Document Title:	Predicting Recidivism Fairly: A Machine Learning Application Using Contextual and Individual Data
Author(s):	Eric L. Sevigny, Ph.D., Thaddeus L. Johnson, Ph.D., Jared A. Greathouse, MS
Document Number:	305036
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

Predicting Recidivism Fairly: A Machine Learning Application Using Contextual and Individual Data

Final Report for the NIJ Recidivism Forecasting Challenge

Eric L. Sevigny, PhD
Thaddeus L. Johnson, PhD
Jared A. Greathouse, MS

Georgia State University

August 31, 2021

Introduction

Community corrections agencies across the U.S. supervise individuals who have been diverted from incarceration to probation or released from prison to parole. Too often, caseloads are large and resources are thin. Services provided to these individuals, including the balance of surveillance and support programming, must be linked to their specific risks and needs to effectively promote desistence from crime. Too much of a poorly targeted intervention or too little of an urgently needed program will do little to improve reintegration and reduce recidivism, and may even make matters worse. Predicting recidivism risk accurately and fairly is therefore critical to effective community corrections sentencing, release, and programming decisions.

Risk assessment instruments typically used by community corrections agencies focus on individual-level risk factors, including prior contacts with the criminal justice system. This reliance can have disparate impacts on minority individuals, whose criminal justice contacts and broader social opportunities may be equally influenced by contextual and individual factors. In an effort to improve the accuracy and fairness of risk assessments, this report directly responds to NIJ's Recidivism Forecasting Challenge by utilizing additional contextual factors characterizing the communities where supervised individuals live. To accomplish this, we integrated a range of contextual metrics with the individual-level Georgia parolee dataset supplied by NIJ, including community-level measures of environment and health, housing, economic opportunity, quality of life, and social and government services. We then assessed and validated various machine learning algorithms using both individual and contextual factors to predict recidivism outcomes. In our validation efforts, we focused on maximizing predictive accuracy across gender and race according to NIJ's judgement criteria. We found that a lasso model performed best overall. This report describes our methodological approach and results for Year 1 recidivism outcomes in the "Small Team" category (our team did not submit Year 2 or 3 forecasts).

Relevant Literature

Prior research on recidivism risk primarily focuses on the influence of individual-level factors, often in context of supervision or reentry programs (Aos et al., 2007; Bonta & Andrews, 2017; Lowenkamp et al., 2010). This body of research suggests that returning to crime or violating supervision conditions are influenced by personal factors such as criminal history, sex, race, and age (Piquero et al., 2013; Wang et al., 2010). However, by definition, static individual-level characteristics are poor targets for intervention and potentially introduce bias into risk assessments.

Berk and Elzarka (2020) observed that criminal justice risk assessments relying on static factors are often conflated with macrolevel processes that could result in inaccurate and unfair risk assessments. For example, disproportionate police contact with residents from disadvantaged neighborhoods is commonly blamed for disparate arrest rates among people of color (Grogger & Ridgeway, 2006; Ridgeway, 2006). Given that police resources and surveillance are regularly concentrated in disorganized areas characterized by structural inequities and criminogenic conditions, which also tend to be places where racial minorities live, work, and play (Braga et al., 2019), disparities embedded in criminal justice data can be carried forward in assessments of recidivism risk.

Further highlighting the importance of contextual factors in understanding recidivism, the social disorganization literature shows community conditions increase recidivism risk in places where justice-involved persons are heavily concentrated (Hipp et al., 2011). According to Berk and Elzarka (2020), incorporating this environmental backcloth into the risk assessment process could achieve more fair and accurate outcomes. Despite this, geographic variation in recidivism remains relatively underexplored. Limited research suggests a nexus of macrolevel factors influences offending through social control mechanisms (Barnett & Mencken, 2002; Mears et al., 2008). For instance, social disorganization may reduce levels of formal and informal social control (Sampson & Groves, 1989). Consequently, justice-involved persons living in socially disorganized areas may have greater challenges remaining crime-free.

While some studies consider potential macrolevel influences in this process (Clark, 2016; Grunwald et al., 2010; Huebner & Pleggenkuhle, 2015; Mears et al., 2008; Onifade et al., 2011; Tillyer & Vose, 2011; Wright et al., 2014), they almost exclusively rely on census-based social disorganization indicators of poverty, racial/ethnic diversity, residential mobility, and family disruption. However, these measures reflect a narrow purview of the larger environmental backdrop. Absent from much of the research integrating community measures is how prosocial places and services, local institutions, and criminogenic establishments impact recidivism risk. Better accounting of these factors would more accurately capture the nature of person-environment interactions within communities.

Gaps in informal social control are often related to the breakdown of community social ties (Shaw & McKay, 2010). Importantly, the presence of local organizations and neighborhood businesses is critical to developing these social bonds (Bursik & Grasmick, 1993; Sampson & Groves, 1989; Sharkey et al., 2017; Triplett et al., 2003; Wo, 2016; Wo et al., 2016). Researchers posit that local institutions mediate effects of structural features on community crime through increased opportunities to congregate, cultivate relationships, and build social cohesion. For example, Sharkey et al. (2017) report long-term decreases in neighborhood crime rates are associated with increases in the number of local nonprofits. Similar benefits have been seen with other institutions, including civic organizations (Lee, 2008), recreation centers (Peterson et al., 2000), coffee shops (Wo, 2016), and religious organizations (Beyerlein & Hipp, 2005).

This is not to say all organizations perform equally across different communities (Slocum et al., 2013).For instance, Drawve and McNeeley’s (2021) study of Minnesota parolees finds that prosocial places (i.e., churches, employment services, and civil and social organizations) were negatively associated with recidivism, but that this relationship was weaker in disadvantaged neighborhoods. Conversely, Doucet and Lee (2015) found the presence of civic institutions decreased violent crime only in impoverished urban neighborhoods.

Research also shows that interventions targeting dynamic risk factors, such as employment or substance misuse problems, can significantly improve reentry and rehabilitation outcomes, even among high-risk individuals. For example, Newton et al.'s (2016) recent systematic review finds that vocational and employment training programs produce particularly positive outcomes among younger and higher-risk parolees and those who access available services sooner upon release. However, they caution that such employment readiness programs must be part of a broader nexus of local services, programs, and opportunities to maximize the odds of successful reentry. Universities and community colleges, for instance, are known to be integral to local and regional workforce development (O'Banion, 2019). In that spirit, Wallace (2015) finds when communities lose educational institutions, neighborhood recidivism increases. She suggests that educational institution losses negatively impact employability of returning citizens and, in turn, the likelihood of reoffending.

Not all institutions are protective in nature. While small businesses are key to cultivating sustainable local economies (Lanning, 2020), neighborhood establishments such as bars and alcohol outlets (Bernasco & Block, 2010; Madensen & Eck, 2008; Peterson et al., 2000), pawnshops (Kubrin & Hipp, 2016), restaurants (Askey et al., 2017), hotels/motels (Krupa et al., 2019), and stadiums (Kurland et al., 2013) can increase both perpetration and victimization opportunities. For example, motivated offenders often seek criminal opportunities in retail centers because of large numbers of potential targets (Brantingham & Brantingham, 1995).

