CS 6603: AI, Ethics, and Society
Fairness Bias Assignment

Cleo Zhang

yzhang3761@gatech.edu

***Abstract*— this assignment will focus on the impact of computing and applying the fairness metrics to "fix" the data that could be used to train algorithms associated with learning from the given credit-based datasets.**

1 SELECTED DATASET OVERVIEW

This report will answer the questions required from the Assignment 5 instructions using the Taiwan Credit Data Set. This dataset has 30,000 observations and 24 variables (1 response variable plus 23 explanatory variables), and I selected the "default payment next month" as my dependent variable.

Table 1 shows the variables in the dataset, related protected class and the associated legal precedence.

Table 1 — Protected class and the associated legal precedence.

Variables	Protected class	Legal precedence
X2: Gender (1 = male; 2 = female).	Sex	Equal Pay Act of 1963; Civil Rights Act of 1964, 1991
X5: Age (year)	Age	Age Discrimination in Employment Act of 1967

2 MANIPULATION ON THE ORIGINAL DATASET

This report will use the following information to split the original dataset into training and testing sets:

- Selected outcome variables: X6 - X11: History of past payment
- Formula to evaluate a customer's creditworthiness:

$$\_\frac{(X6 + X7 + X8 + X9 + X10 + X11)}{6}$$

- Selected protected class attributes: Sex (unprivileged group: Female; privileged group: Male)

After calculating each customer's creditworthiness using the formula above, I will use the maximum and minimum values of calculated creditworthiness to scale the original values to a range of 0-100, where 100 is the maximum value for *Excellent Credit Risk* (i.e., highly likely to pay back a loan) and 0 is the minimum value for *Bad Credit Risk* (i.e., highly possible to default on the loan).

Table 2 shows the overview of the training and testing sets randomly split from the original dataset.

Table 2 – Training and Testing sets.

	Number of Members in Training Set	**Number of Members in Testing Set**
Female	9115	8997
Male	5885	6003
Total	15000	15000

3 BIAS EVALUATION

I can get the following figure by calculating the creditworthiness and populating the histogram.

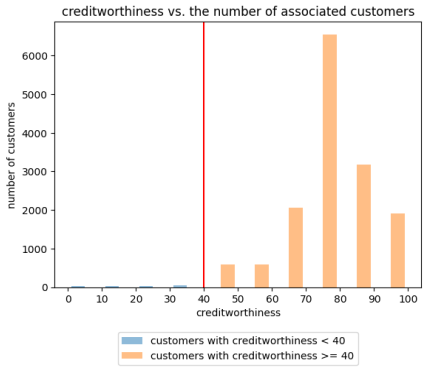

Figure 1 — Creditworthiness vs. the number of associated customers.

In this case, the threshold is 40, attempting to maximize the profit (see details in Table 3). Even if the final total profit is negative, it is the max value I can find. Setting the threshold below 40 or above 40 makes the total profit less.

Table 3 – Profits Calculation Overview.

Approved Loan	Declined Loan	Good Credit Risk	Bad Credit Risk	Customers	Factors	Profits
√		√		3208	+10	32080
√			√	91	-5	-455
	√	√		11669	-3	-35007
	√		√	32	0	0
					Total:	**-3382**

Table 4 documents how many in each group (privileged and unprivileged) received Favorable (i.e. Approved) versus Unfavorable (i.e. Declined) outcomes based on the selected threshold value.

Table 4 – Privileged and Unprivileged Groups received Favorable versus Unfavorable.

	MALE (privileged)	FEMALE (unprivileged)
Favourable	1435	1864
Unfavourable	4450	7251

4 FAIRNESS METRICS ON THE *TRAINING SET*

This section will compute the below metrics for bias evaluation, and the results (difference between male and female groups) are shown in *Figure 2*:

- Disparate Impact: an ideal outcome for this metric is between 0.8 and 1.25.

$$\frac{female\_train\_good\_count}{female\_train\_fgood\_count + female\_train\_bad\_count} \div \frac{male\_train\_good\_count}{male\_train\_good\_count + male\_train\_bad\_count}$$

- Statistical Parity Difference: an ideal result for this metric is between -0.1 and 0.1.

$$\frac{female\_train\_good\_count}{female\_train\_good\_count + male\_train\_bad\_count} - \frac{female\_train\_good\_count}{female\_train\_good\_count + male\_train\_bad\_count}$$

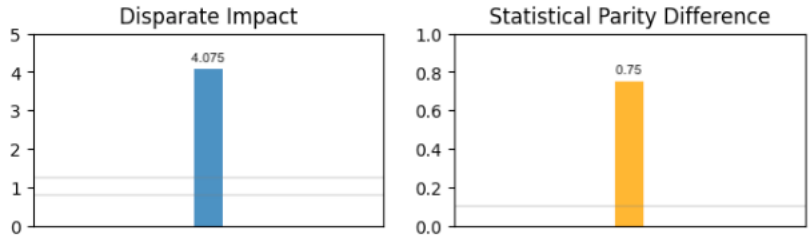

Figure 2 – Disparate Impact and Statistical Parity Difference computed from the training set.

