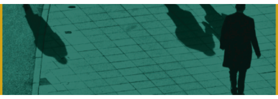
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)
180, *Part* 3, *pp.* 689–722

Interpretable classification models for recidivism prediction

Jiaming Zeng, Berk Ustun and Cynthia Rudin

Massachusetts Institute of Technology, Cambridge, USA

[Received March 2015. Final revision June 2016]

Summary. We investigate a long-debated question, which is how to create predictive models of recidivism that are sufficiently accurate, transparent and interpretable to use for decision making. This question is complicated as these models are used to support different decisions, from sentencing, to determining release on probation to allocating preventative social services. Each case might have an objective other than classification accuracy, such as a desired true positive rate TPR or false positive rate FPR. Each (TPR, FPR) pair is a point on the receiver operator characteristic (ROC) curve. We use popular machine learning methods to create models along the full ROC curve on a wide range of recidivism prediction problems. We show that many methods (support vector machines, stochastic gradient boosting and ridge regression) produce equally accurate models along the full ROC curve. However, methods that are designed for interpretability (classification and regression trees and C5.0) cannot be tuned to produce models that are accurate and/or interpretable. To handle this shortcoming, we use a recent method called supersparse linear integer models to produce accurate, transparent and interpretable scoring systems along the full ROC curve. These scoring systems can be used for decision making for many different use cases, since they are just as accurate as the most powerful black box machine learning models for many applications, but completely transparent, and highly interpretable.

Keywords: Binary classification; Interpretability; Machine learning; Recidivism; Scoring systems

1. Introduction

Forecasting has been used for criminology applications since the 1920s (Borden, 1928; Burgess, 1928) when various factors derived from age, race, prior offence history, employment, grades and neighbourhood background were used to estimate success of parole. Many things have changed since then, including the fact that we have developed machine learning methods that can produce accurate predictive models and have collected large high dimensional data sets on which to apply them.

Prediction of recidivism is still extremely important. In the USA, for example, a minority of individuals commit the majority of the crimes (Wolfgang, 1987): these are the ‘power few’ of Sherman (2007) on which we should focus our efforts. We want to ensure that public resources are directed effectively, be they correctional facilities or preventative social services. Milgram (2014) recently discussed the critical importance of accurately predicting whether an individual who is released on bail poses a risk to public safety, pointing out that high risk individuals are being released 50% of the time whereas low risk individuals are being released less often than they should be. Her observations are in line with long-standing work on clinical *versus* actuarial

Address for correspondence: Jiaming Zeng, Department of Mathematics, Massachusetts Institute of Technology, 355 Massachusetts Avenue, Cambridge, MA 02139, USA.
E-mail: jiaming@alum.mit.edu

0964–1998/17/180689

judgement, which shows that humans, on their own, are not as good at risk assessment as statistical models (Dawes *et al.*, 1989; Grove and Meehl, 1996). This is the reason that several US states have mandated the use of predictive models for sentencing decisions (Pew Center of the States, Public Safety Performance Project, 2011; Wroblewski, 2014).

There has been some controversy about whether sophisticated machine learning methods (such as random forests; see for example Breiman (2001a), Berk *et al.* (2009) and Ritter (2013)) are necessary to produce accurate predictive models of recidivism, or if traditional approaches such as logistic regression or linear discriminant analysis would suffice (see, for example, Tollenaar and van der Heijden (2013), Berk and Bleich (2013) and Bushway (2013)). Random forests may produce accurate predictive models, but these models effectively operate as black boxes, which make it difficult to understand *how* the input variables are producing a predicted outcome. If a simpler, more transparent, but equally accurate predictive model could be developed, it would be more usable and defensible for many decision-making applications. There is a precedent for using such models in criminology (Steinhart, 2006; Andrade, 2009); Ridgeway (2013) argued that a 'decent transparent model that is actually used will outperform a sophisticated system that predicts better but sits on a shelf'. This discussion is captured nicely by Bushway (2013), who contrasted the works of Berk and Bleich (2013) and Tollenaar and van der Heijden (2013). Berk and Bleich (2013) claimed that we need sophisticated machine learning methods because of their substantial benefits in accuracy, whereas Tollenaar and van der Heijden (2013) claimed that 'modern statistical, data mining and machine learning models provides no real advantage over logistic regression and LDA', assuming that humans have done appropriate preprocessing. In this work, we argue that the answer to the question is far more subtle than a simple yes or no.

In particular, the answer depends on how the models will be used for decision making. For each use case (e.g. sentencing, parole decisions and policy interventions), we might need a decision point at a different level of true positive rate TPR and false positive rate FPR (see also Ritter (2013)). Each (TPR, FPR) pair is a point on the receiver operator characteristic (ROC) curve. To determine whether one method is better than another, we must consider the appropriate point along the ROC curve for decision making. As we show, for a wide range of recidivism prediction problems, many machine learning methods (support vector machines (SVMs) or random forests) produce equally accurate predictive models along the ROC curve. However, there are trade-offs between accuracy, transparency and interpretability: methods that are designed to yield transparent models (classification and regression trees (CART); C5.0) cannot be tuned to produce as accurate models along the ROC curve and do not always yield models that are interpretable. This is not to say that interpretable models for prediction of recidivism do not exist. The fact that many machine learning methods produce models with similar levels of predictive accuracy indicates that there is a large class of approximately equally accurate predictive models (called the 'Rashomon' effect by Breiman (2001b)). In this case, there may be interpretable models that also attain the same level of accuracy. Finding models that are accurate and interpretable, however, is computationally challenging.

In this paper, we explore whether such accurate yet interpretable models exist and how to find them. For this, we use a new machine learning method known as a supersparse linear integer model (SLIM) (Ustun and Rudin, 2015) to learn *scoring systems* from data. Scoring systems have been used for many criminal justice applications because they let users make quick predictions by adding, subtracting and multiplying a few small numbers (see, for example, Hoffman and Adelberg (1980), US Sentencing Commission (1987) and Pennsylvania Commission on Sentencing (2012)). In contrast with existing tools, which have been built by using heuristic approaches (see, for example, Gottfredson and Snyder (2005)), the models that are built by

SLIM are fully optimized for accuracy and sparsity, and can handle additional constraints (e.g. bounds on the false positive rate and monotonicity properties for the coefficients). We use SLIM to produce a set of simple scoring systems at different decision points across the full ROC curve and provide a comparison with other popular machine learning methods. Our findings show that the SLIM scoring systems are often just as accurate as the most powerful black box machine learning models, but transparent and highly interpretable.

1.1. Structure

The remainder of this paper is structured as follows. In Section 1.2, we discuss related work. In Section 2, we describe how we derived six recidivism prediction problems. In Section 3, we provide a brief overview of SLIM and describe several new techniques that can reduce the computation that is required to produce scoring systems. In Section 4, we compare the accuracy and interpretability of models produced by the nine machine learning methods on the six recidivism prediction problems. We include additional results that are related to the accuracy and interpretability of models from different methods in Appendix A. An on-line supplement with additional appendices can be found at `http://arxiv.org/abs/1503.07810`.

1.2. Related work

Predictive models for recidivism have been in widespread use in many countries and many areas of the criminal justice system since the early 1920s (see, for example, Borden (1928), Burgess (1928) and Tibbitts (1931)). The use of these tools has been spurred on by continued research into the superiority of actuarial judgement (Dawes *et al.*, 1989; Grove and Meehl, 1996) as well as a desire to use limited public resources efficiently (Clements, 1996; Simon, 2005; McCord, 1978, 2003). In the USA, federal guidelines currently mandate the use of a predictive recidivism measure known as the criminal history category for sentencing (US Sentencing Commission, 1987). Besides the USA, countries that currently use risk assessment tools include Canada (Hanson and Thornton, 2003), the Netherlands (Tollenaar and van der Heijden, 2013) and the UK (Howard *et al.*, 2009). Applications of these tools can be seen in evidence-based sentencing (Hoffman, 1994), corrections and prison administration (Belfrage *et al.*, 2000), informing release on parole (Pew Center of the States, Public Safety Performance Project, 2011), determining the level of supervision during parole (Barnes and Hyatt, 2012; Ritter, 2013), determining appropriate sanctions for violations of parole (Turner *et al.*, 2009), and targeted policy interventions (Lowenkamp and Latessa, 2004).

Our paper focuses on binary classification models to predict general recidivism (i.e. recidivism of any type of crime) as well as crime-specific recidivism (i.e. recidivism for drug, general violence, domestic violence, sexual violence and fatal violence offences). Risk assessment tools for general recidivism include the salient factor score (Hoffman and Adelberg, 1980; Hoffman, 1994), the offender group reconviction scale (Copas and Marshall, 1998; Maden *et al.*, 2006; Howard *et al.*, 2009), the statistical information of recidivism scale (Nafekh and Motiuk, 2002) and the ‘Level of service/case management inventory’ (Andrews and Bonta, 2000). Crime-specific applications include risk assessment tools for domestic violence (see, for example, the spousal abuse risk assessment of Kropp and Hart (2000)), sexual violence (see, for example, Hanson and Thornton (2003) and Langton *et al.* (2007)) and general violence (see, for example, the ‘Historical clinical and risk management’ tool of Webster (1997), or the ‘Structured assessment of violence risk in youth’ tool of Borum (2006)).

The scoring systems that we present in this paper are designed to mimic the form of risk scores that are currently used throughout the criminal justice system—i.e. linear classification models that require users only to add, subtract and multiply a few small numbers to make a prediction (Ustun and Rudin, 2015). These tools are unique in that they allow users to make quick predictions by hand, without a computer, calculator or nomogram (which is a visualization tool for more difficult calculations). Current examples of such tools include the salient factor score (Hoffman and Adelberg, 1980), the criminal history category (US Sentencing Commission, 1987) and the offence gravity score (Pennsylvania Commission on Sentencing, 2012). Our approach aims to produce scoring systems that are fully optimized for accuracy and sparsity without any post-processing. In contrast, current tools are produced through heuristic approaches that primarily involve logistic regression with some *ad hoc* post-processing to ensure that the models are sparse and use integer coefficients (see, for example, the methods that were described in Gottfredson and Snyder (2005)).

Our scoring systems differ from existing tools in that they directly output a predicted outcome (i.e. prisoner i will lapse into crime) as opposed to a predicted probability of the outcome (i.e. the predicted probability that prisoner i will lapse into crime is 90%). The predicted probabilities from existing tools are typically converted into an outcome by imposing a threshold (i.e. classify a prisoner as 'high risk' if the predicted probability of arrest is greater than 70%). In practice, users arbitrarily pick several thresholds to translate predicted probabilities into an ordinal outcome (e.g. prisoner i is 'low risk', if the predicted probability is less than 30%, 'medium risk' if the predicted probability is less than 60% and 'high risk' otherwise). These arbitrary thresholds make it difficult, if not impossible, to assess the predictive accuracy of the tools effectively (Hannah-Moffat, 2013). Netter (2007), for instance, mentioned that 'the possibility of making a prediction error (false positive or false negative) using a risk tool is probable, but not easily determined'. In contrast with existing tools, the scoring systems let users assess accuracy in a straightforward way (i.e. through the true positive rate and true negative rate). Further, our approach has the advantage that it can yield a scoring system that optimizes the class-based accuracy at a particular decision point (i.e. produces the model that maximizes the true positive rate, given a false positive rate of at most 30%).

Our work is related to a stream of research that has aimed to leverage new methods for predictive modelling in criminology. In contrast with our work, much of the research to date has focused on improving predictive accuracy by training powerful black box models such as random forests (Breiman, 2001a) and stochastic gradient boosting (SGB) (Friedman, 2002). Random forests (Breiman, 2001a), in particular, have been used for several criminological applications, including predicting homicide offender recidivism (Neuilly *et al.*, 2011), predicting serious misconduct among incarcerated prisoners (Berk *et al.*, 2006), forecasting potential murders for criminals on probation or parole (Berk *et al.*, 2009), forecasting domestic violence and helping to inform court decisions at arraignment (Berk and Sorenson, 2014). We note that not all studies in black box models used (Berk *et al.*, 2005), for instance, help the Los Angeles Sheriff's Department to develop a simple and practical screener to forecast domestic violence by using decision trees. More recently Goel *et al.* (2016) developed a simple scoring system to help the New York Police Department to address stop and frisk by first running logistic regression, and then rounding the coefficients.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Data and prediction problems

Each problem is a binary classification problem with $N = 33796$ prisoners and $P = 48$ input variables. The goal is to predict whether a prisoner will be arrested for a certain type of crime within 3 years of being released from prison. In what follows, we describe how we created each prediction problem.

2.1. Database details

We derived the recidivism prediction problems in our paper from the 'Recidivism of prisoners released in 1994' database, assembled by the US Department of Justice, Bureau of Justice Statistics (2014). It is the largest publicly available database on prisoner recidivism in the USA. The study tracked 38624 prisoners for 3 years following their release from prison in 1994. These prisoners were randomly sampled from the population of all prisoners released from 15 US states (Arizona, California, Delaware, Florida, Illinois, Maryland, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Texas and Virginia). The population sampled accounts for roughly two-thirds of all prisoners who were released from prison in the USA in 1994. Other studies that use this database include Bhati and Piquero (2007), Bhati (2007) and Zhang *et al.* (2009).

The database is composed of 38624 rows and 6427 columns, where each row represents a prisoner and each column represents a feature (i.e. a field of information for a given prisoner). The 6427 columns consist of 91 fields that were recorded before or during release from prison in 1994 (e.g. date of birth and effective sentence length) and 64 fields that were repeatedly recorded for up to 99 different arrests in the 3-year follow-up period (for example, if a prisoner was rearrested three times within 3 years, three record cycles would be recorded). The information for each prisoner is sourced from record of arrest and prosecution sheets that are kept by state law enforcement agencies and/or the Federal Bureau of Investigation. A detailed descriptive analysis of the database was carried out by statisticians at the US Bureau of Justice Statistics (Langan and Levin, 2002). This study restricted its attention to 33796 of the 38624 prisoners to exclude extraordinary or unrepresentative release cases. To be selected for the analysis of Langan and Levin (2002), a prisoner had to be alive during the 3-year follow-up period and had to have been released from prison in 1994 for an original sentence that was at least 1 year or longer. Prisoners with certain release types—release to custody, detainer or warrant, absent without leave, escape, transfer, administrative release and release on appeal—were excluded. To mirror the approach of Langan and Levin (2002), we restricted our attention to the same subset of prisoners.

This data set has some serious flaws which we point out below. To begin, many important factors that could be used to predict recidivism are missing, and many included factors are sufficiently noisy to be excluded from our preliminary experiments. The information about levels of education is extremely minimal; we do not even know whether each prisoner attended college or completed high school. The information about courses in prison is only an indicator of whether the inmate took any education or vocation courses at all. Also, there is no family history for each prisoner (e.g. foster care) and no record of visitors while in prison (e.g. indicators of caring family members or friends). There is no information about re-entry programmes or employment history. Although some of these factors exist, such as drug or alcohol treatment and in-prison vocational programmes, the data are highly incomplete and therefore have been excluded from our analysis. For example, for drug treatment, less than 14% of the prisoners had a valid entry. The rest were 'unknown'. To include as many prisoners as possible, we chose to exclude factors with extremely sparse information.

