Assignment 1
Algorithmic fairness, accountability and ethics

March 3, 2025

Group:
Malte Sauerland Paulsen - saue
Michella Ravn Søndergaard - micso

Task 1 Classifiers and fairness considerations

Feature engineering
From the open source dataset *folktables* we want to fit models predicting whether or not an individual's total income is above 35.000$. We construct a dataset including the following features, where we highlight the groups of interest when considering fairness:

Categorical:

- SEX: Binary value for male/female
- DIS: Binary value for disability status (with disability/without disability)
- VPS: Binary value for veteran/not veteran
- MAR: Binary value for marital status (married/not married)
- HINS1: Binary value for insurance (government assistance plan)
- HINS2: Binary value for insurance (purchased directly)
- HINS4: Binary value for insurance (through employer or union)
- CIT: 4 groups for citizenship status
- COW: 8 groups for class of worker
- RAC1P: 9 groups for race codes*

**Race could also be interesting, however, we choose to focus on the highlighted.*

Scale:

- AGEP: Age, adults only (17 to 94 years)
- SCHL: Educational attainmant approximate in years (20 lvl range pre school to doctorate)
- ENG: English abilities (5 lvl range limited/none to professional/native)

We choose to **group** in **MAR** feature, such that *'widowed'*, *'separated'*, *'never married'* and *'divorced'* are all seen as *'not married'*. We choose to **group** in **COW** such that all the governmental employees are in one group. All the categorical features are one-hot encoded, as we want to ensure we do not mathematically imply a relationship within the groups.

Age, school and english abilities are scales. Naturally age is a scale of years lived. For school we are given 24 categories, which approximately can be seen as years of school attended.

Approximately as we have some steps being 2-3 years. Within the given school data we choose to **group** *'GED or alternative credential'* with *'Regular highschool'* and *'Master's degree'* with *'Professional degree beyond a bachelor's degree'*. We set **ENG** to be a scale representing how well the person is able to speak english. We choose to **standardize AGE, SCHL** and **ENG** to range between 0 and 1, mainly to avoid numerical issues.

Accuracy
We fit two models, a white box model - Logistic Regression (LR) and a black box model - Random Forest(RF). For each of these fitted models we get corresponding accuracy.

Accuracy	
LR-model	RF-model
76.1%	76.8%

Statistical Parity and Equalised odds
The plots below visualise the statistical Parity (SP) by ratios and Equalised Odds (EO) by True and False positive rates, of our white- and black-box model.

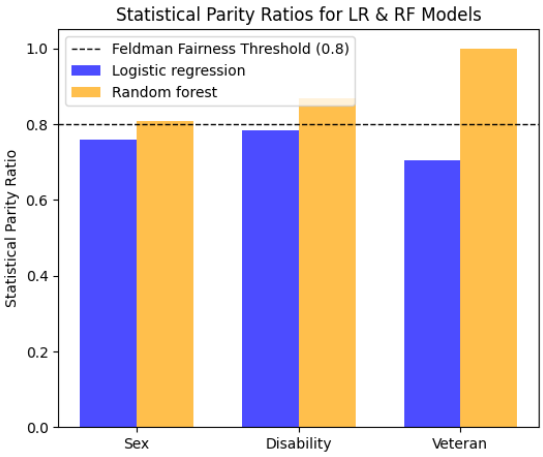

Figure 1: Statistical Parity ratios

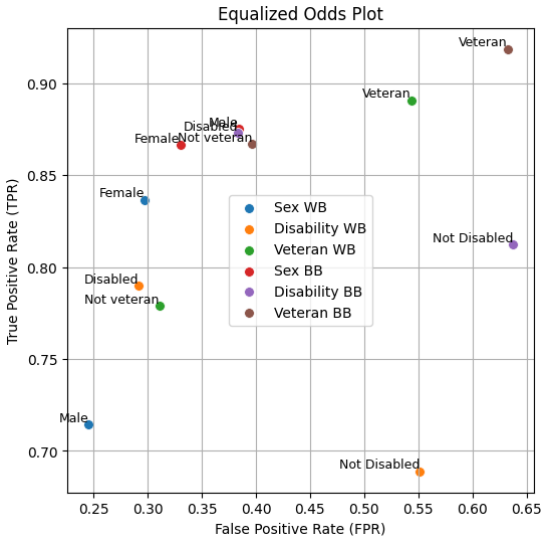

Figure 2: Equalized Odds

As seen in Fig.1, our LR model fail to achieve SP if we use the Feldman Fairness Threshold of 80% discussed in Lecture 2. However our RF model performs better for all groups, slightly for **SEX** and more for **DIS** and **VPS**.

EO is characterized by equal false positive rates and true positive rates within a group. This should result in the data-points for each group and each model (e.g. male/female) in Fig. 2 to lay on top of eachother. This is not the case - and therefore we can conclude that EO is not realised by our classifiers.

Equal statistical parity
We post-process our model output, using the fairlearn library to flip predictions in our predicted data, to achieve statistical parity. We do so for each of the groups **SEX**, **DIS**, and **VPS** and for all three groups at the same time. We observe that the accuracy of the models fluctuate slightly.

	LR	RF
None	76.1%	76.8%
SEX	76.4%	76.4%
DIS	76.1%	76.9%
VPS	76.0%	76.8%
ALL	76.1%	76.3%

Task 2 Explaining models using SHAP

SHAP values for the classifiers

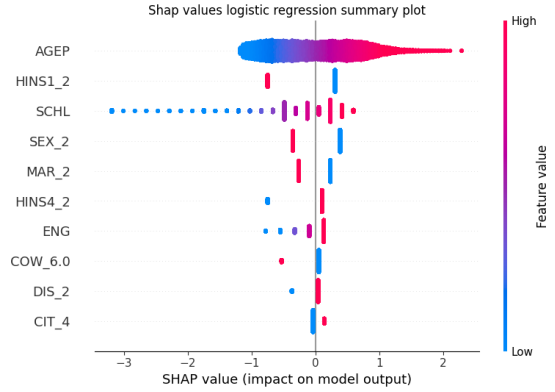

Figure 3: SHAP values for LR

Figure 4: SHAP values for RF

Looking at Fig. 5 and 6, it immediately becomes apparent, that the SHAP values and feature values differ between models. However, we see somewhat the same features having high impact on the prediction. For both models we observe that the age, amount of schooling, sex and insurance-status are the most important predictors when predicting income. The models mostly differ in the numerical size of the shap-values, with the values being much lower for the RF model. This indicates a more balanced impact from all the features. The complexity of the Random Forest balances the impact of the features, since the classification is done through multiple steps, instead of one computation. The simplicity of the linear regression model, in contrasts, results in a few features having a much higher weighting in the prediction.

Which classifier is most suited?
Considering the accuracy, neither of the models are very precise, (≈75% accuracy). Interestingly, the RF model produced a more equal SP ratio at the initial fitting. And both the LR model and RF model somewhat retained their accuracy after post-processing to achieve full SP.

Considering the minimal potential gain from using the Random forest model, We argue that the Linear regression model is more suited for the prediction task, due to the explainability that a white-box model inherently posess. When other metrics are largely the same, we argue that the model that is the most explainable is the best fit for the task.