CS 6603 AI, Ethics, and Society
Fairness and Bias

Kamran Munawar
kmunawar3@gatech.edu

1 EXPLORE THE DATASET

Which dataset did you select?

Taiwan Credit Data Set

How many observations are in the dataset?

30000

How many variables are in the dataset?

24

How many variables are associated with a legally recognized protected class under U.S. law?

SEX: Civil Rights Act of 1964,Equal Pay Act of 1963

AGE: Age Discrimination in Employment Act of 1967

2 STEP 3 – DEFINING CREDITWORTHINESS AND PREPARING THE DATASET

2.1 Formula for creditworthiness

$$\text{Creditworthiness Score} = 0.3 \cdot \left(1 - \frac{\text{Average Repayment Status}}{9}\right) + 0.3 \cdot \left(1 - \frac{\text{Total Outstanding Bill}}{\max(\text{Total Outstanding Bill})}\right) + 0.3 \cdot \left(\frac{\text{Total Payments Made}}{\max(\text{Total Payments Made})}\right)$$

Putting it all together in the final formula:

$$\text{Creditworthiness Score} = 0.3 \cdot \left(1 - \frac{\frac{X_6+X_7+X_8+X_9+X_{10}+X_{11}}{6}}{9}\right) + 0.3 \cdot \left(1 - \frac{X_{12}+X_{13}+X_{14}+X_{15}+X_{16}+X_{17}}{\max(\text{Total Outstanding Bill})}\right) +$$

$$0.3 \cdot \left(\frac{X_{18} + X_{19} + X_{20} + X_{21} + X_{22} + X_{23}}{\max(\text{Total Payments Made})} \right)$$

