Assignment 3.1 Exercise

In this assignment, you will explore the impact of fairness metrics after applying the Disparate Impact Remover (DIR), retrain a Light GBM model with the adjusted data, and compare its metrics (listed below) against the original and reweighted models from the lab tutorial. Afterward, respond to the reflection questions.

Part 1: Experimentation with the DIR

- Modify the levels array for the Disparate Impact Remover to include finer and coarser granularity. (See example in lab/assignment notebook.)
- Identify the best repair level and the Disparate Impact at the best repair level.
- Retrain a Light GBM model on the modified data from the DIR and compute both the performance and fairness metrics.

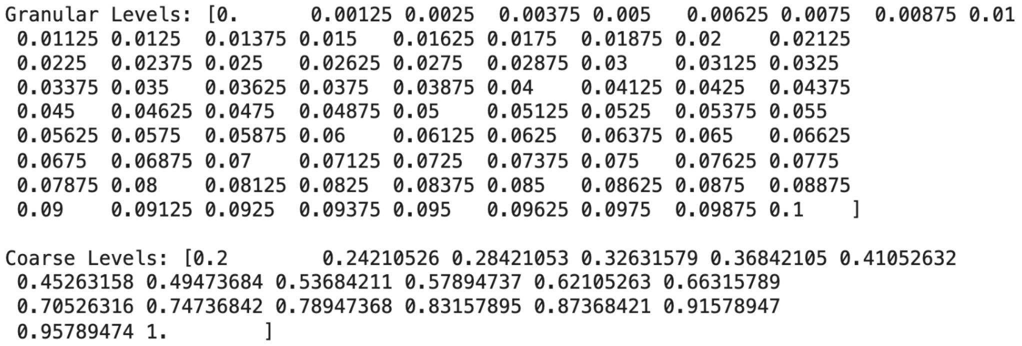
1.1a Code Output for Finer Granularity

Answer

The output is organized as below:

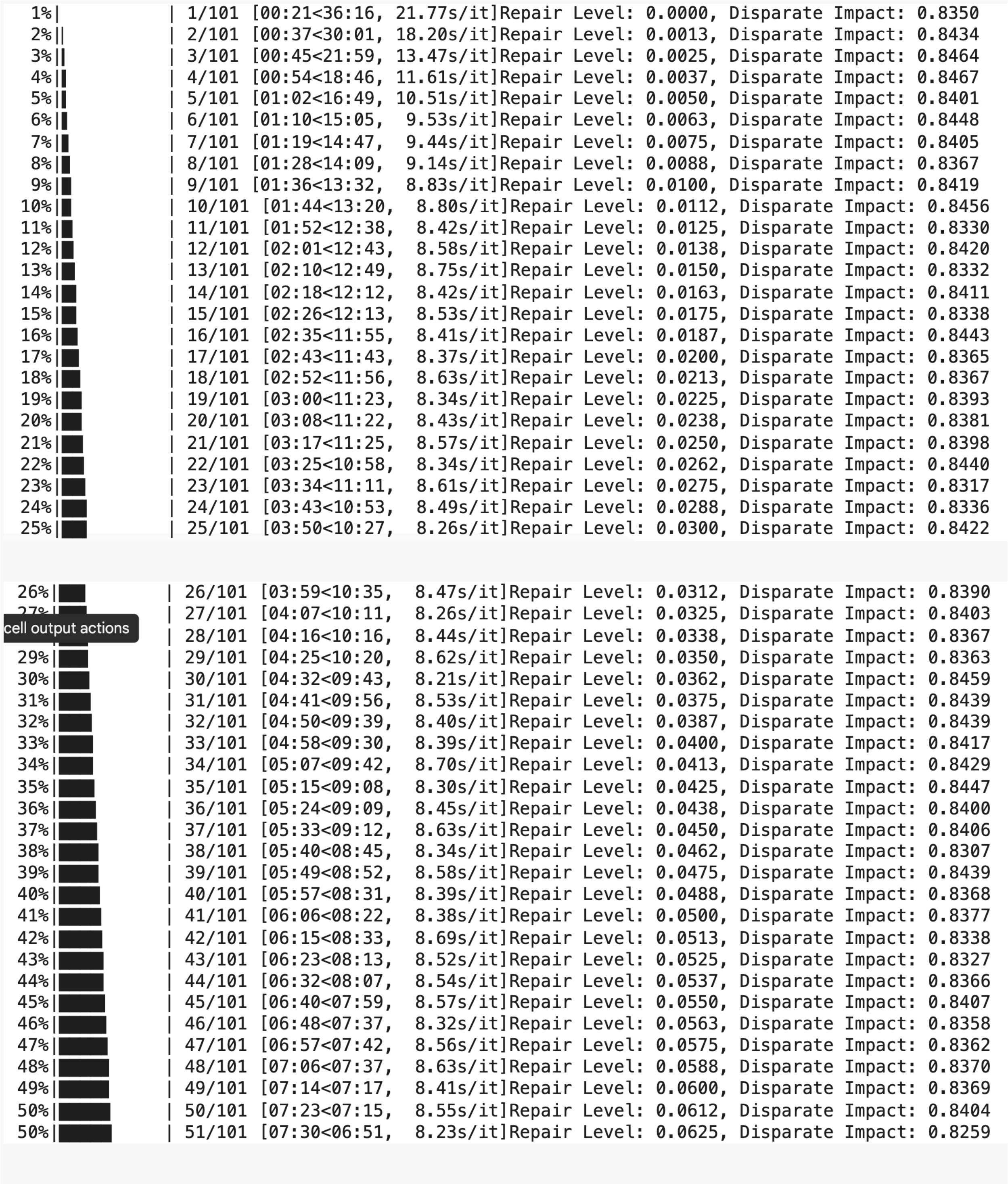
- **Finer Granular and Coarser Levels array Information**
- **Combined Finer Granular and Coarse Disparate Impact Results for different levels**
- **Combined Finer Granular and Coarse Disparate Impact Plots**
- **Finer Granular Disparate Impact Plot**
- **Finer Granular Disparate Impact Results**

Finer Granular and Coarser Levels used for Disparate Impact Remover (DIR) are provided below



