Counterfactual Fairness Analysis for Credit Default

Group Members
a.ntantouri@students.uu.nl
j.d.liutaymar@uu.nl
y.an2@students.uu.nl
y.qu1@students.uu.nl
Utrecht University
Utrecht, the Netherlands

INFOMHCML Teaching Team
d.p.nguyen@uu.nl
h.kaya@uu.nl
h.m.wong@uu.nl
s.p.kaszefskiyaschuk@uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

This project investigates fairness in credit default prediction using an interpretable logistic regression model and counterfactual explanations. We evaluate a credit default classifier trained on the *Default of Credit Card Clients* dataset, focusing on fairness across demographic groups such as gender, age, and marital status. Standard group fairness metrics are first computed to assess disparities in model outcomes. To complement this, we implement an effort-based counterfactual analysis that quantifies how much individuals must change their actionable features to reverse an unfavorable outcome while keeping protected attributes constant. While standard fairness metrics showed that the model favours females, our counterfactual analysis indicated that the model helps males. Specifically, while it is more likely for the model to predict 'default' (unfavourable outcome) for males, our counterfactual analysis showed that when females are predicted to default, it is more difficult for them, although not significantly, to change their outcome to a favourable one. This confirms that fairness is multidimensional. Even when prediction rates appear in favor of one group (in this case, females are in favor based on standard fairness metrics), the effort for changing to a favourable outcome might also be higher for this group (in this case, females require more effort in changing their outcome to 'no default'). Therefore, evaluating fairness using only standard fairness metrics (such as statistical parity or TPR) may overlook disparities embedded in the model's behavior.

KEYWORDS

machine learning, counterfactual fairness, interpretability, logistic regression

ACM Reference Format:
Group Members and INFOMHCML Teaching Team. 2025. Counterfactual Fairness Analysis for Credit Default. In *Proceedings of Utrecht University (INFOMHCML'2025).* Utrecht University, 8 pages.

1 INTRODUCTION

Machine learning (ML) models are increasingly used to automate high-stakes decisions such as credit default, hiring, and insurance pricing. However, these models often raise concerns about fairness, accountability, and transparency, especially when deployed in domains that significantly affect individuals' lives [1]. What seems particularly noteworthy in this analytical context is that in the realm of credit scoring, biased decisions can seemingly reinforce existing socioeconomic disparities, which tends to underscore the importance of interpretable and fair machine learning methods.

Recent years have witnessed a growing emphasis on human-centered machine learning, which not only evaluates model performance but also considers how individuals interact with and are ostensibly affected by predictive systems [9]. Within this broader analytical framework, while traditional fairness metrics like statistical parity and true positive rate tend to provide what might be characterized as a group-level view of equity, they apparently fail to capture the individual burden required to receive favorable outcomes in the majority of cases. What this appears to suggest, therefore, is an increased focus on counterfactual fairness, which seemingly asks: What minimal changes must an individual make to change an outcome, holding protected attributes constant?

What this investigation aims to explore is both standard group fairness and counterfactual fairness in the context of credit default prediction. Given the complexity of these theoretical relationships, we employ an interpretable model—logistic regression—to analyze disparities across gender, age, and marital status. What the analysis tends to support is the implementation of effort-based counterfactual analysis to quantify the difficulty individuals face in changing their outcomes. What these findings seem to point toward is that traditional fairness metrics alone may be misleading, and a more nuanced approach seems necessary to evaluate fairness in real-world ML applications.

2 RELATED WORK

Fairness in machine learning has been extensively studied, especially in sensitive domains such as credit scoring [3, 5]. What the early approaches tend to suggest is a focus on group fairness metrics such as demographic parity and equalized odds [4]. While these metrics seem to be useful for quantifying group-level disparities, they may not adequately reflect the challenges individual users face in altering model predictions within this broader analytical framework.

To address this, what seems to emerge from these findings is that counterfactual fairness has emerged as a powerful alternative that captures individual-level fairness by estimating how much a user must typically change their input features to receive a favorable outcome while keeping protected attributes fixed [8]. What appears particularly significant about these findings is that Wachter et al.

INFOMHCML'2025, 2025, Utrecht, the Netherlands

ACM ISBN xxxxxxx.
https://doi.org/xxxxxxx

[12] introduced a formal method for generating counterfactual explanations in black-box models, while more recent libraries like DiCE have operationalized these ideas into actionable toolkits [10].

Moreover, what the evidence appears to reveal is that interpretability plays a key role in fairness-aware ML. Interpretable models like decision trees or logistic regression tend to provide more insight into decision boundaries and are seemingly preferred in high-stakes contexts [11]. What appears to follow from this analysis is that combining interpretability with counterfactual analysis tends to allow researchers to identify not only if a model is ostensibly biased, but how that bias appears to manifest in practice given the complexity of these theoretical relationships [6].

