01857560

1. Introduction

In this report, we evaluate the trade-off between accuracy and fairness in Machine Learning (ML) classification models. We compare standard ML models (Task 1) with fairness-aware ones (Task 2), focusing on their performance as influenced by a trade-off hyperparameter. Specifically, we employ Logistic regression with the 'libilinear' solver, adjusting the hyperparameter 'C' (inverse of L2 regularizer strength) from 10^{-9} to 10^{2} in logarithmic increments of one order of magnitude. We evaluate the impact on accuracy and fairness, using accuracy and Equal Opportunity Difference (EOD) metrics, respectively. Furthermore, we propose a criterion for selecting an optimal model that balances accuracy and fairness (Task 3). In this study, we use the data from the 2018 American Community Survey (ACS) for Florida, focusing on disability recode ('DIS') as the sensitive variable for our fairness analysis. The target variable for the machine learning task is employment status ('ESR'), transformed into a binary outcome indicating employment. Note that we further pre-processed the data by applying one-hot encoding to all features except 'SCHL', which represents educational attainment and was treated as an ordinal variable. We did this to enhance the models' accuracy and interpretability.

2. Task 1

This task aims to evaluate the performance of standard models identified as the most accurate and fairest based on mean evaluation metrics across five iterations. In each iteration, the ML algorithm is trained on a new, randomly generated 80-20 train-validation split, representing 70% of the entire dataset. The remaining 30% of the data, constitutes the unseen (test) data set aside for final evaluation.

	C=10^{-7}	C=10^{-5}	C=10^{-4}	C=10^{-3}	C=10^{-2}	C=10^{0}
Accuracy	0.7241	0.7356	0.7585	0.7774	**0.7788**	0.7788
EOD	0.8495	0.8495	0.7774	0.6343	0.5770	**0.5675**

Table 1. Performance results of standard Logistic regression models at key C values on validation data.

As we observe from Table 1, Figure 1, and Figure 2, our results reveal an unexpected trend: both accuracy and fairness improve as regularization strength decreases (i.e., C increases), challenging the conventional belief of a trade-off between these two metrics in ML models. Specifically, we observe a notable increase in accuracy from $C = 10^{-5}$ and $C = 10^{-3}$, peaking at $C = 10^{-2}$ with a value of 0.7788, and then stabilizing at higher C values. EOD consistently decreases with increasing C, reaching its best score of 0.5675 at $C = 10^{0}$. This trend is particularly pronounced from $C = 10^{-5}$ to $C = 10^{-2}$, suggesting that high regularization may excessively penalize weights linked to positive outcomes for the less favored class. This observation contradicts the typical expectation that strong regularization reduces overfitting to the majority group, thereby lowering the equal opportunity difference.

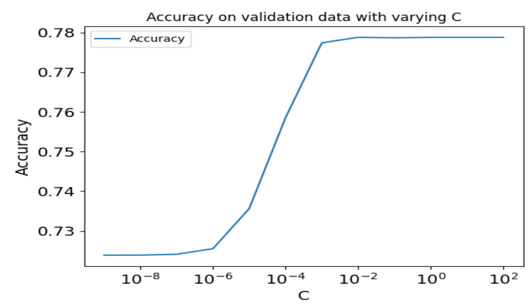

Figure 1. Mean accuracy of standard Logistic regression models across varying C values.

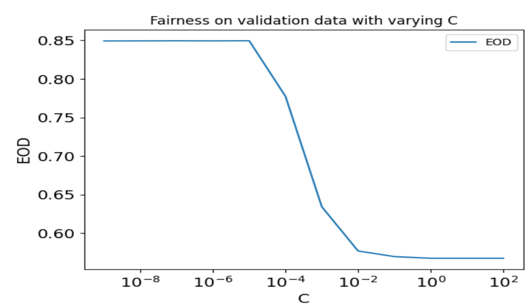

Figure 2. Mean EOD of standard Logistic regression models across varying C values.

	C=10^{-2}
Accuracy	0.7774
EOD	0.5858

Table 2. Performance results of the standard **most accurate** Logistic regression model on the test data.

	C=10^{0}
Accuracy	0.7775
EOD	0.5741

Table 3. Performance results of the standard **fairest** Logistic regression model on the test data.

Upon testing, the most accurate and fairest models displayed comparable performance in both accuracy and fairness. The fairest model slightly outperformed the most accurate model in accuracy, with a marginal difference of approximately 10^{-4}. This minor discrepancy aligns with our model selection criteria, where the most accurate model exhibited only a slightly higher accuracy of about 10^{-5} than the fairest model. Additionally, the fairest model exhibited a marginally lower EOD, indicating a better balance in fairness. Nevertheless, the results do not provide substantial evidence to determine whether or not better generalization could correspond to a fairer model.

