Algorithmic Fairness, Accountability, and Ethics:

Mandatory Assignment 1

Course code: KSALFAE1KU

Chrisanna Kate Cornish ccor@itu.dk
Christian Margo Hansen chmh@itu.dk

Spring Semester 2024

IT UNIVERSITY OF COPENHAGEN

Associated code can be found here: *Equalised Odds Github*

1 Classifiers and Fairness Considerations

Pre-processing

Following the initial pre-processing step form the provided template, sex of the individual was recoded as Male= 0, Female= 1, and health insurance status changed from the numerical labels of 1 for "yes" and 2 for "no" into 1 and 0 respectively. The resulting dataset was left as it was for use in a Random Forest classifier, and a copy further processed for use in a Logistic Regression classifier. Age was scaled with SKLearn's Min-Max Scaler, to avoid being unfairly punished by L2 regularisation. The remaining categorical values were appropriately one-hot encoded, and a selection of these removed to establish a baseline for the categorical features (Table 1).

Omitted Feat.	Description
CIT_1	Born in the US
COW_1	Privately employed
MAR_1	Married
RAC1P_1	White alone
ENG_nan	Speaks only English
SCHL_16	Regular high school diploma

Table 1: Omitted baseline features

Model Accuracy

Each data set was divided using a train-test split, and the classifiers trained on their respective data sets. Both models achieved an accuracy of 77% (Figure 1). This was chosen as an acceptable starting point of the investigation, as the focus of this work is not in maximizing model performance, but rather on investigating explainability, and fairness and its affect on performance. One interesting difference, is that while both models have near identical accuracy, the logistic regression classifier classifies more records to the positive class.

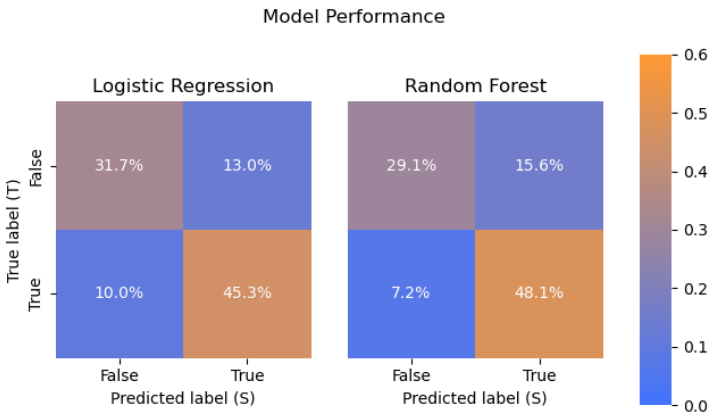


Figure 1: Baseline performance of classifiers

Fairness Metrics

Fairness of the models was evaluated through the metrics of statistical parity, equalised odds, and equalised outcomes. Figure 2 shows that Group 0 tends to be favoured by the logistic regression model. We chose to address this

Group	Base Acc	New Acc
Group 0	0.79	0.77
Group 1	0.75	0.73
Overall	0.77	0.75

Table 2: Accuracy changes

by focusing on the equalised odds metric. Plotting the
ROC curves for both groups revealed an intersection point, which indicated that achieving this fairness metric was possible. Different success thresholds were applied to either group, to align their TPR and FPR as much as possible; for Group 0, we used 0.397, and for Group 1, 0.311. This, not unsurprisingly, reduced the accuracy for each group slightly (Table 2). We also found this slightly improved the statistical parity.

