Discrimination in HR analytics. A fair workflow

Davide Zulato
Matricola 876101

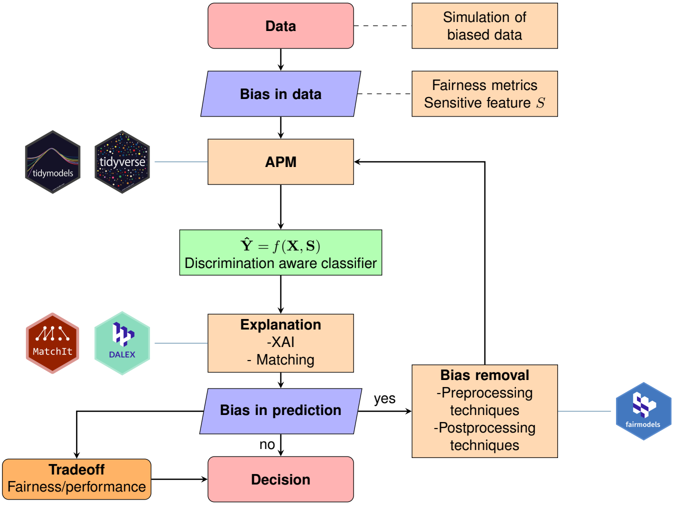
University of Milano-Bicocca, DEMS

Relatore: **Prof. Marco Guerzoni**
Correlatore: **Prof. Matteo Borrotti**

Tesi di laurea magistrale
Scienze statistiche ed economiche
19 January 2023

Outline of the presentation

Workflow


Data
Simulation of biased data
Bias in data
Fairness metrics
Sensitive feature S
tidymodels
tidyverse
APM
$\hat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{S})$
Discrimination aware classifier
MatchIt
DALEX
Explanation
-XAI
- Matching
Bias in prediction
yes
no
Bias removal
-Preprocessing techniques
-Postprocessing techniques
fairmodels
Tradeoff
Fairness/performance
Decision

Literature review & motivation

- HR analytics refers to the use of analysis, data, and systematic reasoning to make decisions regarding the people who are related to the organization [6].
- *Although data algorithms can help to avoid biased human decision-making, they also risk introducing new sources of bias. Algorithms built on inaccurate, biased, or unrepresentative data can produce outcomes biased along lines of race, sex, or other protected characteristics*[4].
- **The reputational-ranking algorithm utilized by a food delivery platform was deemed unfair by tribunale ordinario di Bologna** (2019). The definition of counterfactual fairness was found to be well aligned with the human conception of fairness (Piccininni 2022 [5]).


"L'algoritmo di Deliveroo è discriminatorio": sentenza del Tribunale di Bologna

Accolto il ricorso dei sindacati: "Precedente europeo"

Figure 1: bologna.repubblica.it, 02 GENNAIO 2021

Fairness metrics

Observational criteria: Fairness metrics

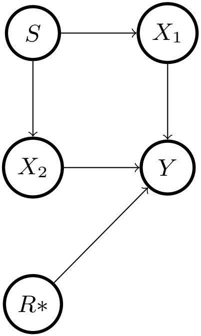
Equal Opportunity	$P(\hat{Y}=0 \mid Y=1, S=S_a) = P(\hat{Y}=0 \mid Y=1, S=S_d)$
Predictive Equality	$P(\hat{Y}=1 \mid Y=0, S=S_a) = P(\hat{Y}=1 \mid Y=0, S=S_d)$
Equalized Odds	$P(\hat{Y}=1 \mid Y=i, S=S_a) = P(\hat{Y}=1 \mid Y=i, S=S_d)$
Predictive Parity	$P(Y=1 \mid \hat{Y}=1, S=S_a) = P(Y=1 \mid \hat{Y}=1, S=S_d)$
Demographic Parity	$P(\hat{Y}=1 \mid S=S_a) = P(\hat{Y}=1 \mid S=S_d)$
AOD	$\frac{1}{2}[(FPR_{S_d} - FPR_{S_a}) + (TPR_{S_d} - TPR_{S_a})]$

Simulation of HR data

An algorithm is only as good as the data it works with [1].

