NYPD Replication

Greg Stoddard   Dylan J. Fitzpatrick   John Greer

This repository contains code to recreate support analysis from our research paper “Predicting Police Misconduct”. In our paper, we study how well a data-driven machine learning system can predict serious instances of police misconduct. The goal of such a system, often called an early intervention system, is to identify officers who exhibit warning signs of a serious negative event and intervene before with support and services before that comes to pass. While our paper focuses primarily on the Chicago Police Department (CPD), this repo uses public data from the New York Police Department (NYPD) to recreate and corroborate the main findings. Read below for a high-level description of results. For more information, please see the paper or contact us.

Data / Methodology

We study this in NYPD using complaint records from New York City’s Citizen Complaint Review Board (CCRB) (which investigates external complaints against NYPD officers), lawsuit data recorded by NYC’s law department, and a roster of NYPD officers. We constructed a panel dataset where the features for each observation includes everything known about an officer’s behavior up to and including year and the outcomes include whether the officer was involved in any negative outcomes in year T and T+1. We focus on prediction years from 2015-2019. The final dataset has between 33,000 - 36,000 observations per year, resulting in a total of 175,000 officer-year observations.

We constructed features (covariates) from an officer’s history of CCRB complaints and lawsuits from the five years prior to the date of prediction. An officer’s history of complaints is represented by the total number of complaints, total number of complaints by type, number of complaints by finding, and complaints by type and finding (eg “number of sustained excessive force complaints”). An officer’s history of lawsuits is represented by the total number of lawsuits, the number of lawsuits by type, and the total monetary payouts in their lawsuits. Each set of features were measured over the prior year, the prior two years, and prior five years in order to allow the models to weight factors differently based on how recently they occurred. In total, there are 117 features per observation.

We defined two outcomes in the NYPD data. The first outcome is whether an officer has a sustained CCRB complaint during the outcome period, which is analogous to the CPD on-duty misconduct outcome since the CCRB only investigates complaints related to use of force, abuse of authority, discourteous policing, and offensive language. This is a rare outcome - only 3% to 4.5% of officers have a sustained complaint over these two-year outcome periods. The second outcome is whether an officer was named in a lawsuit whose payout is $50,000 or greater during the outcome period. For simplicity, we refer to these as expensive lawsuits. This outcome is also rare - only 1-2% of officers are named in an expensive lawsuit over these two-year outcome periods.

As a process note, most datasets have `tax_id` as the unique officer identifier. The only exception is the payroll data which only has names. When we merge that data into the rest of the data, we match on names and drop any names that are duplicates. The result is that officers who have conflict on first-and-last names will not have a known start date. However, in those cases, we're able to impute an officer's start date from the complaint data. Hence we'll have a start date for any officers whose name is unique or who've had a complaint at some point in their career

Results

Although the NYPD is more limited than the Chicago data in terms of the types of prior datasets available (for example, the public NYPD dataset does not have use of force data), we generally find the same conclusions regarding the levels of predictability, the similar performance between machine learning models and simple ranking policies, and the predictive value of prior non-sustained complaints.

Predictability of Machine Learning Models

We begin by documenting how well the machine learning models built on NYPD data predict future sustained complaints and future expensive lawsuits. The plots below show the recall curves of the machine learning models for each outcome. The sustained complaints model shows a similar level of predictability to the CPD models, with the top 5% of officers by predicted risk accounting for 18.6% of all officers who have a sustained complaint in the follow-up period. The 'expensive lawsuit' model shows a higher degree of predictability, with the top 5% of officers by predicted risk accounting for nearly 30% of officers named in an expensive lawsuit during the outcome period.

Comparison of Machine Learning Models and Simple Models

We next compared simple policies like ranking officers either by the prior number of complaints or the prior number of lawsuits to machine learning models. The first table below shows that rank-by-complaints (RBC) captures nearly as many future sustained complaints as the machine learning model that was trained to predict sustained complaints. The gap in performance between ML and RBC is smaller in NYC than Chicago, potentially due to the fact that there are fewer types of officer behavior data in the public NYPD data. The rank-by-lawsuits (RBL) policy, on the other hand, is significantly worse than both RBC and ML.

