From Data to Action: Benchmarking the Accuracy-Interpretability Trade-off of Machine Learning Algorithms for Crime Analysis *

Gaspard Tissandier & Alejandro Gimenez-Santana

2026-02-25

1 Introduction

The use of statistical tools for crime modeling has a long history in criminal justice. Researchers rely on these tools to study crime trends and draw inferences about crime determinants, while practitioners use them to inform operational decisions (Clipper & Selby, 2021). Over the past two decades, this practice has taken a strong algorithmic turn. The growing availability of large scale and highly granular datasets on crime events and their potential determinants, together with advances in machine learning methods, has transformed both research and practice of crime study. Approaches now range from traditional moving average models, often implemented as heat maps and widely used by police departments, to complex machine learning systems that incorporate a large number of predictors and are marketed directly to practitioners (Groff & La Vigne, 2002; Mandalapu et al., 2023).

These developments evolve along two closely related directions. First, they improve inference on crime determinants, which supports a better understanding of crime concentration and the design of public policies. Second, they aim to increase predictive accuracy. In statistics, this distinction echoes the two cultures described by Breiman, 2001, who distinguishes between models that seek to uncover the data generating process and models that prioritize predictive performance. In practice, however, these two perspectives are deeply intertwined. Predictive applications of crime modeling rely on substantive knowledge about crime concentration, while inference oriented studies increasingly adopt tools that originated in predictive settings. Building on this observation, we examine how algorithms used in operational crime modeling can inform policy making and program design.

High performing machine learning models are often complex, large, and difficult for practitioners to interpret, as illustrated by methods such as Random Forest or boosting algorithms. Those algorithms are known to perform best on tabular data, and are widely explored by the criminal justice research community, alongside more traditional models (G. Mohler & Porter, 2018; Rummens & Hardyns, 2020; Wheeler & Steenbeek, 2021). Their strong predictive performance, combined with limited transparency, has led to a central question in the machine learning literature: is the additional complexity of such models justified by meaningful gains in accuracy? More specifically, does a given improvement in predictive performance warrant the use of models that cannot be readily interpreted by

*We thank dearly Pr. Lesia Semenova and Pr. Rayid Ghani for their valuable comments and remarks on this work. We also thank the participants of the Computational Methods in Criminal Justice Settings at the ASC2025.

practitioners, when more transparent alternatives are available?

To address this question, we benchmark a series of models on common crime prediction tasks. We predict crime concentration at the cell-month level for five types of crime in Newark, New Jersey, relying on a standard set of predictors frequently used in the literature and by predictive algorithm developers focusing on crime prediction. These include lagged crime measures, socioeconomic indicators, place based characteristics such as the presence of businesses and public institutions, and weather variables. We compare traditional approaches such as moving averages and kernel density estimation with complex machine learning models such as XGBoost, as well as highly interpretable models including FasterRisk, SIRUS, and Explainable Boosting Machines. Our selection is guided by the objective of comparing models that are standard in crime analysis, models that are widely regarded as high performing, and models explicitly designed to ensure interpretability. Following Murdoch et al., 2019, we adopt a precise definition of interpretability and classify each model according to explicit interpretability criteria. We evaluate performance using standard metrics for classification tasks and pay particular attention to their behavior in settings characterized by sparse and highly concentrated outcomes, which is typical of crime data. Finally, we analyze how each model uses the available predictors and discuss the implications for substantive interpretation.

Our results show that the trade off between accuracy and interpretability is limited for the prediction problems considered. We restrict interpretable models to ten components, such as ten coefficients, ten rules, or ten risk scores, and obtain performance levels only slightly subpar to larger models such as XGBoost or Explainable Boosting Machines. Focusing on aggravated assault and homicide, which are among the most socially harmful forms of crime, we observe that the difference in area under the receiver operating characteristic curve between the best performing algorithm and the least accurate interpretable model is approximately four percentage points.[1] Overall, accuracy ranges from 65 percent to 73 percent across models. Sensitivity varies between 85 percent and 92 percent, implying that the use of highly interpretable models increases the share of missed positive observations by at most seven percentage points relative to the most complex alternatives.

We observe larger performance gaps for traditional approaches such as moving averages or risk terrain modeling, while kernel density estimation delivers competitive results across all crime types. At the same time, high sensitivity often comes at the cost of lower precision, which we illustrate using precision recall curves. We argue that area based metrics should be preferred for model comparison and should be interpreted jointly, for example by considering both the area under the receiver operating characteristic curve and the area under the precision recall curve. Many threshold dependent metrics are sensitive to what we describe as the top k trap, whereby performance appears artificially high when predictions are evaluated only at very high probability thresholds. We document this phenomenon and show how area based measures mitigate this risk. Finally, we demonstrate that interpretable models rely on a broad set of predictors rather than focusing exclusively on past crime levels or on a single category of covariates. By making their structure transparent, these models provide actionable insights for policy design and offer stronger guarantees that predictions reflect meaningful patterns in the data rather than statistical artifacts.

[1] The area under the receiver operating characteristic curve has a probabilistic interpretation and corresponds to the probability that the model assigns a higher score to a randomly chosen positive observation than to a randomly chosen negative observation.

2 Literature review

This literature review aims to equip the reader with knowledge of the different fields this paper explores. This work is relying on the intersection of different disciplines: criminal justice, computer science and statistics. Within each discipline, we draw from different theories, strands and schools of thoughts, which are explained in the present section. We start by reviewing the criminology and criminal justice theories that crime prediction are relying on, and how this practice connect to the machine learning literature. We pursue this review by listing the different research papers exploring crime prediction and crime modeling through statistical methods. Finally, we expose the core principles of interpretable modeling, and present different papers using such kind of methods.

The task of crime prediction has a long history in the criminal justice literature, as ways to explain crime from a causal perspective have long been linked with the ability of a given theory to predict crime occurrences. Later, the evolutions in social sciences methods organized empirical research on potential causes of crime as statistical testing of such theories. However, the predictive aspect of most criminological theories is still present, as a causal explanation of crime is also able to generate prediction of such criminogenic behaviors. The field of criminology has explored many different theories, which were all empirically tested, and thus provide excellent foundations for proposing predictive models for place-based prediction. Most theories on the subject can be operationalized by including variables or proxies which are shown to be a causation for crime. This practice is encouraged in the Machine Learning literature, and known as domain-based variables selection (Kerrigan et al., 2021). We start by introducing the main criminology theories used in crime prediction and modeling.

In this research, we are exploring crime prediction in a place-time approach, thus place-based theories are naturally included in the analysis. Environmental criminology provide a solid framework for potential causation of crime at a local level (P. J. Brantingham, 1982). Routine Activity Theory, proposed by Cohen and Felson, 1979, shows how daily activities shapes crime opportunity, and should be accounted in crime mapping.Al Boni and Gerber, 2016 proposes a modern applications of this theory, by studying daily flows of urban dwellers, and infer crime concentration based on such flows and population concentration in space and time. In the same perspective, P. Brantingham and Brantingham, 1995 propose the concept of crime generators and attractors to study crime concentration. Certain elements of the built environment are considered as crime generators just by their sheer ability to concentrates individuals (transit stations, dense commercial areas, stadiums, for example). On the other hand, crime attractors are drawing offenses by the opportunity they propose (vacant lot, abandoned buildings). Kinney et al., 2008 presents an empirical study of this theory on a Canadian city, explaining the dynamic between attractors and generators. These two theories indicates that certain hub, nodes and ultimately places, will likely concentrates crime. These theory statistically materialize through the concentration of events around certain businesses, public services, as well as public equipments such as parks.

The literature on crime prediction also draw heavily from the static and dynamic structure of crime concentration across space and time. The review and future directions provided by Weisburd, 2015 shows how crime strongly concentrates geographically, a characteristic useful for crime mapping. This fact is known since the beginning of heatmap use by law enforcement agencies, but the extent of this phenomenon structure and dynamic was documented more recently. The law of crime concentration is leveraged by policy-maker and algorithms designers by including variables on the average level of crime in a given location, with heatmaps and density map, but also through more refined means such as

inclusion of multi-level places indicators (using administrative or geographical nesting of places).

Considering the dynamic concentration of crime across space and time, Townsley, 2003 proposes the idea that crimes agglomerates due to the near-repeat hypothesis: a first crime in a given place can increase crime likelihood in the same place and its vicinity, in the time following the first incident. This hypothesis has been heavily explored and modeled in the crime prediction literature (G. O. Mohler et al., 2011; Reinhart & Greenhouse, 2018; Rummens & Hardyns, 2020).

Crime prediction also draw heavily from sociological explanations of crime occurrences, bu socio-economic indicators to crime outcome. Such theories include work on social disorganization, in which social disorganization is proposed as a root causes of crime (Bellair, 2017). Social disorganization indicates a lack of social ties between individuals at a local level (neighborhood, block), consequently weakening informal control, which provide guardianship within community. The initial authors proposes variables such as socio-economic status, population heterogeneity, as well as residential instability as proxies for social disorganization. Following the same reasoning, but based on more recent observations, Sampson et al., 1999 proposes the concept of social efficacy, which goes beyond the structural approach of social disorganization. The authors show that neighborhood, even when exhibiting characteristics associated with social disorganization, can be abundant in collective efficacy, as residents exhibits strong social cohesion.

Crime prediction, as well as the related theories and technologies have been discussed for more than 20 years by the criminal justice literature, as shown by Groff and La Vigne, 2002. Drawing from their work, as well as more recent review, we can classify crime prediction for place-time based applications into two methodological frameworks, using various technical settings.

The first methodological framework can be considered as a-theoretical, as proposed by Groff and La Vigne, 2002: crime occurrences are considered as part of a statistical process, which can be studied without relying on any historical, sociological or behavioral theories. The work of Weisburd, 2015 can be considered as part of this strand, as the author highlights how crime concentrates in different places, and propose explanation, but structurally, the prediction process does not involve other variables than information on past crimes such as their time, type and location. This type of application involves studying hotspots within urban setting (Doyle & Gerell, 2024), which can always be summarized as elaborated heatmaps. It also involve the development of point processes, either solely using past crime dynamics to predict future levels as in G. O. Mohler et al., 2011, or including covariables to provide better performance and stability of the prediction process as in Reinhart and Greenhouse, 2018.

The second methodological framework consists on work drawing on criminology and sociology theories to produce prediction. This strand of research includes the use of standard statistical models such as linear, or generalized linear models, but also more advanced methods such as Machine Learning models. Most of the work in this category aims at benchmarking, or comparing, different prediction methods. For example, the paper of Rummens and Hardyns, 2020 proposes to study the predictive performances of a near-repeat algorithm, a risk terrain modeling algorithm, and a Ensemble learning algorithm, aggregating prediction of a logistic regression and a single layer neural network, using variables linked to environmental criminology, law of crime concentration, and sociological theories. The results indicates that the ensemble algorithm performs almost systematically better than others, but the performance of all the models are rather modest, with the best performing algorithm only able to identify 30% of all the positive instances, and with only 27% of the positively classified instances being right. In a similar fashion, Wheeler and Steenbeek, 2021 compare the performances of a kernel density estimation (drawing on

the law of crime concentration), risk terrain modeling and a Random Forest using a set of variables ranging from presences of businesses and public services, socio-demographic characteristics and place-time indicators. In this application, the Random Forest provides the best performance on the predictive accuracy metric (a ratio of the percentage of correctly predicted crime, over the percentage of predicted areas). This metric is common in the crime prediction literature, but unfortunately the authors don't provide the full range of metrics usually used in the Machine Learning literature, which makes the evaluation of their models rather complicated. All these applications rely on splitting the time and space of the study into a grid: either geographical (cells of different size, ranging from a few meters to a few hundreds meters) or temporal (prediction at the hour, day, month or year scale). As highlighted by Khalfa and Hardyns, 2025; G. Mohler and Porter, 2018, the base unit of analysis can have drastic effect on the performance of the algorithms, and thus this scale should always be clearly precised by researchers. Furthermore, while the predictive accuracy index and its subsequent metrics (predictive efficiency index, recapture rate index, please see Murdoch et al., 2019) are widely used in the criminology community, it is almost impossible to assess properly the performance of an algorithm with only one metric. Thus, the full range of metrics used in criminology but also in machine learning applications should be displayed by the researchers.

