HW 5 - Fairness

95891-A Intro to Artificial Intelligence

Andrew ID: ruomeiw

11/22/2021

1. How many men and how many women are in the data set? How would the answer to this question affect your study of whether the model exhibits gender bias?

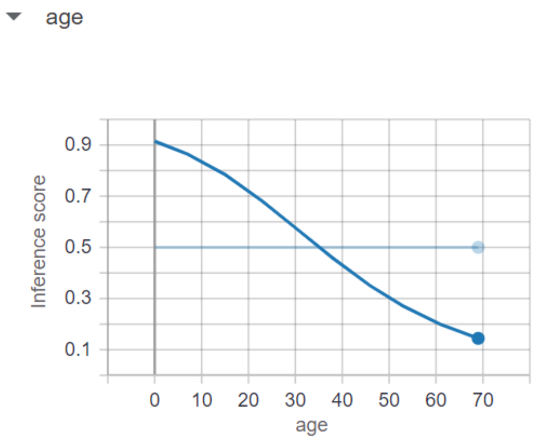
   To find out how many men and how many women are there in the data set, I added below code in the "Read training dataset from CSV" cell.

```
female_count = df[df['sex'] == 'Female'].count()['sex']
male_count = df[df['sex'] == 'Male'].count()['sex']

print('Female Count: ', female_count)
print('Male Count: ', male_count)

----------------------------------------------------------
...
[18316 rows x 40 columns]
Female Count:  3383
Male Count:  14933
```

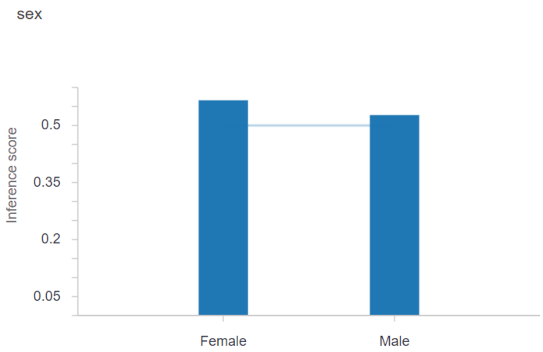
   The output indicates that there are 3383 women and 14933 men in the dataset.

   This answer makes me aware that there is a vast difference between the number of men and women in this dataset, thus, the model trained based on this dataset might be biased towards gender.

2. Look at the partial dependence plots for age and for sex. What do you observe, and can you conclude anything about bias from the partial dependence plots?



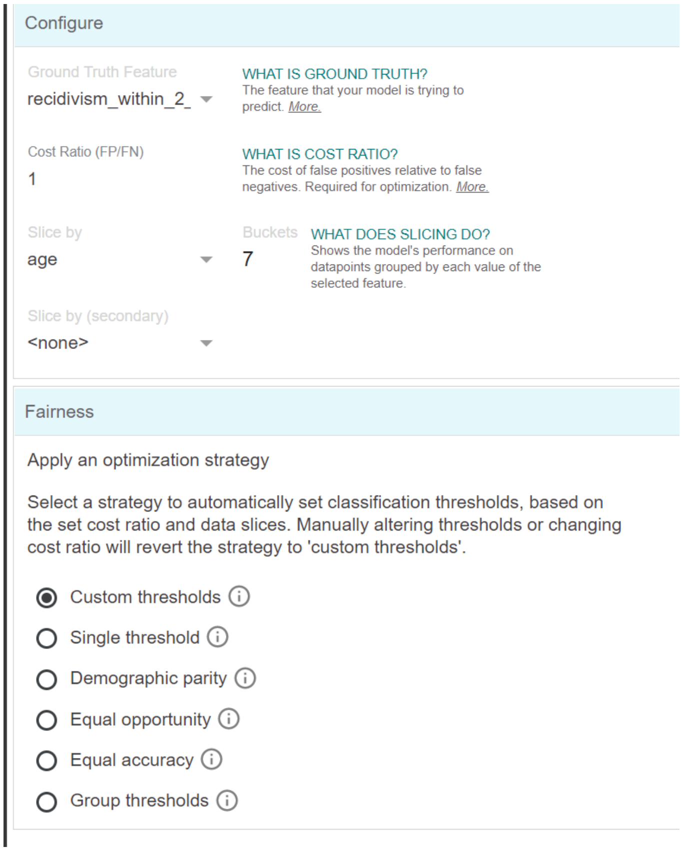
From the partial dependency plot for age, I can see that as age goes up, the inference score decreases. This means that as age increases, the impact of age decreases on the predicted target feature.



From the partial dependency plot for sex, I can see that the inference score of female is higher than the inference score of male. This means that the female gender has bigger impact on the predicted target feature than the male gender.

3. How does accuracy of the model vary with age? What might be a root cause of this variation?

Configuration:



Configure

Ground Truth Feature
recidivism_within_2_

WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN)
1

WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by
age

Buckets
7

WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary)
<none>

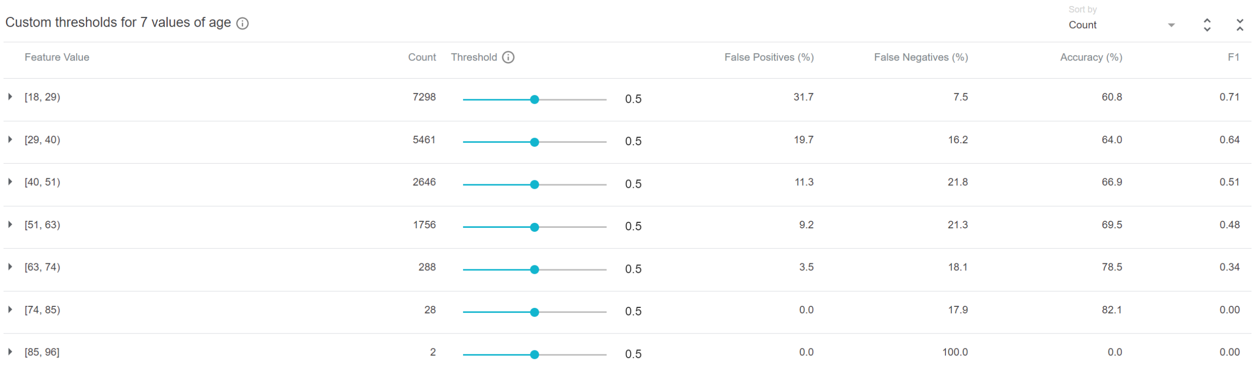
Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

