IT UNIVERSITY OF COPENHAGEN

Mandatory Assignment 1: Fairness something

Constantin-Bogdan Craciun
cocr@itu.dk

Gino Franco Fazzi
gifa@itu.dk

Veron Hoxha
veho@itu.dk

March 11th, 2024

Introduction

This report outlines the findings of employing two classification models to predict whether an individual's total income will surpass $35,000, using data from the 2018 US Census focused on California. Additionally, we assess the fairness of these models concerning gender-specific groups. The report is structured into sections addressing tasks and classifiers.

1 Classifiers and fairness considerations

1.1 Classifiers

White Box model, Logistic Regressor: The first classification model employed in this study was Logistic Regression. Categorical variables were converted into numerical ones using one-hot encoding to facilitate mathematical operations necessary for this model. Additionally, feature scaling (Standardization) was performed to enforce zero-mean and unit-variance for all feature values using the formula: $z = \frac{x-\bar{X}}{s}$, where $\bar{X}$ is the mean of the training samples, and s the standard deviation. This is to ensure uniform ranges across features, aiding algorithm convergence and coefficient comparison. This model achieved an accuracy of **0.77** on the test set.

Black Box model, Random Forest: The second classification model utilized in this study was a Random Forest Classifier. Following similar preprocessing steps as with the logistic regression model, categorical variables were converted into numerical values using one-hot encoding. Unlike logistic regression, scaling of features was unnecessary as Random Forest models are scale-invariant.
This model achieved an accuracy of **0.78** on the test set.

1.2 Fairness metrics

When looking into gender as our group of interest, and analysing fairness across these groups, we see that the neither model achieves any of the fairness metrics (statistical parity, equalized odds, equalized outcomes). Detailed results can be seen in Table 1.

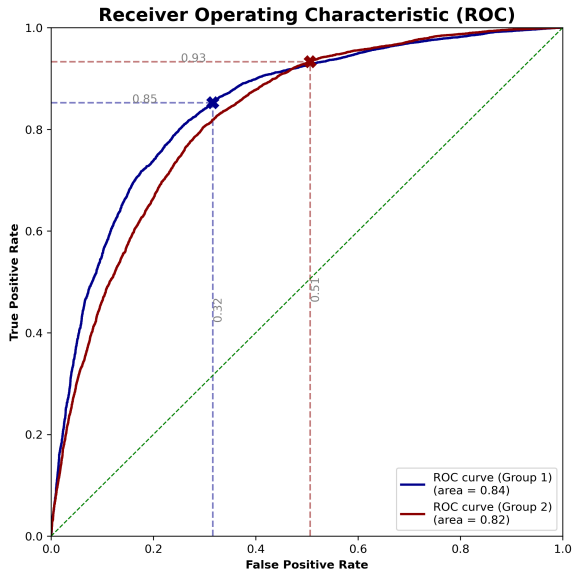
1.3 Intervention Post-Hoc

To fulfill statistical parity on the outcome of the classifiers, a post-processing intervention is applied to adjust predictions for the group with fewer positive outcomes, aligning them with those of the group with more positives. This adjustment is made based on the classifier's probability estimates, favoring predictions closer to the decision threshold[1]. In the case of our Logistic Regression model, the model predicted 13,254 positives for males and 9,615 positives for women (3,640 less). Looking

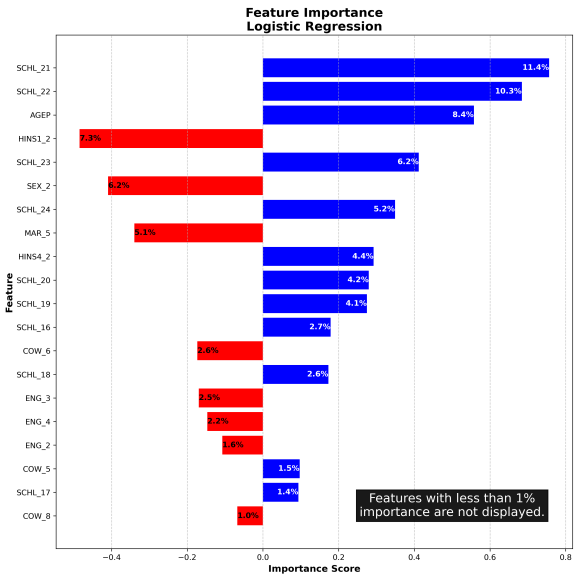
[1]Note that this is equivalent to reducing the threshold for the group with less positive predictions.