Beyond criminogenic spaces, environmental pollutants may also play an important role in crime risk. Prolonged exposure to toxins has a strong connection to criminality (Boutwell et al., 2017; Feigenbaum & Muller, 2016), and childhood lead exposure has been linked to crime at the census tract level (Boutwell et al., 2016). Communal research also shows that access to nature or greenspace improves social cohesion and may exert a crime-suppressing effect (Branas et al., 2016; Mitchell &

Popham, 2008; Shepley et al., 2019). Still, opposite effects have also been observed (Abu-Lughod, 2006; Kim and Hipp, 2018; Li 2008).

Given that most justice-involved persons return to their pre-prison communities (Clear, 2009; Harding et al., 2013), filling the knowledge gap about the role of these broader contextual factors in recidivism takes on added urgency. Further, a major criticism of many risk assessment approaches concerns the overreliance on static risk factors and a failure to take place-based and community resources into account (Hanson, 2018). Although such factors may predict future recidivism, they are inherently poor targets for intervention. Therefore, a deeper understanding of how geographic attributes matter for people under community supervision may improve the accuracy and fairness of risk assessments.

Methods

NIJ released several datasets in multiple stages as part of the Recidivism Forecasting Challenge, including training and test datasets for predicting recidivism. Here, we describe our methodology for forecasting Year 1 recidivism outcomes.

Data and Measures

Georgia Parolee Supervision Dataset

We imported individual-level Year 1 parolee data into Stata MP 16.1 and cleaned it for analysis, including destringing variables and attaching value labels. Along with each parolee's unique ID, place of residence, and recidivism outcomes, 31 individual-level risk factor variables were included in the Year 1 dataset. For each parolee, NIJ mapped their place of residence to the Public Use Microdata Area (PUMA) and then collapsed neighboring PUMAs into 25 larger spatial units (henceforth "Super PUMA" [SPUMA]) to protect parolee confidentiality. Among the 31 individual-level risk measures, four contained missing data. Of these, we excluded *conviction offense* and *risk score* from the set of potential predictors either because of a high proportion of missing data or strong collinearity with other measures. We addressed missingness on the other two variables as follows. First, because gang affiliation was reported only for male parolees, we assumed no verified gang affiliation among women and recoded this indicator as false (0) for female parolees. Second, for the parolee's assigned supervision level, we added an attribute for 'missing' when this was not reported. We performed identical data management operations on both the training and test datasets provided by NIJ. See Table 1, numbered rows 0-29 for variable descriptions.

Added Variables from External Sources

We broadly searched for potentially relevant place-based social, economic, institutional, and environmental risk and protective indicators to merge with the Georgia parolee supervision dataset. We scanned prior literature, examined federal and state agency data tables and reports, and queried various public data repositories. Three criteria informed our data search. First, data needed to be open access or freely available with the understanding that community supervision agencies may have limited resources for data acquisition. Second, data needed to be published serially with the understanding that community risk factors are dynamic and mutable. Third, the data period had to be contemporaneous with the years 2013-2015. We ultimately included 67 contextual measures from a variety of data sources. All measures were aggregated and matched at the SPUMA level of analysis.

The U.S. Center for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI)[1] supplied data on socially vulnerable populations. SVI data are publicly available for the following years: 2000, 2010, 2014, 2016, and 2018. For this study, we used SVI 2014 data, which are based on 2010-2014 American Community Survey (ACS) 5-year estimates (Centers for Disease Control and Prevention & Agency for Toxic Substances and Disease Registry, 2017). We downloaded census tract data for Georgia, and calculated mean indicator values by SPUMA for 15 reported SVI 2014 measures. Note that 11 of 1,966 census tracts contained missing data on *per capita income* (ep_pci), so these did not factor into the aggregate measure. See Table 1, rows numbered 30-44 for SVI variable descriptions.

[1] https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

The U.S. Environmental Protection Agency (EPA) Environmental Justice Screening and Mapping Tool[2] supplied data on community environmental hazards. These data are available for 2015-2020. We used 11 indicators from the 2016 EJSCREEN dataset measuring census tract level air, water, and toxic waste hazards (Environmental Protection Agency, 2017). Most indicators are derived from various EPA toxics monitoring databases for reference years 2011-2015. The exceptions are the *lead paint index*, which is based on the estimated percentage of pre-1960 housing stock from the 2010-2014 ACS, and the *traffic proximity and volume index*, which is based on 2014 US Department of Transportation traffic data. We downloaded census tract data, and then calculated mean indicator values by SPUMA for the 11 reported 2016 EJSCREEN indicators. See Table 1, numbered rows 45-55 for EJSCREEN variable descriptions.

The U.S. Bureau of Transportation Statistics (BTS) Local Area Transportation Characteristics for Households (LATCH) dataset[3] supplied estimates of daily household travel distance (Bureau of Transportation Statistics, 2017). These data were released in 2009 and 2017. The 2017 LATCH data on household transportation statistics were derived using small area estimation techniques based on the 2017 National Household Transportation Survey (NHTS) along with 2012-2016 ASC 5-year estimates. We downloaded census tract data for 2017, and then calculated the mean value by SPUMA for the one measure derived from this source: estimated person miles traveled per day. See Table 1, numbered row 56 for the variable description.

The U.S. Department of Housing and Urban Development (HUD) Affirmatively Furthering Fair Housing Data and Mapping Tool (2017)[4] supplied five socioeconomic and opportunity indexes. AFFH-T data are published periodically, with the most recent sixth iteration released in 2020. We used the third

[2] https://www.epa.gov/ejscreen/download-ejscreen-data
[3] https://www.bts.gov/latch/latch-data
[4] https://www.hud.gov/program_offices/fair_housing_equal_opp/affh

data release for this study. We downloaded census tract data, and then calculated the mean values by SPUMA for the five indexes. See Table 1, numbered rows 57-61 for variable descriptions.

The U.S. Department of Agriculture (USDA) Food Access Research Atlas (2017)[5] supplied a single measure of food accessibility. FARA data were previously published in 2006, 2010, 2015, and 2019. For this study, we used the 2015 FARA, based on 2015 food store directories (i.e., Store Tracking and Redemption System; Trade Dimensions TDLinx) and 2010-2015 ASC 5-year estimates (USDA 2017). We downloaded census tract data, and then calculated the mean value by SPUMA for our selected measure of food accessibility. See Table 1, numbered row 62 for the variable description.

The National Neighborhood Data Archive (NaNDA) is an open data repository from which we obtained a range of contextual measures across different substantive domains.[6] Specifically, we merged NaNDA data on transit stops (Clarke & Melendez, 2019), eating and drinking places (Esposito et al., 2019a), religious and civic/social organizations (Esposito et al., 2019b), law enforcement organizations (Esposito et al., 2020), social services (Finlay et al., 2020d), parks/greenspace (Clarke et al., 2020) healthcare services (Khan et al., 2020), grocery stores (Finlay et al., 2020b), alcohol package stores (Finlay et al., 2020c), dollar stores (Gomez-Lopez et al., 2020), educational services (Finlay et al., 2020a), urbanicity (Miller et al., 2020), and demographics and socioeconomic status (Melendez et al., 2020). In each case, we normalized the measures by population or land area. See Table 1, numbered rows 63-96 for variable descriptions.