Finally, more recent applications are concerned in providing explainable prediction for either case clearance (Campedelli, 2022), or place-time based prediction (Khalfa et al., 2025; Zhang et al., 2022). These papers use complex machine learning models to generate predictions, then propose additional methodology to provide explanations on the decision-process of the algorithms, most of the time based on the variable importance (how much a variable is used to produce the predictions). An important distinction is between explainable machine learning (xAI), which takes complicated models to interpret them in a simple manner, and interpretable machine learning (iML), which train simple models, inherently understandable by practitioners (Rudin, 2019; Rudin et al., 2021). Following Doshi-Velez and Kim, 2017, interpretable models are needed when predictions are used for high-stakes decision, which is most of the time the case regarding any criminal justice matters. Second, they indicate that those models are desirable if the prediction problem is not perfectly mastered i.e the prediction performances are not near perfect, which is also the case as shown in the previous paragraph on different predictive applications and their performances. Amarasinghe et al., 2023 also proposes that iML is useful for model debugging, building public trust, but also to decide whether and how to intervene on a policy matter. This paper scopes is to study algorithms designed for practitioners (community organizers, local stakeholders), which is a situation in which all those elements are of the uttermost importance, especially since the feedback between local citizen and local organizations is strong. However, defining what is interpretable as a model is rather complex, thus we rely on the elements mentioned by Murdoch et al., 2019. A given model is considered as interpretable when it is:

1. Sparse: the span/elements of the model are limited
2. Simulatable: it is possible to reason about the model from a human perspective
3. Modular: meaningful portions of the model can be interpreted independently
4. Domain-based features: the features used are understandable and actionable for the targeted practitioners

The work of Zhang et al., 2022 and Khalfa et al., 2025 are considered as xML applications, while for all the aforementioned reasons we focus this paper on the performances of simple model, such as kernel density estimation, penalized regression, but also iML algorithms.

3 Data

To train the various algorithms used for crime prediction, we use datasets originating from different sources. In this section, we will describe the sources of these datasets, the data preparation as well as the features generation performed before any training of the models.

The objective of the paper is to compare the performance of various predictive algorithms for crime prediction, and to bring their performance in tension with the interpretability of the models. The focus is put on the practitioners side of the prediction: we are interested in daily application of statistical or machine learning algorithms which could be developed by small governmental entities, associations and law enforcement agencies. Thus, it is necessary to settle for datasets which are easily accessible to these entities, either through partnership with private data providers, or by acquiring them through reachable sources such as open data portals and governmental sources. Most of the works on this topic focus on one type of predictive paradigm (and source of data), while many exist (Rummens & Hardyns, 2020). For example, early analyses focused on crime history in a given area to anticipate the next offenses, using the cyclical structure of crime (Doyle & Gerell, 2024). Other papers are focusing on place-based approaches, by identifying buildings, shops, and activities which are more likely to see crime emerge in their surroundings (Wheeler & Steenbeek, 2021). Other methodologies focus on the social sources of crime and use socio-economic data to generate predictions. Finally, more recent works try to gather all these methodologies and data sources, which is also our case (Mandalapu et al., 2023).

We use four main datasets in this paper, all of them being are quite common in crime prediction. Each dataset comes at different scale and frequency of measure, which we accommodate through our data preparation phase, in order to merge all four sources into a single final dataset, then split into a training and test subsets. Our unit of analysis is the cell/month, and each source datasets' variables are imputed at this level. We create a square grid of 1000 feet (300 meters) which splits the city of Newark, and adopt the month as our temporal unit, as most community organizers plan their programming by month.

The first dataset records the historical reported crime in Newark, from 2020 to 2023, to create the predicted outcome as well as features related to past crime levels. This dataset, provided by the Newark Police Department (NPD), display every crime recorded by the NPD, with the exact GPS location, date and time, as well as precision on the type of crime, and various other information. The outcome for all models is the dichotomized crime level at the cell/month level, for each type of crime explored in this research. We restrict the types of crime to robbery, motor vehicle theft, burglary, theft, and assault and homicide joined. These crime types, all defined by the Uniform Crime Reporting (UCR) system[2], are retained due to their relative abundance, and their fairly stable reporting level. Crime are counted by type, month and cell to obtain the crime level for each observations.

The socio-economic variables are provided by the census bureau through the American Community Survey 5-Year Data[3] (ACS). Using the Census API, we pulled variables on population, citizenship composition, employment, family structure, housing and public assistance at the block-group or census tract and year level.

The built environment is described through a places dataset of most businesses, public places and public equipment in Newark. This dataset is supplied by a private data provider, and updated every year. For each place, its GPS location, name and information on the type of place is provided. The type of place is mostly provided by the NAICS code (North

[2]For the types of offenses defined in the UCR system, see https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/offense-definitions

[3]U.S. Census Bureau, "American Community Survey 5-Year Estimates" 2020-2023, http://api.census.gov/data/acs/acs5, accessed on January 1, 2025

$N_{training} = 31565, N_{test} = 7891$	$\#[Y = 1 \mid X_{training}]$	$\#[Y = 1 \mid X_{test}]$
Robbery	1340 (4.2%)	314 (4.0%)
Vehicle theft	7349 (23.3%)	1812 (23.0%)
Burglary	1563 (5.0%)	382 (4.8%)
Theft	2324 (7.4%)	637 (8.1%)
Aggravated assault/Homicide	2929 (9.3%)	716 (9.1%)

Table 1: Distribution of positive instance in training and test set

American Industry Classification System), to which we add a few categories for places which don't fit in the NAICS classification, such as abandoned lot. We count each type of places by cell and year, and duplicate the yearly count for each month of the year. As for the ACS variables, the built environment variables don't exhibit any variation within each year, but change through the years.

We include data on the weather provided by the National Centers for Environmental Information. The NCEI made available a dataset of various weather indicators such as the temperature, precipitation, wind and humidity, recorded by the weather station located at the Newark Liberty Airport with a precision at the hour level.

The final dataset, used for training and testing of the models, consists of observations at the cell/month level, with the crime level, socio-economic indicators, built environment and weather. The dataset spans from 2020 to 2023, with 53 week per year, and 555 cells in Newark, which results in an initial dataset of 117660 observations, divided into training and test set following a 80/20% split.

3.1 Feature engineering and data cleaning

This work is focusing on using simple and interpretable Machine Learning models for crime prediction. Generating interpretable algorithms requires to use model in which the modeling, as well as the features and data structure, are understandable by human agents. This section focuses on the features engineering part of our empirical application, in which we create variables that are used for crime prediction, with a focus on being understandable for the users of the system. While some complex models such as Neural Network, due to their ability to approximate any type of functions, are able to create features based on the provided variables to perform prediction, most interpretable Machine Learning model are not able to perform such task. Thus, the features engineering step in which insightful variables are created is crucial for the performance of the models. Furthermore, domain-based knowledge is highly appreciated for features engineering: field expert have a deep understanding of the studied dynamics, and are able to built features which are highly relevant for prediction tasks. They can also proceed to features selection, to save computation time by discarding features which are known to be not informative for the task at stakes or hardly interpretable by practitioners.

This section focus on the features creation. Details on data cleaning and potential imputation are also provided.

3.1.1 Historical crime data

We use the crime dataset to create the predicted variables. The count of crime by cell and month are Poisson distributed, but with very few units exceeding count higher than 3, which lead us to dichotomize the count of crime. This procedure also allow us to work with classification metrics, which are extremely useful to make sense of the performance of a model, a priority of this paper. The distribution of crime count can be observed on

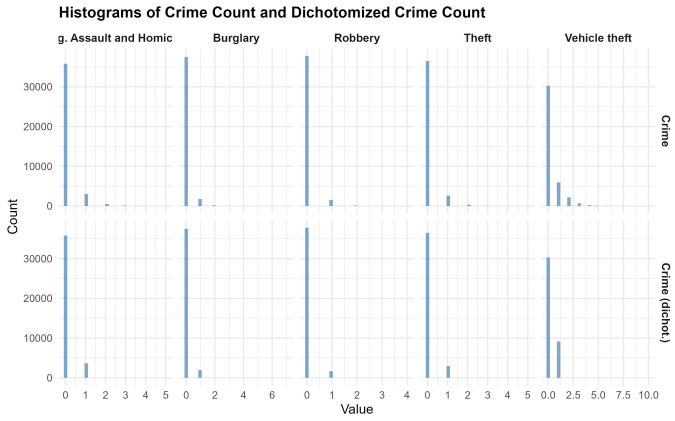


Figure 1: Distribution of crime count across crime type

The models trained for this paper rely on a vast array of data: historical reported crime, built environment with information on shop, businesses, public places and city geographical structure, socio-economic data with variables for housing, employment, family structure, as well as administrative boundaries of neighborhood and wards. However, as shown by the time-series forecasting literature, past realization of crime level in a given place is highly predictive of future crime levels. The intuition as well as empirical exploration of this question show that crime tends to form cycle, and to have a structured evolution. Thus, using regular time interval measure of crime allow the researcher to capture and explore those trend, dynamics and cycle for forecasting.

We take inspiration from this literature for crime forecasting using time-series to generate lagged observations of crime in a given cell. Crime levels are lagged over 6 months, which allow us to capture potential increasing or decreasing dynamics measured over a few months, but keep the temporal horizon coherent for the practitioner and should avoid the introduction of any noise in the form of random correlation between the features and the outcome. While it is understandable and actionable to know that observing a spike of crime in the last month can generate surge in the next month, it will be more complicated to know that an increase in crime 24 months ago can have a negative effect on the current crime level, and is very likely to be due to statistical artifacts rather than actual patterns. Furthermore, keeping the number of variable low helps to output simpler models and keep computation time reasonable. We also include the average of crime levels in each cell over the last 2,3 and 6 months.

Intuitively, if past realization of crime in a given block group are useful for actual level prediction, neighboring realization can also be highly informative. Different dynamics can be observed: all cells in a given neighborhood can exhibits a global increase in crime at the same time, or show contra-cyclical dynamics, where crime activity is moving from one area to the other. These fluctuations can be modeled using past crime level of neighboring units, for a given geographical areas, by choosing the adapted representation function. Certain Machine Learning models such as Neural Network can theoretically approximate any types of function by automating the features generation process, in our case using the 555 cells over up to 40 months. However, the weights of these model used for feature engineering are far from interpretable, which lead us to chose an ex-ante Kernel smoothing summarizing method to represent neighboring realization of past crime levels.

We introduce two types of Kernel: a Gaussian kernel and two square kernels smoothing functions, based on the geographical and time distance to a given cell. The gaussian kernel

generate weights for any observations which are inversely proportional to the distance and time to the observations: closer measures of crime will be given higher weights, following a Gaussian distribution, bi-directional for geographical dimensions, uni-directional for the time dimension. The square Kernel is weighting positively only the direct neighbors of a given block group. The square kernel is expressed as:

$$K_d^{\mathrm{S}}(d_{ij}) = h_d \, \mathbf{1}\{A_{ij} = 1\}. \tag{1}$$

With A_{ij} an indicator of unit i and j being neighbors (sharing a border). The hyper-parameter h_d is chosen as $1/\#[A_{ij} = 1]$, 1 over the number of neighbors. The Gaussian kernel is expressed as:

$$K_{p=d,t}^{\mathrm{G}}(p_{i,j}) = h_p \exp(\frac{-p^2}{l_p^2}) \tag{2}$$

Where p is the distance measure from the observations i to j, h_p will set the integral of the weight function, and l_p can be understood as the "spread" of the kernel: a higher l_p will gives further observations more weight, while smaller l_p will allow the kernel function to focus only on the closest observations. The hyperparameters h_p and l_p are chosen using the Silverman's rule, which divide a measure of the distance variability (the minimum between the sample distance standard deviation or the interquartiles ratio) by $n^{1/5}$. Thus, for a given observations, the kernel propose a prediction such that:

$$\hat{y_i} = \frac{\sum_{j \neq i} K_d K_t y_j}{\sum_i K_d K_t} \tag{3}$$

Usually, the past realization of i are included in the kernel smoothing function, but not in our case. Since past observations of crime in a given cell are already included as lagged variables and as average, they are not included again. This setup is made for interpretability purpose: the lagged variables can be indicative of a cell dynamic, but the kernel smoothing sum will capture the dynamic of a cell with its direct environment, and separating both will be interesting for interpretation purposes.

