4243/5243

Project 4 Group 1

Angel Wang, Arnulfo Andres Trevino, Mansi Singh, Miao Zhang, Nashita Rahman

We were tasked with A1 and A4.

- Learning Fair Representation
- DM (Disparate Mistreatment) and DM-sen (Disparate Mistreatment considering sensitive attributes)

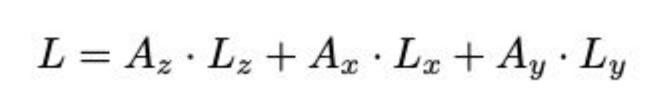
Summary for LFR model

Baseline model: Logistic regression

Goal:

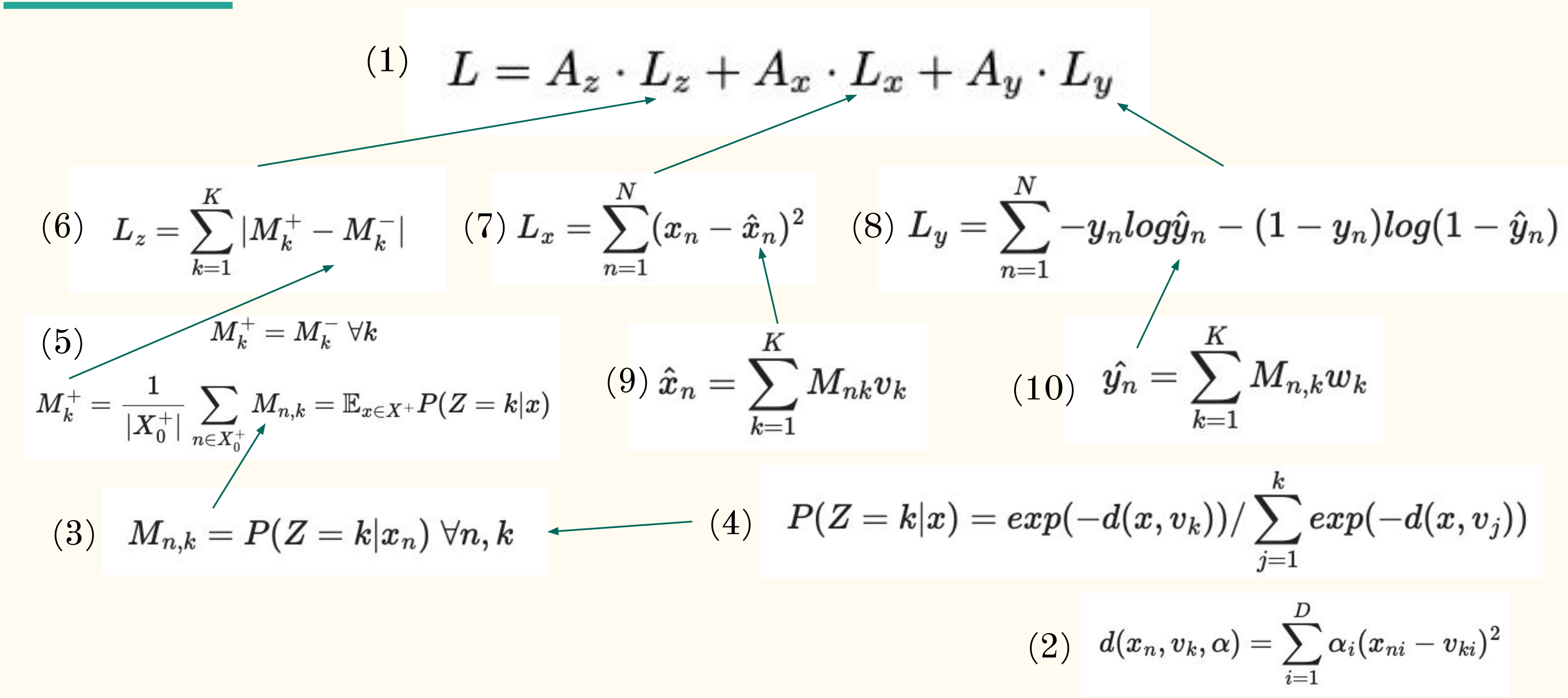
In Al, the authors propose a learning algorithm for fair classification that aims to achieve both group fairness and individual fairness.

This LFR algorithm minimizes an loss function with three terms corresponding to the goals of statistical parity, information preservation, and accurate classification.