Analytic Approach to Supervised Learning and Validation

We explored a range of supervised machine learning algorithms across different software platforms, including Stata, R, Python, and Weka (Brownlee, 2016; Ho et al., 2021; Larose & Larose,

[5] https://www.ers.usda.gov/data-products/food-access-research-atlas/
[6] https://www.openicpsr.org/openicpsr/search/nanda/studies#

2019). Ultimately, we performed analyses using Stata 16.1 MP due to author familiarity with the software. For some commands, Stata served as a wrapper for Python or Java.

The training dataset released by NIJ contained 18,028 observations. To validate model performance, we created a random 70:30 split of the NIJ dataset into gender- and race-balanced training (*n*=13,521) and validation (*n*=4,507) datasets. We assessed the performance of five different machine learning algorithms to predict recidivism outcomes in the validation dataset. These include regularized logit regression (lasso and elastic net), random forests, and boosting (adaptive and gradient).

Lasso logit implements a regularization method for selecting and fitting covariates in a model (Tibshirani, 1996). In selecting a best subset of predictors by forcing coefficients of some variables to zero, lasso seeks to reduce model complexity and overfitting as a solution to the bias-variance tradeoff. More complex models are penalized, reflected by higher parameter values of Λ, as lasso iterates toward a parsimonious solution that minimizes out-of-sample prediction error. We use 10-fold cross-validation (CV) to select the best performing prediction model, implemented using the **`lasso logit`** command in Stata.

Elastic net logit is a generalization of the lasso that may be superior when multicollinearity among predictors is high (Zou & Hastie, 2005). Rather than forcing coefficients of poorly performing predictors exactly to zero, elastic net adds an additional penalty term, α, that shrinks correlated coefficients rather than forcing them to zero completely. When $\alpha = 1$, elastic net resolves to lasso. We also use 10-fold CV to select the best performing prediction model, implemented using the **`elastic net logit`** command.

The random forest algorithm is an ensemble method based on a large number of underlying classification decision trees. Subsets of data and predictors are randomly drawn to build many trees, and the algorithm averages these predictions across the multiple trees (Schonlau & Zou, 2020). Boosting is another ensemble classification method seeking to improve prediction by combining weaker learners so they become stronger. Additional technical details on these various algorithms, and their specific application to recidivism prediction, can be reviewed in several recent treatments (Duwe & Kim, 2015; Ghasemi et al., 2020; Wang et al., 2020). We implement random forest using the **`rforest`** and **`pyforest`** commands, adaptive boosting using **`pyadaboost`**, and gradient boosting using **`pygradboost`**. In all models, we applied the default parameter settings. To assess predictive performance, we applied NIJ criteria for gender-specific accuracy and racial fairness.

Results

Validation Models and Results

Table 2, panel A summarizes the various algorithms and their performance in the validation exercise. The first thing to note is the elastic net produces identical results to the lasso because the model resolved with $\alpha = 1$, which is the default parameter for the lasso. In other words, the more complex and computationally intensive elastic net model offered no advantages over the lasso. Second, lasso outperformed all other models on every performance measure except the male fairness and accuracy index, for which it performed worst.

Based on lasso's overall performance, we selected it as our base prediction model. However, we also assessed alternate training strategies using different subgroups to try to improve prediction outcomes. First, since many risk assessment instruments are gender-specific, we trained and validated the models separately for males and females. Second, following the suggestion of Berk and Elzarka (2020), we trained the model on White parolees and then predicted outcomes for both Blacks and Whites. The rationale is that Blacks can take advantage of "White privilege" embedded in the data. The validation results are presented in Table 2, panel B. In general, subgroup modeling by gender does not improve prediction performance, and may even diminish it. The White trained model performs slightly worse for males and slightly better for females. The biggest differences occur on the fairness and accuracy indexes. For males, the model outperforms the other algorithms. However, the female fairness and accuracy index loses all of its previous advantage. The results are therefore mixed and warrant further investigation. Based on these results, we chose the base lasso model for our Recidivism Forecasting Challenge submission.

Final Model and Results

Final model estimates were developed on the full training sample (*n*=18,028) with a candidate set of 171 covariates. The number of covariates is a function of how lasso enters covariates into the model and how we treated censored numeric variables. First, lasso includes the full set of collinear indicators among the set of candidate predictors. If it did not, and a base category mattered for prediction, then the indicators for all other attributes would need to be included in the model to capture that effect, resulting in an unnecessarily more complex model. Second, we treated censored continuous measures (e.g., 1, 2, 3 or more) as categorical variables because we assumed important individual differences would be masked if we treated top-coded measures as fully interval. Lasso automatically standardizes covariates to have mean 0 and standard deviation 1 to prevent the scale of the covariates from influencing estimation.

We estimated a *k*-fold cross-validated lasso logit model, with *k* = 10. The CV function for logit, the CV mean deviance, is a measure-of-fit statistic that, when minimized, signals the best fitting model. As shown in Figure 1, the lasso grid search across values of lambda (ʎ) achieved a minimum CV mean deviance = 1.1106 at ʎ = 0.0015, yielding a final model with 73 non-zero covariates. The covariates and their standardized coefficients are reported in Table 3. Most of the individual-level covariates have coefficients in the direction we would expect. In general, having fewer prior criminal justice contacts, including arrests and convictions, is associated with lower recidivism, and vice versa. Our decision to treat censored numeric covariates as categorical appears warranted, given the various censored coefficients are always included and tend to be of much higher magnitude than the preceding uncensored value.

Notably, only 12 contextual covariates entered the final model. Several of these measures predicted successful parole outcomes at Year 1, including crowded housing conditions, environmental pollutants (ozone concentration in air and proximity to various pollution sources), racial/ethnic concentrations of poverty, and the number of jails/prisons and emergency relief services in the community. In contrast, several contextual measures predicted higher likelihood of rearrest, including proximity to risk management plan facilities and the number of coffee shops and educational institutions in the community. It is difficult to directly interpret some of these factors—indeed some are perplexing—but it is likely they are correlated with other unobserved factors, especially the findings that link the environmental health of communities to individual-level recidivism outcomes.

We applied this model to predict recidivism outcomes in the test dataset, using Brier scores and a fairness and accuracy index based on differences in Black and White false positive rates. Specifically, the Brier score is calculated by

$$BS = \sum_{t=1}^{n}(f_t - A_t)^2 / N$$

where N is the number of individuals, f_t is the predicted probability of recidivism for individual t, and A_t is the actual recidivism outcome for individual t. Lower Brier scores are better. The fairness and accuracy index (FAI) is calculated by

$$FAI = (1 - BS)(1 - |FP_B - FP_W|)$$

where BS is the Brier score, FP_B is the false positive rate for Blacks, and FP_W is the false positive rate for Whites. Higher FAI scores are better. Table 4 reports the scores for each challenge and our team's place in that challenge.

Discussion

This challenge raised important questions about current practices of recidivism risk assessment for justice-involved persons. Concerns with the improving the accuracy and fairness of these assessments is paramount. We marshalled a broad array of contextual measures and applied a range of machine learning algorithms to the task. The results of our modeling appear promising and warrant further investigation. We feel the metrics and thresholds governing the challenge were reasonable and valid.

Although we understand the need to protect confidentiality and how that informed the aggregation of the parolee's home address into larger spatial areas (i.e., "SPUMAs"), that decision precluded accounting for factors within finer geographic areas such as census tracts. We feel that incorporating contextual data at these more granular levels will benefit future modeling efforts.

