Testing Bias Prevention Techniques on Recidivism Risk Models

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

1 Introduction

Risk assessment algorithms use machine learning to predict future human behaviors. The models are used in both the private and public sector to predict things like loan eligibility and risk of recidivism. Because these algorithms are legally mandated to be "race blind," they output the same results for two individuals of different races with otherwise identical characteristics, which may lead some to argue that the models are "fair" according to the standard statistical definition. However, especially in the case of recidivism risk models, it is important to acknowledge that the makeup of the datasets used to train these risk-assessment models is affected by the biases present in the American criminal justice system.

Arguably, these biases may manifest themselves in the form of higher false positive rates for African-American individuals who are more likely to be falsely considered a "high-risk" for recidivism in the context of one *ProPublica* analysis of the COMPAS algorithm. This is significant, because false positive rates could result in higher rates of incarceration for African-American individuals with no benefit to public safety. Using false positive/negative rates as my main indicators of bias, I used a Naïve Bayes, SVM, Random Forest, and Gradient Boosting classifiers to test threshold manipulation, race-specific modelling, and alternative labelling as techniques to create a fairer algorithm that input demographic and criminal record data and predicted whether or not an individual would be arrested again within two years, while sacrificing as little accuracy as possible.

2 Related Research

There is a significant amount of recent research related to addressing concerns of fairness and bias in the context of the COMPAS algorithm. "The Cost of Fairness" by Corbett-Davies et al addresses in part the issue of predictive inequality, or unequal accuracy across racial groups. By their research the most effective way to ensure conditional statistical parity or predictive equality is to manipulate the decision threshold of an algorithm, which results in a lower true positive rate. In the context of a recidivism risk model, this could result in a threat to public safety.

As risk assessment algorithms and machine learning classifiers in general become more and more prevalent, the issue of calculating and proactively addressing algorithmic bias or disparate impact has been more widely researched. Toon and Calders provide several approaches in their paper "Three naive Bayes approaches for discrimination-free classification", training separate models, threshold manipulation, and the use of an EM algorithm that leverages the latent variables (ones that were mislabeled by the initial algorithm) to reduce bias. I modified and used the former two approaches across all of the models included my experiment. Building on this approach, in "Fairness-aware Learning through Regularization Approach" Kamishima et al incorporate a regularizer term prejudice into Toon and Calders' algorithm, that is intended to remove unintended based the starting differences between the initial probability distributions of the various racial groups represented in a dataset.

Other approaches to mitigating bias focus on altering the data during pre-processing. In their paper, "An algorithm for removing sensitive information: application to race-independent recidivism prediction", James Johndrow and Kristian Lum provide a novel algorithm for addressing the issue of sampling bias in datasets used to train recidivism algorithms and redundant classifiers for sensitive characteristics like race through variable adjustment. In "Certifying and Removing Disparate Impact", Feldman et al also suggest an approach for un-biasing data by balancing the error rates of naïve models. Additionally, they provide a method for calculating the disparate impact caused by an algorithm using both false positive and false negative rates across racial groups. Their disparate impact calculation inspired me to use false negative rates as another metric when evaluating my own experiment, and with more time and data I would have liked to try and implement their versions of both data processing methods.

My project builds on the notion established in this literature that false positive rates can be indicative of disparate impact, and uses variations of some of approaches put forth here to mitigate that impact while maintaining the highest level of accuracy possible. The general prevalence throughout these papers of using false positive rates and comparative racial accuracy as indicators of potential bias caused

me to include those as some of my main metrics.

3 Data and Features

The data for this project comes from the Corrections Department of Broward County, Florida, and is the same set analyzed *ProPublica*'s COMPAS article (Argwin). It is available on Kaggle courtesy of *ProPublica* who obtained it via a public records request. The dataset initially contained 10,892 examples, each corresponding to an individual who was arrested and booked in the county between year and year. In preprocessing I removed duplicates and examples with missing data. It is possible that the incomplete data was made up of individual with less severe charges or short/non-existent sentences, so it is possible that removing these from the dataset impacted the overall experiment. After preprocessing, I was left with 10,176 unique examples, 7,000 of which were used for the training set, 2,000 for the test set, and 1,176 for the development set. Each of these examples is labeled twice, once if the individual was arrested again within two years of being released (recidivism), and a second time if such an arrest occurs for a violent crime (violent recidivism). Every incident of violent recidivism is also counted as an incident of general recidivism.

