Repurposing recidivism models for forecasting police officer use of force

Samira Khorshidi[*], Jeremy G. Carter[*], George Mohler[*]

[*] Indiana University - Purdue University Indianapolis
(sakhors, carterjg, gmohler)@iupui.edu

***Abstract*—We review several concepts and modeling techniques from statistical and machine learning that have been developed to forecast recidivism. We show how these methods might be repurposed for forecasting police officer use of force. Using open Chicago police department use-of-force complaint data for illustration, we discuss feature engineering, construction of black-box models, interpretable forecasts, and fairness.**

***Index Terms*—Use of force, Forecasting, Risk assessment**

I. Introduction

A number of statistical and machine learning methods for forecasting recidivism and informing parole decisions have been developed over the past decade. Decision tree based models such as random forests [2] are used to forecast risk of recidivism and have been shown to lead to a reduction in re-arrests [3]. Interpretable recidivism forecasts [17] provide easy to read risk "score cards" as an alternative to black-box models, given the high-stakes nature of the application. Important questions of fairness arise in constructing models of recidivism. Research has shown that false positive rates are higher for Black individuals in models used in practice [7] and methods have been constructed to mitigate bias [4]. While there has been some debate on the matter, the latest research indicates that algorithms have the potential to be more accurate and fair than humans in making parole decisions [9].

A problem that is mathematically similar, though one that has received less attention from the machine learning community, is forecasting police officer use of force, complaints, and misconduct. Recent research has shown that officers with negative marks on their record (complaints, firearm discharges, etc.) are 3 times more likely to shoot in the line of duty [13]. Other research has shown that officer shootings, misconduct and use of force exhibit network effects, where officers are at greater risk of being involved in these incidents when they socialize with officers who have a history of misconduct and complaints [12, 15, 18]. Statistical methods including survival models [12], logistic regression [18], and point processes [15] have been recently introduced for modeling officer-involved use of force and shootings.

Our goal in this paper is to review some of the machine learning techniques and concepts from criminal justice forecasting and show how they might be repurposed for officer risk assessments. We first provide an overview of the data we use in this paper in Section II. In Section III, we discuss feature engineering of risk factors that are correlated with use of force complaints. In Section IV, we investigate and compare the performance of several standard machine learning models for forecasting use of force complaints. In Section V, we show how interpretable models can be constructed of officer risk assessments and in Section VI we discuss issues of fairness that may arise when constructing models for officer complaints and misconduct.

II. Data on officer use of force in Chicago

In the current study we analyze use of force complaint data from the Chicago Police Department collected through the Citizen's Police Data Project [1]. The data includes demographic features of each officer (age, gender, race), demographic features of the complainant, type of complaint, date the incident occurred, and identifiers for all officers named in the complaint. Overall, we analyze data through the end of 2017 containing 32,351 complaints of excessive use of force involving 14,914 officers. In many cases, multiple officers are named in a single complaint which allows the construction of officer co-complaint networks [12, 18], such as the network shown in Figure 1.

III. Risk factors associated with use of force complaints

To illustrate the construction of officer risk assessment modeling, we use three types of features: 1) demographic features such as officer age, sex and race, 2) network features

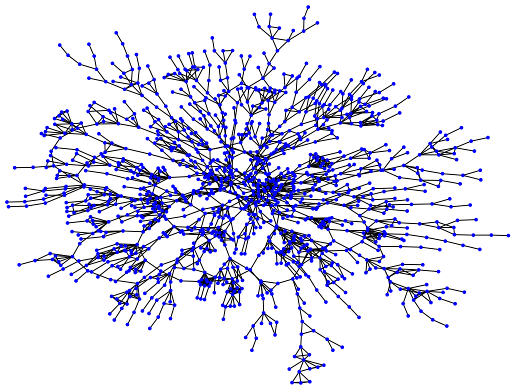

Fig. 1: Network where links between officers with more than 1 common complaint are shown.

