Predicting Police Behavior

David Guo
University of Michigan
davidguo@umich.edu

Abstract

Police stop individuals usually for some law violation. However, police are also prone giving individuals too much or too little punishment. We seek to predict what factors contribute to police stop outcomes. When using stop data from Florida since 2011, our findings show that logistic regression and hidden markov models perform comparably well with an accuracy rate of about 0.7 and 0.75 respectively. This demonstrates the importance of using officer traits in modeling traffic stop behavior. Taking into account officer's attributes and historical actions gives us a better understanding of law enforcement patterns.

1 Introduction

Police departments have approached law enforcement from a predictive standpoint. However, studies, including ProPublica's insight into a crime predicting algorithm COMPAS, have found that because prior police intervention was prone to racial bias, the predictions also ended up biased (1).

To assess civilian reaction to police response, a widely used resource is the Police-Public Contact Survey, which is conducted every 3 years by the Bureau of Justice. In addition to survey pitfalls and response bias or nonresponse, the survey was last conducted in 2011 since it is currently under going improvements. By looking at the raw data, we can get a better picture of certain trends and behaviors that are not necessarily captured in official summarization reports.

2 Related Work

The Southern Coalition for Social Justice, in the Open Data Policing Project, obtains police stop data from North Carolina, Maryland, and Illinois. They do not alter or manipulate the raw data. They use the population and racial demographics from the census to establish a baseline. They examine the frequency of stops based on race, and also break this down by looking also at the frequency for different stop reasons (such as "driving while impaired" or "seat-belt violation"). The report also looks at the average departmental stop rates by police departments and the search rate by ethnicity to compare likelihood of stop cause by comparing the search rates of different ethnic groups and report the contraband hit-rate (searches that results in finding contraband) for each race. (2)

The Stanford Open Policing Project obtained police stop data from 20 states, which also include North Carolina, Maryland, and Illinois. This group took a different approach by cleaning and standardizing the data before performing analysis since the raw data is not recorded uniformly across states. They examine the stop, citation, search, and arrest rates by race, and adjust for driving-age population in each state. Instead of using the hit-rate approach in the Open Data Policing Project, they use a threshold test that uses a hierarchical Bayesian Model in order to account for some of the biases that arise in searches. Although stops are more difficult to conclude that discrimination was a dominating factor, searches for blacks and Hispanics had a lower threshold as compared to white drivers. Another specific problem they also address is the legalization of marijuana and stops, which was found to have a noticeable decrease in stops, but also had similar search discrepancy.(3)

A related problem is interactions that involve police violence. The Mapping Police Violence group looked at the rate at which suspects are killed by race and compared that to the population proportion of that race. They also examined the rate at which different departments kill armed and unarmed suspects. The hit-rate method is also used.(4)

These reports focus on the distinctiveness of the suspects. A survey approach by Pew looked at police officers' perspectives on different police procedures. One of the findings was that "Only about a third (32%) of black officers but about twice the share of whites (62%) report they have become more callous [toward people] since taking the job." "The survey finds that officers who feel they have grown more callous since starting their job are also more likely to endorse the use of aggressive or physically harsh tactics in some situations or in some parts of the community than officers who say they have not grown more callous." (5)

We would like to approach this data set by looking at this approach from police officers. This is different since most of the studies have treated the police officers as a uniform entity. We will try to use a model to predict the rate of behaviors that officers engage in with at stops.

3 Problem Definition and Data

We want to see if there is a good method in predicting officer's behaviors with the given data. We will use the stop outcome as a metric to assess officer behavior because this is ultimately a finalized decision by the officer.

We will use the cleaned data from the Stanford Open Policing Project. Although not all states have responded to requests, the project has about 20 states that are formatted as CSV files. The original data came separately from each state with different formats, such as fixed-width or xls. Of these states, 18 of them include an anonymized ID of officers, where only some contain information about the officer's age, sex, race, and rank.

We examine data from Florida since 2011. There are about 4.8 million entries. The stop outcomes in this state are warning, faulty equipment notice, citation, misdemeanor arrest, and felony arrest. We ignore outcomes that are not available. Relevant variables regarding the officer are age, sex, gender, and rank.

4 Methodology

We assign warning and faulty equipment notice as "light" outcome, while citation, misdemeanor, and felony arrest as a "severe" outcome.