- **Data**:The synthetic Dataset is composed of $n = 10000$ rows and $p = 12$ columns
- X_1 and X_2 represents the set of observable variables
- S is the sensitive feature: S_a for the advantaged group, S_d for the disadvantaged group
- Y is the binary target variable, $Y = 0 \rightarrow Y_{unfav}$ (35%) and $Y = 1 \rightarrow Y_{fav}$ (65%)
- $R*$ is the independent score

Figure 2: Relationship between variables, Directed Acyclic causal Graph



Simulation of HR data

Variable Name	Distribution	Formula	Link
S	$Binomial(\pi)$	$\pi = 0.2$	$identity$
Age	χ^2	$22 + \chi^2(1)$	$identity$
$Interview$	$Poisson(\lambda)$	$\lambda = f(age, S, \eta)$	$identity$
$GitHub\_account$	$Binomial(\pi)$	$\pi = f(S, \eta)$	$logit$
$Proxy$	$Normal(\mu, 2)$	$\mu = f(S, \eta)$	$identity$
$Proxy2$	$Beta(\alpha, \beta)$	$\alpha = f(proxy, age)$	$identity$
X_score	$Normal(\mu, \sigma)$	$\mu = 100, \sigma = 5$	$identity$
$Score$	$Poisson(\lambda)$	$\lambda = f(S)$	$identity$
$Simpson\_score1$	$Normal(\mu, \sigma)$	$\mu = f(S)$	$identity$
$Simpson\_score2$	$Normal(\mu, \sigma)$	$\mu = f(S)$	$identity$
Y	$Binomial(\pi)$	$\pi = f(.)$	$logit$

Bias in Data

Statistical Parity Difference (SPD) is defined as:

$$P(Y = 1|S = S_a) - P(Y = 1|S = S_d) \tag{1}$$

Disparate Impact (DI) is defined as:

$$\frac{P(Y = 1|S = S_d)}{P(Y = 1|S = S_a)} \geq 0.8 \tag{2}$$

the probability that an individual from the group S_d would get $Y = 1$ should be at least 0.8 times the same probability for an individual belonging to the advantaged group S_a.

SPD	**DI**
-0.4217392	0.4146449

Table 1: SPD and DI

APM

Test Data DI is 0.41, the goal is to find the best discrimination-aware classifier ($\hat{Y} = f(X, S)$)

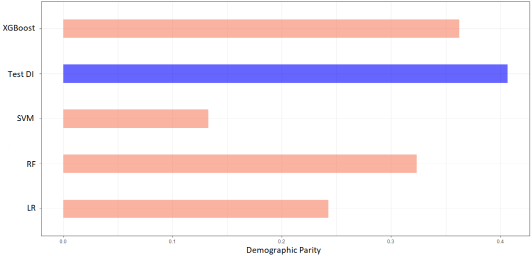


Figure 3: Demographic parity ratio for the models and disparate impact in the test data

Model	DP	Acc
XGBoost	0.36	0.85
LR	0.24	0.83
RF	0.32	0.84
SVM	0.13	0.76

Table 2: Demographic parity ratio and accuracy in test set for the models

XGBoost Model performance

Confusion Matrix								
All (Test set)			$S = 0$			$S = 1$		
	$Y = 0$	$Y = 1$		$Y = 0$	$Y = 1$		$Y = 0$	$Y = 1$
$\hat{Y} = 0$	711	171	$\hat{Y} = 0$	372	130	$\hat{Y} = 0$	339	41
$\hat{Y} = 1$	207	1412	$\hat{Y} = 1$	175	1303	$\hat{Y} = 1$	32	109
Fairness metrics								
Acc	0.849		Acc	0.846		Acc	0.860	
FNR	0.108		FNR	0.090		FNR	0.273	
FPR	0.225		FPR	0.320		FPR	0.086	
Eodds	1.117		Eodds	1.229		Eodds	0.813	
PPV	0.872		PPV	0.882		PPV	0.773	
DP	0.647		DP	0.746		DP	0.271	
TE	0.826		TE	0.743		TE	1.281	

Figure 4: Confusion matrix and fairness metrics by S XGBoost

XAI

Table 3: Test individual with $S = 1$ & $Y = 0$: predicted probability with XGBoost is $\mathbf{0.106}$.

