THE UNIVERSITY OF CHICAGO

Finding Bias in Police Misconduct Investigations Using Differentiated Feature-Sets in Supervised Machine Learning

June 5, 2016

Principal Investigators

Christine Chung
cschung@uchicago.edu

Lauren Dyson
ladyson@uchicago.edu

Leith McIndewar
lmcindewar@uchicago.edu

Abstract

Background

A heated national conversation has emerged around the issue of systemic bias and discrimination within the criminal justice system. After several high profile incidents, the demand for more fairness, transparency, and accountability in the investigation of police misconduct has reached a fever pitch. However, there currently exists a dearth of quantitative methods for detecting unfair bias embedded in these procedures. Our research explores of it is possible to detect possible biases against complainants of certain demographics through different stages in the police misconduct investigative process.

Methods

Our research seeks to explore if there is bias in the investigation process against complainants of a certain race, gender, or age — traits that should not affect the outcome of an investigation. We first build two predictive models with the same base parameters but trained on differentiated feature sets: one that includes complainant demographics, and one that does not. Our hypothesis is that if biasedness against complainants is present in the investigative process, the addition of complainant demographic data will provide a lift over the unbiased model in predicting the outcome of an investigation.

Results

Our results show that considering demographic features improves predictive ability in determining whether or not a complaint is sustained (defined as enough evidence is found to justify disciplinary action against the accused officer”). Our model using complainant demographic information saw a .117 increase in the ROC-AUC score, an improvement of 17% over the baseline of 0.693. We found no improvement from considering complainant demographics in predicting in whether or not an investigation is dropped due to an unsigned affidavit.

Keywords: Machine Learning, Bias, Police, Gradient-Boosting, Discrimination, Criminal Justice

Background and Introduction

In the last several years, high profile incidents such as the deaths of Michael Brown in Ferguson, Freddie Gray in Baltimore, and Tamir Rice in Cleveland have drawn increased attention to the problem of differential policing and treatment of minority communities. In particular, after numerous highly visible incidents of police brutality including the shooting death of Laquan McDonald, scrutiny has focused on the Chicago Police Department. Activists and policymakers have called for more accountability, transparency, and justice. One avenue towards increased accountability is improving the citizen complaint investigation process.

In Chicago, the Independent Police Review Authority (IPRA) is tasked with "[promoting] increased accountability by, and transparency about the work of, the Chicago Police Department."[1] Created in 2007, the board investigates complaints made by citizens regarding officer misconduct. It also investigates uses of force, regardless of whether a complaint is filed. IPRA was created to separate the review process for misconduct from the police department, yet, since the beginning, the board has had significant issues of its own.

A task force created by Chicago Mayor Rahm Emanuel characterized the board as "being opaque, inefficient, ineffective at investigating alleged police misconduct and, ultimately, 'beyond repair.'"[2] On May 14, 2016, Chicago Mayor Emanuel announced plans to dismantle IPRA. Following the task force's recommendations, IPRA will be replaced by a citizens' review board with broader powers and an inspector general with the expanded ability to audit the Chicago Police Department. The goal of IPRA, and soon for the new review board, is to instill trust in the investigative and disciplinary processes of police complaints and misconduct.

Our work investigated if it is possible to identify systematic biases in the outcomes of police misconduct investigations based on the demographics of the complainant. Recent news has predominantly made light of bias in police actions and tactics. We are interested in quantitative methods to detect discrimination in how citizen complaints and concerns are addressed by those tasked with investigating police officers.

Overview of Related Work

Predictive modeling has increasingly been adopted in the criminal justice field. The pressure of budget cuts, increased technological capacity, and a new availability of data at the urban level has led many US police departments to pursue predictive policing strategies (Perry, McInnis, and Smith provide an overview of efforts in this space).[3] However, many of these efforts are focused on anticipating crime activity in order to more efficiently allocate limited policing resources. In particular, two areas have seen rapid adoption in recent years: "hotspotting" approaches to predict where and when crime will occur and "heat list" approaches that predict who is likely to be involved in criminal activity, either as an offender or victim.[4]

Despite the widespread public attention focused on police misconduct specifically, there has been little data available on past complaints and their outcomes. This has been a limitation on the application of machine learning techniques to the area of police misconduct specifically, though early efforts exist. A recent project led by researchers at the University of Chicago in collaboration with the Charlotte-Mecklenberg Police Department has applied machine learning to identify police officers at risk of being involved in adverse interactions while on duty.[5] These models focus on prediction before an incident occurs. Our work differs from this effort in its focus

[1] IPRA Mission Statement (http://www.iprachicago.org/ipra/homepage/about.html)

[2] "Emanuel plans to scrap beleaguered police oversight agency IPRA", Chicago Tribune, May 14, 2016

[3] Perry W, McInnis B, Price C, Smith S, Hollywood J. Predictive POLICING The Role of Crime Forecasting in Law Enforcement Operations [Internet]. The RAND Corporation; 2013. Available from: http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf

[4] Chicago Police Department. CUSTOM NOTIFICATIONS IN CHICAGO - PILOT PROGRAM D13-09. Chicago, Illinois: Chicago Police Department; 2013.

[5] Carton S, Helsby J, Joseph K. et. al. Identifying Police Officers at Risk of Adverse Events (2015). http://www.kdd.org/kdd2016/papers/files/Paper_832.pdf

not on anticipating future incidents of police misconduct, but in uncovering bias in the way complaints have been investigated and adjudicated under the current system.

Approach: Towards A Better Understanding of Bias in Misconduct Complaints

In a talk entitled, “Bias and Ethics in Machine Learning”, researcher Mike Williams writes, “Applied to human beings, machine learning is a formalized method for finding useful stereotypes.”[6] Given this definition, the question becomes whether or not machine learned models can help us determine what those stereotypes are, and whether those stereotypes are ethical to use in practice. This is the premise that undergirds our research.