Outcome variable in my dataset :Y

Selected protected class attribute: Sex

Privileged Group : Male

Unprivileged group: Female

2.2 Members of Protected Class

	Training Dataset	Test Dataset
Male	5882	6006
Female	9118	8994
Total	15000	15000

Table 1—Members of protected class

3 STEP 4 – MAXIMIZING PROFIT

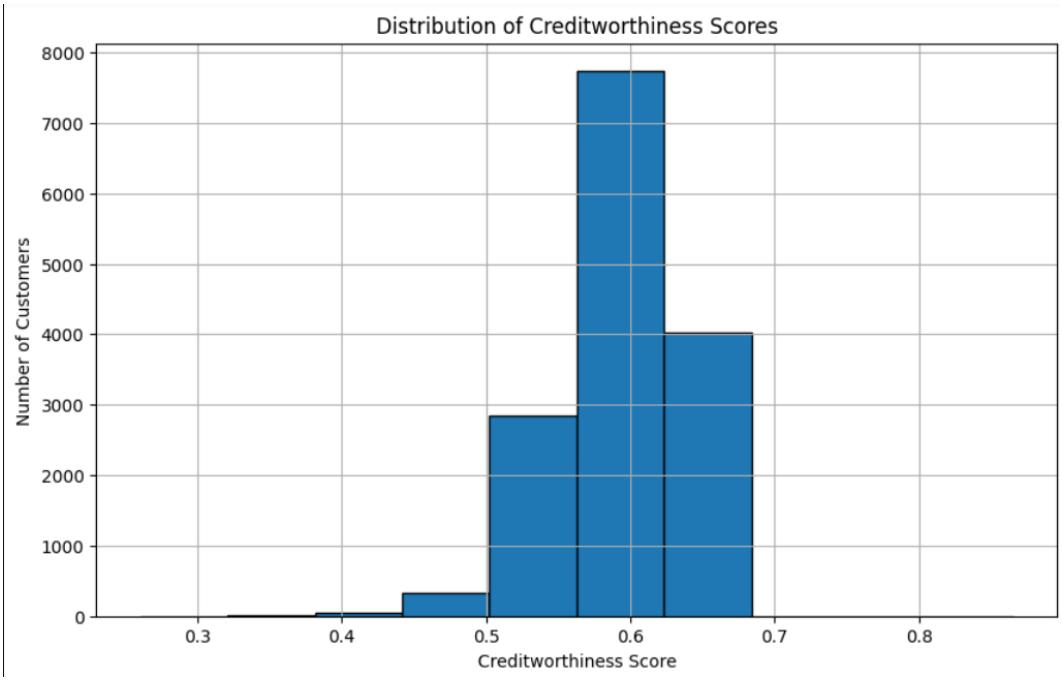

Figure 1—Creditworthiness Score

Threshold 0.40 that maximizes profit of -2613

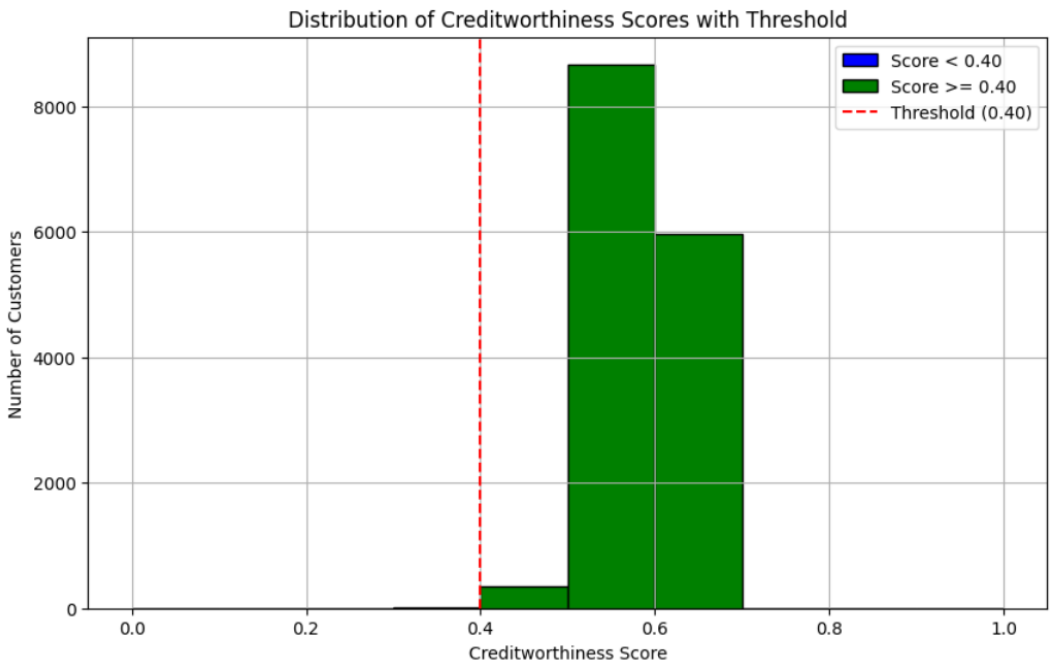

Figure 2—Creditworthiness Scores with Threshold 0.40

Table 2 showing favorable vs Unfavorable outcomes threshold value of 0.40 at maximum profit of -2613

	Total	**Approved**	**Declined**
Male(Privilege)	5882	5877	5
Female(unprivileged)	9118	9110	8
Total	15000	14987	13

Table 2—Favorable vs Unfavorable

4 STEP 5: FAIRNESS METRICS

Fairness Metric	**Computed Diff**	**Range**	**Bias**
Demographic Parity Difference	-0.001344	-0.1 to 0.1	No
Equal Opportunity Difference	-0.001223	-0.1 to 0.1	No

Table 3—Fairness Metrics

Demographic Parity Difference measures the difference in the probability of receiving a favorable outcome (loan approval) between the privileged group (Male) and the unprivileged group (Female).A value of -0.001344 means that males