2.2. Deriving input variables

We provide a summary of the $P=48$ input variables that were derived from the database in

Table 1. Overview of input variables for all prediction problems†

Input variable	$P(x_{ij}=1)$	*Definition*
female	0.06	Prisoner *i* is female
prior_alcohol_abuse	0.20	Prisoner *i* has a history of alcohol abuse
prior_drug_abuse	0.16	Prisoner *i* has a history of drug abuse
age_at_release⩽17	0.00	Prisoner *i* was ⩽17 years old at release in 1994
age_at_release_18_to_24	0.19	Prisoner *i* was 18–24 years old at release in 1994
age_at_release_25_to_29	0.21	Prisoner *i* was 25–29 years old at release in 1994
age_at_release_30_to_39	0.38	Prisoner *i* was 30–39 years old at release in 1994
age_at_release⩾40	0.21	Prisoner *i* was ⩾40 years old at release in 1994
released_unconditional	0.11	Prisoner *i* was released at expiration of sentence
released_conditional	0.87	Prisoner *i* was released on parole or probation
time_served⩽6mo	0.23	Prisoner *i* served ⩽6 months
time_served_7_to_12mo	0.20	Prisoner *i* served 7–12 months
time_served_13_to_24mo	0.23	Prisoner *i* served 13–24 months
time_served_25_to_60mo	0.25	Prisoner *i* served 25–60 months
time_served⩾61mo	0.10	Prisoner *i* served ⩾61 months
infraction_in_prison	0.24	Prisoner *i* has a record of misconduct in prison
age_1st_arrest⩽17	0.14	Prisoner *i* was ⩽17 years old at 1st arrest
age_1st_arrest_18_to_24	0.61	Prisoner *i* was 18–24 years old at 1st arrest
age_1st_arrest_25_to_29	0.10	Prisoner *i* was 25–29 years old at 1st arrest
age_1st_arrest_30_to_39	0.09	Prisoner *i* was 30–39 years old at 1st arrest
age_1st_arrest⩾40	0.04	Prisoner *i* was ⩾40 years old at 1st arrest
age_1st_confinement⩽17	0.03	Prisoner *i* was ⩽17 years old at 1st confinement
age_1st_confinement_18_to_24	0.46	Prisoner *i* was 18–24 years old at 1st confinement
age_1st_confinement_25_to_29	0.18	Prisoner *i* was 25–29 years old at 1st confinement
age_1st_confinement_30_to_39	0.21	Prisoner *i* was 30–39 years old at 1st confinement
age_1st_confinement⩾40	0.12	Prisoner *i* was ⩾40 years old at 1st confinement
prior_arrest_for_drug	0.47	Prisoner *i* was once arrested for a drug offence
prior_arrest_for_property	0.67	Prisoner *i* was once arrested for a property offence
prior_arrest_for_public_order	0.62	Prisoner *i* was once arrested for a public order offence
prior_arrest_for_general_violence	0.52	Prisoner *i* was once arrested for general violence
prior_arrest_for_domestic_violence	0.04	Prisoner *i* was once arrested for domestic violence
prior_arrest_for_sexual_violence	0.03	Prisoner *i* was once arrested for sexual violence
prior_arrest_for_fatal_violence	0.01	Prisoner *i* was once arrested for fatal violence
prior_arrest_for_multiple_types	0.77	Prisoner *i* was once arrested for multiple types of crime
prior_arrest_for_felony	0.84	Prisoner *i* was once arrested for a felony
prior_arrest_for_misdemeanor	0.49	Prisoner *i* was once arrested for a misdemeanour
prior_arrest_for_local_ordinance	0.01	Prisoner *i* was once arrested for local ordinance
prior_arrest_with_firearms_involved	0.09	Prisoner *i* was once arrested for an incident involving firearms
prior_arrest_with_child_involved	0.17	Prisoner *i* was once arrested for an incident involving children
no_prior_arrests	0.12	Prisoner *i* has no prior arrests
prior_arrests⩾1	0.88	Prisoner *i* has at least 1 prior arrest
prior_arrests⩾2	0.78	Prisoner *i* has at least 2 prior arrests
prior_arrests⩾5	0.60	Prisoner *i* has at least 5 prior arrests
multiple_prior_prison_time	0.43	Prisoner *i* has been to prison multiple times
any_prior_jail_time	0.47	Prisoner *i* has been to prison at least once
multiple_prior_jail_time	0.29	Prisoner *i* has been to prison multiple times
any_prior_probation_or_fine	0.42	Prisoner *i* has been on probation or paid a fine at least once
multiple_prior_probation_or_fine	0.22	Prisoner *i* has been on probation or paid a fine multiple times

†Each variable is a binary rule of the form $x_{ij} \in \{0, 1\}$. We list conditions required for $x_{ij} = 1$ under the definition column.

Table 2. Overview of recidivism prediction problems†

Prediction problem	*$P(y_i=1)$ (%)*	*Outcome variable*
`arrest`	59.0	$y_i=1$ if prisoner i is arrested for any offence within 3 years of release from prison
`drug`	20.0	$y_i=1$ if prisoner i is arrested for a drug-related offence (e.g. possession or trafficking) within 3 years of release from prison
`general_violence`	19.1	$y_i=1$ if prisoner i is arrested for a violent offence (e.g. robbery or aggravated assault) within 3 years of release from prison
`domestic_violence`	3.5	$y_i=1$ if prisoner i is arrested for domestic violence within 3 years of release from prison
`sexual_violence`	3.0	$y_i=1$ if prisoner i is arrested for sexual violence within 3 years of release from prison
`fatal_violence`	0.7	$y_i=1$ if prisoner i is arrested for murder or manslaughter within 3 years of release from prison

†The percentages $P(y_i=1)$ do not add up to 100% because a prisoner could be arrested for multiple types of crime at one time (e.g. both drug and public order offences) and could also be arrested multiple times over the 3-year follow-up period.

Table 1. We encoded each input variable as a binary rule of the form $x_{ij}\in\{0,1\}$, $j=1,\ldots,P$, where $x_{ij}=1$ if condition j holds true about prisoner i. This allows a linear model to encode non-linear functions of the original variables. We refer to input variables in the text by using italics (e.g. *female*). All prediction problems in Table 2 and all machine learning methods in Table 4 in Section 4.1.4 use these same input variables.

The final set of input variables is representative of well-known risk factors for recidivism (Bushway and Piehl, 2007; Crow, 2008) and has been used in risk assessment tools since 1928 (see, for example, Borden (1928), Hinojosa *et al.* (2005), Berk *et al.* (2006) and Baradaran (2013)). They include

(a) information about prison release in 1994 (e.g. *time_served*, *age_at_release* and *infraction_in_prison*),
(b) information from past arrests, sentencing and convictions (e.g. *prior_arrests*$\geqslant$*1* and *any_prior_jail_time*),
(c) history of substance abuse (e.g. *alcohol_abuse*) and
(d) gender (e.g. *female*).

(The *prior_arrest* variable does not count the original crime for which they were released from prison in 1994; thus, about 12% of the prisoners have *no_prior_arrests* = 1 even though they were arrested at least once.) These input variables are advantageous because

(a) the information is easily accessible to law enforcement officials (all the above information can be found in state record of arrest and prosecution sheets) and
(b) they do not include socio-economic factors such as race, which would directly eliminate the potential to use these tools in applications such as sentencing.

We note that encoding the input variables as binary values presents many advantages. They produce models that are easier to understand (removing the wide range that is presented by continuous variables), and they avoid potential confusion stemming from coefficients of normalized inputs (for instance, after undoing the normalization for normalized coefficients, a small coefficient might be highly influential if it applies to a variable taking large values). Binarization is especially useful for SLIM as we can fit SLIM models by solving a slightly easier discrete

Table 3. Table of conditional probabilities for all input variables (rows) and prediction problems (columns)†

Input variable	*Probabilities for the following prediction problems:*					
	arrest	drug	general_ violence	domestic_ violence	sexual_ violence	fatal_ violence
female	0.54	0.21	0.11	0.02	0.01	0.0005
prior_alcohol_abuse	0.58	0.18	0.20	0.04	0.03	0.01
prior_drug_abuse	0.61	0.23	0.21	0.03	0.03	0.004
age_at_release$\leqslant$*17*	0.84	0.35	0.31	0.01	0.01	0.04
age_at_release_18_to_24	0.71	0.24	0.25	0.04	0.03	0.01
age_at_release_25_to_29	0.66	0.23	0.21	0.04	0.03	0.01
age_at_release_30_to_39	0.59	0.20	0.17	0.04	0.03	0.01
age_at_release$\geqslant$*40*	0.41	0.12	0.09	0.02	0.03	0.003
released_unconditional	0.65	0.20	0.23	0.06	0.04	0.01
released_conditional	0.58	0.20	0.17	0.03	0.03	0.01
time_served$\leqslant$*6mo*	0.67	0.27	0.19	0.04	0.03	0.01
time_served_7_to_12mo	0.63	0.22	0.19	0.04	0.03	0.01
time_served_13_to_24mo	0.59	0.20	0.17	0.04	0.03	0.01
time_served_25_to_60mo	0.53	0.16	0.17	0.03	0.03	0.01
time_served$\geqslant$*61mo*	0.48	0.11	0.15	0.02	0.04	0.004
infraction_in_prison	0.65	0.19	0.20	0.01	0.04	0.01
age_1st_arrest$\leqslant$*17*	0.73	0.27	0.27	0.04	0.04	0.01
age_1st_arrest_18_to_24	0.64	0.22	0.20	0.04	0.03	0.01
age_1st_arrest_25_to_29	0.47	0.14	0.10	0.02	0.02	0.005
age_1st_arrest_30_to_39	0.34	0.10	0.06	0.02	0.02	0.003
age_1st_arrest$\geqslant$*40*	0.21	0.05	0.03	0.01	0.02	0.002
age_1st_confinement$\leqslant$*17*	0.78	0.28	0.29	0.04	0.04	0.02
age_1st_confinement_18_to_24	0.68	0.24	0.23	0.05	0.04	0.01
age_1st_confinement_25_to_29	0.60	0.20	0.17	0.03	0.03	0.005
age_1st_confinement_30_to_39	0.50	0.16	0.12	0.03	0.02	0.003
age_1st_confinement$\geqslant$*40*	0.34	0.09	0.07	0.01	0.02	0.002
prior_arrest_for_drug	0.68	0.32	0.21	0.04	0.02	0.01
prior_arrest_for_property	0.67	0.24	0.22	0.04	0.03	0.01
prior_arrest_for_public_order	0.65	0.24	0.22	0.04	0.03	0.01
prior_arrest_for_general_violence	0.67	0.25	0.26	0.05	0.04	0.01
prior_arrest_for_domestic_violence	0.66	0.21	0.27	0.13	0.04	0.01
prior_arrest_for_sexual_violence	0.49	0.13	0.16	0.04	0.06	0.01
prior_arrest_for_fatal_violence	0.54	0.19	0.21	0.04	0.03	0.01
prior_arrest_for_multiple_crime_types	0.64	0.23	0.21	0.04	0.03	0.01
prior_arrest_for_felony	0.60	0.21	0.19	0.04	0.03	0.01
prior_arrest_for_misdemeanor	0.69	0.26	0.24	0.06	0.03	0.01
prior_arrest_for_local_ordinance	0.91	0.29	0.43	0.15	0.05	0.02
prior_arrest_with_firearms_involved	0.70	0.30	0.27	0.06	0.03	0.01
prior_arrest_with_child_involved	0.48	0.13	0.14	0.03	0.06	0.01
no_prior_arrests	0.32	0.07	0.08	0.02	0.02	0.003
prior_arrest$\geqslant$*1*	0.63	0.22	0.19	0.04	0.03	0.01
prior_arrest$\geqslant$*2*	0.66	0.23	0.20	0.04	0.03	0.01
prior_arrest$\geqslant$*5*	0.70	0.25	0.22	0.04	0.03	0.01
multiple_prior_prison_time	0.65	0.23	0.19	0.03	0.03	0.01
any_prior_jail_time	0.69	0.25	0.21	0.04	0.03	0.01
multiple_prior_jail_time	0.73	0.27	0.22	0.04	0.03	0.01
any_prior_probation_or_fine	0.67	0.24	0.20	0.04	0.03	0.01
multiple_prior_probation_or_fine	0.71	0.27	0.22	0.05	0.03	0.01

†Each cell represents the conditional probability $P(y=1|x=1)$ where x is the input variable that is specified in the row and y is the outcome variable for the prediction problem specified in the column.

optimization problem when the data contain only binary input variables (as discussed in Section 3.3). In appendix E of the on-line supplement, we explore the change in predictive accuracy if continuous variables are included and show that the changes in performance are minor for most methods. There are some exceptions; for example, the CART and C5.0T methods experienced an improvement of 4.6% for `drug` and SVM radial basis function (RBF) experienced a 7.7% improvement for `fatal_violence`. Yet, even for these methods, no clear improvement is seen across all problems.

2.3. Deriving outcome variables

We created a total of six recidivism prediction problems by encoding a binary outcome variable $y_i \in \{-1, 1\}$ such that $y_i = 1$ if a prisoner is arrested for a particular type of crime within 3 years after being released from prison. For clarity, we refer to each prediction problem in the text by using Courier fount (e.g. `arrest`). We provide details on each recidivism prediction problem in Table 2. These include an arrest for any crime (`arrest`), an arrest for a drug-related offence (`drug`) or an arrest for a certain type of violent offence (`general_violence`, `domestic_violence`, `sexual_violence` and `fatal_violence`).

In the data set, all crime types can be broken down into smaller subcategories (e.g. `fatal_violence` can be broken into six subcategories such as `murder` and `vehicular_manslaughter`). We chose to use the broader crime categories for conciseness and clarity. Indeed, the study by Langan and Levin (2002) also split crimes into the same major categories. We note that the outcomes of violent offences are mutually exclusive, as different types of violence are treated differently within the US legal system. In other words, $y_i = 1$ for `general_violence` does not necessarily imply that $y_i = 1$ for `domestic_violence`, `sexual_violence` or `fatal_violence`).

2.4. Relationships between input and output variables

Table 3 lists the conditional probabilities $P(y = 1 | x_j = 1)$ between the outcome variable y and each input variable x_j for all prediction problems. Using Table 3, we can identify strong associations between the input and output for each prediction problem. These associations can help to uncover insights into each problem and also help to validate predictive models in Section 4.4 qualitatively.

Consider, for instance, the `arrest` problem. Here, we can see that prisoners who are released from prison at a later age are less likely to be arrested (as the probability for arrest decreases monotonically as *age_at_release* increases). This also appears to be so for prisoners who were first confined (i.e. sent to prison) at an older age (see, for example, *age_of_first_confinement*). In addition, we can also see that prisoners with more prior arrests have a higher likelihood of being arrested (as the probability for arrest increases monotonically with *prior_arrest*).

Similar insights can be made for crime-specific prediction problems. In `drug`, for instance, we see that prisoners who were previously arrested for a drug-related offence are more likely to be rearrested for a drug-related offence (32%) than those who were previously arrested for any other type of offence. Likewise, looking at `domestic_violence`, we see that the prisoners with the greatest probability of being arrested for a domestic violence crime are those with a history of domestic violence (13%).

3. Supersparse linear integer models

SLIM is a new machine learning method for creating *scoring systems*—i.e. binary classification models that require users only to add, subtract and multiply a few small numbers to make a prediction (Ustun and Rudin, 2015). Scoring systems are widely used because they allow users to make quick predictions, without the use of a computer, and without extensive training in statistics. These models are also useful because their high degree of sparsity and integer coefficients let users easily gauge the influence of multiple input variables on the predicted outcome (see Section 4.4 for an example). In what follows, we provide a brief overview of SLIM and provide several new techniques to reduce the computation for problems with binary input variables.

3.1. Framework and optimization problem

SLIM scoring systems are linear classification models of the form

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_{j=1}^{P} \lambda_j x_{ij} > \lambda_0, \\ -1 & \text{if } \sum_{j=1}^{P} \lambda_j x_{ij} \leqslant \lambda_0. \end{cases}$$

Here, $\lambda_1, \ldots, \lambda_P$ represent the coefficients (i.e. the 'points' for the input variables $j = 1, \ldots, P$) and λ_0 represents an intercept (i.e. the 'threshold score' that must be surpassed to predict $\hat{y}_i = 1$).

