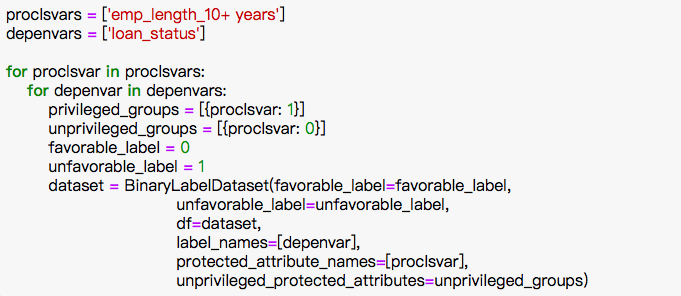
Fairness test

Now we would like to conduct fairness test through AIF 360. We looked for bias to predict the status of loan based on various features in the initial training data. The protected attribute will be “Employment length longer than 10 years”, with “1” and “0”. The values of protected attribute are also the values for the privileged and unprivileged groups respectively.

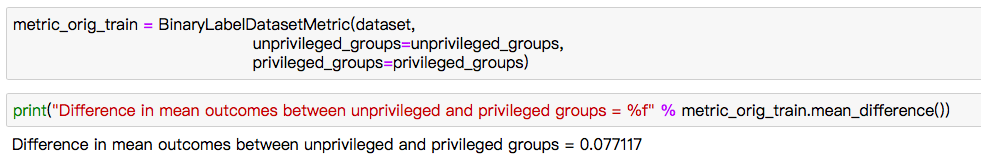
* Load dataset, specifying protected attribute

Firstly, we need to convert the data into a BinaryLabelDataset. Then we need to set two variables for the privileged and unprivileged values for the “Employment length longer than 10 years” attribute, which are key inputs for detecting and mitigating bias.



* Conduct fairness metric on the original training dataset

After identifying the protected attribute ‘Employment length longer than 10 years’ and defining privileged and unprivileged values, we would use AIF 360 to detect bias. More specifically, we would like to compare the percentage of favorable results for the privileged and unprivileged groups respectively. Then we would subtract the former percentage from the latter. A positive value indicates more favorable outcomes for the privileged groups. This is done through the calculation of mean difference on the BinaryLableDatasetMetric class. The code below shows that the difference is 0.077117.



* Reduce bias by transforming the training dataset

From the previous results, we can tell that the privileged group gets 7.77% more positive outcomes based on the training dataset. Then we would conduct pre-processing mitigation to mitigate this bias in the training dataset. More specifically, we used Reweighing algorithm in the aif360 package, which transformed dataset to have more weights in positive outcomes on the protected attribute.



* Check fairness metric on transformed dataset

After the above steps, we can compute whether we remove bias successfully through the BinaryDatasetMetric class again. From the results below, we can notice that the difference in mean outcomes is 0. So we can conclude that the mitigation step is effective.