S	Int	G_a	$Proxy$	$Proxy2$	Age	X_score	$Score$	S_s1	S_s2	Y
1	9	0	4.55	0.09	23	96	8	6.01	2.87	0

Figure 5: Shapley values

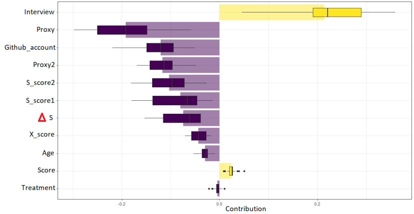
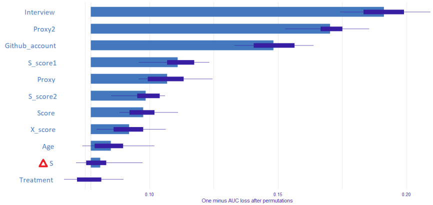


Figure 6: XGBoost Varible Importance Test set



Matching

Figure 7: Assessing Balance: ASMD Method=Full, distance=gbm, link= probit

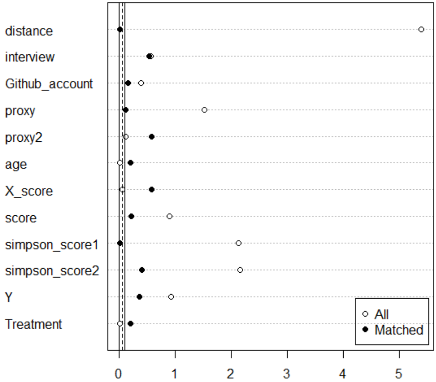
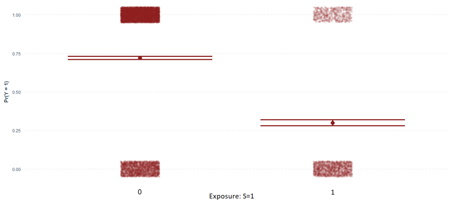


Figure 8: Whe Welch Two Sample t-test of variable Y by variable S in the matched sample revealed mean values of 0.72 and 0.299 for groups 0 and 1, respectively



	$Value$	Se	p_value
RR	-0.273	0.093	0.003**

Removing the Bias

1. **Preprocessing techniques**
   - Reweighting the data [2]

$S_a \wedge Y_{fav}$	0.879
$S_a \wedge Y_{unfav}$	1.313
$S_d \wedge Y_{fav}$	2.119
$S_d \wedge Y_{unfav}$	0.523

   - Disparate impact removal (DIR)
   - Uniform resampling
   - Preferential resampling with generalized least squares to estimate probabilities
2. **Post-processing techniques**
   - Reject Option based Classification pivot (ROC Pivot [3]) with $\theta = 0.1$ and *cutoff* $= 0.5$
   - Ceteris paribus cutoff for the subgroup $S = 1 : S = S_d$ set to 0.13