The values of the coefficients are determined from data by solving a discrete optimization problem that has the form

$$\min_{\boldsymbol{\lambda}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i \neq \hat{y}_i) + C_0 \sum_{j=1}^{P} \mathbb{1}(\lambda_j \neq 0) + \epsilon \sum_{j=1}^{P} |\lambda_j|$$
$$\text{such that } (\lambda_0, \lambda_1, \ldots, \lambda_P) \in \mathcal{L}. \tag{1}$$

Here, the objective directly minimizes the error rate $(1/N)\, \Sigma_{i=1}^{N} \mathbb{1}(y_i \neq \hat{y}_i)$ and directly penalizes the number of non-zero terms $\Sigma_{j=1}^{P} \mathbb{1}(\lambda_j \neq 0)$. The constraints restrict coefficients to a finite set such as $\mathcal{L} = \{-10, \ldots, 10\}^{P+1}$. Optionally, one could include additional operational constraints on the accuracy and interpretability of the scoring system desired.

The objective includes a *tiny* penalty on the absolute value of the coefficients to restrict coefficients to coprime values without affecting accuracy or sparsity. To illustrate the use of this penalty, consider a classifier such as $\hat{y} = \text{sgn}(x_1 + x_2)$. If SLIM minimized only the misclassification rate and the number of terms (the first two terms of the objective), then $\hat{y} = \text{sgn}(2x_1 + 2x_2)$ would have the same objective value as $\hat{y} = \text{sgn}(x_1 + x_2)$ because it makes the same predictions and has the same number of non-zero coefficients. Since coefficients are restricted to a discrete set, we use this *tiny* penalty on the absolute value of these coefficients so that SLIM chooses the classifier with the smallest (coprime) coefficients, $\hat{y} = \text{sgn}(x_1 + x_2)$.

The C_0-parameter represents the maximum accuracy that SLIM is willing to sacrifice to remove a feature from the optimal scoring system. If, for instance, C_0 is set within the range $(1/N, 2/N)$, we would sacrifice the accuracy of one observation to have a model with one fewer feature. Given C_0, we can set the l_1-penalty parameter ϵ to any value

$$0 < \epsilon < \frac{\min(1/N, C_0)}{\max_{\{\lambda_j\}_j \in \mathcal{L}} \sum_{j=1}^{P} |\lambda_j|}$$

so that it does not affect the accuracy or sparsity of the optimal classifier but only induces the coefficients to be coprime for the features that are selected.

SLIM differs from traditional machine learning methods because it directly optimizes accuracy and sparsity without making approximations that other methods make for scalability (e.g. controlling for accuracy using convex surrogate loss functions). By avoiding these approximations, SLIM sacrifices the ability to fit a model in seconds or in a way that scales to extremely large data sets. In return, however, it gains the ability to fit models that are highly customizable, since one could directly encode a wide range of operational constraints in its integer programming (IP) formulation. In this paper, we primarily make use of a simple constraint to limit the number of non-zero coefficients; however, it is also natural to incorporate constraints on class-specific accuracy, structural sparsity and prediction (see Ustun and Rudin (2015)).

In this paper we trained the following version of SLIM, which is different from problem (1) in that it includes class weights and has specific constraints on the coefficients:

$$\begin{aligned}\min_{\boldsymbol{\lambda}} \frac{W^+}{N} \sum_{i\in\mathcal{I}^+} \mathbb{1}(y_i \neq \hat{y}_i) + \frac{W^-}{N} \sum_{i\in\mathcal{I}^-} \mathbb{1}(y_i \neq \hat{y}_i) + C_0 \sum_{j=1}^{P} \mathbb{1}(\lambda_j \neq 0) + \epsilon \sum_{j=1}^{P} |\lambda_j| \\ \text{such that } \sum_{j=1}^{P} \mathbb{1}(\lambda_j \neq 0) \leqslant 8, \qquad (2) \\ \lambda_j \in \{-10, \ldots, 10\} \qquad \text{for} j = 1, \ldots, P, \\ \lambda_0 \in \{-100, \ldots, 100\}.\end{aligned}$$

In this formulation, the constraints restrict each coefficient λ_j to an integer between -10 and 10, the threshold λ_0 to an integer between -100 and 100, the number of non-zero coefficients to at most 8 (i.e. within the range of cognitive entities that humans could handle, as per Miller (1956)). The parameters W^+ and W^- are class-based weights that control the accuracy on positive and negative examples. We typically choose values of W^+ and W^- such that $W^+ + W^- = 2$, so that we recover an error minimizing formulation by setting $W^+ = W^- = 1$. The C_0-parameter was set to a sufficiently small value so that SLIM would not sacrifice accuracy for sparsity: given W^+ and W^-, we can set C_0 to any value

$$0 < C_0 < \min\{W^-, W^+\}/(NP)$$

to ensure this condition. The ϵ-parameter was set to a sufficiently small value so that SLIM would produce a model with coprime coefficients without affecting accuracy or sparsity: given W^+, W^- and C_0, we can set ϵ to any value $0 < \epsilon < C_0 / \max \Sigma_{j=1}^{P} |\lambda_j|$ to ensure this condition.

3.2. General supersparse linear integer model integer programming formulation

Training a SLIM scoring system requires solving an IP problem by using a solver such as CPLEX, Gurobi or Com-OR branch and cut. In general, we use the following IP formulation to recover the solution to the optimization problem (2):

$$\min_{\boldsymbol{\lambda},\mathbf{z},\boldsymbol{\Phi},\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} z_i + \sum_{j=1}^{P} \Phi_j$$

$$\text{such that } M_i z_i \geqslant \gamma - \sum_{j=0}^{P} y_i \lambda_j x_{i,j} \qquad i = 1, \ldots, N \text{ (error on } i\text{)}, \qquad (3a)$$

$$\Phi_j = C_0 \alpha_j + \epsilon \beta_j \qquad j = 1, \ldots, P \text{ (penalty for coefficient } j\text{)}, \qquad (3b)$$

$$-\Lambda_j \alpha_j \leqslant \lambda_j \leqslant \Lambda_j \alpha_j \qquad j = 1, \ldots, P \text{ (} l_0\text{-norm)}, \qquad (3c)$$

$$
\begin{aligned}
-\beta_j \leqslant \lambda_j \leqslant \beta_j \qquad & j=1,\ldots,P\ (l_1\text{-norm}), && \text{(3d)}\\
\lambda_j \in \mathbb{Z} \cap [-\Lambda_j, \Lambda_j] \qquad & j=0,\ldots,P\ (\text{coefficient set}),\\
z_i \in \{0,1\} \qquad & i=1,\ldots,N\ (\text{loss variables}),\\
\Phi_j \in \mathbb{R}_+ \qquad & j=1,\ldots,P\ (\text{penalty variables}),\\
\alpha_j \in \{0,1\} \qquad & j=1,\ldots,P\ (l_0\text{-variables}),\\
\beta_j \in \mathbb{R}_+ \qquad & j=1,\ldots,P\ (l_1\text{-variables}).
\end{aligned}
$$

The constraints in expression (3a) compute the error rate by setting the *loss variables* $z_i = \mathbb{1}(y_i \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_i \leqslant 0)$ to 1 if a linear classifier with coefficients $\boldsymbol{\lambda}$ misclassifies example i (or is close to misclassifying it, depending on the margin γ). This is a *big M constraint* for the error rate that depends on scalar parameters γ and M_i (see, for example, Rubin (2009)). The value of M_i represents the maximum score when example i is misclassified and can be set as $M_i = \max_{\lambda \in \mathcal{L}} (\gamma - y_i \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_i)$, which is easy to compute since $\mathcal{L}$ is finite. The value of γ represents the margin, and the objective is penalized when points are either incorrectly classified or within γ of the decision boundary. How close a point is to the decision boundary (or whether it is misclassified) is determined by $y_i \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_i$. When the features are binary, and since the coefficients are integers, γ can naturally be set to any value between 0 and 1. (In other cases, we can set $\gamma = 0.5$ for instance, which makes an implicit assumption on the values of the features.) The constraints in expression (3b) set the total penalty for each coefficient to $\Phi_j = C_0 \alpha_j + \epsilon \beta_j$, where $\alpha_j := \mathbb{1}(\lambda_j \neq 0)$ is defined by big M constraints in expression (3c), and $\beta_j := |\lambda_j|$ is defined by the constraints in expression (3d). We denote the largest absolute value of each coefficient as $\Lambda_j := \max_{\lambda_j \in \mathcal{L}_j} |\lambda_j|$.

Restricting coefficients to a finite set results in significant practical benefits for the SLIM IP formulation, especially in comparison with other IP formulations that minimize the 0–1-loss and/or penalize the l_0-norm. Without the restriction of λ to a bounded set, we would not have a natural choice for the big M constant, which means that the user chooses one that is very large, leading to a less efficient formulation (see, for example, Wolsey (1998)). For SLIM, the big M constants that were used to compute the 0–1-loss in constraint (3a) are bounded as $M_i \leqslant \max_{\boldsymbol{\lambda} \in \mathcal{L}} (\gamma - y_i \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_i)$, and the big M constant that was used to compute the l_0-norm in constraints (3c) is bounded as $\Lambda_j \leqslant \max_{\lambda_j \in \mathcal{L}_j} |\lambda_j|$. Bounding these constants leads to a tighter linear programming relaxation, which narrows the integrality gap, and improves the ability of commercial IP solvers to obtain a proof of optimality more quickly.

3.3. Improved superparse linear integer model integer programming formulation

The following formulation provides a tighter relaxation of the IP which reduces computation. It relies on the fact that, when the input variables are binary, we are likely to obtain repeated feature values among observations.

$$
\min_{\boldsymbol{\lambda}, \mathbf{z}, \boldsymbol{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{W^+}{N} \sum_{s \in \mathcal{S}} n_s z_s + \frac{W^-}{N} \sum_{t \in \mathcal{T}} n_t z_t + \sum_{j=1}^{P} \Phi_j
$$

$$
\text{such that } M_s z_s \geqslant 1 - \sum_{j=0}^{P} \lambda_j x_{s,j} \qquad s \in \mathcal{S}\ (\text{error on } s), \tag{4a}
$$

$$
M_t z_t \geqslant \sum_{j=0}^{P} \lambda_j x_{t,j} \qquad t \in \mathcal{T}\ (\text{error on } t), \tag{4b}
$$

$$1 = z_s + z_t \qquad \forall\, s, t : \mathbf{x}_s = \mathbf{x}_t,\, y_s = -y_t \text{ (conflicting labels)}, \tag{4c}$$

$$\Phi_j = C_0 \alpha_j + \epsilon \beta_j \qquad j = 1, \ldots, P \text{ (penalty for coefficient} j\text{)}, \tag{4d}$$

$$-\Lambda_j \alpha_j \leqslant \lambda_j \leqslant \Lambda_j \alpha_j \qquad j = 1, \ldots, P \; (l_0\text{-norm}) \tag{4e}$$

$$-\beta_j \leqslant \lambda_j \leqslant \beta_j \qquad j = 1, \ldots, P \; (l_1\text{-norm}), \tag{4f}$$

$$\lambda_j \in \mathbb{Z} \cap [-\Lambda_j, \Lambda_j] \qquad j = 0, \ldots, P \text{ (coefficient set)},$$

$$z_s, z_t \in \{0, 1\} \qquad s \in \mathcal{S},\; t \in \mathcal{T} \text{ (loss variables)},$$

$$\Phi_j \in \mathbb{R}_+ \qquad j = 1, \ldots, P \text{ (penalty variables)},$$

$$\alpha_j \in \{0, 1\} \qquad j = 1, \ldots, P \; (l_0\text{-variables}),$$

$$\beta_j \in \mathbb{R}_+ \qquad j = 1, \ldots, P \; (l_1\text{-variables}).$$

The main difference between this formulation and that in expression (3) is that we compute the error rate of the classifier by using loss constraints that are expressed in terms of the number of *distinct* points in the data set. Here, the set $\mathcal{S}$ represents the set of distinct points with positive labels, and the set $\mathcal{T}$ represents the set of distinct points with negative examples. The parameters n_s (and n_t) count the number of times that a point of type s (or t) is found in the original data set so that $\Sigma_s n_s = \Sigma_{i=1}^N \mathbb{1}(y_i = 1)$, $\Sigma_t n_t = \Sigma_{i=1}^N \mathbb{1}(y_i = -1)$ and $N = \Sigma_s n_s + \Sigma_t n_t$.

The main computational benefits of this formulation are because

(a) we can reduce the number of loss constraints by counting the number of repeated rows in the data set and
(b) we can directly encode a lower bound on the error rate by counting the number of points s and t with identical feature but opposite labels (i.e. $\mathbf{x}_s = \mathbf{x}_t$ but $y_s = -y_t$).

Here benefit (a) reduces the size of the problem that we pass to an IP solver, and benefit (b) produces a much stronger lower bound on the 0–1-loss (in comparison with the linear programming relaxation), which speeds up the progress of branch-and-bound type algorithms. It would be possible to use this formulation on a data set without binary input variables, though it would not necessarily be effective because it could be much less likely for a data set to contain repeated rows in such a setting.

Another subtle benefit of this formulation is that the margin for the negative points is 0 whereas the margin for the positive points is 1. This means that, for positive points, we have a correct prediction if and only if the score is 1 or greater. For negative points, we have a correct prediction if and only if the score is 0 or less. This provides a slight computational advantage since the negative points do not need to have scores below -1 to be correctly classified, which reduces the size of the big M parameter and the coefficient set. For instance, say that we would like to produce a linear model that encodes 'predict rearrest unless a_1 or a_2 are true'. Using the previous formulation with the margin of $\gamma \in (0, 1)$ on both positives and negatives, the optimal SLIM classifier would be 'rearrest = sgn$(1 - 2a_1 - 2a_2)$'. In contrast, the margin of the current formulation is 'rearrest = sgn$(1 - a_1 - a_2)$', which uses smaller coefficients and produces a slightly simpler model.

3.4. Active set polishing

On large data sets, IP solvers may take a long time to produce an optimal solution or to provide users with a certificate of optimality. Here, we present a *polishing* procedure that can be used to improve the quality of solutions locally. For a fixed set of features, this procedure optimizes the values of coefficients.

The polishing procedure takes as input a feasible set of coefficients from the SLIM IP $\boldsymbol{\lambda}^{\text{feasible}}$ and returns a polished set of coefficients $\boldsymbol{\lambda}^{\text{polished}}$ by solving a simpler IP formulation shown in expression (5). The polishing IP optimizes only the coefficients of features that belong to the *active set* of $\boldsymbol{\lambda}^{\text{feasible}}$, i.e. the set of features with non-zero coefficients $\mathcal{A} := \{j : \lambda_j^{\text{feasible}} \neq 0\}$. The coefficients for features that do not belong to the active set are fixed to 0 so that $\lambda_j = 0$ for $j \notin \mathcal{A}$. In this way, the optimization no longer involves feature selection, and the formulation becomes much easier to solve.

$$\min_{\boldsymbol{\lambda},\mathbf{z},\boldsymbol{\Phi},\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{W^+}{N} \sum_{s\in\mathcal{S}} n_s z_s + \frac{W^-}{N} \sum_{t\in\mathcal{T}} n_t z_t \tag{5a}$$

$$\text{such that } M_s z_s \geqslant 1 - \sum_{j\in\mathcal{A}} \lambda_j x_{s,j} \qquad s \in \mathcal{S} \text{ (error on } s\text{)}, \tag{5b}$$

$$M_t z_t \geqslant \sum_{j\in\mathcal{A}} \lambda_j x_{t,j} \qquad t \in \mathcal{T} \text{ (error on } t\text{)}, \tag{5c}$$

$$1 = z_s + z_t \qquad \forall\, s, t : \mathbf{x}_s = \mathbf{x}_t,\, y_s = -y_t \text{ (conflicting labels)}, \tag{5d}$$

$$\lambda_j \in \mathbb{Z} \cap [-\Lambda_j, \Lambda_j] \qquad j \in \mathcal{A} \quad \text{(coefficient set)},$$

$$z_s, z_t \in \{0, 1\} \qquad s \in \mathcal{S},\ t \in \mathcal{T} \quad \text{(loss variables)}.$$

The polishing IP formulation is especially fast to solve to optimality for classification problems with binary input variables because this limits the number of loss constraints. Say for instance that we wish to polish a set of coefficients with only five non-zero variables; then there are at most $|\{-1, 1\}| \times |\{0, 1\}^5| = 64$ possible unique data points, and thus the same number of possible loss constraints.

