



1. [2 points] 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?

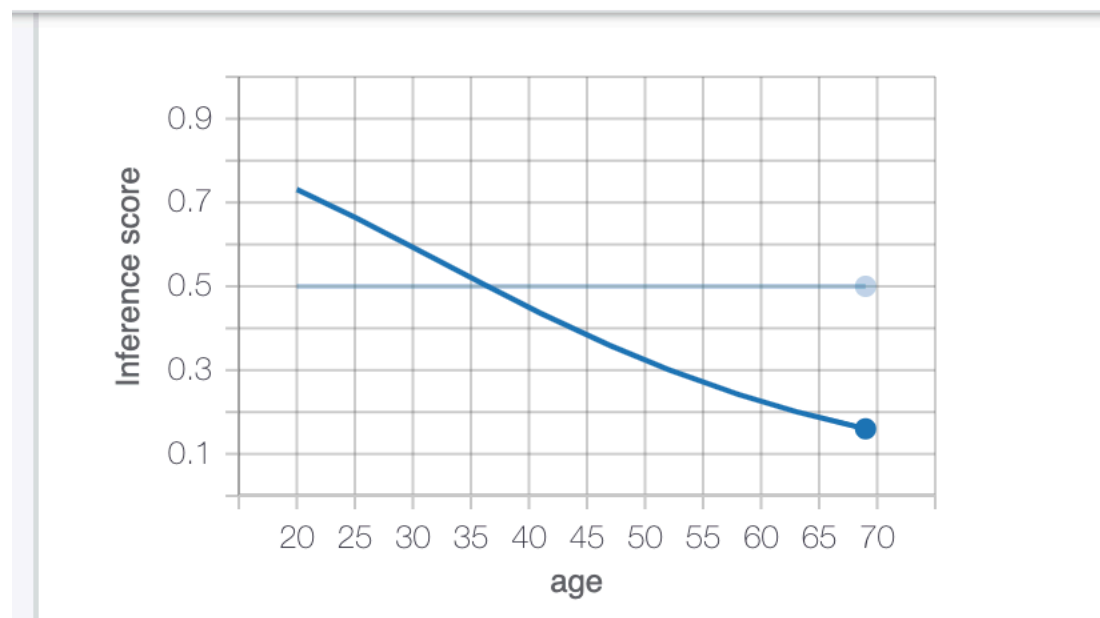
Custom thresholds for 2 values of sex ⓘ

Feature Value	Count	Threshold ⓘ	Sort by Count	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
▶ Male	8179		0.5	22.0	13.6	64.4	0.67
▶ Female	1821		0.5	29.8	12.7	57.4	0.52

As shown in the diagram above, 8179 men and 1821 women are present in the test dataset. If any of the group sample size in the dataset is too small, we cannot reasonably generalize the false positive/ false negative rate of such small sample size into the whole population as the standard deviation of the data would be huge. More data point would be necessary to study the gender bias.

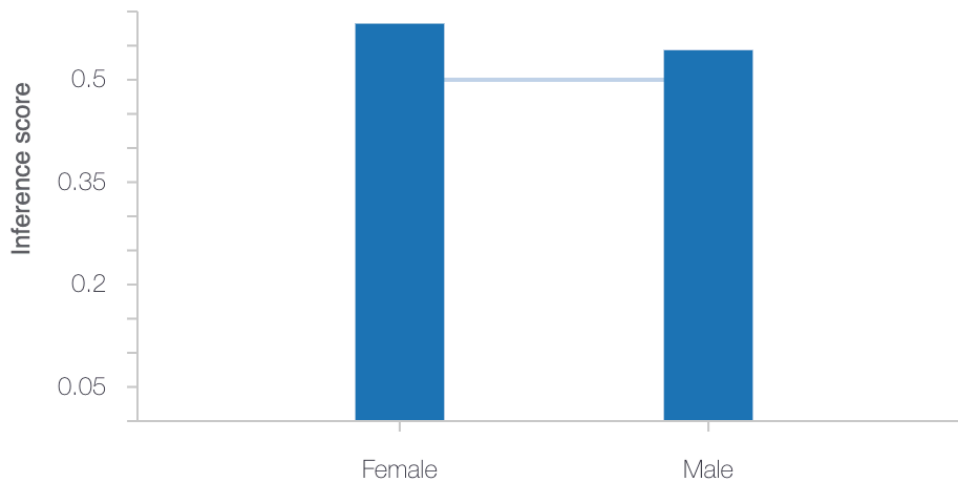
2. [2 points] Look at the partial dependence plots for age and for sex. What do you observe, and can you conclude anything about bias from the partial dependence plots?