Future Considerations

Dividing the challenge into different categories was reasonable, but the low level of participation by students may warrant some reconsideration. For example, running the contest over the summer months may have limited opportunities for student engagement, especially following a period of remote learning at most universities due to Covid-19. Allowing more time between data release and required submission of challenge materials may also have improved participation levels. Our team was constrained by other demands and did not submit Year 2 or Year 3 predictions, though we had planned to when we started the contest.

With respect to future challenges, our team sees value in performing rapid systematic reviews. For emerging crime and justice areas, rapid scoping reviews would be worthwhile. For domains where a critical mass of new experimental or quasi-experimental research has been published, a rapid meta-analysis would be able to summarize that research. Although it would involve considerable effort, we feel a systematic review challenge along these lines would be beneficial to the broader research and policy communities.

References

Aos, S., Miler, M., & Drake, E. (2007). Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates. *Federal Sentencing Reporter, 19*(4), 275-290. https://doi.org/10.1525/fsr.2007.19.4.275

Askey, A. P., Taylor, R., Groff, E., & Fingerhut, A. (2017). Fast food restaurants and convenience stores: Using sales volume to explain crime patterns in seattle. *Crime & Delinquency, 64*(14), 1836-1857. https://doi.org/10.1177/0011128717714792

Barnett, C., & Mencken, F. C. (2002). Social disorganization theory and the contextual nature of crime in nonmetropolitan counties. *Rural Sociology, 67*(3), 372-393. https://doi.org/10.1111/j.1549-0831.2002.tb00109.x

Berk, R., & Elzarka, A. A. (2020). Almost politically acceptable criminal justice risk assessment. *Criminology & Public Policy, 19*(4), 1231-1257. https://doi.org/10.1111/1745-9133.12500

Bernasco, W., & Block, R. (2010). Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. *Journal of Research in Crime and Delinquency, 48*(1), 33-57. https://doi.org/10.1177/0022427810384135

Beyerlein, K., & Hipp, J. R. (2005). Social capital, too much of a good thing? American religious traditions and community crime. *Social Forces, 84*(2), 995-1013. https://doi.org/10.1353/sof.2006.0004

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct*. Routledge.

Boutwell, B. B., Nelson, E. J., Emo, B., Vaughn, M. G., Schootman, M., Rosenfeld, R., & Lewis, R. (2016). The intersection of aggregate-level lead exposure and crime. *Environmental Research, 148*, 79-85. https://doi.org/https://doi.org/10.1016/j.envres.2016.03.023

Boutwell, B. B., Nelson, E. J., Qian, Z., Vaughn, M. G., Wright, J. P., Beaver, K. M., . . . Rosenfeld, R. (2017). Aggregate-level lead exposure, gun violence, homicide, and rape. *PLOS ONE, 12*(11), e0187953. https://doi.org/10.1371/journal.pone.0187953

Braga, A. A., Brunson, R. K., & Drakulich, K. M. (2019). Race, place, and effective policing. *Annual Review of Sociology, 45*(1), 535-555. https://doi.org/10.1146/annurev-soc-073018-022541

Branas, C. C., Kondo, M. C., Murphy, S. M., South, E. C., Polsky, D., & MacDonald, J. M. (2016). Urban blight remediation as a cost-beneficial solution to firearm violence. *American Journal of Public Health, 106*(12), 2158-2164. https://doi.org/10.2105/ajph.2016.303434

Brantingham, P., & Brantingham, P. (1995). Criminality of place. *European Journal on Criminal Policy and Research, 3*(3), 5-26. https://doi.org/10.1007/BF02242925

Brownlee, J. (2016). *Machine learning mastery with Weka: Analyze data, develop models, and work through projects*. Machine Learning Mastery.

Bureau of Transportation Statistics. (2017). *Local area transportation characteristics for households data*. https://tinyurl.com/ye4758hb

Bursik, R. J., & Grasmick, H. G. (1993). Economic deprivation and neighborhood crime rates, 1960-1980. *Law & Society Review, 27*(2), 263-283. https://doi.org/10.2307/3053937

Centers for Disease Control and Prevention, & Agency for Toxic Substances and Disease Registry. (2017). *Social Vulnerability Index, 2014, database Georgia*. https://tinyurl.com/yfgblqjx

Clark, V. A. (2016). Predicting two types of recidivism among newly released prisoners:First addresses as “launch pads” for recidivism or reentry success. *Crime & Delinquency, 62*(10), 1364-1400. https://doi.org/10.1177/0011128714555760

Clarke, P., & Melendez, R. (2019). *National Neighborhood Data Archive (NANDA): Public transit stops by census tract* (United States, 2016-2018; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E111109V1

Clarke, P., Melendez, R., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Parks by census tract* (United States, 2016-2018; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E117921V1

Clear, T. (2009). Incarceration and communities. *Criminal Justice Matters, 75*(1), 26-27. https://doi.org/10.1080/09627250802699749

Department of Housing and Urban Development. (2017). *Affirmatively furthering fair housing*. https://tinyurl.com/yfbm86y3

Doucet, J. M., & Lee, M. R. (2015). Civic communities and urban violence. *Social Science Research, 52*, 303-316. https://doi.org/10.1016/j.ssresearch.2015.01.014

Drawve, G., & McNeeley, S. (2021). Recidivism and community context: Integrating the environmental backcloth. *Journal of Criminal Justice, 73*. https://doi.org/10.1016/j.jcrimjus.2021.101786

Duwe, G., & Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review, 28*(6), 570-600. https://doi.org/10.1177/0887403415604899

Environmental Protection Agency. (2017). *Ejscreen: Environmental justice screening and mapping tool* Version 2017-02-06) [distributor] Environmental Protection Agency. https://tinyurl.com/yduozyue.

Esposito, M., Li, M., Finlay, J., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2019a). *National Neighborhood Data Archive (NANDA): Eating and drinking places by census tract* (United States, 2016-2018; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E115404V2

Esposito, M., Li, M., Finlay, J., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2019b). *National Neighborhood Data Archive (NANDA): Religious, civic, and social organizations by census tract* (United States, 2016-2018; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E115967V2

Esposito, M., Li, M., Finlay, J., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Law enforcement organizations by census tract* (United States, 2016-2018; Version V3) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E115973V3

Feigenbaum, J. J., & Muller, C. (2016). Lead exposure and violent crime in the early twentieth century. *Explorations in Economic History, 62*, 51-86. https://doi.org/https://doi.org/10.1016/j.eeh.2016.03.002

Finlay, J., Li, M., Esposito, M., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2020a). *National Neighborhood Data Archive (NANDA): Education and training services by census tract* (United States, 2003-2017; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E127681V1

Finlay, J., Li, M., Esposito, M., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2020b). *National Neighborhood Data Archive (NANDA): Grocery stores by census tract* (United States, 2003-2017; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E123001V1

Finlay, J., Li, M., Esposito, M., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2020c). *National Neighborhood Data Archive (NANDA): Liquor, tobacco, and convenience stores by census tract* (United States, 2003-2017; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E123541V1

Finlay, J., Li, M., Esposito, M., Gomez-Lopez, I., Khan, A., Clarke, P., & Chenoweth, M. (2020d). *National Neighborhood Data Archive (NANDA): Social services by census tract* (United States, 2016-2018; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E117163V2

Ghasemi, M., Anvari, D., Atapour, M., Stephen wormith, J., Stockdale, K. C., & Spiteri, R. J. (2020). The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice and Behavior, 48*(4), 518-538. https://doi.org/10.1177/0093854820969753