Statistical Parity		
Model	**P(Pred=True | Men)**	**P(Pred=True | Woman)**
LR	**0.643**	0.519
RF	**0.620**	0.544
Equalized Odds		
Model	**True Positive Rate (TPR)**	**False Positive Rate (FPR)**
LR	Men: **0.853** - Women: 0.776	Men: **0.316** - Women: 0.272
RF	Men: **0.839**- Women: 0.808	Men: 0.278 - Women: **0.289**
Equalized Outcomes		
Model	**P(Label=True | Sex, Pred=True)**	**P(Label=True | Sex, Pred=False)**
LR	Men: **0.808** - Women: 0.733	Men: **0.251** - Women: 0.229
RF	Men: **0.824** - Women: 0.729	Men: **0.258** - Women: 0.207

Table 1: Fairness metrics for Logistic Regression (LR) and Random Forest (RF) classifiers. None of the fairness metrics are satisfied.



(a) ROC Curve after statistical parity intervention. Men (blue line), with current probability threshold at 0.5 (TPR=0.84, FPR=0.28). Women (red line), with probability threshold at 0.26 (TPR=0.96, FPR=0.55).



(b) Feature importance for Logistic Regression model. On the X axis the coefficient score. On the bars the percentage of the absolute value of the coefficient in the model.