978-1-7281-6251-520$31.00 ©2020 IEEE

that capture the complaint history of officers in a target officer's social network [12, 18] and 3) Hawkes point process features that capture increases in risk following past incidents [10, 15].

A Hawkes process is defined in terms of its conditional intensity, which can be defined for officer i as,

$$\lambda_i(t) = \mu_i + \sum_{t>t_j^i} \theta w \exp(-w(t - t_j^i)). \tag{1}$$

Here j indexes the use of force incidents in which officer i was involved, μ_i is a baseline rate of incidents, θ is a parameter determining by how much the intensity is elevated after each event and w determines the time scale over which elevated risk decays back to the baseline rate μ_i.

As is done in [10], we take a supervised learning approach and define point process features for a logistic regression (GLM) estimated each month m to forecast the number of excessive use of force complaints for each officer in the following month (using all data historically available).

$$f^i(w, m) = \sum_{m>t_j^i} \theta w \exp(-w(m - t_j^i)) \tag{2}$$

The feature for $f^i(w, m)$ is defined for each officer in the month m currently being forecasted with varying $w = .1, .01, .001$. The coefficients θ in Equation 1 are estimated within the logistic regression.

Similarly, network point process features can be defined by summing point processes over neighboring nodes in the social network:

$$g^i(w, m) = \sum_{m>t_j^k;k\sim i} \theta w \exp(-w(m - t_j^k)). \tag{3}$$

Here $k \sim i$ indicates that k and i are neighboring nodes in the officer excessive use of force network (e.g. officer k and i were co-complainants at some point prior to month m). in addition to the point process features and demographics, we also include count based features including the number of complaints in the past year and number of total historical complaints.

We fit a GLM model to the data to forecast monthly complaints from 2010 to 2017, where the first 50 months are used for training and subsequent months are used for evaluation. In Table I we show the estimated coefficients for the GLM model forecasting the number of complaints for each officer in each month. The point process feature ($w = .001$), neighbor point process feature ($w = .1$), sex, age, and total complaint count are all highly statistically significant. None of the race indicator variables are statistically significant indicating that officers that are male, younger, and have a history of complaints are most likely to have excessive use of force complaints in the near future.

In Figure 2 we display a ROC curve for the GLM model of monthly officer complaints. In Chicago, 1% of officers account for 16% of complaints and 10% of officers account for 72% of complaints.

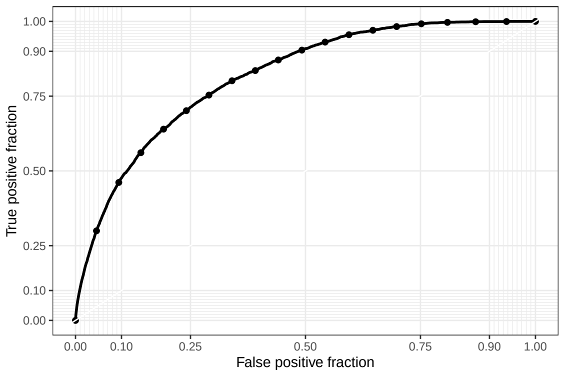

Fig. 2: ROC curve for GLM using Hawkes process, co-complaint, and demograhic features.

TABLE I: Estimated coefficients of GLM model of officer excessive use of force.

	estimate	st. err.	z val.	p-val.
(Intercept)	-1.7038	0.1291	-13.1937	$< 10^{-4}$
neighbor complaint count	-0.0004	0.0032	-0.1291	0.8973
neighbor Hawkes ($w = .1$)	0.2629	0.1000	2.6276	0.0086
neighbor Hawkes ($w = .01$)	-0.0919	0.0429	-2.1424	0.0322
neighbor Hawkes ($w = .001$)	-0.0030	0.0095	-0.3209	0.7483
neighbor complaint/year	0.0055	0.0105	0.5264	0.5986
complaint count	0.0265	0.0072	3.6777	0.0002
Hawkes ($w = .1$)	-0.6086	0.3807	-1.5987	0.1099
Hawkes ($w = .01$)	0.3682	0.1553	2.3714	0.0177
Hawkes ($w = .001$)	0.4662	0.0315	14.8065	$< 10^{-4}$
complaint count year	-0.0368	0.0475	-0.7741	0.4389
Female	-0.3360	0.0534	-6.2858	$< 10^{-4}$
Black	0.1818	0.1053	1.7276	0.0841
Hispanic	0.1234	0.1043	1.1836	0.2366
Native American	0.3161	0.2580	1.2253	0.2205
Race Unknown	-9.0305	72.5484	-0.1245	0.9009
White	0.0448	0.1014	0.4422	0.6583
Age	-0.0858	0.0022	-39.8419	$< 10^{-4}$

