Runyi Yang

1. Introduction

In this coursework, I explore the pressing issue of algorithmic fairness within machine learning models, particularly examining the tension between model accuracy and fairness. The objective is to understand and demonstrate how varying levels of regularization affect this balance. Specifically, this coursework concerns about 3 tasks. One is to analyse whether a standard machine learning model could or not to correspond to a fairer model. The second is to choose a fairness-aware method like reweighing. The last is to based on my observations, propose a model selection strategy that could account for both accuraccy and fairness. Accordingly, I take Logistic Regression (LR) [5] based on scikit-learn library [4] as the base model to analyze accuracy and fairness. To further enhance the depth of this research, two empirical studies are conducted. The first examines the sensitivity of features in relation to fairness outcomes, while the second assesses the generalizability of the model across diverse scenarios.

2. Methods and Evaluation Metrics

In this course work, I train model on LR model and evaluate based on classification accuracy and equal opportunity difference. The dataset used for model training and evaluation is ACS American Community Survey dataset [2] from the folktables [1]. The description of the model and evaluation metrics are discussed below.

2.1. Logistic Regression Model Description

Logistic Regression (LR) is a widely used statistical method for binary classification. It models the probability of a binary response based on one or more predictor variables. The formula for LR is given by:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \tag{1}$$

where $P(Y=1|X)$ is the probability of the event occurring (class 1), X represents the predictor variables, and β_0 and β_1 are the parameters of the model that are learned from the training data.

2.2. Evaluation Metrics

To assess the performance of the classification model, accuracy score is commonly employed as the primary metric. Alongside accuracy, evaluating fairness is crucial, especially in the context of machine learning models. One of the most significant fairness metrics is the Equal Opportunity Difference (EOD), as highlighted by Radovanovic et al. [6].

2.2.1 Accuracy

Accuracy is a fundamental metric in classification tasks which is calculated as the proportion of correct predictions (both true positives and true negatives) to the total predictions made. Mathematically, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{2}$$

In this equation, TP represents the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

2.2.2 Equal Opportunity Difference (EOD)

The Equal Opportunity Difference (EOD) is a crucial measure of fairness in machine learning models, emphasizing the disparity in True Positive Rates (TPR) between different demographic groups. It can be mathematically defined as:

$$\text{EOD} = \text{TPR}_{D=d_1} - \text{TPR}_{D=d_2} \tag{3}$$

The True Positive Rate (TPR), also known as sensitivity, for a given demographic group $D=d$ is defined as:

$$\text{TPR}_{D=d} = \frac{\text{TP}_{D=d}}{\text{TP}_{D=d}+\text{FN}_{D=d}} \tag{4}$$

where $TP_{D=d}$ and $FN_{D=d}$ are the counts of True Positives and False Negatives for the demographic group $D=d$, respectively. Here, $\hat{Y}$ represents the predicted outcome, Y is the actual outcome, and D symbolizes different demographic groups, such as d_1 and d_2. A lower value of EOD signifies a more equitable model, indicating a closer approximation to equal opportunities across different groups.

3. Results

3.1. Task 1: Accuracy Aided Model

The aim of Task 1 is to assess the inherent fairness of a standard Logistic Regression model when applied to a dataset without implementing any specific fairness interventions. I firstly split the data to train-validation-test set. Then I run model selection across the multiple train-validation folds and select model based on the highest accuracy and lowest fairness metrics. I argue that the generalisation is shown on both accuracy and fairness.

3.1.1 Task 1(a): Train-Test Split

In Task 1, the dataset underwent a structured splitting process to ensure effective training and validation of the Logistic Regression model. This process is outlined as follows:

1. **Initial Split:** The dataset was first divided into a training set (70%) and a testing set (30%).

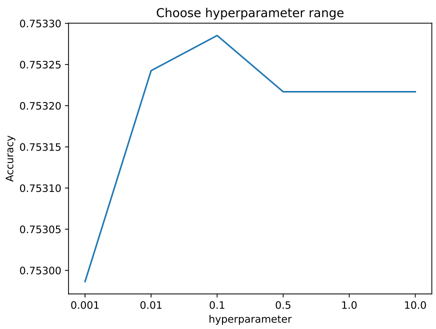

Figure 1. Choose the approximate range of hyperparameter C. Finally, I choose (0.01, 1.0) as the final range to finetune.

2. **Secondary Split:** The training set was further divided into train-train (80%) and train-val (20%) sets.
3. **Repetitions for Robustness:** The train-train/train-val splitting was repeated four more times with shuffling, resulting in five unique sets for model evaluation.

