When to Stop-and-Frisk

Madeline Saviano (msaviano@stanford.edu)
Sarah Tieu (stieu12@stanford.edu)

Predictions Driven by an interest in using machine learning for social impact, we utilized features including gender and race in order to predict which stop-and-frisk incidents would result in non-negative outcomes (i.e. the suspect having been involved in illegal activity that would have justified the stop). We built multiple models and tested many different feature subsets, including multinomial naive bayes, logistic regression, support vector machines, and neural networks. After hyperparameter tuning, we found that many models and feature sets produced similar results, with a few slight exceptions. We then turned our attention to model analysis and discovered that our logistic regression model was more racially biased (i.e. positively labels more minorities) than the police officers themselves. Thus, we incorporated the relatively new field of algorithmic fairness into our results by altering the threshold values for logistic regression so each race's positive violation percentages would mirror the percentages from the 2010 census (a statistical parity approach)[1].

Data We used a dataset compiled by the Stanford Open Policing Project[2], from which we drew over 8 million examples from the state of Washington (80% to training data, 10% each to validation and test data). Each example contains all the information drawn from a single incident and corresponding police report. We used the outcome of the stop as the ground truth, which is positive if a ticket or violation was given.

Features From the raw input data consisting of 34 features (Id state, stop_date, stop_time, location_raw, county_name, county_fips, fine_grained_location, police_department, **driver_gender**, driver_age_raw, **driver_age**, driver_race_raw, **driver_race**, violation_raw, **violation**, **search_conducted**, search_type_raw, search_type, **contraband_found**, **stop_outcome**, is_arrested, violations, officer_id, **officer_gender**, **officer_race**, highway_type, road_number, milepost, **lat**, **lon**, contact_type, enforcements, drugs_related_stop), we removed duplicate features (ex. search_type_raw removed, search_type kept) and removed uncategorizable data (ex. county_name). We started with a set of ten features (**bolded**) and the predicted feature: stop_outcome, a binary outcome of 1 if the outcome was an arrest or a citation and 0 otherwise. Violations were manually mapped to a scale where more severe violations had a larger corresponding integer than minor violations and, for each example, the violation of highest severity was used as the input feature value. For each input feature, we added and removed the feature from our feature set and ran the accuracy; however, the accuracy decreased in all instances, and thus our final feature set consisted of those ten features.

Models We built four different models (multinomial naive bayes, logistic regression, support vector machines, and neural networks) and tuned the hyperparameters of each one to see if the accuracies would differ between the models. For all of the tuned models, the models had about the same validation accuracy, thus we decided to move forward with logistic regression which allows for more manual manipulation of the data.

Neural Network: We decided to use one layer in order to prevent overfitting. After testing different learning rates, batch sizes, and hidden units, the best result had the following hyperparameters: learning rate = 0.005, 256 batch size, and 10 hidden units.

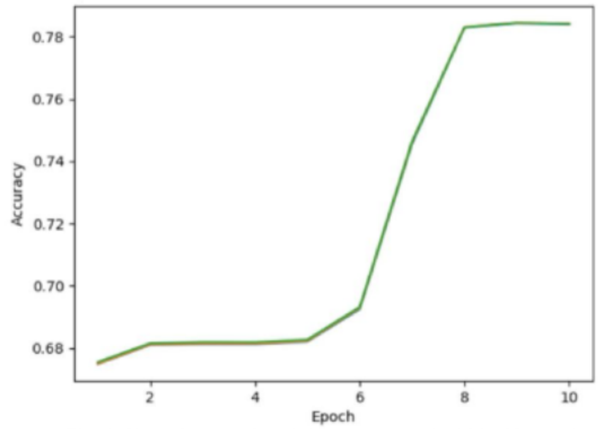

Fig. 1: Neural network training converged within 10 epochs.

Results Training dataset has 6.5 million samples and test dataset has .8 million. Baseline result is from running a naive bayes with only the violation as the input feature.

	Train Accuracy	Test Accuracy
Baseline	0.79204	0.79205
Multinomial Naive Bayes	0.63762	0.63810
Logistic Regression	0.68366	0.68379
Support Vector Machine	0.79704	0.77788
Neural Network	0.77181	0.77256

Discussion As we discovered after training multiple models, our data contains a relatively high level of bias, which could be explained by bias in using police reports generated by the officers themselves (general human error or self-serving justification to provide paperwork to support a stop) or inherent bias in our dataset not possessing enough or the right features to explain the complexity of a stop-and-frisk incident. The fact that our model (logistic regression, if not others) was more racially biased in its labeling than the true violations given by officers was surprising. Furthermore, the act of manually adjusting thresholds based on race feels wrong -- should a single software engineer decide that one race must meet a higher level of suspicion to be pulled over? On the other hand, given that our unaltered logistic regression algorithm was more racially biased implies that software engineers, regardless of whether or not they acknowledge it, have a large impact on potentially devastating social impact issues.

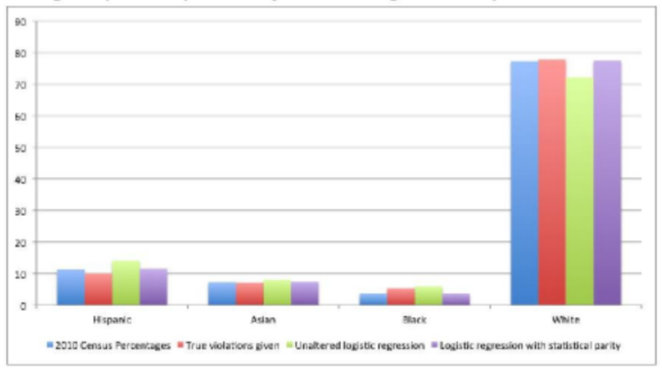

Fig. 2: Altering threshold values to achieve racial statistical parity.

Future We would love to take test accuracy a step further by running our best models on a dataset from another state; however, given that each state has slightly different features, this step would require a large amount of synthesizing between models. We would also attempt to find less biased data by consulting different stop-and-frisk datasets or training on multiple different states from the Stanford Open Policing Project to compare results.

References

[1] Sam Corbe!-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In Proceedings of KDD '17, August 13-17, 2017, Halifax, NS, Canada, 10 pages. DOI: 10.1145/3097983.3098095
[2] E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, V. Ramachandran, C. Phillips, S. Goel. (2017) "A large-scale analysis of racial disparities in police stops across the United States".
[3] Goel, Sharad and Rao, Justin M. and Shroff, Ravi, Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy (March 2, 2015). Annals of Applied Statistics, Vol. 10, No. 1, 365–394, 2016.