IV. Machine learning based forecasts

Next we evaluate several standard machine learning models for the purpose of forecasting monthly excessive use of force complaints. To our knowledge there has been limited work in constructing machine learning models (e.g. decision trees, neural networks) for officer risk assessments, with only one previous study that focuses on using random forests [5].

Here we utilize the H2O Auto-ML workflow [8], which is a unified interface for parameter tuning and comparison of a variety of models including random forest, boosting, and feed-forward neural networks. In comparison to logistic regression, decision tree ensembles (such as random forest and boosting) capture nonlinear interactions that may be present in the data through a sequence of binary if/else conditions that recursively partition the data into risk groups. Feed-foward neural networks also capture non-linear effects in the data, although they do so through a single model (rather than an ensemble) comprised of compositional layers that alternate between a linear mapping and an non-linear sigmoidal activation operation.

We trained autoML using the demographic features, the number of historical complaints, and point processe features to forecast the number of complaints in the each month of the training data. In the experiment we specified the maximum number of models, excluding the Stacked Ensemble models, to be 25. The autoML training took 7 hours and 48 minutes to run on a Lenovo P51s system with 16gb of RAM. In Table II we display the leaderboard results outputted by autoML after training, where we find the best performing model was boosted decision trees (GBM) with the best AUC, logloss, RMSE and MAE. Here we find that the AUC=.82 is much better than random chance (AUC=.5), indicating that the algorithm can correctly identify the more riskier officer out of a randomly chosen pair 82% of the time. We note that the ranking of models in Table II is meant for illustrative purposes and the results depend on the parameters explored by autoML. The leaderboard rankings would be subject to change if autoML was used with a higher maximum number of models or an alternative parameter grid search was employed.

V. Intepretable models

One drawback of the machine learning methods in the previous section is that they are black-box models that can be difficult to explain and interpret. To address this concern in high-stakes applications such as forecasting recidivism, an interpretable "risk-slim" model was introduced in [14] that yields simple, yet accurate risk score cards. The method works by first generating a large candidate set of binary features X and then finding a super-sparse set of integer coefficients θ that solve the following optimization problem:

$$\min_{\theta \in \mathbb{Z}^n} \sum_{i=1}^{N} 1[y_i X_i^T \theta \leq 0] + C_0 \|\theta\|_0 + C_1 \|\theta\|_1.$$

Here the 0-norm encourages sparsity, whereas the 1-norm encourages the coefficients to be small. The problem can be cast as a mixed-integer linear programming problem and solved efficiently using CPLEX [6].

In the training risk-slim algorithm, we processed our dataset by binarizing all categorical and real-valued features. Certain features such as sex and race are already in this form. For real-valued feature f, we defined disjoint intervals U_j covering the domain of f and then constructed binary features $x_f^j = 1\{f \in U_j\}$ based on a feature being contained or not in a given interval.

In Table III we display an example score card for risk-slim trained on Chicago excessive use of force data. The model outputs an interpretable score that is computed by answering yes or no to a series of 5 questions. For example, if an officer is age 30, has 7 complaints total, 3 per year, and 8 complaints per year for officers in their social network, then the score is $3 + 0 + 0 - 2 + 0 = 1$. In Table III we also see that the score is highly correlated with empirical risk, for example for officers with a score of -4 only .03% have a complaint that month, whereas 23% of officers have a complaint the next month when their score is 3.