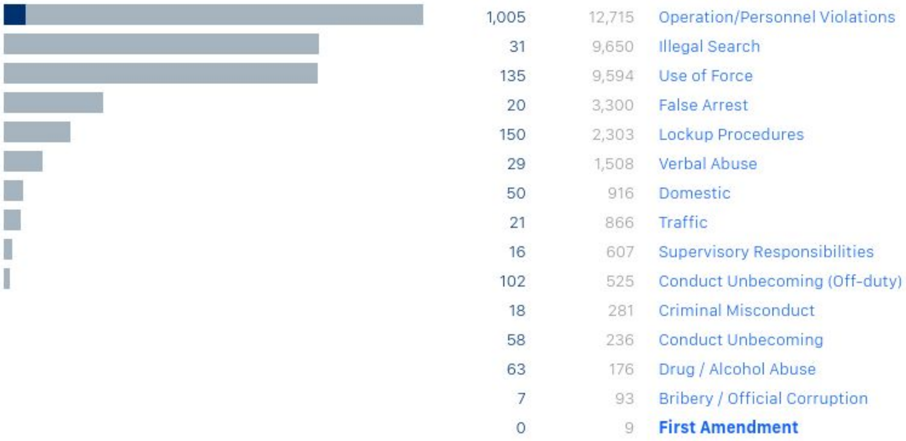
Our research seeks to determine if there is bias in the investigation process against complainants of a certain race, gender, or age — traits that should not affect the outcome of an investigation. We first build two predictive models with the same base parameters but trained on differentiated feature sets: one that includes complainant demographics, and one that does not. Our hypothesis is that if biasedness against complainants is present in the investigative process, the addition of complainant demographic data will provide a lift over the unbiased model in predicting the outcome of an investigation.

The model that does not assume bias is composed of features that might influence a fair investigator’s decision. Of course, the most important feature would be evidence, which we do not have. The unbiased model also includes some “unfair” features that could influence the final outcome of an investigation, but do not have to do with the demographic information of the complainant. These features include whether the incident reported occurred on the weekend and the number of prior complaints and the length of time on the force for the accused officer.

The biased model includes demographic information of the person filing the complaint, and any other information that might be a proxy for demographic information, such as the location of the complaint. Both models include the demographic information of the officer involved in the incident. We define an improvement between the unbiased and biased models as an increase in the average of the Area Under the Receiver Operating Characteristic (ROC-AUC score) for the best performing models using a given classifier over many parameter sets.

Data

The original dataset was obtained through the Freedom of Information Act (FOIA) by the Invisible Institute, an investigative journalism nonprofit based in Chicago. The dataset contains 56,000 misconduct complaint records for approximately 8,500 Chicago police officers. Officer rank, salary, and demographic information and Complainant demographic information was later merged with the original dataset and obtained through



[6] Http://www.slideshare.net/dominodatalab/data-science-popup-austin-privilege-and-supervised-machine-learning

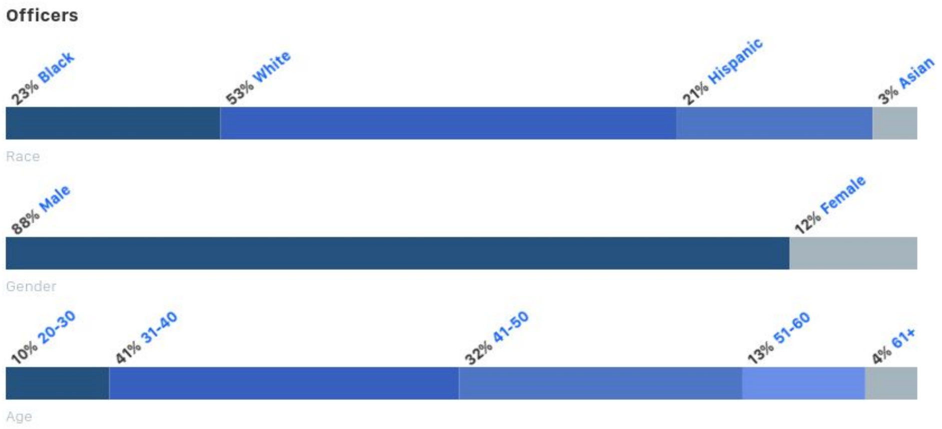
Figure 1. Breakdown of complaint types from the Invisible Institute's Website.

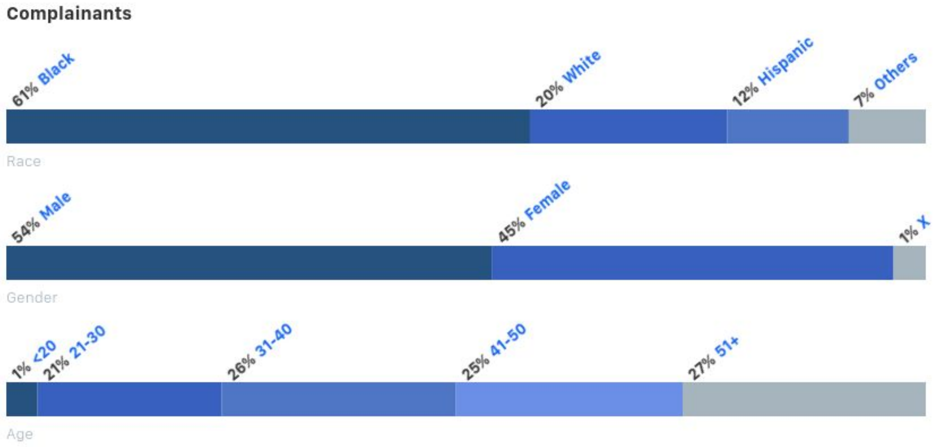
additional FOIA requests or through the City of Chicago's open data portal. We have augmented this with several other data sets in order to aid feature generation. The primary datasets added include crime incident reports, 311 service requests, and ACS/Census data for the city of Chicago. The complete list of data is listed in Table 2 in the Appendix.

Summary Statistics

Half of CPD officers do not accumulate any complaints. 6% of repeat offenders hold 31% of complaints. Accused police officers are overwhelmingly white and male, while complainants are overwhelming black and more evenly split between genders.