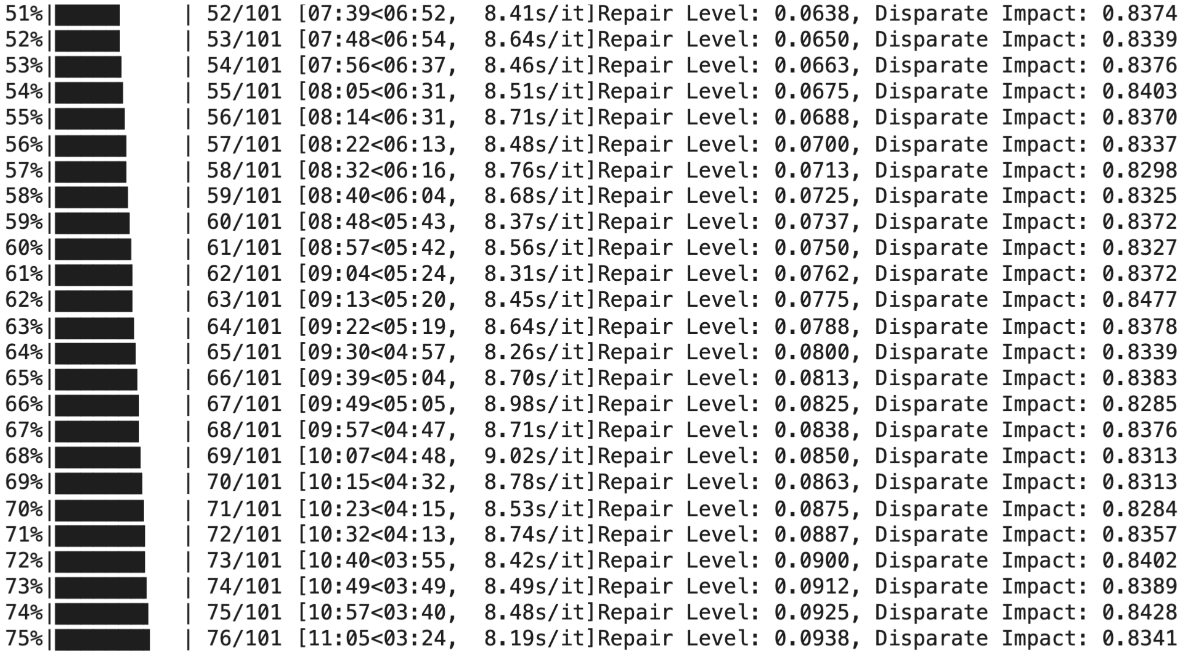
```
Granular Levels: [0.      0.00125 0.0025  0.00375 0.005   0.00625 0.0075  0.00875 0.01
 0.01125 0.0125  0.01375 0.015   0.01625 0.0175  0.01875 0.02    0.02125
 0.0225  0.02375 0.025   0.02625 0.0275  0.02875 0.03    0.03125 0.0325
 0.03375 0.035   0.03625 0.0375  0.03875 0.04    0.04125 0.0425  0.04375
 0.045   0.04625 0.0475  0.04875 0.05    0.05125 0.0525  0.05375 0.055
 0.05625 0.0575  0.05875 0.06    0.06125 0.0625  0.06375 0.065   0.06625
 0.0675  0.06875 0.07    0.07125 0.0725  0.07375 0.075   0.07625 0.0775
 0.07875 0.08    0.08125 0.0825  0.08375 0.085   0.08625 0.0875  0.08875
 0.09    0.09125 0.0925  0.09375 0.095   0.09625 0.0975  0.09875 0.1    ]

Coarse Levels: [0.2        0.24210526 0.28421053 0.32631579 0.36842105 0.41052632
 0.45263158 0.49473684 0.53684211 0.57894737 0.62105263 0.66315789
 0.70526316 0.74736842 0.78947368 0.83157895 0.87368421 0.91578947
 0.95789474 1.        ]
```

```
DIR at different Repair levels are provided below
```

