Fairlearn Parity Constraints for Mitigating Gender Bias in Binary Classification Models – Comparative Analysis

Andrzej Małowiecki(✉) and Iwona Chomiak-Orsa

Wrocław University of Economics and Business, Wrocław, Poland
andrzej.malowiecki@ue.wroc.pl

Abstract. Inequality is one of the problems of the modern world. Discrimination of various kinds can affect many areas of life. The growing importance of data in the modern world makes it all the more important to ensure that the methods used to analyze it do not return results in which unfairness is present. Unfortunately, there may be situations where there is unfairness in the predictions of machine learning models. In recent years, several IT solutions have been developed to mitigate this phenomenon. One of them is Fairlearn, a Python library dedicated to this type of task. This article presents a comparative analysis of parity constraints used in Fairlearn algorithms. The purpose of this article is to identify which of the constraints is best suited for mitigating gender bias in binary classification models. The following research methods were used: literature review, experiment and comparative analysis. The evaluation of constraints will be based on the value of measures: disparity in recall and disparity in selection rate for the column containing information about the person's gender. The values of these measures, achieved by binary classification models in which the Threshold Optimizer algorithm with selected parity constraints was implemented, will be compared in order to identify which of the Fairlearn parity constraints is best suited for mitigating gender bias in binary classification models.

Keywords: Fairlearn · parity constraints · gender bias

1 Introduction

In today's world, decisions made by machine learning models have a significant impact on human life. Therefore, it is crucial that the predictions of the created models are reliable and devoid of various types of social biases. Situations in which the negative decision of the model was mainly influenced by characteristics such as age, gender, race or origin are unfair and against the established sustainability goals.

Among many solutions created in order to mitigate this problem there is Fairlearn, an open-source, community-driven project with an associated Python library. Originally, it was created with the purpose of helping to mitigate unfairness in machine learning models (Dudik et al., 2020). This library provides access to various types of algorithms and parity constraints that can be used on different types of machine learning models.

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 148–154, 2024.
https://doi.org/10.1007/978-3-031-50485-3_13

The purpose of this article is to identify which of the constraints is best suited for mitigating gender bias in binary classification models. The following research methods were used: literature review, experiment and comparative analysis. The evaluation of constraints will be based on the value of measures: disparity in recall and disparity in selection rate for the column containing information about the person's gender. The values of these measures will be compared in order to identify which of the Fairlearn parity constraints is best suited for mitigating gender bias in binary classification models.

2 Gender Bias in Machine Learning Models

Nowadays, the need for sustainable development is being increasingly promoted, especially by the youngest generations (Rzemieniak, Wawer, 2021). In this regard the problem of reducing inequalities becomes one of the biggest challenges of today's world. Among 17 sustainable development goals, 2 of them were established with the purpose of overcoming this problem:

- goal 5 – gender equality,
- goal 10 – reduced inequalities (SDG FUND, 2015).

Nowadays the use of machine learning in decision-making processes is becoming more widespread, covering a variety of fields (Butryn et al., 2021). With the growing role of this technology in everyday life, it is important to ensure that the model adheres to these sustainable development goals. This means that no sensitive characteristics should affect predictions made by them. Unfortunately, there are examples where models tend to make unfair or biased predictions (Barocas, Hardt, Narayanan, 2017, Mittelstadt, Wachter, Russell, 2023, Yang, Wang, Ton, 2023). There are many factors that can affect a model's fairness, including ethnicity, gender, age or race (Mehrabi et al., 2021).

This article focuses on gender bias in machine learning models. This type of unfairness was identified in numerous algorithms. One of the examples is an algorithm that was used in order to deliver ads, designed with the purpose of promoting job opportunities in the Science, Technology, Engineering and Math fields. Even though the ads were supposed to be gender-neutral, the majority of viewers were men. The simple explanation of this behaviour would be that the algorithm just imitated supposed user behaviour, which means that, due to the fact that women were supposedly less likely to click on an ad, it was displayed to fewer of them. This assumption turned out to be incorrect because women were more likely to click on an ad after it was displayed to them (Lambrecht, Tucker, 2019). This means that there were other reasons which can be summarized as a difference in "price" between both demographics. For a given example, women were considered a "prized demographic" due to the fact that they are more likely to engage with advertising than men, even though the stereotypical assumption can be made that these types of ads would not interest them (Lambrecht, Tucker, 2019).

There are two factors that allow users to check whether models learning model are making unfair predictions:

- disparity in predictions – predictions comparison for each group within a selected sensitive feature, measured using selection rate,

- disparity in prediction performance – predictive performance metrics comparison for each group within a selected sensitive feature (Microsoft, 2023).

When any of the disparity values is significant, then the assumption can be made that the given model is lacking fairness. The reason for this may be, e.g., data imbalance, indirect correlation between features or other societal biases. Correct identification of the reason is very important when implementing mitigation methods.

