Programming Assignment 3: Report

CSE 574: Introduction to Machine Learning, Team #: 34

CHARANTEJA KALVA, 50336873

SAI BHARAT KONAKALLA, 50336876

SOUMITH REDDY CHINTHALAPALLY 50336650

Our Model

- Model Choice: Naïve Bayes with Equal Opportunity
- Algorithm Choice: Naïve Bayes
- Fairness Constraint Method : Equal Opportunity
- Secondary Optimization: Financial Cost
- Total cost: $-758,282,364
- Total accuracy: 0.6282586510808053

Values

- Accuracy for African-American: 0.6220379146919431
- Accuracy for Caucasian: 0.6291910181482621
- Accuracy for Hispanic: 0.6483126110124334
- Accuracy for Other: 0.6390532544378699
- FPR for African-American: 0.4935593220338983
- FPR for Caucasian: 0.4441710243002153
- FPR for Hispanic: 0.39274924471299094
- FPR for Other: 0.40487804878048783
- FNR for African-American: 0.28826933193056287
- FNR for Caucasian: 0.29744693940326056
- FNR for Hispanic: 0.29310344827586204
- FNR for Other: 0.2932330827067669
- TPR for African-American: 0.7117306680694371
- TPR for Caucasian: 0.7025530605967394
- TPR for Hispanic: 0.7068965517241379
- TPR for Other: 0.7067669172932332
- TNR for African-American: 0.5064406779661017
- TNR for Caucasian: 0.5558289756997847
- TNR for Hispanic: 0.607250755287009
- TNR for Other: 0.5951219512195122
- Threshold for African-American: 0.3099999999999994
- Threshold for Caucasian: 0.15999999999999925
- Threshold for Hispanic: 0.08999999999999925
- Threshold for Other: 0.06999999999999926

Our model is capable of applying Post-Processing techniques for any metric (not only for race)

Report Essentials

Motivation for creating a new model to replace COMPAS?

- Considering the results of COMPAS, it is observed that the model is showing bias to some races and to balance out this we created our model to even them. We are using different thresholds for different races instead of using same threshold for every sub-group
- When the algorithm was incorrect COMPAS tended to skew very differently for each of these groups.[2]

- This is the base for our motivation to build a Machine Learning model considering some fairness constraint measures to make a model, which minimizes the biases across any metric. i.e. our model does not lean towards a particular sub-group.

Stakeholders in this situation?

- The Judicial system will be a major stakeholder for our model, hoping that our model will help them to minimize the biases that occur when judged by the jury. This can be achieved by considering the statistical values provided by the ML model for the situation.

What biases might exist in this situation?

- There might be some demographic biases in this situation like race, gender, age etc. But the origin for these biases cannot be tracked, that is biases can be induced at any step in the machine learning pipeline like data, action or feedback.
- The biases might be present in the data and if so, the results of model reflect those biases and setting of threshold value of the model may lead to biases. But the algorithm used by the model will not be responsible for the biases reported in the result of the model.
- It might be possible to verify whether biases are present or not, by verifying values of fairness constraint measures like single threshold (ST) and equal opportunity (EO). If threshold values for ST are similar to threshold values of EO we can say that the biases are minimal.
- In our situation, after observing significant difference in FPR values and the threshold values of ST and EO we came to a conclusion that there are biases in the data

Impact of our proposed solution?

- The main purpose of the any legal model is to provide fair justice and Our ML model strives to achieve equality without discriminate biases, so we opted to use Equal Opportunity
- In this situation of involving criminals, our model strives to maintain similar TPR values i.e. people who were correctly predicted to recidivate, which in turn equalizes the FNR which means people who were incorrectly predicted to not recidivate and then went on to commit another crime. Our first priority is to maintain non-discrimination in the predictions and next priority is to maximize the Financial Cost (as our Secondary Optimization Criteria). This helps us to maximize the TNR which in turn minimises the FPR values.

Our proposed solution a better choice than the alternatives

- As explained in the fairness primer and research papers the definition of fairness is not unique and can vary depending on the situation we are dealing,

Blackstone's ratio:: **Better than ten guilty persons escape than one innocent suffer.**
Benjamin Franklin stated it as: **"it is better 100 guilty Persons should escape than that one innocent Person should suffer".**[1]

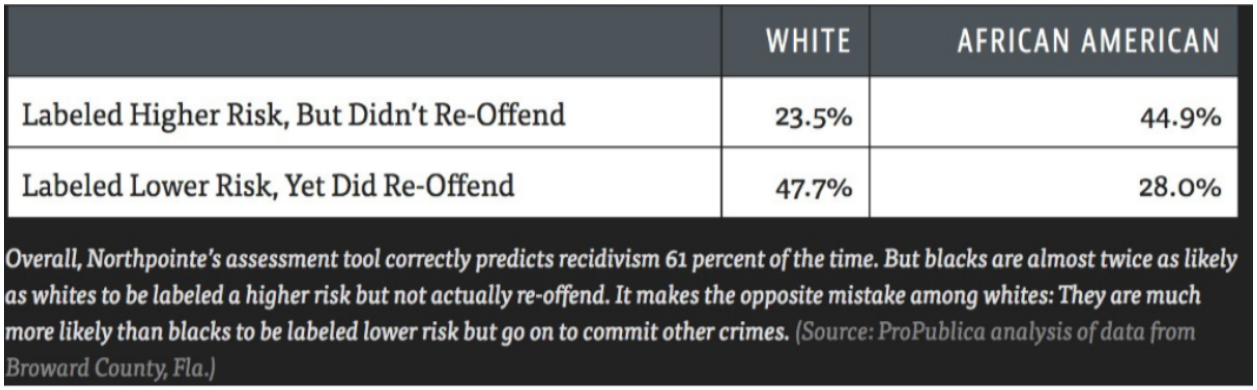
- COMPAS, the algorithm produces much higher false positive rate for black people than white people.[2]
- On the other hand our model thrives to achieve non-discrimination among races.
- Optimizing cost helps our model to decrease the FPR values, which reflects the Benjamin Franklin's statement.

Report Extra Credit

- As humans we can figure out which metrics are sensitive given the situation but the model cannot. So we mention particular metric on which we need to optimize the fairness to eliminate the disparities.
- The assumptions made in the assignment are,
  - The provided list of metrics in the basic model are alone might be sensitive
- The presupposed answers are
  - Race is the major discriminating feature in this situation.
- When it comes to real world, there might be many metrics other than those mentioned in the assignment which are sensitive but cannot be tracked through data
- Our model depends on base rates but the values of base rates are finalized by the fairness constraint methods and data.
- Observing values for different metrics
  - We came to a conclusion that, there are biases across age, gender sub-groups i.e. the model is treating people of age less than 25 are more likely to recidivate than others and also for gender, females are less likely to recidivate than males.

- But if we observe values after applying Equal Opportunity, there were no biases across any metric and TPR, FPR, TNR, FNR values have less difference for all sub-groups compared to other post processing methods. Values used for observations can be found in [5]
- The metrics race, age, gender have significant disparities in the data among the sub-groups.

References and Attachments

1. *https://en.wikipedia.org/wiki/Blackstone%27s_ratio*
2. 

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Figure 1: *COMPAS showing discrimination*

3. The research paper which motivated us to use Equal Opportunity method.
   https://ttic.uchicago.edu/~nati/Publications/HardtPriceSrebro2016.pdf
4. *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*
5. *https://drive.google.com/file/d/1PsoZ9v5p8uXszSrFG11XoAy-6fpXUOH2/view?usp=sharing*