$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

Formula Breakdown



(1) $$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

(6) $$L_z = \sum_{k=1}^{K} |M_k^+ - M_k^-|$$

(7) $$L_x = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2$$

(8) $$L_y = \sum_{n=1}^{N} -y_n log \hat{y}_n - (1 - y_n) log(1 - \hat{y}_n)$$

(5) $$M_k^+ = M_k^- \ \forall k$$

$$M_k^+ = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k} = \mathbb{E}_{x \in X^+} P(Z = k|x)$$

(9) $$\hat{x}_n = \sum_{k=1}^{K} M_{nk} v_k$$

(10) $$\hat{y_n} = \sum_{k=1}^{K} M_{n,k} w_k$$

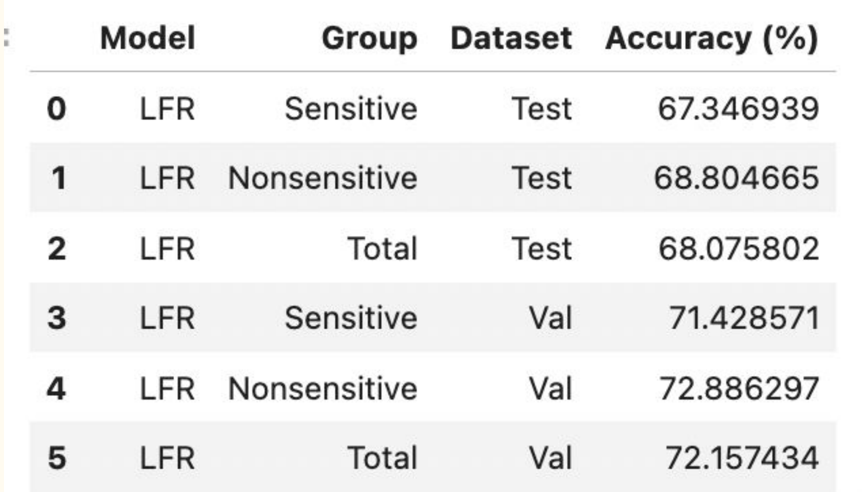
(3) $$M_{n,k} = P(Z = k|x_n) \ \forall n, k$$

(4) $$P(Z = k|x) = exp(-d(x, v_k)) / \sum_{j=1}^{k} exp(-d(x, v_j))$$

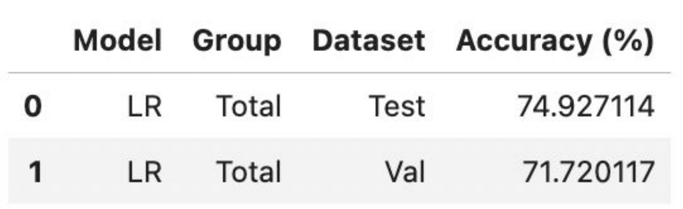
(2) $$d(x_n, v_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{ni} - v_{ki})^2$$

Result:

LFR vs. LR (baseline model)

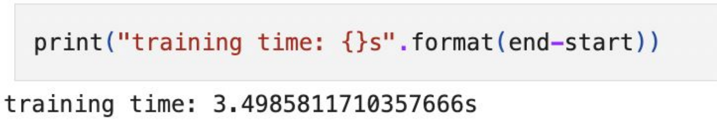


	Model	Group	Dataset	Accuracy (%)
0	LFR	Sensitive	Test	67.346939
1	LFR	Nonsensitive	Test	68.804665
2	LFR	Total	Test	68.075802
3	LFR	Sensitive	Val	71.428571
4	LFR	Nonsensitive	Val	72.886297
5	LFR	Total	Val	72.157434



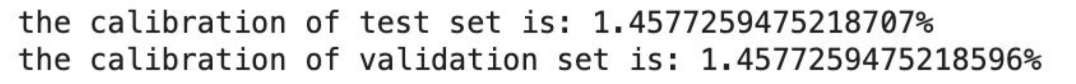
	Model	Group	Dataset	Accuracy (%)
0	LR	Total	Test	74.927114
1	LR	Total	Val	71.720117

LFR training time



```
print("training time: {}s".format(end-start))
```

```
training time: 3.4985811710357666s
```



```
the calibration of test set is: 1.4577259475218707%
the calibration of validation set is: 1.4577259475218596%
```

Sensitive = Caucasian
Nonsensitive = African-American

Summary for DM and DM-sen model

Goal:

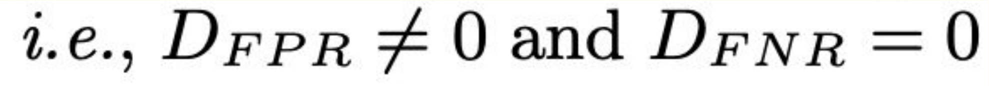
Addresses unfairness in automated decision-making systems, especially classification models.