In our experiments in Section 4, we use the polishing procedure on all the feasible solutions that we find from the earlier formulation. In all cases, we can solve the polishing IP problem to optimality within a few seconds (i.e. an optimality gap of 0.0%).

4. Experimental results

In this section, we compare the accuracy and interpretability of recidivism prediction models from SLIM to models from eight other popular classification methods. In Section 4.1, we explain the experimental set-up that was used for all the methods. In Section 4.2, we compare the predictive accuracy of the methods with the area under the curve values AUC and ROC curves. In Sections 4.3 and 4.4, we evaluate the interpretability of the models. Finally, in Section 4.5, we present the scoring systems that are generated by SLIM.

4.1. Methodology

In what follows we discuss cost-sensitive classification for imbalanced problems and provide an overview of techniques.

4.1.1. Evaluating predictive accuracy for imbalanced problems

The majority of classification problems that we consider are *imbalanced*, where the data contain a relatively small number of examples from one class and a relatively large number of examples from the other.

Imbalanced problems necessitate changes in the way that we evaluate the performance of classification models. Consider, for instance, a heavily imbalanced problem such as `fatal_violence` where only $P(y_i = 1) = 0.7\%$ of individuals are arrested within 3 years of being released from prison. In this case, a method that maximizes overall classification accuracy is likely to produce a trivial model that predicts that no one will be arrested for fatal offences—a result that is not surprising given that the trivial model is 99.3% accurate on the overall population. Unfortunately, this model will never be able to identify individuals who will be arrested for a fatal offence, and therefore it will be 0% accurate on the population of interest.

To provide a measure of classification model performance on imbalanced problems, we assess the accuracy of a model on the positive and negative classes separately. In our experiments, we report the class-based accuracy of each model by using the *true positive rate* TPR, which reflects the accuracy on the positive class, and the *false positive rate* FPR, which reflects the error rate on the negative class. For a given classification model, we compute these quantities as

$$\mathrm{TPR} = \frac{1}{N^+} \sum_{i \in \mathcal{I}^+} \mathbb{1}(\hat{y}_i = 1),$$

$$\mathrm{FPR} = \frac{1}{N^-} \sum_{i \in \mathcal{I}^-} \mathbb{1}(\hat{y}_i = 1),$$

where $\hat{y}_i$ denotes the predicted outcome for example i, N^+ denotes the number of examples in the positive class $\mathcal{I}^+ = \{i : y_i = 1\}$ and N^- denotes the number of examples from the negative class $\mathcal{I}^- = \{i : y_i = -1\}$. Ideally, a classification model should have high TPR and low FPR (i.e. TPR close to 1 and FPR = 0).

Most classification methods can be adapted to yield a model that is more accurate on the positive class, but only if we are willing to sacrifice some accuracy on examples from the negative class, and vice versa. To illustrate the trade-off of classification accuracy between positive and negative classes, we plot all models that are produced by a given method as points on an ROC curve, which plots TPR on the vertical axis and FPR on the horizontal axis. Having constructed an ROC curve, we then assess the *overall* performance of each method by calculating the *area under the ROC curve*, AUC. (We note that AUC is a summary statistic that is frequently misused in the context of classification problems. It is true that a method with AUC = 1 always produces models that are more accurate than a method with AUC = 0. Other than this simple case, however, it is not possible to state that a method with high AUC always produces models that are more accurate than a method with low AUC.) A detailed discussion of ROC analysis in recidivism prediction can be found in the work of Maloof (2003).

4.1.2. Fitting models over the full receiver operating characteristic curve by using a cost-sensitive approach

Different applications require predictive models at different points of the ROC curve. Models for sentencing, for example, need low FPR to avoid predicting that a low risk individual will reoffend. Models for screening, however, need high TPR to capture as many high risk individuals as possible. In our experiments, we use a *cost-sensitive approach* to produce classification models at different points of the ROC curve (see, for example, Berk (2010, 2011)). This approach involves controlling the accuracy on the positive and negative classes by tuning the misclassification costs for examples in each class. In what follows, we denote the misclassification cost on examples from the positive and negative classes as W^+ and W^- respectively. As we increase W^+, the cost of making a mistake on a positive example increases, and we expect to obtain a model that classifies the positive examples more accurately (i.e. with higher TPR). We choose W^+ and W^- so $W^+ + W^- = 2$. Thus, when $W^+ = 2$, we obtain a trivial model that predicts $\hat{y}_i = 1$ and attains TPR = 1. When $W^+ = 0$, we obtain a trivial model that predicts $\hat{y}_i = -1$ that attains FPR = 0.

4.1.3. Choice of classification methods

We compared SLIM scoring systems with models produced by eight popular classification methods, including those previously used for recidivism prediction (see Section 1.2) or those that ranked among the 'top 10 algorithms in data mining' (Wu *et al.*, 2008). In choosing these methods, we restricted our attention to methods that have publicly available software packages and allow users to specify misclassification costs for positive and negative classes. Our final choice of methods includes the following methods.

(a) *C5.0 trees and C5.0 rules*: C5.0 is an updated version of the popular C4.5 algorithm (Quinlan, 2014; Kuhn and Johnson, 2013) that can create decision trees and rule sets.

(b) *CART*: the CART method is a popular method to create decision trees through recursive partitioning of the input variables (Breiman *et al.*, 1984).

(c) *L_1- and L_2-penalized logistic regression* are variants of logistic regression that penalize the coefficients to prevent overfitting (Friedman *et al.*, 2010). L_1-penalized methods are typically used to create linear models that are sparse (Tibshirani, 1996; Hesterberg *et al.*, 2008). The L_2-regularized methods are called 'ridge' regression and are not generally sparse.

(d) *Random forests* are a popular black box method that makes predictions by using a large ensemble of weak classification trees. The method was originally developed by Breiman (2001a) but is widely used for prediction of recidivism (see, for example, Berk *et al.* (2009) and Ritter (2013)).

(e) *SVMs* are a popular black box method for non-parametric linear classification. The RBF kernel lets the method handle classification problems where the decision boundary may be non-linear (see, for example, Cristianini and Shawe-Taylor (2000) and Berk and Bleich (2014)).

(f) *SGB* is a popular black box method that creates prediction models in the form of an ensemble of weaker prediction models (Friedman, 2001; Freund and Schapire, 1997).

4.1.4. Details on experimental design, parameter tuning and computation

We summarize the methods, software and settings that we used in our experiments in Table 4.

For each of the six recidivism prediction problems and each of the nine methods, we constructed ROC curves by running the algorithm with 19 values of W^+. The values of W^+ were chosen to produce models across the full ROC curves. By default, we chose values of $W^+ \in \{0.1, 0.2, \ldots, 1.9\}$ and set $W^- = 2 - W^+$. These values of W^+ were inappropriate for problems with a significant class imbalance as all methods produced trivial models. Thus, for significantly imbalanced problems, such as `domestic_violence` and `sexual_violence`, we used values of $W^+ \in \{1.815, 1.820, \ldots, 1.995\}$. For `fatal_violence`, which was extremely imbalanced, we used $W^+ \in \{1.975, 1.976, \ldots, 1.995\}$.

This set-up requires us to produce a total of 1026 recidivism prediction models (six recidivism problems times nine methods times 19 imbalance ratios). Each of the 1026 models were built on a training set and their performance was assessed out of sample. In particular, a third of the data was reserved as the *test set*. The remaining two-thirds of the data were the *training set*. During training, we used fivefold nested cross-validation (CV) for parameter tuning. Explicitly,

Table 4. Methods, software and free parameters used to train models for all six recidivism prediction problems†

Method	*Software*	*Free parameters and settings*
CART decision trees	rpart (Therneau *et al.*, 2012)	minSplit $\in (3, 5, 10, 15, 20) \times$ CP $\in (0.0001, 0.001, 0.01)$
C5.0T decision trees	c50 (Kuhn *et al.*, 2012)	Default settings
C5.0R decision rules	c50 (Kuhn *et al.*, 2012)	Default settings
Logistic regression (lasso) (L_1-penalty)	glmnet (Friedman *et al.*, 2010)	100 values of L_1-penalty chosen by glmnet
Logistic (ridge) regression (L_2-penalty)	glmnet (Friedman *et al.*, 2010)	100 values of L_2-penalty chosen by glmnet
Random forests	randomForest (Liaw and Wiener, 2002)	sampsize $\in (0.632N, 0.4N, 0.2N) \times$ nodesize $\in (1, 5, 10, 20)$ with unbounded tree depth
SVMs (radial basis kernel)	e1071 (Meyer *et al.*, 2012)	$C \in (0.01, 0.1, 1, 10) \times \gamma \in (1/(10P), 1/(5P), 1/(2P), 1/P, 2/P, 5/P, 10/P)$
SGB (Adaboost)	gbm (Ridgeway, 2006)	shrinkage $\in (0.001, 0.01, 0.1) \times$ interaction.depth $\in (1, 2, 3, 4) \times$ ntrees $\in (100, 500, 1500, 3000)$
SLIM scoring systems	CPLEX 12.6 (Ustun, 2016)	C_0 and ϵ set to find the most accurate model with $\leqslant 8$ coefficients where $\lambda_0 \in \{-100, \ldots, 100\}$ and $\lambda_j \in \{-10, \ldots, 10\}$

†We ran each method for 19 values of W^+ and all combinations of free parameters listed in the table. For each value of W^+, we selected the model that minimized the mean weighted fivefold CV error. The values of W^+ are problem specific (see Section 4.1.4 for details).

the training data were split into five folds, and one of those was reserved as the validation fold. The validation fold was rotated to select free parameter values, and a *final model* was trained on the full training set (two-thirds) with the selected parameter values and its performance was assessed on the test set (a third). The folds were generated once to allow for comparisons across methods and prediction problems. The parameters were chosen during nested CV to minimize the mean weighted fivefold CV error on the training set. Having obtained a set of 19 different models for each method and each problem, we then constructed an ROC curve for that method on that problem by plotting the test TPR and test FPR of the 19 final models.

We trained all baseline methods by using publicly available packages in R 3.2.2 (R Core Team, 2015) without imposing any time constraints. In comparison, we trained SLIM by solving IP problems with the CPLEX 12.6 application program interface in MATLAB 2013a. We solved each IP through the following procedure: we trained the solver on the formulation in Section 3.3 for a total of 4 h on a local computing cluster with 2.7-GHz central processor units. Each time we solved an IP problem we kept 500 feasible solutions and polished them by using the formulation in Section 3.4. We then used the same nested CV procedure as the other methods to tune the number of terms in the final model. Polishing all 500 solutions took less than 1 min of computing time. Thus, the total number of optimization problems that we solved was 500 polishing IP problems times (five folds plus one final model) times six problems times 19 values of $W^+ = 342000$ IP problems.

4.2. Observations on predictive accuracy

We show ROC curves for all methods and prediction problems in Fig. 1 and summarize the test AUC of each method in Table 5. Tables with the training and fivefold CV validation AUCs for all methods are included in Appendix A.

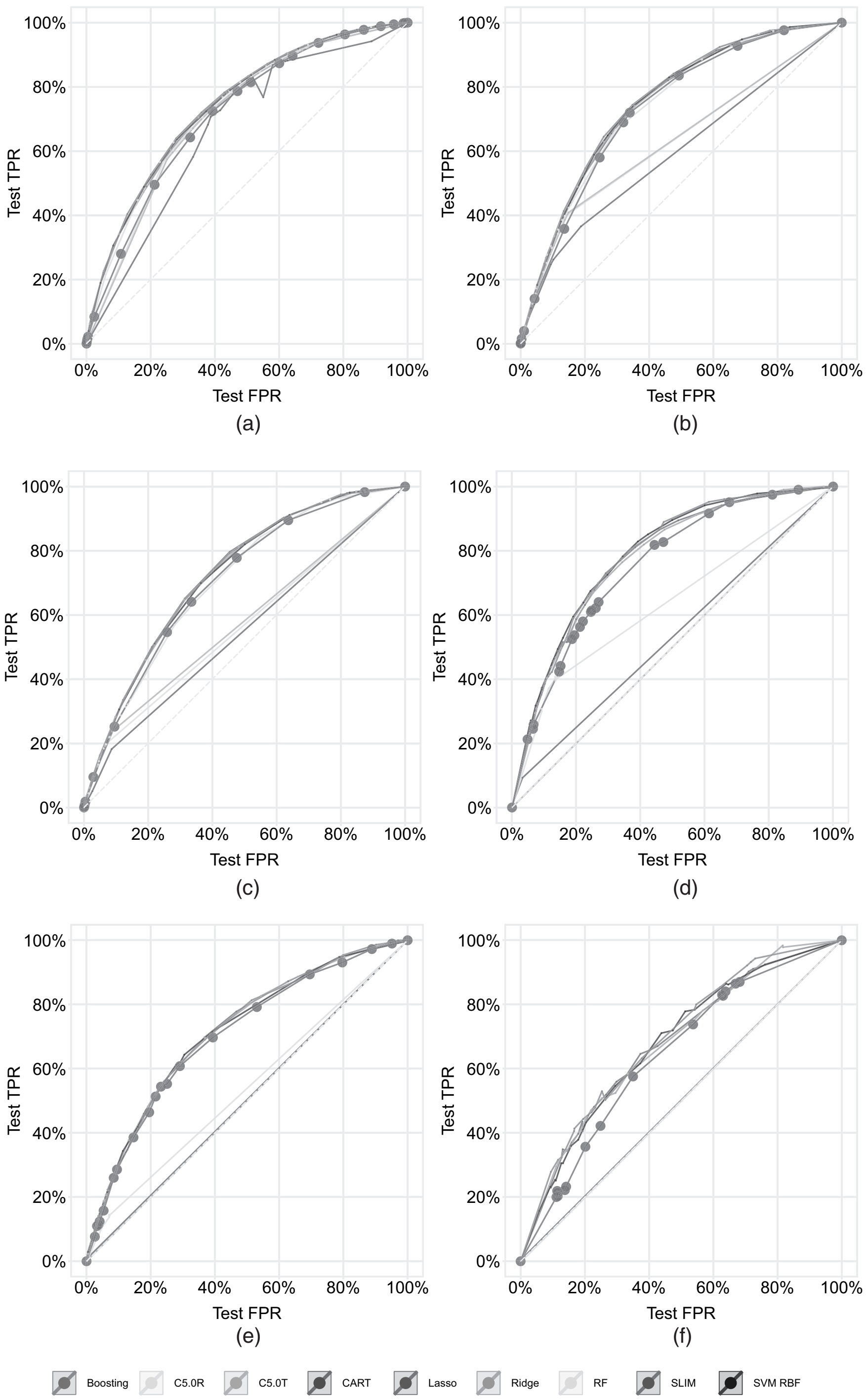


Fig. 1. ROC curves for general recidivism-related prediction problems with test data (all models perform similarly except for the C5.0R, C5.0T and CART models): (a) `arrest` problem; (b) `drug` problem; (c) `general_violence` problem; (d) `domestic_violence` problem; (e) `sexual_violence` problem; (f) `data_violence` problem

Table 5. Test AUC for all methods on all prediction problems†

Prediction problem	*Results for the following methods:*								
	Lasso	*Ridge regression*	*C5.0R*	*C5.0T*	*CART*	*Random forests*	*SVM RBFs*	*SGB*	*SLIM*
arrest	0.72	0.73	0.72	0.72	0.68	0.73	0.72	0.73	0.72
drug	0.74	0.74	0.63	0.63	0.59	0.75	0.73	0.75	0.74
general_violence	0.72	0.72	0.56	0.57	0.56	0.71	0.70	0.72	0.71
domestic_violence	0.77	0.77	0.50	0.50	0.53	0.64	0.77	0.78	0.76
sexual_violence	0.72	0.72	0.50	0.50	0.51	0.54	0.69	0.70	0.70
fatal_violence	0.67	0.68	0.50	0.50	0.50	0.50	0.69	0.70	0.62

†Each cell contains the test AUC.

We make the following important observations, which we believe carry over to a large class of problems beyond prediction of recidivism.