● Custom thresholds ⓘ
○ Single threshold ⓘ
○ Demographic parity ⓘ
○ Equal opportunity ⓘ
○ Equal accuracy ⓘ
○ Group thresholds ⓘ

Metrics:



Custom thresholds for 7 values of age ⓘ

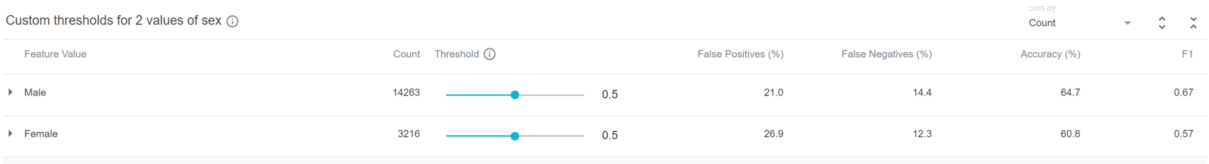
Sort by
Count

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
[18, 29)	7298	0.5	31.7	7.5	60.8	0.71
[29, 40)	5461	0.5	19.7	16.2	64.0	0.64
[40, 51)	2646	0.5	11.3	21.8	66.9	0.51
[51, 63)	1756	0.5	9.2	21.3	69.5	0.48
[63, 74)	288	0.5	3.5	18.1	78.5	0.34
[74, 85)	28	0.5	0.0	17.9	82.1	0.00
[85, 96]	2	0.5	0.0	100.0	0.0	0.00

From the accuracy of the model for different age group, the accuracy increases as age goes up and becomes 0 at the age group [85, 96). Also, we can see that the the youngest group [18, 29) has the highest false positives. As age increases, the false positives gradually go down to 0 for the two oldest age groups, which indicates this model might be biased towards older people and predict them to have lower chances of recidivate within 2 years. The variation here is likely to be caused by the extremely limited data points in the [74, 85) and the [85, 96) age groups.

4. If all thresholds are set at .5 (the default), how do the levels of false positives and negatives vary by sex? If you further slice the data by race, on which particular sex and race combination does the model performs especially poorly? What custom threshold for that combination could you use to bring the accuracy back in line with the other combinations?

Slice by sex:



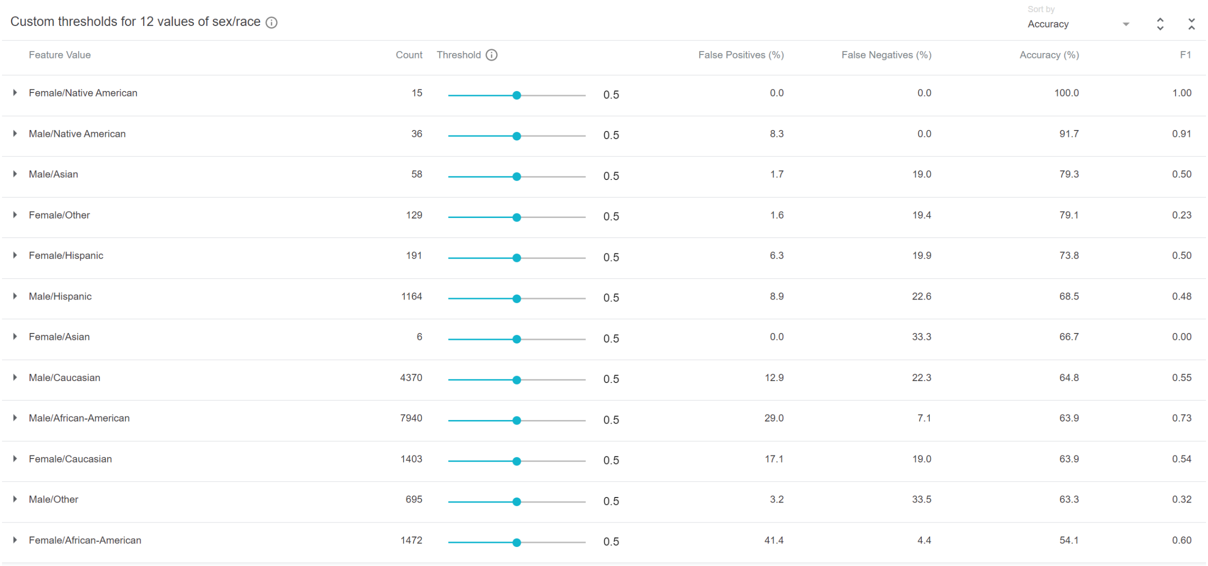
Custom thresholds for 2 values of sex ⓘ

Sort by
Count

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Male	14263	0.5	21.0	14.4	64.7	0.67
Female	3216	0.5	26.9	12.3	60.8	0.57

If all thresholds are set at .5, we can see that males have a lower false positives (21.0%) and a higher false negatives (14.4%); and females have a higher false positives (26.9%) and lower false nagatives (12.3%).