What also appears significant in this context is that recent contributions such as the AI Fairness 360 toolkit [2] provide open-source frameworks for assessing and mitigating algorithmic bias [1]. Similarly, what the 2024 XAI conference proceedings appears to indicate is an emphasis on interpretable neural architectures and fairness metrics in domains like credit assessment, considering the nuanced nature of these findings [9].

In line with our approach, Kuratomi et al. [7] also explore the use of counterfactual explanations to assess fairness by quantifying the 'burden' required for individuals to achieve favorable outcomes. However, our work extends beyond theirs in several meaningful ways. First, rather than relying solely on abstract feature distance as a proxy for effort, our formula explicitly incorporates feature-specific cost weights, which allows us to acknowledge and quantify that changes in features like education or credit limit can carry unequal real-world costs. Second, our project emphasizes inter-group burden comparisons, examining how effort requirements differ by gender, age, and marital status. Third, we incorporate visual explanations of counterfactuals to aid transparency and stakeholder engagement. These contributions aim to bridge the gap between algorithmic fairness assessment and practical, stakeholder-aligned interpretability.

In summary, our project aims to build on prior work by integrating interpretable modeling, group fairness evaluation, and counterfactual analysis to provide a more holistic view of fairness in ML systems used for credit decision-making within these evolving conceptual parameters.

3 METHODS

This section outlines the full pipeline used to evaluate the fairness in credit default prediction. The process consists of three stages: (1) training an interpretable classification model, (2) performing standard fairness analysis based on observed predictions, and (3) conducting counterfactual fairness analysis using effort-based counterfactual explanations.

3.1 Dataset

We used the *Default of Credit Card Clients Dataset* from the UCI repository[1]. It contains data on 30,000 credit card users in Taiwan, with 23 features and a binary target variable indicating default status.

Key Features include:

- **Credit and Demographics:**
  - `LIMIT_BAL`: Client's credit limit (continuous).
  - `SEX`: Gender (categorical: 1=Male, 2=Female).
  - `EDUCATION`: Educational background (ordinal categorical: 1=Graduate school, 2=University, 3=High school, 4=Others, 5-6=Unknown).
  - `MARRIAGE`: Marital status (nominal categorical: 1=Married, 2=Single, 3=Others).
  - `AGE`: Client's age (continuous).
- **Repayment Status (`PAY_0`, `PAY_2` to `PAY_6`):** Ordinal categorical features representing the repayment status over the past six months (from September to April 2005). Values range from -1 (pay duly) to 9 (payment delay for nine months and above), with 0 indicating 'pay duly'.
- **Billing Amounts (`BILL_AMT1` to `BILL_AMT6`):** Continuous features indicating the amount of bill statements for the past six months (September to April 2005).
- **Previous Payment Amounts (`PAY_AMT1` to `PAY_AMT6`):** Continuous features detailing the amounts of previous payments made over the past six months (September to April 2005).

The **target variable** (`default`) is a binary categorical variable indicating whether the client will default on their payment in the next month (1 = Yes, 0 = No). We performed an 80/20 train-test split on the dataset with random seed 42.

3.2 Model Training and Selection

We trained a logistic regression model due to its transparency and interpretability, which are important characteristics when assessing fairness in high-stakes decision-making contexts like credit default.

(1) **Preprocessing:** Features were standardized using z-score normalization. Categorical features were passed through without any transformation.
(2) **Model Selection:** We performed 5-fold cross-validation on the training set using the GridSearchCV library to select the optimal penalties (L1/L2) and regularization strengths (C).
(3) **Final Model:** The best model was retrained on the full training set and we extracted the most important feature coefficients for analysis.

3.3 Standard Fairness Metrics

To evaluate fairness at the group level, we computed common fairness metrics, specifically Demographic Parity, True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR).

We focused particularly on differences between 'gender' groups, but also analyzed fairness metrics per 'age' group, per 'marriage' group and per 'education' group. These metrics reveal whether the model systematically favors or disfavors any group in its predictions.

3.4 Measuring Counterfactual Effort

We define **counterfactual effort** as the normalized, feature-specific cost an individual must incur to change the model's prediction from an unfavorable outcome (`default = 1`) to a favorable one (`no default = 0`). This is done while keeping their protected attributes fixed, focusing on changes they might realistically influence.

[1] https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Specifically, we keep the **protected attributes** `SEX`, `AGE`, and `MARRIAGE` constant during counterfactual generation. Only the **actionable features** `LIMIT_BAL`, `PAY_0`, `PAY_2–6`, `BILL_AMT1–6`, `PAY_AMT1–6`, and `EDUCATION` are allowed to change in order to achieve the desired prediction. These reflect variables a user might realistically influence over time.