(a) All methods did well on the general recidivism prediction problem arrest. In this case, we observe only small differences in predictive accuracy of different methods: all methods other than CART attain a test AUC above 0.72; the highest test AUC of 0.73 was achieved by SGB, ridge regression and random forests. This multiplicity of good models reflects the *Rashomon effect* of Breiman (2001a).

(b) Major differences between methods appeared in their performance on imbalanced prediction problems. We expected different methods to respond differently to changes in the misclassification costs and therefore trained each method over a large range of possible misclassification costs. Even so, it was difficult (if not impossible) to tune certain methods to produce models at certain points of the ROC curve (see, for example, problems with significant imbalance, such as fatal_violence).

(c) the SVM RBFs, SGB, the lasso and ridge regression could produce accurate models at different points on the ROC curve for most problems. SGB usually achieved the highest AUC on most problems (e.g. arrest, drug, general_violence, domestic_violence and fatal_violence). The lasso, ridge regression and the SVM RBFs often produce comparable AUCs. We find that these methods respond well to cost-sensitive tuning, but it is difficult to tune the misclassification costs for highly imbalanced problems, such as fatal_violence, to obtain models at specific points on the ROC curve.

(d) the C5.0T, C5.0R and CART methods could not produce accurate models at different points on the ROC curve on any imbalanced problems. We found that these methods do not respond well to cost-sensitive tuning. The issue becomes markedly more severe as problems become more imbalanced. For drug and general_violence, for instance, these methods could not produce models with high TPR. For fatal_violence, sexual_violence and domestic_violence, these methods almost always produced trivial models that predict $y = -1$ (resulting in AUCs of 0.5). This result may be attributed to the greedy nature of the algorithms that were used to fit the trees, as opposed to the use of tree models in general. The issue is unlikely to be software related as it affects both C5.0 and CART methods and has been observed by others (see, for example, Goh and Rudin (2014)). This problem might not occur if trees were better optimized.

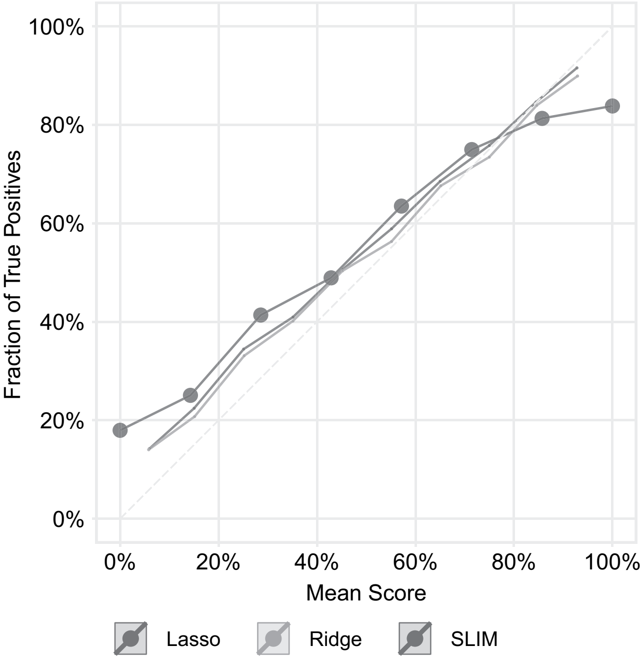


Fig. 2. Risk calibration plot for `arrest` based on test data: we compare three models chosen at a similar decision point, with test FPR $\leqslant$50%; although it is not a risk assessment tool, we see that SLIM is well calibrated

(e) In general, SLIM produced models that are close to or on the efficient frontier of the ROC curve, despite being restricted to a relatively small class of simple linear models (at most eight non-zero coefficients from -10 to 10). Even on highly imbalanced problems such as `domestic_violence` and `sexual_violence`, it responds well to changes in misclassification costs (as expected, by nature of its formulation).

In addition to predictive accuracy, we also examine the risk calibration of the models. Fig. 2 shows the risk calibration for `arrest`, constructed by using the binning method from Zadrozny and Elkan (2002). We include calibration plots for all other problems in an extended version of appendix B of the on-line supplement. We see that SLIM is well calibrated, even though there is no reason why it should be; it is a decision-making tool, not a risk assessment tool. For `arrest`, the lasso and ridge regression are well calibrated; however, they lose this quality once we consider only sparse models (see appendix D of the on-line supplement). This property would also be lost if the lasso and ridge regression coefficients were rounded.

4.3. *Trade-offs between accuracy and interpretability*

In Appendix C, we show that the baseline methods cannot maintain the same level of accuracy as they have in Section 4.2 when their model size was constrained. For the lasso, ridge regression and SLIM, model size is defined as the number of features in the model. For the CART and C5.0 methods, model size is the number of leaves or rules. In fact, we find that the only methods that can consistently produce accurate models along the full ROC curve and also have the potential for interpretability are SLIM and the (non-sparse) lasso.

Tree and rule-based methods such as the CART, C5.0T and C5.0R methods were generally unable to produce models that attain high degrees of accuracy. Worse, even for balanced

Table 6. SLIM scoring system for `arrest`†

Predict arrest for any offence if score > 1

1, *age_at_release_18_to_24*	2 points	⋯
2, *prior_arrests≥5*	2 points	+⋯
3, *prior_arrest_for_misdemeanor*	1 point	+⋯
4, *no_prior_arrests*	−1 point	+⋯
5, *age_at_release≥40*	−1 point	+⋯
Add points from rows 1–5	Score	=⋯

†This model has a test TPR/FPR of 76.6%/44.5%, and a mean fivefold CV TPR/FPR of 78.3%/46.5%.

Table 7. Lasso model for `arrest`, with coefficients rounded to two significant digits†

Predict arrest for any offence if score > 0.31

1, *prior_arrests≥5*	0.63 points	⋯
2, *age_1st_confinement_18_to_24*	0.15 points	+⋯
3, *prior_arrest_for_property*	0.09 points	+⋯
4, *prior_arrest_for_misdemeanor*	0.05 points	+⋯
5, *age_at_release≥40*	−0.20 points	+⋯
Add points from rows 1–5	Score	=⋯

†This model has a test TPR/FPR of 70.9%/43.8%, and a mean fivefold CV TPR/FPR of 72.2%/44.0%.

problems such as `arrest`, where these methods did produce accurate models, the models are complicated and used a very large number of rules or leaves (similar behaviour for C5.0T or C5.0R was also observed by, for instance, Lim *et al.* (2000)). As we show in Appendix C, it was not reasonably possible to obtain a C5.0R, C5.0T or CART model with at most eight rules or eight leaves for almost every prediction problem.

4.4. *On the interpretability of equally accurate transparent models*

To assess the interpretability of different models, we provide a comparison of predictive models produced by the SLIM, lasso and CART methods for the `arrest` problem in Tables 6 and 7 and Fig. 3. This set-up provides a nice basis for comparison as all three methods produce models at roughly the same decision point, and with the same degree of sparsity. For this comparison, we considered any transparent model with at most eight coefficients (the lasso), eight rules (C5.0R) or eight leaves (C5.0T and CART) and had a test FPR of below 50%. We report the models with the minimum weighted test error. Here, neither C5.0R nor C5.0T could produce an acceptable model with at most eight rules or eight leaves, so only models from the SLIM, CART and lasso methods could be displayed. As described before, it is rare for the lasso and CART methods to produce models with a similar degree of accuracy to SLIM when the model size is constrained. We make the following observations.

(a) All three models attain similar levels of predictive accuracy. Test TPR-values ranged between 70% and 79% and test FPR-values ranged between 43% and 48%. There may not be a classification model that can attain substantially higher accuracy.

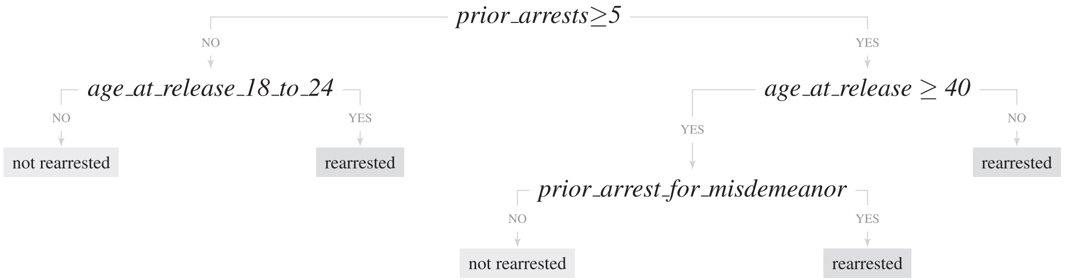


Fig. 3. CART model for `arrest`: this model has a test TPR/FPR of 79.1%/47.9%, and a mean fivefold CV TPR/FPR of 79.9%/48.5%

Table 8. SLIM scoring system for `drug`†

Predict arrest for drug offence if score >7		
1, *prior_arrest_for_drugs*	9 points	$\cdots$
2, *age_at_release_18_to_24*	5 points	$+\cdots$
3, *age_at_release_25_to_29*	3 points	$+\cdots$
4, *prior_arrest_for_multiple_types_of_crime*	2 points	$+\cdots$
5, *prior_arrest_for_property*	1 points	$+\cdots$
6, *age_at_release_30_to_39*	-1 point	$+\cdots$
7, *no_prior_arrests*	-6 points	$+\cdots$
Add points from rows 1–7	Score	$=\cdots$

†This model has a test TPR/FPR of 85.7%/51.1%, and a mean fivefold CV TPR/FPR of 82.3%/49.7%.

(b) The SLIM model uses five input variables and small integer coefficients (see, for example, Table 6). There is a natural rule-based interpretation. In this case, the model implies that, if the prisoner is young (*age_at_release_of_18_to_24*) or has a history of arrests (*prior_arrests*$\geqslant$*5*), he is highly likely to be rearrested. In contrast, if he is relatively older (*age_at_release*$\geqslant$*40*) or has no history of arrests (*no_prior_arrests*), he is unlikely to commit another crime.

(c) The CART model also allows users to make predictions without a calculator. In comparison with the SLIM model, however, the hierarchical structure of the CART model makes it difficult to gauge the relationship of each input variable on the predicted outcome. Consider, for instance, the relationship between age at release and the outcome. In this case, users are immediately aware that there is an effect, as the model branches on the variables *age_at_release*$\geqslant$*40* and *age_at_release_18_to_24*. However, the effect is difficult to comprehend since it depends on prior arrests for misdemeanour: if *prior_arrests*$\geqslant$*5* $=1$ and *age_at_release_18_to_24* $=1$ then the model predicts $\hat{y}=1$; if *prior_arrests*$\geqslant$*5* $=0$ and *age_at_release*$\geqslant$*40* $=0$ then $\hat{y}=1$; however, if *prior_arrests*$\geqslant$*5* $=0$ and *age_at_release*$\geqslant$*40* $=1$ then $\hat{y}=1$ only if *prior_arrest_for_misdemeanor* $=1$. Such issues do not affect linear models such as SLIM and the lasso, where users can immediately gauge the direction and strength of the relationship between an input variable and the predicted outcome by the size and sign of a coefficient. The literature on interpretability in machine learning indicates that interpretability is domain specific; there are some domains where logical models are preferred over linear models, and vice versa (e.g. Freitas (2014)).

Table 9. SLIM scoring system for general_violence†

Predict arrest for general violence offence if score > 7		
1, *prior_arrest_for_general_violence*	8 points	$\cdots$
2, *prior_arrest_for_misdemeanor*	5 points	$+\cdots$
3, *infraction_in_prison*	3 points	$+\cdots$
4, *prior_arrest_for_local_ord*	3 points	$+\cdots$
5, *prior_arrest_for_property*	2 points	$+\cdots$
6, *prior_arrest_for_fatal_violence*	2 points	$+\cdots$
7, *prior_arrest_with_firearms_involved*	1 point	$+\cdots$
8, *age_at_release*$\geqslant$*40*	-7 points	$+\cdots$
Add points from rows 1–8	Score	$=\cdots$

†This model has a test TPR/FPR of 76.7%/45.4%, and a mean fivefold CV TPR/FPR of 76.8%/47.6%.

Table 10. SLIM scoring system for domestic_violence†

Predict arrest for domestic violence offence if score > 3		
1, *prior_arrest_for_misdemeanor*	4 points	$\cdots$
2, *prior_arrest_for_felony*	3 points	$+\cdots$
3, *prior_arrest_for_domestic_violence*	2 points	$+\cdots$
4, *age_1st_confinement_18_to_24*	1 point	$+\cdots$
5, *infraction_in_prison*	-5 points	$+\cdots$
Add points from rows 1–5	Score	$=\cdots$

†This model has a test TPR/FPR of 85.5%/46.0%, and a mean fivefold CV TPR/FPR of 81.4%/48.0%.

Table 11. SLIM scoring system for sexual_violence†

Predict arrest for sexual violence offence if score > 2		
1, *prior_arrest_for_sexual_violence*	3 points	$\cdots$
2, *prior_arrests*$\geqslant$ *5*	1 point	$+\cdots$
3, *multiple_prior_jail_time*	1 point	$+\cdots$
4, *prior_arrest_for_multiple_types_of_crime*	-1 point	$+\cdots$
5, *no_prior_arrests*	-2 points	$+\cdots$
Add points from rows 1–5	Score	$=\cdots$

†This model has a test TPR/FPR of 44.3%/17.7%, and a mean fivefold CV TPR/FPR of 43.7%/19.9%.

4.5. Scoring systems for recidivism prediction

We show a SLIM scoring system for each of the prediction problems that we consider in Tables 8–12. The models are chosen at specific decision points, with the constraint that the fivefold CV FPR$\leqslant$ 50% except for sexual_violence, which is chosen at fivefold CV FPR$\leqslant$ 20%. The models that are presented here may be suitable for screening tasks. To obtain a model that is suitable for sentencing, a point on the ROC curve with a much higher TPR would be needed. We note that these models generalize well from the data set, which is evident by the close match between test TPR/FPR (Table 5) and training TPR/FPR (Table 13 in Appendix A).

Table 12. SLIM scoring system for `fatal_violence`†

Predict arrest for fatal violence offence if score > 4		
1, *age_1st_confinement*⩽*17*	5 points	···
2, *prior_arrest_with_firearms_involved*	3 points	+···
3, *age_1st_confinement_18_to_24*	2 points	+···
4, *prior_arrest_for_felony*	2 points	+···
5, *age_at_release_18_to_24*	1 point	+···
6, *prior_arrest_for_drugs*	1 point	+···
Add points from rows 1–6	Score	=···

†This model has a test TPR/FPR of 55.4%/35.5%, and a mean fivefold CV TPR/FPR of 64.2%/42.4%.

Many of these models exhibit the same 'rule-like' tendencies as discussed in Section 4.4. For example, the model for `drug` in Table 8 predicts that a person will be arrested for a drug-related offence if he or she has ever had any prior drug offences. Similarly, the model for `sexual_violence` in Table 11 effectively states that a person will be rearrested for a sexual offence if and only if he or she has a prior history of sexual crimes. For completeness, we include comparisons with other models in Appendix B.

5. Discussion

Our paper merges two perspectives on modelling recidivism: the first is to obtain accurate predictive models by using the most powerful machine learning tools that are available, and the second is to create models that are easy to use and understand.

We used a set of features that are commonly accessible to police officers and judges, and compared the ability of various machine learning methods to produce models at different decision points across the ROC curve. Our results suggest that it is possible for traditional methods, such as ridge regression, to perform just as well as more modern methods, such as SGB—a finding that is in line with the work of Tollenaar and van der Heijden (2013) and Yang *et al.* (2010). Further, we found that even simple models may perform surprisingly well, even when they are fitting from a heavily constrained space—a finding that is in line with work on the surprising performance of simple models (see, for example, Dawes (1979) and Holte (1993, 2006)).