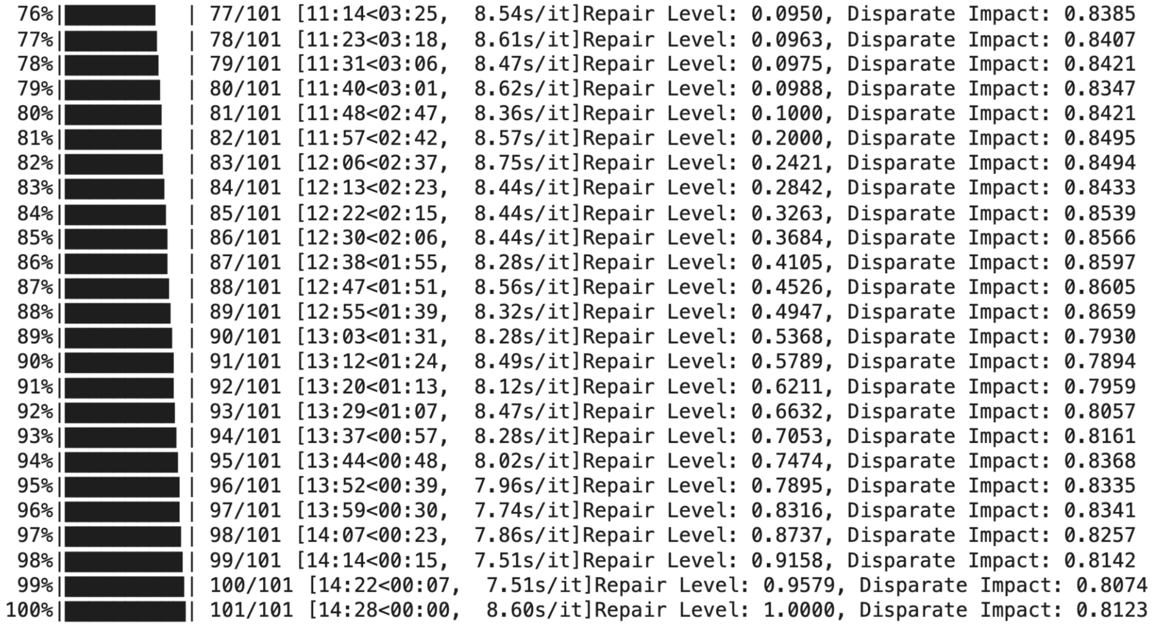


```
  1%|          | 1/101 [00:21<36:16, 21.77s/it]Repair Level: 0.0000, Disparate Impact: 0.8350
  2%|▏         | 2/101 [00:37<30:01, 18.20s/it]Repair Level: 0.0013, Disparate Impact: 0.8434
  3%|▎         | 3/101 [00:45<21:59, 13.47s/it]Repair Level: 0.0025, Disparate Impact: 0.8464
  4%|▍         | 4/101 [00:54<18:46, 11.61s/it]Repair Level: 0.0037, Disparate Impact: 0.8467
  5%|▍         | 5/101 [01:02<16:49, 10.51s/it]Repair Level: 0.0050, Disparate Impact: 0.8401
  6%|▌         | 6/101 [01:10<15:05,  9.53s/it]Repair Level: 0.0063, Disparate Impact: 0.8448
  7%|▋         | 7/101 [01:19<14:47,  9.44s/it]Repair Level: 0.0075, Disparate Impact: 0.8405
  8%|▊         | 8/101 [01:28<14:09,  9.14s/it]Repair Level: 0.0088, Disparate Impact: 0.8367
  9%|▉         | 9/101 [01:36<13:32,  8.83s/it]Repair Level: 0.0100, Disparate Impact: 0.8419
 10%|▉         | 10/101 [01:44<13:20,  8.80s/it]Repair Level: 0.0112, Disparate Impact: 0.8456
 11%|█         | 11/101 [01:52<12:38,  8.42s/it]Repair Level: 0.0125, Disparate Impact: 0.8330
 12%|█▏        | 12/101 [02:01<12:43,  8.58s/it]Repair Level: 0.0138, Disparate Impact: 0.8420
 13%|█▎        | 13/101 [02:10<12:49,  8.75s/it]Repair Level: 0.0150, Disparate Impact: 0.8332
 14%|█▍        | 14/101 [02:18<12:12,  8.42s/it]Repair Level: 0.0163, Disparate Impact: 0.8411
 15%|█▍        | 15/101 [02:26<12:13,  8.53s/it]Repair Level: 0.0175, Disparate Impact: 0.8338
 16%|█▌        | 16/101 [02:35<11:55,  8.41s/it]Repair Level: 0.0187, Disparate Impact: 0.8443
 17%|█▋        | 17/101 [02:43<11:43,  8.37s/it]Repair Level: 0.0200, Disparate Impact: 0.8365
 18%|█▊        | 18/101 [02:52<11:56,  8.63s/it]Repair Level: 0.0213, Disparate Impact: 0.8367
 19%|█▉        | 19/101 [03:00<11:23,  8.34s/it]Repair Level: 0.0225, Disparate Impact: 0.8393
 20%|█▉        | 20/101 [03:08<11:22,  8.43s/it]Repair Level: 0.0238, Disparate Impact: 0.8381
 21%|██        | 21/101 [03:17<11:25,  8.57s/it]Repair Level: 0.0250, Disparate Impact: 0.8398
 22%|██▏       | 22/101 [03:25<10:58,  8.34s/it]Repair Level: 0.0262, Disparate Impact: 0.8440
 23%|██▎       | 23/101 [03:34<11:11,  8.61s/it]Repair Level: 0.0275, Disparate Impact: 0.8317
 24%|██▍       | 24/101 [03:43<10:53,  8.49s/it]Repair Level: 0.0288, Disparate Impact: 0.8336
 25%|██▍       | 25/101 [03:50<10:27,  8.26s/it]Repair Level: 0.0300, Disparate Impact: 0.8422
```

```
 26%|██▌       | 26/101 [03:59<10:35,  8.47s/it]Repair Level: 0.0312, Disparate Impact: 0.8390
 27%|██▋       | 27/101 [04:07<10:11,  8.26s/it]Repair Level: 0.0325, Disparate Impact: 0.8403
                | 28/101 [04:16<10:16,  8.44s/it]Repair Level: 0.0338, Disparate Impact: 0.8367
 29%|██▊       | 29/101 [04:25<10:20,  8.62s/it]Repair Level: 0.0350, Disparate Impact: 0.8363
 30%|██▉       | 30/101 [04:32<09:43,  8.21s/it]Repair Level: 0.0362, Disparate Impact: 0.8459
 31%|███       | 31/101 [04:41<09:56,  8.53s/it]Repair Level: 0.0375, Disparate Impact: 0.8439
 32%|███▏      | 32/101 [04:50<09:39,  8.40s/it]Repair Level: 0.0387, Disparate Impact: 0.8439
 33%|███▎      | 33/101 [04:58<09:30,  8.39s/it]Repair Level: 0.0400, Disparate Impact: 0.8417
 34%|███▎      | 34/101 [05:07<09:42,  8.70s/it]Repair Level: 0.0413, Disparate Impact: 0.8429
 35%|███▍      | 35/101 [05:15<09:08,  8.30s/it]Repair Level: 0.0425, Disparate Impact: 0.8447
 36%|███▌      | 36/101 [05:24<09:09,  8.45s/it]Repair Level: 0.0438, Disparate Impact: 0.8400
 37%|███▋      | 37/101 [05:33<09:12,  8.63s/it]Repair Level: 0.0450, Disparate Impact: 0.8406
 38%|███▊      | 38/101 [05:40<08:45,  8.34s/it]Repair Level: 0.0462, Disparate Impact: 0.8307
 39%|███▊      | 39/101 [05:49<08:52,  8.58s/it]Repair Level: 0.0475, Disparate Impact: 0.8439
 40%|███▉      | 40/101 [05:57<08:31,  8.39s/it]Repair Level: 0.0488, Disparate Impact: 0.8368
 41%|████      | 41/101 [06:06<08:22,  8.38s/it]Repair Level: 0.0500, Disparate Impact: 0.8377
 42%|████▏     | 42/101 [06:15<08:33,  8.69s/it]Repair Level: 0.0513, Disparate Impact: 0.8338
 43%|████▎     | 43/101 [06:23<08:13,  8.52s/it]Repair Level: 0.0525, Disparate Impact: 0.8327
 44%|████▎     | 44/101 [06:32<08:07,  8.54s/it]Repair Level: 0.0537, Disparate Impact: 0.8366
 45%|████▍     | 45/101 [06:40<07:59,  8.57s/it]Repair Level: 0.0550, Disparate Impact: 0.8407
 46%|████▌     | 46/101 [06:48<07:37,  8.32s/it]Repair Level: 0.0563, Disparate Impact: 0.8358
 47%|████▋     | 47/101 [06:57<07:42,  8.56s/it]Repair Level: 0.0575, Disparate Impact: 0.8362
 48%|████▊     | 48/101 [07:06<07:37,  8.63s/it]Repair Level: 0.0588, Disparate Impact: 0.8370
 49%|████▊     | 49/101 [07:14<07:17,  8.41s/it]Repair Level: 0.0600, Disparate Impact: 0.8369
 50%|████▉     | 50/101 [07:23<07:15,  8.55s/it]Repair Level: 0.0612, Disparate Impact: 0.8404
 50%|█████     | 51/101 [07:30<06:51,  8.23s/it]Repair Level: 0.0625, Disparate Impact: 0.8259
```



```
 51%|█████▏    | 52/101 [07:39<06:52,  8.41s/it]Repair Level: 0.0638, Disparate Impact: 0.8374
 52%|█████▏    | 53/101 [07:48<06:54,  8.64s/it]Repair Level: 0.0650, Disparate Impact: 0.8339
 53%|█████▎    | 54/101 [07:56<06:37,  8.46s/it]Repair Level: 0.0663, Disparate Impact: 0.8376
 54%|█████▍    | 55/101 [08:05<06:31,  8.51s/it]Repair Level: 0.0675, Disparate Impact: 0.8403
 55%|█████▌    | 56/101 [08:14<06:31,  8.71s/it]Repair Level: 0.0688, Disparate Impact: 0.8370
 56%|█████▋    | 57/101 [08:22<06:13,  8.48s/it]Repair Level: 0.0700, Disparate Impact: 0.8337
 57%|█████▋    | 58/101 [08:32<06:16,  8.76s/it]Repair Level: 0.0713, Disparate Impact: 0.8298
 58%|█████▊    | 59/101 [08:40<06:04,  8.68s/it]Repair Level: 0.0725, Disparate Impact: 0.8325
 59%|█████▉    | 60/101 [08:48<05:43,  8.37s/it]Repair Level: 0.0737, Disparate Impact: 0.8372
 60%|██████    | 61/101 [08:57<05:42,  8.56s/it]Repair Level: 0.0750, Disparate Impact: 0.8327
 61%|██████▏   | 62/101 [09:04<05:24,  8.31s/it]Repair Level: 0.0762, Disparate Impact: 0.8372
 62%|██████▏   | 63/101 [09:13<05:20,  8.45s/it]Repair Level: 0.0775, Disparate Impact: 0.8477
 63%|██████▎   | 64/101 [09:22<05:19,  8.64s/it]Repair Level: 0.0788, Disparate Impact: 0.8378
 64%|██████▍   | 65/101 [09:30<04:57,  8.26s/it]Repair Level: 0.0800, Disparate Impact: 0.8339
 65%|██████▌   | 66/101 [09:39<05:04,  8.70s/it]Repair Level: 0.0813, Disparate Impact: 0.8383
 66%|██████▋   | 67/101 [09:49<05:05,  8.98s/it]Repair Level: 0.0825, Disparate Impact: 0.8285
 67%|██████▋   | 68/101 [09:57<04:47,  8.71s/it]Repair Level: 0.0838, Disparate Impact: 0.8376
 68%|██████▊   | 69/101 [10:07<04:48,  9.02s/it]Repair Level: 0.0850, Disparate Impact: 0.8313
 69%|██████▉   | 70/101 [10:15<04:32,  8.78s/it]Repair Level: 0.0863, Disparate Impact: 0.8313
 70%|███████   | 71/101 [10:23<04:15,  8.53s/it]Repair Level: 0.0875, Disparate Impact: 0.8284
 71%|███████▏  | 72/101 [10:32<04:13,  8.74s/it]Repair Level: 0.0887, Disparate Impact: 0.8357
 72%|███████▏  | 73/101 [10:40<03:55,  8.42s/it]Repair Level: 0.0900, Disparate Impact: 0.8402
 73%|███████▎  | 74/101 [10:49<03:49,  8.49s/it]Repair Level: 0.0912, Disparate Impact: 0.8389
 74%|███████▍  | 75/101 [10:57<03:40,  8.48s/it]Repair Level: 0.0925, Disparate Impact: 0.8428
 75%|███████▌  | 76/101 [11:05<03:24,  8.19s/it]Repair Level: 0.0938, Disparate Impact: 0.8341
```