The main approach is to take advantage of the time dependent and ordered nature of the data. It is not unreasonable to assume that police officers will be influenced from prior behavior. If we as-

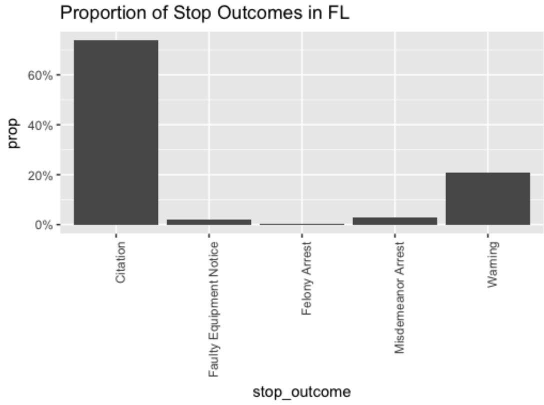

Figure 1: Proportion of Stop Outcomes

sume this, a reasonable approach would be to use a hidden markov model. We use the sequence of stop outcomes as the observed sequence. The possible labels are C (citation), E (Faulty equipment), F (Felony Arrest), and W (Warning). Police officers who only engaged in one stop and police officers with over 10000 stops were not included as a sequence. Then, we simulate sequences from the generated from a model with equal probability of hidden states, and one from the model generated. We then compare the outcomes generated from the two models, using a minimum threshold to deem a sequence as common or not common. If the sequence was categorized differently between the random and HMM model, we labeled it as significant. We calculate the proportion of those that are significant as a measurement of how the HMM predicted sequences in the test data.

5 Evaluation and Results

We will use the training and test error rates to evaluate the effectiveness of the predictor. We used 5-fold cross validation. When we weight each outcome equally (that is, each outcome is equally likely), the average accuracy rate was approximately 0.50. This solution is as if we flipped a coin to pick the outcome. To improve on this approach, we use the proportion of outcomes from the data as the baseline. The average accuracy rate for this approach was about 0.60.

We use logistic regression to determine the outcome, with the dependent variable as the various attributes of the police officer and driver. Different model examples are included in the Appendix. In all the models we tried, the average accuracy rate was approximately 0.70.

The next step of reframing the problem by ex-

# of Hidden States	Log-likelihood (10^6)
2	-2.48251
3	-2.48224
4	-2.33936
5	-2.47789
6	-2.327843
10	-2.30765

Table 1: Log-likelihood values of HMMs with various hidden states

amining the police officer's sequence of stop outcomes, and using a hidden Markov model approach. baseline to compare is if we just use initial, transition, and emission probabilities that are equal for all hidden states. Although 10 hidden states had a higher log-likelihood, for computational feasibility, we used 4 hidden states. We then simulated 1000 sequences of different lengths in both the random and the new HMM model. Due to computational limitations, we could not examine all sequence lengths in the test data (the maximum being 9812), and restricted this to 1000. A sample plot for sequences of length 5 is indicated in figure 2. For sequences shorter than 10, we were able to get an average accuracy rate of approximately 0.75. For longer sequences, performance was poor.

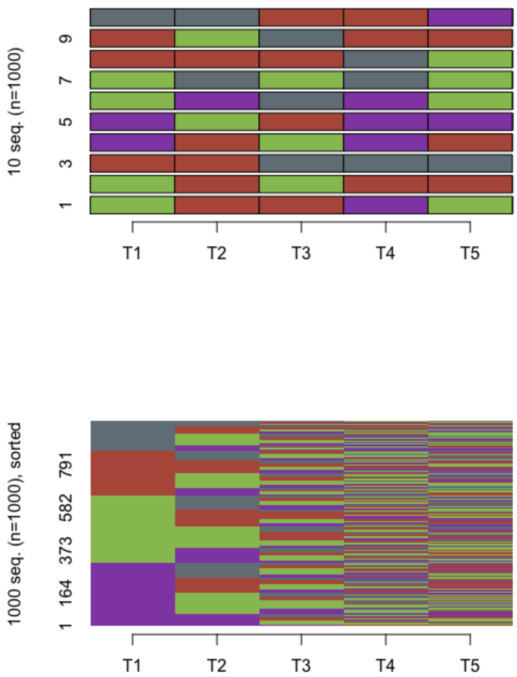

Figure 2: Example of simulation for 5-stop sequences. Top 10 sequences and of all sequences. C: Purple, E: Green, F: Red, W: Grey

6 Discussion

From the logistic regression, the officer's race had more contribution to the outcome result as opposed to the driver's race. Other officer attributes, such as age, were included, but did not adjust the model's fit as much. From an interpretability standpoint, this approach is better than the HMM, as the hidden states may be harder to define.

