Project 4 Group 5

Machine Learning Fairness

Arceneaux, Luke lpa2114@columbia.edu

Ren, Xiaoxue xr2159@columbia.edu

Wei, Jiahao jw4312@columbia.edu

Xia, Weijie wx2281@columbia.edu

Xu, Mingze mx2269@columbia.edu

Zhu, Yiming yz4336@columbia.edu

Models & Algorithms

- Baseline Models: Logistic & SVM
- A2: Maximizing accuracy under fairness constraints (C-SVM and C-LR)
- A6: Handling Conditional Discrimination (LM and LPS)

Data Cleaning and Wrangling

	two_year_recid	id	sex	race	age	age_cat	decile_score	score_text	c_charge_degree	is_recid	is_violent_recid	v_decile_score	v_score_text	priors_count	length_of_stay
1	1	3	Male	African-American	34	25 - 45	3	Low	F	1	1	1	Low	0	10.041667
2	1	4	Male	African-American	24	Less than 25	4	Low	F	1	0	3	Low	4	1.083333
6	1	8	Male	Caucasian	41	25 - 45	6	Medium	F	1	0	2	Low	14	6.291667
8	0	10	Female	Caucasian	39	25 - 45	1	Low	M	0	0	1	Low	0	2.916667
9	1	13	Male	Caucasian	21	Less than 25	3	Low	F	1	1	5	Medium	1	0.958333

	two_year_recid	sex	race	age_cat	c_charge_degree	v_score_text	score_text	priors_count	length_of_stay
0	1	1	0	1	1	0	0	-0.738411	-0.187151
1	1	1	0	0	1	0	0	0.045203	-0.356541
2	1	1	1	1	1	0	1	2.004240	-0.258059
3	0	0	1	1	0	0	0	-0.738411	-0.321875
4	1	1	1	0	1	1	0	-0.542508	-0.358905

```
# According to the instructure, the ratio should be 5:1:1
# It is roughly the same as 5:1:1
train_ratio = 0.7
val_ratio = 0.15
test_ratio = 0.15
```

Metric

- Calibrate_difference
  - calculates the calibration difference between two groups based on their sensitive features, predicted labels, and true labels.
  - absolute difference between the accuracy of the Caucasian group and the African American group.
  - Smaller 👉 better
- p_rule
  - The ratio of the lower percentage to the higher percentage, where the higher percentage is divided by the lower percentage if the African American group has a higher percentage of positive predictions, and vice versa if the Caucasian group has a higher percentage.
  - Closer to 1 👉 better

Baseline Models: Logistic & SVM

	Methods	Set	Accuracy (%)	Calibration(%)	p_rule
0	LR	Val	67.251462	1.973728	0.426691
1	LR	Test	65.730994	0.393226	0.544211
2	SVM	Val	66.783626	2.596923	0.491003
3	SVM	Test	65.847953	2.654994	0.634721

A2: Maximizing accuracy under fairness constraints (C-SVM and C-LR)

- minimizing the loss function subject to a covariance threshold between race (sensitive attribute) and the decision boundary
- C-LR
- C-SVM

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} \quad & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}, \end{aligned} \tag{6}$$

-Continue

	Methods	Set	Accuracy (%)	Calibration(%)	p_rule
0	CLR	Val	48.888889	11.713028	0.997245
1	CLR	Test	49.824561	8.791619	0.998936
2	CSVM	Val	47.485380	10.710206	0.970465
3	CSVM	Test	48.421053	7.956372	0.997649

A6: Handling Conditional Discrimination (LM and LPS)

- Implemented on the dataset to solve the problem of discrimination
  - Example of acceptance rate (CS & medical)
- On our data:
  - Feature: c_charge_degree = F or M
  - F: two_year_recid tend to be 1
  - M: two_year_recid tend to be 0
  - Goal: equal probability
    - $P'(+|ei,f) = P'(+|ei,m) = P^*(+|ei)$, where ei: feature value, f: protected, m:unprotected
- Discrimination is more frequent near the decision boundary

-Continue

- Local Massaging Sampling
  - Change the label if we consider the label is discriminated
- Local Preferential Sampling
  - Remove the ‘wrong’ instances that are close to the decision boundary
  - Duplicate the instances that are ‘right’ and close to the boundary

	Methods	Set	Accuracy (%)	Calibration(%)	p_rule
0	Local Massaging (LR)	Train	69.942341	0.849556	0.992513
1	Local Massaging (LR)	Test	64.561404	1.334673	0.337680
2	Local Preferential Sampling (LR)	Train	69.089997	0.354507	0.973000
3	Local Preferential Sampling (LR)	Test	65.614035	0.094719	0.571877

Summary

	Methods	Accuracy (%)	Calibration(%)	p_rule
0	Baseline LR	65.730994	0.393226	0.544211
1	Baseline SVM	65.847953	2.654994	0.634721
2	CLR	49.824561	8.791619	0.998936
3	CSVM	48.421053	7.956372	0.997649
4	Local Massaging (LR)	64.561404	1.334673	0.337680
5	Local Preferential Sampling (LR)	65.614035	0.094719	0.571877