3.4.1 Methodology.

(1) **Select Rejected Applicants:** Extract instances from the test set predicted as `default`.
(2) **Generate Counterfactuals:** For each rejected instance, generate a counterfactual using the DiCE library such that the prediction changes to `no default`, with minimal changes to actionable features.
(3) **Compute Feature-Specific Effort:** Calculate the change effort using:
   - Feature-specific cost weights c_i
   - Standardized differences for continuous features using the mean μ_i and standard deviation σ_i from training data.
(4) **Aggregate Across Groups:** Compare average total effort across demographic subgroups (e.g., `Male | <30 | Not Married`).

3.4.2 Effort Formula. Let x_i and x_i' denote the values of feature i in the original instance and its counterfactual, respectively. The total counterfactual effort is computed as:

$$\text{Total Effort} = \sum_{i \in \mathcal{F}} \text{Effort}_i$$

where $\mathcal{F}$ is the set of actionable features, and:

$$\text{Effort}_i = \begin{cases} c_i \cdot \left| \frac{x_i - \mu_i}{\sigma_i} - \frac{x_i' - \mu_i}{\sigma_i} \right|, & \text{if } i \in \mathcal{F}_{\text{cont}} \\ c_i \cdot |x_i - x_i'|, & \text{if } i \in \mathcal{F}_{\text{cat}} \end{cases} \quad (1)$$

- c_i is the predefined cost of changing feature i
- $\mathcal{F}_{\text{cont}}$ is the set of continuous features
- $\mathcal{F}_{\text{cat}}$ is the set of categorical or ordinal features (encoded as integers)

3.4.3 Fairness Criterion. We define a model as **counterfactually fair** if the average counterfactual effort is approximately equal across sensitive groups:

$$\mathbb{E}[\text{Effort} \mid \text{Group A}] \approx \mathbb{E}[\text{Effort} \mid \text{Group B}]$$

This ensures that no protected group must undergo significantly greater effort to receive favorable outcomes.

3.4.4 Implementation Details. We implemented counterfactual generation using the DiCE library with constraints to ensure realistic and meaningful changes. The cost weight for each actionable feature is set to 1, to treat each feature equally in terms of its contribution to the total effort. Standardization parameters are derived from the training data. To account for variability in counterfactual generation and enhance the robustness of our analysis, the process of generating counterfactuals and calculating efforts was repeated 5 times, and the results were aggregated (thus, 5 counterfactuals were generated for each instance). The link to the code repository with the full implementation can be found in Appendix C.

4 RESULTS

This section explains the metrics of the trained logistic regression model and the outcomes of two fairness analyses. Concerning the fairness analyses, the first one is more general and concerns the fairness of the model, according to the gender groups. The second one quantifies and explains the effort needed to change the result from "default" (1) to "no default" (0). These outcomes are divided according to the number of runs performed, and therefore the number of counterfactuals generated and analyzed for each instance.

For each case, a heatmap and two horizontal histograms were generated. The map is a detailed overview, with 160 cells, that shows clearly how the effort values change for each group and each of the 20 mutable features. The first histogram collects all the values depending on the subgroups considered and averages them, ranking them based on the total average effort. The second histogram is a macro vision of the effort, based on the two genders, and provides a more robust overview of the effort disparity across genders.

4.1 Interpretable Logistic Regression Model

Before quantifying the performance in classifying the instances, the best parameters for this model were found using the GridSearchCV library. After that, a confusion matrix and a classification report have been created.

4.1.1 Hyperparameter Tuning. To ensure optimal performance, hyperparameter tuning was done using GridSearchCV, which employs a grid search cross-validation approach. The best parameters found for the model are:

(1) **classifier__C: 0.1**
It represents the regularization strength, and this value indicates that the model found a good balance between fitting the training data and maintaining simplicity to avoid memorizing noise.

(2) **classifier__penalty: l1**
l1 refers to the type of regularization, in this case, Lasso regularization. This kind of approach performs feature selection, which consists in nullifying some feature values while keeping the attribute of interest. In addition, it creates sparsity that is improves the model's ability in identifying the important attributes. Indeed, it allowed to recognize PAY_0 as one of the most impactful features among all.

(3) **classifier__solver: liblinear**
This last variable specifies that *liblinear* algorithm should be used in the optimization problems where the dataset is small and l1 regularization is implemented, like in this case.