Our study shows that there may be major advantages of using SLIM for prediction of recidivism, as it can dependably produce a simple scoring system that is accurate and interpretable on any decision point along the ROC curve. Interpretability is crucial for many of the high stakes applications where recidivism prediction models are being used. In such applications, it is not enough for the decision maker to know what input variables are being used to train the model, or how individual input variables are related to the outcome; decision makers should know how the model combines all the input variables to generate its predictions, and whether this mechanism aligns with their ethical values. SLIM not only shows this mechanism but also accommodates constraints that are designed to align the prediction model with the ethical values of the decision maker.

In comparison with current machine learning methods, the main drawback of running SLIM is increased computation involved in solving an IP problem. For this, we proposed two new techniques to reduce computation involved in training high quality SLIM scoring systems:

(a) a polishing procedure that improves the quality of feasible solutions that are found by an IP solver and
(b) an IP formulation that makes it easier for an IP solver to provide a certificate of optimality.

In our experiments, the time that is required to train SLIM was ultimately comparable with the time that is required to train random forests or SGB. However, it was still significant compared with the time that is required for other methods such as the CART, C5.0 methods and penalized logistic regression. In theory, the computation that is required to find an optimal solution to the SLIM integer programme is 'NP hard', meaning that the run time increases exponentially with the number of features. In practice, the run time depends on several factors, such as the number of samples, the number of dimensions, the underlying ease of the classification and how the data are encoded. Since most criminological problems cannot by nature involve massive data sets (since each observation is a person), and since computer speed of solving millions of instructions per second also increases exponentially, it is possible that mathematical programming techniques like SLIM are well suited to criminological problems that are substantially larger and more complex than the problem that is discussed in this work.

Acknowledgement

Jiaming Zeng and Berk Ustun contributed equally to this work.

Appendix A: Additional results on predictive accuracy

To supplement the experimental results in Section 4.2, we include the training and fivefold CV results. Table 13 shows the training AUC performance for all methods on all prediction problems, and Table 14 shows the fivefold CV AUC performance for all methods. A table of test AUCs for all methods on all prediction problems can be found in Table 5.

Appendix B: Model-based comparisons

In Section 4, we included a comparison of transparent models produced for the `arrest` problem. Here, we include a similar comparison for all other recidivism prediction problems.

The models and calibration plots that are shown here correspond to the *best* models that we produced by using the lasso and ridge regression (i.e. the models that were plotted as points in Fig. 1). We omit CART results and C5.0 models are shown because all models that were produced were either trivial or

Table 13. Training AUC for all methods on all regression prediction problems

Prediction problem	*Results for the following methods:*								
	Lasso	*Ridge regression*	*C5.0R*	*C5.0T*	*CART*	*Random forests*	*SVM RBFs*	*SGB*	*SLIM*
`arrest`	0.73	0.73	0.73	0.73	0.81	0.73	0.87	0.75	0.72
`drug`	0.74	0.73	0.65	0.66	0.76	0.73	0.85	0.77	0.73
`general_violence`	0.71	0.71	0.58	0.59	0.77	0.71	0.84	0.74	0.71
`domestic_violence`	0.77	0.77	0.50	0.50	0.75	0.64	0.88	0.81	0.76
`sexual_violence`	0.71	0.71	0.50	0.50	0.84	0.55	0.86	0.77	0.71
`fatal_violence`	0.75	0.74	0.50	0.50	0.50	0.51	0.90	0.84	0.73

Table 14. Fivefold CV AUC for all methods on all prediction problems†

Prediction problem	*Results for the following methods:*								
	Lasso	*Ridge regression*	*C5.0R*	*C5.0T*	*CART*	*Random forests*	*SVM RBF*	*SGB*	*SLIM*
`arrest`	0.72	0.73	0.71	0.71	0.67	0.73	0.71	0.73	0.72
	0.72–0.74	0.72–0.74	0.71–0.73	0.70–0.72	0.66–0.69	0.72–0.74	0.70–0.72	0.72–0.74	0.71–0.73
`drug`	0.73	0.73	0.62	0.62	0.59	0.73	0.72	0.74	0.72
	0.72–0.74	0.71–0.74	0.61–0.64	0.61–0.64	0.58–0.60	0.72–0.74	0.71–0.73	0.72–0.74	0.71–0.73
`general_violence`	0.71	0.71	0.56	0.57	0.56	0.70	0.69	0.71	0.70
	0.70–0.71	0.70–0.71	0.55–0.57	0.55–0.59	0.55–0.58	0.69–0.71	0.69–0.70	0.70–0.71	0.69–0.71
`domestic_violence`	0.76	0.76	0.50	0.50	0.53	0.63	0.76	0.77	0.75
	0.75–0.79	0.75–0.78	0.50–0.50	0.50–0.50	0.51–0.54	0.59–0.66	0.74–0.78	0.75–0.79	0.72–0.78
`sexual_violence`	0.70	0.69	0.50	0.50	0.51	0.54	0.67	0.68	0.68
	0.68–0.74	0.66–0.74	0.50–0.50	0.50–0.50	0.50–0.51	0.53–0.55	0.63–0.70	0.65–0.72	0.66–0.72
`fatal_violence`	0.66	0.67	0.50	0.50	0.50	0.51	0.67	0.67	0.65
	0.59–0.74	0.62–0.75	0.50–0.50	0.50–0.50	0.50–0.52	0.50–0.53	0.63–0.73	0.61–0.74	0.61–0.69

†We report the fivefold CV mean validation AUC. The ranges underneath each cell represent the fivefold CV minimum and maximum.

contained too many leaves to be printed. For any given problem, the models operate at similar decision points (TPR) and are constrained to the same FPR-criteria as in Section 4.5.

The calibration plots will appear to be flat for problems with significant class imbalance. Typically, a well-calibrated classifier on a problem without class imbalance should fall on the $x=y$ line. However, because the y-axis is defined as $P\{y=1|s(x)=s\}$, where s is the predicted score of a model, the slope of the graph will be less than $P(y=1)$ by definition. Therefore, for a highly imbalanced problem such as `fatal_violence`, where $P(y=1)=0.7\%$, the plot will be flat.

B.1. drug

This is the SLIM model for `drug`. This model has a test TPR/FPR of 85.7%/51.1%, and a mean fivefold CV TPR/FPR of 82.3%/49.7%:

9.00 *prior_arrest_for_drugs*	+ 5.00 *age_at_release_18_to_24*	+ 4.00 *age_at_release_25_to_29*
+ 3.00 *prior_arrest_for_multiple_types_of_crime*	+ 1.00 *prior_arrest_for_property*	− 6.00 *no_prior_arrests*
− 1.00 *age_at_release_30_to_39*	− 7.00	

This is the best lasso model for `drug`. This model has a test TPR/FPR of 82.0%/45.9%, and a mean fivefold CV TPR/FPR of 81.2 %/45.9%:

1.14 *prior_arrest_for_drugs*	+ 0.18 *prior_arrest_for_multiple_types_of_crime*	+ 0.17 *prior_arrest_for_misdemeanor*
+ 0.19 *prior_arrest_for_other_violence*		
+ 0.16 *age_at_release_18_to_24*	+ 0.14 *prior_arrests⩾5*	+ 0.13 *age_1st_confinement_18_to_24*
+ 0.12 *prior_arrest_for_public_order*	+ 0.10 *prior_arrest_with_firearms_involved*	
+ 0.06 *age_1st_arrest⩽17*	+ 0.04 *multiple_prior_jail_time*	+ 0.08 *any_prior_jail_time*
+ 0.03 *multiple_prior_prison_time*	+ 0.03 *any_prior_prb_or_fine*	+ 0.04 *drug_abuse*
− 0.25 *prior_arrest_for_sexual*	− 0.23 *age_at_release_30_to_39*	− 0.62 *age_at_release⩾40*
− 0.11 *prior_arrest_with_child_involved*	− 0.08 *alcohol_abuse*	− 0.12 *time_served_25_to_60mo*
− 1.11 × 10^{-03} *time_served⩾61mo*	− 1.01	− 0.07 *age_1st_confinement⩾40*
+ 0.27 *prior_arrest_for_property*	+ 0.26 *time_served⩽6mo*	

This is the best ridge regression model for `drug`. This model has a test TPR/FPR of 84.0%/48.2%, and a mean fivefold CV TPR/FPR of 83.1%/48.4%:

0.91 *prior_arrest_for_drugs*	+ 0.25 *time_served⩽6mo*	+ 0.17 *prior_arrest_for_misdemeanor*
+ 0.21 *prior_arrest_for_multiple_types_of_crime*	+ 0.20 *prior_arrest_for_property*	
	+ 0.17 *age_1st_confinement_18_to_24*	+ 0.14 *prior_arrests⩾5*
+ 0.17 *prior_arrest_for_other_violence*	+ 0.12 *age_at_release_25_to_29*	+ 0.11 *drug_abuse*
+ 0.13 *prior_arrest_with_firearms_involved*	+ 0.09 *age_1st_arrest⩽17*	+ 0.08 *age_1st_confinement⩽17*
+ 0.11 *prior_arrest_for_public_order*	+ 0.07 *multiple_prior_jail_time*	+ 0.07 *age_at_release⩽17*
+ 0.08 *any_prior_jail_time*	+ 0.06 *released_unconditonal*	+ 0.05 *any_prior_prb_or_fine*
+ 0.06 *multiple_prior_prison_time*	+ 0.04 *time_served_7_to_12mo*	+ 0.04 *multiple_prior_prb_or_fine*
+ 0.05 *prior_arrests⩾2*	+ 0.01 *age_1st_confinement_25_to_29*	+ 0.01 *released_conditonal*
+ 0.02 *prior_arrests⩾1*	+ 1.76 × 10^{-03} *age_1st_arrest_18_to_24*	+ 9.58 × 10^{-04} *prior_arrest_for_fatal_violence*
+ 2.52 × 10^{-03} *prior_arrest_for_felony*	− 0.25 *prior_arrest_for_sexual*	
− 0.33 *age_at_release⩾40*	− 0.15 *time_served_25_to_60mo*	− 0.19 *age_1st_confinement⩾40*
− 0.16 *prior_arrest_with_child_involved*	− 0.10 *prior_arrest_for_domestic_violence*	− 0.14 *alcohol_abuse*
− 0.13 *time_served⩾61mo*		− 0.09 *age_at_release_30_to_39*
− 0.05 *age_1st_arrest⩾40*	− 0.04 *female*	− 0.04 *infraction_in_prison*
− 0.03 *age_1st_arrest_30_to_39*	− 0.02 *age_1st_confinement_30_to_39*	− 0.02 *no_prior_arrests*
− 4.71 × 10^{-03} *prior_arrest_for_local_ord*	− 4.45 × 10^{-03} *time_served_13_to_24mo*	− 2.23 × 10^{-03} *age_1st_arrest_25_to_29*
− 1.09	+ 0.24 *age_at_release_18_to_24*	

B.2. general_violence

This is the SLIM model for `general_violence`. This model has a test TPR/FPR of 76.7%/45.4%, and a mean fivefold CV TPR/FPR of 76.8%/47.6%:

8 *prior_arrest_for_other_violence*	+ 5 *prior_arrest_for_misdemeanor*	+ 3 *infraction_in_prison*
+ 3 *prior_arrest_for_local_ord*	+ 2 *prior_arrest_for_property*	+ 2 *prior_arrest_for_fatal_violence*
+ *prior_arrest_with_firearms_involved*	− 7 *age_at_release⩾40*	− 7

This is the best lasso model for `general_violence`. This model has a test TPR/FPR of 79.7%/45.5%, and a mean fivefold CV TPR/FPR of 77.3%/45.7%:

$$\begin{array}{llllll}
 & 0.90\ \textit{prior\_arrest\_for\_other\_violence} & + & 0.35\ \textit{prior\_arrest\_for\_property} & + & 0.28\ \textit{prior\_arrest\_for\_misdemeanor} \\
+ & 0.28\ \textit{age\_at\_release\_18\_to\_24} & + & 0.24\ \textit{prior\_arrest\_for\_public\_order} & + & 0.20\ \textit{age\_1st\_arrest}\leqslant\textit{17} \\
+ & 0.20\ \textit{released\_unconditonal} & + & 0.17\ \textit{age\_1st\_confinement\_18\_to\_24} & + & 0.16\ \textit{alcohol\_abuse} \\
+ & 0.14\ \textit{prior\_arrest\_for\_fatal\_violence} & + & 0.14\ \textit{age\_1st\_confinement}\leqslant\textit{17} & + & 0.10\ \textit{prior\_arrest\_for\_felony} \\
+ & 0.10\ \textit{prior\_arrests}\geqslant\textit{5} & + & 0.10\ \textit{prior\_arrest\_with\_firearms} & + & 0.10\ \textit{age\_1st\_arrest\_18\_to\_24} \\
+ & 0.09\ \textit{infraction\_in\_prison} & & \quad\textit{\_involved} & + & 0.03\ \textit{time\_served\_7\_to\_12mo} \\
+ & 2.89\times10^{-03}\ \textit{prior\_arrest\_for\_drugs} & + & 0.04\ \textit{time\_served}\leqslant\textit{6mo} & - & 0.41\ \textit{female} \\
- & 0.27\ \textit{age\_at\_release\_30\_to\_39} & - & 0.72\ \textit{age\_at\_release}\geqslant\textit{40} & - & 0.07\ \textit{age\_1st\_confinement}\geqslant\textit{40} \\
- & 0.05\ \textit{age\_1st\_arrest}\geqslant\textit{40} & - & 0.15\ \textit{prior\_arrest\_with\_child\_involved} & - & 1.84\times10^{-03}\ \textit{age\_1st\_confinement} \\
- & 1.19 & - & 0.01\ \textit{time\_served\_25\_to\_60mo} & & \quad\textit{\_30\_to\_39}
\end{array}$$

This is the best ridge regression model for `general_violence`. This model has a test TPR/FPR of 81.4%/48.1%, and a mean fivefold CV TPR/FPR of 80.0%/48.5%:

$$\begin{array}{llllll}
 & 0.62\ \textit{prior\_arrest\_for\_other\_violence} & + & 0.27\ \textit{age\_at\_release\_18\_to\_24} & + & 0.24\ \textit{prior\_arrest\_for\_property} \\
+ & 0.23\ \textit{prior\_arrest\_for\_misdemeanor} & + & 0.19\ \textit{age\_1st\_confinement\_18\_to\_24} & + & 0.18\ \textit{prior\_arrest\_for\_public\_order} \\
+ & 0.17\ \textit{age\_1st\_arrest}\leqslant\textit{17} & + & 0.14\ \textit{prior\_arrest\_for\_multiple\_types} & + & 0.13\ \textit{released\_unconditonal} \\
+ & 0.13\ \textit{prior\_arrests}\geqslant\textit{5} & & \quad\textit{\_of\_crime} & + & 0.12\ \textit{prior\_arrest\_with\_firearms} \\
+ & 0.11\ \textit{age\_1st\_confinement}\leqslant\textit{17} & + & 0.13\ \textit{prior\_arrest\_for\_felony} & & \quad\textit{\_involved} \\
+ & 0.10\ \textit{prior\_arrest\_for\_fatal\_violence} & + & 0.11\ \textit{alcohol\_abuse} & + & 0.10\ \textit{age\_at\_release\_25\_to\_29} \\
+ & 0.07\ \textit{prior\_arrest\_for\_domestic\_violence} & + & 0.09\ \textit{infraction\_in\_prison} & + & 0.08\ \textit{age\_1st\_arrest\_18\_to\_24} \\
+ & 0.05\ \textit{prior\_arrest\_for\_local\_ord} & + & 0.05\ \textit{drug\_abuse} & + & 0.05\ \textit{time\_served}\leqslant\textit{6mo} \\
+ & 0.03\ \textit{prior\_arrests}\geqslant\textit{2} & + & 0.04\ \textit{time\_served\_7\_to\_12mo} & + & 0.04\ \textit{age\_at\_release}\leqslant\textit{17} \\
+ & 0.01\ \textit{prior\_arrest\_for\_drugs} & + & 0.03\ \textit{multiple\_prior\_prb\_or\_fine} & + & 0.02\ \textit{multiple\_prior\_jail\_time} \\
- & 0.20\ \textit{female} & + & 3.41\times10^{-03}\ \textit{no\_prior\_arrests} & - & 0.32\ \textit{age\_at\_release}\geqslant\textit{40} \\
- & 0.12\ \textit{age\_1st\_arrest}\geqslant\textit{40} & - & 0.18\ \textit{age\_1st\_confinement}\geqslant\textit{40} & - & 0.12\ \textit{prior\_arrest\_with\_child\_involved} \\
- & 0.08\ \textit{age\_at\_release\_30\_to\_39} & - & 0.11\ \textit{age\_1st\_arrest\_30\_to\_39} & - & 0.09\ \textit{age\_1st\_confinement\_30\_to\_39} \\
- & 0.04\ \textit{time\_served\_25\_to\_60mo} & - & 0.05\ \textit{age\_1st\_arrest\_25\_to\_29} & - & 0.04\ \textit{prior\_arrest\_for\_sexual} \\
- & 0.03\ \textit{age\_1st\_confinement\_25\_to\_29} & - & 0.03\ \textit{time\_served}\geqslant\textit{61mo} & - & 0.03\ \textit{released\_conditonal} \\
- & 5.89\times10^{-03}\ \textit{multiple\_prior\_prison\_time} & - & 0.02\ \textit{any\_prior\_prb\_or\_fine} & - & 0.02\ \textit{time\_served\_13\_to\_24mo} \\
- & 1.13 & - & 3.60\times10^{-03}\ \textit{any\_prior\_jail\_time} & - & 3.47\times10^{-03}\ \textit{prior\_arrests}\geqslant\textit{1}
\end{array}$$

