CSE 574: Assignment 3

Model used: Naive Bayes (Categorical)
Secondary criteria: Cost
Total Accuracy:
Training: 69. 128 %
Testing: 67. 162 %

Algorithm: Equal Opportunity

Total Cost: -$673,320,500.00
Training: $ -536,249,750
Testing: $ -137,070,796

Motivation

- The COMPAS model correctly predicted recidivism at similar rates for both white (59%) and black defendants (63%).
- But, the algorithm was biased as it tended to skew very differently for each of these groups [2].
    - White defendants who re-offended within two years were mistakenly labelled low-risk almost twice as often as their black counterparts.
    - Additionally, black defendants who did not recidivate were rated as high-risk at twice the rate of comparable white defendants.
- The charts presented by ProPublica showed that white defendants were skewed towards lower-risk as compared to black defendants.
- We are trying to address this problem by equalizing the number of true predicted recidivists on a given number of known recidivists in each of the races.

Stakeholders

- The primary stakeholders would be criminals being considered for pretrial, sentencing, or parole.
- Law makers, judges, and executioners would also be stakeholders in this as users of this system.
- Financially, the government and taxpayers are the stakeholders.
- The people who could potentially be affected due to a false negative slipping away.

Biases

ML algorithms generally function through the extraction and manipulation of mathematical encodings of patterns present within a set of data. Bias and discrimination are interwoven in the data itself as they are acquired from the patterns of the society. Even if race is not an input feature, other features are correlated with race. As an example of such correlation, an inmate's number of priors and previous time in jail are correlated with race because originally black inmates were more likely to be jailed for the same crimes as white inmates and are arrested at a higher rate for less serious crimes. These biases can be perpetuated in machine learning models as present in the data.

Impact of our solution

Firstly, our model does not take only accuracy into account as accuracy does not give the impact of overall skew as well as race-wise skew. For instance, the accuracy may be the same in case the total sum of true positives and true negatives is the same. But at the same time, if accuracy is skewed towards more number of false negatives than false positives, it decreases the overall cost if the cost associated with false negatives and false positives is different. Our model has an advantage that it does not allow the accuracy to skew in the undesirable range. Since the negatives associated with more number of criminals on the street are often calculated to be more serious than the negatives for someone being falsely incarcerated, one should follow whatever is in the interest of the society as a whole. Since it is always a tradeoff, the tradeoff can be made in a more calculated way.
Secondly, we minimize the bias across the races as our solution accounts for an equal number of true predicted recidivists on a given number of known recidivists in each of the races. This is done by applying equal opportunity threshold to a Categorical Naive bayes model. The major impact will be maximizing the total profit while keeping equal opportunity for all

the races. This ensures that the accuracy is always skewed for minimizing false negatives, hence safer society and minimum loss.

Why this solution

The metrics often considered for fairness are TPR, FPR, PPV and NPV. Our solution optimizes the total profit while keeping equal opportunity as a constraint. This can be justified in terms of given “cost” for True Positives, False Positives, True negatives and False Negatives. We know that TP + FN = 1 and FP + TN = 1. It is always a tradeoff and **when we maximize accuracy, either the number of false positives or false negatives increases**. **But since we know the (financial) cost associated with each of them, our model optimizes cost** to reduce their negative impacts on society rather than just being more accurate.
As our group's mission revolves around making fair decisions while also considering financial concerns, it makes sense to use cost as our secondary optimization rather than blindly optimizing accuracy.
From the constraints perspective, we keep the true positives rate within 1% for all the races. Following [1], unlike demographic parity or predictive parity, equalized odds allows prediction to depend on sensitive attributes but only through the target variable Y. It allows us to directly predict Y, but prohibits abusing sensitive attributes as a proxy for Y. **Equalized odds enforces** that the accuracy is equally high in all demographics, punishing models that perform well only on the majority. In this way our solution addresses the issue with the COMPAS model.

$$\text{Obj Func} = \text{Max cost} \quad TP*(-\$60{,}076) + TN * (\$23{,}008) + FP * (-\$110{,}076) + FN * (-\$202{,}330)$$
$$\text{Constraint - for all classes } TPR_i = TPR$$

If we assume TP = x, FN = 1 - x, FP = y, TN = 1 - y, substituting in the eq, $142x - $133y - 142. This shows that objective function is skewed towards keeping TP (x) slightly higher than it is for keeping FP (y) lower. Now, we know that equal opportunity can slightly reduce overall accuracy, as no predictor can have a higher TPR than the worst performer of the lot. But since we already have an objective function which takes care of the slight decrease in accuracy, it tends to somewhat neutralize the effect as the skew is already in place for true positives to work better.
This is clearly reflected in the results as shown in the tables below that for our market model gives the best result in terms of cost and equal opportunity over all the available models evaluated. Our model performs best in terms of maximum equal opportunity and giving best profit at the same time.

Extra Justifications

Our metric for comparison is equal opportunity and we consider this more appropriate than other methods as it provides equal true positive rates across the races which was the original problem posed by the COMPAS system. TPR was low for white-defendants and FPR was high for black-defendants. Equalizing TPR or FPR could be the possible solutions in this case, but given the loss associated with FP is less than FN (1-TP), it makes more sense to equalize TPR as we already have a skew to balance out the negative impact of enforcing equal opportunity constraint as it slightly decreases overall accuracy towards that of the lowest TPR class.

There are certain assumptions like, the variance and biases of different models are considered equal as we do not process the training and test data separately. This information remains hidden as we postprocess concatenated data from the models.