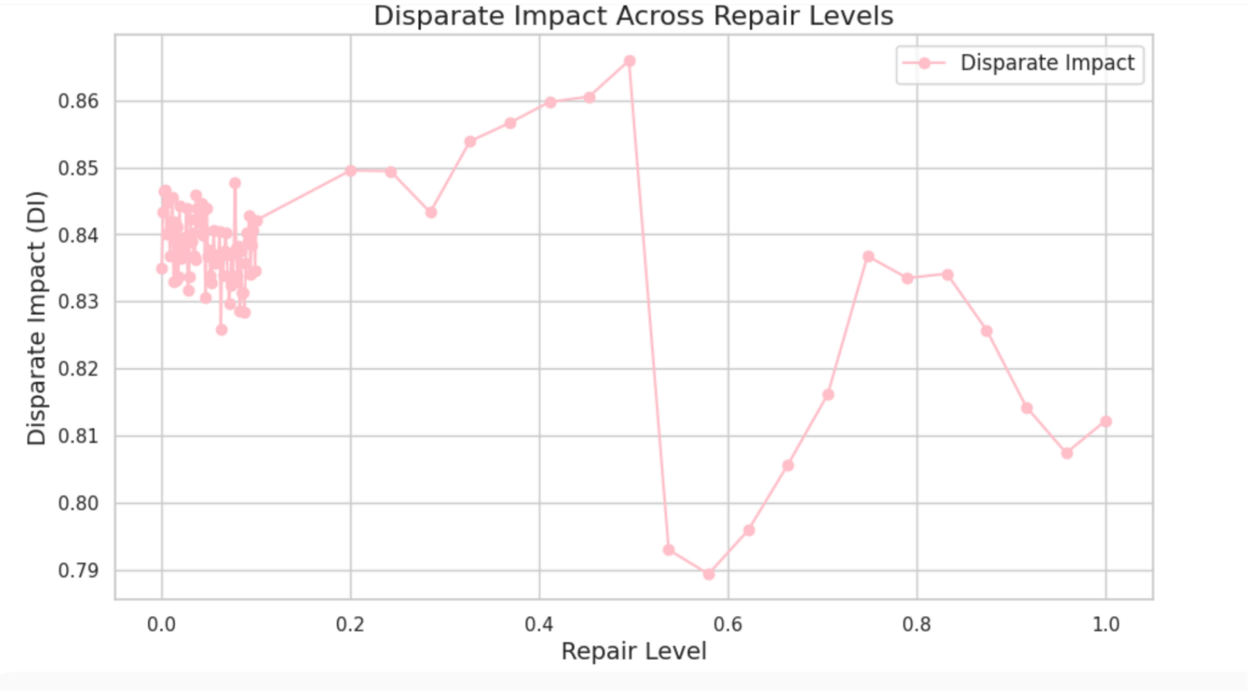


```
 76%|███████▌  | 77/101 [11:14<03:25,  8.54s/it]Repair Level: 0.0950, Disparate Impact: 0.8385
 77%|███████▋  | 78/101 [11:23<03:18,  8.61s/it]Repair Level: 0.0963, Disparate Impact: 0.8407
 78%|███████▊  | 79/101 [11:31<03:06,  8.47s/it]Repair Level: 0.0975, Disparate Impact: 0.8421
 79%|███████▉  | 80/101 [11:40<03:01,  8.62s/it]Repair Level: 0.0988, Disparate Impact: 0.8347
 80%|████████  | 81/101 [11:48<02:47,  8.36s/it]Repair Level: 0.1000, Disparate Impact: 0.8421
 81%|████████  | 82/101 [11:57<02:42,  8.57s/it]Repair Level: 0.2000, Disparate Impact: 0.8495
 82%|████████▏ | 83/101 [12:06<02:37,  8.75s/it]Repair Level: 0.2421, Disparate Impact: 0.8494
 83%|████████▎ | 84/101 [12:13<02:23,  8.44s/it]Repair Level: 0.2842, Disparate Impact: 0.8433
 84%|████████▍ | 85/101 [12:22<02:15,  8.44s/it]Repair Level: 0.3263, Disparate Impact: 0.8539
 85%|████████▌ | 86/101 [12:30<02:06,  8.44s/it]Repair Level: 0.3684, Disparate Impact: 0.8566
 86%|████████▌ | 87/101 [12:38<01:55,  8.28s/it]Repair Level: 0.4105, Disparate Impact: 0.8597
 87%|████████▋ | 88/101 [12:47<01:51,  8.56s/it]Repair Level: 0.4526, Disparate Impact: 0.8605
 88%|████████▊ | 89/101 [12:55<01:39,  8.32s/it]Repair Level: 0.4947, Disparate Impact: 0.8659
 89%|████████▉ | 90/101 [13:03<01:31,  8.28s/it]Repair Level: 0.5368, Disparate Impact: 0.7930
 90%|█████████ | 91/101 [13:12<01:24,  8.49s/it]Repair Level: 0.5789, Disparate Impact: 0.7894
 91%|█████████ | 92/101 [13:20<01:13,  8.12s/it]Repair Level: 0.6211, Disparate Impact: 0.7959
 92%|█████████▏| 93/101 [13:29<01:07,  8.47s/it]Repair Level: 0.6632, Disparate Impact: 0.8057
 93%|█████████▎| 94/101 [13:37<00:57,  8.28s/it]Repair Level: 0.7053, Disparate Impact: 0.8161
 94%|█████████▍| 95/101 [13:44<00:48,  8.02s/it]Repair Level: 0.7474, Disparate Impact: 0.8368
 95%|█████████▌| 96/101 [13:52<00:39,  7.96s/it]Repair Level: 0.7895, Disparate Impact: 0.8335
 96%|█████████▌| 97/101 [13:59<00:30,  7.74s/it]Repair Level: 0.8316, Disparate Impact: 0.8341
 97%|█████████▋| 98/101 [14:07<00:23,  7.86s/it]Repair Level: 0.8737, Disparate Impact: 0.8257
 98%|█████████▊| 99/101 [14:14<00:15,  7.51s/it]Repair Level: 0.9158, Disparate Impact: 0.8142
 99%|█████████▉| 100/101 [14:22<00:07,  7.51s/it]Repair Level: 0.9579, Disparate Impact: 0.8074
100%|██████████| 101/101 [14:28<00:00,  8.60s/it]Repair Level: 1.0000, Disparate Impact: 0.8123
```