Gomez-Lopez, I., Esposito, M., Li, M., Finlay, J., Khan, A., Clarke, P., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Dollar stores by census tract* (United States, 2003-2017; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E123802V1

Grogger, J., & Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association, 101*(475), 878-887. https://doi.org/10.1198/016214506000000168

Grunwald, H. E., Lockwood, B., Harris, P. W., & Mennis, J. (2010). Influences of neighborhood context, individual history and parenting behavior on recidivism among juvenile offenders. *Journal of Youth and Adolescence, 39*(9), 1067-1079. https://doi.org/10.1007/s10964-010-9518-5

Hanson, R. K. (2018). Long-term recidivism studies show that desistance is the norm. *Criminal Justice and Behavior, 45*(9), 1340-1346. https://doi.org/10.1177/0093854818793382

Harding, D. J., Morenoff, J. D., & Herbert, C. W. (2013). Home is hard to find: Neighborhoods, institutions, and the residential trajectories of returning prisoners. *The ANNALS of the American Academy of Political and Social Science, 647*(1), 214-236. https://doi.org/10.1177/0002716213477070

Hipp, J. R., Jannetta, J., Shah, R., & Turner, S. (2011). Parolees' physical closeness to social services: A study of California parolees. *Crime & Delinquency, 57*(1), 102-129. https://doi.org/10.1177/0011128708322856

Ho, A. T. Y., Huynh, K. P., Jacho-Chávez, D. T., & Rojas-Baez, D. (2021). Data science in Stata 16: Frames, lasso, and python integration. *Journal of Statistical Software, 98*. https://doi.org/10.18637/jss.v098.s01

Huebner, B. M., & Pleggenkuhle, B. (2015). Residential location, household composition, and recidivism: An analysis by gender. *Justice Quarterly, 32*(5), 818-844. https://doi.org/10.1080/07418825.2013.827231

Khan, A., Li, M., Finlay, J., Esposito, M., Gomez-Lopez, I., Clarke, P., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Health care services by census tract* (United States, 2016-2018; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E120907V2

Krupa, J. M., Boggess, L. N., Chamberlain, A. W., & Grubesic, T. H. (2019). Noxious housing: The influence of single room occupancy (SRO) facilities on neighborhood crime. *Crime & Delinquency, 67*(9), 1404-1428. https://doi.org/10.1177/0011128719875701

Kubrin, C. E., & Hipp, J. R. (2016). Do fringe banks create fringe neighborhoods? Examining the spatial relationship between fringe banking and neighborhood crime rates. *Justice Quarterly, 33*(5), 755-784. https://doi.org/10.1080/07418825.2014.959036

Kurland, J., Johnson, S. D., & Tilley, N. (2013). Offenses around stadiums: A natural experiment on crime attraction and generation. *Journal of Research in Crime and Delinquency, 51*(1), 5-28. https://doi.org/10.1177/0022427812471349

Lanning, K. (2020). A case for american economic reform: Small businesses and inclusive economies. *Local Development & Society, 1*(1), 5-14. https://doi.org/10.1080/26883597.2020.1833549

Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*. Wiley.

Lee, M. R. (2008). Civic community in the hinterland: Toward a theory of rural social structure and violence. *Criminology, 46*(2), 447-478. https://doi.org/10.1111/j.1745-9125.2008.00115.x

Lowenkamp, C. T., Makarios, M. D., Latessa, E. J., Lemke, R., & Smith, P. (2010). Community corrections facilities for juvenile offenders in Ohio: An examination of treatment integrity and recidivism. *Criminal Justice and Behavior, 37*(6), 695-708. https://doi.org/10.1177/0093854810363721

Madensen, T. D., & Eck, J. E. (2008). Violence in bars: Exploring the impact of place manager decision-making. *Crime Prevention and Community Safety, 10*(2), 111-125. https://doi.org/10.1057/cpcs.2008.2

Mears, D. P., Wang, X., Hay, C., & Bales, W. D. (2008). Social ecology and recidivism: Implications for prisoner reentry. *Criminology, 46*(2), 301-340. https://doi.org/10.1111/j.1745-9125.2008.00111.x

Melendez, R., Clarke, P., Khan, A., Gomez-Lopez, I., Li, M., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Socioeconomic status and demographic characteristics of census tracts* (United States, 2008-2017; Version V2) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E119451V2

Miller, S., Melendez, R., & Chenoweth, M. (2020). *National Neighborhood Data Archive (NANDA): Urbanicity by census tract* (United States, 2010; Version V1) [distributor] Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E130542V1

Mitchell, R., & Popham, F. (2008). Effect of exposure to natural environment on health inequalities: An observational population study. *The Lancet, 372*(9650), 1655-1660. https://doi.org/10.1016/S0140-6736(08)61689-X

Newton, D., Day, A., Giles, M., Wodak, J., Graffam, J., & Baldry, E. (2016). The impact of vocational education and training programs on recidivism: A systematic review of current experimental evidence. *International Journal of Offender Therapy and Comparative Criminology, 62*(1), 187-207. https://doi.org/10.1177/0306624X16645083

O'Banion, T. U. (2019). A brief history of workforce education in community colleges. *Community College Journal of Research and Practice, 43*(3), 216-223. https://doi.org/10.1080/10668926.2018.1547668

Onifade, E., Petersen, J., Bynum, T. S., & Davidson, W. S. (2011). Multilevel recidivism prediction: Incorporating neighborhood socioeconomic ecology in juvenile justice risk assessment. *Criminal Justice and Behavior, 38*(8), 840-853. https://doi.org/10.1177/0093854811407026

Peterson, R. D., Krivo, L. J., & Harris, M. A. (2000). Disadvantage and neighborhood violent crime: Do local institutions matter? *Journal of Research in Crime and Delinquency, 37*(1), 31-63. https://doi.org/10.1177/0022427800037001002

Piquero, A. R., Jennings, W. G., Diamond, B., & Reingle, J. M. (2013). A systematic review of age, sex, ethnicity, and race as predictors of violent recidivism. *International Journal of Offender Therapy and Comparative Criminology, 59*(1), 5-26. https://doi.org/10.1177/0306624X13514733

Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology, 22*(1), 1-29. https://doi.org/10.1007/s10940-005-9000-9

Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology, 94*(4), 774-802. https://doi.org/10.1086/229068

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3-29. https://doi.org/10.1177/1536867X20909688

Sharkey, P., Torrats-Espinosa, G., & Takyar, D. (2017). Community and the crime decline: The causal effect of local nonprofits on violent crime. *American Sociological Review, 82*(6), 1214-1240. https://doi.org/10.1177/0003122417736289

Shaw, C. R., & McKay, H. D. (2010). Juvenile delinquency and urban areas: A study of rates of delinquency in relation to differential characteristics of local communities in American cities

(1969). Martin A. Andresen, Paul J. Brantingham, & J. Bryan Kinney (Eds.), *Classics in environmental criminology*, (pp. 87-124). CRC Press. https://doi.org/10.4324/9781439817803-9

Shepley, M., Sachs, N., Sadatsafavi, H., Fournier, C., & Peditto, K. (2019). The impact of green space on violent crime in urban environments: An evidence synthesis. *International Journal of Environmental Research and Public Health, 16*(24), 5119.

