# Balancing Classifier Fairness with Public Safety in Traffic Stops CS 229 Final Project

**Vikul Gupta** vikulg@stanford.edu
Stanford University, Department of Computer Science

**Kuhan Jeyapagrasan** kuhanj@stanford.edu
Stanford University, Department of Computational and Mathematical Engineering

**Jaydeep Singh** jaydeeps@stanford.edu
Stanford University, Department of Mathematics

## 1. Introduction

The issue of algorithmic fairness has recently come to the forefront of machine learning, as classifiers increasingly propose decision rules in applications ranging from loan approval, criminal risk estimation, and job application review. In particular, the authors in (Dwork et al., 2011) and (Zemel et al., 2013) explore how different social expectations for group and individual fairness manifest as constraints on learned decision rules, and how such formulations suggest an inevitable accuracy/fairness trade-off in classification.

This issue of fairness has been explored at length by (Corbett-Davies et al., 2017), in which the authors study racial disparities in algorithms assigning risk levels to defendants awaiting trial. Our group develops this analysis in the context of predictive policing, drawing data from the Stanford Open Policing Project to analyze the trade-off between public safety and fairness in police traffic stops (5harad, 2017). We apply Naive Bayes and Logistic Regression to the analysis and prediction of traffic stop outcomes in Connecticut traffic stop data, given categorical and quantitative features as input. The resulting model is shown to have poor performance on fairness metrics - to address this, we apply regularization during learning alongside threshold post-processing, to promote logistic classification compliant with statistical parity and predictive equality fairness constraints. We show that while misclassification accuracy and loss increases are a necessary result of fairness optimization, for our dataset this optimization can be done with proportionally small increase in classification error.

## 2. Related Work

Authors in (Corbett-Davies et al., 2017) starkly illustrate the theoretical and practical challenges underlying algorithmic fairness. Working with the infamous COMPAS prisoner risk-prediction dataset, the authors observe how unconstrained algorithms both reflect and amplify discrepancies between sensitive groups - in this case, race - within a dataset, leading to classifiers that violate social notions of fairness. In particular, our study applies the group fairness metrics outlined in this original study.

Authors in (Adler et al., 2016) introduce a framework for quantifying the reliance of classifiers on a given sensitive attribute, and argue for the importance of studying Balanced Error Rates (as opposed to classification accuracy alone) in studying fair classification problems. The authors also suggest a method, "Gradient Feature Auditing," with applications to dataset manipulation in pursuance of fairness. However, for real-world datasets such as police traffic stop data, we find such direct data-level intervention insufficiently general.

Authors in (Zemel et al., 2013) and (Berk et al., 2017) explore the challenging problem of rectifying discriminatory classification. Both papers avoid dataset manipulation, preferring to intervene during the learning process via regularizers. While (Zemel et al., 2013) aims to produce a compressed representation of data compliant with statistical parity, (Berk et al., 2017) more directly promotes equity in classifier prediction across sensitive groups, in the form of dual individual/group regularization. The latter approach stems most directly from conceptualizations of fairness as "treating similar individuals similarly," as defined in the seminal paper (Dwork et al., 2011), and therefore is most generalizable to a variety of fairness tasks.

We adopt the approach of (Berk et al., 2017) for the study of the Stanford Open Policing dataset. The original analysis of the dataset observed variances in stop rates across driver race, gender, and age (Pierson et al., 2017). Applying the regularization strategies of (Berk et al., 2017), we work to rectify stop rate and false positive imbalances in classifiers trained on this dataset.

