Interpretable Classification Models for Recidivism Prediction

Jiaming Zeng[†], Berk Ustun[†], Cynthia Rudin

[†]These authors contributed equally to this work.

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 (SVM, SGB, Ridge Regression) produce equally accurate models along the full ROC curve. However, methods that designed for interpretability (CART, 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 (SLIM) 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: recidivism, machine learning, interpretability, scoring systems, binary classification

1. Introduction

Forecasting has been used for criminology applications since the 1920s (Borden, 1928; Burgess, 1928) when various factors derived from age, race, prior offense history, employment, grades, and neighborhood 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 datasets on which to apply them.

Recidivism prediction is still extremely important. In the United States, 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 if 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 while low-risk individuals are being released less often then they should be. Her observations are in line with longstanding work on clinical versus actuarial judgment, 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 U.S. 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 as to whether sophisticated machine learning methods (such as random forests, see e.g., Breiman, 2001b; Berk et al., 2009; 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 e.g., Tollenaar and van der Heijden, 2013; Berk and Bleich, 2013; 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) argues 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 contrasts the works of Berk and Bleich (2013) and Tollenaar and van der Heijden

(2013). Berk and Bleich (2013) claim we need sophisticated machine learning methods due to their substantial benefits in accuracy, whereas Tollenaar and van der Heijden (2013) claim that "modern statistical, data mining and machine learning models provides no real advantage over logistic regression and LDA," assuming that humans have done appropriate pre-processing. 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, policy interventions), one 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 if one method is better than another, one 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, 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 (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 recidivism prediction 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 2001a). In this case, there may exist 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. To this end, 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 that have used for many criminal justice applications because they let users make quick predictions by adding, subtracting and multiplying a few small numbers (see e.g., Hoffman and Adelberg, 1980; U.S. Sentencing Commission, 1987; Pennsylvania Commission on Sentencing, 2012). In contrast to existing tools, which have been built using heuristic approaches (see e.g., Gottfredson and Snyder, 2005), the models built by SLIM are fully optimized for accuracy and sparsity, and can handle additional constraints (e.g., bounds on the false positive rate, 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 6 recidivism prediction problems. In Section 3, we provide a brief overview of SLIM and describe several new techniques that can reduce the computation required to produce scoring systems. In Section 4, we compare the accuracy and interpretability of models produced by the 9 machine learning methods on the 6 recidivism prediction problems. We include additional results related to the accuracy and interpretability of models from different methods in the Appendix.

1.2. Related Work

Predictive models for recidivism have been in widespread use in different countries and different areas of the criminal justice system since the early 1920s (see e.g., Borden, 1928; Burgess, 1928; Tibbitts, 1931). The use of these tools has been spurred on by continued research into the superiority of actuarial judgment (Dawes et al., 1989; Grove and Meehl, 1996) as well as a desire to efficiently use limited public resources (Clements, 1996; Simon, 2005; McCord, 1978, 2003). In the U.S., federal guidelines currently mandate the use of a predictive recidivism measure known as the Criminal History Category for sentencing (U.S. Sentencing Commission, 1987). Besides the U.S., countries that currently use risk assessment tools include Canada (Hanson and Thornton, 2003), the Netherlands (Tollenaar and van der Heijden, 2013), and the U.K. (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 parole violations (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 offenses). 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 e.g., the Spousal Abuse Risk Assessment of Kropp and Hart, 2000), sexual violence (see e.g., Hanson and Thornton, 2003; Langton et al., 2007), and general violence (see e.g., Historical Clinical and Risk Management tool of Webster et al. 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 – that is, linear classification models that only require users 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 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 (SFS) (Hoffman and Adelberg, 1980), the Criminal History Category (CHC) (U.S. Sentencing Commission, 1987), and the Offense Gravity Score (OGS) (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 e.g., the methods 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 recidivate) as opposed to an predicted probability of the outcome (i.e. the predicted probability that prisoner i will recidivate 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 $> 70\%$). 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 $< 30\%$, “medium risk” if the predicted probability is $< 60\%$, and “high risk” otherwise). These arbitrary threshholds make it difficult, if not impossible, to effectively assess the predictive accuracy of the tools (Hannah-Moffat, 2013). Netter (2007), for instance, mentions 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 to 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 is can yield a scoring system that optimizes the class-based accuracy at a particular decision point (i.e., produce 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 modeling in criminology. In contrast to 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, 2001b) and stochastic gradient boosting Friedman (2002). Random forests (Breiman, 2001b), 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 help inform court decisions at arraignment (Berk and Sorenson, 2014). We note that not all studies in used black-box models: Berk et al. (2005), for instance, help the Los Angeles Sheriff’s Department develop a simple and practical screener to forecast domestic violence using decision trees. More recently, (Goel et al., 2015), developed a simple scoring system to help the New York Police Department address stop and frisk by first running logistic regression, and then rounding the coefficients.

