Adversarial Debiasing –

Adversarial debiasing involves two neural networks – one that is trained on the original recidivism data and attempts to predict accurate sentencing, another that attempts to predict the sensitive variable S, here, race. The goal would be for cross-entropy loss, or the dissimilarity between predictions and results in a logistic regression, to be small for the first algorithm (L1), to preserve accuracy, and large for the second algorithm (L2), to preserve fairness. The objective loss function is therefore proportional to positive L1 and proportional to negative L2. Wadsworth’s particular dataset improved both in fairness and accuracy with adversarial debiasing. Brian Hu Zhang et al. identified theoretical conditions for convergence between the results of the two algorithms and found that if the dataset converges, it must satisfy the new fairness constraint. However, a significantly larger dataset than the one I had access to was required for convergence.

Regularization –

Regularization requires a series of matrix multiplications with different probabilities carried out relative to the non-sensitive variables and the sensitive variable to create an objective function that would minimize redundant information present in both the sensitive variable and the outcome. This process is only possible for logistic regression, as it requires calculations of the probability of the target variable. Because this objective function requires, essentially, penalizing violations of conditional procedural accuracy equality (equalized errors relative to actual sentencing values), it might trade off with conditional use accuracy equality (equalized errors relative to predicted sentencing values), which might create worse predicted sentences across the board. Richard Berk et al. raise concerns that a concave objective function might return a localized solution that does not maximize fairness, and every case of unfairness is often difficult to quantify such that an overzealous regularizer might put a thumb on the scale.

Re-weighting –

Re-weighting explicitly reverses the implicit effect of race in algorithms, which requires that this weight be quantified. Faisal Kamiran and Toon Calders describe a form of positive bias that results when the expected probability is higher than the observed probability of, in this case, a high-risk designation. The inverse is true for negative bias. The new weight would target those divergences from expectation by multiplying each weight with the ratio of expected to observed probability. Berk et al.’s criticism of intervention applies here, but they suggest a workaround that involves ranking inputs from low to high certainty. Low certainty inputs can be reversed with little accuracy differential, so they might be more suitable for alteration. This is the approach I used to correct for bias.