Slocum, L. A., Rengifo, A. F., Choi, T., & Herrmann, C. R. (2013). The elusive relationship between community organizations and crime: An assessment across disadvantaged areas of the south bronx. *Criminology, 51*(1), 167-216. https://doi.org/10.1111/1745-9125.12001

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tillyer, M. S., & Vose, B. (2011). Social ecology, individual risk, and recidivism: A multilevel examination of main and moderating influences. *Journal of Criminal Justice, 39*(5), 452-459. https://doi.org/10.1016/j.jcrimjus.2011.08.003

Triplett, R. A., Gainey, R. R., & Sun, I. Y. (2003). Institutional strength, social control and neighborhood crime rates. *Theoretical Criminology, 7*(4), 439-467. https://doi.org/10.1177/13624806030074003

U.S. Department of Agriculture. (2017). *Food environment atlas*. https://tinyurl.com/ybeabno3

Wallace, D. (2015). Do neighborhood organizational resources impact recidivism? *Sociological Inquiry, 85*(2), 285-308. https://doi.org/https://doi.org/10.1111/soin.12072

Wang, J., Hu, J., Shen, S., Zhuang, J., & Ni, S. (2020). Crime risk analysis through big data algorithm with urban metrics. *Physica A: Statistical Mechanics and its Applications, 545*. https://doi.org/10.1016/j.physa.2019.123627

Wang, X., Mears, D. P., & Bales, W. D. (2010). Race-specific employment contexts and recidivism. *Criminology, 48*(4), 1171-1211. https://doi.org/10.1111/j.1745-9125.2010.00215.x

Wo, J. C. (2016). Community context of crime:A longitudinal examination of the effects of local institutions on neighborhood crime. *Crime & Delinquency, 62*(10), 1286-1312. https://doi.org/10.1177/0011128714542501

Wo, J. C., Hipp, J. R., & Boessen, A. (2016). Voluntary organizations and neighborhood crime: A dynamic perspective. *Criminology, 54*(2), 212-241. https://doi.org/10.1111/1745-9125.12101

Wright, K. A., Kim, B., Chassin, L., Losoya, S. H., & Piquero, A. R. (2014). Ecological context, concentrated disadvantage, and youth reoffending: Identifying the social mechanisms in a sample of serious adolescent offenders. *Journal of Youth and Adolescence, 43*(10), 1781-1799. https://doi.org/10.1007/s10964-014-0173-0

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Table 1. Variables Used in Analysis

No.	Description	Variable Name	Source	Format	Values
Outcome					
0	Recidivism arrest in year 1	recid_yr1	NIJ	Binary	0 = no, 1 = yes
Predictors					
1	Gender	gender	NIJ	Binary	0 = female, 1 = male
2	Race	race	NIJ	Binary	0 = White, 1 = Black
3	Age at release	ager	NIJ	Ordinal	1 = 18-22, 2 = 23-27, 3 = 28-32, 4 = 33-37, 5 = 38-42, 6 = 43-47, 7 = 48+
4	Education level	educ	NIJ	Ordinal	1 = less than HS diploma, 2 = HS diploma, 3 = at least some college
5	# dependents	dependents	NIJ	Censored Numeric	0, 1, 2, 3+
6	Prison years	prisyrs	NIJ	Ordinal	1 = less than 1 year, 2 = 1-2 years, 3 = greater than 2 to 3 years, 4 = more than three years
7	Gang affiliated	gang	NIJ	Binary	0 = false, 1 = true
8	First supervision level	level	NIJ	Nominal	1 = standard, 2= high, 3 = specialized, 4 = missing
9	# prior felony arrests	pa_felony	NIJ	Censored Numeric	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10+
10	# prior misdemeanor arrests	pa_misd	NIJ	Censored Numeric	0, 1, 2, 3, 4, 5, 6+
11	# prior violent arrests	pa_violent	NIJ	Censored Numeric	0, 1, 2, 3+
12	# prior property arrests	pa_property	NIJ	Censored Numeric	0, 1, 2, 3, 4, 5+
13	# prior drug arrests	pa_drug	NIJ	Censored Numeric	0, 1, 2, 3, 4, 5+
14	# prior technical violation arrests	pa_techv	NIJ	Censored Numeric	0, 1, 2, 3, 4, 5+
15	Any prior domestic violence arrest	pa_dv	NIJ	Binary	0 = false, 1 = true
16	Any prior gun charge arrest	pa_gun	NIJ	Binary	0 = false, 1 = true
17	# prior felony convictions	pc_felony	NIJ	Censored Numeric	0, 1, 2, 3+
18	# prior misdemeanor convictions	pc_misd	NIJ	Censored Numeric	0, 1, 2, 3, 4+
18	Any prior violent conviction	pc_violent	NIJ	Binary	0 = false, 1 = true
20	# prior property convictions	pc_property	NIJ	Censored Numeric	0, 1, 2, 3+
21	# prior drug convictions	pc_drug	NIJ	Censored Numeric	0, 1, 2+
22	Any prior technical violation conviction	pc_techv	NIJ	Binary	0 = false, 1 = true

No.	Description	Variable Name	Source	Format	Values
23	Any prior domestic violence conviction	pc_dv	NIJ	Binary	0 = false, 1 = true
24	Any prior gun charge conviction	pc_gun	NIJ	Binary	0 = false, 1 = true
25	Any prior parole revocation	prev_par	NIJ	Binary	0 = false, 1 = true
26	Any prior probation revocation	prev_prob	NIJ	Binary	0 = false, 1 = true
27	Mental health/substance abuse treatment parole conditions	cond_mh_sa	NIJ	Binary	0 = false, 1 = true
28	Cognitive skills/education programming conditions	cond_cog_ed	NIJ	Binary	0 = false, 1 = true
29	Other parole conditions	cond_other	NIJ	Binary	0 = false, 1 = true
30	% of persons below poverty line	ep_pov	CDC/SVI	Numeric	10.35 to 29.52
31	% of civilians (age 16+) unemployed	ep_unemp	CDC/SVI	Numeric	7.82 to 15.40
32	Per capita income	ep_pci	CDC/SVI	Numeric	17,848.47 to 42,292.55
33	% of persons with no high school diploma (age 25+)	ep_nohsdp	CDC/SVI	Numeric	8.29 to 24.77
34	% of persons aged 65 and older	ep_age65	CDC/SVI	Numeric	8.57 to 17.71
35	% of persons aged 17 and younger	ep_age17	CDC/SVI	Numeric	18.27 to 27.98
36	% of civilian noninstitutionalized population with a disability	ep_disabl	CDC/SVI	Numeric	7.65 to 19.68
37	% of single parent households with children under 18	ep_sngpnt	CDC/SVI	Numeric	7.13 to 14.95
38	% minority (all persons except White, non-Hispanic)	ep_minrty	CDC/SVI	Numeric	13.56 to 71.37
39	% of persons (age 5+) who speak English “less than well”	ep_limeng	CDC/SVI	Numeric	0.69 to 7.36
40	% of housing in structures with 10 or more units	ep_munit	CDC/SVI	Numeric	0.85 to 37.71
41	% of mobile homes	ep_mobile	CDC/SVI	Numeric	0.64 to 30.87
42	% of occupied housing units with more people than rooms	ep_crowd	CDC/SVI	Numeric	1.80 to 3.75
43	% of households with no vehicle	ep_noveh	CDC/SVI	Numeric	2.90 to 18.01
44	% of persons in institutionalized and noninstitutionalized group quarters	ep_groupq	CDC/SVI	Numeric	0.37 to 6.62
45	Lifetime cancer risk from inhalation of air toxics (per million people)	cancer	EPA/ EJSCREEN	Numeric	39.88 to 58.88
46	Air toxics respiratory hazard index	resp	EPA/ EJSCREEN	Numeric	1.19 to 3.61
47	Diesel particulate matter concentration in air ($\mu g/m^3$)	dslpm	EPA/ EJSCREEN	Numeric	0.26 to 1.80
48	Small particulate matter (2.5 microns or less) concentration in air ($\mu g/m^3$)	pm25	EPA/ EJSCREEN	Numeric	7.95 to 10.76
49	Ozone concentration in air (parts per billion)	ozone	EPA/ EJSCREEN	Numeric	35.32 to 49.33
50	Traffic proximity and volume index (vehicles by distance)	ptraf	EPA/ EJSCREEN	Numeric	46.95 to 2,419.56