2. Data and Prediction Problems

Each problem is a binary classification problem with $N = 33,796$ 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 U.S. Department of Justice, Bureau of Justice Statistics (2014). It is the largest publicly available database on prisoner recidivism in the United States. The study tracked 38,624 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 U.S. states (Arizona, California, Delaware, Florida, Illinois, Maryland, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Texas, and Virginia). The sampled population accounts for roughly two-thirds of all prisoners that were released from prison in the U.S. in 1994. Other studies that use this database include: Bhati and Piquero (2007); Bhati (2007); Zhang et al. (2009).

The database is composed of 38,624 rows and 6,427 columns, where each row represents a prisoner and each column represents a feature (i.e. a field of information for a given prisoner). The 6,427 columns consist of 91 fields that were recorded before or during release from prison in 1994 (e.g., date of birth, effective sentence length), and 64 fields that were repeatedly recorded for up to 99 different arrests in the 3 year follow-up period (e.g., if a prisoner was rearrested three times with 3 years, there would be three record cycles recorded). The information for each prisoner is sourced from record-of-arrest-and-prosecution (RAP) sheets kept by state law enforcement agencies and/or the FBI. A detailed descriptive analysis of the database was carried out by statisticians at the U.S. Bureau of Justice Statistics (Langan and Levin, 2002). This study restricted its attention to 33,796 of the 38,624 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/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 dataset 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 noisy enough to be excluded from our preliminary experiments. The information about education levels 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 reentry programs or employment history. While some of these factors exist, such as drug or alcohol treatment and in-prison vocational programs, the data is highly incomplete and therefore 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 derived from the database in 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 nonlinear functions of the original variables. We refer to input variables in the text using italicized font (e.g., *female*). All prediction problems in Table 2 and all machine learning methods in Table 4 use these same input variables.

The final set of input variables are representative of well-known risk factors for recidivism (Bushway and Piehl, 2007; Crow, 2008) and have been used in risk assessment tools since 1928 (see e.g., Borden, 1928; Ricardo H. Hinojosa et al., 2005; Berk et al., 2006; Baradaran, 2013). They include: 1) information about prison release in 1994 (e.g., *time_served*, *age_at_release*, *infraction_in_prison*); 2) information from past arrests, sentencing, and convictions (e.g., *prior_arrests*$\geq$*1*, *any_prior_jail_time*);[1] 3) history of substance abuse (e.g., *alcohol_abuse*) 4) gender (e.g., *female*). These input variables are advantageous because: a) the

[1] 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.

information is easily accessible to law enforcement officials (all above information can be found in state RAP sheets); b) they do not include socioeconomic 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 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 optimization problem when the data only contains binary input variables (as discussed in Section 3.3). In Appendix E, 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, CART and C5.0T experienced an improvement of 4.6% for `drug` and SVM 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 6 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 using typewriter font (e.g., `arrest`). We provide details on each recidivism prediction problems in Table 2. These include: an arrest for any crime (`arrest`); an arrest for a drug-related offense (`drug`); or an arrest for a certain type of violent offense (`general_violence`, `domestic_violence`, `sexual_violence`, `fatal_violence`).

In the dataset, all crime types can be broken down into smaller subcategories (e.g., `fatal_violence` can be broken into 6 subcategories such as `murder`, `vehicular_manslaughter`, etc.). We chose to use the broader crime categories for the sake of conciseness and clarity. Indeed, the study by Langan and Levin (2002) also split the crimes into the same major categories. We note that the outcomes of violent offenses are mutually exclusive, as different types of violence are treated differently within the U.S. legal system. In other words, $y_i = +1$ for `general_violence` does not necessarily imply $y_i = +1$ for `domestic_violence`, `sexual_violence`, `fatal_violence`).

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

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

Table 2. Overview of recidivism prediction problems. 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 offenses), and could also be arrested multiple times over the 3 year follow-up period.