From 2011 - 2014, 28.6% of cases were dropped due to a missing affidavit. 36.5% were dropped due to an unknown officer. Together, that amounts to 65% of accusations that were never investigated. Further, only 4.2% of investigations were sustained, meaning that the "the allegation is supported by sufficient evidence to justify disciplinary action".





Data Cleaning

In total, we dropped 37,689 observations from the original dataset, decreasing the total number of observations in our base dataset from 56,384 to 18,695. We kept observations that fell within March 2011 - December 2014. These were the most reliable data, as data outside of the time range came from a different database and were difficult to resolve. In addition, we dropped any rows that did not have a recorded incident date or location, since

Figure 2. Breakdown of officer demographic information (top) and complainant demographic information (bottom) from the Invisible Institute's website[7].

many of our features rely on the time of the incident and the location of the incident. Please refer to Table 1 in the appendix for a detailed list of changes to the dataset.

The missing age of officers and complainants were imputed with the mean. For complainants, there were 33 in the dataset over the age of 100, including 15 over the age of 900. These ages were removed and later imputed with the mean. Missing races and genders were denoted with a separate dummy variable (e.g. Gender:Nan, Race:Nan). This, in part, was to account for the possibility that there was some selection bias about when data was omitted.

We made predictions at two different stages of the investigative process, and therefore, used two different subsets of data (see the Methodology section). For the first stage, our data included 18,694 observations, representing 12,825 unique complaints. The number of total observations is higher because some complaints featured more than one complainant. Of these, approximately 6,503 or 34.79% were dropped due to not having signed an affidavit, the predicted variable. In the second stage, we saw 600 observations or 4.92% where the finding was 'Sustained', the variable of interest in Stage 2. In total, there were 12,191 observations and 9,301 unique complaints.

There are two issues presented here. First, by adding rows for each complainant on an individual complaint, we are increasing the weight or power of complaints with more than one complainant. We made this decision because we thought that having a row for each individual complainant was essential to determining if bias existed against complainants. Secondly, especially in Stage 2, we encounter a problem of class imbalance. We discuss our techniques for addressing this in the Methodology section.

Feature Generation

We generated a set of features for each allegation made against an officer based on the details of the complaint, demographic information of the officer, history of the officer, investigation information, and area of the incident.

Table 1. Features Generated

The parentheses indicates which feature-sets included the category, further explained in the Methodology section.

Officer Features (Strict, Lenient, Full-Demographic):

- Length of time on the force at the time of complaint
- Prior complaints made against that officer at time of complaint
- A weighted *degree-centrality score* based on the number of times an officer had been co-accused with another officers[8]

Allegation Features (Strict, Lenient, Full-Demographic):

[7] https://cpdb.co/

[8] We created a graph with officers as nodes and being named on the same complaint with another officer as an edge. Being named on more than one complaint with the same officer increased the weight of that edge. A degree centrality score was generated by dividing an individual officer's number of edges by the total possible number of edges in the graph.

- Number of other police misconduct complaints made two weeks, three months, six months and one year prior to the complaint date by distance (radius of 500, 1000, and 2500 meters) from the complaint location

Complaint Features (Strict, Lenient, Full-Demographic):

- A dummy variable indicating if the allegation category involved physical force
- A dummy variable indicating whether the investigator is a police officer
- Whether the complaint was made on a weekday or weekend
- Travel time from incident location to IPRA office via car and public transit[9]

Location-based Features (Lenient, Full-Demographic):

- Number of 311 complaints made two weeks, three months, and six months prior to the complaint date by distance (of 500 meters, 1000 meters, and 2500 meters) from the complaint location, broken down by type of complaint (rodents, sanitation, alley lights out, graffiti, street lights out, vacant buildings, garbage, abandoned vehicles, and potholes)
- The number of crime reports made two weeks, three months, and six months prior to the complaint date by distance (of 500 meters, 1000 meters, and 2500 meters) from the complaint location, broken down by type of crime (robbery, assault, drug abuse)

Census-tract Features (Lenient, Full-Demographic):

- Population by census tract broken down by age group, race, poverty level, and educational attainment

Methodology

We examined outcomes at two stages in the complaint review process:

- **Stage 1:** Complaints are dropped if no affidavit is signed by the complainant
- **Stage 2:** An outcome of ‘Sustained’ is determined at the conclusion of the investigation (possible outcomes are unsustained, unfounded, exonerated, or unknown)[10]

While there are multiple possible outcomes at each of these stages, we coded the outcome as a binary variable: for stage 1, whether or not an affidavit was signed, and for stage 2, whether or not the complaint was ultimately sustained.

We created three feature-sets to make predictions at two different stages of the investigative process. This resulted in six different datasets that we used to make predictions. The criteria used to determine the three datasets was how easily could you infer the demographic information of the complainant using that data.

For each stage, we created three versions of the dataset (see features table in Data section for more details):

- **Strict Non-Demographic:** Includes officer-related features and complaint-related features, including allegation counts, travel time to IPRA office, prior complaints, and more. The intent of this feature-set is to exclude all information that explicitly or implicitly relates to the demographics of the complainant. This is of particular concern in Chicago, where neighborhoods are highly segregated and any location-based features (e.g. crime rate and 311 counts) may be highly predictive of a complainant’s race.
- **Lenient Non-Demographic:** Includes the above, plus geographic-related features, 311 counts, census tract-related features, and police beat.
- **Full-Demographic:** Includes all of the above, plus complainant-related demographic information and all generated features.

[9] Complainants need to sign an affidavit in person at the IPRA office in order for the complaint to be investigated. It may be that the more difficult it is to sign the affidavit, the less likely the complaint will be investigated.