## 3. Dataset

For analysis, we drew data from the Stanford Open Policing Project (5harad, 2017). We selected Connecticut data for its rich and complete feature set, yielding 318, 669 raw examples. Relevant features included county identification code (”fips code”), driver race, age, gender, violation type, and stop violation. We binned the categorical variables into separate features using indicator variables, yielding 28 total features. We chose to remove training examples with missing fields, and normalize features to be mean zero, unit variance, yielding a training set size of 50, 000, cross validation set of size 10, 000 and test set of size 10, 000. The target variable is the traffic stop outcome, which we binned into a positive class (indicative of driver given ”ticket”, ”summons”, or ”arrest”) and negative class (driver given ”verbal warning” or ”written warning”. Overall, the positive class comprises 76.2% of the training set, and 76.0% of the test set, indicating a class imbalance whose effects must be taken into account during analysis.

## 4. Preliminary Data Analysis

### 4.1. Understanding the Dataset

After repeating some of the analysis done in (Pierson et al., 2017) to gain an intuition for the data, we performed the following additional analysis on how various features are affected by race and gender. Observe that all y-axes in Figures 1 and 2 are either normalized values for race or gender (the populations were obtained from 2010 Census Data). First, we analyzed stop outcomes, as shown in Figure 1. We see that the proportion of the Hispanic population that is arrested (0.254%), is equivalent to that of the Black population and greater than that of the White population, despite the Hispanic population being stopped less frequently. Meanwhile, the White population receives far more warnings than the Black and Hispanic populations (disproportionate to the proportion of total stops). Second, we analyzed stop durations, as shown in Figure 2. We see that the Black and Hispanic populations are stopped for much longer periods of time than White populations. For instance, 0.237% of Hispanics and 0.278% of Blacks are stopped for more than 30 minutes, whereas only 0.131% of Whites are stopped for that long.

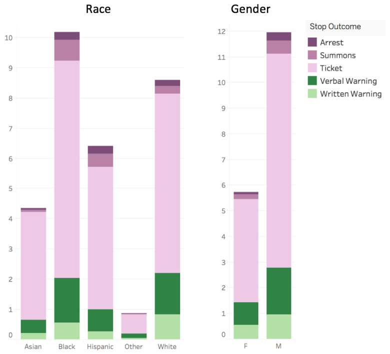


*Figure 1.* Stop Outcomes by Race and Gender

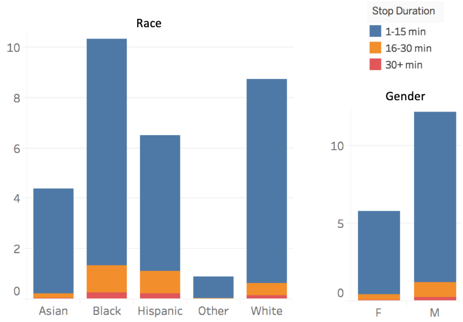


*Figure 2.* Stop Durations by Race and Gender

### 4.2. Feature Analysis

To gain further insight into the features, we ran PCA in 4 different ways: (1) all observations, all features, and the stop outcome, (2) all observations and all features, (3) observations where the stop outcome is 1 and all features, and (4) observations where the stop outcome is 1 and all features. In all 4 cases, the first principal component explained no more than 13% of vari-

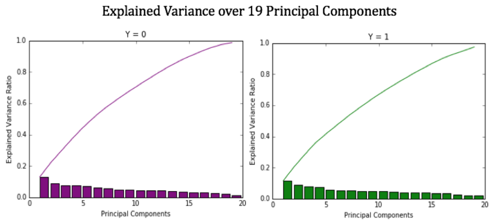


*Figure 3.* PCA Results on Training Dataset Given Y

ance, suggesting feature generation using PCA would be undesirable. The latter 2 tests took 20 principal components for the cumulative explained variance to be greater than 99% (Figure 4.2). This observation suggests that each feature $x_i$ given y is relatively independent of the other features $x_j$ given y, justifying the Naive Bayes assumption.