Prediction Problem	**P**$(y_i = +1)$	**Outcome Variable**
`arrest`	59.0%	$y_i = +1$ if prisoner i is arrested for any offense within 3 years of release from prison
`drug`	20.0%	$y_i = +1$ if prisoner i is arrested for drug-related offense (e.g., possession, trafficking) within 3 years of release from prison
`general_violence`	19.1%	$y_i = +1$ if prisoner i is arrested for a violent offense (e.g., robbery, 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

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 this table, we can identify strong associations between the input and output for each prediction problem. These associations can help uncover insights into each problem and also help qualitatively validate predictive models in Section 4.4.

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 the case for prisoners who were first confined (i.e., sent to prison or jail) at an older age (see e.g., *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 offense are more likely to be rearrested for a drug-related offense (32%) than those who were previously arrested for any other type of offense. 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%).

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

Input Variable	**Prediction Problem**					
	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≤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≥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≤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≥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≤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≥40	0.21	0.05	0.03	0.01	0.02	0.002
age_1st_confinement≤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≥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≥1	0.63	0.22	0.19	0.04	0.03	0.01
prior_arrest≥2	0.66	0.23	0.20	0.04	0.03	0.01
prior_arrest≥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

3. Supersparse Linear Integer Models

A *Supersparse Linear Integer Model* (SLIM) is a new machine learning method for creating *scoring systems* – that is, binary classification models that only require users 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:

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

Here, $\lambda_1, \ldots, \lambda_P$ represent the coefficients (i.e. the “points” for input variables $j = 1, \ldots, P$), and λ_0 represents an intercept (i.e. the “threshold score” that has to 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 following form:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \neq \hat{y}_i\right] + C_0 \sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right] + \epsilon \sum_{j=1}^{P} |\lambda_j| \\ \text{s.t.} \quad & (\lambda_0, \lambda_1, ..., \lambda_P) \in \mathcal{L}. \end{aligned} \tag{1}$$

Here, the objective directly minimizes the error rate $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \neq \hat{y}_i\right]$ and directly penalizes the number of non-zero terms $\sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right]$. 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 desired scoring system.

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 $\hat{y} = \text{sign}\,(x_1 + x_2)$. If SLIM only minimized the misclassification rate and the number of terms (the first two terms of the objective), then $\hat{y} = \text{sign}\,(2x_1 + 2x_2)$ would have the same objective value as $\hat{y} = \text{sign}\,(x_1 + x_2)$ 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, $\hat{y} = \text{sign}\,(x_1 + x_2)$.

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 datasets. 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 into its integer programming 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 than (1) in that it includes class weights, and has specific constraints on the coefficients:

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{\lambda}} \dfrac{W^+}{N} \displaystyle\sum_{i\in\mathcal{I}^+} \mathbb{1}\left[y_i \neq \hat{y}_i\right] + \dfrac{W^-}{N} \displaystyle\sum_{i\in\mathcal{I}^-} \mathbb{1}\left[y_i \neq \hat{y}_i\right] + C_0 \displaystyle\sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right] + \epsilon \displaystyle\sum_{j=1}^{P} |\lambda_j| & \\
\text{s.t.} \quad \displaystyle\sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right] \leq 8 & (2) \\
\qquad \lambda_j \in \{-10, \ldots, 10\} \text{ for } j = 1...P & \\
\qquad \lambda_0 \in \{-100, \ldots, 100\}. &
\end{array}
$$

In the formulation above, 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 to at most 8 (i.e., within the range of cognitive entities 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

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

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 $0 < \epsilon < C_0 / \max \sum_{j=1}^{P} |\lambda_j|$ to ensure this condition.

3.2. General SLIM IP Formulation

Training a SLIM scoring system requires solving an integer programming (IP) problem using a solver such as CPLEX, Gurobi, or CBC. In general, we use the following IP formulation to recover the solution to the optimization problem (2):

$$
\begin{array}{lrcllrl}
\min\limits_{\boldsymbol{\lambda},\mathbf{z},\boldsymbol{\Phi},\boldsymbol{\alpha},\boldsymbol{\beta}} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} z_i & + & \displaystyle\sum_{j=1}^{P} \Phi_j & & & \\
\text{s.t.} & M_i z_i & \geq & \gamma - \displaystyle\sum_{j=0}^{P} y_i \lambda_j x_{i,j} & i = 1...N & \textit{error on } i & \text{(3a)} \\
& \Phi_j & = & C_0 \alpha_j + \epsilon \beta_j & j = 1...P & \textit{penalty for coef } j & \text{(3b)} \\
& -\Lambda_j \alpha_j & \leq & \lambda_j \leq \Lambda_j \alpha_j & j = 1...P & \ell_0\textit{-norm} & \text{(3c)} \\
& -\beta_j & \leq & \lambda_j \leq \beta_j & j = 1...P & \ell_1\textit{-norm} & \text{(3d)} \\
& \lambda_j & \in & \mathbb{Z} \cap [-\Lambda_j, \Lambda_j] & j = 0...P & \textit{coefficient set} & \\
& z_i & \in & \{0, 1\} & i = 1...N & \textit{loss variables} & \\
& \Phi_j & \in & \mathbb{R}_+ & j = 1...P & \textit{penalty variables} & \\
& \alpha_j & \in & \{0, 1\} & j = 1...P & \ell_0 \textit{ variables} & \\
& \beta_j & \in & \mathbb{R}_+ & j = 1...P. & \ell_1 \textit{ variables} &
\end{array}
$$