[10] http://www.cityofchicago.org/content/dam/city/depts/cpb/PoliceDiscipline/AllegMiscond201505.pdf

We then ran a parameter search for find the best model to predict the outcome for each stage using the strict non-demographic dataset. For parameters, we considered the following:

Oversampling methods versus no oversampling: Because of the imbalanced nature of this dataset, we tested several approaches to artificially increasing the proportion of minority class cases in the training dataset used to fit each model. The approaches tested were Random Minority Over-sampling with Replacement, Synthetic Minority Over-sampling Technique (SMOTE),[11] and no oversampling.

Classifiers: We iterated over parameter grids for Random Forest, AdaBoost, KNN, Decision Tree, Naive Bayes, and Gradient Boosting classifiers. In early trials, Random Forest, AdaBoost and Gradient Boosting consistently performed better with higher ROC-AUC scores. This is not surprising since boosting algorithms are well-suited to classification tasks where there is an imbalance in class representation. We focused efforts on more exhaustive parameter tuning for these three classifiers:

- **Random Forest:** Number of estimators to include in the forest, maximum depth of trees in the forest, maximum features to include, minimum samples to split on
- **ADABoost:** Boosting algorithm (SAMME or SAMME.R), number of estimators, base estimator (decision trees with varying parameters of criterion to split on, maximum depth, minimum samples to split on, and class weight of balanced or none), learning rate
- **Gradient Boosting:** Number of estimators, learning rate, subsample, maximum depth

Iterating over these parameters, we evaluated models using three-fold temporal cross-validation with a tack-one-on method (see Figure 4). We selected the model that had the best performance on average ROC Area-Under-the-Curve (AUC) score across the three validation folds on the strictly non-demographic dataset. The best performing models by that definition are:

Stage 1: Best performing model using Strict Non-Demographic Data	**Stage 2: Best performing model using Strict Non-Demographic Data**
Preprocessing: Random Oversampling Classifier: Gradient Boosting ● N_estimators: 100 ● Subsample: 0.1 ● learning_rate: 0.01 ● Max_depth: 5	Preprocessing: Random Oversampling Classifier: Gradient Boosting ● N_estimators: 10 ● Subsample: 0.1 ● learning_rate: 0.001 ● Max_depth: 5

Figure 3. Parameters for best performing models in Stage 1 and Stage 2 for the unbiased feature-set

Then, we ran a series of five trials to fit and test this model on the strictly non-demographic, lenient non-demographic, and demographic datasets, using the three-fold temporal cross-validation described above. To evaluate if including the demographic features improved the performance of the models in any significant way, we compared the performance across datasets on average ROC area-under-curve, precision/recall, and average accuracy. The results are discussed in the next section.

In addition to comparing the model that performed best on the non-demographic data with the same parameters across the three datasets, a secondary goal was to evaluate if a significantly better predictive model could be built using demographic information. We ran parameter searches to find the best-performing models overall using the demographic datasets and compared the top performers across feature sets.

[11] SMOTE: Synthetic Minority Over-sampling Technique, Chawla et. al. (2011) https://www.jair.org/media/953/live-953-2037-jair.pdf

Results

The best models for Stage 1 over 5 trials obtained an accuracy of 0.757, 0.752, 0.769 for the strict, lenient, and full-demographic data sets. A naive model that always predicted an affidavit would be signed would achieve .65 accuracy with perfect precision but no recall. In our second stage, the best models obtained an accuracy of 0.85, 0.79, 0.85. An untrained model that always predicted sustained would have obtained .95 accuracy with perfect precision but no recall. Our models had reduced accuracy over this naive hypothetical model but increased ability to predict the minority class (see precision-recall curves in Table 1).

When comparing the performance of the models that performed best with strictly non-demographic information with those given additional demographic information, the results were different for the Stage 1 task and the Stage 2 task. For the Stage 1 task of predicting complaints that would be dropped due to no affidavit, there was no improvement in average AUC when additional demographic information was provided. The AUC remained consistent at 0.766. However, for the Stage 2 task of predicting which complaints would ultimately be sustained, there was in increase of 0.117 in ROC AUC when additional demographic information was provided from a baseline of 0.693 for the unbiased model.

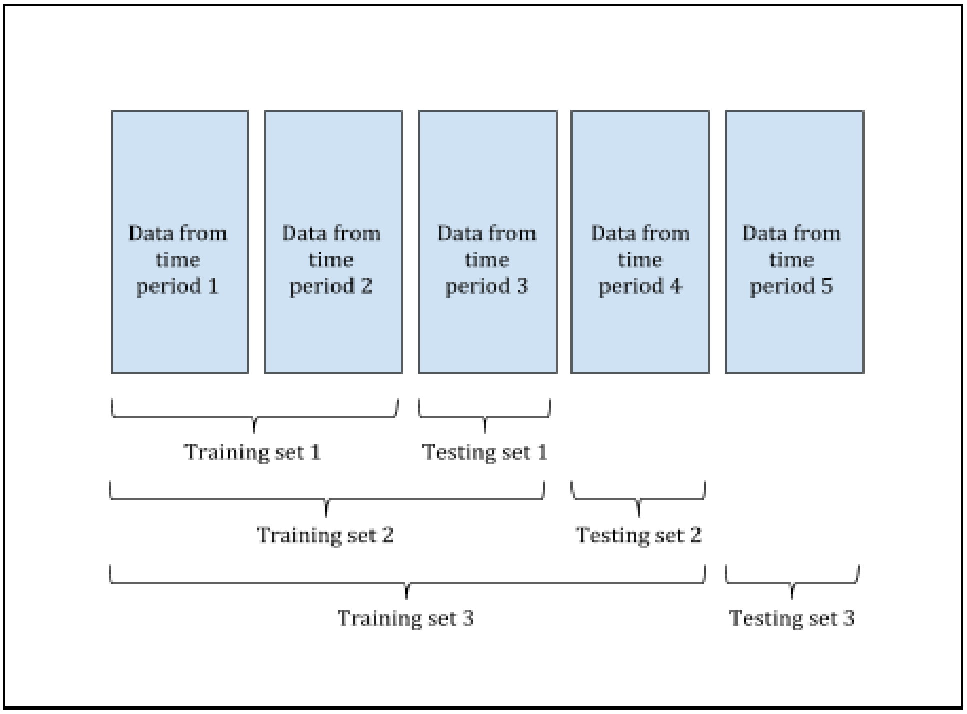


Figure 4. Temporal Cross-Validation — Tack-One-On Method**.**

	Stage 1: Average ROC AUC	**Stage 2: Average ROC AUC**
Strict Non-Demographic	0.766	0.693
Non-Demographic	0.752	0.754
Demographic	0.766	0.81
Difference between Strict Non-Demographic and Demographic	**0.0** (0% improvement)	**.117** (17% improvement)