3 Fairlearn Overview

Fairlearn was originally started in 2018 as a Python package created for the purpose of a connected research paper by Miro Dudik with the aim of providing data scientists with a toolkit to mitigate unfairness in their machine learning models (Dudik et al., 2020).

The basis of fairness consists of two types of algorithms that allow unfairness mitigation in machine learning models:

- postprocessing algorithms – algorithms that transform predictions created by a trained model, e.g., Threshold Optimizer, which establishes different decision thresholds for each group within a selected sensitive feature so that the model complies with the selected constraint,
- reduction algorithms – algorithms that iteratively re-weight data points and retrain the model in order for the final version of it to have the best performance and at the same time comply with the selected constraint, e.g., Exponentiated Gradient or Grid Search (Dudik et al., 2020).

Both types have advantages and disadvantages that make their use vary depending on the given use case. In general, reduction algorithms are more flexible and compliant due to the fact that they allow the use of a wider range of metrics and do not require access to sensitive features during deployment (which often can be forbidden by the law) (Dudik et al., 2020). On the other hand, postprocessing algorithms are easier and faster to use due to the fact that there is no need to make any changes to the model, just to its predictions.

Besides algorithms, Fairlearn consists of other features that are designed with the purpose of helping people detect and mitigate unfairness in machine learning models, e.g., special metrics like selection rate, which is used to measure the proportion of positive predictions for each group within a selected sensitive feature (Microsoft, 2023, Pandey, 2022).

4 Fairlearn Parity Constraints

In Fairlearn, parity constraints are constraints that a model has to satisfy in order for it to be considered fair. There are different types of constraints that are designed in a way so that the user is able to choose whichever is best suitable for the given machine learning task and specific fairness criteria[1].

[1] Available parity constraints are mostly algorithm-agnostic, which means that they should be able to work with both types of Fairlearn algorithms. One of the exceptions is error rate parity, which is a constraint example that works only with reduction algorithms.

Among many available Fairlearn parity constraints, ones that will be used for the purpose of this article are:

- Demographic parity – constraint designed to assure that an equal number of positive predictions is being made for each group within a selected sensitive feature,
- True positive rate parity – constraint designed to assure that a comparable proportion of true positive predictions is being made for each group within a selected sensitive feature,
- False positive rate parity – constraint designed to assure that a comparable proportion of false positive predictions is being made for each group within a selected sensitive feature,
- Equalized odds – constraint designed to assure that a comparable proportion of true positive and false positive predictions is being made for each group within a selected sensitive feature (Dudik et al., 2020).

Some of the selected constraints are designed for the purpose of reducing specific unfairness factors, e.g., the use of demographic parity should reduce mainly the disparity in the model's predictions, while equalized odds should concentrate on reducing the disparity in its prediction performance. A conducted experiment should give an answer if that will be the case in this instance.

5 Comparative Analysis of Fairlearn Parity Constraints for Mitigating Gender Bias in Binary Classification Models

Machine learning models are used for the purpose of supporting the decision-making process in many areas, such as financial services, marketing or health care. Companies in every industry have the opportunity to benefit from the use of this technology in decision-making by using its models in the hiring process. As machine learning algorithms are increasingly being used at every stage of this process, it is all the more important to ensure that the decisions they make are not unfair (Schumann et al., 2020).

5.1 Experiment Overview

For the following experiment, "Utrecht Fairness Recruitment dataset" dataset was selected. This dataset was created by Sieuwert van Otterloo, AI researcher at Vrije Universiteit and Utrecht University of Applied Sciences. The owner of it is Utrecht ICT Institute, which has made it available on Kaggle with a license: CC BY-SA 4.0 (Kaggle, 2023).

Selected dataset contains data on recruitment decisions of 4 companies. It consists of over 500 candidates who are described using attributes such as gender, age, nationality, sports background, university grade and previous working experience. A number of sensitive features (such as gender, age or nationality) makes this dataset an appropriate choice for the experiment.

The experiment will be conducted according to the following procedure:

1. Creation of a baseline binary classification model, using the decision tree algorithm.
2. Calculation of evaluation metrics for the baseline model.
3. Calculation of the disparity in evaluation metrics for gender groups in the baseline model.
4. Addition of a balancing index to the training dataset.
5. Creation of ThresholdOptimizer instances for selected parity constraints.
6. Calculation of evaluation metrics for ThresholdOptimizer instances.
7. Calculation of the disparity in evaluation metrics for gender groups in the ThresholdOptimizer instances.

5.2 Baseline Model

The first stage of the experiment was to create a binary classification model, using the decision tree algorithm. Before that, selected dataset was prepared for the given task, which means that all non-numerical attributes were converted using LabelEncoder and all rows with gender values different than "male" or "female" were removed[2]. It is important to mention that the selected dataset is imbalanced.