The constraints in (3a) compute the error rate by setting the *loss variables* $z_i = \mathbb{1}\left[y_i \boldsymbol{\lambda}^T \boldsymbol{x}_i \leq 0\right]$ 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 e.g., Rubin, 2009). The value of M_i represents the maximum score when example i is misclassified, and can be set as $M_i = \max_{\boldsymbol{\lambda}\in\mathcal{L}}(\gamma - y_i \boldsymbol{\lambda}^T \boldsymbol{x}_i)$ 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 $y_i\boldsymbol{\lambda}^T\boldsymbol{x}_i$. 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 (3b) set the total penalty for each coefficient to $\Phi_j = C_0\alpha_j + \epsilon\beta_j$, where $\alpha_j := \mathbb{1}\left[\lambda_j \neq 0\right]$ is defined by Big-M constraints in (3c), and $\beta_j := |\lambda_j|$ is defined by the constraints in (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 to 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 the user chooses one that is very large, leading to a less efficient formulation (see e.g., Wolsey, 1998). For SLIM, the Big-M constants used to compute the 0–1 loss in constraint (3a) is bounded as $M_i \leq \max_{\boldsymbol{\lambda}\in\mathcal{L}}(\gamma - y_i\boldsymbol{\lambda}^T\boldsymbol{x}_i)$, and the Big-M constant used to compute the ℓ_0-norm in constraints (3c) is bounded as $\Lambda_j \leq \max_{\lambda_j\in\mathcal{L}_j} |\lambda_j|$. Bounding these constants lead to a tighter LP 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 SLIM IP 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 get repeated feature values among observations.

$$
\begin{array}{lrcl l r}
\min\limits_{\boldsymbol{\lambda},\mathbf{z},\boldsymbol{\Phi},\boldsymbol{\alpha},\boldsymbol{\beta}} & \dfrac{W^+}{N}\displaystyle\sum_{s\in\mathcal{S}} n_s z_s & + & \dfrac{W^-}{N}\displaystyle\sum_{t\in\mathcal{T}} n_t z_t + \sum_{j=1}^{P}\Phi_j & & \\
\text{s.t.} & M_s z_s & \geq & 1 - \displaystyle\sum_{j=0}^{P}\lambda_j x_{s,j} & s \in \mathcal{S} & \textit{error on s}\,\text{(4a)} \\
 & M_t z_t & \geq & \displaystyle\sum_{j=0}^{P}\lambda_j x_{t,j} & t \in \mathcal{T} & \textit{error on t}\,\text{(4b)} \\
 & 1 & = & z_s + z_t & \forall s,t : \boldsymbol{x}_s = \boldsymbol{x}_t, y_s = -y_t & \textit{conflicting labels}\,\text{(4c)} \\
 & \Phi_j & = & C_0\alpha_j + \epsilon\beta_j & j = 1...P & \textit{penalty for coef j}\,\text{(4d)} \\
 & -\Lambda_j\alpha_j & \leq & \lambda_j \leq \Lambda_j\alpha_j & j = 1...P & \ell_0\textit{-norm}\,\text{(4e)} \\
 & -\beta_j & \leq & \lambda_j \leq \beta_j & j = 1...P & \ell_1\textit{-norm}\ \text{(4f)} \\
 & \lambda_j & \in & \mathbb{Z}\cap[-\Lambda_j, \Lambda_j] & j = 0...P & \textit{coefficient set} \\
 & z_s, z_t & \in & \{0,1\} & s \in \mathcal{S}\ t \in \mathcal{T} & \textit{loss variables} \\
 & \Phi_j & \in & \mathbb{R}_+ & j = 1...P & \textit{penalty variables} \\
 & \alpha_j & \in & \{0,1\} & j = 1...P & \ell_0\ \textit{variables} \\
 & \beta_j & \in & \mathbb{R}_+ & j = 1...P. & \ell_1\ \textit{variables}
\end{array}
$$

