller than 250 would indeed be more efficient (especially when using greedy heuristic), but would not aggregate enough information for the hypothesis test to be effective. On the other hand, values tending higher than that would start to become very computationally intensive to the point of unfeasibility for extensive use and analysis.
- **Perturbation probability: 0.3** (chance of non-sensitive feature to be perturbed randomly)
- **Perturbation noise level: 0.05** (Gaussian noise scaled to value being perturbed)
    - Both perturbation parameters have been set to provide a low-to-moderate injection of noise into input features without completely distorting inputs to the point where they no longer represent the original individual.
- **Bias detection threshold (for improved design): 0.05** (difference between pair of cases)
    - The average absolute difference in prediction score between test pairs was found to be generally ~0.02, reflecting general variability due to sensitive attribute flips for this model. The bias detection threshold of 0.05 was selected to catch non-trivial, significant deviations. Capturing shifts in prediction of greater than twice the average is more likely to reflect systematic underlying bias, without also capturing random noise. This ensures that flagged cases represent meaningful deviations from expected model stability.

Evaluation Metrics

2 main evaluation metrics will be used:

- **IDI Ratio**
  Reasoning has been discussed in previous sections
- **Wilcoxon Signed Rank Test**
  A statistical method is needed to discover systematic biases that will not necessarily manifest as outright prediction changes. For that purpose, the Wilcoxon signed rank test is applied to the signed differences (not absolute differences) between the prediction scores for pairs of test cases. The null hypothesis of the test posits that the median of these signed differences is 0, i.e. any observed variation in predictions when sensitive attributes are changed is purely random noise, and on average the model does not consistently treat any one group differently. The null hypothesis is rejected when the median of signed differences is *not* 0. If median is not 0, a consistent directional shift in model output is exists when a sensitive attribute is changed, providing sufficient evidence to reject the null hypothesis.

Why Wilcoxon?

Paired t-test was an alternative for statistical testing, but Wilcoxon is easily the right choice. This is because:

- **It is non-parametric**: the distribution of signed differences is not assumed to be normal, unlike with t-test. In fact, the distribution could well be heavily skewed.
- **The focus is median, rather than mean**: this makes it more robust to outliers, which can be common especially in biased systems that can exhibit extreme deviations.
- **It tests for directional effects:** the aim is to determine whether sensitive attribute change causes consistent directional changes in paired observations, rather than identifying mean differences. The directional component is more appropriate in a scenario with concerns regarding fairness.

Experiments

Objective: to test whether the improved tool more effectively reveals model bias than the baseline. We quantitatively compare performance of improved tool to baseline using two evaluation metrics and two different datasets. Specified parameters are used to ensure effective comparisons.

Results (COMPAS dataset and model):