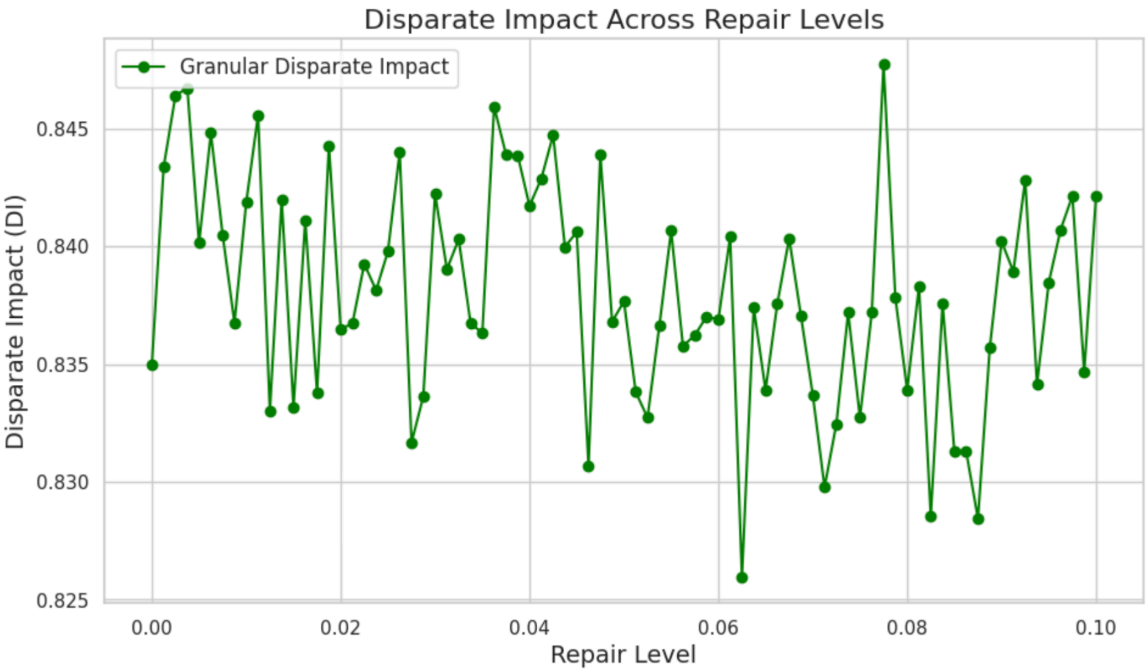
Results Interpretation

The results show that small fairness adjustments (Granular Levels) have minimal impact on Disparate Impact (DI), while moderate repair levels (0.3 - 0.5) improve fairness significantly. However, excessive fairness repair (>0.5) reduces DI, suggesting diminishing returns and potential performance degradation. The optimal fairness correction lies within moderate repair levels, balancing bias mitigation without harming accuracy.