Figure 1.
Rates of True Positives for Recidivism Labels

	Recidivism Rate	Violent Recidivism Rate
Overall	33%	8%
Black	39%	9%
White	27%	6%
Hispanic/Latino	25%	6%

Because of privacy restrictions, the features included in this dataset are very different from the features used for training the actual COMPAS risk algorithm, which come from a 137 question survey individuals in the county fill out when they are booked in jail. My experiment used 10 features containing basic demographics, quantitative information about the individual's past criminal record (separated by juvenile and adult), and data corresponding to the current charge against them. I used race as a feature only in my baseline tests, and then removed it before training all of the other models. I also removed the name feature instead of processing it, because in some cases including it would create a redundant classifier for race or sex (e.g. the last Hernandez turns up in 38 examples from the dataset, each with the race feature equal to Hispanic/Latino) (Kilbertus).

Two of features used (charge category and violent, nonviolent, or drug charge) were generated using the Charge Description category from the initial dataset.

Figure 2.
Feature Breakdown by Category

Feature Category	**Features Included**
Demographics	Sex, Age, Race
Past Record	Juvenile felony count, Juvenile other count, Adult priors count
Current Charge	Length of sentence, Charge degree, Charge category, Violent, nonviolent, or drug charge

Initial poor performance caused me to reexamine the Charge Description feature, which initially included 502 unique categories, each representing to a legally distinct charge that corresponded to few examples in the dataset. I manually classified these into 32 broader categories, and saw an improvement in the overall accuracy of the models I was using. Removing descriptors like "aggravated" and quantities of drugs found on an individual runs the risk of overgeneralizing the current charges, but my hope was that the charge degree feature, which stated if the crime committed was classified as a contempt violation, a misdemeanor, or a felony, and of what degree, would mitigate this. I also included a feature that indicated whether the current charge was a violent, non-violent, or drug-related charge, as determined by Broward County's own classification system.

Figure 3.
Most Common Charges

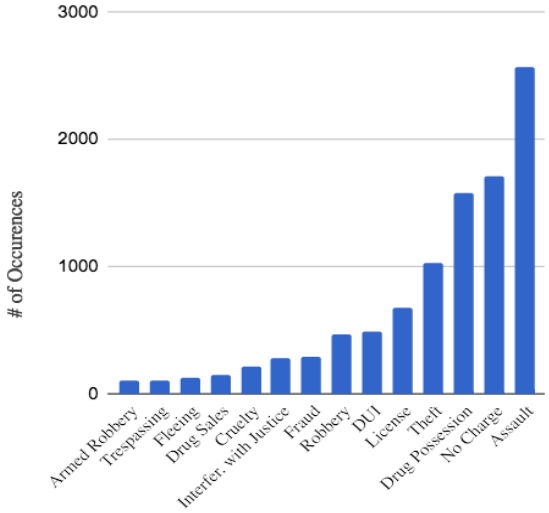


Figure 4.
Prevalence of Charge Types

	Violent charges as % of all charges	Non-Violent	Drug
Overall	**29%**	**54%**	**17%**
Black	27%	56%	17%
White	30%	51%	19%
Hispanic/Latino	31%	54%	13%

As shown above, the majority of the current charges correspond to nonviolent offenses, followed by violent crimes, and drug offenses. By far the most common charge was assault, followed by drug possession, theft, license violation, and driving under the influence. The No Charge category corresponds to individual who were arrested and booked, but no formal charges were filed against them.

4 Methods

I chose to work with several different models because model selection and design can be an effective way of lowering false positive rates (Grgic-Hlaca), and the outcome of different bias-prevention techniques could vary from model to model.

4.1 Naïve Bayes

I chose Naïve Bayes as my basic model because the threshold manipulation and separate model techniques I used in my experiment were drawn from Toon and Calder's "Three naïve bayes approaches for discrimination free classification," and I was interested in seeing how they would perform on the original model when applied to this problem. As mentioned before, Naïve Bayes relies on the strong assumption that each feature distribution is independent from the others. Training the Naïve Bayes is equivalent to maximizing the joint likelihood

$$\prod_{i=1}^{m} p(x^{(i)}, y^{(i)})$$

where x represents the feature vector of some example i and y represents the binary label.

4.2 Support Vector Machine with Radial Basis Function Kernel

The second model I chose was a Support Vector Machine with hinge loss and radial basis function kerne I because of its generally robust performance on nonlinear classification tasks. Training this SVM model is equivalent to minimizing the regularized risk,

$$J_\lambda(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\left[1 - y^{(i)}K^{(i)^T}\alpha\right]_+ + \frac{\lambda}{2}\alpha^T K\alpha$$

where K is the kernel function

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

. SVMs use kernel functions to map the input of a problem into a higher dimension feature space where the data is separable by a hyperplane. In this case, our algorithm used the Radial Basis Function, which is well suited for classification tasks with small datasets and performed better than the polynomial kernel in initial testing.

