Homework No. 6 - Fairness

Paulina Kaczynska

December 8, 2022

The goal of this homework is to learn about fairness measures and their implementation in Dalex. I train models on Adult Income Dataset from Kaggle, which is a classification dataset, where people belong either to the category of income bigger than 50$ or smaller.

1 Task 1

Group fairness, also called demographic parity, is a measure given by the ratio of the model qualifying a record if it belongs to the protected group versus if it does belong to the privileged group. It is $\frac{0.65}{0.5} = 1.3$.

Equal opportunity (equal true positive rate for both groups) is $\frac{0.75}{0.5} = 1.5$. False positive rate is $\frac{0.25}{0.5} = 2$.

Predictive rate parity coefficients are following: positive:

$$\frac{\frac{60}{65}}{0.5} = 1.84$$

and negative:

$$\frac{\frac{20}{35}}{0.5} = 1.14$$

.

2 Task2

2.1 Fairness of XGBooster

Firstly, I train XGBooster using model from xgboost package. It achieved recall of 0.54, precision of 0.81 and accuracy: 0.86. Then I apply Dalex fairness module, which provides following statistics of the model for each protected category: True Positive Rates (TPR), Accuracy Equality Ratio (ACC), Positive Predictive Value (PPV), False Positive Rates (FPR) and Statistical Parity Ratio.

I apply it to gender, race and age. In the first two cases choosing protected and privileged groups is not problematic: in the first case privileged are man and in the second white people. Equal opportunity ratio, predictive equality ratio and statistical parity ratio are smaller than one for categories treated as discriminated on the job market excluding people of Asian, Pacific or Islander origin.True positive ratio does not exceed the limit of 0.8.

However, finding conditions for the privileged group based on age was more problematic. According to the first intuition, older people (close to retirement period) are discriminated against on the job market, since they are more prone to losing jobs and less likely to find a new one. However, it is not something that will affect the salary and especially not the total income (including for example capital gains), since people included in this statistic already have the job. Indeed, people older than 50 had the measures skewed in their favour. Final choice of the protected category because of age were people younger than 30.

	TPR	ACC	PPV	FPR	STP
female	0.86	1.13	1.02	0.22	0.32
Amer-Indian-Eskimo	0.8	1.07	1.06	0.28	0.43
Asian-Pac-Islander	1.28	0.98	0.85	2.72	1.62
Black	0.84	1.09	1.05	0.26	0.38
Other	0.7	1.07	0.74	0.63	0.37
younger than 30	0.49	1.19	1.09	0.02	0.02

According to 4/5 rule, the model is unfair, since it scores outside the range in more than two categories for every protected category. Especially the prediction for young people is very biased.

2.2 Fairness of neural network

In the next step I train two layer multi-layer perceptron using function from sklearn module. It achieved recall 0.45, precision 0.79 and accuracy 0.84, so similar performance to XGBooster.

	TPR	ACC	PPV	FPR	STP
female	0.72	1.13	0.94	0.28	0.29
Amer-Indian-Eskimo	0.62	1.02	0.56	1.46	0.63
Asian-Pac-Islander	1.36	1.0	0.91	2.22	1.61
Black	0.65	1.09	1.02	0.24	0.3
Other	0.55	1.11	1.26	-	0.17
younger than 30	0.59	1.2	0.71	0.22	0.14

In terms of fairness it scored worse than XGBooster. The metrics are not much different, but it scores worse in terms of 4/5 rule in more metrics than XGBooster. For example for Indian and Eskimo people it scores badly in 4 metrics out of 5 instead of 2. The prediction in terms of age is not as unfair as for the XGBooster.

