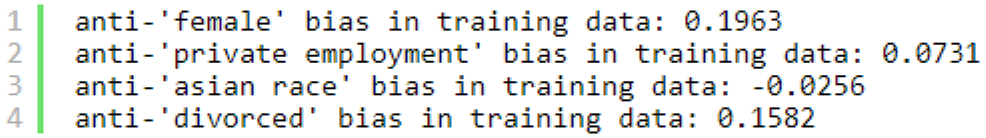
General Notes:

- Decided to go with a statistical parity test because it is easy to understand and implement.
- One downfall with this test is it is only looking at group fairness, not individual. If we want to measure this later we can find another test, but we think our primary focus at this point is bias against certain groups.
- In simple terms, statistical parity measures group fairness by comparing the proportion of a protected group vs the proportion of the population that are receiving some positive or negative classification. If statistical parity is achieved, these proportions should be equal.
- Statistical parity is calculated as the difference between the probability that a person from our protected group is put into a certain group (high risk for example) and the probability that a person from the unprotected group is put into the same group.
- The output of a statistical parity test will take on a value (-1,1). A negative value would indicate bias in favor of one group, while a positive would be biased in favor of the other.
- AI fairness suggests that if the absolute value of the result is <0.1, the test is fair. We are mostly just looking to show one method is less biased than another, but this is still a good thing to note.
- Our overall goal in using the statistical parity test is to run it for our improved model and get an output that is closer to zero. This will suggest that we have improved group fairness.

From article:

- Outlined and described in [One definition of algorithmic fairness: statistical parity]
- This article walks through a statistical parity test looking at bank loan approval, and provides some snippets of code which the development team could use.
    - Given function labelBias() computes the bias of a certain set of labels on a dataset
        - Returns the difference in probability between the protected portion of the dataset and everything else
    - Given function signedBias() computes the bias for a specific hypothesis on a dataset (may not be needed)
    - Running the given code gives an output of something like:



```
anti-'female' bias in training data: 0.1963
anti-'private employment' bias in training data: 0.0731
anti-'asian race' bias in training data: -0.0256
anti-'divorced' bias in training data: 0.1582
```

        - The positive values indicate bias against quoted groups.