Fairness in Machine Learning

Role: Employee of Publicly Traded Corporation

CSE 574: Introduction to Machine Learning - PA 3

May 8, 2020

Group # 4

Anjali Bachani
Abhishek Mishra
Adithya L. Narayanan

To replace Northpointe's COMPAS system, we present to the market, a system which uses a **Naive Bayes classifier** and enforces **Equal Opportunity** across racial lines to ensure maximum fairness possible within the scope of the project requirements. We use the equal opportunity algorithm to ensure that the **optimal accuracy (secondary optimization)** that the data offers can be achieved. The following reasons make our market model an able competitor to, if not better than COMPAS:

Part 1: Motivation
COMPAS claims to achieve Predictive Parity when base rates of recidivism for races are accounted for. However, this approach falls short morally. Each individual should be assessed from a unique standpoint and bringing a prior probability of recidivism into account due to a person's race being more or less likely to recidive on an average is not a morally sound approach. The data that drives this prior is bound to carry the decades of bias that races have faced. Even if this were acceptable, predictive parity only ensures that of all the people the algorithm labels as positive (likely to recidivate), the proportion who are actually going to recidivate is the same for every race. This doesn't ensure fairness in the real world, since the true positive rate or TPR (proportion of people likely to recidivate that are correctly identified) and the false positive rate or FPR (proportion of people unlikely to recidivate that are wrongly misclassified) can be quite different for every race and therefore cannot stand up to public scrutiny if the algorithm skews differently across races and allows a leniency for a one/some. We wish to account for as many of these rate metrics as we can in ensuring fairness, so that there isn't a particular race, in the real world, that is being dealt an unfair hand in chances of rightly or wrongly being identified as positive.

Part 2: Stakeholders
The people with interests vested in this situation are the public itself- the very people this model is supposed to be fair to, across racial lines, the publicly traded company that we work for, who have a vested interest in ensuring a product that can win them further contracts to stay profitable, and the government/justice system which has a primary objective of ensuring that the freedom of the innocent are upheld, while punishing as many guilty as possible (in this case, ensuring that every unlikely to recidivate person should optimally be classified as negative while ensuring that as many true positives are classified i.e. a low false positive rate, and a high true positive rate). Ideally, no innocent person's rights should be compromised, at any cost.

Part 3: Bias
In most machine learning scenarios like this, bias is almost inevitable since models rely on data. In our situation, this data comes from past data on recidivism which cannot be guaranteed to be free from bias since it was a product of human decision making in the justice system and these biases will definitely manifest. By training models on such data, it is possible that the biases existent in the data will manifest in predictions and be biased against a certain racial group. There is also scope for a bias in the number of cases that are even recorded across racial lines, with some races facing more scrutiny than others. This can in turn also lead to class imbalance, and thereby, poor predictive properties.

Part 4: Impact
Keeping these points in mind, in a utopia, we believe false positive rates should be the first condition to be equalised, but since that remains outside the scope of the project since specificity is unavailable to us as an algorithm, using equal opportunity as out measure of fairness enforcer, we strive to ensure that the model doesn't identify a higher proportion of people correctly (likely to recidivate) from one particular race by equating over this metric using a unique threshold for each race. By achieving this, we also achieve equality over the proportion of people who are likely to recidivate but aren't identified across all races, and ensure that the algorithm is not going after a particular race unfairly when identifying positives correctly of negatives incorrectly, i.e, every likely person, who is actually likely to recidivate, has an equal and thereby fair chance, of being identified, regardless of which race he or she belongs to.

Single threshold is the simplest approach to designing a fair approach as it excludes characteristics pertaining to each group, and therefore has no dependence on group membership. However, not taking characteristics of each group into consideration and evaluating them using the same threshold can also lead to discriminatory behavior. For instance, in Broward county, FL, woman with COMPAS score of seven recidivate less than 50% whereas men with same score recidivate more than 60% [1]. By acknowledging the predictive value of gender, one can create a decision rule that detains fewer women and ignoring this information would be discrimnitation against women. Therefore, when characteristics like gender or race add predictive value, excluding these attributes and evaluating the model using a single threshold will lead to unjustified disparate impacts. Maximum profit is clearly the most unfair solution as defining different thresholds for each group in order to achieve maximum accuracy can be very pessimistic and holding different groups to different standards is a key trait of being unfair. Instead, it makes sense to look at the difference between the true and false positive rates that each classification criteria achieves for each group.

Demographic Parity is a flawed solution since it labels some individuals from a certain racial group positively or negatively without much more context just to ensure that the proportions of individuals labeled across groups match [2]. In our case, by setting thresholds such that the proportion of people labeled recidivists are same in each racial group, we actually do wrongly classify some individuals as not recidivists in some racial groups. By allowing such individuals to have a better chance at any further justice proceedings, we are in fact, over time, widening the gap between such a group and the group based on whom the threshold is set since we will account for most of the people likely to recidivate from this group. Predictive parity ensures that an arbitrary individual from every racial group has an equal chance of being correctly labeled a recidivist. However, predictive parity stops here. By allowing for a further measure of fairness using the true positive and false negative rates (FNR), equal opportunity achieves what predictive parity does by allowing for similar classifications of recidivists across racial lines, but while ensuring that every likely person, who is actually likely to recidivate, has an equal and thereby fair chance, of being identified, regardless of which race he or she belongs to, unlike predictive parity which too accounts for equal probability of true positives, but only amongst those it it has labeled positive already [3]. By ensuring equality in the population's positives itself, equal opportunity ensures that every person likely to recidivate is given a rope of the same length in not being identified. It can therefore be said that the model doesn't target a race by correctly identifying more positives from that group.

Part 5: Choice

Since we choose the equal opportunity postprocessing method for our model, we observe equal TPR (therefore, equal FNR) across all races. However, the FPR is slightly higher for African Americans comparatively. We need to understand that using this simple model, we can't uncover the problems in decision-making for the criminal justice system or provide a truly viable solution. The tradeoff between FPR and FNR at the end is a social choice and the onus lies on the decision maker to alleviate this. We have to weigh the options of letting dangerous criminals go free against imprisoning innocent people. Comparing all classification parity results from our model, we choose the method that strikes the balance of errors and provides the lowest FPR.

References

[1] Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv, cs.CY.*

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

[3] Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* . (pp. 1-7). IEEE.