Figure 5: Evaluation of Strict Non-Demographic Best Performing Model with Additional Demographic Feature Sets

This improvement suggests that complainant demographic information does in fact have predictive value of whether a misconduct complaint will be ultimately sustained. However, this is not enough to imply bias in the adjudication process. It is difficult to isolate the effects of demographic information and to attribute any increase in predictive ability to systematic bias. See the Challenges section for further discussion.

In addition to the AUC-ROC generally improving, we also examined changes in the precision and recall curves as more demographic information was added to the model. For the Stage 2 prediction tasks, at low levels of recall (<0.25), additional demographic data raised precision (see Chart 1). However, at mid-levels of recall (0.5), additional demographic information had little effect on precision. For Stage 1 prediction tasks, there was no significant difference at any point along the precision-recall curve between including strictly non-demographic features and including demographic features.

Table 1: Comparison Between Best Strict Non-Demographic Model Performance on Stage 2 Task[12]

<table>
<tr>
<td>

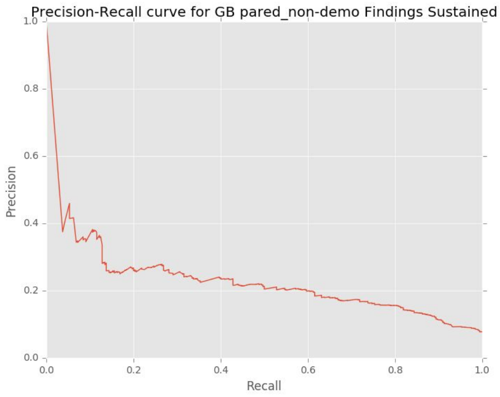
Precision/Recall Curves:
Predicting Stage 2 Outcome (complaint sustained) using Gradient Boosting model with Random Oversampling selected for best performance on Strict Non-Demographic Data

</td>
<td>

ROC Curves:
Predicting Stage 2 Outcome (complaint sustained) using Gradient Boosting model with Random Oversampling selected for best performance on Strict Non-Demographic Data