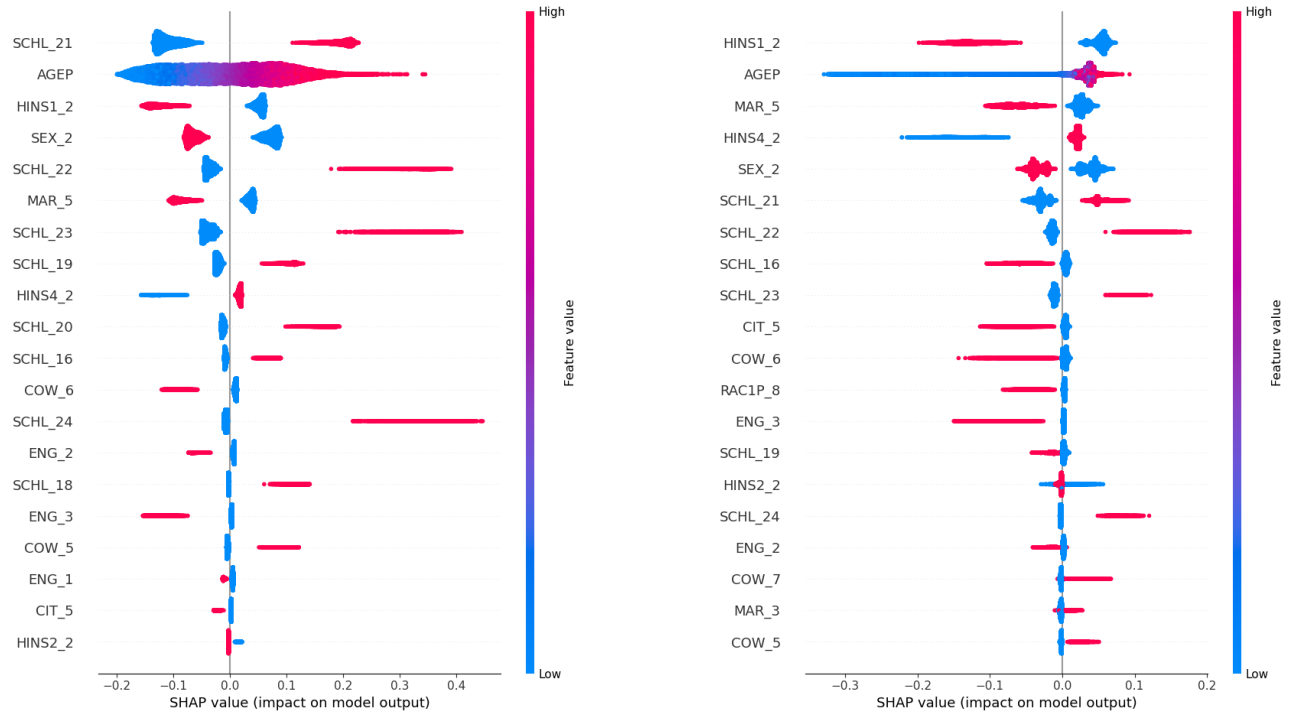
Figure 1

at the predicted probabilities outputted by the model, we flip the prediction of the first 3,640 negatives (probabilities in descending order), resulting in a total positive predictions of 13,254 for both genders. The underlying assumption is that adjusting predictions closest to the threshold mitigates the unfairness introduced by altering predictions for individual observations.
While this post-hoc method achieves statistical parity, it comes at the cost of a two-point reduction in accuracy. Moreover, there is an observed increase in the True Positive Rate (TPR), indicating a higher rate of correctly identified positive cases, but also an increase in the False Positive Rate (FPR), indicating a rise in misclassifications by the model. The resulting ROC curve after this intervention can be seen in Figure 1a.

2 Explaining white-box models

To decipher the outcomes of the logistic regression model, we examine the coefficients of the trained model, focusing on the most influential predictors (see Figure 1b. The two primary predictors are associated with education levels. Specifically, possessing a Bachelor's Degree (SCHL_21) or a Master's Degree (SCHL_22), compared to having no education (baseline: No schooling completed), positively correlates with predictions of high income. While this aligns with our expectations, we note that the weight of a Bachelor's degree exceeds that of a Master's degree, contrary to our initial assumptions (both of which rank significantly higher than holding a PhD). The third feature in descending importance is Age. This also follows our intuition, since we can associate age as a factor of increased experience and accumulated capital, and thus higher income. In fourth place we found the first negatively associated feature: HINS1_2 (lack of insurance coverage through a current or former employer or union). This observation resonates with our understanding, as the absence of such insurance may suggest unstable employment. A final mention is to sex (ranked 6^{th} with 6.2% importance), where being a woman affects negatively the probabilities of high income.

Counterfactual example: Using the feature importance ranking, we identify two observations with minimal variance in their features (from most to least relevant) but differing prediction outcomes. This examination leads us to observation 374998 (predicted Positive) and observation 140871 (predicted Negative), which are identical except for one feature: HINS2_2 (lack of insurance purchased directly from an insurance company). Thus, if observation 140871 had acquired insurance from a company, it would have been classified as Positive.



(a) Logistic Regression classifier.

(b) Random Forest classifier.

Figure 2: Summary plots (SHAP module) with most important features on average.

3 Model-agnostic explanations

Using a permutation-based approach, we calculate the Shapley values[2], and we ascertain that, in the Logistic Regression model, the feature SCHL_21 (Bachelor's Degree) primarily drives positive predictions, consistent with findings in Section 2. Subsequently, Age (AGEP) emerges as the second most influential feature, indicating that younger individuals are less likely to surpass $35k, aligning with earlier observations. However, SCHL_22 (Master's Degree) ranks lower in importance on average (5th place). In contrast, in the Random Forest model, HINS1_2 (lack of insurance through employer or union) becomes the most influential factor for predicting negative outcomes, while Age remains a significant force behind positive predictions. Detailed summary plots are available in Figure 2[3]. Notably, both models highlight being identified as female (SEX_2) as one of the top 5 drivers for negative outcomes, corroborating findings from logistic regression coefficients and contributing to understanding the observed fairness metrics discrepancies.

4 Reflection

This study underscores the critical importance of auditing decision-making models for fairness. In this investigation, we uncovered the inherent biases within our models, revealing a troubling association between being female and lower income. While this may reflect broader societal inequities, it is imperative to rectify such biases within our models.

Our study's conclusions highlight the superiority of White-Box models for the task at hand. Despite their marginally lower accuracy, the significant gains in explainability make them preferable. Even with the aid of model-agnostic explanation tools, the complexity of models like Random Forest ensembles renders their results challenging to interpret. The intricate interactions between features and inherent randomness in their outcomes pose significant obstacles for human comprehension, further emphasizing the value of transparent and interpretable models in decision-making contexts.

[2]This was done using the SHAP module. See documentation here.

[3]A wider catalogue of Shap plots can be found in the repository code.