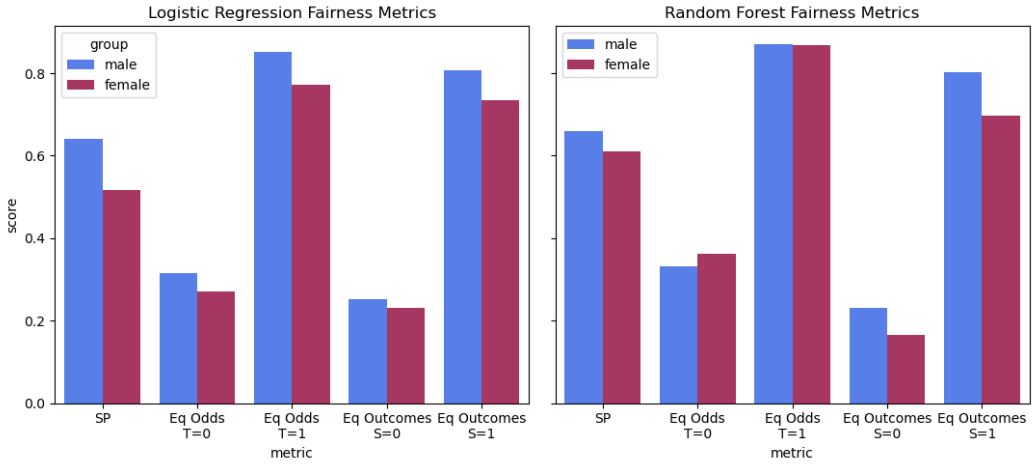


Figure 2: Summary of fairness measures. A pair of equal bars indicates fairness for that particular measure

2 Explaining the White-box Model

We calculated the odds ratio for each feature based on the model's coefficients (Table 3). This showed that age (*AGEP*) was overwhelmingly the most significant feature in determining the prediction. This fits with our intuition, that the older someone is, the more likely they are to be earning a higher wage, due to experience and seniority. Other important features included the higher education levels, and again this makes sense as it (hopefully) makes higher earning jobs more attainable.

Feature	Coef.	Odds
AGEP	2.877	17.759
SCHL_24.0	2.015	7.501
SCHL_23.0	1.986	7.286
SCHL_22.0	1.797	6.030
SCHL_21.0	1.296	3.655
...	...	...

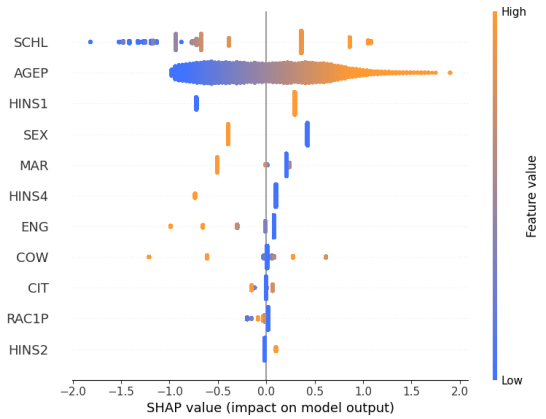
Table 3: Sample of logistic regression model weights

Exploring Counterfactual Data Points

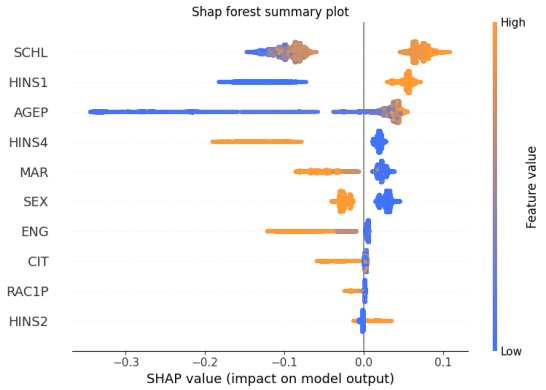
To further explore the influence of the *AGEP* feature, we selected a data point that had narrowly failed to satisfy the probability threshold. The age of that individual was increased by an arguably insignificant amount, from 26 to 26.008, which successfully flipped the decision from *False* to *True*. This effectively highlights the risk of the model placing such importance on a single feature. The case could also be made that age in itself is not a

significant feature, but due to the correlation it has to other influential features such as education, its importance is erroneously inflated.

3 Model-Agnostic Explanations



(a) SHAP summary plot for LR model



(b) SHAP summary plot of RF model

Figure 3: SHAP summary plots for each model

Summary plots of the SHAP[1] values were made, to illustrate the influence of each feature in either model. These findings mostly aligned with those from the logistic regression coefficients; notably that age is a relatively powerful determining feature, with larger age values primarily having a positive effect, as clearly shown in Figure 3b. Other than that, higher attained forms of education are also among the more influential features. The added benefit of the SHAP summary plots, is the indication of the varying impact a feature has based on its value.

4 Reflection

Taking both the performance and fairness metrics into consideration, it can be argued that the logistic regression model is the more appropriate choice. The baseline fairness metrics are not significantly different to make one model clearly superior over the other. That being said, the ability to tweak the thresholds of the logistic regression model in order to satisfy specific fairness metrics is a vital advantage. Additionally, it is much easier to investigate individual data points within the logistic regression model, and understand why a certain decision was made, and tweak the values that might sway said decision one way or the other. The extent of the transparency and explainability enabled by this model is why we consider it to be the better choice in this context.

[1] SHapley Additive exPlanations