Further slice by race:



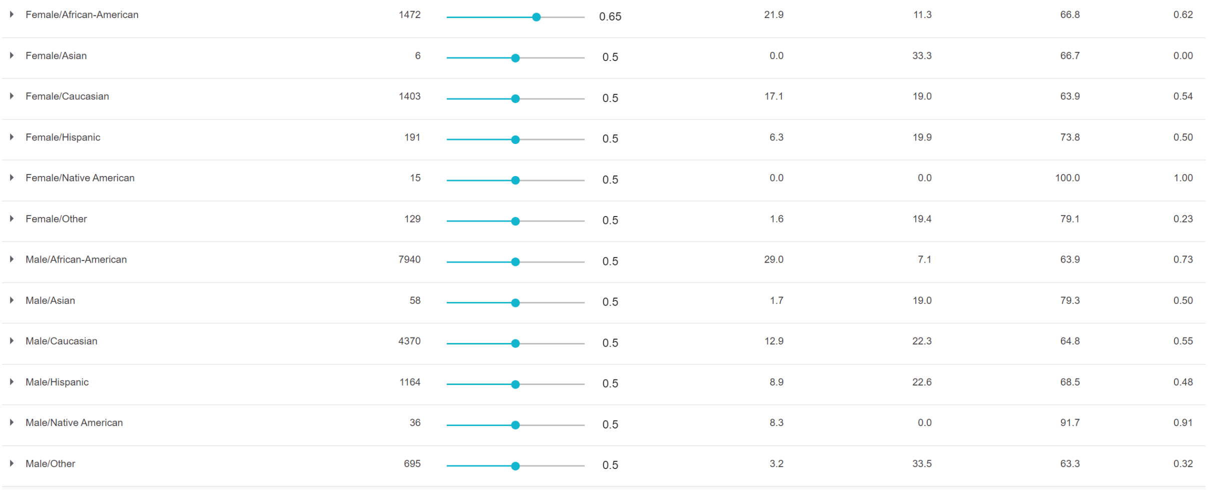
Custom thresholds for 12 values of sex/race

Sort by
Accuracy

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/Native American	15	0.5	0.0	0.0	100.0	1.00
Male/Native American	36	0.5	8.3	0.0	91.7	0.91
Male/Asian	58	0.5	1.7	19.0	79.3	0.50
Female/Other	129	0.5	1.6	19.4	79.1	0.23
Female/Hispanic	191	0.5	6.3	19.9	73.8	0.50
Male/Hispanic	1164	0.5	8.9	22.6	68.5	0.48
Female/Asian	6	0.5	0.0	33.3	66.7	0.00
Male/Caucasian	4370	0.5	12.9	22.3	64.8	0.55
Male/African-American	7940	0.5	29.0	7.1	63.9	0.73
Female/Caucasian	1403	0.5	17.1	19.0	63.9	0.54
Male/Other	695	0.5	3.2	33.5	63.3	0.32
Female/African-American	1472	0.5	41.4	4.4	54.1	0.60

In the “Female/African-American” sex and race combination, the model performs especially poorly and renders the lowest accuracy (54.1%) and the highest false positives (41.4%) among all.

Changing the custom threshold:



Female/African-American	1472	0.65	21.9	11.3	66.8	0.62
Female/Asian	6	0.5	0.0	33.3	66.7	0.00
Female/Caucasian	1403	0.5	17.1	19.0	63.9	0.54
Female/Hispanic	191	0.5	6.3	19.9	73.8	0.50
Female/Native American	15	0.5	0.0	0.0	100.0	1.00
Female/Other	129	0.5	1.6	19.4	79.1	0.23
Male/African-American	7940	0.5	29.0	7.1	63.9	0.73
Male/Asian	58	0.5	1.7	19.0	79.3	0.50
Male/Caucasian	4370	0.5	12.9	22.3	64.8	0.55
Male/Hispanic	1164	0.5	8.9	22.6	68.5	0.48
Male/Native American	36	0.5	8.3	0.0	91.7	0.91
Male/Other	695	0.5	3.2	33.5	63.3	0.32

By trying different the custom thresholds, I found adjusting the threshold to about 0.65 could bring the accuracy back in line with the other combinations. The accuracy for “Famale/African-American” after the change is now 66.8%, closer to other female race combination groups.

5. What is the difference between demographic parity, equal opportunity, and equal accuracy? For slicing the data by sex and race, how varied are the thresholds to achieve the best results for each of the three fairness constraints?

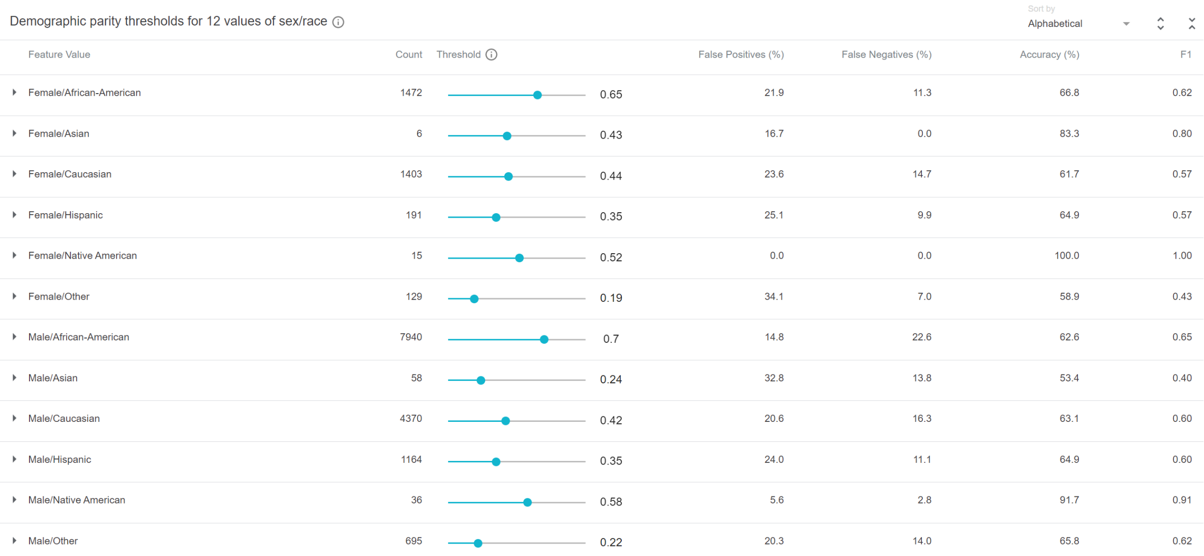
   Demographic parity, equal opportunity, and equal accuracy are three different definitions of fairness.