3.1.2 Task 1(b): Model with the Highest Accuracy

In the initial stage of the experimentation, the hyperparameter C for the Logistic Regression model was selected from a broad range. This preliminary selection was based on evaluating the model's accuracy scores across different C values. The results of this initial analysis are illustrated in Fig. 1, where the approximate optimal interval for C was identified. Following this preliminary assessment, a narrowed range of C between 0.01 and 1.0 was chosen for subsequent experiments to fine-tune the model's performance.

As seen in Fig. 2, the best model with the highest accuracy is with the hyperparameter $C = 0.08$. The accuracy and the fairness score on the test set is reportes in Tab. 1. The accuracy is 0.75054 and the EOD score is 0.60537.

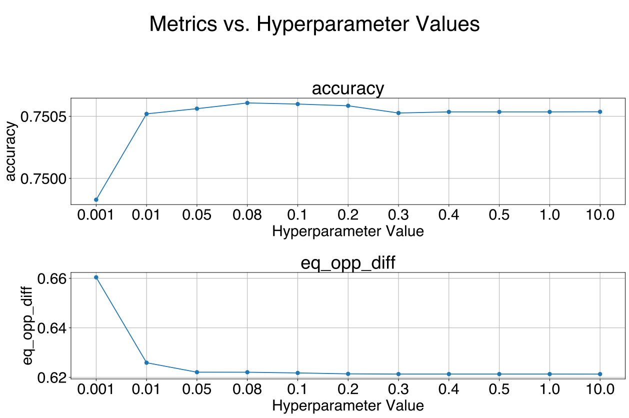

Figure 2. Model selection based on accurcacy and EOD scores on the training and validation set.

Table 1. Comparison of Models on Test Set without concerning fairness

	C	Accuracy	EOD
Accuracy-Based Model	0.08	0.75054	0.60537
EOD-Based Model	0.3	0.75064	0.60541

3.1.3 Task 1(c): Model with the Best fairness metric

As seen in Fig. 2, the best model with the lowest EOD score is with the hyperparameter $C = 0.3$. The results show a very similar result when choosing the model based on the accuracy. Besides, the accuracy of Accuracy-Based Model is slightly lower than the EOD-Based Model and the EOD value of Accuracy-Based Model is higher than the EOD-based model. This is because the difference and unbalance of the data between train-validation set and the test set.

Since the model performs similar on accuracy and fairness metric, I could give a rough conclusion based on the experimental results that the standard model could demonstrate the generalisation on both accuracy and fairness.

3.2. Task 2: Fairness Aided Model

The objective of Task 2 is to apply a fairness-aware preprocessing method to the data before training the Logistic Regression model. The chosen method for this study is reweighing [3]. This technique aims to mitigate bias by assigning different weights to the instances in the training data, aiming to equalize the impact of the positive and negative classes across protected and unprotected groups.

To make the model aware of fairness, the reweighing strategy is adopted. The loss function is adjusted:

$$L_{\text{reweighted}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_i \Big[y_i \log(\sigma(\mathbf{x}_i^\top \theta)) + (1-y_i)\log(1-\sigma(\mathbf{x}_i^\top \theta)) \Big] \quad (5)$$

3.2.1 Task 2(a): Model with the Highest Accuracy

Similar analysis as in Task 1(a), the performance metrics, specifically accuracy and EOD, were evaluated on the test set and are depicted in Figure 3.

The results demonstrate a marked decrement in the EOD score indicating the model better performance in fairness, in contrast to the model that does not account for fairness in Task 1. Concurrently, a marginal decrease in accuracy was observed, that is acceptable.

According to the best average accuracy score on the validation sets, the best model hyperparameter is $C = 1e-4$, with the accuracy and EOD score on the test set 0.71984 and 0.03572 respectively.

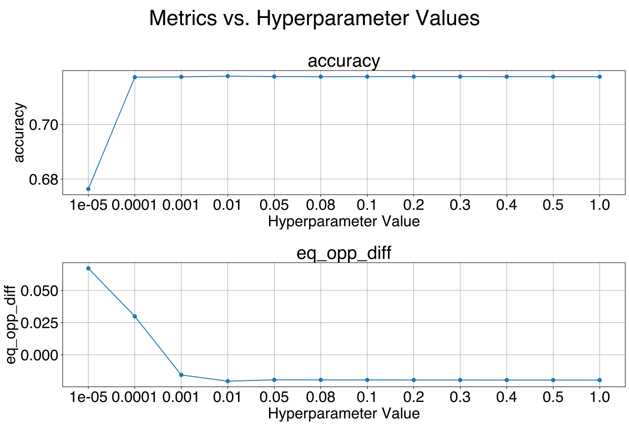

Figure 3. Model selection based on accurcacy and EOD scores on the training and validation set.