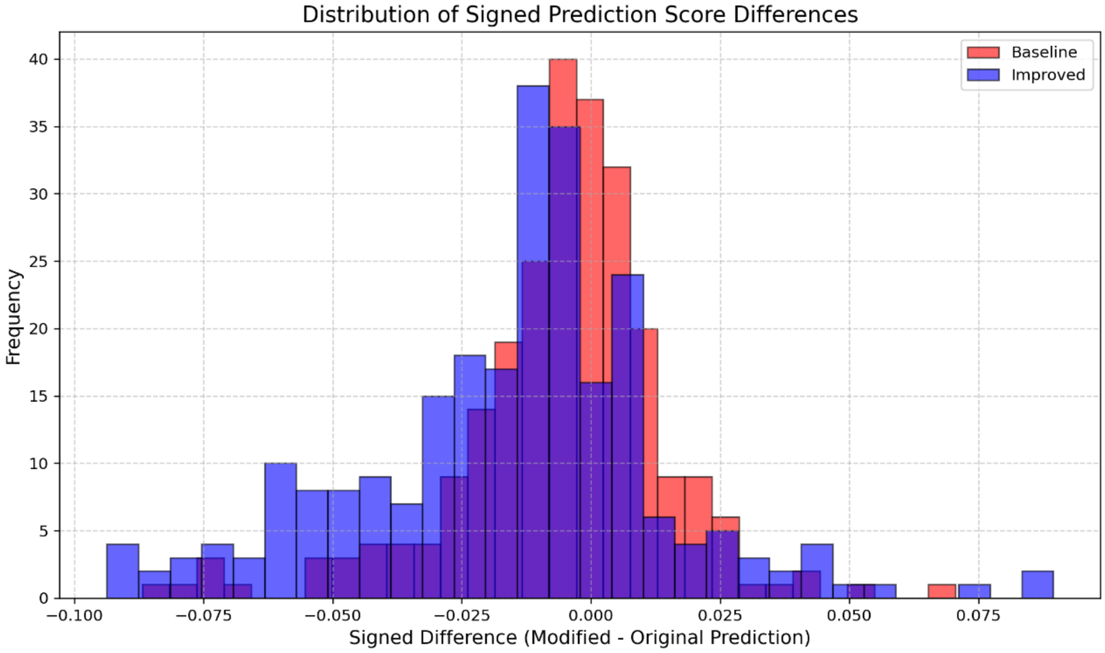
Metric/value of interest	**Baseline tool**	**Improved tool**
IDI ratio	0.001 (3/250 cases)	0.088 (22/250)
Wilcoxon test statistic	11924.0	8377.0
Wilcoxon p-value	0.001	1.69e-10
Average signed difference	-0.003	-0.009
Median signed difference	-0.003	-0.011

IDI Ratio: The baseline tool identifies 3 out of 250 cases as discriminatory, while the improved tool flags 22 out of 250 cases, nearly a *9x increase* in IDI ratio (0.001 vs 0.088). This shows that the greedy heuristic combined with threshold-based detection is significantly more effective at flagging biased instances that otherwise go unnoticed with the random search baseline.

Wilcoxon test:

- While p-values for both tools reject the null hypothesis, the improved tool's p-value tends to 0, providing overwhelmingly stronger evidence than the baseline tool that changing the sensitive attribute systematically alters predictions. In contrast, the baseline p-value still indicates significance, but with much, much less statistical confidence.
- Further, the test statistic is significantly lower for the improved tool (8377 vs 11924), showing stronger deviation from h0 i.e. more consistent bias between test pairs. Given that the average and signed differences are also larger for the improved tool compared to the baseline, it again reinforces the more substantial shift that is revealed due to the targeted bias search enabled by the greedy heuristic.

Signed differences histogram



Observations:

- The baseline tool (red) shows a narrow distribution centred around zero, with most predictions falling in the -0.02…0.02 range.
- The improved tool (blue) is overlaid, showing a wider distribution skewed to the left. A larger proportion of values falls below -0.025.

The randomness of race reassignment is evident in the baseline distribution, as the direction and magnitude of deviation varies inconsistently. This is why the distribution is symmetric and narrow, largely centred around 0, suggesting random search reveals largely inconsequential shifts in predictions.

The improved tool deliberately prioritises maximum score shifts, uncovering the model's most discriminatory behaviour; hence the wider and left-skewed distribution, showing a consistent decrease in prediction scores when attributes are flipped using the greedy heuristic. This skew reveals the model's bias in reducing prediction (i.e. lowering perceived recidivism risk) for certain racial groups. Clear non-random, directional bias in model behaviour is visualised by the asymmetry, further supporting that the improved tool's targeted strategy acts as a deliberate stress test that successfully exposes consistent worst-case bias, a characteristic lacked by the baseline.

Further, the skew confirms the integrity of the Wilcoxon test. The baseline tool's symmetry illustrates the null hypothesis (i.e. median closer to zero with no apparent systematic effect). The improved tool's skewed distribution clearly implies a non-zero median, supporting the test's higher confidence in rejecting the null hypothesis when the improved tool is used.