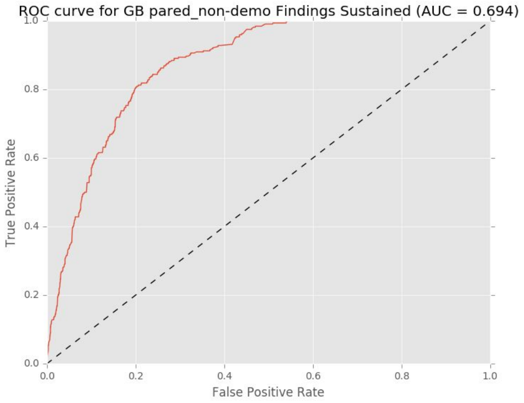
</td>
</tr>
<tr>
<td>

Performance using Strict Non-demographic Data



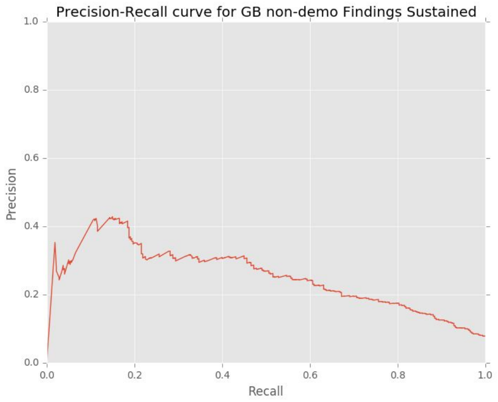
</td>
<td>

Performance using Strict Non-demographic Data



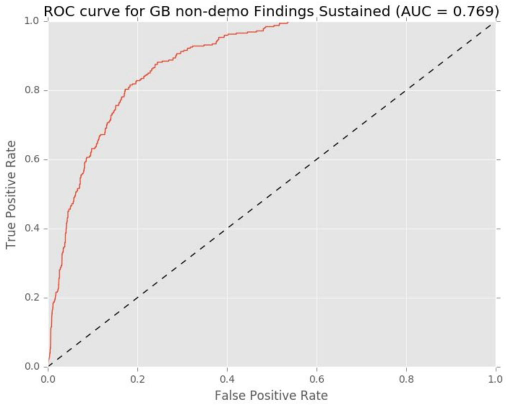
</td>
</tr>
<tr>
<td>

Performance using Non-demographic Data



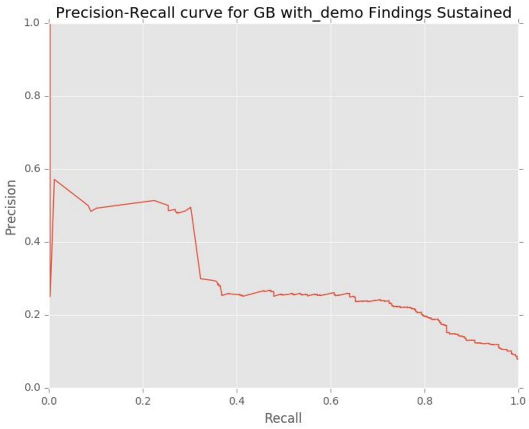
</td>
<td>

Performance using Non-demographic Data



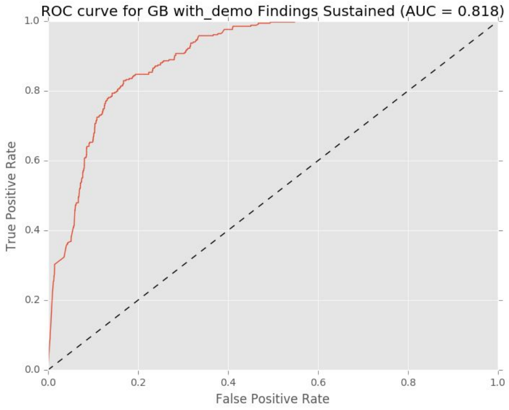
</td>
</tr>
<tr>
<td>

Performance using Demographic Data



</td>
<td>

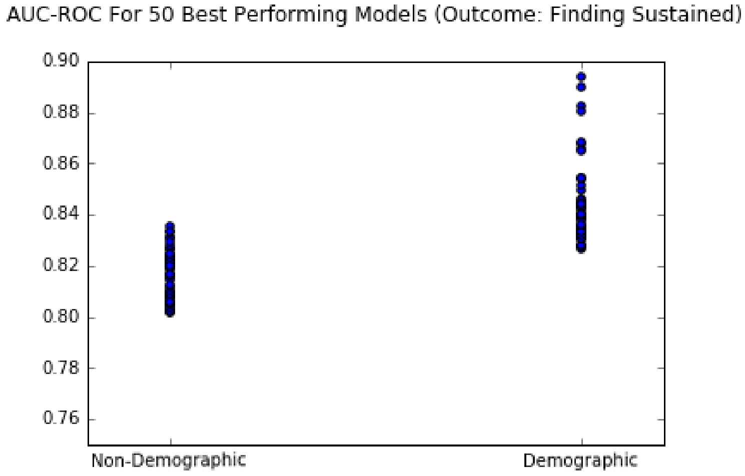
Performance using Demographic Data



</td>
</tr>
</table>

[12] See appendix for Stage 1 Task comparison.

We focused on the AUC-ROC as a balanced measure of the performance of the models. We also generated plots for the precision versus recall. For these we compared the metrics of the separately best performing models generated using including and excluding complainant demographic information. This compared whether any combinations of models and parameters performed better when additional information about the complainant was included. The figure to the left shows that the best model/parameter combinations performed better when the data included demographic information. On average the fifty best performing models with demographic information included had an AUC-ROC of 0.8445, while those without demographic information averaged 0.8164.



Best Performing Model AUC-ROC for “Finding Sustained” **All top-performing models were gradient boosting classifiers with random oversampling*	
Strictly Non-Demographic	Demographic
0.835	0.894
0.833	0.889
0.833	0.882
0.831	0.880
0.830	0.868
Average: 0.832	**Average: 0.883**

Figure 5: Evaluation of Strict Non-Demographic Best Performing Model with Additional Demographic Feature-sets

Discussion

Because of the limitations of the data, it is difficult to draw strong conclusions about the source of the improvement in model performance. While these models serve as a first step to identifying potential bias in the investigation and adjudication process of police complaints, we cannot definitively say that there is systemic discrimination. Instead this should be taken as an indication that the process may be susceptible to the same biased tendencies that plague other law enforcement practices.

Policy Recommendations

Further investigation would be needed to determine whether the trends found here are consistent in a larger sample of citizen complaint investigations. A simple, yet substantial, step would be to improve the collection and availability of information about citizen complaints and the subsequent investigations. Missing and incorrect data about the complaints process only contributes to the sense that review boards operate in a way opaque to public input and scrutiny. This perception can affect trust in the process whether it is ultimately fair or not.

After it is confirmed that a bias exists, a second step would be to determine the cause of the bias. Are citizens of different demographics less able to file a complaint and provide evidence in a way that leads to a sustained accusation? Do the individual investigators harbor explicit or implicit racial or demographic biases? Are there physical or geographic barriers involved with investigations that disproportionately affect differently locations, and therefore, different demographics? These are all potential reasons for the potential bias we found in Stage 2. In order to design an effective intervention, we need to know what the actual reason or reasons are.

Additionally, this kind of evaluation could be performed regularly by the police department or independent oversight body. As the institutions with the best access to this data and the most at stake in the outcome, organizations such as IPRA or its successor would benefit from routinely evaluating the complaint process. Any indication that there are routine differences in the outcomes of investigations based on complainant information should lead to review and reform by the responsible body.

Ultimately the goal should be to increase the transparency of the investigation process. Police accountability is an important part of building trust between officers and communities they serve. A fully independent investigatory body with input from citizens would go a long way towards restoring faith in the how the police department responds to citizen complaints. Many of the steps taken in the last month reflect this need to improve the public opinion of the Chicago Police Department and its oversight.

Limitations, Caveats, and Future Work

One of the major limitations of our work proved to be the quality and completeness of the data. As discussed above, the original dataset was much larger than the subset we ended up working with and training models on. Outside of the time window that we used, most of the data was very unreliable and incomplete.

Another significant challenge is data leakage, or the possibility that certain variables associated with a complaint may have hidden information about the ultimate outcome of the case. One example is the complaint category feature, which indicates what type of misconduct is alleged in the complaint or if the type is unknown. After data exploration, it became clear that cases are only assigned a complaint category after an affidavit is signed. Knowing that a case has a category also tells you that the case made it past the affidavit stage, but this data would not be available before that event had occurred. Therefore, having that information is an unfair way of "seeing into the future." We addressed this particular issue by eliminating the "Category" features when predicting whether a complaint would be dropped due to no affidavit, but it is likely that there are other sources of data leakage that we were unable to detect without more complete insight into the complaint adjudication process and how data is gathered throughout.

Additionally, it is difficult to isolate the effect of demographic information. Many features that might at first glance appear to be independent of demographics, such as the number of pothole requests opened in the last six months nearby, may have correlations with race or other demographic indicators. This is of particular concern in Chicago, where neighborhoods are highly segregated by race.[13] Almost any information related to geographic location is likely correlated with race.