4.3 Random Forest

Random forest classifiers fit a number of decision tree predictors on independently selected subsets of the data, then combine the predictions from all of the trees into the final model. Random forests allow us to examine which features were most influential in shaping the model, which is useful in this experiment because it allows us to determine whether demographic features, prior record, current charge information, or some combination of the three had the greatest effect on the final model. I was also drawn to of its generally good performance on unbalanced data given the racial disparities in our dataset, and because of its tendency to avoid overfitting.

4.4 Gradient Boosting with Regularization

Gradient boosting seeks to minimize some loss function, in this case, the binomial deviance loss function,
by incrementally improving the model using gradient descent with line search, where the update formula is

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

and γ_m is given by

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

I tried several parameters to regularize my gradient boosting model by restricting the learning rate and the subsample size, before choosing 0.1 and 0.5 respectively.

5 Experiment and Results

For all of the aforementioned models, I made variations either pre- or post-processing with the intention of mitigating the disparity in false positive rates for White and African-American individuals. I trained each model separately, and then used overall and race-specific false positive rate, false negative rate, precision, and accuracy as metrics to compare the results.

5.1 Model Variations

Baseline

For the baseline model, I included all of the features in Figure 2 and made no pre- or post-processing adjustments besides the data processing described in Section 3. For the Random Forest, Gradient Boosting, and SVM classifiers, I used the baseline model and the development set to tune the parameters to improve accuracy while avoiding overfitting. I used the same parameters in all the following variations of the given model.

Race Removed

Although it is sometimes collected, it is generally a violation of the Civil Rights Act to explicitly include race as a feature in risk assessment algorithms. I was interested in seeing how this slightly more realistic model would compare the baseline, especially in terms of race-specific false positive rates. To most accurately simulate the performance of actual risk algorithms, data from all included racial groups was used to train and test all of the models with the exception of the deliberately separate ones. Because white and black individuals make up a large majority of the dataset (), and their false positive rates were the most dramatically different in the baseline models, our race based analysis focuses primarily on these groups.

Threshold Manipulation

This technique is described by both Toon and Calders and Corbett-Davies et al. In one case, post-processing the decision threshold for all examples is increased from the standard 0.5 to a level at which the false positive rates for African-American and White groups are almost equal. In the other, only the threshold for African-American individuals is increased until the group's false positive rate matches that for White individuals using the same model. Similar techniques are described by both Toon and Calders and Corbett-Davies et al.

Race-Specific Models

Another technique described by Toon and Calders, this variation involved splitting the dataset by race and training and testing each model on the separate data. To account for the fact that there are fewer white individuals in this dataset than African-Americans, I normalized the sizes of the race-specific test and training sets for this model.

Alternative Labelling

For this variation, I substituted the violent recidivism labels for the general recidivism ones used for all other models. Using the actual COMPAS algorithm, African-American individuals on this dataset were 77% more likely to be falsely assigned a high risk score when predicting for violent crimes than white individuals, as opposed to 45% for general recidivism, and I was interested to see how my results compared (Angwin).

5.2 Results and Discussion

Figure 5.
Comparative accuracy across all models

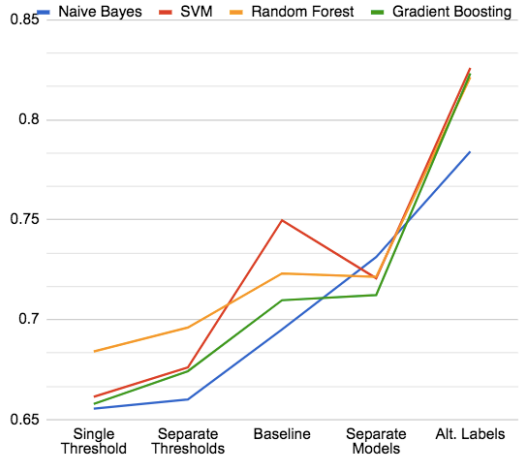


Figure 6.
False positive rates for alternative labelling models

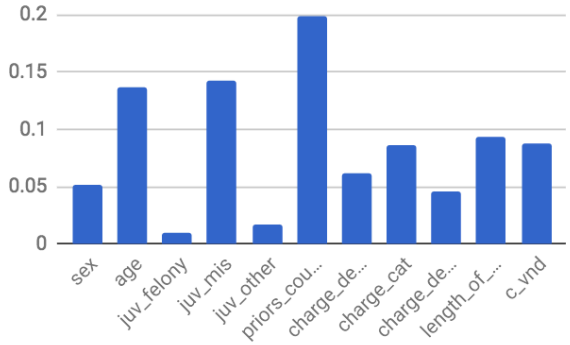
	Naïve Bayes	SVM	Random Forest	Gradient Boosting
Overall	0.09432	0.07821	0.08369	0.08572
African-American	0.12928	0.08937	0.10261	0.09431
White	0.06353	0.06854	0.07834	0.07672

By far, the most successful modification was the use of alternative labels, which both reduced bias across all models and increased accuracy. This intuitively makes sense to some extent, on a national level for minor re-offenses like drug possession or license violations certain groups are more likely to be

arrested than others although the crimes are committed at approximately the same rate (Corbett-Davies), so any patterns found in those re-arrests might be less likely to hold than for a crime like aggravated assault.

