Interpretable and Fair Mechanisms for Abstaining Classifiers - Appendix

Daphne Lenders[1,2], Andrea Pugnana[3], Roberto Pellungrini[4], Toon Calders[1,2], Dino Pedreschi[3], and Fosca Giannotti[4]

[1] Adrem Data Lab, University of Antwerp, Antwerp, Belgium
[2] DigiTax, University of Antwerp, Antwerp, Belgium
daphne.lenders@uantwerpen.be
[3] KDD Lab, University of Pisa, Pisa, Italy
[4] KDD Lab, Scuola Normale Superiore, Pisa, Italy

A Illustrative Example of IFAC's Rejection Process

In Figure 1 we see how our selective classification model IFAC behaves on one instance $\mathbf{x}$ of ACSIncome. In this example, a base classifier predicts that a $\mathbf{x}$ has a low income with a probability of 74.17%. To decide whether to keep this original prediction, IFAC starts by analysing if the prediction falls under any global patterns of unfairness it has recorded. In this case, the instance falls under the group of women, working in Sales aged between 60 and 69, that is marked as potentially discriminated. The reason why it is marked as such is that on a separate dataset, the ratio of negative prediction labels for this subgroup is much lower when the sensitive part describing this subgroup (in this case their sex) is negated. To illustrate: on this separate dataset the base-classifier predicted a negative decision label 90% of the time for the group women, working in Sales and aged between 60 and 69, as opposed to 40% for the same group of *non-female* instances. Given this high difference, the first global fairness check has failed, and the rejector proceeds with an individual fairness analysis. Here it makes use of the Situation Testing algorithm, and compares the positive label ratios of $\mathbf{x}$'s most similar instances from the reference group (i.e. white men), with the positive label ratios of $\mathbf{x}$'s most similar instances from the non-reference group. In doing so, it can make a more fine-grained fairness analysis, and not just assess the classifiers' behaviour on the group of people working in Sales and aged between 60 and 69; but also take into account other features, like peoples' education level or marital status. We observe here that even if individuals are similar regarding all legally grounded features, their sensitive characteristics still influence the ratio of positive decision labels, which is 2/3rd for our reference group white men and 0 for our non-reference group. Because this difference is quite large the local fairness test fails and the overall prediction is deemed as unfair. To then decide whether to perform a fairness intervention or reject the prediction, the rejector checks if the prediction probability of 74.17% falls above $t\_unfair\_certain$. In this case, it does, meaning that our prediction is unfair but certain. Hence, the rejector rejects the original low-income prediction. As a next step, this rejection and the explanation behind why the original prediction was

considered unfair can be passed on to a human decision-maker. This person can use their domain knowledge as well as the explanation behind the rejection, to form a new decision for the instance in question. For instance, they may review the instances that were used for the similarity analysis in the individual fairness check, and determine if these instances were similar enough to the instance in question to draw discrimination conclusions from. Further, the list of subgroups that the classifier behaves favourably/discriminatory on can serve to increase an expert's general understanding of the base classifier, and may be even adapted by them to incorporate their domain knowledge.

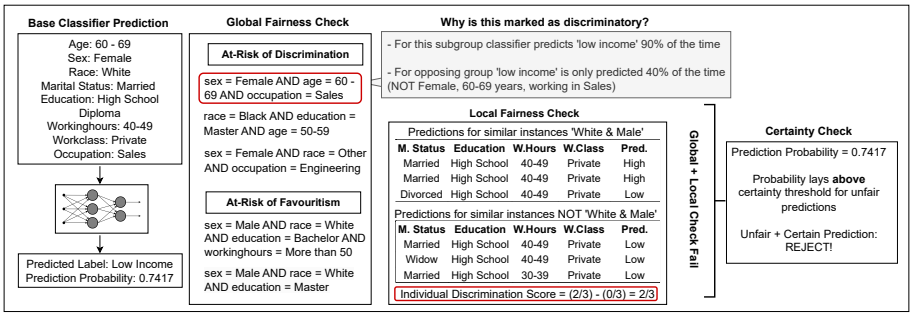

Fig. 1: An illustrative example of how a low-income prediction for a woman from ACSIncome is deemed as discriminatory and subsequently rejected by our model