The main difference between this formulation and the one in (3) is that we compute the error rate of the classifier using loss constraints that are expressed in terms of the number of *distinct* points in the dataset. 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 a point of type s (or t) are found in the original dataset so that $\sum_s n_s = \sum_{i=1}^{N} \mathbb{1}\left[y_i = +1\right]$, $\sum_t n_t = \sum_{i=1}^{N} \mathbb{1}\left[y_i = -1\right]$, and $N = \sum_s n_s + \sum_t n_t$.

The main computational benefits of this formulation are due to the fact that: (i) we can reduce the number of loss constraints by counting the number of repeated rows in the dataset; and (ii) we can directly encode a lower bound on the error rate by counting the number of points s, t with identical feature but opposite labels (i.e., $\boldsymbol{x}_s$ = $\boldsymbol{x}_t$ but $y_s = -y_t$). Here (i) reduces the size of the problem that we pass to an IP solver, and (ii) produces a much stronger lower bound on the 0–1 loss (in comparison to the LP relaxation), which speeds up the progress of branch-and-bound type algorithms. Note that it would be possible to use this formulation on a dataset without binary input variables, though it would not necessarily be effective because it could be much less likely for a dataset to contain repeated rows in such a setting.

Another subtle benefit of this formulation is that the margin for the negative points is 0 while 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 ≥ 1. For negative points, we have a correct prediction if and only if the score ≤ 0. 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 we would to produce a linear model that encode: "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 = sign$(1 - 2a_1 - 2a_2)$." In contrast, the margin of the current formulation is: "rearrest = sign$(1 - a_1 - a_2)$", which uses smaller coefficients, and produces a slightly simpler model.

3.4. Active Set Polishing

On large datasets, IP solvers may take long time to produce an optimal solution or provider 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 a simpler IP formulation shown in (5). The polishing IP only optimizes the coefficients of features that belong to the *active set* of $\boldsymbol{\lambda}^{\text{feasible}}$: that is, the set of features with nonzero coefficients $\mathcal{A} := \left\{j : \lambda_j^{\text{feasible}} \neq 0\right\}$. The coefficients for features that do not belong to the active set are fixed to zero 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.

$$
\begin{array}{llcllr}
\min\limits_{\boldsymbol{\lambda},\mathbf{z},\boldsymbol{\Phi},\boldsymbol{\alpha},\boldsymbol{\beta}} & \frac{W^+}{N}\sum\limits_{s\in\mathcal{S}} n_s z_s & + & \frac{W^-}{N}\sum\limits_{t\in\mathcal{T}} n_t z_t & & \text{(5a)} \\
\text{s.t.} & M_s z_s & \geq & 1 - \sum\limits_{j\in\mathcal{A}} \lambda_j x_{s,j} & s \in \mathcal{S} \quad \textit{error on s} & \text{(5b)} \\
 & M_t z_t & \geq & \sum\limits_{j\in\mathcal{A}} \lambda_j x_{t,j} & t \in \mathcal{T} \quad \textit{error on t} & \text{(5c)} \\
 & 1 & = & z_s + z_t & \forall s,t : \boldsymbol{x}_s = \boldsymbol{x}_t, y_s = -y_t \quad \textit{conflicting labels} & \text{(5d)} \\
 & \lambda_j & \in & \mathbb{Z} \cap [-\Lambda_j, \Lambda_j] & j \in \mathcal{A} \quad \textit{coefficient set} & \\
 & z_s, z_t & \in & \{0,1\} & s \in \mathcal{S}\ t \in \mathcal{T}. \quad \textit{loss variables} &
\end{array}
$$

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 5 nonzero 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 of the feasible solutions we find from the earlier formulation. In all cases, we can solve the polishing IP to optimality within a few seconds (i.e. a MIPGAP of 0.0%).

4. Experimental Results

In this section, we compare the accuracy and interpretability of recidivism prediction models from SLIM to models from 8 other popular classification methods. In Section 4.1, we explain the experimental setup used for all the methods. In Section 4.2, we compare the predictive accuracy of the methods with the AUC values and ROC curves. In Section 4.3 and 4.4, we evaluate the interpretability of the models. Finally, in Section 4.5, we present the scoring systems generated by SLIM.