Another barrier is the imbalanced nature of this dataset. Sustained complaints are a very rare outcome, making up approximately 3% of all complaints. By simply predicting an unsustained outcome all the time, the model can achieve a high accuracy rate of about 97%. Predicting rare events is a notorious challenge in machine learning

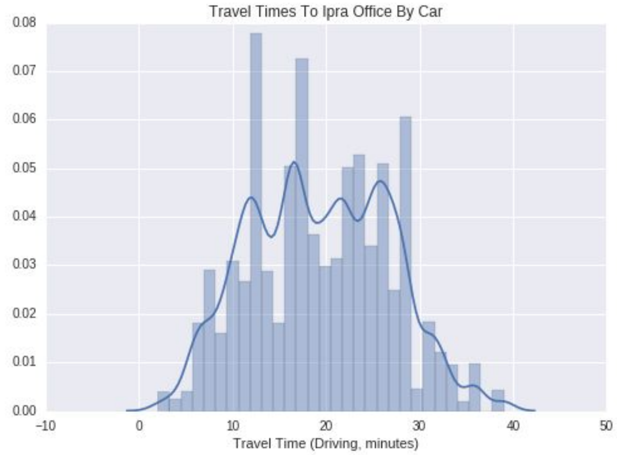
[13] "Segregation declines in Chicago, but city still ranks high, census data show", Chicago Tribune, January 4, 2016

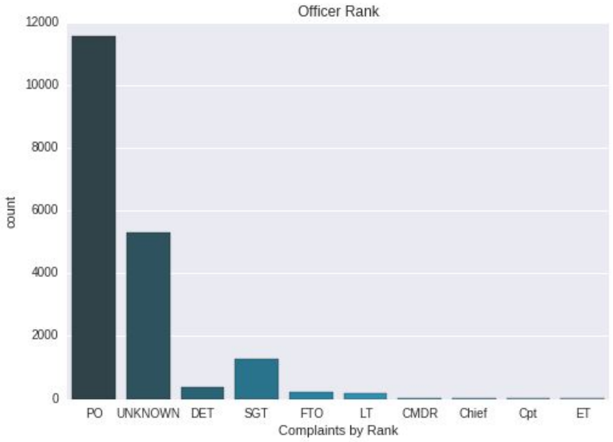
and recent research has explored methods to improve precision and recall in predicting rare events. One technique is known as oversampling, in which cases from the underrepresented class are replicated to artificially “balance” the dataset. The model is trained on this fabricated dataset that contains a higher proportion of the minority class.

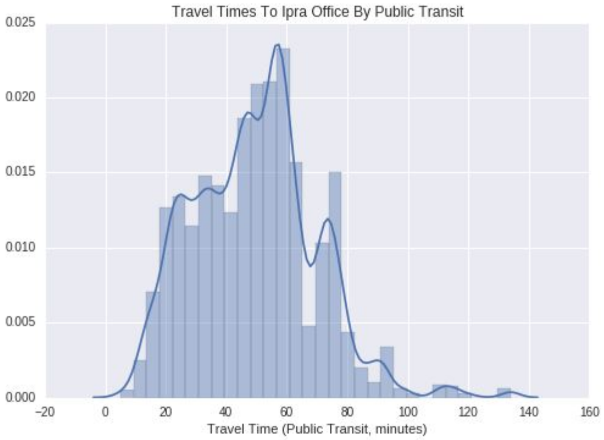
As mentioned above, future work should focus on acquiring more and better data about police complaints. Models would greatly benefit from additional data, especially given the imbalance of sustained complaints in the current dataset.