No.	Description	Variable Name	Source	Format	Values
51	Lead paint index (% pre-1960 housing stock)	leadpnt	EPA/ EJSCREEN	Numeric	4.02 to 31.72
52	Proximity to risk management plan facilities (# facilities/distance)	prmp	EPA/ EJSCREEN	Numeric	0.13 to 0.86
53	Proximity to hazardous waste treatment, storage, or disposal facilities (# facilities/distance)	ptsdf	EPA/ EJSCREEN	Numeric	0.00 to 0.16
54	Proximity to National Priorities List (NPL) Superfund sites (# facilities/distance)	pnpl	EPA/ EJSCREEN	Numeric	0.00 to 0.18
55	Proximity to major direct water pollutant dischargers (# facilities/distance)	pwdis	EPA/ EJSCREEN	Numeric	0.08 to 0.49
56	Average weekday household person-miles traveled per day	est_pmiles	BTS/ LATCH	Numeric	46.11 to 68.46
57	Environmental health index (↑ better environmental quality)	haz_idx	HUD/ AFFH-T	Numeric	15.86 to 64.79
58	Jobs proximity index (↑ more employment opportunities)	jobs_idx	HUD/ AFFH-T	Numeric	21.96 to 85.70
59	Labor market engagement index (↑ more labor force participation and human capital)	lbr_idx	HUD/ AFFH-T	Numeric	22.57 to 73.64
60	Low transportation cost index (↑ lower cost of transportation)	tcost_idx	HUD/ AFFH-T	Numeric	2.30 to 75.37
61	% of tracts designated racially/ethnically-concentrated areas of poverty	recaps	HUD/ AFFH-T	Numeric	0.00 to 20.62
62	% of tract housing units that are without vehicle and beyond 1/2 mile from supermarket	lahunvhalfshare	USDA/ FARA	Numeric	2.29 to 10.34
63	# transit stops per square mile (2016-2018)	count_ntm_ stops	NaNDA/ 111109-V1	Numeric	0.00 to 26.05
64	# full service restaurants per 100,000 (2013)	count_sales_ 722511	NaNDA\ 115404-V2	Numeric	80.83 to 313.50
65	# fast food restaurants per 100,000 (2013)	count_sales_ 722513	NaNDA\ 115404-V2	Numeric	40.75 to 85.27
66	# coffee shops per 100,000 (2013)	count_sales_ 722515	NaNDA\ 115404-V2	Numeric	5.01 to 28.34
67	# bars per 100,000 (2013)	count_sales_ 722410	NaNDA\ 115404-V2	Numeric	5.01 to 63.89
68	# religious organizations per 100,000 (2013)	count_8131	NaNDA/ 115967-V2	Numeric	161.09 to 367.85
69	# civic/social organizations per 100,000 (2013)	count_8134	NaNDA/ 115967-V2	Numeric	31.47 to 168.06
70	# police stations per 100,000 (2013)	count_922120	NaNDA\ 115973-V3	Numeric	2.31 to 14.33
71	# jails/prisons per 100,000 (2013)	count_922140	NaNDA\ 115973-V3	Numeric	0.85 to 14.33

No.	Description	Variable Name	Source	Format	Values
72	# child/youth services per 100,000 (2013)	count_624110	NaNDA\ 117163-V2	Numeric	2.76 to 11.19
73	# elderly/disability services per 100,000 (2013)	count_624120	NaNDA\ 117163-V2	Numeric	3.45 to 12.97
74	# individual/family services per 100,000 (2013)	count_624190	NaNDA\ 117163-V2	Numeric	36.84 to 125.80
75	# community food services per 100,000 (2013)	count_624210	NaNDA\ 117163-V2	Numeric	0.00 to 1.07
76	# temporary shelter services per 100,000 (2013)	count_624221	NaNDA\ 117163-V2	Numeric	0.00 to 3.15
77	# emergency relief services per 100,000 (2013)	count_624230	NaNDA\ 117163-V2	Numeric	0.18 to 2.97
78	# vocational rehabilitation services per 100,000 (2013)	count_624310	NaNDA\ 117163-V2	Numeric	1.79 to 14.17
79	# child daycare services per 100,000 (2013)	count_624410	NaNDA\ 117163-V2	Numeric	27.9 to 122.22
80	% parks/greenspace (2018)	tot_park_area_ sqmiles	NaNDA\ 117921-V1	Numeric	0.04 to 22.12
81	# mental health physicians per 100,000 (2013)	count_621112	NaNDA\ 120907-V2	Numeric	2.50 to 22.13
82	# mental health practitioners per 100,000 (2013)	count_621330	NaNDA\ 120907-V2	Numeric	1.33 to 50.72
83	# outpatient care centers per 100,000 (2013)	count_6214	NaNDA\ 120907-V2	Numeric	5.33 to 24.86
84	# supermarkets/grocery stores per 100,000 (2013)	count_445110	NaNDA\ 123001-V1	Numeric	43.38 to 101.43
85	# beer/wine/liquor stores per 100,000 (2013)	count_4453	NaNDA\ 123541-V1	Numeric	4.21 to 20.88
86	# tobacco stores per 100,000 (2013)	count_443991	NaNDA\ 123541-V1	Numeric	0.67 to 5.72
87	# dollar stores per 100,000 (2013)	count_452319	NaNDA\ 123802-V1	Numeric	4.23 to 21.33
88	# community colleges per 100,000 (2013)	count_6112	NaNDA\ 127681-V1	Numeric	0.00 to 4.44
89	# four-year colleges and professional schools per 100,000 (2013)	count_6113	NaNDA\ 127681-V1	Numeric	2.54 to 41.83
90	# cosmetology/barber schools per 100,000 (2013)	count_611511	NaNDA\ 127681-V1	Numeric	0.00 to 3.23
91	# apprenticeship schools per 100,000 (2013)	count_611513	NaNDA\ 127681-V1	Numeric	0.00 to 5.34
92	# trade/technical schools per 100,000 (2013)	count_611519	NaNDA\ 127681-V1	Numeric	0.33 to 8.45
93	Urbanicity index (2010)	ruca7	NaNDA\ 130542-V1	Numeric	1.00 to 4.47
94	Disadvantage index (2008-2012)	disadvantage08 _12	NaNDA\ 119451-V2	Numeric	6.67 to 26.28
95	Affluence index (2008-2012)	affluence08_12	NaNDA\ 119451-V2	Numeric	21.32 to 51.87

No.	Description	Variable Name	Source	Format	Values
96	Ethnic/immigrant concentration (2008-2012)	ethnicimmigrant 08_12	NaNDA\ 119451-V2	Numeric	2.59 to 20.25
Notes: With the exception of the NIJ measures, the base variable names are repeated from the original data source. For the NaNDA data, the specific study number is noted, as well as the data year(s) used.					