Combined Finer Granular and Coarse Disparate Impact Plots



Finer Granular Disparate Impact Results



```
Repair Level: 0.0000, Disparate Impact: 0.8350
Repair Level: 0.0013, Disparate Impact: 0.8434
Repair Level: 0.0025, Disparate Impact: 0.8464
Repair Level: 0.0037, Disparate Impact: 0.8467
Repair Level: 0.0050, Disparate Impact: 0.8401
Repair Level: 0.0063, Disparate Impact: 0.8448
Repair Level: 0.0075, Disparate Impact: 0.8405
Repair Level: 0.0088, Disparate Impact: 0.8367
Repair Level: 0.0100, Disparate Impact: 0.8419
Repair Level: 0.0112, Disparate Impact: 0.8456
Repair Level: 0.0125, Disparate Impact: 0.8330
Repair Level: 0.0138, Disparate Impact: 0.8420
Repair Level: 0.0150, Disparate Impact: 0.8332
Repair Level: 0.0163, Disparate Impact: 0.8411
Repair Level: 0.0175, Disparate Impact: 0.8338
Repair Level: 0.0187, Disparate Impact: 0.8443
Repair Level: 0.0200, Disparate Impact: 0.8365
Repair Level: 0.0213, Disparate Impact: 0.8367

Repair Level: 0.0225, Disparate Impact: 0.8393

Repair Level: 0.0238, Disparate Impact: 0.8381

Repair Level: 0.0250, Disparate Impact: 0.8398

Repair Level: 0.0262, Disparate Impact: 0.8440

Repair Level: 0.0275, Disparate Impact: 0.8317

Repair Level: 0.0288, Disparate Impact: 0.8336

Repair Level: 0.0300, Disparate Impact: 0.8422

Repair Level: 0.0312, Disparate Impact: 0.8390

Repair Level: 0.0325, Disparate Impact: 0.8403

Repair Level: 0.0338, Disparate Impact: 0.8367

Repair Level: 0.0350, Disparate Impact: 0.8363

Repair Level: 0.0362, Disparate Impact: 0.8459

Repair Level: 0.0375, Disparate Impact: 0.8439

Repair Level: 0.0387, Disparate Impact: 0.8439

Repair Level: 0.0400, Disparate Impact: 0.8417

Repair Level: 0.0413, Disparate Impact: 0.8429

Repair Level: 0.0425, Disparate Impact: 0.8447

Repair Level: 0.0438, Disparate Impact: 0.8400

Repair Level: 0.0450, Disparate Impact: 0.8406

Repair Level: 0.0462, Disparate Impact: 0.8307

Repair Level: 0.0475, Disparate Impact: 0.8439

Repair Level: 0.0488, Disparate Impact: 0.8368

Repair Level: 0.0500, Disparate Impact: 0.8377

Repair Level: 0.0513, Disparate Impact: 0.8338

Repair Level: 0.0525, Disparate Impact: 0.8327

Repair Level: 0.0537, Disparate Impact: 0.8366

Repair Level: 0.0550, Disparate Impact: 0.8407

Repair Level: 0.0563, Disparate Impact: 0.8358

Repair Level: 0.0575, Disparate Impact: 0.8362
Repair Level: 0.0588, Disparate Impact: 0.8370

Repair Level: 0.0600, Disparate Impact: 0.8369

Repair Level: 0.0612, Disparate Impact: 0.8404

Repair Level: 0.0625, Disparate Impact: 0.8259

Repair Level: 0.0638, Disparate Impact: 0.8374

Repair Level: 0.0650, Disparate Impact: 0.8339

Repair Level: 0.0663, Disparate Impact: 0.8376

Repair Level: 0.0675, Disparate Impact: 0.8403

Repair Level: 0.0688, Disparate Impact: 0.8370

Repair Level: 0.0700, Disparate Impact: 0.8337

Repair Level: 0.0713, Disparate Impact: 0.8298

Repair Level: 0.0725, Disparate Impact: 0.8325

Repair Level: 0.0737, Disparate Impact: 0.8372

Repair Level: 0.0750, Disparate Impact: 0.8327

Repair Level: 0.0762, Disparate Impact: 0.8372

Repair Level: 0.0775, Disparate Impact: 0.8477

Repair Level: 0.0788, Disparate Impact: 0.8378

Repair Level: 0.0800, Disparate Impact: 0.8339

Repair Level: 0.0813, Disparate Impact: 0.8383

Repair Level: 0.0825, Disparate Impact: 0.8285

Repair Level: 0.0838, Disparate Impact: 0.8376

Repair Level: 0.0850, Disparate Impact: 0.8313

Repair Level: 0.0863, Disparate Impact: 0.8313

Repair Level: 0.0875, Disparate Impact: 0.8284

Repair Level: 0.0887, Disparate Impact: 0.8357

Repair Level: 0.0900, Disparate Impact: 0.8402

Repair Level: 0.0912, Disparate Impact: 0.8389

Repair Level: 0.0925, Disparate Impact: 0.8428

Repair Level: 0.0938, Disparate Impact: 0.8341

Repair Level: 0.0950, Disparate Impact: 0.8385
Repair Level: 0.0963, Disparate Impact: 0.8407

Repair Level: 0.0975, Disparate Impact: 0.8421

Repair Level: 0.0988, Disparate Impact: 0.8347

Repair Level: 0.1000, Disparate Impact: 0.8421
```