The Disparate Impact is 4.075, which implies a much higher benefit for the unprivileged group (female) instead for the historically advantaged group of interest (male). A similar conclusion can be drawn from the Statistical Parity Difference metric, whose value is 0.75.

5 MITIGATE THE BIAS IN THE TRAINING DATASET

To mitigate the bias indicated in the previous section, I will experiment with the below method:

- Update the formula to calculate the creditworthiness for the unprivileged group (female) but keep the formula the same for male customers.

$$-\frac{(X6 + X7 + X8 + X9 + X10 + X11)}{6} \times 0.25$$

- Set different thresholds for different groups – 70 for females and 40 for males – to approve the loan or not. *Figure 3* is the histogram associated with Good Credit Risk versus Bad Credit Risk as a function of Sex subgroups with threshold highlights.

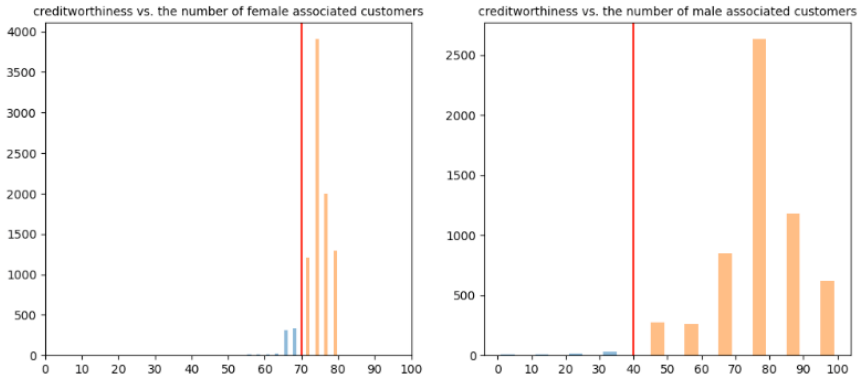

Figure 3 – Mitigated creditworthiness by groups.

Table 5 shows the profit calculation after the mitigation. I have tried a few threshold combinations, and -17164 is the max value among all attempts, significantly lower than the original total profit computed in Section 3.

Table 5 – Post-mitigation Profits Calculation Overview.

Approved Loan	Declined Loan	Good Credit Risk	Bad Credit Risk	Customers	Factors	Profits
√		√		2104	+10	21040
√			√	1195	-5	-5975
	√	√		10743	-3	-32229
	√		√	958	0	0
					Total:	**-17164**

6 FAIRNESS METRIC DIFFERENCES BETWEEN THE ORIGINAL AND POST-MITIGATION DATASETS

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

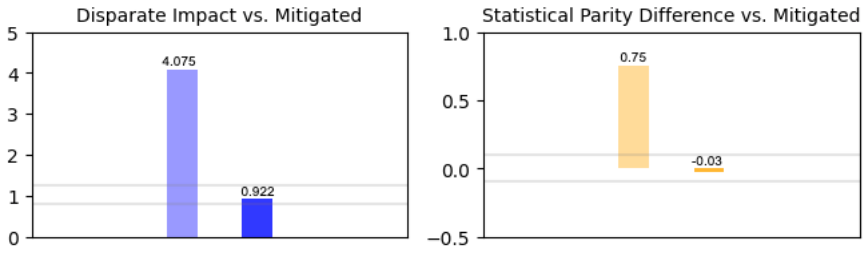

Figure 4 – Compare Disparate Impact/Statistical Parity Differences between the original and mitigated dataset.

Adding a coefficient of 0.25 to the original formula effectively mitigated the initial bias towards the male group. The female group's creditworthiness value becomes one-quarter of the initial 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 received a positive advantage, and the mitigation step disadvantaged the female group.

Even though this method helps mitigate the existing bias, it also leads to some issues. By applying the extra factor to the formula calculating the female group's creditworthiness, the values become less than they are. As a result, some female group members who should have received the loan will be unable to get it, which is unfair to them.