Minority Report: ML Fairness in Criminality Prediction

Dominick Lim
djlim@stanford.edu

Torin Rudeen
torinmr@stanford.edu

1. Introduction

1.1. Motivation

Machine learning is used more and more to make decisions which can have a large impact on people's lives, ranging from deciding whether someone will get a job to whether they will be investigated by the police. Given this, just as it is important to ensure that a traditional job application process or police investigation is free of discrimination based on race or gender, it is also important to ensure that a machine learning algorithm is also free of similar discrimination.

One might naively assume that a machine learning algorithm cannot discriminate racially or otherwise, since it is simply a mathematical algorithm. This is true in a trivial sense, but misleading: while an algorithm itself may not discriminate, it can reflect or even magnify discrimination in the data it is trained on. This has the potential to create social feedback loops, where data based on a discriminatory source (i.e. criminal justice data, where racial minorities are disproportionately the subject of police attention [1]) is used to feed an algorithm which will perpetuate this discrimination.

1.2. Related Work

Given this and similar concerns, awareness about potential discrimination in machine learning has grown over the past years. Fairness, Accountability, and Transparency in Machine Learning[1] is an organization that holds a yearly workshop of the same name to explore this topic.

Discrimination-aware classification is concerned with the construction and use of classifiers learned from discriminatory or biased data, and was first introduced by Pedreschi et al. [6]. Kamiran et al. [4] studied a large sample of data on Dutch youth, and found that a naive classifier resulted in astronomically higher false positive rates for members of minority groups than for non-minorities. They presented a number of different algorithms for counteracting this effect, based on changing the training data or the objective function.

1.3. Experimental Setup

In our work, we used a dataset (described fully below) with detailed life-history information about several thousand United States youth, including criminal behavior. To emulate a hypothetical policing application, we set the following prediction task: Given all the data on an individual up through their 24th year, predict whether they would be incarcerated in their 25th year.

We decided to begin by implementing a traditional non-discrimination-aware classifier (simple Naive Bayes). Then, we analyzed the behavior of the classifier on two of the race categories used in the dataset (Black and Non-Black/Non-Hispanic). Finally, we implemented several different approaches to try to make the classifier more fair.

2. Data and Features

2.1. National Longitudinal Survey of Youth 1997

The data used in this study is derived from the National Longitudinal Survey of Youth 1997 (NLSY97) [3] provided by the United State Department of Justice. NLSY97 collected age-based calendar year variables on arrests and incarcerations, self-reported criminal activity, substance use, demographic variables and relevant variables from other domains on individuals each year for up to 12 years (ages 14-25).

NLSY97 was created to be representative of United States youth in 1997 who were born between the years of 1980 and 1984. The NLSY97 cohort comprises two independent probability samples: a cross-sectional sample and an oversample of Black and/or Hispanic or Latino respondents. The cohort was selected using these two samples to meet the survey design requirement of providing sufficient numbers of Black and Hispanic or Latino respondents for statistical analysis.

In this study, a dataset was created from the 2,977 individuals in a subsample of NLSY97 who were asked self-reported illegal activity questions in the calendar year that they turned 20 and beyond (not necessarily the same calendar year for every person in the study). We chose this subsample of NLSY97 because people not in this subsample were missing answers to these self-reported illegal activity questions.

2.2. Features

The dataset contained a total of 350 features, including:

- Demographic data (race, gender, etc.)

[1] https://www.fatml.org

- Health information
- Information about family relations
- Socioeconomic status
- Income and employment information for each year
- Educational information
- Self-reported information on various types of crimes for each year
- Arrests, convictions, and incarcerations for each year.

2.2.1. Preprocessing. We began by removing individuals who passed away before the end of the study or did not report data in the final year of the study.

Next, we found that while some of the data were categorical (e.g.: race/ethnicity, sex), other data were not (e.g.: household income, poverty ratio, test scores). Therefore, non-categorical data were bucketed into five buckets of equal size whose ranges were determined by the minimum and maximum values of the attribute across all individuals in the dataset. We also replaced all missing values in the dataset with the constant, -1.

After cleaning, the dataset contained 2,814 individuals. The racial breakdown of these individuals is 30.2% Black, 21.3% Hispanic/Latino, 1.1% Mixed (Non-Hispanic), 47.4% Non-Black/Non-Hispanic. 6.2% of the individuals were incarcerated in the year they turned 25.