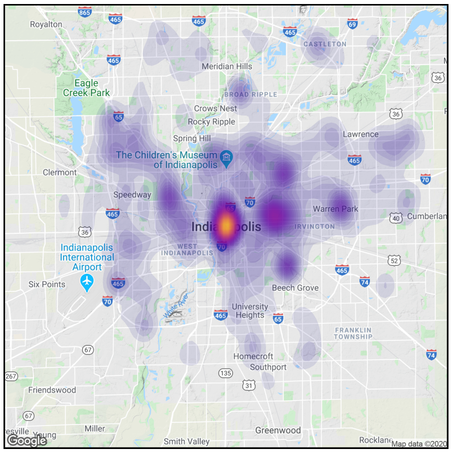

Fig. 3: Density of use of force incidents in Indianapolis, which are concentrated Downtown and on the near east side.

VI. Fairness

Reasearch into fairness of machine learning models of recidivism has shown that models can be heavily biased, with false positive rates mismatched across racial or other protected groups [7]. Similar mismatches in false positive rates or other fairness metrics have the potential to arise in forecasts of officer use of force and misconduct. For example, suppose we define the high-risk threshold for the GLM model in Section III above to be $p = .3$ and that we define the protected variable to be age (either above or below 35). In that case we find the false positive rate of the model for age≥35 to be 80.0%, whereas the false positive rate for officers with age<35 is 89.8%. In this case the model is biased against younger officers and is more likely to flag them as higher risk when they are not.

However, one can argue that the stakes, at least internally (with respect to the officer) are lower, where a high risk score might result, for example, in additional procedural justice training. In comparison, a false positive in the case of a parole decision leads to much greater harm to the individual who is kept imprisoned, even though they would not have committed a crime upon release. We also note that internal fairness considerations (within the police department) need to be weighed against bias of outcomes that are external to the department. In officer involved fatal shootings, Black civilians are twice as likely as white civilians to be unarmed [11]. In Indianapolis, Black citizens are the subject of 53% of use of force incidents, while comprising 28% of the population. Furthermore these incidents are concentrated in certain geographies in the city (see Figure 3). New algorithms are likely needed for balancing external, city-wide fairness of policing with internal fairness of risk assessments.

VII. Conclusion

We reviewed several machine learning concepts and methods and showed how they can be repurposed for forecasting police excessive use of force. Using features on officer

TABLE II: Auto-ML leaderboard results when trained on officer excessive use of force.

Method	**AUC**	**Logloss**	**RMSE**	**MSE**
GBM-grid–1-AutoML	$\mathbf{8.22 \cdot 10^{-1}}$	$\mathbf{1.61 \cdot 10^{-2}}$	$\mathbf{4.93 \cdot 10^{-2}}$	$\mathbf{2.43 \cdot 10^{-3}}$
GBM-5-AutoML	$8.00 \cdot 10^{-1}$	$1.62 \cdot 10^{-2}$	$4.96 \cdot 10^{-2}$	$2.46 \cdot 10^{-3}$
StackedEnsemble-AllModels-AutoML	$7.94 \cdot 10^{-1}$	$1.75 \cdot 10^{-2}$	$4.97 \cdot 10^{-2}$	$2.47 \cdot 10^{-3}$
GLM-1-AutoML	$7.79 \cdot 10^{-1}$	$1.70 \cdot 10^{-2}$	$4.94 \cdot 10^{-2}$	$2.44 \cdot 10^{-3}$
DRF-1-AutoML	$7.33 \cdot 10^{-1}$	$1.99 \cdot 10^{-2}$	$5.16 \cdot 10^{-2}$	$2.67 \cdot 10^{-3}$
StackedEnsemble-BestOfFamily-AutoML	$7.29 \cdot 10^{-1}$	$1.81 \cdot 10^{-2}$	$4.96 \cdot 10^{-2}$	$2.46 \cdot 10^{-3}$
XRT-1-AutoML	$7.16 \cdot 10^{-1}$	$2.02 \cdot 10^{-2}$	$5.12 \cdot 10^{-2}$	$2.62 \cdot 10^{-3}$
DeepLearning-1-AutoML	$6.48 \cdot 10^{-1}$	$1.91 \cdot 10^{-2}$	$5.16 \cdot 10^{-2}$	$2.67 \cdot 10^{-3}$
DeepLearning-grid–3-AutoML	$5.12 \cdot 10^{-1}$	$2.01 \cdot 10^{-2}$	$\mathbf{4.93 \cdot 10^{-2}}$	$\mathbf{2.43 \cdot 10^{-3}}$