Table 2. Preliminary Model Comparison on Validation Sample

Classification Model	Command	Male Brier Score	Female Brier Score	Average Brier Score	Male Fairness & Accuracy Index	Female Fairness & Accuracy Index
Panel A						
Lasso	lasso logit	0.1962	0.1634	0.1798	-0.2974	0.5354
Elastic Net	elastic net logit	0.1962	0.1634	0.1798	-0.2974	0.5354
Adaptive Boosting	pyadaboost	0.3096	0.2292	0.2694	0.1450	-0.0154
Random Forest	pyforest	0.3165	0.2292	0.2729	0.1435	-0.0154
Gradient Boosting	pygradboost	0.3167	0.2256	0.2712	0.1435	-0.0155
Random Forest	rforest	0.3243	0.2220	0.2732	0.1419	-0.0156
Panel B						
Lasso (by Gender)	lasso logit	0.1961	0.1652	0.1807	-0.2974	0.5343
Lasso (Train on Whites)	lasso logit	0.1991	0.1635	0.1813	0.1682	-0.0167

Table 3. Final Lasso Results Predicting Rearrest within 1 Year Post-Release

Variable	Standardized Coefficient
No Gang Affiliation	-0.2112
Age Group	
18-22	0.2838
23-27	0.2561
28-32	0.1142
38-42	-0.0408
43-47	-0.0790
48 or older	-0.2047
Supervision Level	
Standard	-0.0409
High	0.0054
Education	
Less than HS diploma	-0.0260
High School Diploma	0.0168
Prison Years	
Less than 1 year	0.1058
Greater than 2 to 3 years	-0.0498
More than 3 years	-0.0541
Female	-0.1221
# dependents	
3 or more	-0.0252
# prior felony arrests	
1	-0.1611
2	-0.0783
3	-0.0388
6	0.0102
7	-0.0033
8	0.0450
9	0.0307
10+	0.1291
# prior misdemeanor arrests	
0	-0.0581
1	-0.0114
2	-0.0077
5	0.0132
6+	0.0753
# prior violent arrests	
0	-0.0036
3+	0.0191

Variable	Standardized Coefficient
# prior property arrests	
0	-0.0713
1	-0.0178
3	0.0264
4	0.0372
5+	0.1197
# prior drug arrests	
1	0.0112
2	-0.0091
3	0.0054
5+	-0.0173
# prior technical violation arrests	
0	-0.0411
4	0.0319
5+	0.0985
No prior domestic violence arrests	-0.0138
No prior gun charge arrests	0.0073
# prior felony convictions	
0	-0.0275
2	0.0025
3+	0.0033
# prior misdemeanor convictions	
0	-0.0283
4+	0.0441
No prior violent conviction	-0.0356
# prior property convictions	
0	-0.0077
3+	0.0375
# prior drug convictions	
0	0.0198
2+	-0.0083
No prior technical violation conviction	0.0092
No prior domestic violence conviction	0.0101
No prior parole revocations	-0.0836
No prior probation revocations	0.0153
No mental health/substance abuse treatment parole conditions	-0.1135
No cognitive skills/education programming conditions	-0.0529
% occupied housing units with more people than rooms	-0.0017
Ozone concentration in air (parts per billion)	-0.0583
Proximity to risk management plan facilities	0.0266

Variable	Standardized Coefficient
Proximity to hazardous waste treatment, storage, or disposal facilities	-0.0164
Proximity to National Priorities List (NPL) Superfund sites	-0.0078
Proximity to major direct water pollutant dischargers	-0.0057
% tracts designated racially/ethnically-concentrated areas of poverty	-0.0006
# jails/prisons per 100,000	-0.0655
# coffee shops per 100,000	0.0132
# emergency relief services per 100,000	-0.0284
# community colleges per 100,000	0.0083
# four-year colleges and professional schools per 100,000	0.0585

Table 4. Final Model Performance Metrics, Year 1, Small Team

Classification Model	Command	Male Brier Score	Female Brier Score	Average Brier Score	Male Fairness & Accuracy Index	Female Fairness & Accuracy Index
Lasso	lasso logit	0.1920 (3rd)	0.1555 (4th)	0.1738 (3rd)	0.8080 (2nd)	0.2871 (n/a)

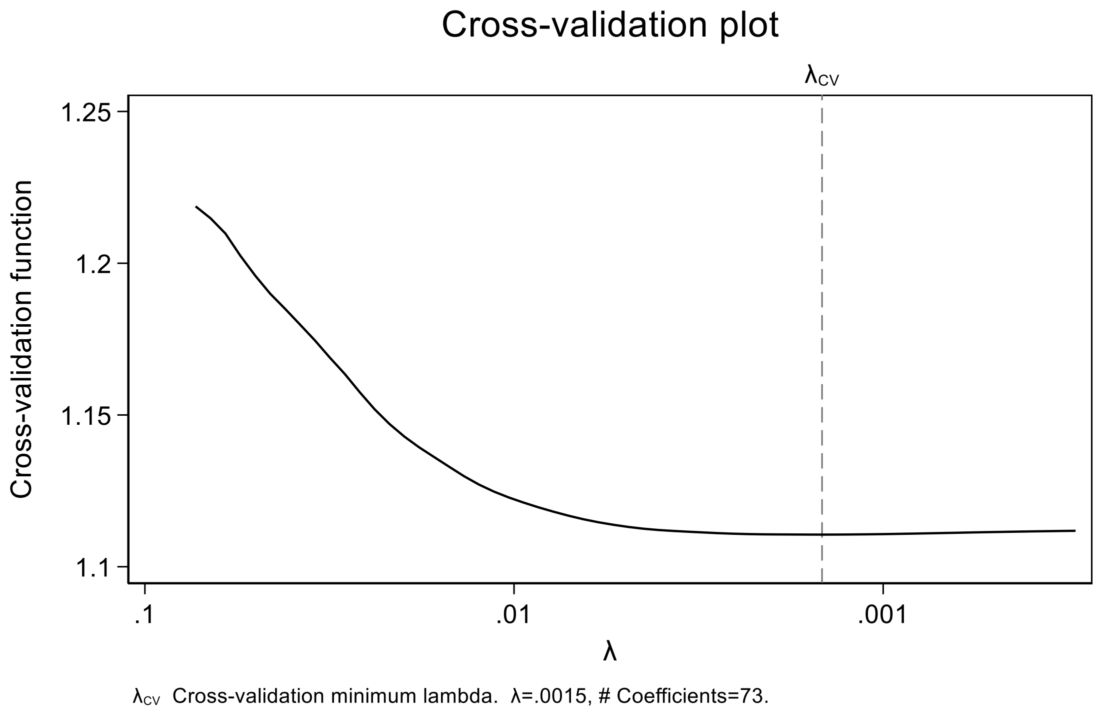

Figure 1. Cross Validation Plot