The worst accuracy performance came from the threshold manipulation techniques. The threshold manipulation models artificially matched the false positive rates at a given level, which ranged from 0.0765 to 0.1223, and so excelled that metric. However, they had the lowest level of accuracy and positive prediction rate of any of the models. This is to be expected based on Corbett-Davies, because artificially raising the threshold, either for one group or for all, chooses to ignore the true positives near the original decision threshold (0.5), and in doing so sacrifices accuracy in exchange for a lower bias metric. Overall, the model with the best performance was the Support Vector Machine, which had the most accurate baseline test and responded well to most of the modifications.

Figure 7.
Random Forest feature significance averaged over non-baseline plots



Running a feature analysis on each of the Random Forest classifiers indicates that the most influential features were priors_count, length_of_sentence, juvenile_misdemeanor_count, and age. The least influential were juvenile_felony_count (possibly because it contained so few non-zero values) and charge degree, which is interesting because one might expect that a more serious charge degree to indicate a higher chance of recidivism. This indicates that criminal record is generally weighed more heavily than demographic information when the model is optimized, although the demographic data clearly played a role. For the baseline classifier, race was the fifth most influential feature, indicating that including it affected the outcome somewhat but not substantially.

6 Conclusion

This research indicates that race blind algorithms alone are not enough to prevent disparate racial impact in recidivism risk models, and demonstrates that it may be possible to reduce this impact without sacrificing accuracy by examining the potential biases present in the problems that we are trying to solve and the data we use to solve them. It intuitively makes sense that the alternative labels performed the best out of all of the methods. Nationally, crimes like drug possession or license violations are committed at similar rates across racial groups, but the arrest rates for the offenses are skewed toward minority groups. As a result, the feature distribution for who goes on to commit those crimes is likely less representative than the distribution for something like aggravated assault, resulting in a worse prediction. Interestingly, this was at odds with the COMPAS predictions, which had a greater disparity between white and black false positive rates for violent recidivism than for nonviolent recidivism (Larson). This may suggest the possibility that a lower threshold may have been used in their algorithm when predicting violent recidivism. My results indicate that when using machine learning to solve problems,

7 Future Work

Moving forward, I would be interested in experimenting with omitting features (for example, all demographic information), and observing how that impacts the metrics used above. With respect to model selection, I would like to test combinations of the model variations I tried here where possible and compare the results. With more time, I would also work on doing more to vary the actual logic of the algorithms as described in Toon and Calders and Bechavod and Ligett. I would also be interested in acquiring another recidivism dataset with similar labels and features and replicating the experiment to see if the patterns I observed hold. If I had access to the data and resources, I would try modifying the experiment for a similar problem that could be more challenging to classify because of greater distributional overlap between risk groups, like loan eligibility assessment.

8 Acknowledgements:

Thank you to Professors Andrew and Dan, as well as all of the TAs for their help and guidance in conducting this experiment and putting together this report.

9 References:

Corbett-Davies, Sam, et al. "Algorithmic decision making and the cost of fairness." *arXiv preprint arXiv:1701.08230* (2017).

Berk, Richard, et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *arXiv preprint arXiv:1703.09207* (2017).

Calders, Toon, and Sicco Verwer. "Three naive Bayes approaches for discrimination-free classification." *Data Mining and Knowledge Discovery* 21.2 (2010): 277-292.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." *arXiv preprint arXiv:1609.07236* (2016).

Bechavod, Yahav, and Katrina Ligett. "Learning Fair Classifiers: A Regularization-Inspired Approach." *arXiv preprint arXiv:1707.00044* (2017).

Grgic-Hlaca, Nina, et al. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning." (2018).

Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 2012.

Kilbertus, Niki, et al. "Avoiding Discrimination through Causal Reasoning." arXiv preprint arXiv:1706.02744 (2017).

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Angwin, Julia et al. "Machine Bias." *ProPublica*, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Larson, Jeff et al. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*. 23 May 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Jones E, Oliphant E, Peterson P, *et al.* SciPy: Open Source Scientific Tools for Python, 2001, http://www.scipy.org/ [Online; accessed 2017-12-13].

John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.5