# Step 1-2

In this assignment, we will look at the impact of computing and applying fairness metrics to “fix” data that could be used to train algorithms associated with learning from credit-based data sets.

* Which dataset did you select?
  + **Taiwan Credit Data Set**
* How many observations are in the dataset?
  + **30000 observations**
* How many variables in the dataset?
  + **24 (1 response variable plus 23 explanatory variables)**
* Which variables did you select as your dependent variables?
  + **default payment next month (response variable)**
* How many and which variables in the dataset are associated with a legally recognized protected class? Of those variables associated with a protected class, what is the associated legal precedence/law it falls under as discussed in the lectures?

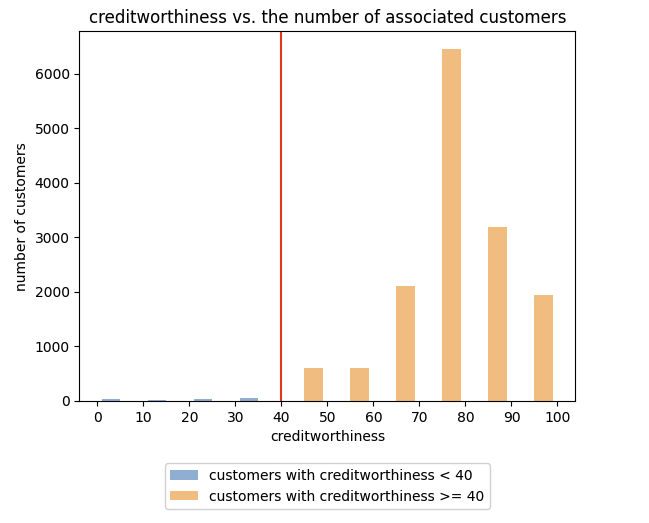
|  |  |  |
| --- | --- | --- |
| **variables** | **protected class** | **legal precedence** |
| X2: Gender  (1 = male; 2 = female). | Sex | Equal Pay Act of 1963;  Civil Rights Act of 1964, 1991;  Equal Credit Opportunity Act |
| X5: Age  (year). | Age | Age Discrimination in Employment Act of 1967;  Equal Credit Opportunity Act |

# Step 3

* Selected outcome variable:
  + X6 - X11: History of past payment and Y: default payment next month
* Formula to evaluate each customer =
* *Excellent Credit Risk (i.e. highly likely to pay back a loan)* VS. *Bad Credit Risk (i.e. highly likely to default on loan)*:
  + 100 – 0
  + sklearn MinMaxScaler to rescale the results to [0,100]
* Selected protected class attribute: Sex
  + unprivileged group: Female
  + privileged group: Male

|  |  |  |
| --- | --- | --- |
|  | Number of Members in Training Set | Number of Members in Testing Set |
| Female | 9061 | 9051 |
| Male | 5939 | 5949 |
| Total | 15000 | 15000 |

# Step 4



To compute profits, assume, in this case:

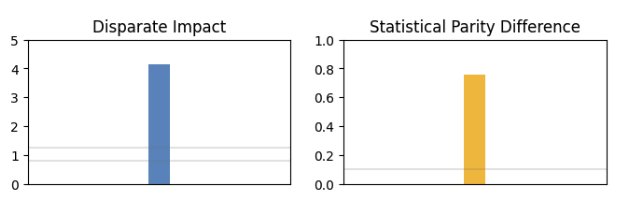
* Approved Loan/Good Credit Risk = +10 Profit
* Approved Loan/Bad Credit Risk = -5 Profit
* Declined Loan/Good Credit Risk = -3 Profit
* Decline Loan/Bad Credit Risk = 0 Profit

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Approved Loan | Declined Loan | Good Credit Risk | Bad Credit Risk | Customers | Profit | Total Profits |
| √ |  | √ |  | 3253 | +10 | 32530 |
| √ |  |  | √ | 77 | -5 | -385 |
|  | √ | √ |  | 11627 | -3 | -34881 |
|  | √ |  | √ | 43 | 0 | 0 |

|  |  |  |
| --- | --- | --- |
|  | MALE (privileged) | FEMALE (unprivileged) |
| Favorable | 1426 | 1904 |
| Unfavorable | 4513 | 7157 |

# Step 5

* Disparate Impact: the ratio of the rate of the favorable outcome of the unprivileged group to the rate of favorable outcome for the privileged group. Ideally the value should be 1. Fairness for this metric is between 0.8 and 1.25.
* Statistical Parity Difference: the difference between rate of favorable outcomes for unprivileged group (female) and the rate of favorable outcome for the privileged group (male). Fairness for this metric is between -0.1 and 0.1.



Step 6  
1) define two different creditworthiness formulas for the unprivileged versus privileged groups; OR,

2) define different threshold values for approving a loan even if they are considered a bad credit risk.

* For female:

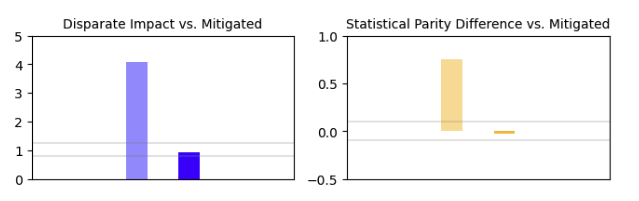
To compute profits, assume, in this case:

* Approved Loan/Good Credit Risk = +10 Profit
* Approved Loan/Bad Credit Risk = -5 Profit
* Declined Loan/Good Credit Risk = -3 Profit
* Decline Loan/Bad Credit Risk = 0 Profit

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Approved Loan | Declined Loan | Good Credit Risk | Bad Credit Risk | Customers | Profit | Total Profits |
| √ |  | √ |  | 2495 | +10 | 24950 |
| √ |  |  | √ | 862 | -5 | -155 |
|  | √ | √ |  | 11632 | -3 | -33078 |
|  | √ |  | √ | 617 | 0 | 0 |

Step 7

Compared with the metrics (Disparate Impact and Statistical Parity Difference) calculated in Step 5, the difference between the unprivileged group and privileged group is decreased using the new threshold and formula, as shown in Figure 4.



* Statistical Parity Difference: 99.3% - 99.0% = 0.3%
* Disparate Impact: 99.3%/99.0%=1

Adding an extra coefficient of 0.25 to the original formula was effective to mitigate the original bias towards the male group.

The female group's creditworthiness value becomes one quarter of the original value due to the coefficient of 0.25, while the male group's creditworthiness value remains the same. Therefore, in relative terms, the male group receives a positive advantage, and the female group was disadvantaged by the mitigation step.

Even though this method is helpful to mitigate the existing bias, it leads to some issues as well. By applying the extra factor to the formula calculating the female group’s creditworthiness, the values become less than they actually are. As a result, some female group members who should have received the loan will not actually be able to get the loan, which is unfair to them.

* **historically disadvantaged group of interest - (unprivileged group: Female; privileged group: Male)**