4.1.2 Confusion Matrix. Figure 1 is the confusion matrix that shows how many correct and incorrect predictions were made by the model (1 is 'default' and 0 is 'no default'):

- **True Negatives:** The number of 'no default' instances classified correctly is 4552.
- **False Positives:** The amount of 'no default' instances classified incorrectly as 'default' is 135. The value is relatively

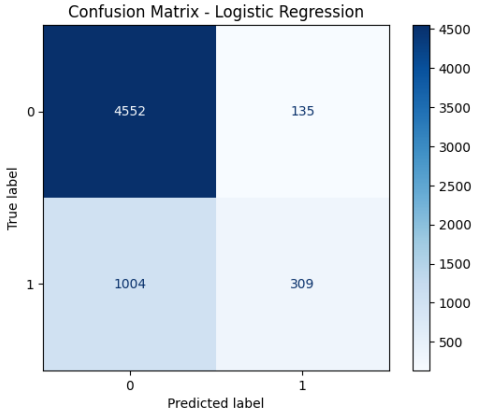

Figure 1: Confusion Matrix

small and therefore acceptable, since the cases of false 'default' predictions are few.

- **False Negatives:** 1004 "default" instances were predicted incorrectly as 'no default' by the model. This represents a major issue as the model struggles in identifying many true default cases while indicating that the model is conservative in predicting 'default', leading to a tendency to over-predict 'no default'.
- **True Positives:** At last, 309 instances were classified correctly as 'default'.

4.1.3 Classification Report. After finding and applying the best hyperparameters, Table 1 displayed the classification report for the model.

Class	Precision	Recall	F1-score	Support
0	0.82	0.97	0.89	4687
1	0.70	0.24	0.35	1313
Accuracy			0.81	6000
Macro Avg	0.76	0.60	0.62	6000
Weighted Avg	0.79	0.81	0.77	6000

Table 1: Classification Report

- **Precision:** Precision measures the proportion of positive identifications that were correct. By looking at the table, we can notice that for class 0 ('no default') the performance is 0.82, which is a high score. For class 1 ('default') the value is slightly lower (0.70), which is generally acceptable.

- **Recall:** Recall measures the proportion of true positives that were actually identified. For class 0, the score is very high (0.97), compared to class 1 where the score is really low (0.24). This is important because while the model is fairly precise when it predicts a 'default' outcome (70% precision), it misses a large proportion of instances that truly should default (only identifies 24% of them). Therefore, it classifies instances as 'no default' even when they should be 'default'.

- **F1-score:** Regarding this metric, for class 0, a strong balance between the model's precision (0.82) and high recall (0.97) is reflected by its high F1-score (0.89). Instead, for class 1, where we have a low recall (0.24) and a moderate precision (0.7), the low value for this metric (0.35) indicates that the model struggles to effectively identify instances that will truly be 'default', making its predictions for the 'default' class less reliable overall.

- **Accuracy:** Overall, the proportion of correct predictions made by the model is high across all classes and reaches 0.81, nevertheless, it happens in the context of class imbalance: class 0 constitutes approximately 78% of the dataset. The model's high accuracy is primarily driven by its excellent performance on this prevalent 'no default' class, potentially masking its poor performance on the minority 'default' class.

4.1.4 Most influential features. Before assessing fairness, we examine the logistic regression coefficients to understand the model's decision logic. Table 2 shows the 10 most influential features by absolute weight.

Feature	Coefficient
PAY_0	0.576
PAY_AMT2	-0.226
BILL_AMT1	-0.219
MARRIAGE	-0.162
PAY_AMT1	-0.139
SEX	-0.123
LIMIT_BAL	-0.108
EDUCATION	-0.102
PAY_2	0.089
AGE	0.072

Table 2: Top 10 most influential features in the logistic regression model.

`PAY_0` has the strongest positive impact, indicating recent payment delays greatly increase default risk. Payment and bill amounts (e.g., `PAY_AMT2`, `BILL_AMT1`) have protective effects (negative coefficients). Notably, protected attributes like `SEX` and `MARRIAGE` also carry substantial weight. This suggests that while financial behavior dominates, demographic factors still influence outcomes, raising fairness concerns explored in the next section.

4.2 Model Fairness Metrics by Gender

The purpose of this analysis is to provide an overview of the differentiated performance of the model, while considering just the protected attribute 'gender' (male and female).

As seen in Figure 2, the performance of the model is similar between the groups 'Male' and 'Female' for most metrics, suggesting a good trend towards overall parity. Nevertheless, some differences emerge, especially when the model is required to correctly identify positive instances (TPR and FNR).

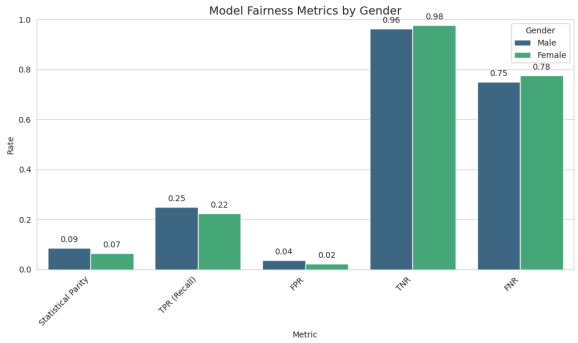

Figure 2: Fairness metrics of the model divided by gender

- **Statistical Parity:** Statistical Parity establishes if the number of individuals receiving a positive outcome is the same between protected groups. In the considered scenario, the metric for males is 0.09 and for females is 0.07. and this close-to-zero difference (0.02) suggests that there's a balanced distribution of the positive outcomes, between the two genders.