Regarding the square weighting function, we introduce two version of this Kernel: the weighted sum of all the neighboring cells over 1 and 3 months. This Kernel should allow us to capture any geographically localized dynamic of crime, much more than the Gaussian Kernel which is weighting observations with a far greater reach in distance and time. The standard deviations for the Gaussian multivariate distribution are 766 feet for the distance dimension, and 2.2 months for the temporal dimension. Interestingly, the average block size in Newark is around 500 feet, thus the Silverman's rule weight are mostly concentrated on observations within a block and 2 months of the considered observation.

3.1.2 Built environment

The features indicating the presence of certain types of place in a given cell/year are integers presenting the number of such places. Since these indicators are most of the time comprised between 1 and 3, we dichotomize all the place features for them to reflect only the presence or absence of such places. This step also aims at simplifying the interpretation: it is easier to explain that the presence of a certain business or service is positively or negatively correlated with the crime outcome than to adopt a marginal explanation using the number of places. The frequency of each type of place is presented in figure 3

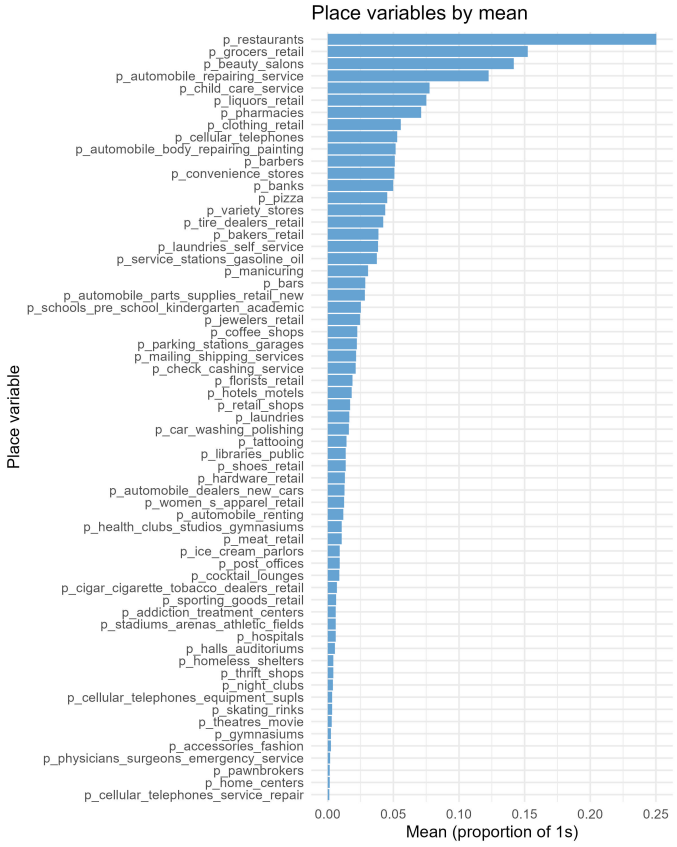


Figure 2: Distribution of place features across cells

3.1.3 Socio-economic data

The socioeconomic indicators used for prediction are drawn from the American Community Survey 5-year estimates, which allow us to have an accurate picture of the situation in each block-group regarding family and citizenship structure, tenant/owner distribution, economic situation as well as rent level and recourse to social services. Since our base geographical unit is the 1000 feet cell, we proceed to an areal weighted interpolation (Prener et al., 2019). This process, by assuming an homogeneous distribution of the variable's density across block group and census tract, distribute the value of the higher geographical units (block group) to the lower units (1000ft cell) by following the share of surface occupied by the cell within the block group (or the census tract for certain variable). While the homogeneous distribution of each variable is a strong assumption, we can assume that the census unit are not designed randomly, and aim at representing fairly homogeneous units of population. Crime data and the final dataset are concatenated at the monthly level, thus we duplicate each ACS yearly estimates for each month of the dataset's span. Thus, within one year for a given cell, the socio-economic variables will have no variance. Most variables included and described in 2 are available at the block-group/year level. However, for certain indicators, providing values at the block-group level would raises privacy concerns. In this situation, the indicators are provided at the census-tract level, and we impute the census-track value down to the cell level, following the same method of areal weighted interpolation. The full table with the ACS code and description can be found in appendix, in 10.

Certain observations have missing values, which are replaced using a median imputation. This solution is chosen to avoid discarding to much observations, and because for certain indicators with missing values (median earning and median rent, the distribu-

Name	Code (project)
Total population	a_median_age
Median age (total)	a_median_age_male
Median age (male)	a_median_age_female
Median age (female)	a_population
In labor force, employed	a_born_us
In labor force, unemployed	a_naturalization
In labor force, total	a_not_us_citizen
Median earnings (households)	a_native_pop
Family households, total	a_foreign_born
Family households, married couple	a_family
Households, male householder, no wife	a_married_couple
Households, female householder, no husband	a_male_householder
Households, single householder (male and female combined)	a_female_householder
Households, non-single householder	a_single_householder
Housing units by tenure, owner vs renter	a_nonsingle_householder
Foreign-born, total	a_median_earnings
Foreign-born, naturalized citizen	a_food_stamp
Foreign-born, non-citizen	a_employed
Foreign-born, total (alt)	a_unemployed
Native-born, total	a_notin_labor_force
Median gross rent (housing)	a_total_tenure
Households receiving food stamp/SNAP	a_median_rent

Table 2: American Community Survey variables

tion is strongly right-skewed. In this situation, taking a mean imputation would pull the missing observations assigned values toward high and non representative values. Finally, we standardize each variable down to a z-score. This standardization is preferred to 0-1 standardization, mostly due to the presence of a few very high values.

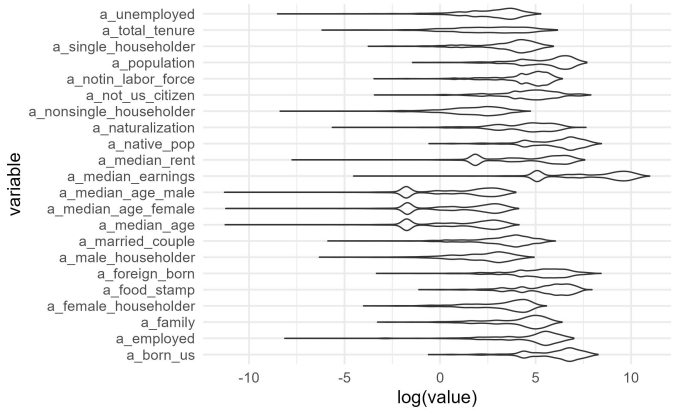


Figure 3: Distribution of ACS features across cells (log)

3.1.4 Weather data

We average the daily measures for the precipitation level, wind speed, minimum-maximum-average temperature at the monthly level, and duplicate the city observation for each cell. Thus, the weather data are showing variation between months, but not between cell. We also perform a z-score standardization, for the same reason as for the ACS features.

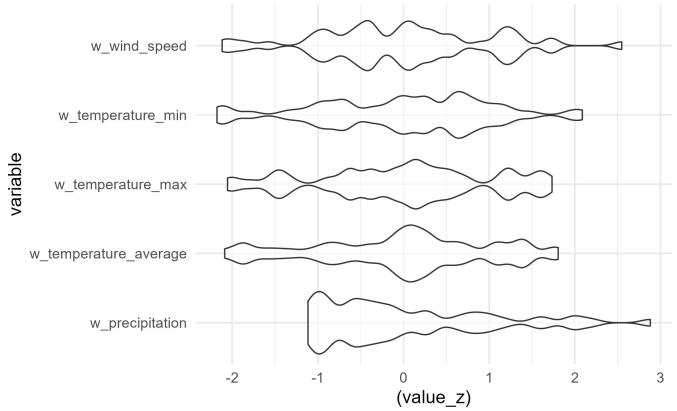


Figure 4: Distribution of weather features across cells

4 Methodology

As highlighted in the introduction and literature review, most crime prediction applications tends to be very heterogeneous in their way to organize the benchmark of models, while it is however essential to be as precise as possible regarding the methodological choices done regarding the variables used, the data-processing and features generation steps, but also which model are used, how are they trained, fine-tuned and ultimately selected. In this section, we present our theoretical reasoning for selecting the best performing models, describe the selected models, and finally explain how we achieve interpretability of those models.

4.1 Objective and adapted metrics

The main focus of this paper is to evaluate models which are able to produce interpretable output for practitioners, which differs substantially from past works and requires an adapted evaluation framework. Most Machine Learning models for public policy applications focus on providing the best prediction possible, which are then used for various goal which can, most of the time, be rooted down to constrained resources allocation, public workers advising or process optimization. However, as seen in the introduction and in the literature review, providing interpretable models is essential for criminal justice questions, as the decision taken using those algorithms are considered high-stakes, transparency is essential for just processes, and we are still unsure about the quality of the predictions (Doshi-Velez & Kim, 2017). For crime analysis through crime prediction models, and as most algorithms don't show accuracy good enough to explore the "algorithmic modeling" culture solely, the objective of the models is then not only to perform as good as possible, but to offer as much insights as possible.

Allocation objectives gather various policy applications such as dispatching social resources and community organizers in a city's neighborhoods, discover which segment of the population is more likely to experience an adverse event and thus should be targeted by a given policy, for example. Most of those allocations are done with a resources efficiency maximization objective: since resources are constrained, the policymakers want these resources to be as useful as possible. For example, if a given neighborhood is experiencing disproportionately high level of domestic violence, concentrating non-profit organizations effort in this neighborhood can ensure the most efficient allocation of their resource. In this situation, it appears that the policymakers is concerned with identifying units with the higher risk of ADVERSE effect, which lead them to use top-k metrics: being able to

allocate k units of resources, the decision-maker wants to allocate these units to the observations with the highest risk propensity in the whole distribution. If this objective was adopted for this benchmark, top-k metrics would be adopted: Recall for example, on the observations with the highest predicted probability of crime occurrence. Subsequently, this metric can be used to select models, and to fine-tune algorithms, ensuring that the parameters used for training are maximizing the chosen metric. While using top-k metrics can provide good predictive performance on a subset of the dataset, it will only be indicative of the accuracy of the algorithm on this subset, and might perform poorly on the rest. Furthermore, this choice encounter a recurrent problem in crime prediction. As crime tends to concentrate in time-place clusters (G. O. Mohler et al., 2011; Weisburd, 2015), for example with increase of burglaries around winter holidays, increase in assault around stadium for sport events, some positive observations will have high predicted probabilities only using a rudimentary subset of place and time indicators. However, those easy-to-predict and high predicted probability observations are not informative from a policy perspective: while it is easy to predict that somebody will get robbed in Time Square tonight, it doesn't inform the policymaker in a meaningful manner. This reasoning leads us to defend the use of alternative metrics than usual top-k metrics for ranking the performance of the selected models.

The objective described in the last paragraph target resource allocation, while the objective of our models is to inform policy-making. Thus, we motivate the choice of alternative metrics than those commonly used in crime prediction. We are looking for a metric which reflect how well a model is able to distinguish between positive and negative observations for all the observations in the dataset. A metric covering the entirety of the dataset has to provide a measure for every threshold τ chosen, such that $\hat{y_i} = \mathbf{1}[Pr(y_i) > \tau]$.

More specifically, each positive observation classified as so can potentially offer meaningful information in the relationship it exhibit between predictors and the outcome, which encourage us to pick a metric that maximize the True Positive Rate, as having a high TPR indicates that our model is able to correctly identify positive observations, and eventually extract interesting pattern out of it.

$$TPR = \frac{tp}{tp + fn} \tag{4}$$

Conversely, a model which minimize its False Positive Rate is also minimizing the number of negative observations classified as positive. Such model will then introduce a minimum number of uninformative observations in the positively classified pool, thus minimizing potential noise.

$$FPR = \frac{fp}{fp + tn} = 1 - TNR \tag{5}$$

Those requirements lead us to pick the Area Under the ROC Curve (ROC AUC) as the best metrics for interpretable models, as this metric indicates, for every threshold τ, the probability that a randomly drawn observation $y_i = 0$ will have a lower probability than a randomly drawn observation $y_j = 1$, such that $AUC = Pr(\hat{y_j} > \hat{y_i})$. While this metric is useful for ranking the trained models from an interpretable perspective, it is itself tenuous to make sense of from a practitioners perspective. Thus, we also include an analysis of a range of classification metrics, based on the confusion matrix. Working with a confusion matrix implies the dichotomization of the predicted probabilities into positive and negative predictions around a threshold τ. In our case, we pick two threshold rules:

- Youden: the threshold maximizing the Youden's index (defined as $J = TPR + TNR - 1$) is picked. This is equivalent to picking the threshold proposing the maximum vertical distance between the ROC curve and the diagonal random guess line

- Top 10%: the threshold classifying exactly 10% of the observations as positive is picked. This is equivalent to what is chosen usually in a resource allocation setup

Those two methods to define a probability threshold are common and well accepted in the literature: the top 10% methods ensure that we are selecting only what a constrained budget for programming can allow, while the Youden method is more flexible, and balance the performance between true positive rate and true negative rate. The confusion matrix is defined as:

Observed	Predicted	
	0	1
0	True Negative (tn)	False Positive (fp)
1	False Negative (fn)	True Positive (tp)

Table 3: Confusion matrix for binary classification ($y \in \{0, 1\}$).