Key Issue: Highlights the risk of disparate mistreatment (different misclassification rates across social groups)

- Introduces "disparate mistreatment," a comprehensive fairness metric that quantifies the adverse consequences of both false positives and negatives across demographic subgroups.
- Demonstrates the effectiveness of the proposed method on synthetic and real-world datasets, balancing fairness and accuracy.

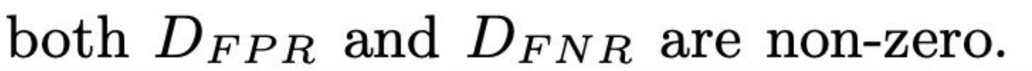
Fairness Constraints

- Disparate mistreatment on only false positive rate or false negative rate



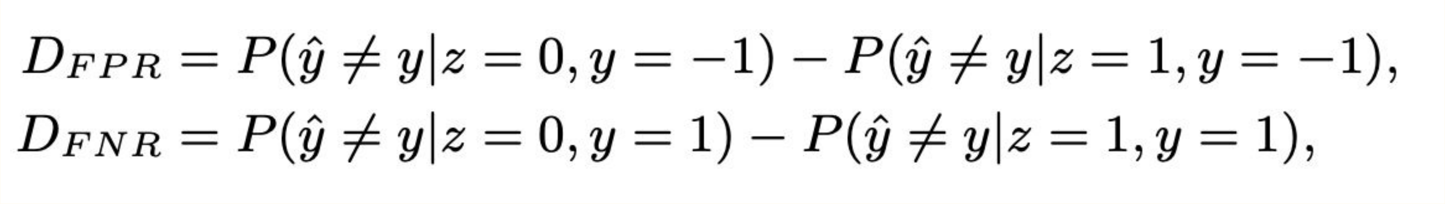
i.e., $D_{FPR} \neq 0$ and $D_{FNR} = 0$

- Disparate mistreatment on both false positive rate and false negative rate



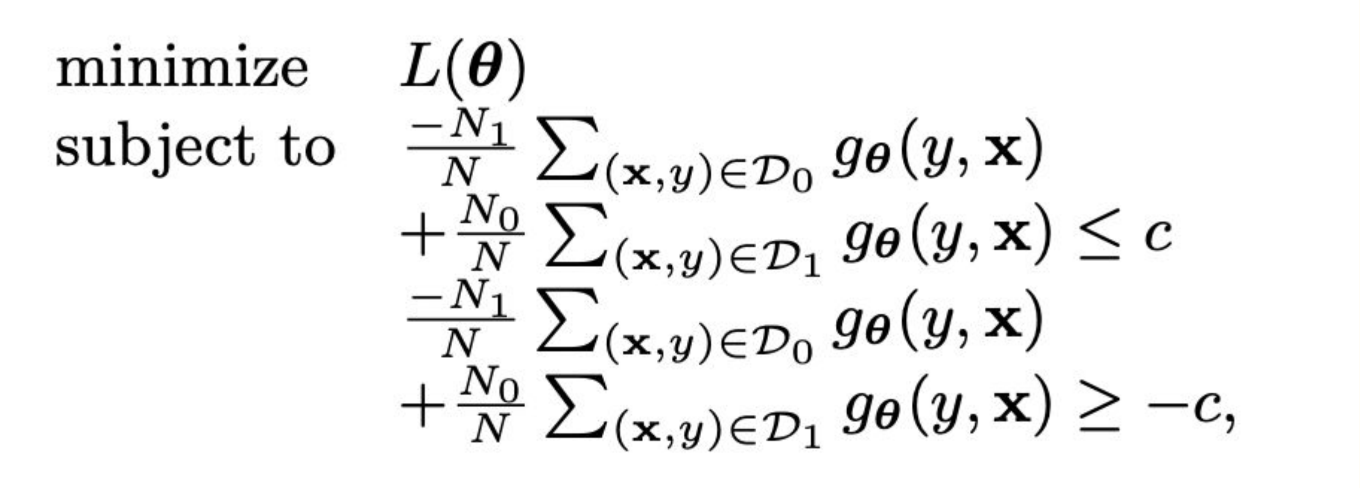
both D_{FPR} and D_{FNR} are non-zero.