- **True Positive Rate:** True Positive Rate, or Recall, quantifies how many true positive instances have been correctly identified by the model. In this case, the TPR is 0.25 for men and 0.22 for females and even if the difference is relatively small (0.03), the model seems more efficient in predicting correctly the true positive instances inside the group Male. Nonetheless, both values are low, and this suggests the model might struggle in identifying the true positives, regardless of the gender.

- **False Positive Rate:** False Positive Rate computes the proportion of negative instances that have been incorrectly classified as positive. The values on the graph for this metric show 0.04 for men and 0.02 for women. The difference is very small and overall, the values are low for both genders, meaning that the errors made by the model are few, especially for the second protected group.

- **True Negative Rate:** True Negative Rate shows the proportion of negative instances that have been correctly identified as negative by the model. The metric is relatively high and similar for both groups: indeed, we have 0.96 and 0.98, for men and women respectively. This suggests that the model is efficient in identifying correctly the negative instances for both genders, for the group 'Female'.

- **False Negative Rate:** False Negative Rate computes the proportion of positive instances that are wrongly classified as negative instances. As seen, the values are 0.75 for men and 0.78 for women, which are high for both genders. This means the model struggles identifying a great portion of true positive values, especially when treating instances related to the group 'Female'.

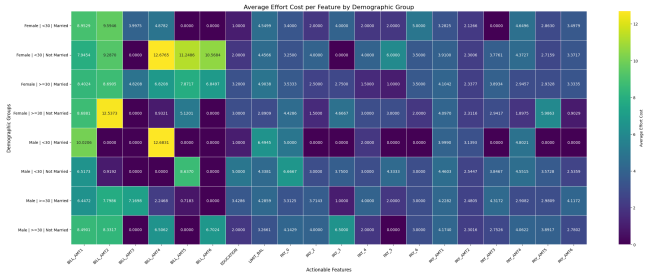

Figure 3: Heatmap with run = 1

In conclusion, the model seems to favor sensitivity (TPR) for males and specificity (TNR) for females. This suggests a potential bias in how the model assigns errors. For males, the model is more lenient in predicting positives (higher TPR but also higher FPR). For females, the model is more conservative in predicting positives (higher TNR but also higher FNR).

4.3 Counterfactual Analysis

This segment of the paper considers in detail the effort required to change an outcome from 'default' (1) to 'granted' (0). The results of this study are presented in two different sections that coincide with the number of runs performed to generate the counterfactuals, which are then used to quantify the effort.

4.3.1 With run = 1. In this section, the results that are taken into consideration are the ones obtained by generating a single counterfactual per instance. This approach provides an initial estimate of the effort required to change the outcome.

A. **Heatmap**
Figure 3 shows a matrix where the cells contain the distribution of the average effort cost associated with modifying specific features, for each of the eight groups, depending on the 20 features. The rows represent various demographic groups, sorted by gender, age, and marital status (e.g., 'Female | <30', 'Male | >=30 | Not Married', etc.). The columns correspond to the mutable and actionable features of the model, such as 'BILL_AMT'. Based on the generation of one counterfactual for each instance, an average effort is computed and depending on the value and the protected features involved, a color intensity is associated with the corresponding cell.
A preliminary analysis of the heatmap reveals several observations. Firstly, some features require consistently more effort across multiple demographic groups. For instance, 'BILL_AMT' features – especially 'BILL_AMT3' (e.g., 12.8766 for 'Female | <30 | Not Married' and 12.6831 for 'Male | <30'), and 'BILL_AMT4' (e.g., 11.2465 for 'Female | <30 | Not Married') – show one of the highest effort costs, suggesting that changing the amount of bill statement is the most challenging pathway to change an outcome. Instead, features like 'EDUCATION' and 'LIMIT_BAL' often present relatively low or zero effort costs, indicating they might be easier to adjust. At last, some patterns emerge: for instance, the 'Female |

<30 | Not Married' and 'Male | <30' groups exhibit high effort values in critical 'BILL_AMT' features. In addition, the group 'Male | >=30 | Not Married' also shows a high effort for 'BILL_AMT2' (12.5373). All these insights showcase how relevant some barriers are for some demographic segments and features to be overcome.