We can then define all the metrics that we will implement using the two threshold described previously:

- Accuracy: $\frac{tp+tn}{n}$, percentage of observations correctly classified. Highly sensible to classes imbalance
- Sensitivity (or true positive rate, recall, hit rate): $\frac{tp}{tp+fn}$, out of the positive observations, how many were correctly classified
- Specificity (or true negative rate): $\frac{tn}{tn+fp}$, out of the negative observations, how many were correctly classified
- Precision (or positive predictive value): $\frac{tp}{tp+fp}$, out of the positively predicted observations, how many were correctly classified

Finally, the also implement the Predictive Accuracy Index, used extensively in crime prediction setting, and relying on the actual crime count per cells, and not only the confusion matrix. The PAI is using a hotspot approach, and compared the number of crime predicted over the total number of crime, to the number of observation positively predicted, over the total number of observation.

$$PAI = \frac{n/N}{c/C} \in [0, \infty) \tag{6}$$

With n the number of crime predicted, N the total number of crime, c the number of observation positively predicted, C the total number of observation (here cell per month).

We explained at the beginning of this section, that certain metrics are sensible to the "top-k" question, which we explain more in depth in this paragraph. As presented before, the AUC for the ROC of Precision-Recall curves are not sensible to the threshold chosen by the researcher, as they provide a metric for the entirety of the possible τ values, and not only a fixed threshold. However, we saw that in many applications, and especially for constrained resource allocation, the policymaker has a certain number of units to allocate, and thus will only select the top k units as positive, by choosing the adapted threshold τ such that $\#[Pr(y_i = 1) > \tau_k] = k$. When we pick such threshold, we are interested in a restrained number of metrics, written metric@k, such as the Precision or Recall, but also the PAI. We can define those metric as:

$$Precision@k = \frac{\#[y_i = 1 | Pr(y_i = 1) > \tau_k]}{k} \tag{7}$$

$$Recall@k = \frac{\#[y_i = 1 | Pr(y_i = 1) > \tau_k]}{[\#y_i = 1]} \tag{8}$$

The precision@k indicates, within the top k elements, how many are actually positive, while the recall@k indicate, among those top k elements, how many are within the positive elements. Finally, the PAI does not need any adaptation, as the denominator of the formula already include the k elements selected as a ratio of the total number of elements selectable.

4.2 Models

Our benchmark is organized to be as close as possible from actual applications and cases encountered by practitioners when training models for crime analytic, but also to implement more complex models which can provide interesting performance and/or interpretable outcomes. We train three main types of models:

- Standard crime analytics: those models are well known and regularly used by practitioners, they represent good reference point in terms of accuracy and interpretability. In this category, we include Moving Average (MA), Kernel Density Estimation (KDE), and Risk Terrain Modeling (RTM) through penalized logistic regression
- Standard Machine Learning: those models are well integrated in the Machine Learning literature, and widely used for many type of applications. In this category are included L1 Penalized Logistic Regression (PLR), and an eXtreme Gradient Boosting (XGB)
- Interpretable Machine Learning: those models are more recent, and designed to be inherently interpretable. We include Explainable Boosting Machine (EBM), FasterRisk and Stable and Interpretable RUle Set (SIRUS)

All the models mentioned above, beside belonging to different trends and culture of statistical modeling, also rely on various mechanisms and predictive paradigms. We describe briefly each algorithm, how it is trained, and the potential fine-tuning operated. The theoretical framework on which each model relies is also described, as well as the type of data used for its training.

Moving Average: moving average have a long history in crime analytics. A MA is averaging the crime level observed in a given unit across past time units (in our case 6 months), and use this average to formulate prediction on future units. This type of methodology can be assimilated to simple heatmap using cells as their base unit, and thus is quite usual for practitioners. The heuristic of this simple model, similar to what is described by Weisburd, 2015 with the law of concentration, but also by G. O. Mohler et al., 2011 with space-time clustering of crime, relies on the idea that crime concentrates in places and time, and thus past observations of the same geographical unit are helpful to predict future levels. We implement the Moving Average as a variable in a logistic regression, including fixed effects for each census tracts.

Kernel Density Estimation: KDE methods are common in crime analytics, especially grounded in the work on crime concentration. Very linked to hotspot analysis, KDE is a weighted average of the level of observed crime in neighboring units (here cells and month) to provide an expected level for the considered unit. The weighting is done using a Gaussian kernel (two-sided for the longitude and latitude, one sided for the temporal dimension). As

a bandwidth (or standard deviation), which is the value for the kernel standard deviation, we use 2000 feet and 6 months. The KDE is included in a logistic regression alongside the MA and fixed effects for each census tracts.

Risk Terrain Modeling: RTM recently grew to be one of the most used crime analytics and prediction algorithm by public institutions and law enforcement agencies. More than its algorithm, the main concept of RTM is to focus on places around which crime concentrates, to both provide insights and generate crime predictions. This method relies on penalized regression to select the most predictive places for the predicted crime. The initial software relies on Poisson L1 penalized regression, however since the outcome used in this analysis is binarized, we adapt the RTM by training a L1 penalized logistic regression with only place features.

L1 penalized logistic regression (or Lasso): L1 penalty are highly appreciated in high dimensional setup, as all the features included in the regression might not be helpful for prediction. The L1 penalty will force certain features parameters to be equal to 0, which let us with the most predictive subset of features. For this model, we include all the features selected through domain-based knowledge, including the places features as with RTM, but also including socio-economic features, crime and weather features, and various indicators (cell number, census tract, month and year). We don't fine-tune this model, as we use penalty which are corresponding to precise number of coefficients.

eXtreme Gradient Boosting: Boosting algorithms are commonly recognized as the best performing algorithms on tabular data. XGBoost is a tree-based method, which is building small models sequentially, the next tree being trained on observations misclassified by the previous tree. Due to this robust architecture, XGBoost acquired a strong appreciation as the most accurate algorithm for complex data, and is included in this benchmark as a reference point, to compare the performance of simpler algorithms. The algorithm is fine-tuned by maximizing the AUC ROC.

Explainable Boosting Machine: This recent algorithm is also based on the boosting concept, but the boosting is used to built a Generalized Additive Model. The GAM are building a specific function for each features, proposing a better flexibility than standard regression, but also a fully interpretable setup. For each features, the practitioner can use the built functions to inform policymaking. EBM is also able to include pairwise interaction between features. For more details, see Lou et al., 2013.

FasterRisk: This method build risk score, score that are multiplied with a restricted set of features to provide easily doable prediction. RiskSLIM (Ustun & Rudin, 2019), and its latest implementation FasterRisk (used in this paper) uses a lattice cutting plane algorithm to solve the complex optimization problem of finding optimal integer score. This algorithm proposes the user to tailor the final model as needed, especially regarding the number of score to integrate. For more details, see Liu et al., 2022.

Stable and Interpretable RUle Set: Building on the flexibility and strong accuracy of Random Forest, SIRUS compute the probability of a given split in the forest's trees to appear, and aggregate the most used splits. These splits are then included in a rules set used for prediction, providing both simple and stable (not sensible to data perturbations) models. For more details, see Bénard et al., 2020.

4.3 Implementing interpretability

The notion of interpretability is highly discussed in the machine learning and public policy literature, mostly because qualifying a given model as interpretable is highly context dependent, based on the initial knowledge of the examiner, and is multi-factorial. We take here the stance of following what has been discussed in the machine learning literature, as well as in the criminology literature for most of the features selection.

Murdoch et al., 2019 proposes the following points for a given model to interpretable:

- Sparsity: a model is considered sparse if the number of elements to be taken into account to understand the prediction process is limited. The main question reside in selecting the appropriate number of elements, to which we answer by following Rudin et al., 2021 who indicate that 10 elements is more or less for a given model to be mentally conceived and exploited by an untrained person. In our situation, we argue that most practitioners using crime analytics are used to work with such models, discuss crime-related dynamics, but still choose to restrain the most interpretable model to 10 elements

- Simulatability: a model is considered simulatable if "a user is able to internally simulate and reason about its entire decision-making process". This conditions is highly linked to the first one: a model is easily simulatable if it is also sparse (a decision tree for example is not satisfying this condition if the number of nodes is too large).

- Modularity: a model is considered modular if "meaningful portion of its prediction process can be interpreted independently". Some good examples of modular model include regressions, generalized additive models, risk and rules based models. Most tree-based models can hardly be considered as modular.

- Domain-based feature engineering: the features used by the model should be understandable and actionable by the target audience

The common point of all the models trained for the benchmark is the set of features they use: using work in criminology as well as our domain knowledge, we picked only features that are either known to be explainable of crime outcomes, or actionable by practitioners through local programming. An example of such selection is the presence of lawyers office in a given cell: while this feature is a good predictor for certain crime emergence, it is complicated to mobilize this pattern in a policymaking framework, and thus is not included. Most models are trained with the entire set of predictors beside Moving Average, which only use the lagged crime variables, Kernel Density which are using neighboring units past crime occurrences, and RTM which is using only place-based features.

Certain model described in the previous subsection are inherently interpretable, due to their simple architecture and straightforward training (or even absence of training). The moving average as well as the kernel density parameters (temporal horizon for the moving average, temporal and geographic horizon for the kernel density) are directly plugged-in and don't require any training. The logistic regressions the MA and KDE values are introduced in are straightforward and don't require any fine-tuning. Ultimately, the reduced number of parameters to be considered for doing prediction (the moving average, kernel density, and corresponding census tract associated parameters) make them highly interpretable models.

The second range of model, that can be qualified as standard Machine Learning models, are able to process large array of features, and rely on different mechanics to produce prediction. In this category are included L1 penalized logistic regression (RTM with place features only and the regression using the features selected) as well as XGBoost. The L1 regression are using a penalty to reduce the number of variables to include only meaningful predictors of the outcome, thus ensuring a certain sparsity. We also select penalty level such that the regression is limited to 10 features maximum. Those regression have a simple mechanics: the variance of the outcome associated to the predictor is isolated to compute a correlation coefficient for each feature, everything else equals. Thus, they are also modular and simulatable. XGBoost however relies on a vast number of tree to produce prediction,

and thus is not constrained to be interpretable, especially because the good performance of this algorithm relies on the large number of tree built.

Finally, we restrict the number of elements to 10 for FasterRisk (10 risk scores) and SIRUS (10 rules), while EBM does not propose such function and thus cannot be deemed sparse. However, all those models are both modular, as each risk score, rule or function can be interpreted in isolation, and are also simulatable, as the mechanics to produce prediction is explicit and can be represented (risk scores/rules cards for FasterRisk and SIRUS, plotting of the functions for EBM). The characteristics of each model regarding its interpretability is summarized in table 4.

Model	**Sparsity**	**Simulatability**	**Modularity**
MA	Yes	Yes	Yes
MA + KDE	Yes	Yes	Yes
RTM	Not really	Yes	Yes
PLR	Yes	Yes	Yes
XGB	No	No	No
EBM	No	Yes	Yes
FasterRisk	Yes	Yes	Yes
SIRUS	Yes	Yes	Yes

Table 4: Intepretability of selected models

5 Results

The results section is divided in three parts: following the methodology section on models evaluation, we start by ranking the algorithm following their Area Under the Curve of the ROC curve (AUC ROC), and weight their performances by detailing their accuracy on the task at stake using the Precision-Recall curve (AUC PR). Then, we develop the accuracy as well as the potential use of those models in real prediction setting by discussing a large range of metrics proposed by the binary classification setup we adopted. Finally, we show the output of the interpretable models, discuss which features are mostly used and how, and compare how each models perform prediction.