# 5. Methods

## 5.1. Quantifying Fairness and Accuracy

Following authors in (Corbett-Davies et al., 2017), we assess the fairness of a model with respect to a sensitive group $S$ with class labels $\{S_i\}$ by the metrics of statistical parity and predictive equality. Define $d : \mathbb{R}^p \mapsto [0, 1]$ where $d(x)$, the decision rule is the probability that action $a_1$ is taken. Define $g(x)$ to be the group to which the individual with features $x$ belongs. Then, we have statistical parity defined as

$$E[d(X)|g(X)] = E[d(X)]$$

and predictive equality defined as

$$E[d(X)|Y = 0, g(X)] = E[d(X)|Y = 0].$$

If a model predicts false positive rates $\mathrm{FP}_i$ and positive classification rate $\mathrm{PCR}_i$ for label $S_i$, statistical parity requires the overall variance $\mathrm{var}(\mathrm{PCR}_i)$ to be small, and predictive equality requires $\mathrm{var}(\mathrm{FP}_i)$ small. Intuitively, statistical parity requires a fair model to predict that individuals of varying races, genders, and ages are equally likely to be in the positive class, and predictive equality that the accuracy of the model be fairly constant across sensitive labels.

Recognizing data imbalance between positive and negative classes, we choose to judge classification accuracy via the balanced error rate (BER), which more heavily weights classifier error on minority classes. If a model has overall confusion matrix parameters $FP$, $TP$, $FN$, $TN$, the balanced error is defined as

$$\mathrm{BER} = \frac{1}{2}\left(\frac{FP}{FP + TN} + \frac{FN}{FN + TP}\right).$$

## 5.2. Model Selection and Baseline

We created a two-class classification problem from our dataset (determining the outcome of traffic stops based on the other available data). Due to its ubiquity and frequent usage in industry, we chose no-regularization logistic regression as our baseline model. The other reason for doing so was to be able to test fairness issues as they would appear in common classification problems via regularization (in the form of individual and group penalty), and post-regularization threshold manipulation. For training, mini-batch gradient descent with a batch size of 20 was used to update coefficients. Mini-batch gradient descent was chosen due its ability to benefit from the robustness of stochastic gradient descent and the efficiency of batch gradient descent. This process involved iteratively updating the parameters (the coefficients of the input vector) by subtracting a multiple of the mini-batch loss function's gradient. Once training was conducted, testing involved inputting the dot product of the optimal coefficients and the input vector to the sigmoid function, whose output is a probability from 0 to 1. To determine positive or negative classification, this output was compared with the threshold value of 0.5.

We also tested a Naive Bayes baseline due to the results of our exploratory data analysis, which showed that for both classification classes, training data showed no signs of correlation, thus confirming the Naive Bayes assumption that the predictor features are independent given their prediction class. We used a Multi-variate Bernoulli Naive Bayes model, as all of our feature vectors were binary (all our predictor features were indicator variables). The Naive Bayes model is trained using Maximum Likelihood Estimation (as it has a closed form for Naive Bayes). For testing, we calculated the probability of each class by multiplying the probabilities of the individual features given the class - whichever class had the highest overall probability was the predicted class for any given traffic stop.

## 5.3. Fairness Regularization

The authors in (Berk et al., 2017) build on notions of fairness given in (Dwork et al., 2011), in which fairness is broadly captured as equitable treatment of individuals from different protected classes (e.g. race,gender, or age), conditioned on the individuals' stop outcome. Observe that both statistical parity and predictive equality are group level fairness constraints, requiring only that discrimination is negligible averaged over a sensitive label. Thus the authors distinguish between group fairness and individual fairness (the latter being more stringent), and propose regularizers of use for both purposes. Let $\theta_i$ be model parameters, $S_i$, $i = 1, \ldots, n$ be class labels for a protected class, and $x_i \in S_i$ be feature sets. Then we define the regularizers

$$L_{\text{individual}} = \frac{1}{|S_1|...|S_n|} \sum_{\substack{x_i \in S_i \\ x_j \in S_j \\ y(x_i) = y(x_j) \\ i < j}} (\theta \cdot x_i - \theta \cdot x_j)^2$$

$$L_{\text{group}} = \left( \frac{1}{|S_1|...|S_n|} \sum_{\substack{x_i \in S_i \\ x_j \in S_j \\ y(x_i)=y(x_j) \\ i<j}} \theta \cdot x_i - \theta \cdot x_j \right)^2$$

