Fairness and proxy variables: Identifying and Mitigating Information Leakage for Fair Data Science

Gilad Aharoni Yonatan Lahat

May 1, 2024

Abstract

Machine learning models may discriminate against different groups such as immigrants and African Americans. In this work we will present a method for improving the fairness of models and reducing their bias against protected classes. The method we will present identifies a limited set of features in the dataset that can be used to predict who belongs to the protected class and removes this set of features. This method succeeds in reducing the discrimination against the protected class, but it damages the accuracy of the model.

1 Problem description

Data science models may discriminate against minorities such as African Americans or immigrants because of different features that exist in the dataset that contain information related to those groups. We will call the group that we do not want to discriminate against protected class. Even if we do not use a feature that is directly related to the protected class, it is still possible that the model will be able to identify the protected class with the help of other features that exist in the dataset and as a result it will be biased against them and give them a worse prediction than the rest of the population. We are trying in this project to identify those features whose removal will reduce the bias of the model against the protected class.

2 Solution overview

Our solution is to create an automatic mechanism that knows how to choose which features to remove in order to avoid discrimination against the protected class. This mechanism finds features that can be used to predict most accurately which people belong to the protected class. In this project we tried two approaches in order to find a limited set of features that can be used to predict the protected class.

The first approach can be called ”Brute Force”. This approach is to train a model that predicts whether a person belongs to the protected class and check the quality of the model for each subset of features of the selected size. This approach gives the subgroup that predicts in the best way, but its computational cost is high. This approach requires training and testing an order of magnitude of $O(n^m)$ models where n is the number of features and m is the size of the subgroup we would like to select. This calculation is based on the assumption that m is significantly smaller than n.

The second approach is based on the method of ruckstiess et al [1] . This method is a more efficient method for selecting features in terms of runtime complexity, but it may reach a solution that is not the best. In this method, the feature that alone best predicts the target variable is first selected. After that, they look for a feature to add to it so that together they will predict the target variable in the best way. Continue like this until you reach the requested number of features.

We will use this method to select a subset of features that predicts well who belongs to the protected class. The number of models required to train and test with this approach is significantly smaller compared to the first approach. It is required to train and test an order of magnitude of $O(n*m)$ models where n is the number of features and m is the size of the subgroup we would like to select.

3 Experimental evaluation

In our evaluation of fairness, we utilize the disparate impact ratio, which quantifies the difference in positive outcome probabilities between the unprivileged and privileged groups. This metric is calculated as:

$$\text{Disparate Impact Ratio} = \frac{P(R=1|A=\text{unprivileged})}{P(R=1|A=\text{privileged})} \tag{1}$$

Where:

- $P(R=1|A=\text{unprivileged})$ represents the probability of the target feature (R) being positive given membership in the unprivileged group (A).

- $P(R=1|A=\text{privileged})$ represents the probability of the target feature (R) being positive given membership in the privileged group (A).

3.1 Interpretation:

- If the disparate impact ratio equals 1, it indicates that there is no disparity between the unprivileged and privileged groups regarding positive outcome probabilities.
- A value less than 1 suggests that the privileged group has a higher probability of positive outcomes compared to the unprivileged group, indicating potential unfairness or bias.
- Conversely, a value greater than 1 indicates that the unprivileged group has a higher probability of positive outcomes compared to the privileged group, which could also suggest a form of bias or unfairness.

By employing this metric, we aim to identify and mitigate any information leakage or biases present in our data science models, particularly focusing on promoting fairness across different demographic groups.

Secondary metric for evaluation performance is the AUROC (Area Under the Receiver Operating Characteristic) score. It is a metric used to evaluate the performance of a classification model. It quantifies how well the model can distinguish between different classes. A higher AUROC indicates better separability between positive and negative classes, while an AUROC of 0.5 suggests no class separation capacity. Importantly, the AUROC is particularly useful in binary classification scenarios where the class distribution is imbalanced—meaning that the number of instances in one class (e.g., the positive class) significantly outweighs the other class (e.g., the negative class). We uses this metric to evaluate the performance of our models in the target feature prediction.