Definition of D_FPR and D_FNR(z is sensitive attribute)



$$D_{FPR} = P(\hat{y} \neq y | z = 0, y = -1) - P(\hat{y} \neq y | z = 1, y = -1),$$
$$D_{FNR} = P(\hat{y} \neq y | z = 0, y = 1) - P(\hat{y} \neq y | z = 1, y = 1),$$

DCCP



$$\begin{array}{ll}\text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \\ & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c,\end{array}$$

By applying DCCP, training decision boundary-based classifiers would not suffer from disparate mistreatment.

Algorithm:

Steps:

1. Load the dataset and preprocess it.
2. Set the sensitive attribute and encode labels.
3. Standardize the features.
4. Split the dataset into training and testing sets.
5. Implement fairness constraints as part of the model training.
6. Train the model using these constraints.
7. Evaluate model performance on accuracy and fairness metrics like FPR and FNR for each Group.

Constraints and Considerations:

- The algorithm integrates fairness by binding the model to a constraint that the difference in error rates between groups is minimized.
- The threshold for acceptable disparity in error rates is not fixed; it can be adjusted depending on the level of fairness the situation demands.

Evaluation

Accuracy Evaluation

Overall Accuracy:

- The overall accuracy (`trainScore` and `testScore`) is calculated as the proportion of correct predictions in the training and test datasets, respectively. This is a standard metric for evaluating the performance of classification models.

Accuracy Difference Across Groups:

- `acc_difTrain` and `acc_difTest` measure the absolute difference in accuracy between two groups (defined by `Z_train` and `Z_test`) in the training and test datasets, respectively. This metric helps in understanding if the model's performance is consistent across different groups or if it is biased towards one group.

Evaluation

Fairness Evaluation

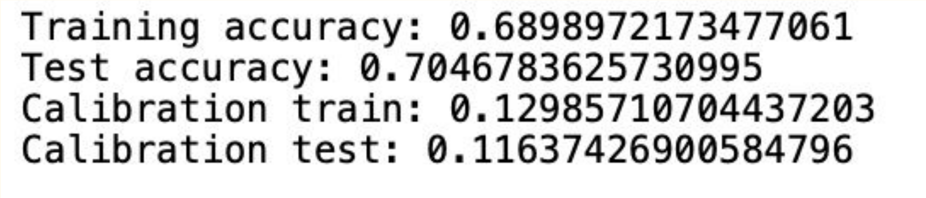
Group-specific Metrics:

- The function calculate_group_metrics computes metrics for each group separately in both the training and test datasets. These metrics might include group-specific accuracy, FPR, FNR, etc. This is crucial for assessing the fairness of the model.

Interpreting Fairness Metrics:

- The fairness of the model is evaluated by analyzing how these metrics differ between groups. Significant differences may indicate potential fairness issues.

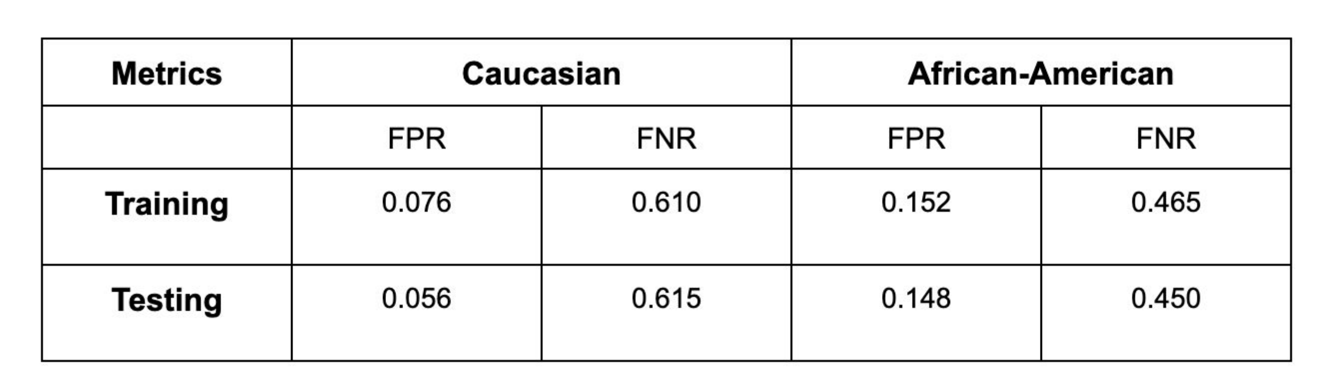
Results



```
Training accuracy: 0.6898972173477061
Test accuracy: 0.7046783625730995
Calibration train: 0.12985710704437203
Calibration test: 0.11637426900584796
```