B. **Average Effort by Subgroup**

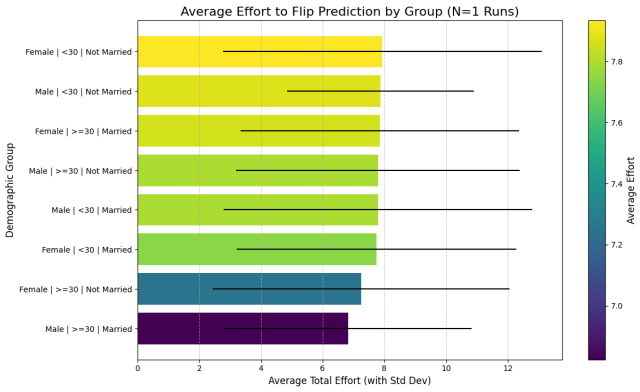

Figure 4: Average Effect by Subgroup with run = 1

Figure 4 pictures a horizontal histogram: on the Y-axis, all eight demographic groups are ranked based on the average effort, while on the X-axis, the length represents the average total effort, including the standard deviation bars, which represents the variability within each group.
As seen from this graph, the 'Male | >=30 | Married' group consistently shows the lowest average total effort, followed by 'Female | >=30 | Not Married', with both groups indicating a comparatively lower burden to achieve the wanted outcome. Instead, groups like 'Female | <30 | Not Married' and 'Male | <30 | Not Married' appear they require a higher average total effort. These results show that younger and unmarried people, regardless of the gender, struggle in flipping the outcome. Regarding the standard deviation, the 'Male | >=30 | Married' has one of the highest, which suggests a broader range of individual efforts within that group.

C. **Average Effort by Gender**
Figure 5 represents a simplified and summarized view of the average total effort, aggregated according to the gender: on the Y-axis the two groups – 'Male' and 'Female' - can be distinguished, while on the X-axis, the Average Total Effort is quantified, using standard deviation bars. From the histogram, the 'Female' group shows a slightly higher average total effort (7.75 circa) compared to the 'Male' group (7.50 circa). This difference suggests that on average, females need a marginally greater effort to change their status to 'granted'. Instead, if the standard deviation is taken into account, both groups show a significant variability in the effort.

4.3.2 With run = 5. The following segment refers to the results obtained by generating five counterfactuals from each instance.

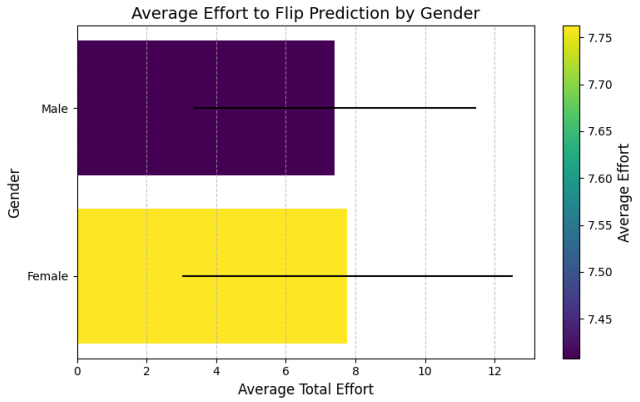

Figure 5: Average Effort by Gender with run = 1

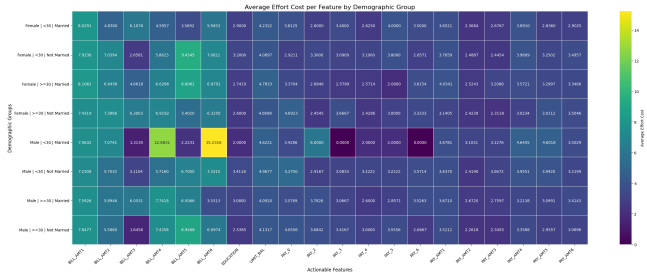

Figure 6: Heatmap with run = 5

This approach contributes to the initial estimation, offering a more complete and robust view of the efforts involved.

A. **Heatmap**
Similar to the 'run = 1' analysis, Figure 6 represents a heatmap where rows represent the eight demographic groups, and columns denote the mutable and actionable features. However, the cells collect the aggregated average effort required to modify that specific feature for individuals within that category to flip the outcome. This different and multi-run approach aims to reduce the sensitivity derived by using a single counterfactual and provide a more robust estimate of the effort.
As seen in the previous case with 'run = 1', the general patterns of effort still show consistency. Features like 'BILL_AMT' (especially 'BILL_AMT3', 'BILL_AMT4', and 'BILL_AMT5') still come up as the highest effort-demanding across most demographic groups. For instance, 'BILL_AMT3' for 'Male | <30' shows an average effort of 13.3100, and 'BILL_AMT4' for the same group is 22.2231, consistent with their high effort in the 1-run analysis. Similarly, 'BILL_AMT4' for 'Female | <30 | Not Married' (9.4345) and 'BILL_AMT5' for 'Female | <30 | Married' (9.8398) also remain in the same range of effort. On the other hand, 'EDUCATION' and 'LIMIT_BAL' maintain a lower effort in both runs. At last, the range of average effort cost increases up to 15.2316, compared to the maximum in 1-run, which is 12.6831.