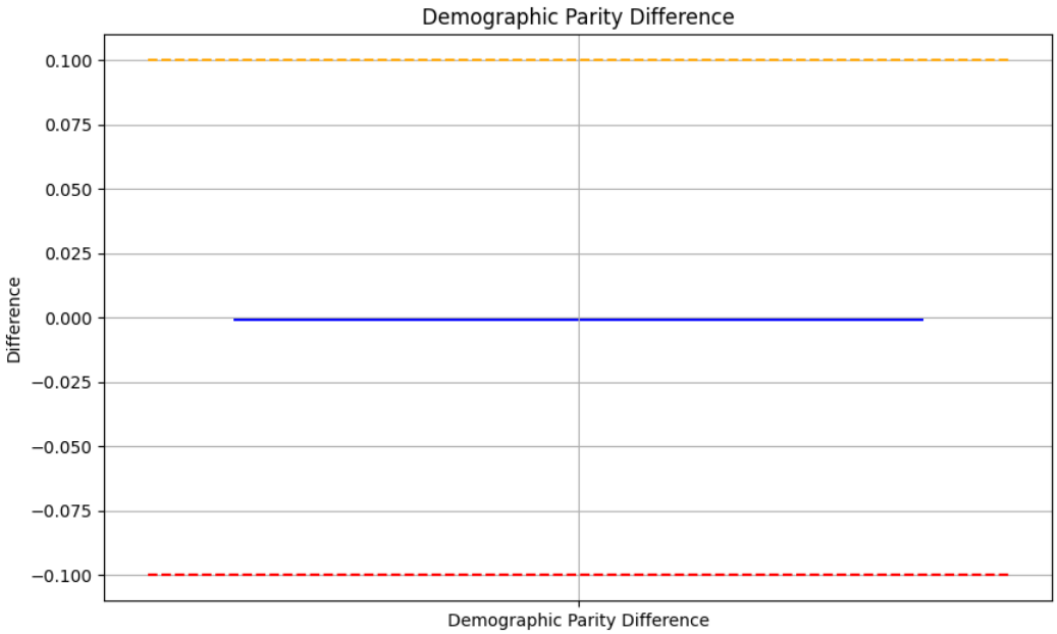

Figure 3—Demographic parity difference

Figure 4—Equal opportunity difference

(privileged group) are 0.1344% less likely to receive loan approval compared to females (unprivileged group).Since the computed difference is within the acceptable range (-0.1 to 0.1), it indicates no bias.This suggests there is no significant bias for or against the privileged group (Male) or the unprivileged group (Female) in terms of the likelihood of loan approval. Equal Opportunity Difference measures the difference in the true positive rates (TPR) for receiving a favorable outcome between the privileged group (Male) and the unprivileged group (Female).A value of -0.001223 means that males (privileged group) have a 0.1223% lower true positive rate for receiving loan approval compared to females (unprivileged group).Since the computed difference is within the acceptable range (-0.1 to 0.1), it indicates no bias.This suggests there is no significant bias for or against the privileged group (Male) or the unprivileged group (Female) in terms of the true positive rate for loan approval. Both fairness metrics, Demographic Parity Difference and Equal Opportunity Difference, indicate no significant bias for or against the privileged group (Male) or the unprivileged group (Female). The computed differences are well within the acceptable ranges, suggesting fair outcomes in terms of both the likelihood of loan approval and the true positive rate for loan approvals. This indicates that the loan approval process does not disproportionately favor or disadvantage either group.

5 STEP 6 – MITIGATE BIAS IN THE TRAINING SET

Sex	Computed Difference	Threshold	Indicates Bias
Male(Privilege)	-0.000230	0.37	No
Female(Non Privilege)	-0.000230	0.27	No

Table 4—Demographic Parity Difference

Total Profit: -2517

The threshold values for loan approval are set differently for the privileged and unprivileged groups. For the privileged group, males, the threshold value is 0.37, whereas for the unprivileged group, females, the threshold value is 0.27. These thresholds indicate the creditworthiness scores above which loans are approved for each group.The total profit, based on these threshold values, is calculated to be -2517. This suggests that, overall, the losses from denying loans that were approved in the original dataset or approving loans that were denied in the original dataset outweigh the gains from correctly approving loans. These results high-