Removing the Bias

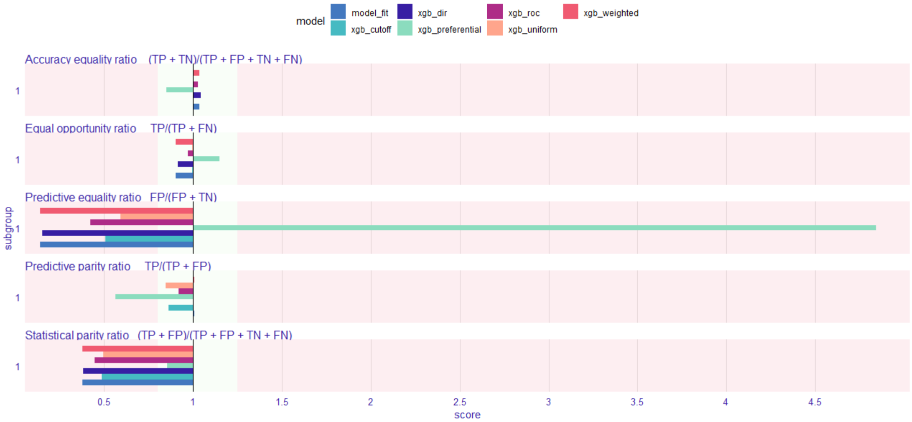


Figure 9: XGBoost bias reduction on training set

Tradeoff Fairness-Performance

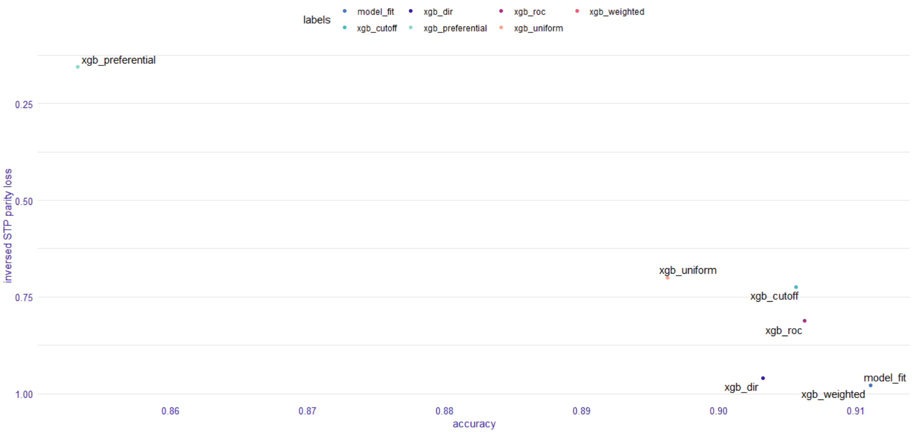


Figure 10: XGBoost bias reduction tradeoff performance-fairness

Conclusion

- **Model matters**: The performance of different discrimination-aware classifiers may vary when considering a protected class, highlighting the importance of selecting an appropriate model.
- It is important to understand the prediction of a black box model, particularly in a human resources context, so we also performed a explainable artificial intelligence (XAI) analysis.
- **Fairness** comes at the cost of performance.
- In order to address the various instances of unfairness that may occur during the human resource management process, it is essential to approach HR analytics from a multidisciplinary perspective.
- Future research could aim to utilize counterfactual methods in conjunction with domain expertise to further improve the analysis.

Grazie per l'attenzione

Thank You

Code and Data: https://github.com/DavideZulato/Tesi-2022

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[2] Faisal Kamiran and Toon Calders. "Data preprocessing techniques for classification without discrimination". In: *Knowledge and information systems* 33.1 (2012), pp. 1–33.

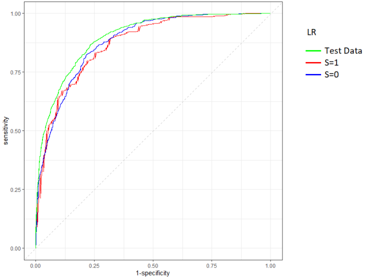
[3] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-Aware Classification". In: *2012 IEEE 12th International Conference on Data Mining*. 2012, pp. 924–929. DOI: 10.1109/ICDM.2012.45.

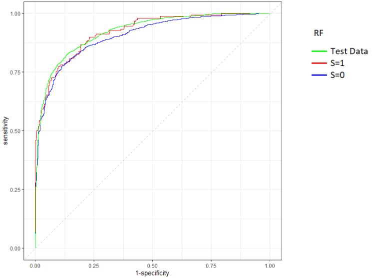
[4] Pauline T Kim. "Data-driven discrimination at work". In: *Wm. & Mary L. Rev.* 58 (2016), p. 857.

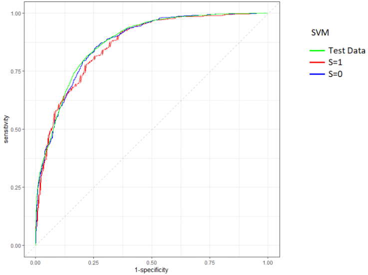
[5] Marco Piccininni. "Counterfactual fairness: The case study of a food delivery platform's reputational-ranking algorithm". In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.1015100. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1015100.

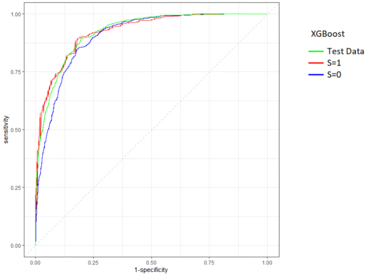
[6] Sjoerd Van den Heuvel and Tanya Bondarouk. "The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support". In: *Journal of Organizational Effectiveness: People and Performance* (2017).