B Proof: Setting slift threshold

In our methodology we select the discriminatory association rules used by IFAC, by checking for which of the rules the following property holds:

$$conf_{\mathbf{X}}((A,B) \rightarrow Y_v) - slift_{\mathbf{X}}((A,B) \rightarrow Y_v) < 0.5 \tag{1}$$

Which in the context of binary classification is true *iff*:

$$conf_{\mathbf{X}}((\neg A,B) \rightarrow Y_v) < conf_{\mathbf{X}}((\neg A,B) \rightarrow \neg Y_v) \tag{2}$$

Intuitively, this means that we only select the subgroups $\{A, B\}$ for which negating the sensitive part of the group ($\{\neg A, B\}$) yields a higher confidence for value Y_v w.r.t. the other value $\neg Y_v$.

Proof. Recalling the definition of $conf_{\mathbf{X}}((A,B) \rightarrow Y_v)$ as $P(Y_v|(A,B))$ we have that:

$$\begin{aligned} P(Y_v|(A,B)) - slift_{\mathbf{X}}((A,B) \rightarrow Y_v) &< 0.5 \\ P(Y_v|(A,B)) - (P(Y_v|(A,B)) - P(Y_v|(\neg A,B))) &< 0.5 \\ P(Y_v|(\neg A,B)) &< 0.5 \\ 2P(Y_v|(\neg A,B)) &< 1 \end{aligned} \tag{3}$$

For binary classification we can write $1 = P(Y_v|(\neg A,B)) + P(\neg Y_v|(\neg A,B))$ which yields:

$$\begin{aligned} 2P(Y_v|(\neg A,B)) &< P(Y_v|(\neg A,B)) + P(\neg Y_v|(\neg A,B)) \\ P(Y_v|(\neg A,B)) &< P(\neg Y_v|(\neg A,B)) \\ conf_{\mathbf{X}}((\neg A,B) \rightarrow Y_v) &< conf_{\mathbf{X}}((\neg A,B) \rightarrow \neg Y_v) \end{aligned} \tag{4}$$

C Full Fairness Results

In Table 1 and 2 we display the full fairness results for ACSIncome and WisconsinRecidivism for each classifier-methdology combination.

			M. Wh.	F. Wh.	M. Bl.	F. Bl.	M. Oth.	F. Oth.	Range	Std.
RF	FNR	FC	.33±.03	.57±.03	.57±.09	.60±.11	.44±.18	.59±.22	.27	.11
		UBAC	.26±.03	.54±.04	.61±.11	.67±.10	.30±.18	.54±.26	.40	.17
		IFAC	.37±.04	.44±.06	.57±.08	.49±.11	.41±.17	.52±.25	**.20**	**.08**
	FPR	FC	.24±.03	.10±.01	.12±.04	.05±.01	.08±.07	.05±.05	.19	.07
		UBAC	.20±.03	.06±.01	.07±.03	.02±.01	.07±.08	.03±.04	.18	.07
		IFAC	.18±.03	.11±.01	.10±.04	.04±.02	.08±.07	.05±.05	**.14**	**.05**
	Pos. Ratio	FC	.43±.02	.17±.01	.17±.03	.09±.01	.18±.07	.13±.07	.34	.12
		UBAC	.43±.03	.13±.01	.12±.03	.05±.02	.16±.07	.10±.07	.38	.13
		IFAC	.36±.02	.20±.01	.16±.03	.09±.02	.17±.08	.15±.07	**.27**	**.09**
NN	FNR	FC	.34±.03	.52±.04	.60±.08	.69±.09	.40±.22	.56±.22	.35	.13
		UBAC	.24±.04	.56±.06	.63±.09	.75±.10	.38±.22	.42±.26	.50	.18
		IFAC	.35±.04	.47±.07	.60±.08	.60±.14	.38±.22	.44±.29	**.25**	**.11**
	FPR	FC	.19±.02	.06±.01	.07±.03	.03±.01	.04±.04	.07±.04	.16	.06
		UBAC	.15±.02	.03±.01	.04±.03	.01±.01	.02±.03	.03±.04	.13	.05
		IFAC	.13±.01	.06±.01	.06±.03	.03±.02	.02±.03	.07±.04	**.11**	**.04**
	Pos. Ratio	FC	.40±.02	.15±.01	.14±.03	.07±.01	.15±.05	.16±.05	.34	.11
		UBAC	.40±.02	.09±.01	.10±.03	.03±.01	.12±.06	.11±.06	.37	.13
		IFAC	.33±.02	.15±.01	.12±.03	.07±.01	.12±.06	.15±.05	**.27**	**.09**
XGB	FNR	FC	.29±.03	.57±.05	.57±.09	.62±.07	.36±.14	.52±.25	.33	.13
		UBAC	.20±.03	.62±.07	.65±.12	.80±.08	.16±.16	.43±.28	.65	.26
		IFAC	.33±.03	.47±.06	.61±.10	.62±.11	.38±.15	.40±.26	**.29**	**.12**
	FPR	FC	.19±.02	.05±.01	.07±.02	.04±.01	.08±.05	.03±.04	.16	.06
		UBAC	.14±.02	.02±.01	.03±.02	.02±.01	.03±.04	.02±.02	.12	.05
		IFAC	.11±.02	.06±.01	.06±.01	.03±.01	.06±.06	.02±.04	**.09**	**.03**
	Pos. Ratio	FC	.42±.02	.13±.01	.14±.02	.08±.02	.19±.06	.13±.07	.34	.12
		UBAC	.41±.02	.08±.02	.09±.03	.04±.01	.15±.07	.10±.06	.38	.14
		IFAC	.32±.02	.15±.01	.12±.03	.06±.02	.16±.06	.13±.07	**.27**	**.09**