Result Interpretation of Finer Granular Changes

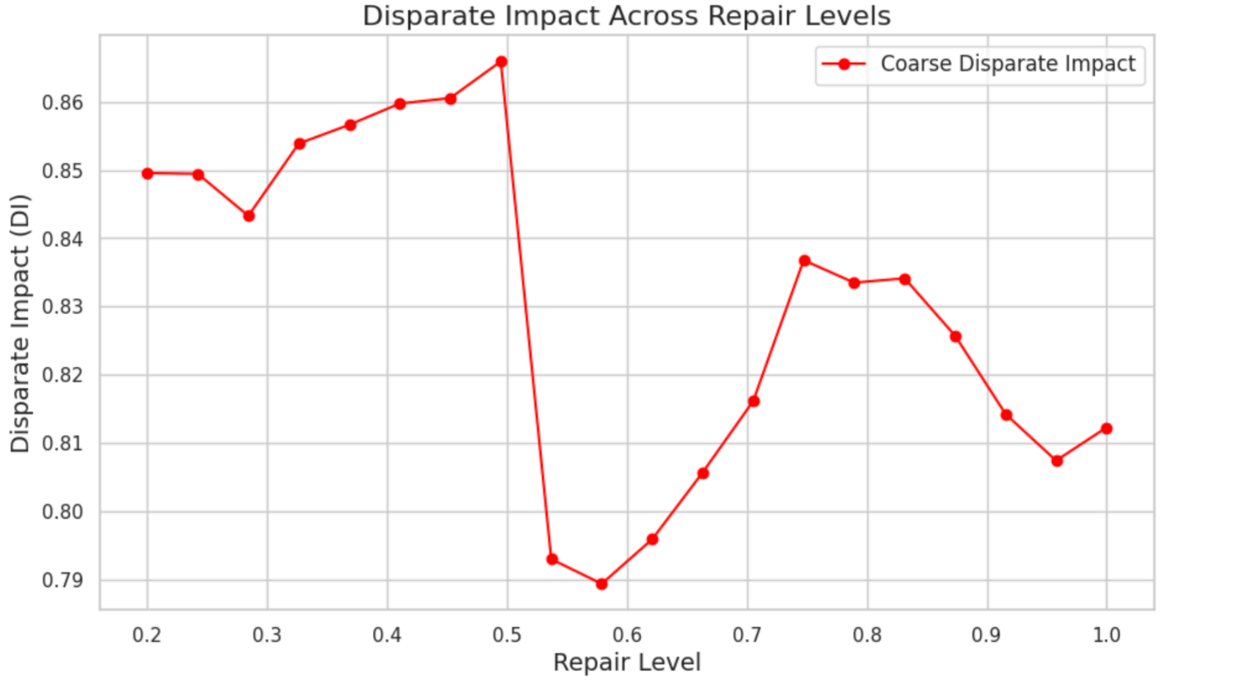
DI fluctuates significantly, indicating that small changes in fairness repair can lead to inconsistent effects. There is no clear increasing or decreasing trend, and some repair levels improve DI while others reduce it. The DI values stay within a narrow range (0.825 - 0.846), suggesting that fine-grained adjustments may not always yield stable fairness improvement**s**.

1.1b. Code Output for Coarser Granularity

Answer:

The output is organized as below:

- **Coarse Disparate Impact Plots**
- **Coarse Disparate Impact Results**



Coarse Disparate Impact Results

```
Repair Level: 0.2000, Disparate Impact: 0.8495
Repair Level: 0.2421, Disparate Impact: 0.8494

Repair Level: 0.2842, Disparate Impact: 0.8433

Repair Level: 0.3263, Disparate Impact: 0.8539

Repair Level: 0.3684, Disparate Impact: 0.8566

Repair Level: 0.4105, Disparate Impact: 0.8597

Repair Level: 0.4526, Disparate Impact: 0.8605

Repair Level: 0.4947, Disparate Impact: 0.8659

Repair Level: 0.5368, Disparate Impact: 0.7930

Repair Level: 0.5789, Disparate Impact: 0.7894

Repair Level: 0.6211, Disparate Impact: 0.7959

Repair Level: 0.6632, Disparate Impact: 0.8057

Repair Level: 0.7053, Disparate Impact: 0.8161

Repair Level: 0.7474, Disparate Impact: 0.8368

Repair Level: 0.7895, Disparate Impact: 0.8335

Repair Level: 0.8316, Disparate Impact: 0.8341

Repair Level: 0.8737, Disparate Impact: 0.8257

Repair Level: 0.9158, Disparate Impact: 0.8142

Repair Level: 0.9579, Disparate Impact: 0.8074

Repair Level: 1.0000, Disparate Impact: 0.8123
```

Coarse Result Interpretation

Initially, DI remains relatively stable but then increases slightly before experiencing a sharp **drop around the 0.5 repair level**. After this decline, DI gradually recovers but does not fully regain its earlier peak. This suggests that **moderate repair levels improve fairness**, but excessive adjustments (beyond a certain threshold) can have unintended negative effects on fairness. The fluctuations indicate that the **relationship between repair levels and fairness is non-linear**, requiring careful tuning to balance both fairness and performance.

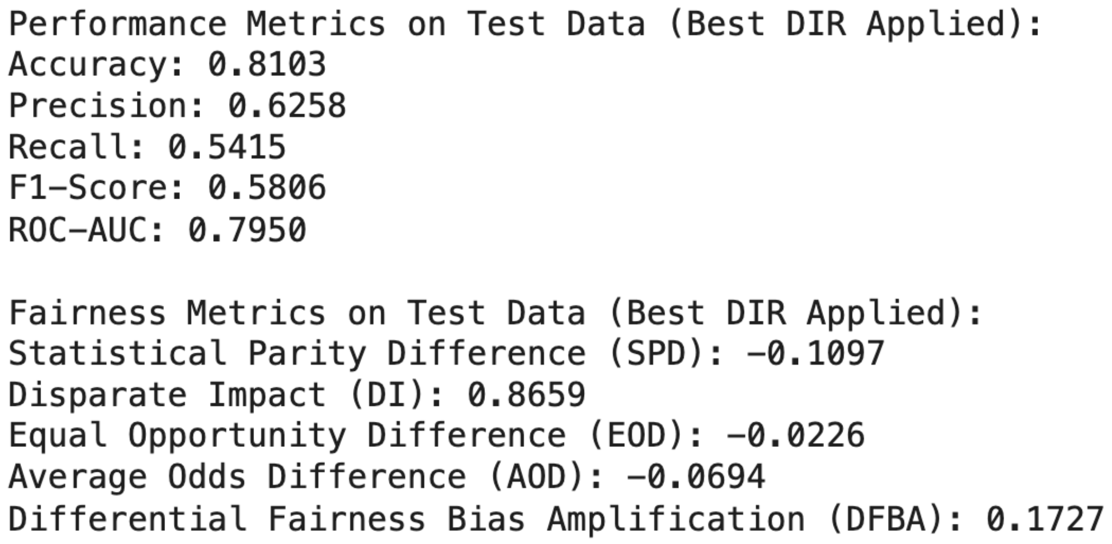
1.2. Code Output for Best Repair Level and Disparate Impact at Best Repair Level