The result section is mostly focused on the aggravated assault and homicide (merged together) predictions, as it is easier and more straightforward to present the entire set of result only for one task. The results for the other types of crime are reported in the appendix. We picked this type of crime because it is not the task with the best accuracy of all crime types, which allow to be explicit and moderate about the performances of the models. Furthermore, violent crimes are always a pressing issue, and understanding potential policy lever is essential.

5.1 Ranking models

In this section, we present the ROC curves as well as the Precision-Recall curves for all models trained to predict aggravated assaults and homicides. In the Methodology section, we proposed that the ROC curves and their corresponding Area Under Curve (AUC ROC) is a useful metric to rank models based on their ability to classify observations correctly in an interpretable framework, for its ability to reflect the performance of the models on the entire dataset, and not only on the top-k observations, which tend to be easier to classify properly due to structural crime concentration laws. To properly carry out the objective of training and interpreting the prediction models, we analyze both the output of the models (how each model is doing prediction) as well as the predicted probability (probability for a

given observation to be positive). Most of the metrics usually employed for crime prediction use a set threshold τ to classify observations such that $\hat{y_i} = \mathbf{1}[Pr(y_i = 1) > \tau]$. However, the AUC ROC allows the researcher to compute the true positive rate and false positive rate for every possible threshold τ. Thus, this metric is useful to get a sense of the overall performance of the models, which is especially useful in case of concentrated prediction. For the prediction problem at stake, as crime tend to cluster in space and time, certain observations are easy to classify correctly, but the performance of the model on those is not our main interest, as we would like to gain knowledge of the prediction process over the whole dataset, including harder to classify observations (i.e not easily identifiable with time or space indicators). This idea is based on the inclusion of those models in already existing policy setups: practitioners are already equipped with statistical analysis models, as well as more or less accurate heuristics. Thus, a model providing information already acquired will not be adopted, for the cost of adoption exceeds its benefit.

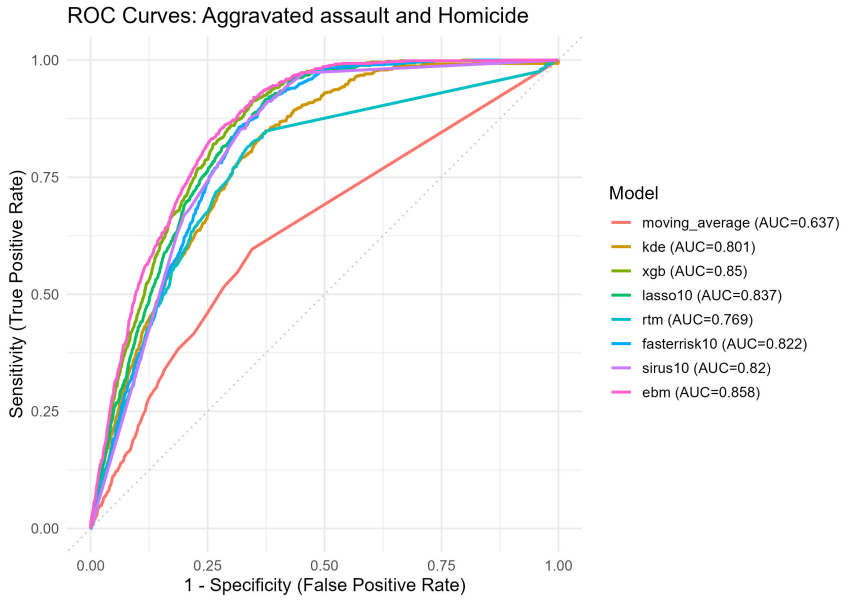


Figure 5: ROC Curves for aggravated assault and homicide

The ROC curves are presented in figure 5. The further the curve is from the 45 degrees line (random classifications), the more accurate the model is at classifying correctly positive from negative observations. First of all, we can observe that, beside the moving average (MA) and RTM, most models are exhibiting similar accuracy, without having a model clearly prevailing on the others. The linear section of the curve for the MA and RTM are due to an absence of observations with a predicted probability below a certain threshold: as we progress toward higher false positive rate, the threshold for a positive prediction is lowered (at $(1, 1)$, we have $\tau = 0$, since all observations are classified as positive, and $TPR = FPR$), thus if no observation has a predicted probability lower than this threshold, the curve is simply linearly interpolated to the $(1, 1)$ point. The AUC ROC are reported next to the models' names. Regarding AUC ROC, the least accurate model over the whole distribution is the MA, which also happens to be the simplest one, indicating that, while decent performance can be achieved with basic models, it is not possible to fully approximate the statistical process of crime occurrences with such a simple tool. However, we can observe that, by adding the kernel density estimation to the moving average (and census tracts average), we reach an AUC of 0.8, close to the best performing model (EBM, with an AUC of 0.86) or models including more potential features such as RTM or penalized regression (noted Lasso here). It is worth noticing that, while regularly recognized as the

best model for tabular data, XGBoost does not propose the best performance on this task, as it is EBM, also a boosting model, which is showing the best AUC ROC, proving again the efficiency of such type of algorithms for complex prediction tasks.

While EBM was developed to be an interpretable model, its lack of sparsity does not permit it to be included in the fully interpretable models pool. The most interpretable models, L1 penalized regression (reported as LASSO10), SIRUS (SIRUS10) and FasterRisk (here denominated FasterRISK10), restricted to 10 elements, exhibit an AUC of 0.84, 0.82 and 0.822 respectively, which is close to the performance of way more complex and dense model such as XGBoost (AUC of 0.85) and EBM (AUC of 0.86), and largely over-perform models such as MA and RTM.

This first set of result indicate that using interpretable models, either due to their simplicity or interpretable by design, is not so costly from an accuracy perspective. The ROC AUC can be interpreted as the probability for the considered model to attribute a higher predicted probability to a positive observation against a negative observation, both taken at random. Thus, between the worst performing of the interpretable by design models (SIRUS10 with 0.82) and the best model (EBM), only 4 additional percent of the observations will be misclassified by the interpretable model.

Following certain rules of thumb, an AUC of 0.8 can be considered as good performances for a model. However, this type of rule are not context-specific, and most of the time defined for balanced dataset (same proportion of positive and negative classes), and then need to be discussed properly in situ. First, the AUC is highly dependent on the class distribution: models trained on imbalanced dataset can exhibit large ROC AUC even without correctly classifying positive observations (our class of interest) in a significant proportion. However, this does not disqualify our model ranking method: all our models are trained on the same dataset, and the training and test datasets exhibit the same imbalance, thus it is valid to discriminate between models using the same data, but cross case study comparison are not straightforward with such metric. In this situation, as the negative class is way more represented (90.7% of the training data), even if lowering the threshold τ admit a certain amount of false positive, the false positive rate stay low, as the total number of negative observations is important. Thus, a good way to assess the accuracy of our models without being misled by the strong classes imbalance, and with metrics useful for policymaker, is to compute the AUC for the Precision-Recall curve, and use it as a complementary assessment.

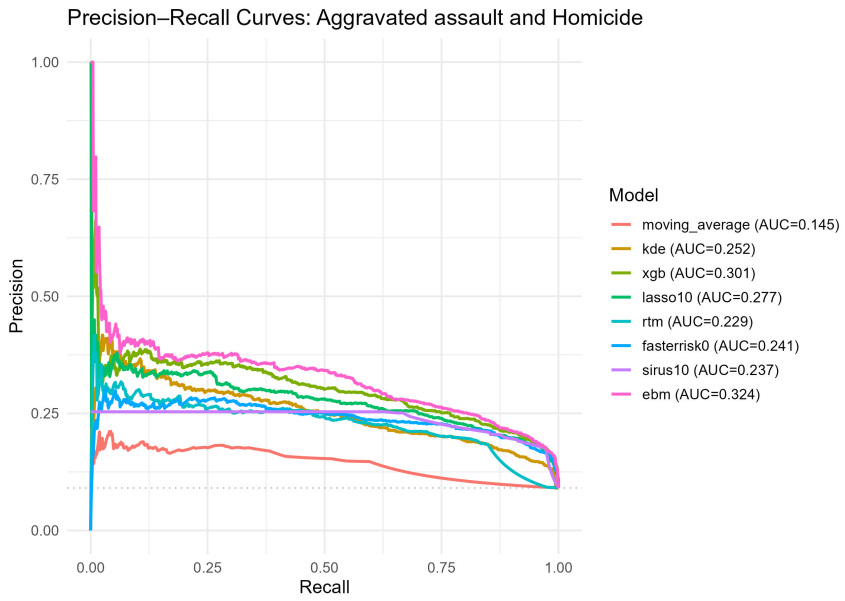


Figure 6: PR Curves for aggravated assault and homicide

The PR curve is plotting the precision (the number of true positive among positive predictions) against the recall/sensitivity (how many positive predictions among the positive observations) for every threshold τ. This curve gives a way more nuanced view of the models, as the metrics are reflecting the concern of practitioners without suffering from class imbalance: how accurate is the model on the class of interest (here positive), can one base intervention on it, how much resource might be waisted by following the model's predictions (i.e what will be the false positive rate for a corresponding recall rate). On the plot 6, we can observe that EBM is still the most accurate model, and MA the least accurate. However, even by using the most accurate model, EBM, practitioners cannot use these predictions "out of the shelf" to inform policymaking and programming. If a policymaker set the threshold τ to achieve a recall of 50% with EBM's prediction, meaning that 50% of the positive observations are discovered by the model, then the precision of the model will be of around 30%, meaning that only 30% of the positive predictions are actually true positives. In this situation, it is hardly possible to base resources allocation solely on the picked model, because either the practitioners might miss many occurrences if they chose to adopt a high threshold corresponding to a strong precision, or capture most of the positive observation with a low threshold, but will potentially "waste" many intervention and resource, represented by the low precision (i.e high false positive rate).

Rank	Agg. assault + Hom.		Burglary		Robbery		Theft		Veh. theft	
	Model	AUC	Model	AUC	Model	AUC	Model	AUC	Model	AUC
1	ebm	0.86	ebm	0.84	ebm	0.83	ebm	0.83	ebm	0.88
2	xgb	0.85	xgb	0.83	xgb	0.82	xgb	0.83	xgb	0.87
3	lasso10	0.84	lasso10	0.82	risk10	0.81	risk10	0.82	lasso10	0.86
4	risk10	0.82	risk10	0.81	lasso10	0.81	lasso10	0.82	risk10	0.86
5	sirus10	0.82	sirus10	0.80	kde	0.81	kde	0.80	kde	0.85
6	kde	0.80	kde	0.80	sirus10	0.78	sirus10	0.76	MA	0.84
7	rtm	0.77	rtm	0.75	rtm	0.76	rtm	0.76	sirus10	0.83
8	MA	0.64	MA	0.66	MA	0.65	MA	0.73	rtm	0.77

Table 5: Model ranking based on AUC ROC

Finally, we present the models ranking based on their AUC ROC for all type of crime in table 5. We can observe that EBM and XGBoost systematically outperform others algorithms, proving again the accuracy of boosting techniques. However, we can observe

that the full interpretable models roaster (LASSO10, FasterRISK10 and SIRUS10) are occupying the 3rd to 5th rank for burglary as well, with performances close to the boosting models. This is even more remarkable as XGBoost is composed of 500 trees, and EBM is including a function for all the variables possible in the dataframe, thus the interpretable models, with only 10 elements, are competing with the boosting models composed of hundreds of functions or trees. On the simpler models roaster, KDE is performing reasonability well for Burglary, Robbery and Vehicle theft, outlying the importance of concentration and space-time clustering for such types of crime.

Now that we have assessed that, from an AUC ROC perspective, most interpretable models provide decent performances and actually capture signal, and thus, exploiting the prediction process should provide insights for policy making. However, as proposed in the methodology section, we cannot only rely on one metric to assess the performance of the models, thus we detail the full range of metrics for all the models in the next section.