Table 1: Full Fairness Results Income Prediction

			White	**Black**	**Other**	Range	Std.
RF	FNR	BC	.20 ± .01	.34 ± .02	.26 ± .02	.14	.07
		USC	.14 ± .01	.27 ± .02	.25 ± .02	.13	.07
		FSC	.14 ± .01	.24 ± .02	.24 ± .02	.10	.05
	FPR	BC	.61 ± .02	.51 ± .02	.55 ± .05	.09	.05
		UBAC	.66 ± .02	.53 ± .03	.54 ± .05	.13	.07
		IFAC	.64 ± .02	.56 ± .03	.56 ± .06	.08	.05
	Pos. Ratio	FC	.72 ± .01	.59 ± .01	.65 ± .03	.13	.07
		UBAC	.79 ± .01	.63 ± .02	.66 ± .03	.15	.08
		IFAC	.77 ± .01	.66 ± .02	.67 ± .03	.11	.06
NN	FNR	FC	.22 ± .01	.38 ± .02	.30 ± .02	.17	.08
		UBAC	.20 ± .01	.34 ± .02	.27 ± .02	.14	.07
		IFAC	.20 ± .01	.33 ± .02	.26 ± .02	.13	.06
	FPR	FC	.58 ± .02	.44 ± .02	.51 ± .06	.14	.07
		UBAC	.56 ± .02	.42 ± .02	.50 ± .05	.14	.07
		IFAC	.55 ± .02	.43 ± .02	.51 ± .05	.12	.06
	Pos. Ratio	BC	.70 ± .01	.53 ± .01	.62 ± .03	.17	.09
		UBAC	.71 ± .01	.55 ± .01	.63 ± .03	.16	.08
		IFAC	.70 ± .01	.56 ± .01	.64 ± .03	.14	.07
XGB	FNR	FC	.20 ± .01	.33 ± .03	.26 ± .02	.14	.07
		UBAC	.14 ± .01	.28 ± .02	.23 ± .02	.14	.07
		IFAC	.14 ± .01	.28 ± .02	.23 ± .02	.14	.07
	FPR	FC	.60 ± .01	.46 ± .03	.57 ± .03	.15	.07
		UBAC	.65 ± .02	.47 ± .04	.51 ± .03	.18	.09
		IFAC	.64 ± .02	.46 ± .04	.51 ± .03	.18	.09
	Pos. Ratio	BC	.72 ± .01	.56 ± .02	.67 ± .02	.16	.08
		UBAC	.78 ± .01	.60 ± .02	.66 ± .02	.18	.09
		IFAC	.78 ± .01	.60 ± .02	.67 ± .02	.18	.09