Demographic parity **means the outcome should be independent of the protected attribute, and it requires similar percentages of data points from each slice are predicted as positive classifications. Demographic parity optimize a threshold per slice based on the specified cost ratio, ensuring the different slices achieve demographic parity.**

Equal opportunity **means that among those data points with the positive ground truth label, there is a similar percentage of positive predictions in each slice. Equal opportunity measures whether the people who should qualify for an opportunity are equally likely to do so regardless of their group membership. Equal opportunity optimize a threshold per slice based on the specified cost ratio, ensuring the different slices achieve equal opportunity.**

Equal accuracy **means that there is a similar percentage of correct predictions in each slice, and it optimize a threshold per slice based on the specified cost ratio, ensuring the different slices achieve equal accuracy.**

Demographic Parity:

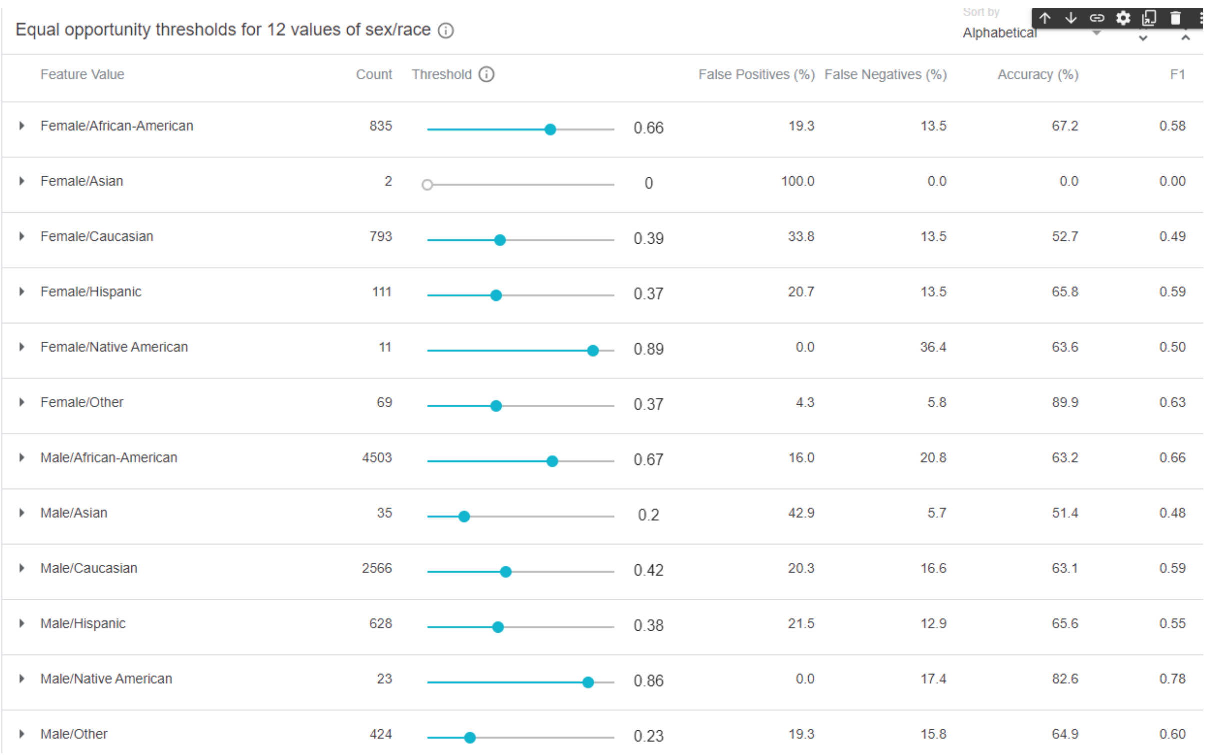


Demographic parity thresholds for 12 values of sex/race

Sort by: Alphabetical

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	1472	0.65	21.9	11.3	66.8	0.62
Female/Asian	6	0.43	16.7	0.0	83.3	0.80
Female/Caucasian	1403	0.44	23.6	14.7	61.7	0.57
Female/Hispanic	191	0.35	25.1	9.9	64.9	0.57
Female/Native American	15	0.52	0.0	0.0	100.0	1.00
Female/Other	129	0.19	34.1	7.0	58.9	0.43
Male/African-American	7940	0.7	14.8	22.6	62.6	0.65
Male/Asian	58	0.24	32.8	13.8	53.4	0.40
Male/Caucasian	4370	0.42	20.6	16.3	63.1	0.60
Male/Hispanic	1164	0.35	24.0	11.1	64.9	0.60
Male/Native American	36	0.58	5.6	2.8	91.7	0.91
Male/Other	695	0.22	20.3	14.0	65.8	0.62

Equal Opportunity:

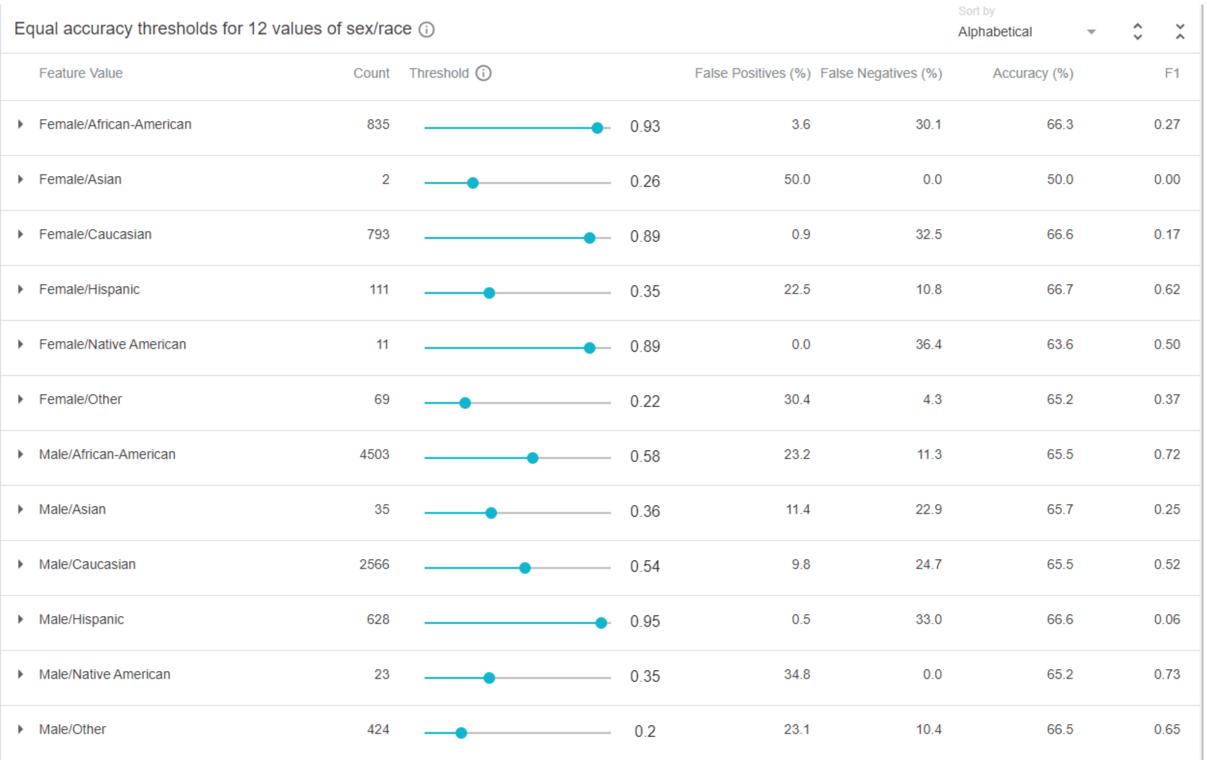


Equal opportunity thresholds for 12 values of sex/race

Sort by: Alphabetical

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	835	0.66	19.3	13.5	67.2	0.58
Female/Asian	2	0	100.0	0.0	0.0	0.00
Female/Caucasian	793	0.39	33.8	13.5	52.7	0.49
Female/Hispanic	111	0.37	20.7	13.5	65.8	0.59
Female/Native American	11	0.89	0.0	36.4	63.6	0.50
Female/Other	69	0.37	4.3	5.8	89.9	0.63
Male/African-American	4503	0.67	16.0	20.8	63.2	0.66
Male/Asian	35	0.2	42.9	5.7	51.4	0.48
Male/Caucasian	2566	0.42	20.3	16.6	63.1	0.59
Male/Hispanic	628	0.38	21.5	12.9	65.6	0.55
Male/Native American	23	0.86	0.0	17.4	82.6	0.78
Male/Other	424	0.23	19.3	15.8	64.9	0.60

Equal Accuracy:



Equal accuracy thresholds for 12 values of sex/race

Sort by: Alphabetical

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	835	0.93	3.6	30.1	66.3	0.27
Female/Asian	2	0.26	50.0	0.0	50.0	0.00
Female/Caucasian	793	0.89	0.9	32.5	66.6	0.17
Female/Hispanic	111	0.35	22.5	10.8	66.7	0.62
Female/Native American	11	0.89	0.0	36.4	63.6	0.50
Female/Other	69	0.22	30.4	4.3	65.2	0.37
Male/African-American	4503	0.58	23.2	11.3	65.5	0.72
Male/Asian	35	0.36	11.4	22.9	65.7	0.25
Male/Caucasian	2566	0.54	9.8	24.7	65.5	0.52
Male/Hispanic	628	0.95	0.5	33.0	66.6	0.06
Male/Native American	23	0.35	34.8	0.0	65.2	0.73
Male/Other	424	0.2	23.1	10.4	66.5	0.65

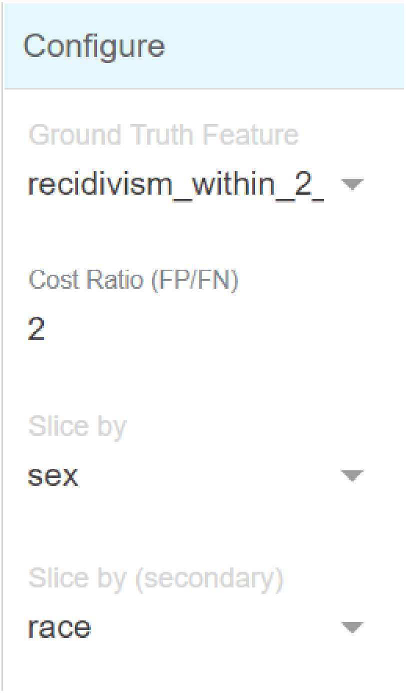
I can see that there are a variety of adjustments made by the fairness measures when the data is partitioned by sex and race.

The thresholds for the combination groups "Female/African-American", "Female/Native American", "Male/African-American", and "Male/Native American" increased while the thresholds for the other groups fell in terms of demographic parity. The changes in accuracy did not follow any clear pattern, with some groups' thresholds rising while accuracy fell or remained the same, and other groups' thresholds falling while accuracy rose.