4.1. Methodology

In what follows we discuss cost-sensitive classification for imbalanced problems, 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 $\mathrm{P}(y_i = +1) = 0.7\%$ 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 no one will be arrested for fatal offenses – 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 that will be arrested for a fatal offense, and therefore 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 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 negative class. For a given classification model, we compute these quantities as

$$TPR = \frac{1}{N^+} \sum_{i \in \mathcal{I}^+} \mathbb{1}\left[\hat{y}_i = +1\right] \quad \text{and} \quad FPR = \frac{1}{N^-} \sum_{i \in \mathcal{I}^-} \mathbb{1}\left[\hat{y}_i = +1\right],$$

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 produced by a given method as points on a *receiver operating characteristic* (ROC) curve, which plots the TPR on the vertical axis and the 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).[2] 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 ROC Curve 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 in order to avoid predicting that a low-risk individual will reoffend. Models for screening, however, need high TPR in order 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 e.g., 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 that $W^+ + W^- = 2$. 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 to 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:

- **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.

[2] We note that AUC is a summary statistic that is frequently misused in the context of classification problems. It is true that a method that 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.

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

- L_1 **and** L_2**-Penalized Logistic Regression**: 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" and are not generally sparse.

- **Random Forests**: A popular black-box method that makes predictions using a large ensemble of weak classification trees. The method was originally developed by Breiman (2001b) but is widely used for recidivism prediction (see e.g., Berk et al., 2009; Ritter, 2013).

- **Support Vector Machines**: A popular black-box method for non-parametric linear classification. The Radial Basis Function (RBF) kernel lets the method to handle classification problems where the decision-boundary may be non-linear (see e.g., Cristianini and Shawe-Taylor, 2000; Berk and Bleich, 2014).

- **Stochastic Gradient Boosting**: A popular black-box method that create 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 6 recidivism prediction problems and each of the 9 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 setup requires us to produce a total of 1,026 recidivism prediction models (6 recidivism problems × 9 methods × 19 imbalance ratios). Each of the 1,026 models were built on a training set and their performance was assessed out-of-sample. In particular, 1/3 of the data was reserved as the *test set*. The remaining 2/3 of the data was the *training set*. During training, we used 5-fold nested cross-validation (5-CV) for parameter tuning. Explicitly, the training data were split into 5 folds, and one of those 5 was reserved as the validation fold. The validation fold was rotated in order to select free parameter values, and a *final model* was trained on the full training set (2/3) with the selected parameter values and its performance was assessed on the test set (1/3). The folds were generated once to allow for comparisons across methods and prediction problems. The parameters were chosen during nested cross validation to minimize the mean weighted 5-CV validation 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 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 integer programming problems (IP) with the CPLEX 12.6 API in MATLAB 2013a. We solved each IP through the following procedure: (i) we trained the solver on the formulation in Section 3.3 for a total of 4 hours on a local computing cluster with 2.7GHz CPUs. Each time we solved a IP we kept 500 feasible solutions, and polished them using the formulation in Section 3.4. We then used the same nested cross-validation procedure as the other methods to tune the number of terms in the final model. Polishing all 500 solutions took less than one minute of computing time. Thus, the total number of optimization problems we solved were 500 polishing IP's × (5 folds + 1 final model) × 6 problems × 19 values of W^+ = 342,000 integer programming problems.

4.2. Observations on Predictive Accuracy

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

Table 4. Methods, software and free parameters used to train models for all 6 recidivism prediction problems. 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 5-CV validation error. The values of W^+ are problem-specific (see Section 4.1.4 for details)

Method	**Acronym**	**Software**	**Free Parameters and Settings**
CART Decision Trees	CART	**rpart** (Therneau et al., 2012)	minSplit $\in (3, 5, 10, 15, 20) \times$ CP $\in (0.0001, 0.001, 0.01)$
C5.0 Decision Trees	C5.0T	**c50** (Kuhn et al., 2012)	default settings
C5.0 Decision Rules	C5.0R	**c50** (Kuhn et al., 2012)	default settings
Logistic Regression (L_1-Penalty)	Lasso	**glmnet** (Friedman et al., 2010)	100 values of L_1-penalty chosen by **glmnet**
Logistic Regression (L_2-Penalty)	Ridge	**glmnet** (Friedman et al., 2010)	100 values of L_2-penalty chosen by **glmnet**
Random Forests	RF	**randomForest** (Liaw and Wiener, 2002)	sampsize $\in (0.632N, 0.4N, 0.2N) \times$ nodesize $\in (1, 5, 10, 20)$ with unbounded tree depth
Support Vector Machines (Radial Basis Kernel)	SVM RBF	**e1071** (Meyer et al., 2012)	$C \in (0.01, 0.1, 1, 10) \times$ $\gamma \in (\frac{1}{10P}, \frac{1}{5P}, \frac{1}{2P}, \frac{1}{P}, \frac{2}{P}, \frac{5}{P}, \frac{10}{P})$
Stochastic Gradient Boosting (Adaboost)	SGB	**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	SLIM	**CPLEX 12.6** (Ustun, 2016)	C_0 and ϵ set to find most accurate model with ≤ 8 coefficients where $\lambda_0 \in \{-100, \dots, 100\}$ and $\lambda_j \in \{-10, \dots, 10\}$