After preparation, the dataset was split into training and test sets at a ratio of 2:1. The training set was used during the process of learning the decision tree model. The test set was used to evaluate models using selected evaluation metrics: selection rate, accuracy, recall and precision. After evaluation, disparities in evaluation metrics for gender groups in the baseline model were calculated. Table 1 presents results of the baseline model evaluation.

Table 1. Baseline model evaluation results.

	selection rate	accuracy	recall	precision
female	0.252632	0.826316	0.655172	0.659722
male	0.330567	0.825726	0.710037	0.799163
disparity	0.077935	−0.000590	0.054865	0.139441

According to obtained evaluation results, there is a disparity between the values of selection rate and recall, which means that the model is slightly biased.

5.3 Comparative Analysis

After successful evaluation of the baseline model, the next step was to implement a balancing index into the dataset. This index is used to ensure that in the input there is an equal number of samples that produce the result of 0 and 1. The new, balanced dataset was split again into train and test sets at the same ratio as before.

[2] There was only one different value for gender – "other". It was removed due to the fact that there were fewer observations of it: 83 to 2127 for "male" and 1790 for "female".

The newly created train set was used to train ThresholdOptimizer instances, which were also using the originally trained model. Each of the 4 instances had different parity constraints implemented: demographic parity, true positive rate, false positive rate and equalized odds.

Created models were evaluated using selected evaluation metrics. After evaluation, disparities in evaluation metrics for gender groups were calculated for each of the models. Table 2 presents a comparison of disparities in selection rate and recall between all the models, including the baseline.

Table 2. Comparison of disparities in selection rate and recall between all the models.

	disparity in selection rate	disparity in recall
Baseline model	0.077935	0.054865
Demographic parity	0.006064	0.050224
True positive rate	0.0924	0.081528
False positive rate	0.088149	0.070914
Equalized odds	0.083999	0.049147

5.4 Summary

Among the selected constraints, demographic parity achieved the lowest value of disparity in selection rate and the second lowest value of disparity in recall. The lowest value of disparity in recall was achieved by equalized odds constraint. Besides demographic parity, none of the other constraints achieved a lower value of disparity in selection rate than the baseline model. Additionally, the true positive rate and false positive rate constraints achieved higher value of disparity in recall than in baseline models.

Selected evaluation criteria indicated demographic parity as the most suitable parity constraint for a given use case. The true positive rate and false positive rate parities were indicated as unsuitable due to the fact that they achieved higher values of both metrics than in the baseline model.

6 Conclusions

In this article the results of comparative analysis of Fairlearn parity constraints in binary classification models were presented. Created decision tree models were compared using disparity in selection rate and disparity in recall measures.

The comparative analysis indicated demographic parity constraint as most suitable for the given use case. The use of equalized odds can also be advised as the disparity in recall achieved by this constraint was better than in the baseline model. The true positive rate and false positive rate constraints achieved worse results than the baseline model, so the application of them for a given use case is not advised.

In future publications the scope of compared Fairlearn features could be expanded to include comparison of different algorithms for different machine learning models (not only binary classification but also regression). Additionally, it could be worth trying to detect and mitigate model unfairness in different areas, such as corporate credit risk analysis or markets selection.

References

Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. Nips tutorial **1**, 2017 (2017)

Bird, S., et al.: Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR2020–32 (2020)

Butryn, B., Chomiak-Orsa, I., Hauke, K., Pondel, M., Siennicka, A.: Application of Machine Learning in medical data analysis illustrated with an example of association rules. Procedia Comput. Sci. **192**, 3134–3143 (2021)

Kaggle (2023). https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitmentdataset. Accessed 15 Jul 2023

Lambrecht, A., Tucker, C.: Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manage. Sci. **65**(7), 2966–2981 (2019)

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)

Microsoft (2023). https://learn.microsoft.com/en-us/training/modules/detect-mitigate-unfairness-models-with-azure-machine-learning/2-consider-model-fairness. Accessed 15 Jul 2023

Mittelstadt, B., Wachter, S., Russell, C.: The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default (2023). arXiv preprint arXiv:2302.02404

Pandey, H.: Comparison of the usage of Fairness Toolkits amongst practitioners: AIF360 and Fairlearn (2022)

Rzemieniak, M., Wawer, M.: Employer branding in the context of the company's sustainable development strategy from the perspective of gender diversity of generation Z. Sustainability **13**(2), 828 (2021)

Schumann, C., Foster, J., Mattei, N., Dickerson, J.: We need fairness and explainability in algorithmic hiring. In: International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2020)

SDG FUND. Sustainable development goals (2015). https://www.un.org/sustainabledevelopment/inequality

Yang, M., Wang, J., Ton, J.F.: Rectifying unfairness in recommendation feedback loop. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28–37 (2023)