For equal opportunity, the threshold of the same combination groups “Female/African-American”, “Female/Native American”, “Male/African-American”, and “Male/Native American” increased and the rest groups decreased. There is again no any specific trend in the changes in accuracy.

For equal accuracy, the threshold of more combination groups increased - “Female/African-American”, “Female/Asian”, “Female/Caucasian”, “Female/Native American”, “Male/African-American”, “Male/Caucasian”, “Male/Hispanic”, while the remainder of the groupings shrank. The accuracy was drawn to be close to a centroid (about 66%) but was still inconsistent. Once more, there was no discernible trend in the changes in accuracy, as some groups' thresholds rose while accuracy decreased or remained the same, while other groups' thresholds fell while accuracy climbed.

6. If you vary the cost ratio to weight false positives twice as much as false negatives, how does this affect the achievable accuracy under each of the fairness constraints? Is one fairness constraint more suitable for this data set when the cost ratio is asymmetric?



Configure

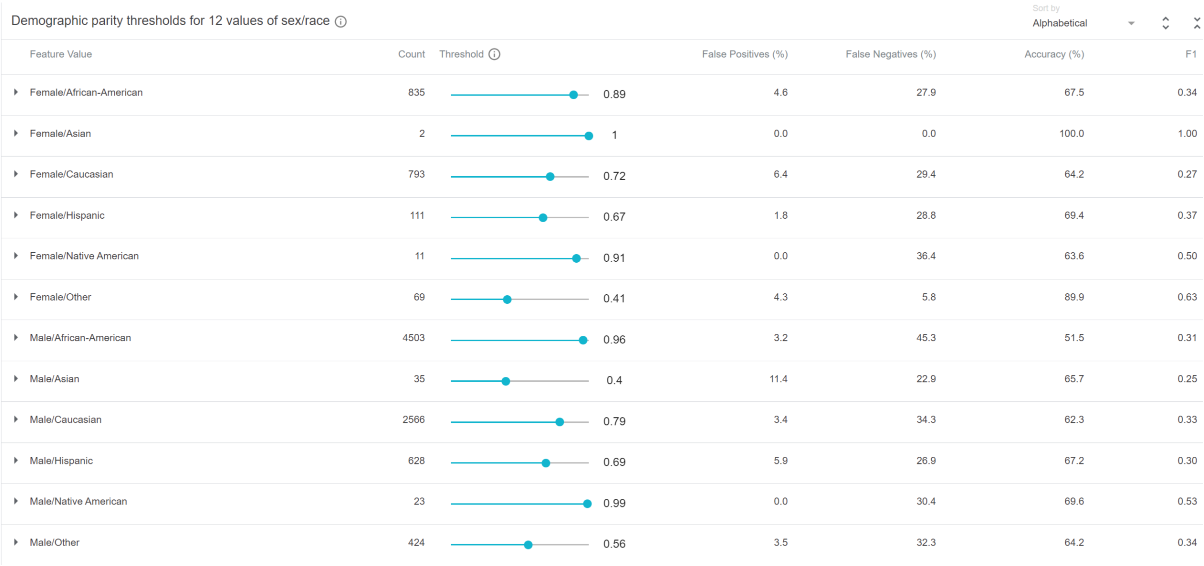
Ground Truth Feature
recidivism_within_2_

Cost Ratio (FP/FN)
2

Slice by
sex

Slice by (secondary)
race

Demographic Parity:

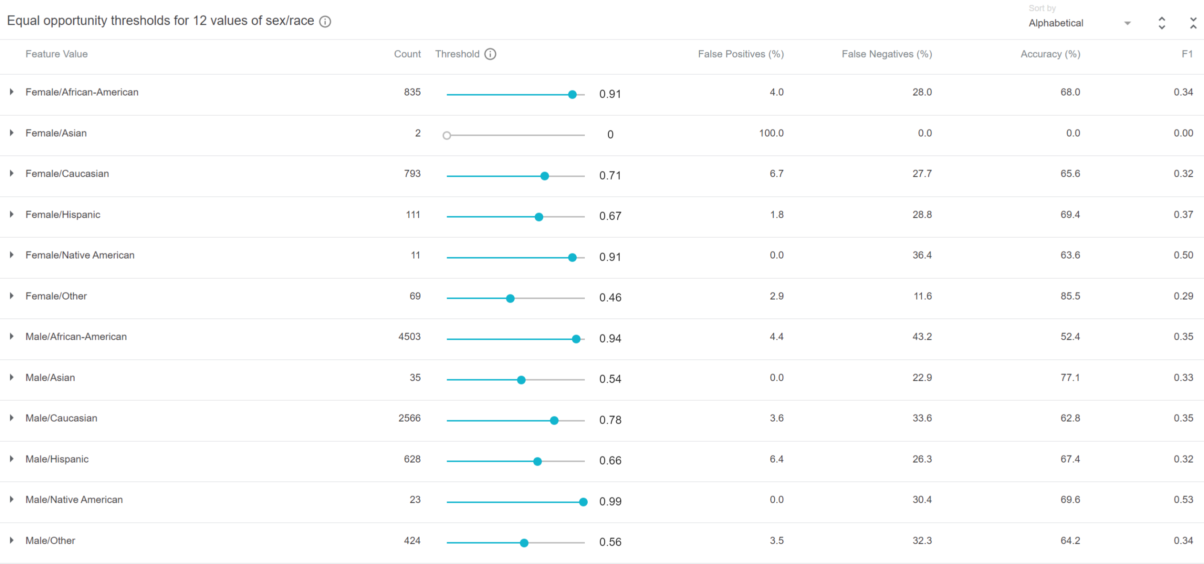


Demographic parity thresholds for 12 values of sex/race

Sort by: Alphabetical

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	835	0.89	4.6	27.9	67.5	0.34
Female/Asian	2	1	0.0	0.0	100.0	1.00
Female/Caucasian	793	0.72	6.4	29.4	64.2	0.27
Female/Hispanic	111	0.67	1.8	28.8	69.4	0.37
Female/Native American	11	0.91	0.0	36.4	63.6	0.50
Female/Other	69	0.41	4.3	5.8	89.9	0.63
Male/African-American	4503	0.96	3.2	45.3	51.5	0.31
Male/Asian	35	0.4	11.4	22.9	65.7	0.25
Male/Caucasian	2566	0.79	3.4	34.3	62.3	0.33
Male/Hispanic	628	0.69	5.9	26.9	67.2	0.30
Male/Native American	23	0.99	0.0	30.4	69.6	0.53
Male/Other	424	0.56	3.5	32.3	64.2	0.34