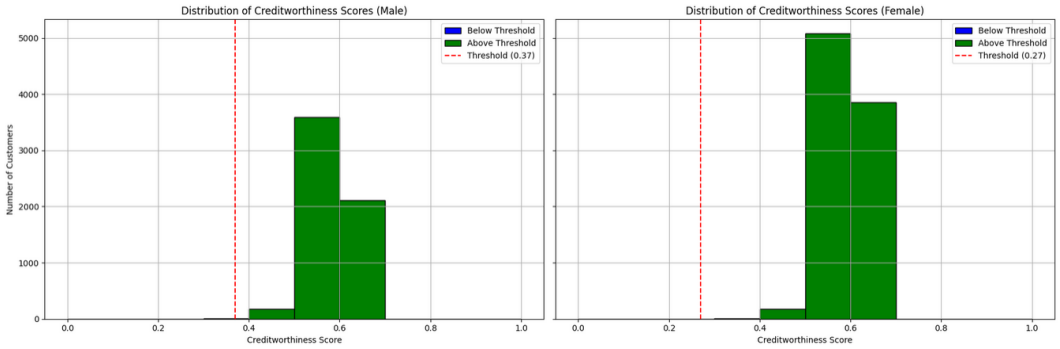

Figure 5—Distribution of Creditworthiness Scores

light the financial impact of the chosen thresholds on the loan approval system's profitability. Adjusting these thresholds could potentially improve profitability, but such adjustments must carefully consider fairness and the avoidance of bias.

	Total	**Approved**	**Declined**
Male(privileged)	5882	5880	2
Female(unprivileged)	9118	9117	1
Total	15000	14997	3

Table 5—Approved vs Declined using different threshold value

6 STEP 7 – POST BIAS MITIGATION ANALYSIS

6.1 For each of the fairness metrics selected in Step 5, discuss if there were any differences in the outcomes for the privileged versus unprivileged group?

Both metrics, Demographic Parity Difference and Equal Opportunity Difference, show that there is no significant bias for or against either males or females. The computed differences are comfortably within the acceptable limits, indicating that the loan approval process is fair in terms of both the likelihood of approval and the true positive rate. Thus, the process does not disproportionately favor or disadvantage either group.

6.2 Was the mitigation step in Step 6 effective and for whom? Did any group receive an advantage? Was any group disadvantaged by the mitigation step?

To ensure fairness in loan approvals, the mitigation step involved using different thresholds for the privileged group (males) and the unprivileged group (females).

Initially, both groups had nearly identical approval rates, suggesting that the system was fair. With equal thresholds, the number of declined applications differed slightly (5 for males and 8 for females), indicating minimal discrepancy.

By lowering the threshold for females to 0.27 compared to 0.37 for males, the computed differences for both Demographic Parity and Equal Opportunity stayed within the acceptable range of -0.1 to 0.1, signifying no bias. This approach ensured that both groups had similar approval rates, thereby maintaining fairness.

The unprivileged group (females) received an advantage through a lower threshold, ensuring their approval rate matched that of males. Despite the higher threshold for males, their approval rate remained very high, indicating that they were not significantly disadvantaged. The computed differences confirm that males were not unfairly treated by this adjustment.

6.3 Identify any issues that would arise if this method was used to mitigate bias. Justify your reasoning.

Using different thresholds can be an effective short-term strategy to mitigate bias, but it introduces several significant issues that must be carefully considered and addressed. For instance, it might inadvertently create a new form of bias against privileged groups, leading to resentment and claims of unfair treatment. Implementing different standards for different groups can erode the perception of a fair and transparent process, as consistent criteria are essential for maintaining trust in decision-making systems.

Moreover, different thresholds add layers of complexity, making the system harder to manage and more prone to errors. This complexity can undermine the reliability and fairness of the system. Additionally, different thresholds might conflict with legal standards of equal treatment and face ethical challenges.