There was some consideration of using a time series approach in terms of an aggregate as opposed to a sequence approach for each officer. However, with the time series approach, dependence between different officers would need to be included and would create a complex model. With the sequence approach, we thought it could be feasible without considering the interactions. In terms of interpretability, we can treat the 4 hidden states as the main racial categories of White, Black, Hispanic, and Asian. Since the sequences took into account a form of temporal dependence, it performed slightly better than the logistic regression. This only worked for shorter sequences, which may mean that police behavior's stop outcome is only dependent on the short term.

Future studies can try to apply this method to different states besides Florida to find if there is a more general stop outcome. Another attribute, officer rank, was too unclear to clean, as many abbreviations were ambiguous and could stand for many different ranks. If this attribute was cleaned, different ranked police officers could be tested for different outcome rates. For those with more computational power, the 10 hidden state markov model could also be tested. Although there was a code for county and police department, many police officers were not confined to one county or the police department was at the state level, so analysis for county differences was not sufficient.

Acknowledgments

Thanks to Professor Jurgens for additional guidance on this project.

References

[1] How We Analyzed the COMPAS Recidivism Algorithm https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[2] Open Data Policing. https://opendatapolicing.com/

[3] 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"

[4] Mapping Police Violence. https://mappingpoliceviolence.org/planning-team/

[5] Pew Research Center, Jan. 2017, "Behind the Badge: Amid protests and calls for reform, how police view their jobs, key issues and recent fatal encounters between blacks and police."

A Supplemental Material

Logistic Regression

```
Coefficients:          Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.722048   0.154322  17.639  < 2e-16 ***
officer_raceAsian      0.295812   0.017065  17.334  < 2e-16 ***
officer_raceBlack      0.342217   0.004782  71.561  < 2e-16 ***
officer_raceHispanic   0.233369   0.005145  45.360  < 2e-16 ***
officer_raceOther      1.003556   0.018387  54.581  < 2e-16 ***
officer_raceWhite      0.028904   0.003121   9.261  < 2e-16 ***
driver_raceAsian      -1.506068   0.154728  -9.734  < 2e-16 ***
driver_raceBlack      -1.543421   0.154332 -10.001  < 2e-16 ***
driver_raceHispanic   -1.256797   0.154333  -8.143 3.84e-16 ***
driver_raceOther      -1.450468   0.154526  -9.387  < 2e-16 ***
driver_raceWhite      -1.806715   0.154312 -11.708  < 2e-16 ***

Coefficients:          Estimate Std. Error  z value Pr(>|z|)
(Intercept)            2.847742   0.519093    5.486 4.11e-08 ***
officer_raceAsian      0.224184   0.020435   10.971  < 2e-16 ***
officer_raceBlack      0.353261   0.005633   62.709  < 2e-16 ***
officer_raceHispanic   0.241745   0.005781   41.816  < 2e-16 ***
officer_raceOther      0.908148   0.020638   44.005  < 2e-16 ***
officer_raceWhite      0.042791   0.003603   11.875  < 2e-16 ***
driver_raceAsian      -1.223480   0.519251   -2.356  0.01846 *
driver_raceBlack      -1.237016   0.519089   -2.383  0.01717 *
driver_raceHispanic   -0.970267   0.519089   -1.869  0.06160 .
driver_raceOther      -1.141636   0.519168   -2.199  0.02788 *
driver_raceWhite      -1.452854   0.519080   -2.799  0.00513 **
driver_age            -0.011581   0.000100 -115.780  < 2e-16 ***
```

HMM Model

```
Initial probabilities :
State 1 State 2 State 3 State 4
0.0737  0.3174  0.4083  0.2006
Transition probabilities :
                        to
from      State 1 State 2 State 3 State 4
State 1  0.2335  0.3773  0.0028   0.386
State 2  0.2877  0.0433  0.3423   0.327
State 3  0.2065  0.4987  0.0779   0.217
State 4  0.0326  0.5122  0.2679   0.187
Emission probabilities :
                                  symbol_names
state_names      C      E      F      W
State 1        0.0118 0.303 0.362 0.3230
State 2        0.1745 0.333 0.176 0.3165
State 3        0.4615 0.372 0.157 0.0095
State 4        0.1795 0.330 0.365 0.1257
```