Results (KDD dataset and model)

To check generalisability of solution, the tools were also applied to the KDD income prediction dataset and model. Again, the model is tested for bias when changing the race attribute, and the only aspect of the experimental setup to be changed is the bias detection threshold. The average absolute difference in prediction of this model is slightly lower (~0.01…0.02), and therefore the threshold is set to 0.04, in line with the COMPAS model where the threshold is set slightly more than double the mean.

Metric/value of interest	**Baseline tool**	**Improved tool**
IDI ratio	0.032 (8/250 cases)	0.116 (29/250)
Wilcoxon test statistic	10973.0	8345.0
Wilcoxon p-value	0.0099	3.51e-9
Average signed difference	-0.0001	-0.009
Median signed difference	-0.0044	-5.19e-6

The IDI ratio more than triples when using the improved tool, and the Wilcoxon p-value sharply deviates towards 0, indicating strong evidence of systematic prediction deviation following race flips. Despite the median remaining small in both cases, the increase in average signed difference and the lower p-value nevertheless illustrates consistent bias in the model. The results suggest that the improved tool remains effective when applied to different contexts.

Reflection

Despite the credibly improved results of the improved tool, limitations exist in design and implementation.

- The scope for the greedy heuristic has not been extended for binary attributes like gender or fields with larger discrete option ranges such as age. The improvement granted by the greedy heuristic therefore is inconsistent across diverse attributes. Extending the heuristic to work on continuous attributes could be achieved by sampling a certain number of values in sub-sections of the range of discrete options. Furthermore, a multi-attribute greedy heuristic could be developed to consider combinations of several sensitive attributes, which is a good way to include testing of binary attributes as well.
- Threshold-based IDI computation could be considered arbitrary, as it is selected based on empirical observation of relevant data. No parameter tuning or validation was used to optimise/justify the threshold, and so it could well be over- or under-fitting the data. If too low it can incorrectly label random noise as bias, and if too high it can ignore true bias in the system. A good improvement would be a dynamic threshold that tailors itself to the model and data, deriving from the distribution of differences *under non-sensitive perturbations* as the baseline.
- Random perturbation method is basic. There is no consideration beyond scaling noise to the value of the attribute, so despite the fact that it largely works as intended it may create unrealistic inputs, or incorrectly apply noise to important values like IDs, locations, etc.

- The Wilcoxon test assumes all input differences are not zero, but the code does not validate this. In this context the assumption was true, but this does not bode well for generalisability.

Conclusion

The obtained results clearly support the conclusion that the improved tool outperforms the baseline in detection of both individual and systematic bias in given models. The use of the greedy heuristic and threshold-based bias detection resulted in almost 9x more individual instances of bias being flagged (IDI ratios 0.088 vs 0.001), and the Wilcoxon signed-rank test provided strong statistical evidence of consistent non-trivial directional deviation in prediction scores when the race attribute is modified. The p-value deviating so far that it tends to zero provides strong confidence in the presence of systematic bias that the baseline fails to capture.

Further analysis of the visualised signed difference distributions complements these conclusions, with the improved solution clearly illustrating a broader left-skewed distribution while the baseline remains clustered around zero.

Artifact

https://github.com/VishnuSankarakumar/ModelFairnessTesting/

References

Amit Giloni, Edita Grolman, Tanja Hagemann, Ronald Fromm, Sebastian Fischer, Yuval Elovici, Asaf Shabtai. 2020. “BENN: Bias Estimation Using Deep Neural Network.” *https://arxiv.org/abs/2012.12537*.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, Deepak Agarwal. 2020. “Evaluating Fairness Using Permutation Tests.” *https://arxiv.org/abs/2007.05124*.

Jieyu Zhao, Kai-Wei Chang. 2020. “LOGAN: Local Group Bias Detection by Clustering.” *https://arxiv.org/abs/2010.02867*.

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. “Machine Bias.” *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*.