TABLE III: Interpretable model of officer excessive use of force risk.

Age ≤ 40	3 points	+.....
Age < 55 and ≥ 40	2 points	+.....
# complaints ≤ 5	-1 points	+.....
# complaints < 5 per year	-2 points	+.....
# complaints in officer network < 5 per year	-1 points	+.....
ADD POINTS FROM ROWS 1 to 5	SCORE	= .....

Score	-4	-3	-2	-1	0	1	2	3
Risk	$< 1\%$	$< 1\%$	$< 1\%$	1%	2%	5%	16%	23%

demographics, social networks, and complaint history, we showed that black-box models can capture a high percentage of complaints when flagging a small percentage of officers. Super-sparse integer models can be constructed that produce easy-to-interpret score cards. Finally, issues of fairness may arise both internally within the police department, as well as externally in terms of disparate impacts of policing towards certain groups and geographical areas.

Recent research has shown that procedural justice training can reduce use of force incidents by 6.4% and complaints by 10% [16]. Statistical and machine learning models may help facilitate procedural justice and implicit bias training, by helping to determine when officers need more training or other interventions to reduce risk of excessive use of force.

Future research should focus on incorporating dynamics into officer risk assessments, as risk has been shown to propagate over time in social networks [15], designing methods that can mitigate overall (city-wide) bias and harm caused by policing, while balancing fairness across citizen groups, geographies, and internally within a police department, and transitioning research to practice through field implementations and trials.

VIII. Acknowledgements

Support from NSF grants SCC-1737585 and ATD-1737996.

References

[1] Citizens police data project, 2020.

[2] Richard Berk. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

[3] Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, 2017.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[5] Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 67–76, 2016.

[6] IBM ILOG Cplex. V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53):157, 2009.

[7] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[8] H2O.ai. *H2O AutoML*, June 2017. H2O version 3.30.0.1.

[9] Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652, 2020.

[10] George Mohler and Michael D Porter. Rotational grid, pai-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(5):227–236, 2018.

[11] Justin Nix, Bradley A Campbell, Edward H Byers, and Geoffrey P Alpert. A bird's eye view of civilians killed by police in 2015: Further evidence of implicit bias. *Criminology & Public Policy*, 16(1):309–340, 2017.

[12] Marie Ouellet, Sadaf Hashimi, Jason Gravel, and Andrew V Papachristos. Network exposure and excessive use of force: Investigating the social transmission of police misconduct. *Criminology & Public Policy*, 18(3):675–704, 2019.

[13] Greg Ridgeway. Officer risk factors associated with police shootings: a matched case–control study. *Statistics and Public Policy*, 3(1):1–6, 2016.

[14] Berk Ustun, Stefano Traca, and Cynthia Rudin. Supersparse linear integer models for interpretable classifica-

tion. *arXiv preprint arXiv:1306.6677*, 2013.

[15] George Wood, Daria Roithmayr, and Andrew V Papachristos. The network structure of police misconduct. *Socius*, 5:2378023119879798, 2019.

[16] George Wood, Tom R Tyler, and Andrew V Papachristos. Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences*, 117(18):9815–9821, 2020.

[17] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

[18] Linda Zhao and Andrew V Papachristos. Network position and police who shoot. *The ANNALS of the American Academy of Political and Social Science*, 687(1):89–112, 2020.