B.3. domestic_violence

This is the SLIM model for `domestic_violence`. This model has a test TPR/FPR of 85.5%/46.0%, and a mean fivefold CV TPR/FPR of 81.4%/48.0%:

$$\begin{array}{llllll}
 & 4\ \textit{prior\_arrest\_for\_misdemeanor} & + & 3\ \textit{prior\_arrest\_for\_felony} & + & 2\ \textit{prior\_arrest\_for\_domestic\_violence} \\
+ & \textit{age\_1st\_confinement\_18\_to\_24} & - & 5\ \textit{infraction\_in\_prison} & - & 3
\end{array}$$

This is the best lasso model for `domestic_violence`. This model has a test TPR/FPR of 87.0%/45.8%, and a mean fivefold CV TPR/FPR of 84.5%/45.8%:

$$\begin{array}{llllll}
 & 0.88\ \textit{prior\_arrest\_for\_misdemeanor} & + & 0.73\ \textit{prior\_arrest\_for\_domestic\_violence} & + & 0.73\ \textit{prior\_arrest\_for\_felony} \\
+ & 0.66\ \textit{prior\_arrest\_for\_other\_violence} & + & 0.54\ \textit{released\_unconditonal} & + & 0.32\ \textit{age\_1st\_confinement\_18\_to\_24} \\
+ & 0.24\ \textit{multiple\_prior\_prb\_or\_fine} & + & 0.21\ \textit{alcohol\_abuse} & + & 0.17\ \textit{prior\_arrest\_for\_sexual} \\
+ & 0.16\ \textit{prior\_arrests}\geqslant\textit{5} & + & 0.16\ \textit{prior\_arrest\_with\_firearms\_involved} & + & 0.08\ \textit{age\_at\_release\_18\_to\_24} \\
+ & 0.06\ \textit{no\_prior\_arrests} & + & 0.05\ \textit{time\_served\_7\_to\_12mo} & + & 0.03\ \textit{prior\_arrest\_for\_property} \\
+ & 0.01\ \textit{age\_1st\_arrest\_18\_to\_24} & + & 0.01\ \textit{prior\_arrest\_for\_public\_order} & - & 1.09\ \textit{infraction\_in\_prison} \\
- & 0.54\ \textit{age\_at\_release}\geqslant\textit{40} & - & 0.47\ \textit{drug\_abuse} & - & 0.40\ \textit{multiple\_prior\_prison\_time} \\
- & 0.31\ \textit{prior\_arrest\_with\_child\_involved} & - & 0.28\ \textit{multiple\_prior\_jail\_time} & - & 0.26\ \textit{female} \\
- & 0.20\ \textit{age\_1st\_confinement}\geqslant\textit{40} & - & 0.16\ \textit{any\_prior\_jail\_time} & - & 0.07\ \textit{age\_1st\_arrest\_30\_to\_39} \\
- & 0.07\ \textit{any\_prior\_prb\_or\_fine} & - & 0.06\ \textit{prior\_arrest\_for\_drugs} & - & 0.06\ \textit{time\_served}\geqslant\textit{61mo} \\
- & 4.48\times10^{-04}\ \textit{time\_served\_25\_to\_60mo} & - & 1.04 & &
\end{array}$$

This is the best ridge regression model for `domestic_violence`. This model has a test TPR/FPR of 87.0%/47.7%, and a mean fivefold CV TPR/FPR of 85.2%/47.5%:

0.76 prior_arrest_for_misdemeanor
+ 0.54 prior_arrest_for_felony
+ 0.27 multiple_prior_prb_or_fine
+ 0.18 alcohol_abuse
+ 0.15 prior_arrest_for_local_ord
+ 0.10 prior_arrest_for_property
+ 0.08 age_at_release_30_to_39
+ 0.07 age_1st_arrest_18_to_24
+ 0.05 time_served⩽6mo
+ 3.08×10^{-03} age_1st_confinement_30_to_39
− 0.39 multiple_prior_prison_time
− 0.25 multiple_prior_jail_time
− 0.19 any_prior_jail_time
− 0.10 any_prior_prb_or_fine
− 0.08 prior_arrest_for_drugs
− 0.04 released_conditonal
− 1.01
+ 0.59 prior_arrest_for_other_violence
+ 0.40 released_unconditonal
+ 0.21 prior_arrest_for_sexual
+ 0.18 prior_arrests⩾5
+ 0.12 age_at_release_25_to_29
+ 0.10 prior_arrest_for_fatal_violence
+ 0.07 prior_arrest_for_multiple_types_of_crime
+ 0.07 prior_arrest_for_public_order
+ 0.05 time_served_13_to_24mo
− 0.86 infraction_in_prison
− 0.36 age_at_release⩾40
− 0.25 female
− 0.14 time_served⩾61mo
− 0.10 age_1st_arrest⩾40
− 0.06 age_1st_confinement_25_to_29
− 0.04 age_at_release⩽17
+ 0.57 prior_arrest_for_domestic_violence
+ 0.27 age_1st_confinement_18_to_24
+ 0.19 prior_arrest_with_firearms_involved
+ 0.17 age_at_release_18_to_24
+ 0.11 time_served_7_to_12mo
+ 0.10 no_prior_arrests
+ 0.07 age_1st_arrest⩽17
+ 0.05 age_1st_arrest_25_to_29
+ 0.05 prior_arrests⩾2
+ 0.40 drug_abuse
− 0.26 prior_arrest_with_child_involved
− 0.24 age_1st_confinement⩾40
− 0.12 age_1st_arrest_30_to_39
− 0.10 prior_arrests⩾1
− 0.05 time_served_25_to_60mo
− 0.02 age_1st_confinement⩽17

B.4. sexual_violence

This is the SLIM model for `sexual_violence`. This model has a test TPR/FPR of 44.3%/17.7%, and a mean fivefold CV TPR/FPR of 43.7%/19.9%:

3 prior_arrest_for_sexual
− 2 no_prior_arrests
+ prior_arrests⩾5
− prior_arrest_for_multiple_types_of_crime
+ multiple_prior_jail_time
− 2

This is the best lasso model for `sexual_violence`. This model has a test TPR/FPR of 46.9%/18.1%, and a mean fivefold CV TPR/FPR of 43.7%/17.9%:

1.10 prior_arrest_for_sexual
+ 0.27 prior_arrest_for_felony
+ 0.12 prior_arrest_for_property
+ 0.03 age_1st_confinement⩽17
− 0.58 female
− 0.05 any_prior_prb_or_fine
− 0.01 prior_arrest_for_misdemeanor
+ 0.40 prior_arrest_for_other_violence
+ 0.19 prior_arrest_with_child_involved
+ 0.09 prior_arrest_for_public_order
+ 0.02 age_1st_arrest⩽17
− 0.25 age_at_release⩾40
− 0.05 drug_abuse
− 5.85×10^{-03} age_1st_confinement_30_to_39
+ 0.27 age_1st_confinement_18_to_24
+ 0.19 infraction_in_prison
+ 0.07 prior_arrests⩾5
+ 8.11×10^{-04} prior_arrest_for_fatal_violence
− 0.23 prior_arrest_for_drugs
− 0.01 time_served_25_to_60mo
− 1.63

This is the best ridge regression model for `sexual_violence`. This model has a test TPR/FPR of 48.6%/19.3%, and a mean fivefold CV validation TPR/FPR of 44.9%/19.4%:

0.92 prior_arrest_for_sexual
+ 0.28 prior_arrest_with_child_involved
+ 0.14 prior_arrest_for_property
+ 0.12 prior_arrests⩾5
+ 0.07 time_served⩾61mo
+ 0.06 any_prior_jail_time
+ 0.04 multiple_prior_prb_or_fine
+ 0.03 released_unconditonal
+ 7.60×10^{-03} prior_arrests⩾1
− 0.25 prior_arrest_for_drugs
− 0.11 any_prior_prb_or_fine
− 0.09 age_1st_arrest⩾40
− 0.05 time_served_25_to_60mo
− 0.04 time_served_7_to_12mo
− 0.02 time_served⩽6mo
− 7.46×10^{-03} no_prior_arrests
− 1.47
+ 0.35 prior_arrest_for_other_violence
+ 0.20 age_1st_confinement_18_to_24
+ 0.14 prior_arrest_for_public_order
+ 0.10 prior_arrest_for_fatal_violence
+ 0.07 age_1st_arrest⩽17
+ 0.05 age_at_release_30_to_39
+ 0.03 time_served_13_to_24mo
+ 0.02 age_1st_arrest_18_to_24
+ 6.27×10^{-03} age_at_release⩽17
− 0.16 age_at_release⩾40
− 0.11 age_1st_confinement_30_to_39
− 0.07 prior_arrest_for_misdemeanor
− 0.04 prior_arrests⩾2
− 0.03 prior_arrest_for_multiple_types_of_crime
− 0.02 age_1st_confinement_25_to_29
− 5.79×10^{-03} age_1st_arrest_25_to_29
+ 0.30 prior_arrest_for_felony
+ 0.18 infraction_in_prison
+ 0.13 age_1st_confinement⩽17
+ 0.07 age_at_release_18_to_24
+ 0.07 prior_arrest_for_local_ord
+ 0.04 age_at_release_25_to_29
+ 0.03 released_conditonal
+ 9.63×10^{-03} age_1st_arrest_30_to_39
− 0.37 female
− 0.11 age_1st_confinement⩾40
− 0.09 drug_abuse
− 0.06 multiple_prior_jail_time
− 0.04 alcohol_abuse
− 0.02 prior_arrest_for_domestic_violence
− 0.02 multiple_prior_prison_time
− 4.60×10^{-03} prior_arrest_with_firearms_involved

B.5. fatal_violence

This is the SLIM model for `fatal_violence`. This model has a test TPR/FPR of 55.4%/35.5%, and a mean fivefold CV TPR/FPR of 64.2%/42.4%:

5 *age_1st_confinement≤17*
+ 2 *prior_arrest_for_felony*
− 4
+ 3 *prior_arrest_with_firearms_involved*
+ *age_at_release_18_to_24*
+ 2 *age_1st_confinement_18_to_24*
+ *prior_arrest_for_drugs*

This is the best lasso model for `fatal_violence`. This model has a test TPR/FPR of 68.9%/44.5%, and a mean fivefold CV TPR/FPR of 67.6%/42.4%:

1.52 *age_1st_confinement≤17*
+ 0.73 *age_at_release_18_to_24*
+ 0.60 *prior_arrest_for_fatal_violence*
+ 0.39 *prior_arrest_for_drugs*
+ 0.35 *age_1st_arrest≤17*
+ 0.28 *no_prior_arrests*
+ 0.20 *multiple_prior_prison_time*
+ 0.11 *any_prior_prb_or_fine*
+ 0.04 *age_1st_arrest_18_to_24*
− 0.70 *drug_abuse*
− 0.42 *released_conditonal*
− 0.34 *prior_arrest_for_misdemeanor*
− 0.24 *multiple_prior_jail_time*
− 0.08 *age_at_release_30_to_39*
− 2.00
+ 1.47 *age_at_release≤17*
+ 0.69 *alcohol_abuse*
+ 0.54 *age_1st_confinement_18_to_24*
+ 0.38 *age_1st_confinement_25_to_29*
+ 0.34 *prior_arrest_for_public_order*
+ 0.26 *age_1st_arrest_25_to_29*
+ 0.19 *prior_arrest_for_property*
+ 0.07 *time_served_7_to_12mo*
− 2.69 *age_1st_arrest≥40*
− 0.55 *infraction_in_prison*
− 0.39 *prior_arrests≥2*
− 0.33 *prior_arrest_with_child_involved*
− 0.16 *released_unconditonal*
− 0.08 *prior_arrest_for_domestic_violence*
+ 1.12 *prior_arrest_for_felony*
+ 0.66 *prior_arrests≥5*
+ 0.47 *prior_arrest_with_firearms_involved*
+ 0.35 *prior_arrest_for_other_violence*
+ 0.31 *prior_arrest_for_multiple_types_of_crime*
+ 0.24 *age_1st_confinement_30_to_39*
+ 0.18 *prior_arrest_for_sexual*
+ 0.07 *time_served≤6mo*
− 1.68 *female*
− 0.50 *time_served≥61mo*
− 0.36 *age_at_release≥40*
− 0.29 *multiple_prior_prb_or_fine*
− 0.13 *time_served_13_to_24mo*
− 0.02 *prior_arrests≥1*

This is the best ridge regression model for `fatal_violence`. This model has a test TPR/FPR of 62.2%/34.0%, and a mean fivefold CV TPR/FPR of 60.1%/33.0%:

0.55 *prior_arrest_for_felony*
+ 0.39 *age_1st_arrest≤17*
+ 0.35 *prior_arrest_with_firearms_involved*
+ 0.26 *prior_arrest_for_public_order*
+ 0.19 *age_at_release≤17*
+ 0.15 *time_served_7_to_12mo*
+ 0.10 *any_prior_prb_or_fine*
+ 0.06 *no_prior_arrests*
+ 0.03 *prior_arrest_for_local_ord*
− 0.35 *drug_abuse*
− 0.28 *age_1st_confinement≥40*
− 0.19 *multiple_prior_jail_time*
− 0.16 *age_at_release_30_to_39*
− 0.14 *age_1st_confinement_30_to_39*
− 0.06 *age_at_release_25_to_29*
− 0.01 *prior_arrest_for_domestic_violence*
− 1.33
+ 0.54 *age_1st_confinement≤17*
+ 0.39 *prior_arrest_for_fatal_violence*
+ 0.29 *prior_arrest_for_other_violence*
+ 0.25 *alcohol_abuse*
+ 0.16 *multiple_prior_prison_time*
+ 0.14 *time_served≤6mo*
+ 0.08 *prior_arrest_for_sexual*
+ 0.06 *time_served_25_to_60mo*
− 0.51 *female*
− 0.30 *infraction_in_prison*
− 0.25 *time_served≥61mo*
− 0.17 *prior_arrest_with_child_involved*
− 0.15 *released_conditonal*
− 0.12 *age_1st_arrest_30_to_39*
− 0.06 *age_1st_confinement_25_to_29*
− 0.01 *any_prior_jail_time*
+ 0.45 *age_at_release_18_to_24*
+ 0.35 *prior_arrests≥5*
+ 0.29 *prior_arrest_for_drugs*
+ 0.24 *prior_arrest_for_multiple_types_of_crime*
+ 0.16 *prior_arrest_for_property*
+ 0.12 *age_1st_confinement_18_to_24*
+ 0.06 *released_unconditonal*
+ 0.05 *age_1st_arrest_25_to_29*
− 0.42 *age_at_release≥40*
− 0.29 *age_1st_arrest≥40*
− 0.20 *multiple_prior_prb_or_fine*
− 0.16 *prior_arrest_for_misdemeanor*
− 0.14 *prior_arrests≥2*
− 0.07 *time_served_13_to_24mo*
− 0.06 *prior_arrests≥1*
− 8.27×10^{-03} *age_1st_arrest_18_to_24*

Appendix C: Additional results on the trade-off between accuracy and interpretability

In the experiments in Section 4, we used SLIM to fit models from a highly constrained space (i.e. models with at most eight non-zero integer coefficients between −10 and 10). Here, we present evidence to show that baseline methods cannot attain the same level of accuracy or risk calibration when they are used to fit models from a slightly less constrained model space (i.e. a model with at most eight non-zero coefficients, eight leaves or eight rules).