ROCs by S for the models LR, RF, SVM, XGB









XGBoost details

Table 4: Optimal parameters for the XGBoost model when using different preprocessing techniques. Model tuning was performed using a 10-fold cross-validation on a grid 20×4

Preprocessing	**min_n**	**tree_depth**	**learning_rate**	**loss_reduction**
P1 Accuracy	38	11	0.0198722	0.1080567
P2 AUC	10	3	0.0705904	0.0662725
P3 Accuracy	20	12	0.0063106	$3.45 \cdot 10^{-6}$

P1 preserves the most the original data, P3 applies PCA with 5 principal componenets

Table 5: XGBoost performances P3

Metric	**estimate train**	**estimate test**
accuracy	0.91	0.85
bal_accuracy	0.87	0.83
specificity	0.93	0.89
precision	0.88	0.80
recall	0.82	0.78
kap	0.76	0.67

Model comparison

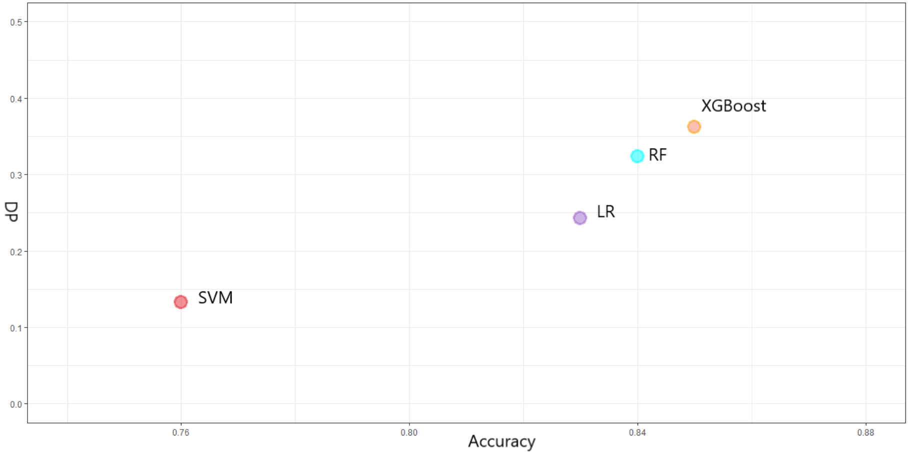


Figure 11: Accuracy and demographic parity in test set

Covariate balance

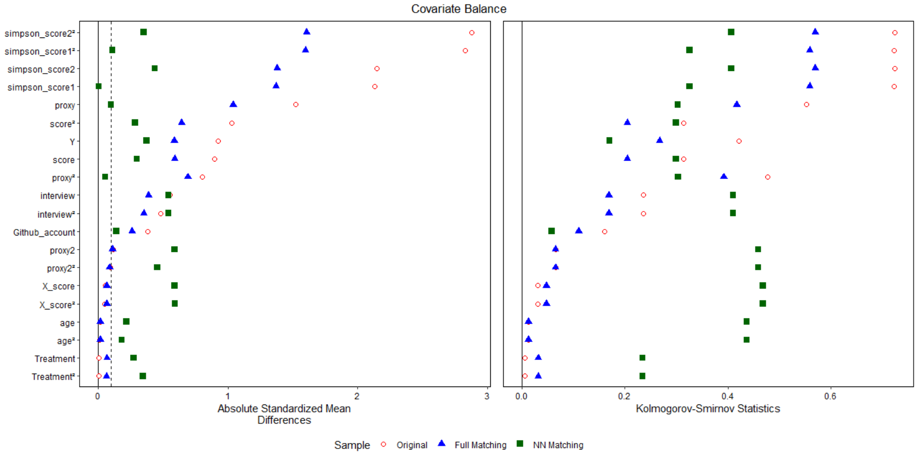


Figure 12: Matching: covariate balance comparison

Details on tradeoff

FPR	PPV	TPR	STP	Acc	Model
0.01	0.89	0.03	0.52	0.89	xgb_cutoff (S_d = 0.13)
0.87	0.91	0.03	0.81	0.91	xgb_roc
0.52	0.90	0.00	0.70	0.90	xgb_uniform
1.58	0.85	0.14	0.16	0.85	xgb_preferential
1.96	0.91	0.10	0.98	0.91	xgb_weighted
1.96	0.91	0.10	0.98	0.91	model_fit
1.89	0.90	0.09	0.96	0.90	xgb_dir