5.2 Understanding the performance

We stated in the previous section that boosting models are the most accurate regarding their ability to separate classes, but we also shown that, while AUC ROC is a good metric to understand if the models are capturing signal through the entire training dataset, it is also quite limited to fully understand the overall performance of the models, as the comparison the AUC PR introduced does not translate easily in a policy implementation framework. This is mostly due to two reasons:

1. The AUC ROC gives a probabilistic interpretation within an artificially balanced population (one positive against one negative observation)

2. The interpretation of the AUC ROC is probabilistic, and does not refer the the outcome's unit (count of crime, or proportion of crime count)

Thus, in this section, we extend the analysis of the ranked models with an additional range of metrics, to understand the performance of each model in a policy setting. This step is crucial, as using only one metric to assess the model is most of the time misleading, and transparency in data-analysis is essential for policy implementation. Most classification metrics relies on the predicted class instead of the predicted probabilities, thus we present results for two probability thresholds τ: Youden (maximizing the sum of Sensitivity and Specificity) and top 10% (the metrics are computed over the whole dataset, with only the top 10% classified as positive), in order to highlight how those metrics evolve based on the chosen setup. Most of crime prediction application report various metrics for a top-k setup (top 1,5 or 10%), however, most metrics can appear inflated based on the top-k reporting only. Thus, presenting the range of metrics for two thresholds ensure an appropriate understanding of the metrics' behaviors[4].

[4] $Accuracy = \frac{tp+tn}{n}$, $Sensitivity = Recall = \frac{tp}{tp+fn}$, $Specificity = \frac{tn}{tn+fp}$, $Precision = \frac{tp}{tp+fp}$, $PAI = \frac{c/C}{k/N}$

			Youden						Top 10%			
Rank	Model	AUC	τ	Acc.	Sens.	Spec.	Prec.	PAI	τ	Sens.	Prec.	PAI
1.00	EBM	0.86	0.50	0.73	0.85	0.72	0.23	2.57	0.79	0.39	0.36	3.92
2.00	XGB	0.85	0.35	0.68	0.91	0.65	0.21	2.29	0.79	0.37	0.34	3.74
3.00	LASSO10	0.84	0.39	0.65	0.92	0.63	0.20	2.17	0.69	0.33	0.30	3.27
4.00	RISK10	0.82	0.42	0.70	0.86	0.68	0.21	2.34	0.83	0.31	0.28	3.06
5.00	SIRUS10	0.82	0.16	0.68	0.88	0.66	0.20	2.25	0.20	0.31	0.28	3.07
6.00	KDE	0.80	0.09	0.65	0.85	0.63	0.18	2.04	0.24	0.32	0.29	3.17
7.00	RTM	0.77	0.34	0.68	0.81	0.67	0.20	2.16	0.75	0.28	0.26	2.84
8.00	MA	0.64	0.08	0.65	0.60	0.65	0.15	1.62	0.13	0.19	0.17	1.90

Table 6: Models' metrics based on Youden and top10% thresholds

In table 6, the models have been ranked following their AUC ROC. We precise, for the Youden criterion, the threshold in chosen, accuracy, sensitivity (or recall), specificity, precision and PAI. For the Top 10% criterion, we present the corresponding threshold, sensitivity@10%, precision@10% and PAI.

First and foremost, we can observe that the accuracy of all models is within a interval of 8 percentage points, but with a wider corresponding interval for the AUC ROC. This result indicates that, the accuracy will not be the best metric in this situation, as the class imbalance generate high accuracy. The worst model is the Moving Average, which classify 65% of the observation correctly, while the best, EBM, classify 73% of the test dataset correctly.

For the highly interpretable models (LASSO10, RISK10 and SIRUS10), the sensitivity is ranging from 86% to 92% of the positive observations detected. It indicates that those simple models were able to correctly detect almost all the positive observations. Regarding sensitivity, LASSO10 is the best performing algorithm, following by XGBoost with a sensitivity of 0.91. This result shows that, using a sparse models with 10 coefficients, LASSO10 was able to correctly identify more positive observations that XGBoost, a dense models composed of many thousands of trees. The performance of LASSO10 does not come at the cost of sacrificing on the specificity (ability of the models the correctly classify negative observations) or the precision (how many positively predicted observations are correctly classified). LASSO10 and XGBoost exhibit very similar performance on those metrics: they were able to correctly classify respectively 63% and 65% of the negative observations. However, they both suffer from low precision: only 20% of the positive predictions are actually crime occurrences. This last result indicate that, while the associated models are uncovering relationships between the predictors and the outcomes, using the predictions to inform policy making can lead to 80% of false positive, thus begetting considerable policy resource waste. KDE and RTM also propose competitive performance with a sensitivity of 0.85 and 0.81. KDE, with a very simple structure, also provide decent detection of positive and negative observations, but a slightly subpar precision of 0.18.

A common metric in the criminal justice literature concerned with crime prediction, forecasting or modeling is the Predictive Accuracy Index, which propose a convenient weighting of the sensitivity by the ratio of positively predicted units. For this indicator, we use the actual crime count per cell-month, instead of a binarized variable. The PAI is computed as the ratio of detected crime over the total number of crime, over the proportion of positively predicted units. A PAI greater than 1 indicates that a greater fraction of crime are detected than the fraction of units investigated. The best performing algorithm in term of PAI is EBM, with a PAI of 2.57, followed by RISK10 with 2.34, and XGBoost with 2.29. All the PAI are greater than 2, beside MA with a PAI of 1.62. The PAI offers a different view of the trained models, based on the actual crime distribution, and is more tailored for communication with policymakers.

The second part of the table help to highlights how the indicators are behaving when only the top 10% of the probability distribution is selected as positive. This practice is common in policymaking, as crime prediction is often considered as employed for resource allocation, which are, by design, constrained. For this second threshold, we present the sensitivity, precision and PAI, with a selection threshold higher than with the Youden statistics.

First of all, mechanically, the sensitivity is lower, as we are classifying as positive fewer observations. However, we can observe that by doing so, the sensitivity varies between roughly 30 to 40%, with only 10% of the observation classified as positive. This indicates that our models are efficient to capture signal from the positive observations, as up to 40% of the positive observations are within the top 10% predicted probabilities. Furthermore, we can observe that the Precision drastically increased, which again highlight the mechanism we explained before: by focusing only on the top predicted observations, we are actually selecting easier to predict units, guarantying a better precision. This behavior is illustrated in figure 7.

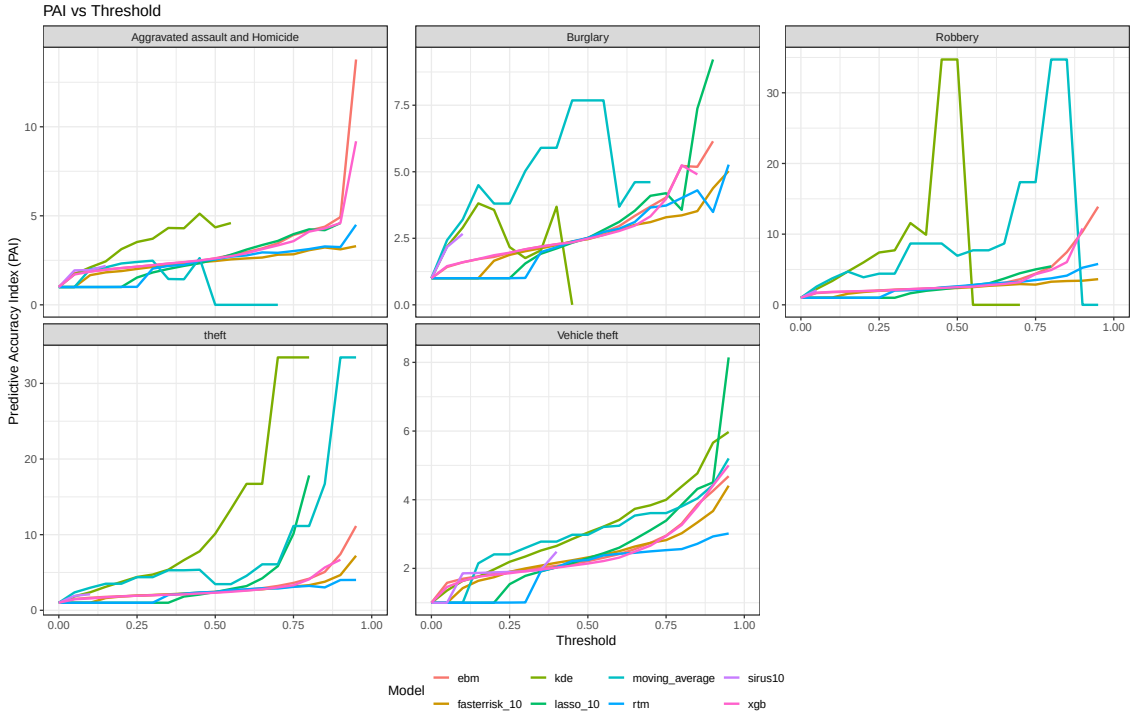


Figure 7: PAI as a function of the threshold

The first noticeable behavior of the PAI as a function of the chosen threshold is that the metric is not defined for every values of the threshold: since we cannot normalize the PAI to a given value (like the precision, standardized to 1 when the chosen threshold is greater than every predicted probabilities), above a certain threshold, this metric is not indicative of the performance of the algorithm anymore. Most importantly, we can observe that the relationship between the PAI and τ is strongly increasing: selecting fewer observations lead to increasingly higher PAI. This phenomenon is particularly visible for crime types with many occurrences such as theft, or vehicle theft. Since the PAI is defined as the ratio of observed crime over total crime, with the number of selected area over the total number of areas, as long as the second derivative of the numerator in τ is higher than those of the denominator (which is very common, following the law of crime concentration), we will observe this phenomenon. To conclude, researchers and practitioners could be tempted to report only top 1, 5 or 10% of the predicted probability distribution, thus obtaining very

high PAI, but this result is mostly misleading on the ability of the algorithm to model the data, and alternative metrics have to be used to complement the analysis of the algorithm's performances.

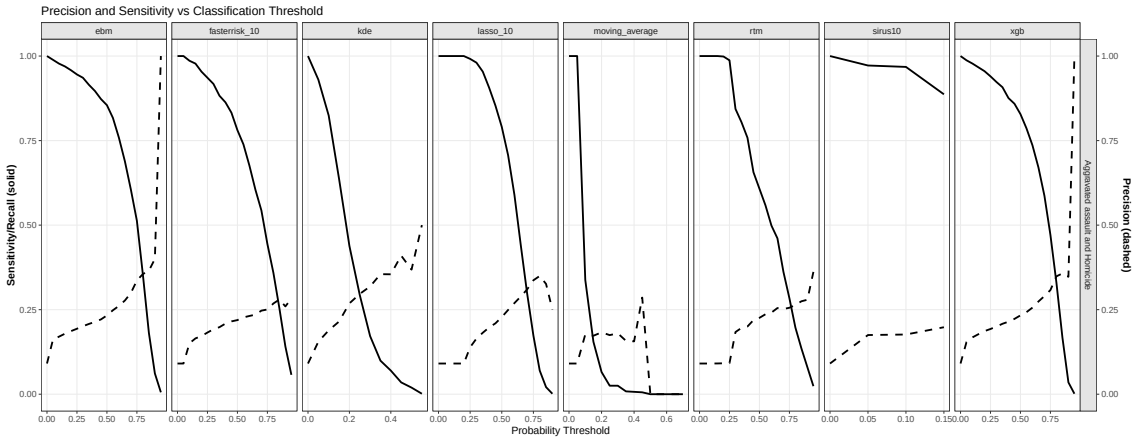


Figure 8: Precision and Sensibility as a function of the threshold

In figure 8, for algorithm outputting predicted probabilities distributed over $[0, 1]$[5], we can observe the same phenomenon as in table 7: the precision, represented by the dashed line, is an increasing function of the selection threshold. However, as the precision increase strongly, the sensitivity also decrease starkly. Thus, it is possible to select threshold which will raise very high precision, but predicting almost no positive observations, as illustrated in the PR Curves. This situation is particularly visible for EBM, KDE and XGB.

5.3 Interpretation of the models

Figure 9: Spatial distribution of aggravated assault and homicide across models.

(a) Observed (b) Moving Average (c) KDE + Moving Avg (d) RTM

(e) LASSO10 (f) FasterRisk10 (g) XGBoost (h) EBM

[5]It is important to note that most sparse models predicted probabilities are not distributed over [0,1]. This can be explained by the restricted number of predictors used, reducing the number of possible mapping of the predictors to the outcome. In this case, the probabilities will be distributed on a smaller interval, but this does not impact the prediction, as we are using relative threshold (such as Youden) and not absolute threshold (0.5 is commonly used to split de predicted probabilities between positive and negative observation).