While all the trends are generally stable, the 5-run aggregation provides more average effort values, while potentially smoothing out outliers that might have had an impact on the 1-run estimations.

B. **Average Effort by Subgroup**

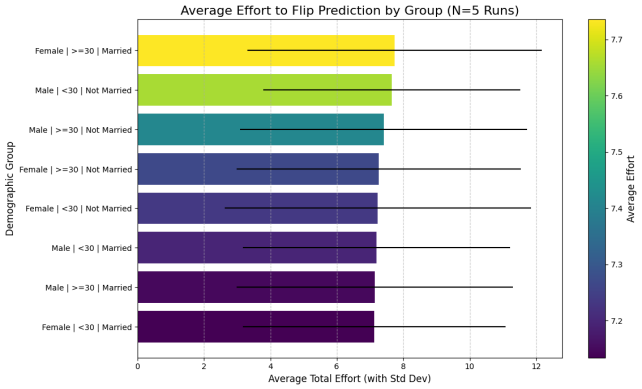

Figure 7: Average Effect by Subgroup with run = 5

As with 1-run, the histogram in Figure 7 represents the average effort cost depending on the subgroup we are considering.

Like for the heatmap, a consistency between the two graphs for 1-run and 5-run is remarkable. Groups such as 'Female | >=30 | Married' and 'Male | >=30 | Not Married' continue to exhibit the lowest average efforts, suggesting an easier path to the 'granted' outcome for these segments. Instead, demographic groups like 'Female | <30 | Not Married' and 'Male | <30 | Not Married' still require a higher average total effort, reinforcing the conclusion that younger and unmarried individuals usually face a greater burden. Finally, the standard deviation bars show variability persists, but the effort is more robustly defined, and the range of average effort cost narrows a bit, becoming more precise.

Overall, this approach seems to confirm what was found in 1-run, while providing stronger assessment of the true cost required.

C. **Average Effort by Gender**

Figure 8 shows the average total effort required to flip the outcome, aggregated by gender. This provides a high-level comparison between 'Male' and 'Female' groups.

Consistent with the macro analysis made in the 1-run case and represented in Figure 3, if we consider multiple counterfactuals, the 'Female' group continues to exhibit an average total effort equal to 7.47, which is slightly higher than the one for 'Male' group, that corresponds to 7.42. Even if the absolute difference between the two groups remains small, this persistent trend across both runs suggests that females require marginally greater effort to achieve the wanted result. At last, the standard deviation bars indicate that the variability in effort is still present in both groups, but the

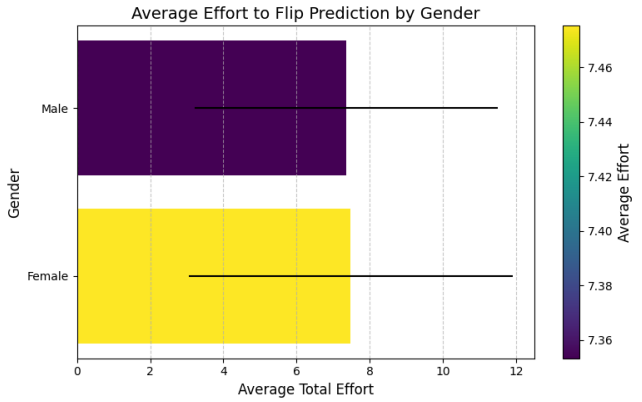

Figure 8: Average Effort by Gender with run = 5

central tendency of the average effort remains stable and consistent with previous findings.

Overall, the aggregating approach with 5 counterfactuals per instance (run=5) reveals how the small disparity between males and females in the average effort required to flip their outcome from 'default' to 'no default' is a persistent characteristic of the model.

5 DISCUSSION AND CONCLUSION

Fairness in machine learning involves not only predicting outcomes but also how accessible favorable results are to different groups. Our results suggest that while traditional group fairness metrics (such as TPR) show relatively small differences between male and female groups, both groups have high false negative rates. It indicates an overall limitation of the model in identifying the true defaulters.

However, standard fairness metrics alone do not capture the full picture. The standard fairness metrics show that the model is more "harsh" on males in terms of classification frequency, but the counterfactual fairness analysis shows that the model is more "rigid" on females in terms of actionable changes. In other words, while males are more likely to be predicted to default, when females are predicted to default, it is more difficult for them to change their outcome (shown by the higher cost in the counterfactual fairness analysis). This suggests fairness interventions should consider both outcome rates and effort costs, since improving one metric alone may worsen the other.