▼ age



We could observe a decreasing inference score with the increase in age from the partial dependence plot for age. It means older people have lower probability to be tagged as high-risk criminals by the model.

sex



We could observe that males have lower inference score than females from the partial dependence plot for sex, indicating higher recidivism rate predicted for females.

3. [2 points] How does accuracy of the model vary with age? What might be a root cause of this variation?

Custom thresholds for 5 values of age ⓘ				Sort by Alphabetical ▾ ⓘ ⌵ ⌴			
Feature Value	Count	Threshold ⓘ		False Positives (%)	False Negatives (%)	Accuracy (%)	F1
▶ [18, 34)	5908	<div><div></div></div>	0.5	29.4	9.0	61.6	0.70
▶ [34, 49)	2659	<div><div></div></div>	0.5	17.5	19.5	63.0	0.52
▶ [49, 65)	1308	<div><div></div></div>	0.5	10.0	20.9	69.0	0.48
▶ [65, 80)	122	<div><div></div></div>	0.5	3.3	17.2	79.5	0.19
▶ [80, 96]	3	<div><div></div></div>	0.5	0.0	66.7	33.3	0.00

Based on the image above, the accuracy rate of the model generally increases with age. There is a sudden drop in accuracy score for [80,96] as there are only 3 people in the group, which is a biased representation of the group. The general rise in accuracy might be caused by the fact that people with older age have less tendency to commit crime again in two years. This conjecture is further supported by the decrease in false positive rate with the increase in ages; the model is making fewer positive predictions for older groups.

4. [3 points] 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?

Feature Value	Count	Threshold ⓘ		False Positives (%)	False Negatives (%)	Accuracy (%)	F1
▶ Female	1821	<div><div></div><div></div></div>	0.5	29.8	12.7	57.4	0.52
▶ Male	8179	<div><div></div><div></div></div>	0.5	22.0	13.6	64.4	0.67













If all thresholds are set to 0.5, female have higher false positive rate and lower false negative rate than male. Female/African-American has the lowest accuracy score of 50.5%, just slightly better than random guess. I would increase the threshold to about 0.62 to reach an accuracy score of 61.2% which is similar to the accuracy score for the entire population (63.1%).

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	10000	<div><div></div><div></div></div> 0.5	23.4	13.5	63.1	0.65

5. [2 point] 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?












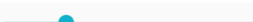
Demographic parity means that similar percentages of datapoints from each slice are predicted as positive classifications.

Demographic parity thresholds for 12 values of sex/race ⓘ












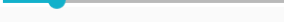
Feature Value	Count	Threshold ⓘ	Fals
▶ Female/African-American	835		0.63
▶ Female/Asian	2		0.29
▶ Female/Caucasian	793		0.42
▶ Female/Hispanic	111		0.36
▶ Female/Native American	11		0.59
▶ Female/Other	69		0.19
▶ Male/African-American	4503		0.7
▶ Male/Asian	35		0.25
▶ Male/Caucasian	2566		0.41
▶ Male/Hispanic	628		0.35
▶ Male/Native American	23		0.84
▶ Male/Other	424		0.22

Equal opportunity means that among those datapoints with the positive ground truth label, there is a similar percentage of positive predictions in each slice.

Equal opportunity thresholds for 12 values of sex/race ⓘ

Feature Value	Count	Threshold ⓘ	False
▶ Female/African-American	835		0.67
▶ Female/Asian	2		0
▶ Female/Caucasian	793		0.4
▶ Female/Hispanic	111		0.41
▶ Female/Native American	11		0.92
▶ Female/Other	69		0.35
▶ Male/African-American	4503		0.71
▶ Male/Asian	35		0.24
▶ Male/Caucasian	2566		0.43
▶ Male/Hispanic	628		0.4
▶ Male/Native American	23		0.93
▶ Male/Other	424		0.24

Equal accuracy means that there is a similar percentage of correct predictions in each slice.

Feature Value	Count	Threshold ⓘ	False Po
▶ Female/African-American	835		0.91
▶ Female/Asian	2		0.29
▶ Female/Caucasian	793		0.88
▶ Female/Hispanic	111		0.34
▶ Female/Native American	11		0.92
▶ Female/Other	69		0.21
▶ Male/African-American	4503		0.58
▶ Male/Asian	35		0.41
▶ Male/Caucasian	2566		0.55
▶ Male/Hispanic	628		0.95
▶ Male/Native American	23		0.43
▶ Male/Other	424		0.19

Comparing the results, the direction of adjustment for threshold is the same for demographic parity and equal opportunity while the adjustment extent is different. To reach equal accuracy, the adjustments needed are quite different from the previous two.