Table 2: Full Fairness Results Recidivism Prediction

D WisconsinRecidivism Results with Less Strict Unfairness Selection

In Figure 2 we see the results of a Random Forest classifier combined with the different abstention methods on WisconsinRecidivism. For the local fairness check as executed with Situation Testing we now set the threshold t to 0.0. Intuitively this means, that regardless of the local fairness results any instance falling under a global pattern of discrimination will be considered as unfair (the situation testing results can still be used as extra information for a human reviewer). We see here that with this less strict unfairness selection, IFAC reduces FNR, FPR and PDR differences across demographics more than when using t = 0.3.

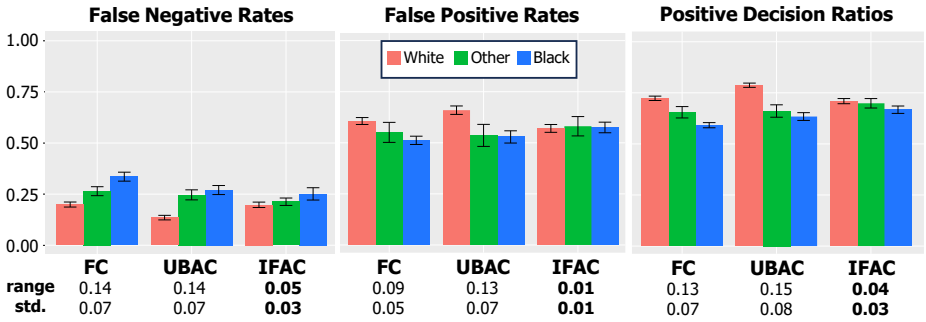

Fig. 2: Caption

E Effects of c and w_u

In Figure 3 we display the effects of both the coverage parameter c and the unfair-reject-weight w_u on the accuracy as well as the fairness of our abstention method IFAC. We compare the results with a regular uncertainty based abstaining classifier (UBAC) and a full covage (FC) one.

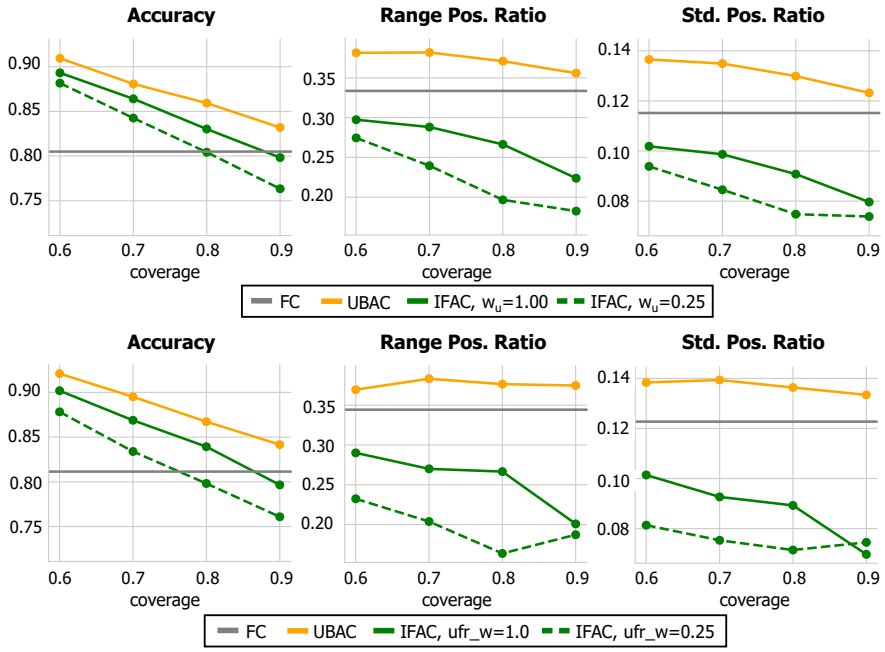

Fig. 3: ACSIncome effect of different values for c and w_u on abstention methods combined with Neural Network (above) and XGBoost