Milestone 5

For modelling the prediction task on COMPAS dataset, we initially implemented a neural network with 3 layers; however, we realized that the neural network model is unnecessarily complex given the smaller size of our dataset (about 9000 records) and the fact that logistic regression can find the global minimum due to its convex nature as opposed to non-convex nature of neural network. With neural network, we obtained best accuracy of roughly 71% on both training and test dataset, whereas, with logistic regression model, we obtain an accuracy of roughly 76% on both train and test dataset. Before applying the debiasing algorithm, we obtained the following confusion matrix on validation dataset (15% of records) with a logistic regression classifier i.e.

Validation Confusion Matrix Without Debiasing

African-Americans	
TN:464	FP:34
FN:192	TP:139
FPR=0.068, FNR=0.594	

Caucasians	
TN:400	FP:12
FN:112	TP:44
FPR=0.021, FNR=0.684	

We note that the model has a higher false positive rate (0.068 vs 0.021) and lower false negative rate (0.594 vs 0.684) for African-Americans than Caucasians. In other words, both of these differences in error rates are biased against African-Americans since negative outcome i.e. not recidivate is a favorable outcome.

We haven't yet finished implementing our logistic regression with adversarial debiasing but here is how we will be implementing it. First, we will encode output of adversary $\hat{z}$ (prediction of the sensitive attribute which in our case will be the race variable) as follows:

$s = \sigma((1 + |c|)\sigma^{-1}(\hat{y}))$ $\qquad\qquad$ $\hat{z} = w_2.[s, sy, s(1-y)] + b$

Where w_2 are the adversarial parameters to learn. We aim to update the parameters W of the predictor model using the following equation:

$\nabla_W L_P - \alpha \nabla_W L_A - \text{proj}_{\nabla_W L_A} \nabla_W L_P$

where L_P, L_A denote predictor's and adversary's logistic loss respectively. At each training step, we will simultaneously update parameters of predictor and adversary and will stop when updates lead to very small changes in the loss functions.

After meeting with Professor Sontag today about our project, we were given suggestions to make our project include more depth and background. We have largely focused on adversarial debiasing as an alternative to logistic regression; Professor Sontag recommended that we take a look at a recent paper on equal opportunity and affirmative action to explore other methods and have more contextualization. In addition, Professor Sontag suggested us to consider other datasets outside of COMPAS, as recent papers reveal that modeling on just the 1-2 most relevant features and discarding all other information also performs well.