6. [4 points] 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?












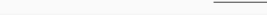
After changing the cost ratio of false positives to false negatives to 2, I observed the results as followed. The thresholds are adjusted in a way to reduce false positive rate.

The change in cost ratio decreases the achievable accuracy under each of the fairness constraints as it reduces false positive rate at the cost of accuracy.

Equal accuracy is not suitable as its false positive rates remain high with this adjusted cost ratio although the overall accuracy rate might be higher than that of the other two fairness constraints. Comparing equal opportunity and demographic parity, equal opportunity has slightly higher overall accuracy score, hence more suitable for this data set when the cost ratio is asymmetric.

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		3.8	28.6	67.5	0.32
Female/Asian	2		0.0	0.0	100.0	1.00
Female/Caucasian	793		5.4	30.3	64.3	0.24
Female/Hispanic	111		1.8	28.8	69.4	0.37
Female/Native American	11		0.0	36.4	63.6	0.50
Female/Other	69		4.3	8.7	87.0	0.47
Male/African-American	4503		2.7	46.6	50.7	0.28
Male/Asian	35		2.9	22.9	74.3	0.31
Male/Caucasian	2566		2.8	35.2	62.0	0.31
Male/Hispanic	628		5.1	28.3	66.6	0.26
Male/Native American	23		0.0	30.4	69.6	0.53
Male/Other	424		4.0	34.7	61.3	0.27

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	<div><div></div><div></div></div> 0.88	4.3	27.9	67.8	0.34
Female/Asian	2	<div><div></div><div></div></div> 0	100.0	0.0	0.0	0.00
Female/Caucasian	793	<div><div></div><div></div></div> 0.66	7.1	27.2	65.7	0.34
Female/Hispanic	111	<div><div></div><div></div></div> 0.63	1.8	28.8	69.4	0.37
Female/Native American	11	<div><div></div><div></div></div> 0.92	0.0	36.4	63.6	0.50
Female/Other	69	<div><div></div><div></div></div> 0.42	2.9	11.6	85.5	0.29
Male/African-American	4503	<div><div></div><div></div></div> 0.93	4.1	43.5	52.4	0.35
Male/Asian	35	<div><div></div><div></div></div> 0.56	0.0	22.9	77.1	0.33
Male/Caucasian	2566	<div><div></div><div></div></div> 0.74	3.6	33.5	62.9	0.35
Male/Hispanic	628	<div><div></div><div></div></div> 0.62	7.0	26.1	66.9	0.32
Male/Native American	23	<div><div></div><div></div></div> 0.99	0.0	30.4	69.6	0.53
Male/Other	424	<div><div></div><div></div></div> 0.48	4.2	32.1	63.7	0.35

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	<div><div></div><div></div></div> 1	0.0	36.2	63.8	0.00
Female/Asian	2	<div><div></div><div></div></div> 0.29	50.0	0.0	50.0	0.00
Female/Caucasian	793	<div><div></div><div></div></div> 0.71	6.1	30.3	63.7	0.24
Female/Hispanic	111	<div><div></div><div></div></div> 0.37	20.7	15.3	64.0	0.56
Female/Native American	11	<div><div></div><div></div></div> 0.92	0.0	36.4	63.6	0.50
Female/Other	69	<div><div></div><div></div></div> 0.2	33.3	4.3	62.3	0.35
Male/African-American	4503	<div><div></div><div></div></div> 0.65	16.9	19.5	63.6	0.67
Male/Asian	35	<div><div></div><div></div></div> 0.39	14.3	22.9	62.9	0.24
Male/Caucasian	2566	<div><div></div><div></div></div> 0.65	5.7	30.6	63.7	0.42
Male/Hispanic	628	<div><div></div><div></div></div> 0.35	25.5	11.6	62.9	0.55
Male/Native American	23	<div><div></div><div></div></div> 0.43	34.8	0.0	65.2	0.73
Male/Other	424	<div><div></div><div></div></div> 0.48	4.2	32.1	63.7	0.35

7. [5 points] 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? Include code for this question.

Before:

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	10000	<div><div></div><div></div></div> 0.5	23.4	13.5	63.1	0.65

After:

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

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
▼ All datapoints	10000	<div><div></div><div></div></div> 0.5	23.1	12.7	64.1	0.66

After I exclude race and sex from the inputs to the model, the accuracy of the model increases.

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

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
▼ All datapoints	10000	<div><div></div><div></div></div> 0.5	21.1	13.3	65.6	0.66

After I add in the feature is_violent_recid, the accuracy score is increased to 65.6%.