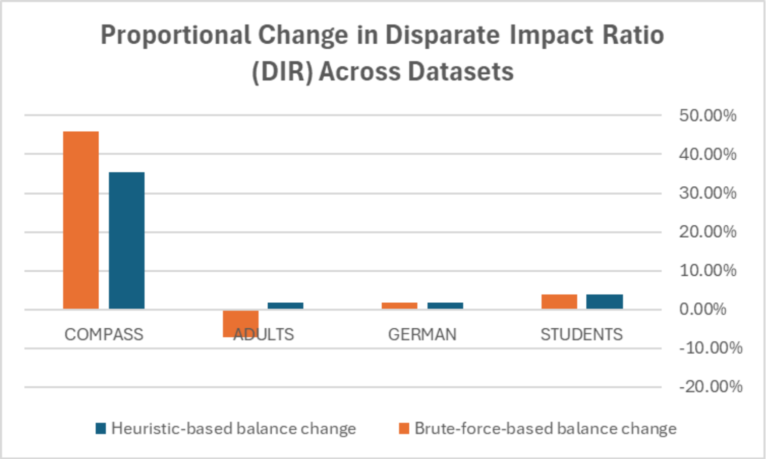

Figure 1: Proportional Change in Disparate Impact Ratio (DIR) Across Datasets

3.2 Results

Table 1: Fairness Evaluation Results

Dataset	Initial DIR	Heuristic-based DIR	Brute-force-based DIR
Students [2]	0.78	0.81	0.81
COMPAS [3]	0.65	0.88	0.95
German Credit [4]	0.94	0.96	0.96
Adults salary [5]	0.81	0.82	0.75

The results appear inconclusive. While both methods improved the initial DIR (Disparate Impact Ratio), it seems that their effectiveness varies based on the dataset rather than the approach itself. Notably, an exceptional case arises in the Adults Salary dataset, where the brute force method yields a lower DIR compared to the initial value. Upon closer examination of the selected proxy features, we observe that only the German Credit dataset shares a feature chosen by both methods. In contrast, the remaining datasets exhibit entirely different feature selections between the two approaches.

Table 2: AUROC Comparison for Feature Removal Approaches

Dataset	Initial	Heuristic-based	Brute-force-based
Students	0.76	0.76	0.71
COMPAS	0.73	0.59	0.54
German Credit	0.77	0.77	0.78
Adults salary	0.73	0.73	0.7

When examining the AUROC records of models before and after feature selection methods, a notable drawback emerges: feature selection methods designed for fairness can harm model performance. Consequently, the reliability of the model—whether biased or not—suffers. Take the COMPAS dataset, for instance, which contains only 9 features. In this scenario, the feature selection process renders the model nearly ineffective. Additionally, our initial assumption—that each dataset includes at least 3 proxy features—may lead us astray, resulting in inaccurate data. Ultimately, the delicate balance between fairness and accuracy can significantly impact the overall success of the model.

In light of the COMPAS dataset's smaller feature set compared to the others, we conducted an additional experiment. This time, both methods were employed to remove only 2 proxy features.

Table 3: AUROC and DIR Comparison for features removal approaches in COMPAS Data, removing only 2 features

Approach	DIR	AUROC Score
Initial	0.66	0.74
Heuristic-based	0.89	0.64
Brute-force-based	0.92	0.55

We observe that there is no significant improvement in the metrics. The trade-off between fairness and accuracy remains relevant here: while these approaches may enhance fairness, they often come at the cost of decreased model accuracy.

4 Related work

Both the work of Salazar et al [6] and the work of Grgic Hlaca et al [7] deal with feature selection for fair decision making as in this work.

In the work of Salazar et al [6] , as in our work, they propose an algorithm that chooses which features to remove to improve the fairness of the model. Salazar et al [6] propose a different mechanism than ours to choose which features to remove. They choose features whose removal from the set of features improves the model's fairness metric the most significantly. We used an indirect method to choose which features to remove. In their article, Salazar et al [6] also dealt with the feature engineering part and they constructed new features with the help of transformations performed on combinations of features.

In the work of Grgic Hlaca et al [7] , they deal extensively with the fairness of the process and not only with the fairness of the outcome. In the article by Grgic Hlaca et al [7] , several measures were proposed for people's discomfort with using various features that are based on a survey in which people are asked several questions about the fairness of each feature. We did not address this issue and only tried to improve the fairness of the model's outcome. According to our method, the model will use features that people think are unfair to use if the outcome of the model is not biased against the protected class when we use these features.

5 Conclusion

The method we proposed improves the fairness and reduces the discrimination of the model against the protected class. We saw this in the experiments we performed on most of the datasets.

As we have seen this method harms the accuracy of the model and in some cases it can significantly harm the accuracy of the model. When we ran our method on the COMPAS dataset we got a model with an extremely low auc score. The explanation for this is probably that the model that was trained on the COMPAS dataset used few features to begin with, so removing 2 or 3 features particularly hurt it.

References

[1] Thomas Rückstieß, Christian Osendorfer, and Patrick Van Der Smagt. Sequential feature selection for classification. In *AI 2011: Advances in Artificial Intelligence: 24th Australasian Joint Conference, Perth, Australia, December 5-8, 2011. Proceedings 24*, pages 132–141. Springer, 2011.

[2] Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TG7T.

[3] Jeff Larson and Marjorie Roswell. *propublica/compas-analysis*. 6 2017.

[4] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[5] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[6] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment*, 14(9):1694–1702, 2021.

[7] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.