Yielding the following overall logistic loss function:

$$L_{\text{overall}} = L_{\text{logistic}} + \mu L_{\text{individual}} + \nu L_{\text{group}} + \lambda \|\theta\|^2$$

where $\mu$,$\nu$, and $\lambda$ are hyperparameters that require tuning. The authors observe that these loss functions are convex, and thus can be optimized using gradient descent.

Moreover, as in (Berk et al., 2017), we define a price of fairness (PoF) as follows: for a given $\alpha \in [0, 1]$, let $L_{\text{logistic}}(\theta^\star)$ be the unconstrained, optimal value of logistic loss, and for either individual or group loss, let $\theta_\alpha$ be the model minimizing logistic loss, subject to $L_{\text{individual/group}}(\theta_\alpha) \leq \alpha L_{\text{individual/group}}(\theta^\star)$. Then the ratio of logistic loss increase at a given $\alpha$ value intuitively corresponds to the "price" of demanding a degree of fairness. Plotting this ratio across varying $\alpha$ yields a PoF plot. Observe by construction the graph is downward sloped as a function of $\alpha$.

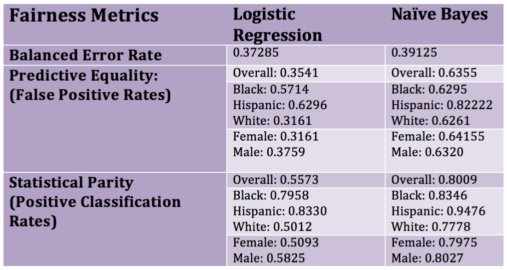
### 5.4. Post-training Threshold Regularization

Once the loss function has been defined and the model trained, one additional hyperparameter can be tuned: the threshold value. Using the BER provides the same threshold for all groups. However, as presented in (Corbett-Davies et al., 2017), each group must have its own threshold value to optimize for fairness. Thus, we compute two separate sets of threshold values given the training set, the parameters, and the BER threshold. For statistical parity, the overall positive classification rate is computed using the given values, and each group's threshold is computed such that that group's positive classification rate is the same as the overall rate. Predictive equality follows similarly except that we set the overall and group false positive rates equal.

## 6. Results

### 6.1. Baseline Results

The below table displays how our baseline models fare on the two measures of model fairness and their balanced error rates. The higher false positive and positive classification rates for the race feature in both models suggest violations of statistical parity and predictive equality. The discrepancies are not as stark across age groups and genders.



| Fairness Metrics | Logistic Regression | Naïve Bayes |
|---|---|---|
| Balanced Error Rate | 0.37285 | 0.39125 |
| Predictive Equality:<br>(False Positive Rates) | Overall: 0.3541 | Overall: 0.6355 |
| | Black: 0.5714<br>Hispanic: 0.6296<br>White: 0.3161 | Black: 0.6295<br>Hispanic: 0.82222<br>White: 0.6261 |
| | Female: 0.3161<br>Male: 0.3759 | Female: 0.64155<br>Male: 0.6320 |
| Statistical Parity<br>(Positive Classification Rates) | Overall: 0.5573 | Overall: 0.8009 |
| | Black: 0.7958<br>Hispanic: 0.8330<br>White: 0.5012 | Black: 0.8346<br>Hispanic: 0.9476<br>White: 0.7778 |
| | Female: 0.5093<br>Male: 0.5825 | Female: 0.7975<br>Male: 0.8027 |

Table 1: Baseline Balanced Error, Statistical Parity, and Predictive Equality Results for Logistic Regression and Naive Bayes baselines