Equal Opportunity:

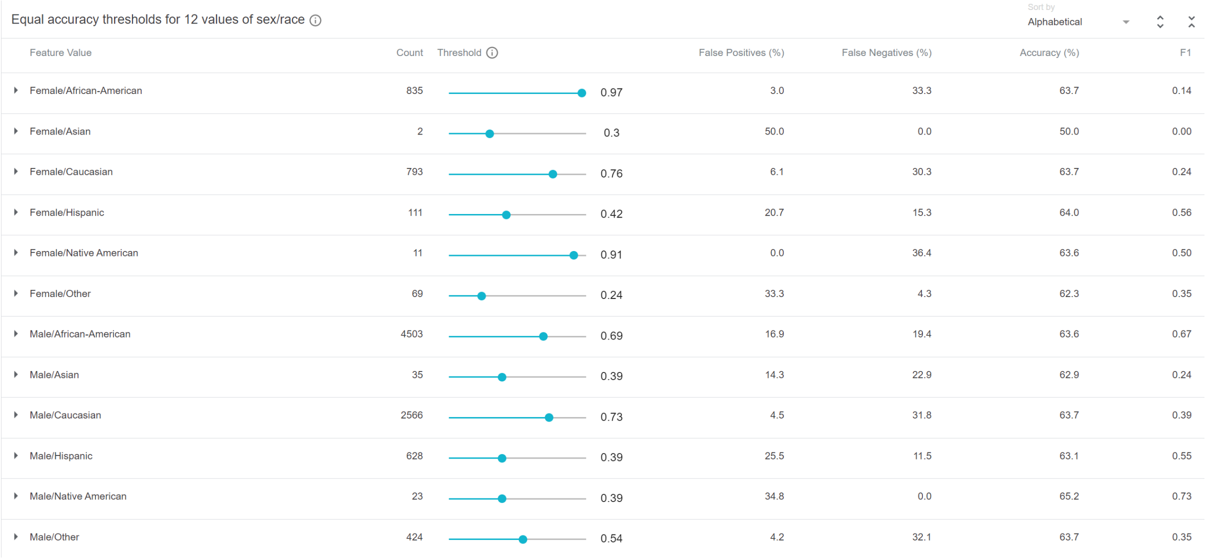


Equal opportunity thresholds for 12 values of sex/race

Sort by: Alphabetical

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	835	0.91	4.0	28.0	68.0	0.34
Female/Asian	2	0	100.0	0.0	0.0	0.00
Female/Caucasian	793	0.71	6.7	27.7	65.6	0.32
Female/Hispanic	111	0.67	1.8	28.8	69.4	0.37
Female/Native American	11	0.91	0.0	36.4	63.6	0.50
Female/Other	69	0.46	2.9	11.6	85.5	0.29
Male/African-American	4503	0.94	4.4	43.2	52.4	0.35
Male/Asian	35	0.54	0.0	22.9	77.1	0.33
Male/Caucasian	2566	0.78	3.6	33.6	62.8	0.35
Male/Hispanic	628	0.66	6.4	26.3	67.4	0.32
Male/Native American	23	0.99	0.0	30.4	69.6	0.53
Male/Other	424	0.56	3.5	32.3	64.2	0.34

Equal Accuracy:



Equal accuracy thresholds for 12 values of sex/race ⓘ

Sort by Alphabetical

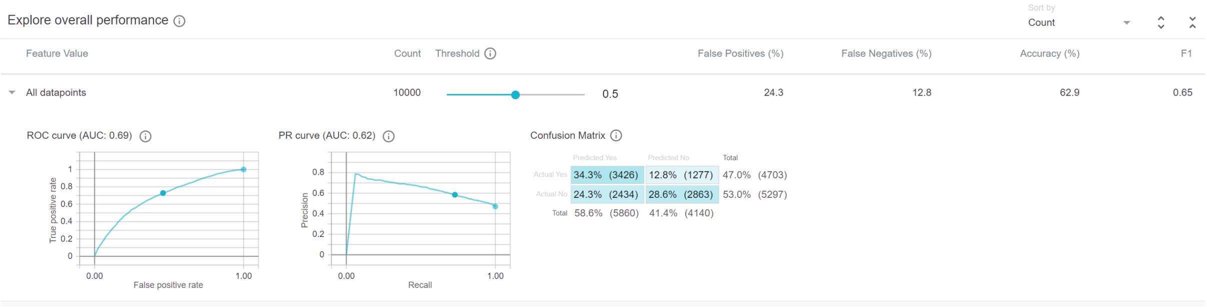
Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
Female/African-American	835	0.97	3.0	33.3	63.7	0.14
Female/Asian	2	0.3	50.0	0.0	50.0	0.00
Female/Caucasian	793	0.76	6.1	30.3	63.7	0.24
Female/Hispanic	111	0.42	20.7	15.3	64.0	0.56
Female/Native American	11	0.91	0.0	36.4	63.6	0.50
Female/Other	69	0.24	33.3	4.3	62.3	0.35
Male/African-American	4503	0.69	16.9	19.4	63.6	0.67
Male/Asian	35	0.39	14.3	22.9	62.9	0.24
Male/Caucasian	2566	0.73	4.5	31.8	63.7	0.39
Male/Hispanic	628	0.39	25.5	11.5	63.1	0.55
Male/Native American	23	0.39	34.8	0.0	65.2	0.73
Male/Other	424	0.54	4.2	32.1	63.7	0.35

The results of setting the cost ratio to 2 are illustrated above. Increasing the cost ratio generally decreased the percentage of false positives since false positives in this model were more costly.