3. Task 2

This task follows the same procedure as in Task 1, but evaluates fairness-awar models instead. For a fairness-aware method, we apply the reweighing pre-processing method developed by Kamiran and Calders.

	C=10^{-7}	C=10^{-5}	C=10^{-4}	C=10^{-3}	C=10^{-2}	C=10^{0}
Accuracy	0.7134	0.7220	0.7376	**0.7470**	0.7453	0.7446
EOD	0.1531	0.0983	0.0256	**-0.0153**	-0.0243	-0.0278

Table 4. Performance results of fairness-aware Logistic regression models at key C values on validation data.

As we observe from Table 4, Figure 3, and Figure 4, there is a notable increase in accuracy from $C = 10^{-7}$ and $C = 10^{-3}$, peaking at $C = 10^{-3}$ with a value of 0.7470, and then stabilizing at higher C values with a slightly lower accuracy. There is, however, a minimal variation across the accuracy values, with a coefficient of variation of 2.04%, suggesting that the reweighing method has likely optimized the dataset for accuracy. In contrast, fairness varies more significantly across different C values, with the most substantial improvement occurring from $C = 10^{-7}$ to $C = 10^{-3}$. The fairest outcome, with the lowest absolute EOD, is at $C = 10^{-3}$ with a value of 0.0153. Notably, for C values larger than 10^{-4}, the EOD becomes negative, indicating a shift in bias towards the other group. This highlights the impact of regularization on fairness, distinct from its effect on accuracy, likely due to reweighing reducing inherent biases and enhancing the model's fairness sensitivity to regularization changes.

	C=10^{-3}
Accuracy	0.7440
EOD	0.0181

Table 5. Performance results of the fairness-aware **most accurate & fairest** Logistic regression model on the test data.

Table 5 reveals that the selected model, which is both the most accurate and fairest, has an accuracy 4.30% lower than the most accurate normal model and an EOD value 96.85% lower than the fairest standard model. These results indicate the effectiveness of the reweighing method in significantly improving fairness with only a marginal reduction in accuracy compared to the standard model. The trade-off between accuracy and fairness becomes evident when comparing the test data performance of our standard models with that of our fairness-aware models.

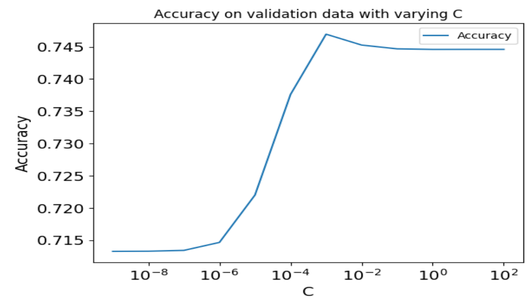

Figure 3. Mean accuracy of fairness-aware Logistic regression models across varying C values.

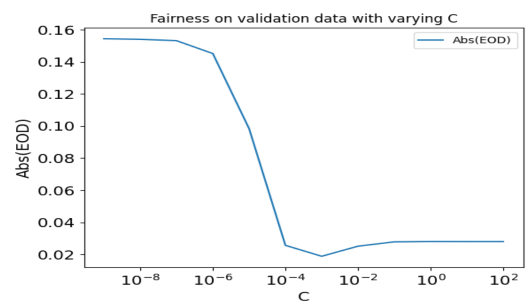

Figure 4. Mean absolute EOD of fairness-aware Logistic regression models across varying C values.

4. Task 3

This task proposes a strategy that balances accuracy and fairness to identify optimal models, both with and without fairness-aware methods (specifically reweighing) applied. This strategy introduces a new evaluation metric, termed 'fairacc', which combines accuracy and EOD in the following way:

$$\begin{aligned} \text{fairacc} &= 0.75 \times \min(\text{accuracy}, 1 - |\text{EOD}|) \\ &\quad + 0.25 \times \max(\text{accuracy}, 1 - |\text{EOD}|) \end{aligned}$$

where the weights 0.75 and 0.25 correspond to the importance assigned to accuracy and fairness accordingly, ensuring their sum equals 1 to maintain the metric's range within [0,1]. The absolute value of EOD is used to indicate bias regardless of direction, and $1 - |\text{EOD}|$ aligns its range and interpretation with that of accuracy.

Regarding the standard model selection, as illustrated in Table 6 and Figure 5, we observe a notable improvement in performance as the value of C increases, particularly from $C = 10^{-5}$ to $C = 10^{-2}$, and reaches its peak at $C = 10^{0}$ with a fairacc score of 0.5191. This trend aligns with expectations, as a larger weight is given to fairness and, recalling from Task 1, fairness decreases with higher C values. The chosen model, at $C = 10^2$, registers a fairacc of 0.5138 on test data.