Appendix

A. Additional Graphs and Summary Data







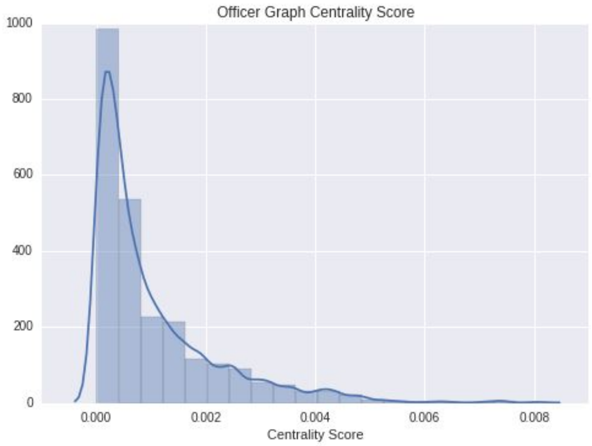


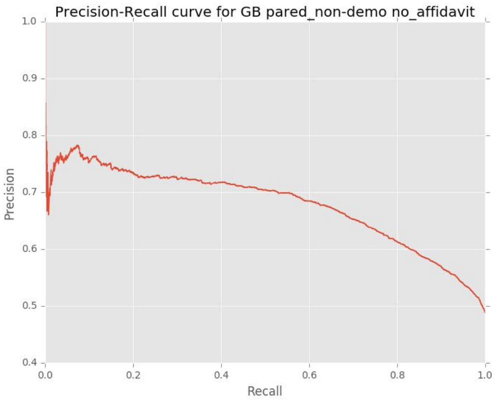
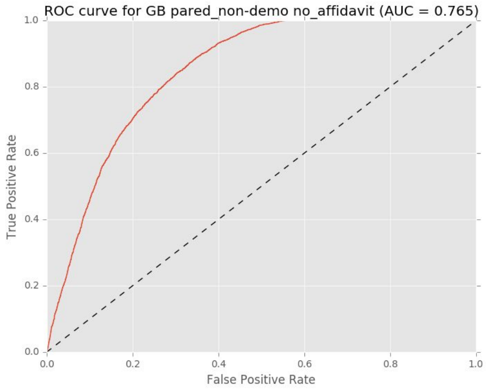
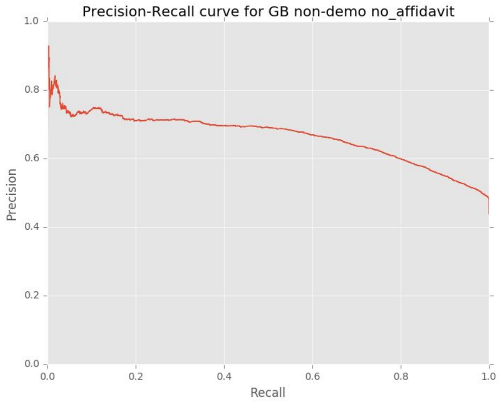
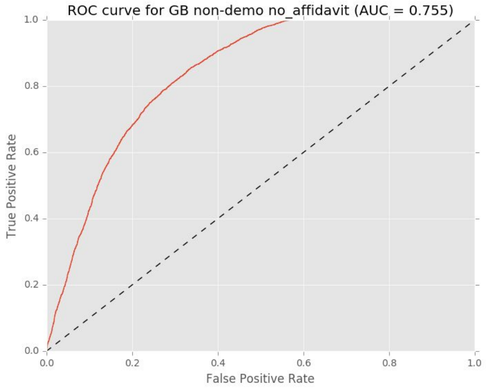
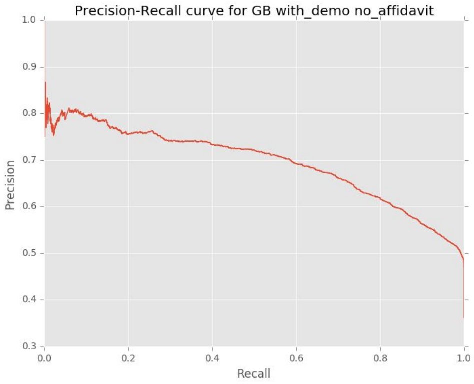
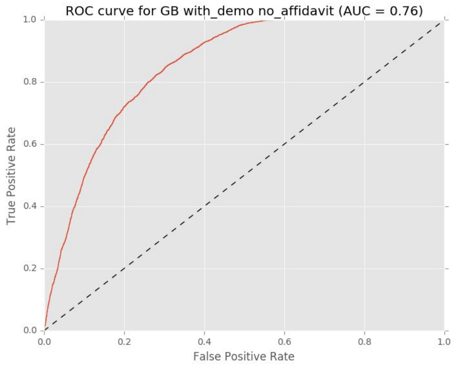
Table 2: Removed Rows and Imputed Values

Description	**Reason**	**Rows Affected**	**Rows Remaining**
Table: Allegations	Pre-Cleaning	N/A	56,384
Table: Allegations **Column:** Incident_Date **Criteria:** Deleted Null Rows	Many of our features are generated based on the time of the incident. Further, we are only looking at the period between 2011 - 2014. Without the incident date, we had no idea of knowing if the incident fell within the time frame. Lastly, for rows without an incident date, there was also very little other information associated with the allegation.	27,796	28,588
Table: Allegations **Column:** Incident_Date **Criteria:** Deleted Years > 2014	The most reliable data comes from the period from March 2011 - December 2014. Records from before that period are sparse, and records from after that period contain what seem like duplicates from the 2011 - 2014 period.	2,438	26,150
Table: Allegations **Column:** Officer ID **Action:** Changed officer null values to 0	Changed Officer Null Values to 0 in order to set primary keys on allegations table to crid and officer_id.	7,008	26,150
Table: Allegations **Column:** City **Action:** Deleted null rows	Since the bulk of features are location-based, we removed records where there was no complete address associated with the complaint.	7,336	18,814
Table: Allegations **Column:** Address **Action:** Deleted null rows	Deleted rows with complaint address outside of Chicago (for example, some addresses were in Wisconsin)	119	18,695

Table 3: Complete List of Data Sources

	Source	Description	Year
1	Chicago City Data Portal	311 Data: Abandoned Property	2011 - 2014
2	Chicago City Data Portal	311 Data: Abandoned Vehicles	2011 - 2014
3	Chicago City Data Portal	311 Data: Alley Lights	2011 - 2014
4	Chicago City Data Portal	311 Data: Garbage Carts	2011 - 2014
5	Chicago City Data Portal	311 Data: Graffiti	2011 - 2014
6	Chicago City Data Portal	311 Data: Potholes	2011 - 2014
7	Chicago City Data Portal	311 Data: Rodent Baiting	2011 - 2014
8	Chicago City Data Portal	311 Data: Sanitation Code Complaints	2011 - 2014
9	Chicago City Data Portal	311 Data: Street Lights	2011 - 2014
10	Chicago City Data Portal	311 Data: Tree Debris	2011 - 2014
11	Chicago City Data Portal	311 Data: Tree Trims	2011 - 2014
12	Chicago City Data Portal	311 Data: Rodent Baiting	2011 - 2014
13	US Census Bureau	ACS Census Demographic Data	2012
14	Chicago City Data Portal	Chicago Police Beat Shapefiles	2012
15	Chicago City Data Portal	Chicago Census Tract Shapefiles	2012
16	Chicago City Data Portal	Chicago Crime Data	2011 - 2014
17	FOIA/The Invisible Institute	Chicago Police Complaint/Allegation Data	2011 - 2014

Table 4: Comparison Between Best Strict Non-Demographic Model Performance on Stage 1 Task

Precision/Recall Curves: *Predicting Stage 1 Outcome (no affidavit) using Gradient Boosting model with Random Oversampling selected for best performance on Strict Non-Demographic Data*	**ROC Curves:** *Predicting Stage 1 Outcome (no affidavit) using Gradient Boosting model with Random Oversampling selected for best performance on Strict Non-Demographic Data*
Performance using Strict Non-demographic Data  	*Performance using Strict Non-demographic Data* 
Performance using Non-demographic Data  	*Performance using Non-demographic Data* 
Performance using Demographic Data  	*Performance using Demographic Data*  ROC curve for GB with_demo no_affidavit (AUC = 0.76) True Positive Rate False Positive Rate 1.0 0.8 0.6 0.4 0.2 0.0 0.0 0.2 0.4 0.6 0.8 1.0