3. Analyzing a Simple Model

We trained Naive Bayes on the training set, consisting of 70% of the data ($N = 1969$). We then ran prediction on the training set and on a test set consisting of the remaining 30% of the data ($N = 845$). As an initial check to see whether or not our algorithm was overfitting, we compared precision and recall on the training set versus the test set. On the training set, precision was 26.5% and recall was 77.2%. On the test set, precision was 28.4% and recall was 70.0%. The similarity between these numbers gave us confidence that the Naive Bayes model was not overfitting the data.[2]

We then began examining the performance of the algorithm on subjects of different races. To begin with, we looked at the overall rate of positives (i.e. instances where the subject was incarcerated during their 25th calendar year). The table below shows the rate of positive values in the data, and in the predictions made by the algorithm.

	Black	Non-Black/Non-Hispanic
Ground-truth positive	15%	4%
Predicted positive	27%	12%

Overall, Black subjects are substantially more likely to be incarcerated than Non-Black/Non-Hispanic subjects, and the algorithm's predictions are in line with this. Furthermore, the algorithm predicts positive at a rate 2-3 times higher than the ground truth, both overall and in every race category.

	Black	Non-Black/Non-Hispanic
Precision	0.400	0.1837
Recall	0.737	0.600
False Positive	0.195	0.102
False Negative	0.263	0.400

The table above shows a variety of measures we computed to better understand the performance of the algorithm on different races. Focusing on comparing Black to Non-Black/Non-Hispanic performance, we found a very surprising result here: Both precision and recall were substantially higher for Black subjects, with the difference being particularly large for precision. In other words, a Black predicted "criminal"[3] was more than twice as likely to be an actual "criminal" as a Non-Black/Non-Hispanic predicted criminal.

Looking at the data differently, however, showed a seemingly contradictory result: the false positive rate[4] for Black subjects was also significantly higher than for Non-Black/Non-Hispanics. In other words, an "innocent" (i.e. ground truth negative) Black subject was twice as likely to be falsely accused by the algorithm as an "innocent" Non-Black/Non-Hispanic subject.

At first, these two findings seem mutually contradictory. The first seems to suggest the algorithm is less likely to falsely accuse a Black subject, the second that it is more likely. The answer to this paradox lies in the different base rates between the two groups. The fraction of true positives is greater for Blacks than for Non-Black/Non-Hispanics, and the number of predicted positives is also greater, but not by as much. So, while the number of false positives is low compared to the number of predicted positives (high precision), it is still higher when compared to the population of true negatives.

We think there is an interesting and important lesson here. Take the example of NYPD's much criticized "stop and frisk" policy. Critics say that the stop and frisk has a huge disparate impact on minorities, leading to huge numbers of innocent Black and Hispanic people being stopped, and disproportionately arrested for minor crimes (like possession of small amounts of marijuana). Meanwhile, defenders[5] claim that stops only disproportionately impact minorities because they mirror actual patterns of violent crime in New York City, and that the minority suspects they stop are not more likely to be innocent than the non-minority suspects. Our results, if applied more broadly, suggest that both can be correct in a mathematical sense: When a police advocate says stop and frisk is not biased, they are saying that the "algorithm" has high precision. When a minority citizen or advocate says stop and frisk is biased, they are saying it has a high false positive rate. As we have shown, it is quite possible, and even natural, for both to be the case.

[2] Which we expected, since Naive Bayes is a relatively high bias (in the ML sense), low variance model.

[3] That is, someone the algorithm thought likely to be incarcerated in their 25th year, based on knowledge about previous years.

[4] I.e. the percentage of ground truth negative examples for which the algorithm predicted positive.

[5] Like former mayor Michael Bloomberg in this editorial: [2]

3.1. False Positives as a Fairness Criterion

Given this analysis, we decided it may be fruitful to look at the false positive rate as an alternative criterion for fairness, as opposed to just looking at the prediction rate. Two reasons this may be a good fairness criterion are that it corresponds to a real world phenomenon which is universally regarded as unfair, and that it is less likely to be criticized as "reverse discrimination" than an algorithm which simply equalizes the prediction rates between different classes.

In the rest of this paper, we will compare various methods which attempt to decrease the false positive rate on the minority class to match that on the majority class. Specifically, we will look at methods which change the algorithm's behavior only for the minority class, leaving it unchanged for the majority class. Since it is trivially easy to decrease the false positive rate on an algorithm by simply making it predict positive less often, we will benchmark our various approaches by the false negative rates[6] they incur. In other words, we will consider an algorithm to be better than another if it decreases the false positive rate for minorities to the same leve as the majority class at the expense of a smaller increase in the false negative rate than the other algorithm.