	C=10^{-9}	C=10^{-5}	C=10^{-4}	C=10^{-3}	C=10^{-2}	C=10^{0}
Fairacc	0.2940	0.2968	0.3566	0.4686	0.5120	**0.5191**

Table 6. Fairacc results of standard Logistic regression models at key C values on validation data.

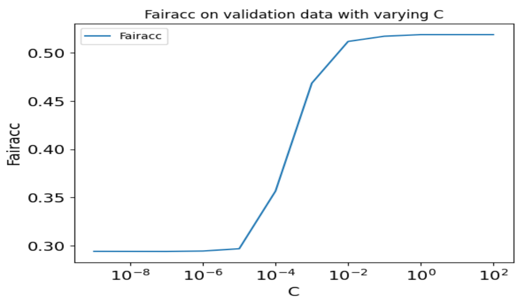

Figure 5. Mean fairacc of standard Logistic regression models across varying C values.

	C=10^{-9}	C=10^{-6}	C=10^{-5}	C=10^{-4}	C=10^{-3}	C=10^{0}
Fairacc	0.7464	0.7497	0.7669	0.7968	**0.8055**	0.8014

Table 7. Performance results of fairness-aware Logistic regression models at key C values on validation data.

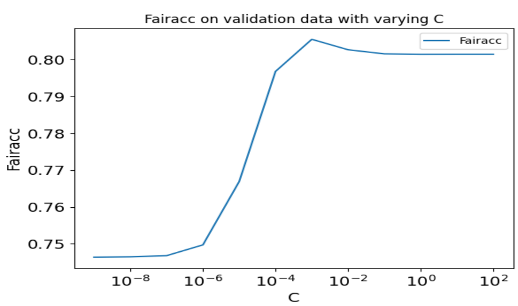

Figure 6. Mean fairacc of fairness-aware Logistic regression models across varying C values.

Regarding the fairness-aware model selection, as shown in Table 7 and Figure 6, there is less variation in fairacc compared to the fairacc of the standard models, with the most significant increase occurring from $C = 10^{-6}$ to $C = 10^{-3}$, where it peaks and then slightly declines for larger C values. Fairness-aware models consistently outperform standard models when evaluated with the fairacc metric, attributed to reweighing's significant reduction in fairness values with minimal impact on accuracy. The chosen model, at $C = 10^{-3}$, achieves a fairacc of 0.8035 on the test data.

5. Conclusion

Intriguingly, across all the experiments, C values ranging from 10^{-5} to 10^{-3} seem to have the most significant impact on the model performance, both in terms of accuracy and fairness, possibly indicating the optimal regularisation range. Additional considerations in our experiments include the concept of intersectionality, where individuals have multiple, overlapping identities that may lead to distinct experiences of discrimination or privilege[1]. Assessing ML model fairness based on a single sensitive variable, though useful, overlooks these complexities. For example, a model might be fair in terms of gender or disability when evaluated separately, but could fail to address the unique needs of intersectional groups, such as women with disabilities. This underscores the necessity of incorporating intersectionality into ML fairness research. Moreover, our approach employs the equal opportunity difference metric, focusing on equalizing true positive rates across groups. A more holistic approach might involve the equalized odds metric, which considers both true and false positive rates.

6. Extra Section

In this section, we perform model selection using the Florida dataset and evaluate the selected model on the Texas dataset. Although using the entire Florida dataset for model selection could yield more robust models, we retain the same data splits as in the previous section to ensure fair comparisons when assessing final model performance on both Florida and Texas test data. So we use the models 1-6 produced in the previous sections, and test them on the Texas state dataset.

Model	1	2	3	4	5	6
C	10^{-2}	10^{0}	10^{-3}	10^{-3}	10^{0}	10^{-3}
Accuracy	0.7588	0.7583	0.7294	0.7294	-	-
EOD	0.5660	0.5537	0.0292	0.0292	-	-
Fairacc	-	-	-	-	0.5243	0.7898

Table 8. Performance results of the 6 chosen Logistic regression models on the test data.

The results from Table 8 are largely consistent with those of our main study, with some minor deviations. Notably, the average accuracy of the first four models on Texas data is 2.20% lower than on Florida data. This slight difference, although not significant, may indicate differences in employment criteria between the two states, as the model optimized for Florida shows slightly reduced effectiveness when test on Texas. On the other hand, the mean absolute EOD for these models is 1.50% lower in Texas, implying similar yet potentially more balanced demographic structures for individuals with disabilities in Texas compared to those in Florida.

References

[1] U. Gohar and L. Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. 2023. 3