In the real-world there is a bias in the data due to its correlation with sensitive attributes. This is also evident from the fact that even though COMPAS did not take race into consideration, racial biases creeped in. Another example of such a correlation is stated in the biases section above.
The risk in our model is measured in terms of the financial and general impact on the society. This is handled by using cost as the secondary optimization criteria, thus minimizing the worst impact possible.

Splitting by gender gives similar results as race, indicating a fair model. The accuracy when split on age is much less owing to a lack of data in much of the category splits.

References

[1] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

[2] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.

Tables and Comparisons

Table 1 (NBC)

NBC	Maximum Profit				Demographic Parity				Predictive Parity				Single Threshold				Equal Opportunity			
	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot
TPR	76.17%	70.78%	50.43%	50.37%	65.75%	69.61%	74.57%	75.19%	84.58%	68.19%	54.31%	50.38%	78.06%	70.26%	64.22%	63.16%	70.75%	70.99%	70.69%	70.68%
FPR	54.50%	44.02%	19.34%	18.54%	44.41%	43.92%	44.42%	44.39%	67.93%	42.66%	23.86%	20.49%	58.10%	44.42%	32.63%	29.76%	49.01%	45.49%	39.26%	40.49%
PPV/Pos prob for demo					56.43%	56.77%	56.84%	56.51%	61.61%	61.51%	61.46%	61.47%								
Acc	62.76%	62.93%	68.21%	69.23%	61.32%	62.84%	63.41%	63.31%	61.64%	62.77%	67.14%	68.05%	62.26%	62.92%	66.07%	67.46%	62.11%	62.75%	64.83%	63.91%
Tot Cost	-$750,610,206.00				-$765,265,554.00				-$757,430,622.00				-$754,522,380.00				-$760,001,220.00			

Table 2 (NN)

NN	Maximum Profit				Demographic Parity				Predictive Parity				Single Threshold				Equal Opportunity			
	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot
TPR	78.27%	71.61%	57.58%	33.08%	60.81%	65.21%	62.07%	67.66%	93.84%	73.82%	50.43%	36.09%	78.27%	71.60%	57.77%	52.63%	100.00%	100.00%	100.00%	100.00%
FPR	52.33%	43.37%	30.21%	10.73%	42.64%	38.91%	39.57%	44.39%	82.85%	49.52%	25.67%	16.09%	52.33%	43.32%	30.21%	27.80%	100.00%	100.00%	100.00%	100.00%
PPV/Pos prob for demo					52.87%	52.06%	48.84%	53.55%	59.34%	59.85%	57.92%	59.26%								
Acc	64.89%	64.12%	64.83%	67.45%	59.30%	63.15%	61.10%	60.35%	60.33%	62.15%	64.47%	65.08%	64.49%	64.11%	64.83%	64.50%	56.30%	50.00%	41.20%	39.34%
Tot Cost	-$733,279,770.00				-$777,229,038.00				-$768,936,708.00				-$734,375,070.00				-$910,659,204.00			

Table 3 (SVM)

SVM	Maximum Cost				Demographic Parity				Predictive Parity				Single Threshold				Equal Opportunity			
	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot
TPR	85.69%	74.04%	65.51%	33.08%	70.59%	74.23%	75.43%	75.19%	93.58%	80.13%	42.22%	36.84%	86.74%	80.13%	71.12%	66.17%	74.91%	75.70%	75.43%	75.19%
FPR	62.92%	47.12%	34.74%	10.24%	48.68%	47.52%	51.36%	49.76%	81.49%	53.92%	20.54%	16.10%	65.69%	53.92%	41.99%	34.15%	51.25%	50.33%	51.17%	49.76%
PPV/Pos prob for demo					61.02%	60.87%	61.27%	59.76%	59.68%	59.77%	59.04%	59.76%								
Acc	64.45%	63.45%	65.36%	67.46%	62.17%	63.34%	59.68%	60.06%	60.78%	63.10%	64.12%	65.38%	63.83%	63.10%	63.41%	65.98%	63.48%	62.69%	59.50%	60.05%
Tot Cost	-$738,440,370.00				-$759,052,674.00				-$757,161,864.00				-$743,930,460.00				-$757,870,974.00			

Table 4 Market Model (Our model - Categorical Naive Bayes)

Custom	Maximum Profit				Demographic Parity				Predictive Parity				Single Threshold				Equal Opportunity			
	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot	AA	Ca	Hi	Ot
TPR	84.00%	70.69%	57.76%	57.89%	67.60%	70.69%	75.00%	74.43%	84.16%	70.69%	40.09%	47.36%	83.32%	70.68%	45.69%	45.11%	71.01%	70.68%	69.82%	71.43%
FPR	50.71%	33.46%	20.54%	19.02%	31.32%	33.47%	36.86%	37.07%	51.32%	33.47%	13.29%	14.63%	49.97%	33.47%	14.20%	13.66%	34.03%	33.46%	32.62%	29.26%
PPV/Pos prob for demo									59.68%	59.77%	59.04%	59.76%								
Acc	68.83%	68.61%	70.52%	71.89%	68.07%	68.61%	68.03%	67.46%	68.78%	68.60%	67.14%	71.30%	68.78%	68.61%	69.27%	70.12%	68.81%	68.61%	68.38%	71.01%
Tot Cost	-$669,408,282.00				-$679,004,802.00				--673,609,554				$671,932,962.00				-$673,320,500.00			

Values that are more than (Average + 5) for that particular metric in a comparison of races

Values that are less than (Average - 5) for that particular metric in a comparison of races

Prepared By:

Sahil Kapahi
sahilkap@buffalo.edu
50317075

Salil Dabholkar
salilsan@buffalo.edu
50321748

Nihar Patel
nihardil@buffalo.edu
50318506