First, we present the spatial distribution of the predictions and the observed level of aggravated assault and homicide. It is essential to precise that the temporal dimension is missing from those maps, but is essential to fully perceive how each models "captured" the data generation process. The first map is the average of crime occurrence per cell in the test dataset, while all the other maps present the average prediction per cell. We can observe that the best performing models i.e the complex or interpretable models subsets understood the spatial distribution of the aggravated assault and homicide. The LASSO seems to have the best distribution, while certain algorithms as EBM and XGBoost exhibit high predicted probabilities for a lot more cells.

In this rest of the section, we propose to analyze the different interpretable models trained by examining the models themselves, and not only their outcome. This task is useful from two main perspectives:

- Understand, by the variables selection, their ordering, and the weights, coefficients or scores attached to each, how the model generated the prediction. This step ensure accountability regarding the prediction process. Additionally, the verification of the models ensure a sound exploitation of the data, and that the predictions are not relying on statistical artifact.
- By extracting the correlations and patterns, the practitioners can derive policy lever from the model itself. The practitioners also ensure the right interpretation of such patterns, and field knowledge is thus needed

As described in the Methodology section, the models which are directly interpretable are the Moving Average, the Kernel Density Estimation, LASSO, FasterRisk and SIRUS. We provide both the main intuition for the interpretation of each models, as well as the actual interpretation of each models for the prediction of aggravated assault and homicide. We start by detailing the regression based models (MA, KDE and LASSO). While not regression model themselves, we integrated the MA and KDE outputs as predictors in a logistic regression, to ensure the best weighting of their values.

- Moving Average: the average of the lagged crime count over the last 6 months, and included as a predictor in a logistic regression
- Kernel Density Estimation: the Gaussian multivariate density for a given cells, using 6 months, and 300 feet/100 meters as the standard deviation of the multivariate Gaussian distribution. It is important to note that, the distribution is one-sided for the time dimension (only taking into account past observations), and included as a predictor in a logistic regression, alongside the moving average and a census tract fixed effect, averaging the level of crime per cells
- LASSO: the whole subset of variables (crime, places, socio-economic indicators, weather) are included in a penalized logistic regression. The penalty parameter is restricted to include maximum 10 coefficients

In table 7, all three regressions are presented. We can observe that the average lag on 6 months has a positive coefficient for both MA and KDE, indicating that the law of crime concentration is verified in our dataset: observing higher level of crime in a given cell in the past months is associated with higher probability of crime occurrence in the current month. Surprisingly, when including the KDE with the MA as well as a census tracts fixed effect, the coefficient for MA stays positive, but the coefficient the KDE is negative. It indicate that observing higher level of crime in the surrounding cells in the past months

Table 7: Logistic and LASSO Model Estimates for Aggravated Assault and Homicide

	(1) MA	(2) MA + KDE + FE	(3) LASSO
Intercept	-2.360	-7.168	-0.421
avg_lag	0.839	0.151	—
kde_score	—	-0.031	—
Born in US	—	—	0.071
Single householder	—	—	0.068
Median age (female)	—	—	0.062
Kernel crime	—	—	0.314
Female householder	—	—	0.201
Kernel crime neighbors	—	—	0.093
Median earnings	—	—	0.092
Grocers / Retail	—	—	0.058
Observations	31,565	31,565	31,565
Null Deviance	19,500	19,500	—
Residual Deviance	18,810	15,360	—
AIC	18,820	15,540	—
Census Tract FE	No	Yes	No

Notes: Models 1 and 2 are logistic regressions. Model 2 includes census tract fixed effects (coefficients omitted). Model 3 reports logistic regression with a L1 penalty estimates.

decrease the propensity for crime in the current month and cell. It is important to note that the Gaussian kernel for the KDE and the Gaussian Kernel included in other models are not similar: we used a standard deviation of 6 months for the temporal dimension, with still 500 feet for the geographical dimension, while the later kernel are set with 2 months and 500 feet. Furthermore, the later kernel don't include the current cell in the computation of the density, while the KDE does.

The interpretation of LASSO coefficient is less straightforward due to the penalty constraint. When the optimization is operated with a penalty within the objective function, the coefficient of non-zero predictors cannot be interpreted directly as everything else equal correlation, but as an ordering of importance. This result is explained in Tibshirani, 1996, as the sign of the parameters can change from the OLS estimation (which, under certain conditions, guarantee that the estimators are the best linear unbiased estimators) to the LASSO estimation. However, within the subset of predictors selected, and with respect to the predictions operated, the coefficients are straightforward in their sign: a positive coefficient increase the predicted probability, and inversely. In our case, we can observe that all coefficients are positive, thus every variables increase is associated with an increase in crime likelihood. The most important coefficient is the share of US born in the given cells, followed by the share of single householders and the median age of female. Thus, a cell with higher share of US born will see a higher propensityof crime, as well cell with a higher percentage of single householder. We can also observe that the median age of female influences positively the crime likelihood. One could argue that it is a well-observed phenomenon, in which neighborhood with higher share of female only householder have higher crime concentration. However, we can see that the share of female householder is also included in the model, and thus those are two distinct effects. Interestingly, two kernels are included as predictors (Gaussian and neighbors kernel), indicating that higher concentration of crime in the neighboring cells and previous months will increase crime likelihood. This result differs from the one observed with the KDE values as predictors.

Table 8: Risk Score Card — Assault/Homicide Model

Panel A: Point Assignment	
Variable	Points
Born in US	1
Single householder	1
Automobile repairing service	1
Kernel crime	1
Kernel crime neighbors 2	1
Census tract 34013001900	2
Cell number 42	5
Cell number 766	5

Panel B: Score to Risk Mapping					
Score	Risk	Score	Risk	Score	Risk
0	27.6%	6	87.3%	12	99.2%
1	38.2%	7	91.8%	13	99.5%
2	50.0%	8	94.7%	14	99.7%
3	61.8%	9	96.7%	15	99.8%
4	72.4%	10	97.9%	16	99.9%
5	80.9%	11	98.7%	17	99.9%

Notes: Total score equals the sum of assigned points. Risk values represent predicted probabilities of assault/homicide conditional on the score.

The FasterRisk model provides a sparse and fully operational scorecard representation of the model (Table 8). In contrast to LASSO, which selects variables through an L1 penalty but leaves coefficients on a continuous scale, FasterRisk constrains coefficients to small integers and directly optimizes predictive performance under this interpretability constraint. The selected variables partially overlap with those identified by LASSO. In particular, Born in US, Single householder, Kernel crime, and Kernel crime neighbors are retained in both models, confirming that demographic structure and spatial crime concentration are central predictors of assault and homicide risk. However, FasterRisk excludes several socioeconomic variables selected by LASSO, such as Median age (female) and Median earnings, while introducing Automobile repairing service and two highly localized identifiers (one census tract and two specific cells) with relatively large point allocations. This difference illustrates the tension between the two approaches: whereas LASSO distributes weight across multiple correlated demographic indicators, FasterRisk concentrates predictive power on a smaller subset and captures residual spatial heterogeneity through discrete geographic indicators. Importantly, all selected predictors contribute positively to the score, consistent with the LASSO signs, reinforcing the conclusion that higher values of these characteristics are associated with higher predicted crime likelihood. The score-to-risk mapping shows a steep gradient, meaning that the accumulation of a few risk factors is sufficient to move a cell into a very high predicted probability range. Compared to LASSO, FasterRisk sacrifices some granularity in coefficient magnitude but gains transparency and operational usability, while broadly confirming the substantive drivers identified by the penalized regression.

The SIRUS model departs more substantially from the LASSO specification by emphasizing threshold effects and interactions rather than additive contributions. The decision rules reported in Table 9 repeatedly select Median age, Median age (female), Female householder, Median earnings, and Kernel crime as splitting variables. Notably, demographic and socioeconomic variables dominate the upper part of the rule list, indicating that much of the predictive structure can be explained by discontinuities around specific standardized thresholds. In contrast with LASSO, where all selected coefficients are positive and interpreted as monotonic contributions, SIRUS reveals strong non-nonlinearities: for example, cells below certain median age or female age thresholds exhibit extremely low predicted risk (around 0.4–2%), while those above the thresholds jump to risk levels near 18–20%. This suggests that the relationship is not merely additive but characterized by regime changes. Moreover, while LASSO includes both kernel-based measures and multiple socioeconomic indicators simultaneously, SIRUS shows that combinations of age-related variables define particularly low-risk strata, highlighting interaction effects that are not explicit in the penalized logistic model. Interestingly, SIRUS does not rely on the Born in US variable emphasized by LASSO and FasterRisk, suggesting that once non-linear splits on age, earnings, and household composition are introduced, that variable adds limited additional discrimi-

Table 9: SIRUS Decision Rules — Assault/Homicide Model

	Rule Condition	Risk	n
	Baseline (full sample)	0.0928	31,565
|	Median age (female) < -0.0109	0.0185	18,932
	Median age (female) ≥ -0.0109	0.2040	12,633
|	Female householder < -0.119	0.0186	18,931
	Female householder ≥ -0.119	0.2040	12,634
|	Median age < -0.0209	0.0186	18,937
	Median age ≥ -0.0209	0.2040	12,628
|	Median earnings < 0.0316	0.0196	18,939
	Median earnings ≥ 0.0316	0.2030	12,626
|	Kernel crime < 0.293	0.0344	22,095
	Kernel crime ≥ 0.293	0.2290	9,470
|	Median age < -0.407	0.0039	15,778
	Median age ≥ -0.407	0.1820	15,787
|	Median age (female) < -0.426	0.0043	15,788
	Median age (female) ≥ -0.426	0.1810	15,777
|	Median age < -0.407 & Median age (female) < -0.0109	0.0039	15,778
	Median age < -0.407 & Female householder < -0.119	0.0038	15,766

Notes: Reported values correspond to terminal-node predicted risk and sample size. Continuous variables are standardized as z-scores.

natory power. Overall, SIRUS puts into perspective the LASSO findings: it confirms the importance of demographic structure and spatial crime concentration but reframes them as threshold-driven and interaction-based patterns rather than purely additive effects.

6 Conclusion

This paper examined whether the increasing complexity of machine learning models for crime prediction is justified by substantial gains in predictive performance. Using a common benchmark across several crime types in Newark, New Jersey, we compared traditional approaches, highly flexible machine learning algorithms, and explicitly interpretable models under the same data structure and evaluation framework.

Our results suggest that the trade off between accuracy and interpretability is limited for the prediction of spatial and temporal crime concentration. Interpretable models constrained to a small number of coefficients, rules, or risk scores achieve performance levels close to those of more complex algorithms. Although the most flexible models occasionally obtain the highest aggregate metrics, the magnitude of these gains remains modest. From a practical perspective, small improvements in predictive performance may not justify the loss of transparency, especially in policy environments where explanation, accountability, and communication are central.

We also highlight the importance of carefully selecting evaluation metrics in settings characterized by rare and highly concentrated outcomes. Area based measures, interpreted jointly, provide a more reliable assessment of model performance than threshold dependent metrics taken in isolation. These methodological choices directly shape how practitioners interpret model outputs and allocate limited resources.

Overall, our findings support a balanced approach to algorithmic crime modeling. When interpretable models deliver comparable predictive performance, they offer a compelling alternative to more complex systems. Advancing crime analytics therefore does not necessarily require greater model complexity, but rather a careful alignment between predictive accuracy, transparency, interpretability, and practical usefulness.