Paste your screenshot here. (Choose the single best out of all repair levels for finer/coarser. This will be the DI closest to 1.0.)

Answer

```
Best Repair Level: 0.4947

Disparate Impact at Best Repair Level: 0.8659
```

1.3. Code Output for Light GBM Performance Metrics and Fairness Metrics



```
Performance Metrics on Test Data (Best DIR Applied):
Accuracy: 0.8103
Precision: 0.6258
Recall: 0.5415
F1-Score: 0.5806
ROC-AUC: 0.7950

Fairness Metrics on Test Data (Best DIR Applied):
Statistical Parity Difference (SPD): -0.1097
Disparate Impact (DI): 0.8659
Equal Opportunity Difference (EOD): -0.0226
Average Odds Difference (AOD): -0.0694
Differential Fairness Bias Amplification (DFBA): 0.1727
```

Part 2: Short-Response Questions

Question:

2.1. How does the Disparate Impact (DI) change with finer granularity? Coarser granularity?

Answer:

Disparate Impact (DI) measures whether an AI model disproportionately favors or disadvantages certain groups. It is calculated as:DI = Selection Rate of Unprivileged Group / Selection Rate of Privileged GroupA DI score closer to 1.0 indicates fairness, while a lower DI suggests bias.

Effect of Finer Granularity on DI:

Finer granularity breaks groups into smaller subcategories for a more detailed assessment.

Example: A bank categorizes borrowers into Male (10% default rate) and Female (15% default rate), resulting in DI = 0.67. If age is added as a factor:

Males 18-30: 12%, 31-50: 9%, 51+: 7%

Females 18-30: 18%, 31-50: 14%, 51+: 11%

Now, DI varies by age, revealing hidden biases. Younger females (DI = 0.67) face more discrimination than older females (DI = 0.78).

Effect of Coarser Granularity on DI:

Coarser granularity merges subgroups, making DI appear more stable but potentially hiding disparities.

Example: If all females are grouped together (15% default rate), age-based bias is masked, and DI remains 0.67 without showing variation.

Conclusion

Finer Granularity → **More precise DI but higher fluctuation.**

Coarser Granularity → **Smoother DI but may hide subgroup biases.**

Balanced Approach → **Detects bias while minimizing noise.**

Question:

2.2. Based on your findings, which model would you recommend (Original, Reweighted, or DIR)? Justify your choice by balancing performance and fairness.

Answer:

The Reweighted Model is the best recommendation because: It achieves the best fairness score (DI closest to 1.0). It still maintains good performance without a major drop in accuracy or precision. It balances both fairness and model effectiveness better than the Original and DIR models.

The results for the all 3 models (Original, Reweighted and DIR) are provided below:

--- Original Model ---

```
Performance Metrics:
  accuracy: 0.8153
  precision: 0.6400
  recall: 0.5454
  f1: 0.5889
  roc_auc: 0.8003
Fairness Metrics:
  SPD: -0.0862
  DI: 0.8972
  EOD: 0.0164
  AOD: -0.0163
  Fairness Score (DI deviation): 0.1028

--- Reweighted Model ---
Performance Metrics:
  accuracy: 0.8111
  precision: 0.6275
  recall: 0.5437
  f1: 0.5826
  roc_auc: 0.7960
Fairness Metrics:
  SPD: -0.0351
  DI: 0.9565
  EOD: 0.0164
  AOD: -0.0163
  Fairness Score (DI deviation): 0.0435
--- DIR Model ---
Performance Metrics:
  accuracy: 0.8103
  precision: 0.6258
  recall: 0.5415
  f1: 0.5806
  roc_auc: 0.7950
Fairness Metrics:
  SPD: -0.1097
  DI: 0.8659
  EOD: -0.0226
  AOD: -0.0694
  Fairness Score (DI deviation): 0.1341

Recommended Model: Reweighted Model
```

Question:

2.3. Consider a scenario where you are responsible for deploying a machine learning model that shows a fairness-accuracy trade-off.

The following reflection explores the ethical dilemmas in machine learning, emphasizing the trade-off between fairness and accuracy. Key considerations include organizational values, societal impact, and the specific application context.

Deploying machine learning models that strike a balance between fairness and accuracy requires a principled approach aligned with ethical standards, social responsibility, and business objectives. Organizations must determine their priorities: if efficiency and profitability are the focus, accuracy is paramount for precise predictions; if inclusivity and ethical responsibility take precedence, fairness must be prioritized to mitigate bias and discrimination. For instance, fairness is crucial in hiring models to foster diversity, whereas fraud detection systems prioritize accuracy to minimize financial risk.

Societal values such as justice and accessibility also influence this balance, as AI models can either reinforce or mitigate systemic inequalities. In high-stakes fields like healthcare, finance, or law enforcement, fairness is essential to prevent harm to marginalized groups. A healthcare model might prioritize accuracy for reliable diagnoses, while a loan approval system must carefully balance fairness and accuracy to promote inclusivity without exposing financial institutions to undue risk.

This trade-off presents ethical challenges: focusing too much on fairness may compromise accuracy (e.g., overlooking fraudulent activity), whereas neglecting fairness can perpetuate biases. Transparency and accountability—supported by frameworks like GDPR—help ensure AI decisions remain explainable and ethically sound.

Ultimately, responsible AI development requires ongoing efforts to optimize both fairness and accuracy, with continuous monitoring to adapt to evolving ethical, societal, and business considerations. A well-designed model should not only deliver strong performance but also uphold principles of equity, inclusivity, and justice.