Table 2. Comparison of Models concerning fairness on Test Set

	C	Accuracy	EOD
Accuracy-Based Model	0.01	0.71984	0.03572
EOD-Based Model	0.001	0.71744	0.02807

3.2.2 Task 2(b): Model with the Best fairness metric

According to the best EOD score on the validation sets, the best model hyperparameter is $C = 1e - 3$, with the accuracy and EOD score on the test set 0.71744 and 0.02807 respectively.

Utilizing a methodology analogous to that employed in Task 1(a), the performance metrics, specifically accuracy and the EOD, were evaluated on the test set and are depicted in Figure 3. The results demonstrate a marked diminution in the EOD for the model incorporating fairness considerations, in contrast to the model that does not account for fairness. Concurrently, a marginal decrease in accuracy was observed. These findings underscore a discernible trade-off between fairness and accuracy in the model's performance, highlighting that enhancements in fairness are often accompanied by a slight compromise in accuracy. This trade-off is pivotal in the context of developing models that are not only effective but also ethically sound and equitable.

3.3. Task 3: Accuracy-Fairness Trade-off

In this task, I propose to add more criterion than only EOD to describe the fairness.

Statistical Parity Difference (SPD)

$$\text{SPD} = P(\hat{Y} = 1|D = d_1) - P(\hat{Y} = 1|D = d_2) \tag{6}$$

SPD measures the difference in the likelihood of a positive outcome between two demographic groups.

Average Odds Difference (AOD)

$$\begin{aligned}\text{AOD} = \frac{1}{2}&[(\text{FPR}D = d_1 - \text{FPR}D = d_2)\\ &+ (\text{TPR}D = d_1 - \text{TPR}D = d_2)]\end{aligned} \tag{7}$$

AOD evaluates fairness by averaging the differences in False Positive and True Positive Rates across groups.

Disparate Impact (DI)

$$\text{DI} = \frac{P(\hat{Y} = 1|D = d_1)}{P(\hat{Y} = 1|D = d_2)} \tag{8}$$

DI assesses fairness by comparing the ratio of positive outcomes between different demographic groups.

Based on these metrics, I selected the best model as seed in Fig. 3.

3.3.1 Task 3(a): Model with the Highest Accuracy

As the similar analysis in the former 2 tasks. The scores are shown in Tab. 3. The accuracy-based model shows the best on the accuracy but worse than others on the EOD score, with accuracy of 0.72005 and EOD of 0.036442.

Table 3. Comparison of Models concerning fairness on Test Set based on more evaluation metrics

	C	Accuracy	EOD
Accuracy-Based Model	0.09	0.72005	0.03642
EOD-Based Model	0.01	0.71984	0.03572
AOD-Based Model	0.05	0.72000	0.03651
DI-Based Model	0.08	0.72005	0.03638
SPD-Based Model	0.1	0.72003	0.03647

3.3.2 Task 3(b): Model with the Best fairness metric

The EOD-based model shows the best on the EOD score but the worst on the accuracy, with accuracy of 0.071984 and EOD of 0.03572.

4. Additional Research

4.1. Model selection on data for Florida and test on Texas

The model was trained utilizing RW on the Florida dataset, followed by testing on the Texas dataset, and following the training pipeline as same as that in Task 2. The outcomes of this process are detailed in Table 4. Analysis of the results indicates that the accuracy-based model demonstrates superior performance compared to the EOD-based model. Notably, the model predicated on accuracy exhibits enhanced cross-dataset generalization capabilities when compared to its EOD counterpart.

Table 4. Training on Florida data and test on Texas.

	C	**Accuracy**	**EOD**
Accuracy-Based Model	0.01	0.70114	0.00751
EOD-Based Model	0.001	0.70109	0.00764

4.2. Remove sensitive features

The model training follows the training pipeline as same as that in Task 2 and removes the sensitive features. After removing sensitive features, the model seems meaningless on evaluation metrics of EOD. Thus I just select the highest accuracy model with C=0.08 as in Task 1 and get the accuracy of 0.73291 .

References

[1] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021. 1

[2] D. H. Griffin and P. J. Waite. American community survey overview and the role of external evaluations. *Population Research and Policy Review*, 25:201–223, 2006. 1

[3] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012. 2

[4] O. Kramer and O. Kramer. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53, 2016. 1

[5] S. Menard. *Applied logistic regression analysis*. Number 106. Sage, 2002. 1

[6] S. Radovanović and M. Ivić. Enabling equal opportunity in logistic regression algorithm. *Management: Journal of Sustainable Business & Management Solutions in Emerging Economies*, 28(2), 2023. 1