4. Threshold-Based Fairness

We first implemented the method described by Kamiran et al. [4] that performed best on their dataset, where the decision boundary is moved for the minority class (Black). Still using Naive Bayes, we moved the decision threshold back and forth to generate a variety of different models, and then picked the one which created a false positive rate (on the training set) for Black subjects closest to that for the majority class (Non-Black/Non-Hispanic). This gave the following results:

	Black	Non-Black/Non-Hispanic
False Positive	0.0930	0.102
False Negative	0.526	0.400

So, the false positive rate was cut in half, but at the expense of doubling the false negative rate. Also see Figures 1a and 1b, which show the false positive and false negative rates for different thresholds, on the training and test sets. These graphs show that the threshold method generalized fairly well, with performance on the test set being very close to performance on the training set.

5. Feature Selection-Based Fairness

We next implemented a method of our own devising, which worked by feature selection in the form of **stepwise regression**. Forward feature selection builds a feature set by greedily adding the best feature not already in the set at each step until there are no more features to add. Similarly, backward feature elimination removes the worst feature in the set at each step until there are no more features in the set. We wanted to create a feature selection (or elimination) criterion which would attempt to reach a target false positive rate, while keeping false negatives as low as possible.

We recognized that this was similar to a constrained optimization problem, so we decided to use a **penalty method** to convert the constrained optimization problem into an unconstrained optimization problem by adding a penalty term equal to the square of the deviation from the desired constraint. This gave us the following cost function:

$$FN_B + \gamma(FP_B - FP_T)^2$$

Where FN_B and FP_B are the false negative and false positive rates of the predictions on Black subjects, FP_T is the target false positive rate, and γ is a hyper parameter: larger γ means more weight is placed on reaching the desired false positive goal.

For all of the results below we used the following scheme: train Naive Bayes on the training dataset, and then predict on the test set, using only the selected features if predicting on a Black subject, but using all features if predicting on a Non-Black/Non-Hispanic subject. This doesn't require training separate models, since in Naive Bayes the learned weights for each feature are independent of the other features. We chose this approach because, like the thresholding method, this algorithm is tweaked only for the minority class; thus, the false positive rate on Non-Black/Non-Hispanic subjects will not shift during the optimization process.

5.1. Choosing γ

We wanted to prioritize minimizing the difference between FP_B and FP_T when they are not "close enough".

We reasoned that the difference between false positive rates can be at most 0.1 (10%) for the false positive rates to still be deemed "close". So when FP_B and FP_T differ by 0.1 or more, $\gamma(FP_B - FP_T)^2$ should be greater than any value that FN_B can take on. By definition, no error rate can be greater than 1. Therefore, γ must satisfy the following inequality:

$$\gamma(0.1)^2 \geq 1$$

$$\gamma \geq 100$$

We also reasoned that if the difference between false positive rates is 0.001 (0.1%) or less, we should care an order of magnitude less about making the false positive rates closer than we do about minimizing FN_B. Since 0.1 is a respectable false negative rate (and 0.01 is an order of magnitude less than 0.1), this sentiment is captured by the following inequality:

$$\gamma(0.001)^2 \leq 0.01$$

$$\gamma \leq 10000$$

Therefore, we ran the algorithms described below with the following values of γ: 100, 300, 1000, 3000, 10000. We

[6] I.e. percentage of ground-truth positives predicted negative.

selected the feature subset that minimized the cost function for each of the runs (values of γ). Then, we computed a normalized cost function to select the best overall feature subset, where our normalized cost function was the cost function using the largest value of γ that we tried (10000):

$$FN_B + 10000(FP_B - FP_T)^2$$

5.2. Selecting The Best Feature Subset

First, we trained a Naive Bayes model on our training set and used the model to predict on the cross-validation set.[7] Next, we used the predictions to calculate our FP_T, the false positive rate of Non-Black/Non-Hispanic subjects in the cross-validation set.

At each step of our stepwise regression algorithms, for each feature subset considered, we trained a Naive Bayes model on the training set using only the given feature subset. Then, we evaluated each of the models by using it to predict on the Black subjects in our cross-validation set and calculate FN_B and FP_B for our cost function. Next, we added (or removed) the feature that minimized our cost function and moved on to the next step.

Finally, after all of the steps were completed, we chose the feature subset that minimized the cost function over all steps as the best feature subset.