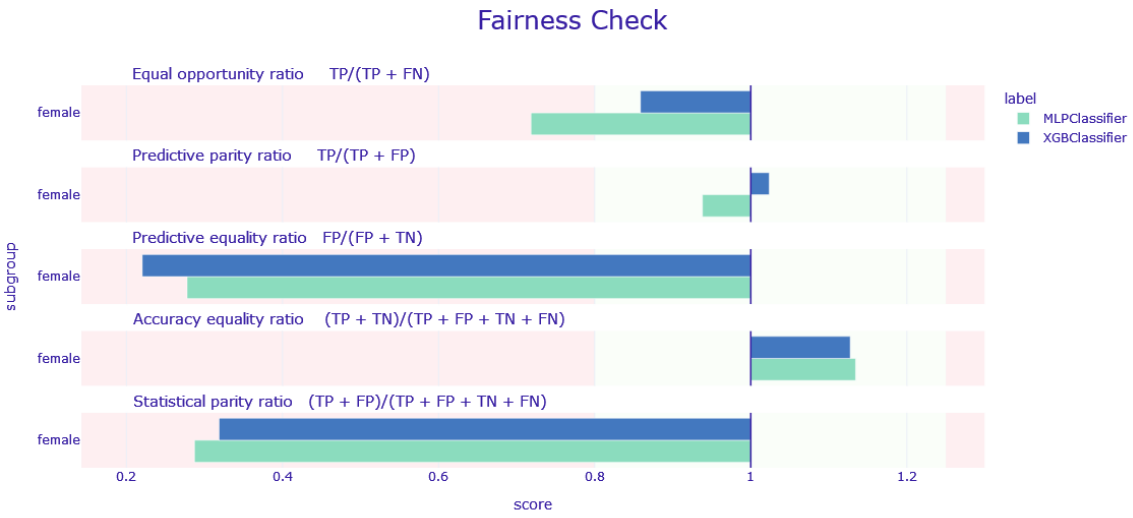
The metrics' values for both models can be seen on the Figure 1.

2.3 Improving fairness

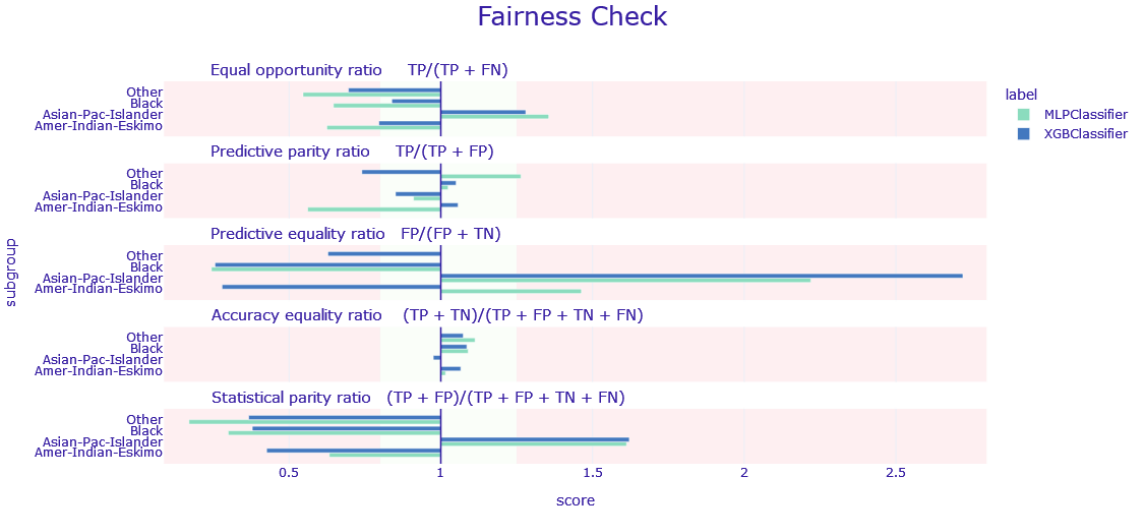
Three procedures were implemented to improve the fairness of the XGBoost model in case of gender: resample, reweight and ROC pivot. Their metrics can be seen on the figure 2. In resample method only part of the dataset is used to train the model in order to achieve more representative dataset. As can be seen on the figure 2, this method had the greatest effect on the fairness measures. False Positive ratio was especially skewed: from 0.22 it grew to over five. Two other methods, reweight and ROC pivot, improved fairness measures without making them too big and exceeding 4/5 range on the other side. Formally, all models still exceed the condition of staying in 4/5 range in at least two metrics, but now the measures are close to the range.