DIR in action

$$\bar{F}_s^{-1}(\alpha) = (1-\lambda)F_s^{-1}(\alpha) + \lambda(F_A)^{-1}(\alpha) \tag{3}$$

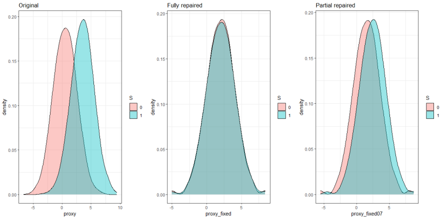


Figure 13: DIR for $Proxy$ kernel density plot

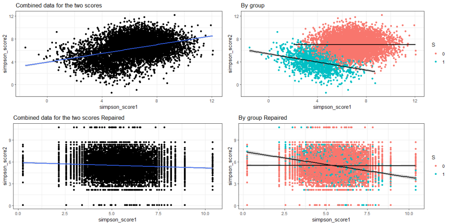


Figure 14: *Simpson's Scores* repaired $\lambda = 1$

CPC

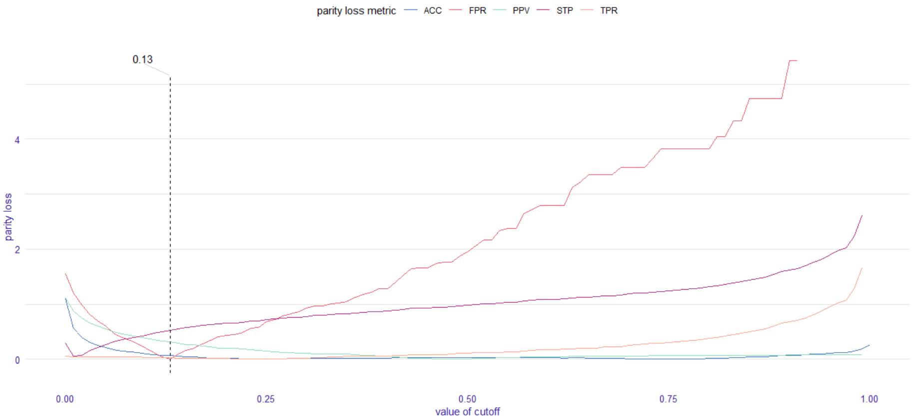


Figure 15: Ceteris paribus cutoff based on $S = 1$

Fairness in test set

Figure 16: XGBoost model train e test

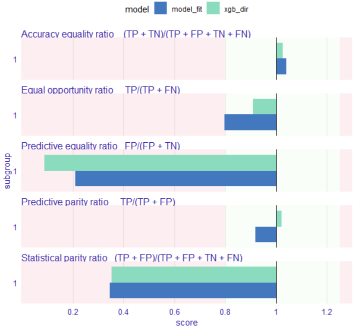


Figure 17: XGBoost DIR Test set $\lambda = 1$

GBM

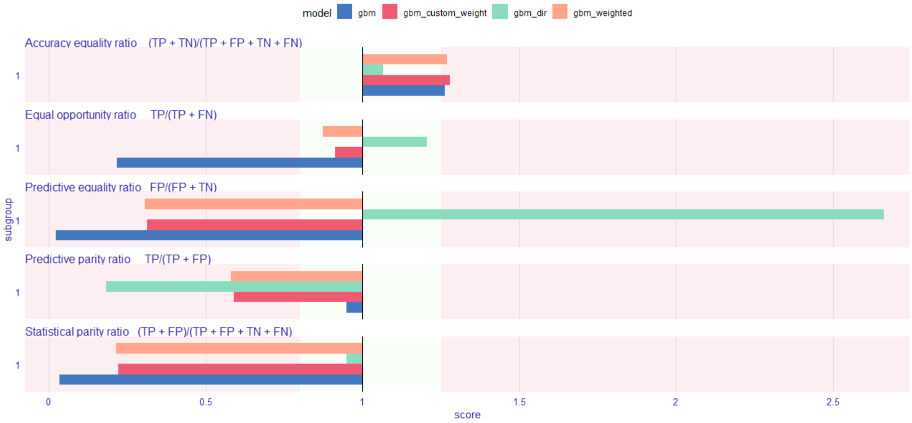


Figure 18: GBM reweighted and DIR