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

- 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, and RF. This multiplicity of good models reflects the *Rashomon effect* of Breiman (2001b).

- 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 e.g., problems with significant imbalance, such as `fatal_violence`).

- SVM RBF, SGB, Lasso and Ridge were able to 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`, `fatal_violence`). Lasso, Ridge, and SVM RBF 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 get models at specific points on the ROC curve.

- C5.0T, C5.0R and CART were unable to 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 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, and has been observed by others (see e.g., Goh and Rudin, 2014). This problem might not occur if trees were better optimized.

- 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 8 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. Figure 2 show the risk calibration for `arrest`, constructed using the binning method from Zadrozny and Elkan (2002). We include calibration plots for all other problems in Appendix B. We see that SLIM is well-calibrated, even though there is no reason it should be; it is a decision-making tool, not a risk assessment tool. For `arrest`, Lasso and Ridge are well-calibrated; however, they lose this quality once we consider only sparse models (see Appendix D). This property would also be lost if the Lasso and Ridge coefficients were rounded.

Table 5. Test AUC for all methods on all prediction problems. Each cell contains the test AUC.

Prediction Problem	**Lasso**	**Ridge**	**C5.0R**	**C5.0T**	**CART**	**RF**	**SVM RBF**	**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

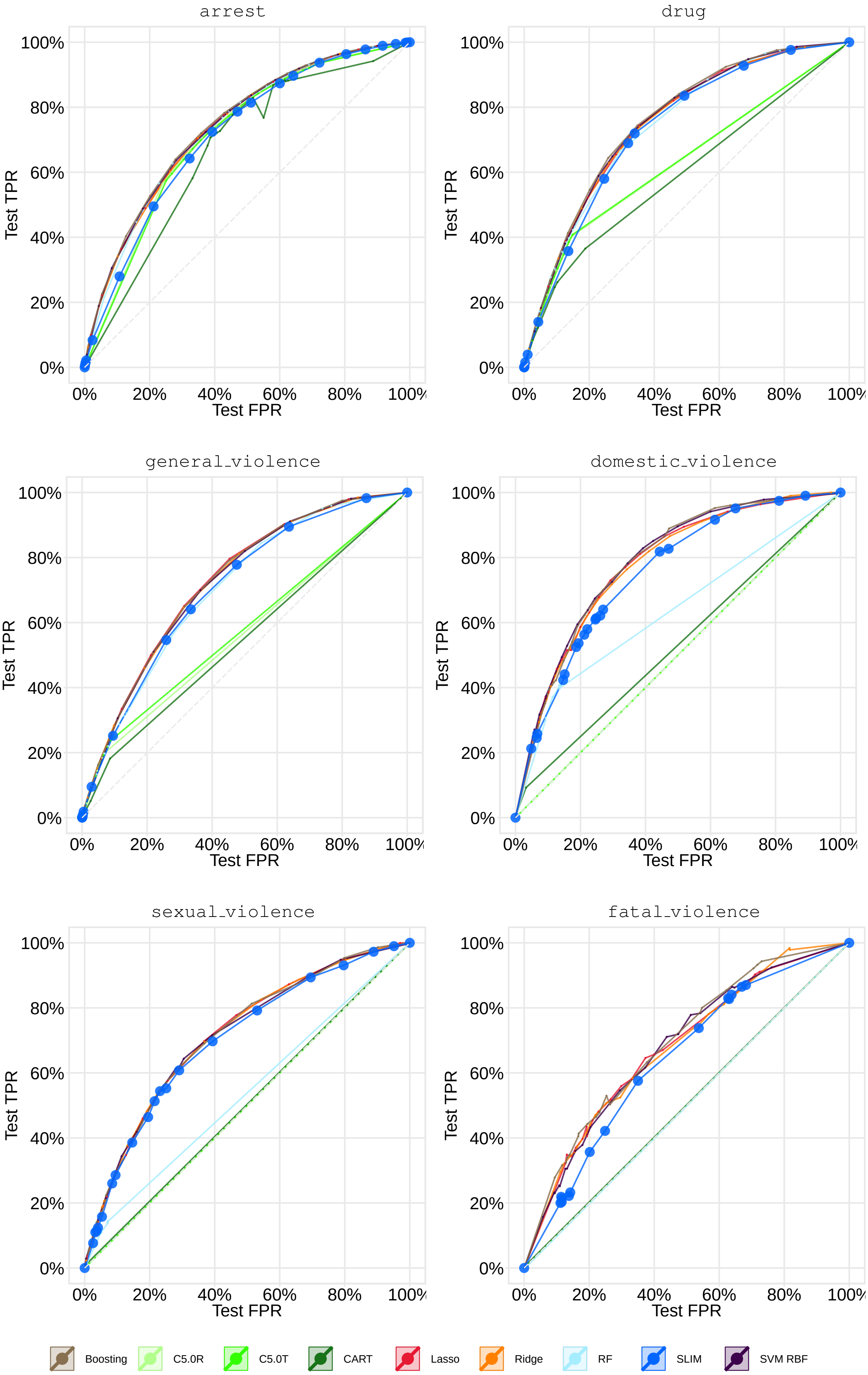

Fig. 1. ROC curves for general recidivism-related prediction problems with test data. We plot SLIM models using large blue dots. All models perform similarly except for C5.0R, C5.0T, and CART.

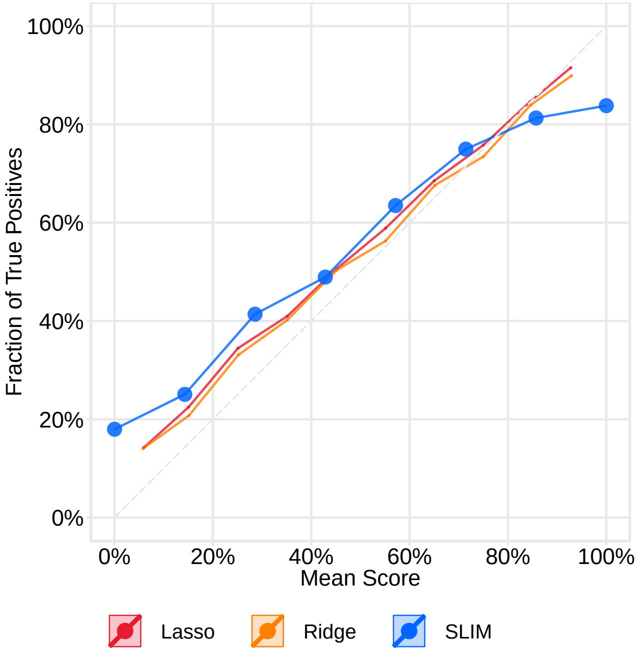