2.4 Dependence of fairness on the model's performance

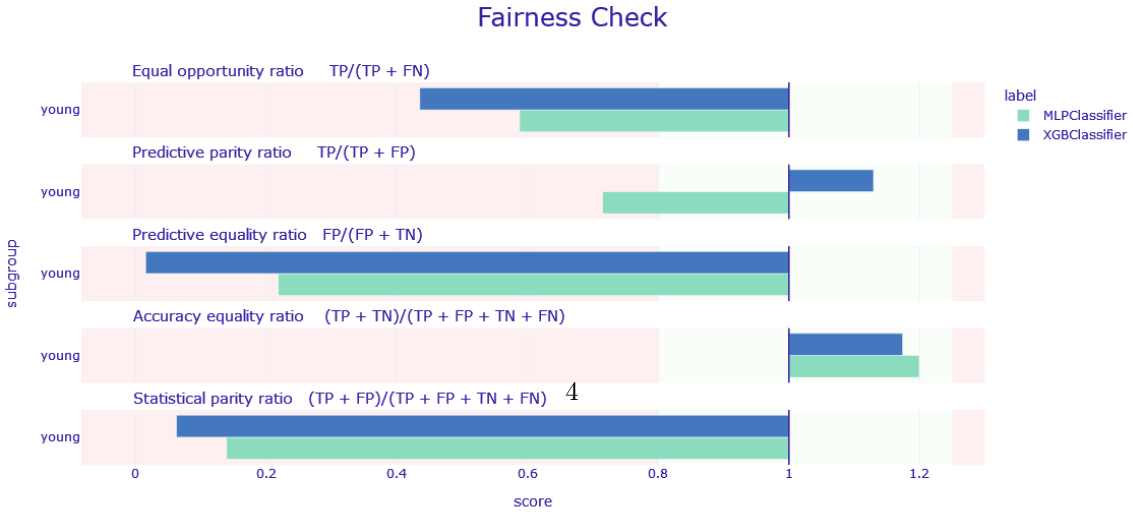
In terms of model's performance every bias mitigation technique impacted it negatively. However, reweight's and ROC pivot's effect is very small and resampling impacted it vastly. Resampling had recall of 0.42, precision of 0.8 and accuracy of 0.83. Reweight and ROC pivot achieved similar results of recall 0.51-0.52, precision 0.8 and accuracy 0.85. Summing up, reweight and ROC pivot improved fairness and did not impact model performance much, while resampling lead to significantly worse performance, while also exceeding the $[0.8, 1.25]$ on the other side.



(a) gender



(b) race



(c) age

Figure 1: Fairness of prediction for XGBoost and MLPClassifier.

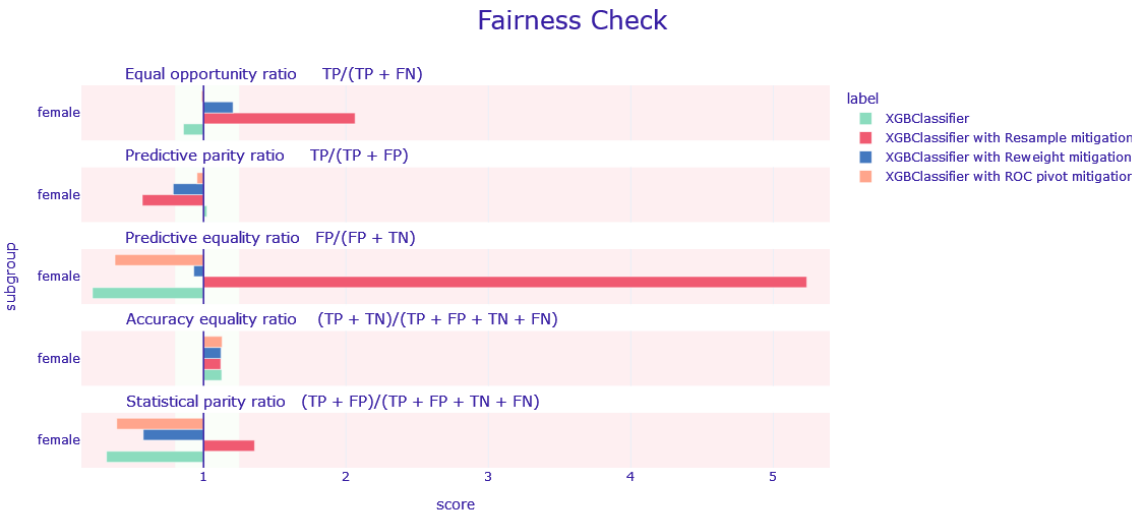


Figure 2: Comparison of models with method