Testing Bias Prevention Techniques on Recidivism Risk Models

Claudia McKenzie, Mathematical and Computational Science, claudi10@stanford.edu

PREDICTING

Risk assessment algorithms are used in the private and public sector to predict human behavior, but recidivism models like the COMPAS have been criticized for producing biased outputs [1]. I used model selection, threshold manipulation, separate models for sensitive characteristics, and alternative labels that may represent less biased observations to mitigate potential racial bias (measured via false positive rates and other metrics) [2]. I found that Random Forest Classifiers produced the best results with the least bias, and using alternative labels made the biggest impact on bias while increasing accuracy.

DATA AND FEATURES

The data for this project comes from the Broward County Florida Corrections Department, When duplicates and delinquent data are removed, the dataset contains 10,179 examples, each corresponding to an arrest case, labeled with whether or not the individual was arrested again in the following 2 years, and whether the re-arrest was for a violent crime.

I used 16 features, which included demographic data, prior criminal records, and categories regarding nature of the crime. Some demographic data was baseline tests, and then removed to better simulate actual risk assessment algorithms. Initially the crime description features were too specific (502 unique). Processing them manually into 32 broadly descriptive categories ("Assault and Battery") significantly improved performance, as well as a variable based on Broward County's own system of separating violent, nonviolent, and drug related crimes.[1]

MODELS

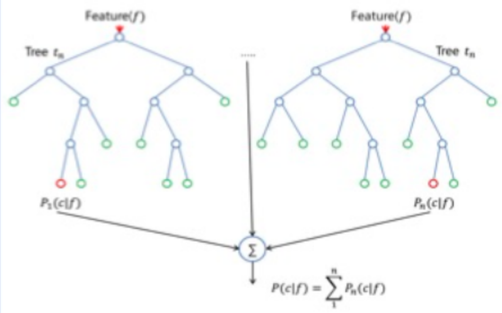
Gradient Boosting with Regularization

$$\underset{\beta}{argmin} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m))$$

Support Vector Classification with Radial Basis Function Kernel

$$K(x,x') = exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$$

Random Forest



Naïve Bayes

$$\hat{y} = \underset{y}{argmax}\ P(y) \prod_{i=1}^{n} P(x_i|y)$$

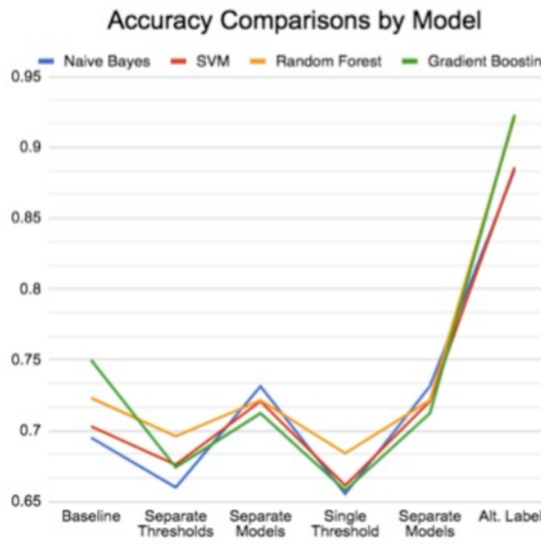
RESULTS

Some interesting results from selected models are shown below. In general, the more complex models performed better, with lower false positive rates, but the techniques performed similarly on each .

Model:	Test acc.	Train acc.	FP Race == AA	FP Race == White
NB baseline	0.6977	0.6965	0.2634	0.1203
NB alt. labels	0.8846	0.8850	0.0514	0.0328
RF alt threshold	0.6301	0.6156	0.1593	0.1638
GB separate	0.7594	0.7397	0.0903	0.0778

DISCUSSION

Using alternative labelling was by far the most effective technique for every model, reducing bias while also increasing accuracy. Manipulating a single threshold, which essentially negated bias by choosing the decision criteria to match the false positive rates, cost the most in terms of accuracy.



FUTURE WORK

Moving forward, I would analyze which features had the greatest effect on model accuracy. Calders and Verwer [2] describe dat manipulation techniques to help mitigate redundant classifiers for sensitive characterist which would be interesting to apply here. I wo also like to do a similar analysis on a risk assessment dataset with greater distributiona overlap (i.e loan applications).

References

[1] S. Corbett-Davies et al. "The Cost of Fairness." *eprint arXiv:1701.08230.* [Online]. Available: https://arxiv.org/pdf/1701.08230.pdf, Jan 2017 [Accessed Dec, 2017].

[2] T. Calders and S. Verwer. (July 2010) "Three naïve Bayes approaches for discrimination-classification." *Data Mining and Knowledge Discovery* [Online]. Vol. (issue), Available: https://link.springer.com/article/10.1007/s10618-010-0190-x [Accessed Dec. 10, 2017].