Beyond the fairness metrics discussed earlier, the logistic regression model's overall performance also reveals important limitations in the context of fairness. While the model achieves relatively high accuracy (0.81) and strong precision for the majority class ('no default'), it performs poorly in identifying the minority class ('default'), with a recall as low as 0.24 and an F1-score of only 0.35. This imbalance suggests that many individuals who are likely to default are misclassified as safe. If true defaulters are not correctly identified, the evaluation of who faces unjust outcomes becomes less reliable. Those issues would distort fairness assessments.

Building on these, our counterfactual analysis focuses on measuring the effort individuals need to obtain a positive prediction under fixed protected attributes. By keeping protected attributes

(e.g. gender, age, marital status) fixed, and allowing only actionable features to change, we measured the counterfactual effort required to move from a negative to a positive decision. The results consistently showed that younger, unmarried individuals (especially females) faced higher average effort. While these differences were not large, they persisted across both single (run=1) and multiple (run=5) counterfactual runs, revealing fairness gaps not captured by standard metrics.

This confirms that fairness is multidimensional. Even when prediction rates appear balanced, some groups may face more barriers in achieving 'no default'. Therefore, evaluating only outcome fairness (such as statistical parity or TPR) may overlook structural disadvantages embedded in the model's behavior.

Future work could improve the current approach in several ways. First, counterfactual analysis could be restricted to true positive cases. In other words, generating counterfactuals not for all the instances that the model predicted as 1 (default), but only for the true positive instances. This could better reflect the fairness criterion of equal opportunity and avoid bias introduced by misclassified instances. Additionally, the current effort calculation treats all features equally, but in practice, some changes (like increasing credit limit or changing education level) may be more difficult than others. In order to assign realistic, feature-specific cost weights would make the effort metric more accurate and better reflect individual burdens across groups.

A ETHICS STATEMENT

This project uses a publicly available dataset: the Default of Credit Card Clients Dataset from the UCI repository. It contains anonymized financial and demographic information of credit card clients. No new data was collected, and no personally identifiable information is included in this dataset. Therefore, no direct ethics board approval (e.g., IRB) was required. However, we acknowledge that even with anonymized data, fairness and downstream use remain important ethical concerns.

Our analysis focuses on model fairness across demographic groups, especially gender, age, and marital status. While the aim is to promote transparency and fairness in credit risk prediction, the model may still reinforce or reveal existing societal inequalities. For example, higher counterfactual effort observed in certain groups (e.g., younger, unmarried females) could reflect systemic disadvantages. If such models are deployed in real-world credit decision-making without proper oversight, they may contribute to indirect discrimination or reduced financial access for vulnerable groups.

Moreover, our model assumes equal cost for changing all actionable features. In reality, financial and social constraints may limit individuals' ability to adjust certain variables (e.g., education level, credit limit). It raises ethical concerns about representational fairness and real-world feasibility.

To mitigate these concerns, our project emphasizes transparency by reporting group-specific outcomes and advocating for the inclusion of effort-based fairness in model evaluation.

B CONTRIBUTIONS

Table 3 shows the contributions of each member to the project.

Contribution	Alexia	David	Yating	Yixuan
Presentation	✓		✓	
Code	✓	✓		
Poster			✓	✓
Introduction				✓
Related Work				✓
Methods	✓			
Results		✓		
Conclusion			✓	

Table 3: Team Contributions

C CODE REPOSITORY

The link to the code repository is the following: https://github.com/alexia-nt/Human-Centered-Machine-Learning. The python notebook contains additional graphs for the reader, like dataset counts and additional graphs regarding the counterfactual fairness analysis.

REFERENCES

[1] Rachel KE Bellamy, Kuntal Dey, Michael Hind, and et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943* (2018).

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[3] Richard Berk, Hoda Heidari, and et al. 2017. Fairness in criminal justice risk assessments: The state of the art. In *Sociological Methods & Research*.

[4] Cynthia Dwork, Moritz Hardt, Toni Pitassi, and et al. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*.

[5] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*.

[6] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[7] Alejandro Kuratomi, Evaggelia Pitoura, Panagiotis Papapetrou, Tony Lindgren, and Panayiotis Tsaparas. 2022. Measuring the burden of (un) fairness using counterfactuals. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 402–417.

[8] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*.

[9] L. Longo, S. Lapuschkin, and C. Seifert. 2024. *Explainable Artificial Intelligence: Second World Conference, XAI 2024, Valletta, Malta, July 17–19, 2024, Proceedings, Part I.* Springer. https://books.google.com/books?id=mD0TEQAAQBAJ.

[10] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[11] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[12] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* (2017).