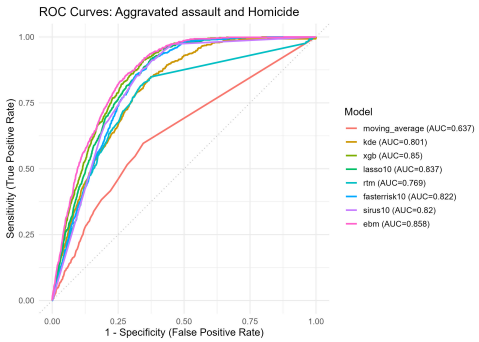
7 Appendix

A Variables description

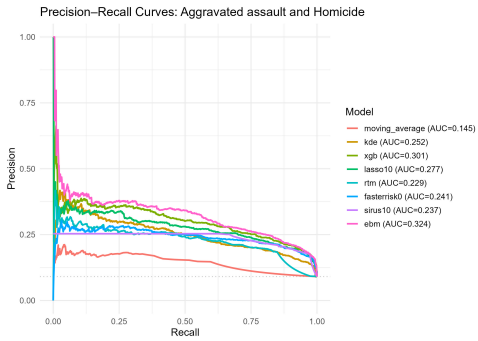
Name	Code (project)	Code (ACS)	Description
Total population	a_median_age	B01003_001E	Total population estimate
Median age (total)	a_median_age_male	B01002_001E	Median age of the population
Median age (male)	a_median_age_female	B01002_002E	Median age of male population
Median age (female)	a_population	B01002_003E	Median age of female population
In labor force, employed	a_born_us	B23025_007E	Number of employed persons aged 16+
In labor force, unemployed	a_naturalization	B23025_004E	Number of unemployed persons aged 16+
In labor force, total	a_not_us_citizen	B23025_005E	Total in labor force aged 16+
Median earnings (households)	a_native_pop	B20002_001E	Median earnings for households
Family households, total	a_foreign_born	B11001_008E	Number of total family households
Family households, married couple	a_family	B11001_009E	Number of married-couple family households
Households, male householder, no wife	a_married_couple	B11001_002E	Number of family households with male householder, no wife present
Households, female householder, no husband	a_male_householder	B11001_003E	Number of family households with female householder, no husband presen
Households, single householder (male and female combined)	a_female_householder	B11001_005E	Number of family households with single householder (no spouse)
Households, non-single householder	a_single_householder	B11001_006E	Number of family households with non-single householder (spouse present
Housing units by tenure, owner vs renter	a_nonsingle_householder	B25003A_001E	Number of housing units, tenure (owner/renter)
Foreign-born, total	a_median_earnings	B05001_002E	Number of persons born in United States
Foreign-born, naturalized citizen	a_food_stamp	B05001_005E	Number of persons who are naturalized U.S. citizens
Foreign-born, non-citizen	a_employed	B05001_006E	Number of persons who are non U.S. citizens
Foreign-born, total (alt)	a_unemployed	B05002_002E	Number of foreign-born persons
Native-born, total	a_notin_labor_force	B05002_013E	Number of native-born persons
Median gross rent (housing)	a_total_tenure	B25031_001E	Median gross rent for renter-occupied housing units
Households receiving food stamp/SNAP	a_median_rent	B22001_001E	Number of households receiving SNAP / food stamps

Table 10: American Community Survey variables and description

B Models curves

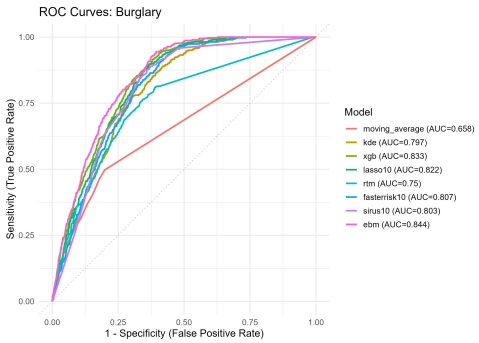


(a) ROC Curve: Vehicle theft

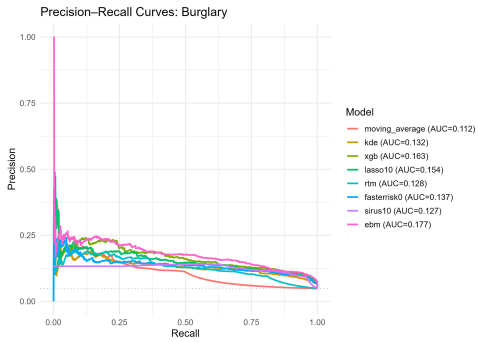


(b) PR Curve: Vehicle theft

Figure 10: ROC and PR curves for Vehicle theft

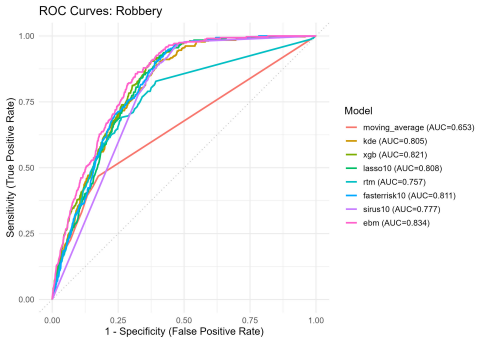


(a) ROC Curve: Vehicle theft

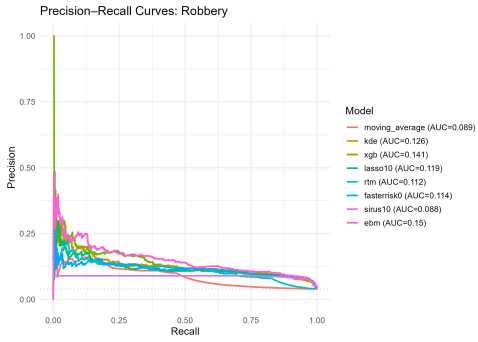


(b) PR Curve: Vehicle theft

Figure 11: ROC and PR curves for Vehicle theft

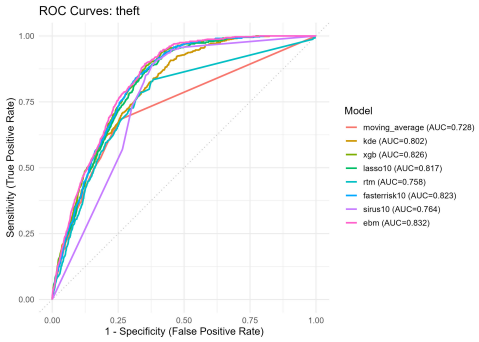


(a) ROC Curve: Vehicle theft

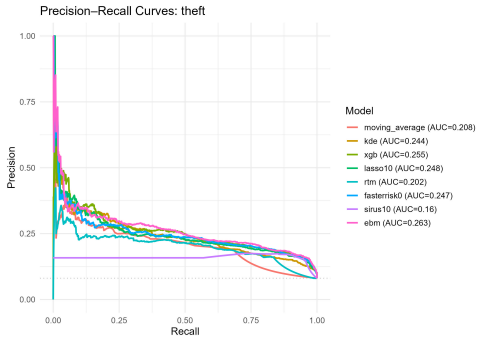


(b) PR Curve: Vehicle theft

Figure 12: ROC and PR curves for Vehicle theft

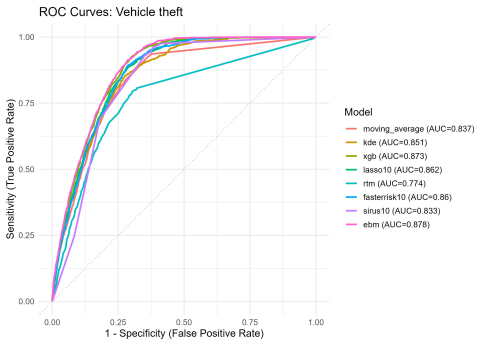


(a) ROC Curve: Vehicle theft

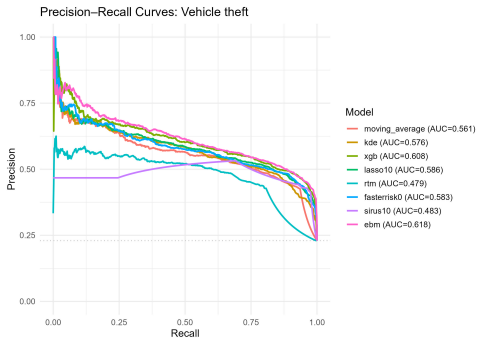


(b) PR Curve: Vehicle theft

Figure 13: ROC and PR curves for Vehicle theft



(a) ROC Curve: Vehicle theft



(b) PR Curve: Vehicle theft

Figure 14: ROC and PR curves for Vehicle theft

References

Al Boni, M., & Gerber, M. S. (2016). Predicting crime with routine activity patterns inferred from social media. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 001233–001238. https://doi.org/10.1109/SMC.2016.7844410

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, *5*, e5. https://doi.org/10.1017/dap.2023.2

Bellair, P. (2017, July 27). Social disorganization theory. In *Oxford research encyclopedia of criminology and criminal justice*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264079.013.253

Bénard, C., Biau, G., Veiga, S. d., & Scornet, E. (2020, December 16). SIRUS: Stable and interpretable RUle set for classification. https://doi.org/10.48550/arXiv.1908.06852

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

Brantingham, P. J. (Ed.). (1982). *Environmental criminology*. Sage.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. Retrieved February 23, 2026, from http://www.jstor.org/stable/2676681

Campedelli, G. M. (2022). Explainable machine learning for predicting homicide clearance in the united states. *Journal of Criminal Justice*, *79*, 101898. https://doi.org/10.1016/j.jcrimjus.2022.101898

Clipper, S., & Selby, C. (2021). Crime prediction/forecasting. In *The encyclopedia of research methods in criminology and criminal justice* (pp. 458–462). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119111931.ch94

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*(4), 588. https://doi.org/10.2307/2094589

Doshi-Velez, F., & Kim, B. (2017, March 2). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608

Doyle, M. C., & Gerell, M. (2024). Assessing crime history as a predictor: Exploring hotspots of violent and property crime in malmö, sweden. *International Criminal Justice Review*, 10575677241230915. https://doi.org/10.1177/10575677241230915

Groff, E. R., & La Vigne, N. G. (2002). Forecasting the future of predictive crime mapping. *Crime Prevention Studies*, *13*, 29–58.

Kerrigan, D., Hullman, J., & Bertini, E. (2021). A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction*, *5*(12), 73. https://doi.org/10.3390/mti5120073

Khalfa, R., & Hardyns, W. (2025). Comparing machine learning-based crime predictions across micro-geographic units: Street segments, rectangular grids, and hexagonal grids. *Applied Spatial Analysis and Policy*, *18*(3), 79. https://doi.org/10.1007/s12061-025-09683-1

Khalfa, R., Theinert, N., & Hardyns, W. (2025). Comparing XAI techniques for interpreting short-term burglary predictions at micro-places. *Computational Urban Science*, *5*(1), 27. https://doi.org/10.1007/s43762-025-00185-x

Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., & Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment*, *34*(1), 62–74. https://doi.org/10.2148/benv.34.1.62

Liu, J., Zhong, C., Li, B., Seltzer, M., & Rudin, C. (2022). Fasterrisk: Fast and accurate interpretable risk scores. https://arxiv.org/abs/2210.05846

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623–631. https://doi.org/10.1145/2487575.2487579

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access*, *11*, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

Mohler, G., & Porter, M. D. (2018). Rotational grid, PAI-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(5), 227–236. https://doi.org/10.1002/sam.11389

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Prener, Christopher, Revord, & Charles. (2019). areal: An R package for areal weighted interpolation. *Journal of Open Source Software*, *4*(37). https://doi.org/10.21105/joss.01221

Reinhart, A., & Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *67*(5), 1305–1329. https://doi.org/10.1111/rssc.12277

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021, July 10). Interpretable machine learning: Fundamental principles and 10 grand challenges. https://doi.org/10.48550/arXiv.2103.11251

Rummens, A., & Hardyns, W. (2020). Comparison of near-repeat, machine learning and risk terrain modeling for making spatiotemporal predictions of crime. *Applied Spatial Analysis and Policy*, *13*(4), 1035–1053. https://doi.org/10.1007/s12061-020-09339-2

Sampson, R. J., Morenoff, J. D., & Earls, F. (1999). Beyond social capital: Spatial dynamics of collective efficacy for children. *American Sociological Review*, *64*(5), 633. https://doi.org/10.2307/2657367

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved February 16, 2026, from http://www.jstor.org/stable/2346178

Townsley, M. (2003). Infectious burglaries. a test of the near repeat hypothesis. *British Journal of Criminology*, *43*(3), 615–633. https://doi.org/10.1093/bjc/43.3.615

Ustun, B., & Rudin, C. (2019, September 17). Learning optimized risk scores. https://doi.org/10.48550/arXiv.1610.00168

Weisburd, D. (2015). THE LAW OF CRIME CONCENTRATION AND THE CRIMINOLOGY OF PLACE*. *Criminology*, *53*(2), 133–157. https://doi.org/10.1111/1745-9125.12070

Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, *37*(2), 445–480. https://doi.org/10.1007/s10940-020-09457-7

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., & Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, *94*, 101789. https://doi.org/10.1016/j.compenvurbsys.2022.101789