Table 15 shows the test AUC of each method when they are used to fit a model with a model size of 8 or less. Trivial models of size 1 have also been omitted. Table 16 shows the percentage change in test AUC for the methods due to the model size restriction. For all models other than SLIM, the predictive accuracy was compromised with the size constraint. We see that C5.0R and C5.0T cannot produce a suitably sparse

Table 15. Test AUC on all prediction problems when transparent methods are restricted to models with at most eight coefficients, eight leaves or eight rules

Prediction problem	*Results for the following methods:*				
	Lasso	*C5.0R*	*C5.0T*	*CART*	*SLIM*
arrest	0.70	—	—	0.66	0.72
drug	0.71	—	—	0.50	0.74
general_violence	0.70	0.50	0.50	0.50	0.71
domestic_violence	0.74	—	—	0.50	0.76
sexual_violence	0.70	—	—	0.50	0.70
fatal_violence	0.60	—	—	0.50	0.62

Table 16. Percentage in test AUC with respect to SLIM's model on all prediction problems when transparent methods are restricted to models with at most eight coefficients, eight leaves or eight rules

Prediction problem	*Results (%) for the following methods:*				
	Lasso	*C5.0R*	*C5.0T*	*CART*	*SLIM*
arrest	−3.8	—	—	−2.8	0.0
drug	−4.0	—	—	−15.7	0.0
general_violence	−2.2	−11.0	−12.7	−10.3	0.0
domestic_violence	−4.1	—	—	−5.4	0.0
sexual_violence	−2.2	—	—	−1.8	0.0
fatal_violence	−11.2	—	—	0.0	0.0

model for some of the problems since their implementation does not provide control over model sparsity. Note that we have omitted results for ridge regression because it could not produce a model with fewer than eight coefficients for all prediction problems (see Section 4.4 for explanation).

References

Andrade, J. T. (2009) *Handbook of Violence Risk Assessment and Treatment: New Approaches for Mental Health Professionals*. New York: Springer.

Andrews, D. A. and Bonta, J. (2000) *The Level of Service Inventory—Revised*. Toronto: Multi-Health Systems.

Baradaran, S. (2013) Race, prediction, and discretion. *G. Wash. Law Rev.*, **81**, 157–222.

Barnes, G. C. and Hyatt, J. M. (2012) Classifying adult probationers by forecasting future offending. *Technical Report*. National Institute of Justice, US Department of Justice, Washington DC.

Belfrage, H., Fransson, R. and Strand, S. (2000) Prediction of violence using the hcr-20: a prospective study in two maximum-security correctional institutions. *J. Forens. Psychiatr.*, **11**, 167–175.

Berk, R. (2010) Balancing the costs of forecasting errors in parole decisions. *Alb. Law Rev.*, **74**, 1071–1085.

Berk, R. (2011) Asymmetric loss functions for forecasting in criminal justice settings. *J. Quant. Crim.*, **27**, 107–123.

Berk, R. A. and Bleich, J. (2013) Statistical procedures for forecasting criminal behavior. *Crim. Publ. Poly*, **12**, 513–544.

Berk, R. and Bleich, J. (2014) Forecasts of violence to inform sentencing decisions. *J. Quant. Crim.*, **30**, 79–96.

Berk, R. A., He, Y. and Sorenson, S. B. (2005) Developing a practical forecasting screener for domestic violence incidents. *Evaln Rev.*, **29**, 358–383.

Berk, R. A., Kriegler, B. and Baek, J.-H. (2006) Forecasting dangerous inmate misconduct: an application of ensemble statistical procedures. *J. Quant. Crim.*, **22**, 131–145.

Berk, R., Sherman, L., Barnes, G., Kurtz, E. and Ahlman, L. (2009) Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *J. R. Statist. Soc.* A, **172**, 191–211.

Berk, R. A. and Sorenson, S. D. (2014) Machine learning forecasts of domestic violence to help inform release decisions at arraignment. *Technical Report*. University of Pennsylvania, Philadelphia.

Bhati, A. S. (2007) Estimating the number of crimes averted by incapacitation: an information theoretic approach. *J. Quant. Crim.*, **23**, 355–375.

Bhati, A. S. and Piquero, A. R. (2007) Estimating the impact of incarceration on subsequent offending trajectories: deterrent, criminogenic, or null effect? *J. Crimnl Law Crim.*, 207–253.

Borden, H. G. (1928) Factors for predicting parole success. *J. Am. Inst. Crimnl Law Crim.*, 328–336.

Borum, R. (2006) *Manual for the Structured Assessment of Violence Risk in Youth (SAVRY)*. Odessa: Psychological Assessment Resources.

Breiman, L. (2001a) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman, L. (2001b) Statistical modeling: the two cultures. *Statist. Sci.*, **16**, 199–231.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984) *Classification and Regression Trees*. Boca Raton: CRC Press.

Burgess, E. W. (1928) Factors determining success or failure on parole. Illinois Committee on Indeterminate-Sentence Law and Parole, Springfield.

Bushway, S. D. (2013) Is there any logic to using logit. *Crim. Publ. Poly*, **12**, 563–567.

Bushway, S. D. and Piehl, A. M. (2007) The inextricable link between age and criminal history in sentencing. *Crime Delinq.*, **53**, 156–183.

Clements, C. B. (1996) Offender classification: two decades of progress. *Crimnl Just. Behav.*, **23**, 121–143.

Copas, J. and Marshall, P. (1998) The offender group reconviction scale: a statistical reconviction score for use by probation officers. *Appl. Statist.*, **47**, 159–171.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

Crow, M. S. (2008) The complexities of prior record, race, ethnicity, and policy: Interactive effects in sentencing. *Crimnl Just. Rev.*, **33**, 502–523.

Dawes, R. M. (1979) The robust beauty of improper linear models in decision making. *Am. Psychol.*, **34**, 571–582.

Dawes, R. M., Faust, D. and Meehl, P. E. (1989) Clinical versus actuarial judgment. *Science*, **243**, 1668–1674.

Freitas, A. A. (2014) Comprehensible classification models: a position paper. *Explorns Newslett.*, **15**, 1–10.

Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, **29**, 1189–1232.

Friedman, J. H. (2002) Stochastic gradient boosting. *Computnl Statist. Data Anal.*, **38**, 367–378.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Goel, S., Rao, J. M. and Shroff, R. (2016) Precinct or prejudice?: Understanding racial disparities in New York city's stop-and-frisk policy. *Ann. Appl. Statist.*, **10**, 365–394.

Goh, S. T. and Rudin, C. (2014) Box drawings for learning with imbalanced data. In *Proc. 20th Special Interest Group on Knowledge Discovery and Data Mining Conf. Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.

Gottfredson, D. M. and Snyder, H. N. (2005) The mathematics of risk classification: changing data into valid instruments for juvenile courts; ncj 209158. Office of Juvenile Justice and Delinquency Prevention, Washington DC.

Grove, W. M. and Meehl, P. E. (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical–statistical controversy. *Psychol. Publ. Poly Law*, **2**, 293–323.

Hannah-Moffat, K. (2013) Actuarial sentencing: an 'unsettled' proposition. *Just. Q.*, **30**, 270–296.

Hanson, R. K. and Thornton, D. (2003) Notes on the development of static-2002. Department of the Solicitor General of Canada, Ottawa.

Hesterberg, T., Choi N. H., Meier, L. and Fraley, C. (2008) Least angle and l_1 penalized regression: a review. *Statist. Surv.*, **2**, 61–93.

Hinojosa, R. H. *et al.* (2005) A comparison of the federal sentencing guidelines criminal history category and the U.S. Parole Commission salient factor score. *Technical Report*. US Sentencing Commission.

Hoffman, P. B. (1994) Twenty years of operational use of a risk prediction instrument: the United States parole commission's salient factor score. *J. Crimnl Just.*, **22**, 477–494.

Hoffman, P. B. and Adelberg, S. (1980) The salient factor score: a nontechnical overview. *Fed. Probn*, **44**, 44.

Holte, R. C. (1993) Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.*, **11**, 63–91.

Holte, R. C. (2006) Elaboration on two points raised in "Classifier technology and the illusion of progress". *Statist. Sci.*, **21**, 24–26.

Howard, P., Francis, B., Soothill, K. and Humphreys, L. (2009) OGRS 3: the revised offender group reconviction scale. *Technical Report*. Ministry of Justice, London.

Kropp, P. R. and Hart, S. D. (2000) The spousal assault risk assessment (sara) guide: reliability and validity in adult male offenders. *Law Hum. Behav.*, **24**, 101–118.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*. New York: Springer.

Kuhn, M., Weston, S., Coulter, N. and Quinlan, R. (2012) C50: C5.0 decision trees and rule-based models. *R Package Version 0.1.0-013*. (Available from `http://CRAN.R-project.org/package=C50`.)

Langan, P. A. and Levin, D. J. (2002) Recidivism of prisoners released in 1994. *Fed. Sentncng Rep.*, **15**, 58–65.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L. and Hansen, K. T. (2007) Actuarial assessment of risk for reoffense among adult sex offenders evaluating the predictive accuracy of the static-2002 and five other instruments. *Crimnl Just. Behav.*, **34**, 37–59.

Liaw, A. and Wiener, M. (2002) Classification and regression by randomforest. *R News*, **2**, no. 3, 18–22.

Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, **40**, 203–228.

Lowenkamp, C. T. and Latessa, E. J. (2004) Understanding the risk principle: how and why correctional interventions can harm low-risk offenders. *Top. Commty Correctns*, 3–8.

Maden, A., Rogers, P., Watt, A., Lewis, G., Amos, T., Gournay, K. and Skapinakis, P. (2006) Assessing the utility of the offenders group reconviction scale-2 in predicting the risk of reconviction within 2 and 4 years of discharge from English and Welsh medium secure units. *Final Report to the National Forensic Mental Health Research Programme*.

Maloof, M. A. (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proc. Workshp Learning from Imbalanced Data Sets II*, vol. 2, pp. 2–1.

McCord, J. (1978) A thirty-year follow-up of treatment effects. *Am. Psychol.*, **33**, 284–289.

McCord, J. (2003) Cures that harm: unanticipated outcomes of crime prevention programs. *Ann. Am. Acad. Polit. Socl Sci.*, **587**, 16–30.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2012) *e1071. R Package Version 1.6-1*. Department of Statistics, Technische Universität Wien, Vienna. (Available from `http://CRAN.R-project.org/package=e1071`.)

Milgram, A. (2014) Why smart statistics are the key to fighting crime. *Ted Talk*, Jan.

Miller, G. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.*, **63**, 81–97.

Nafekh, M. and Motiuk, L. L. (2002) The statistical information on recidivism, revised 1 (SIR-R1) scale: a psychometric examination. Research Branch, Correctional Service of Canada, Ottawa.

Netter, B. (2007) Using group statistics to sentence individual criminals: an ethical and statistical critique of the Virginia Risk Assessment Program. *J. Crimnl Law Crim.*, **97**, 699–729.

Neuilly, M.-A., Zgoba, K. M., Tita, G. E. and Lee, S. S. (2011) Predicting recidivism in homicide offenders using classification tree analysis. *Hom. Stud.*, **15**, 154–176.

Pennsylvania Commission on Sentencing (2012) Risk/needs assessment project interim report 4: development of risk assessment scale. *Report*. Pennsylvania Commission on Sentencing.

Pew Center of the States, Public Safety Performance Project (2011) Risk/needs assessment 101: science reveals new tools to manage offenders. Pew Center of the States, Washington DC.

Quinlan, J. R. (2014) *C4. 5: Programs for Machine Learning*. New York: Elsevier.

R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ridgeway, G. (2006) gbm: generalized boosted regression models. *R Package Version, 1(3)*.

Ridgeway, G. (2013) The pitfalls of prediction. *Natn. Inst. Just. J.*, **271**, 34–40.

Ritter, N. (2013) Predicting recidivism risk: new tool in Philadelphia shows great promise. *Natn. Inst. Just. J.*, **271**, 4–13.

Rubin, P. A. (2009) Mixed integer classification problems. In *Encyclopedia of Optimization*, pp. 2210–2214. New York: Springer.

Sherman, L. W. (2007) The power few: experimental criminology and the reduction of harm. *J. Exptl Crim.*, **3**, 299–321.

Simon, J. (2005) Reversal of fortune: the resurgence of individual risk assessment in criminal justice. *A. Rev. Law Socl Sci.*, **1**, 397–421.

Steinhart, D. (2006) Juvenile detention risk assessment: a practice guide to juvenile detention reform. Annie E. Casey Foundation, Baltimore.

Therneau, T., Atkinson, B. and Ripley, B. (2012) rpart: recursive partitioning. *R Package Version 4.1-0*. (Available from `http://CRAN.R-project.org/package=rpart`.)

Tibbitts, C. (1931) Success or failure on parole can be predicted: a study of the records of 3,000 youths paroled from the Illinois State Reformatory. *J. Crimnl Law Crim.*, **22**, no. 11, 11–50.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Tollenaar, N. and van der Heijden, P. G. M. (2013) Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J. R. Statist. Soc.* A, **176**, 565–584.

Turner, S., Hess, J. and Jannetta, J. (2009) Development of the California Static Risk Assessment Instrument (CSRA). Center for Evidence-Based Corrections, University of California at Irvine, Irvine.

US Department of Justice, Bureau of Justice Statistics (2014) Recidivism of prisoners released in 1994. Inter-university Consortium for Political and Social Research. (Available from http://doi.org/10.3886/ICPSR03355.v8.)

US Sentencing Commission (2012) Criminal history and criminal livelihood, November 1987. In *Guidelines Manual*, ch. Four. Washington DC: US Sentencing Commission.

Ustun, B. (2016) slim_for_matlab v0.1. (Available from http://dx.doi.org/10.5281/zenodo.47964.)

Ustun, B. and Rudin, C. (2015) Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.*, **102**, 349–391.

Webster, C. D. (1997) HCR-20: assessing risk for violence. *Technical Report*. Mental Health, Law, and Policy Institute, Simon Fraser University, Burnaby.

Wolfgang, M. E. (1987) *Delinquency in a Birth Cohort*. Chicago: University of Chicago Press.

Wolsey, L. A. (1998) *Integer Programming*, vol. 42. New York: Wiley.

Wroblewski, J. J. (2014) Annual letter. Criminal Division, US Department of Justice, Washington DC.

Wu, X., Kumar, V., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D. and Steinberg, D. (2008) Top 10 algorithms in data mining. *Knowl. Inform. Syst.*, **14**, 1–37.

Yang, M., Wong, S. C. P. and Coid, J. (2010) The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychol. Bull.*, **136**, 740–767.

Zadrozny, B. and Elkan, C. (2002) Transforming classifier scores into accurate multiclass probability estimates. In *Proc. 8th Special Interest Group on Knowledge Discovery and Data Mining Int. Conf. Knowledge Discovery and Data Mining*, pp. 694–699. New York: Association for Computing Machinery.

Zhang, Y., Zhang, L. and Vaughn, M. S. (2009) Indeterminate and determinate sentencing models: a state-specific analysis of their effects on recidivism. *Crime Delinq.*, **60**, 693–715.