### 6.2. Fairness Regularization Results

We implemented mini-batch gradient descent with the composite loss function consisting of logistic loss, individual/group loss, and L2 regularized loss. We used a mini-batch size of 20 to provide stable, efficiently computable gradients for the loss function, with enough members of each protected class label in a given batch for the impact of fairness regularizers to be observable. We set a learning rate of 2 (reduced from higher values, after finding difficulty balancing the competing effects of different losses), with a penalty factor of .5 applied to the learning rate every $2,000$ iterations to promote convergence. To evaluate the effects of various hyper parameter settings, we ran approximately 6 iterations with only L2 regularization, and approximately 150 trials with regularization, with sensitive features ranging across age, gender, and race. We note that while most models display comparable loss on the training and cross validation set, indicating good generalizability and little overfitting, some models achieve lower loss on the cross validation set, suggesting the optimization algorithm may not be exploring all regions of parameter space. Future study should focus on optimizing the algorithms for efficiently optimizing these fairness regularizers.

With the resulting model parameters, we computed the optimal threshold for positive class identification by selecting the threshold minimizing the Balanced Error Rate on the cross validation set. This threshold then yielded the model predictions on the test set, from which we computed the variance in FPR and PCR across sensitive attributes, as compared to a logistic regression baseline. Interestingly, for age, only about 11.9% of trials reduced the variance in FPR and PCR, 10.7 for gender, and for age 16% reduced FPR, while 100% reduced PCR. These results suggest our loss function requires more careful tuning of hyper parameters to determine the optimal range, with particular choices required for each sensitive attribute. The trials that successfully reduced rates are displayed graphically in figures (with L2 regularization set to

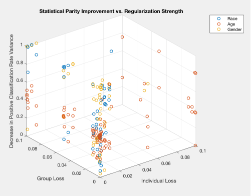


*Figure 4.* Improvement on statistical parity metric with respect to baseline, over range of hyper parameter settings

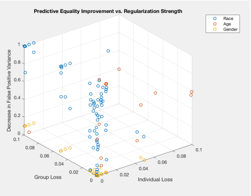


*Figure 5.* Improvement on predictive equality metric with respect to baseline, over range of hyper parameter settings

Price of Fairness Diagrams

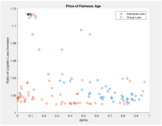


*Figure 6.* Age

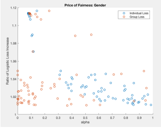


*Figure 7.* Gender

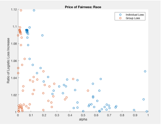


*Figure 8.* Race

zero, for simplicity) 4 and 5. One can immediately observe that increasing group and individual penalties together were responsible for the most improvement in fairness, decreasing FP variance and PCR variance on a gradient up to 100%. We note that the extremes correspond to models outputting identical results for each training example (and therefore uniformly fair), so in practice appropriate intermediate penalties should be taken.

Following the discussion above, we compute the Price of Fairness graphs for regularized regression, displaying the results in 6, 7, and 8. Race most starkly displays the loss decreasing as a function of alpha, indicating there is the highest price to pay in accuracy if one enforces race-based constraints on our data.

### 6.3. Threshold Regularization Results

Once the optimal parameters had been determined for this wide set of hyperparameter choices, the thresholds for statistical parity (SP) and predictive equality (PE) were calculated for a select few choices. The following data (6.3) displays the false positive (FP), positive classification (PC), and balanced error (BE) rates for $u = v = \lambda = 0$ for the BER, SP, and PE thresholds. Observe that all false positive and positive classification rates are lower for the two newly computed thresholds compared with the BER threshold. In addition, while the overall FP and PC rates for these two threshold arrays are the same, the group rates are much lower for SP thresholds than PE thresholds. However, the FP and PC rates do not appear to be close for groups, suggesting that there is data mismatch between the train and test datasets. For the BE rates, overall PE and SP thresholds seem to do better overall, but worse with specific groups.