5.3. Forward Feature Selection

At each step in our forward feature selection algorithm, we add the feature f to the current feature set $\mathcal{F}$ which satisfies the following equation (on the cross-validation set):

$$\arg\min_{f \notin \mathcal{F}} \Big(FN_B(\mathcal{F} \cup f) + \gamma(FP_B(\mathcal{F} \cup f) - FP_T)^2 \Big)$$

This approach generated the false positive and false negative values shown in Figures 1c and 1d. We chose the feature subset that had the lowest cost when evaluated on the cross-validation set. In the event that multiple feature subsets had the same cost, we chose the smallest subset. This gave the following results on the test set (using the original feature set for prediction on Non-Black/Non-Hispanic subjects):

	Black	Non-Black/Non-Hispanic
False Positive	0.0744	0.102
False Negative	0.421	0.400

5.4. Backward Feature Elimination

Naturally, we also implemented backward feature elimination on the same objective, at each step removing the feature f from $\mathcal{F}$ satisfying:

$$\arg\min_{f \in \mathcal{F}} \Big(FN_B(\mathcal{F} \setminus f) + \gamma(FP_B(\mathcal{F} \setminus f) - FP_T)^2 \Big)$$

This generated the false positive and false negative values shown in Figures 1e and 1f,[8] Again, we chose the feature subset that had the lowest cost when evaluated on the cross-validation set. However, this time, we picked the largest subset size (fewest features removed) if multiple feature subsets had the same cost. This gave the following results on the test set (again using the original feature set for prediction on Non-Black/Non-Hispanic subjects):

	Black	Non-Black/Non-Hispanic
False Positive	0.0698	0.102
False Negative	0.368	0.400

As the graphs on the next page show, both forward feature selection and backward feature elimination led to a substantial degree of overfitting: false negative rates especially were far lower on the cross-validation set (performance on which was used to select the feature subsets) than on the test set. Fortunately, the performance on the test set was still superior to thresholding for both kinds of feature selection, so the technique did in fact generalize in a meaningful way.

One method we used to limit the degree of overfitting from feature selection was to choose the first encountered feature subset size which met the false positive criterion (i.e. the smallest subset size for forward feature selection, and the largest subset size for backward). As discussed in [5], feature selection can lead to overfitting by the sheer number of different possible feature selections it can explore, and so terminating the algorithm as early as possible can reduce the degree of overfitting. This can be seen clearly in Figures 1e and 1f: on the cross-validation set, our chosen point (denoted by the green line) is at the rightmost edge of a series of feature subsets with equivalent performance on the test set. However, we can see that the chosen subset size had much better generalization error than the smaller subset sizes. A similar but less dramatic effect can be seen on the forward feature selection plots.