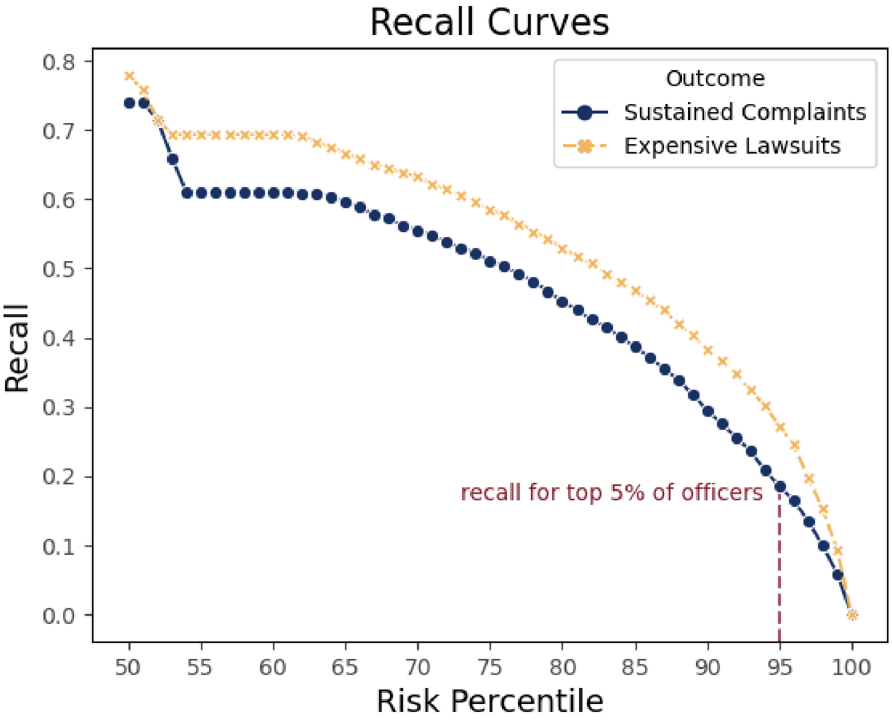

Figure 1: Recall Plots

	method	outcome	recall	true_positives
0	rank by complaints	sustained complaints	0.18	193.00
1	rank by lawsuits	sustained complaints	0.11	120.04
2	machine learning	sustained complaints	0.19	202.80

Machine Learning vs. Simple Policy - Sustained Complaints Outcome

The table below shows the machine learning model trained to predict future expensive lawsuits is more accurate than either the RBC or RBL policies, but both policies successfully identify a high-risk group of officers. The top of 5% of officers by predicted lawsuit risk account for 27.5% of all officers named in a future lawsuit (a rate that's nearly 6x times higher than the average officer) while the top 5% of officers identified by RBL or RBC account for 21 or 22% of officers named in an expensive lawsuit (which is roughly 4x higher than the average officer).

	method	outcome	recall	true_positives
0	rank by complaints	expensive lawsuits	0.22	149.96
1	rank by lawsuits	expensive lawsuits	0.20	138.16
2	machine learning	expensive lawsuits	0.27	185.00

Machine Learning vs. Simple Policy - Expensive Lawsuits Outcome

Simple ranking policies are fairly competitive with complex machine learning models. This may be particularly important for small police departments that might lack the sample size or resources to build a complex predictive model.

Predictive Value of Non-Sustained Complaints

We finally repeated the experiment of testing whether records of non-sustained complaints have predictive value. Specifically, we constructed two statistical models for each outcome. The "all complaints" model used information derived from all complaints filed against an officer in the five years prior to the date of prediction, while the "only sustained complaints" model only had access to information from complaints that were sustained.

This experiment shows that non-sustained complaints carry predictive signals about the risk, which echoes our findings on the CPD data. When flagging the top 5% of officers by predicted risk of a future sustained complaint, the 'All complaints' model correctly flags 18.6% of officers with a sustained complaint in the outcome period while the 'Only sustained complaints' model flags 11.4% of officers with a sustained complaint during the outcome period. This is a relative drop in accuracy of 38%, and translates to 80 fewer correctly flagged officers per year. Similarly, the machine learning model trained to predict future expensive lawsuits suffers a relative drop in accuracy of 38% when limiting prior features to only sustained complaints, which translates to 56 fewer correctly flagged officers per year.

	method	outcome	recall	true_positives
0	Model with all prior complaints	sustained complaints	0.19	202.60
1	Model with only sustained complaints	sustained complaints	0.11	119.00

Predictive Value of Sustained Complaints - Sustained Complaints Outcome

	method	outcome	recall	true_positives
0	Model with all prior complaints	expensive lawsuits	0.24	164.80
1	Model with only sustained complaints	expensive lawsuits	0.15	99.20

Predictive Value of Sustained Complaints - Expensive Lawsuits Outcome