The calibration train and test values indicate the difference in accuracy between the two groups (Caucasian and African-American) in the training and test datasets, respectively.

These values are essential for assessing the fairness of your model.

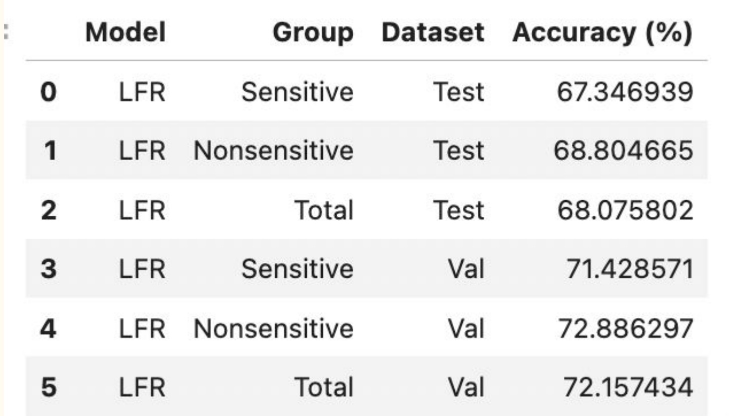


Metrics	**Caucasian**		**African-American**	
	FPR	FNR	FPR	FNR
Training	0.076	0.610	0.152	0.465
Testing	0.056	0.615	0.148	0.450

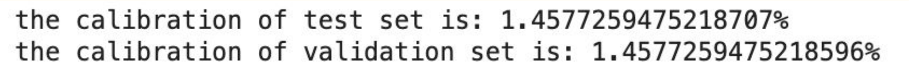
Comparing the Two

Comparison & Conclusion:

A1

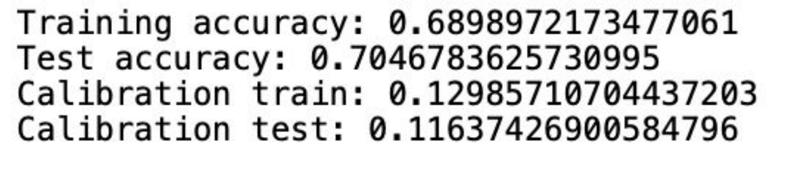


	Model	Group	Dataset	Accuracy (%)
0	LFR	Sensitive	Test	67.346939
1	LFR	Nonsensitive	Test	68.804665
2	LFR	Total	Test	68.075802
3	LFR	Sensitive	Val	71.428571
4	LFR	Nonsensitive	Val	72.886297
5	LFR	Total	Val	72.157434

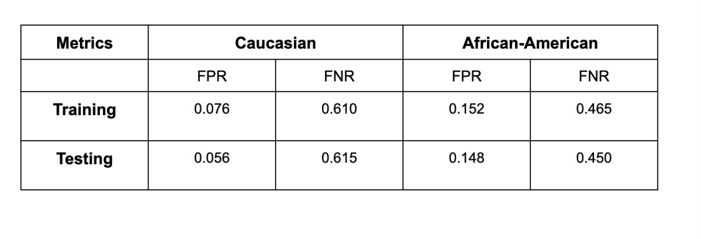


```
the calibration of test set is: 1.4577259475218707%
the calibration of validation set is: 1.4577259475218596%
```

A4



```
Training accuracy: 0.6898972173477061
Test accuracy: 0.7046783625730995
Calibration train: 0.12985710704437203
Calibration test: 0.11637426900584796
```



Metrics	Caucasian		African-American	
	FPR	FNR	FPR	FNR
Training	0.076	0.610	0.152	0.465
Testing	0.056	0.615	0.148	0.450

Thank You for Listening