As we can see from these findings, when the cost ratio is asymmetric, neither equal opportunity nor equal accuracy are suitable for this data set because some of the false positives in these two fairness constraints increased rather than decreased. The demographic accuracy constraint is the most appropriate one for this data set's fairness requirements since it produces fewer false positives than a symmetric cost ratio and has the highest average accuracy.

7. Exclude race and gender from the inputs to the model and retrain. Does accuracy go down? Does adding any of the other features in the data that were excluded from the original model (beyond the original 7 that were included) improve accuracy?

Before excluding race and gender:



Explore overall performance ⓘ

Sort by Count

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	10000	0.5	24.3	12.8	62.9	0.65

Confusion Matrix ⓘ

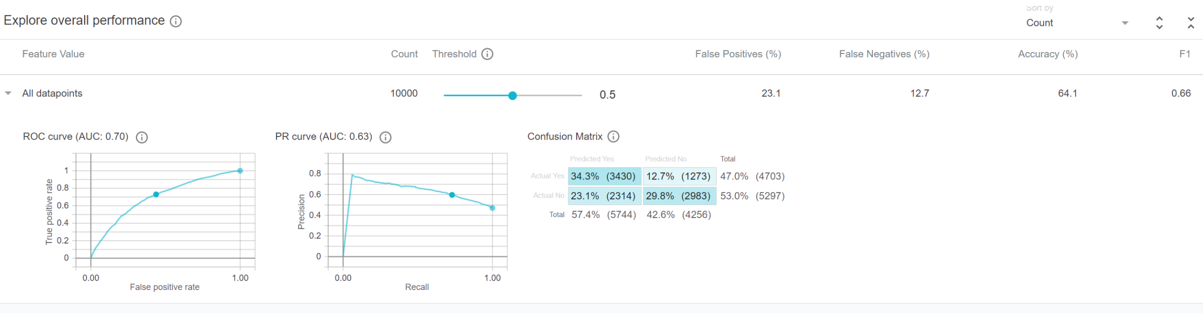
	Predicted Yes	Predicted No	Total
Actual Yes	34.3% (3426)	12.8% (1277)	47.0% (4703)
Actual No	24.3% (2434)	28.6% (2863)	53.0% (5297)
Total	58.6% (5860)	41.4% (4140)	

Excluding race and gender and retrained the model:

I added below code in the "Read training dataset from CSV" cell to exclude "race" and "sex" and revised the code about input features.

```
df = df.loc[:, ~df.columns.isin(['race', 'sex'])]
print(df)
```

```
input_features = ['age', 'priors_count', 'juv_fel_count', 'juv_misd_count', 'juv_other_count']
```



Explore overall performance

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	10000	0.5	23.1	12.7	64.1	0.66

ROC curve (AUC: 0.70)

PR curve (AUC: 0.63)

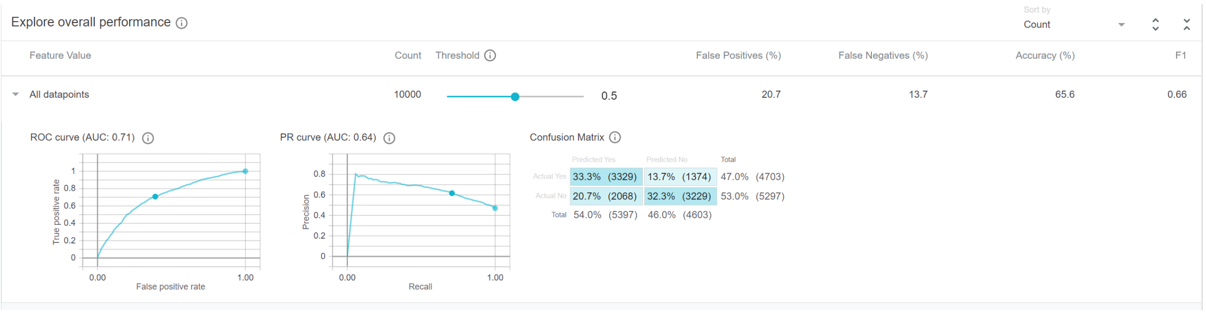
Confusion Matrix

	Predicted Yes	Predicted No	Total
Actual Yes	34.3% (3430)	12.7% (1273)	47.0% (4703)
Actual No	23.1% (2314)	29.8% (2983)	53.0% (5297)
Total	57.4% (5744)	42.6% (4256)	

After excluding race and sex and retrained the model, the accuracy did not go down but went up to 64.1%.

The features I added are “is_violent_recid” and “vr_charge_degree”. The reason why I picked these two is that I think whether crimes they conducted are violent and how serious were the charges should be important factors in this model. However, during the training process, the “vr_charge_degree” could not be interpreted due to its String datatype. So I just went with “is_violent_recid”.

```
input_features = ['age', 'priors_count', 'juv_fel_count', 'juv_misd_count', 'juv_other_count', 'is_violent_recid']
```



Explore overall performance

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	10000	0.5	20.7	13.7	65.6	0.66

ROC curve (AUC: 0.71)

PR curve (AUC: 0.64)

Confusion Matrix

	Predicted Yes	Predicted No	Total
Actual Yes	33.3% (3329)	13.7% (1374)	47.0% (4703)
Actual No	20.7% (2068)	32.3% (3229)	53.0% (5297)
Total	54.0% (5397)	46.0% (4603)	

After retaining the model, the accuracy increased to 65.6%.