Classification models can *increase* bias found in the training data. Consider the example of a univariate tree-based model, trying to predict whether a donor will give a large gift.

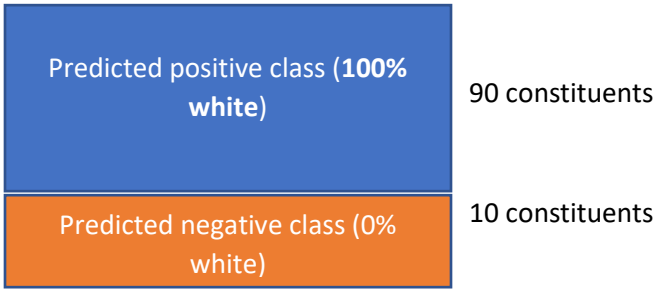
Scenario 1

Let us imagine that our only predictor in this case is race. It is easy to see how bias will be increased. Imagine that our training data consists of 100 constituents, 50 of whom have given large gifts. It is easy to imagine that, of those who have given large gifts, a larger proportion will be white than in the general donor base. Let's assume 10% of our donor base is non-white overall. If the positive class is 95% white, the negative class must be 85% white.

Training Data:

Now, a model trying to predict class based solely on race will predict that white constituents will give a large gift, and non-white constituents will not. Therefore, in the predicted data, 100% of predicted large gift givers will be white. This makes sense. This is actually exactly what we want machine learning models to do: to make heuristic decisions based on limited information.

Predicted Scores:



This is relatively straightforward. And this is exactly why many do not use race variables in their models. However, we will see that even *only using variables correlated with race*, without using race variables specifically, the model will still increase bias through predictions.

Scenario 2

Imagine that now, the only variable we train our model with is wealth. To keep it simple, we will assume this is a binary variable, taking on values of "High" and "Low," and it splits our data evenly. In our positive class, we expect there to be a higher concentration of High wealth individuals than in our data overall. As long as there is a relatively higher concentration of this variable in the positive class, and this class has disproportionately more white individuals, our model will increase bias. Consider the extreme case where overall in our data, half the population is white, and just 51% of the positive class is high wealth. If 3/5 of the High wealth individuals are white, 3/5 of our predictions will be white! While only half of our data was white.

For this example we will assume our data has a total of 200 constituents, to keep the calculations simple. Let's say we have 60 High wealth, white individuals, 40 High wealth non-white individuals, 60 Low wealth, non-white individuals, and 40 Low wealth white individuals.

Now assume *our positive class is actually majority non-white*. This is possible. They can be all 40 non-white High wealth individuals, 11 white High wealth individuals, and 49 non-white Low wealth individuals.

The negative class is then 11 non-white Low wealth individuals, 49 white High wealth individuals, and 40 white Low wealth individuals.

Training Data:

Positive class (51% High wealth, **89% non-white**)	100 constituents total 51 high wealth, 49 low wealth
Negative class (49% High wealth, 11% non-white)	100 constituents total 49 high wealth, 51 low wealth

Predicted Scores:

Predicted positive class (100% High wealth, **3/5 white**)	100 constituents total 100 high wealth, 0 low wealth; 40 non-white, 60 white
Predicted negative class (0% High wealth, 2/5 white)	100 constituents total 0 high wealth, 100 low wealth; 60 non-white, 40 white

The overall population is 50% white, the positive training samples are 11% white, and yet the positive predictions are **60% white**. We can see that *statistical parity* has not been preserved, and there is a *disparate impact* on the non-white group.