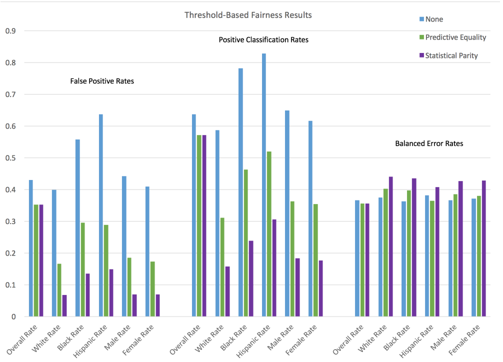


*Figure 9.* False positive, positive classification, and balanced error rates for balanced error, predictive equality, and statistical parity thresholds

## 7. Conclusion and Future Work

In our analysis of the Stanford Open Policing Project data, we aimed to study the impact of regularization schemes and post-training threshold processing on reducing variances in FPR and PCR among classifiers trained to predict traffic stop outcome. We were able to reproduce the results of (Berk et al., 2017), illustrating the inherent tradeoff between fairness with respect to natural individual and group losses, and model accuracy. Moreover, we illustrated that despite operating with a loss function not explicitly targeting false positive rates or positive classification rates, fairness regularization is well suited for enforcing common fairness metrics, such as statistical parity and predictive equality. Post-training threshold regularization demonstrated an overall improvement in both fairness and accuracy, but no change in group fairness and a decrease in group accuracy.

However, as authors in (Dwork et al., 2011) observe, group fairness metrics such as statistical parity and predictive inequality are insufficient to capture the strongest notions of individual fairness. Future analyses of this dataset and related ones should look at losses in individual fairness arising from optimizing for FPR or PCR alone, and should consider regularization schemes that best address the challenges of individual fairness. Further analysis should also include finding a model that optimizes accuracy metrics. Utilizing the fairness techniques presented here to analyze such a model would prove noteworthy.

## Contributions

Vikul Gupta (vikulg): Exploratory Data Analysis, Principal Component Analysis, Threshold-based Fairness
Kuhan Jeyapragasan (kuhanj): Data Classification/Analysis/Cleaning, Naive Bayes, Logistic Regression, Method Comparison
Jaydeep Singh (jaydeeps): Preliminary Planning/Agenda, Data Classification/Cleaning, Logistic Regression, Regularization

## References

5harad. Stanford open policing project. https://github.com/5harad/openpolicing, 2017.

Adler, Philip, Falk, Casey, Friedler, Sorelle A, Rybeck, Gabriel, Scheidegger, Carlos, Smith, Brandon, and Venkatasubramanian, Suresh. Auditing black-box models for indirect influence. *IEEE International Conference on Data Mining*, abs/1701.08230, 2016. URL https://arxiv.org/abs/1602.07043.

Berk, Richard, Heidari, Hoda, Jabbari, Shahin, Joseph, Matthew, Kearns, Michael J., Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. A convex framework for fair regression. *CoRR*, abs/1706.02409, 2017. URL http://arxiv.org/abs/1706.02409.

Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, Goel, Sharad, and Huq, Aziz. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017. URL http://arxiv.org/abs/1701.08230.

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard S. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. URL http://arxiv.org/abs/1104.3913.

Pierson, Emma, Simoui, Camelia, Overgoor, Jan, Corbett-Davies, Sam, Ramachandran, Vignesh, and cheryl Phillips Sharad Goel. A large-scale analysis of racial disparities in police stops across the united states. *CoRR*, 2017. URL https://arxiv.org/abs/1706.05678.

Zemel, Richard, Wu, Yu, Swersky, Kevin, Pitassi, Toniann, and Dwork, Cynthia. Learning fair representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–325–III–333. JMLR.org, 2013. URL http://dl.acm.org/citation.cfm?id=3042817.3042973.