6. Conclusion and Future Work

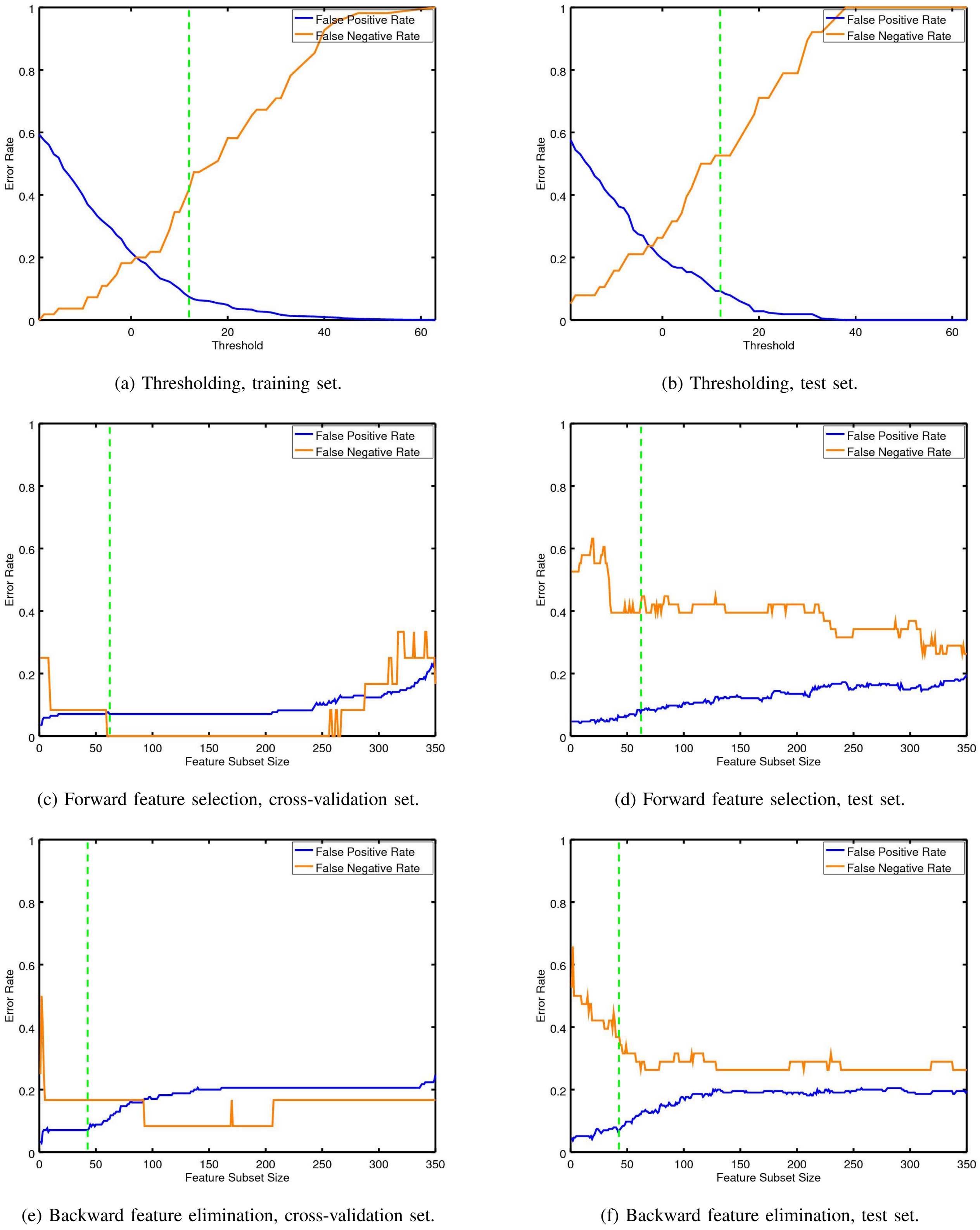
We examined the performance of Naive Bayes in a criminal justice application on a real-world dataset, and found that it produced substantially disparate false positive rates for different racial groups. Then, we applied several techniques to try to equalize the false positive rate: a simple thresholding technique, and then both forward feature selection and backward feature elimination.

We found that although feature selection introduced more overfitting than thresholding, it also produced substantially better performance on the test set, as measured by the false negative rate for Black subjects once the false positive rate had been equalized.

In the future, we would like to extend this work to see if similar feature selection methods are effective when used with more sophisticated algorithms, such as boosting or SVMs.

[7] For the feature-selection section of this paper, we divided the non-test data into a training set (70% of the original training set) and a cross-validation set (30% of the original training set).

[8] Note that these graphs should be read from right to left, since the algorithm was taking features away rather than adding them.



(a) Thresholding, training set.

(b) Thresholding, test set.

(c) Forward feature selection, cross-validation set.

(d) Forward feature selection, test set.

(e) Backward feature elimination, cross-validation set.

(f) Backward feature elimination, test set.

Figure 1: False positive and false negative rates (on Black subjects) for various algorithms.

References

[1] M. M. Alexander and M. Alex, The new Jim Crow: Mass incarceration in the age of colorblindness, 2nd ed. Jackson, TN: Distributed by Perseus Distribution, 2010.

[2] M. R. Bloomberg, “Michael Bloomberg: ‘Stop and frisk keeps New York safe,” in Washington Post, Washington Post, 2013. [Online]. Available: https://www.washingtonpost.com/opinions/michael-bloomberg-stop-and-frisk-keeps-new-york-safe/2013/08/18/8d4cd8c4-06cf-11e3-9259-e2aafe5a5f84_story.html. Accessed: Nov. 22, 2016.

[3] Bureau of Labor Statistics, U.S. Department of Labor, “National Longitudinal Survey of Youth 1997”. Produced by the *National Opinion Research Center, the University of Chicago* and distributed by the *Center for Human Resource Research, The Ohio State University*, 2013.

[4] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan, “Classifying Socially Sensitive Data Without Discrimination: An Analysis of a Crime Suspect Dataset” in *2012 IEEE 12th International Conference on Data Mining Workshops*, 2012.

[5] Loughrey, John, and Pdraig Cunningham. ”Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search.” Proceedings of the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. 2005.

[6] D. Pedreschi, S. Ruggieri, and F. Turini, “Discrimination-aware data mining” in *Proc. of ACM SIGKDD Conference on Knowledge and Data Mining*, 2008.