Fig. 2. Risk calibration plot for `arrest` based on test data. We compare 3 models chosen at a similar decision point, with test FPR$\leq 50\%$. Although it is not a risk assessment tool, we see that SLIM is well calibrated.

4.3. Trade-offs Between Accuracy and Interpretability

In Appendix C, we show that the baseline methods are unable to maintain the same level of accuracy as they have in Section 4.2 when their model size was constrained. For Lasso, Ridge and SLIM, model size is defined as the number of features in the model. For CART and C5.0, model size is the number of leaves or rules. In fact, we find the only methods that can consistently produce accurate models along the full ROC curve and also have the potential for interpretability are SLIM and (non-sparse) Lasso.

Tree and rule-based methods such as CART, C5.0T and C5.0R were generally unable to produce models that attain high degrees of accuracy. Worse, even for balanced problems such as `arrest`, where these methods did produce accurate models, the models are complicated and use a very large number of rules or leaves (similar behavior for C5.0T/C5.0R is 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/CART model with at most 8 rules or 8 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 SLIM, Lasso and CART for the `arrest` problem in Figures 3–5. This setup 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 8 coefficients (Lasso), 8 rules (C5.0R) or 8 leaves (C5.0T, 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 8 rules or 8 leaves, so only models from SLIM, CART and Lasso could be displayed. As described before, it is rare for Lasso and CART to produce models with a similar degree of accuracy to SLIM when model size is constrained. We make the following observations:

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

- The SLIM model uses 5 input variables and small integer coefficients (see e.g., Figure 3). 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*$\geq$*5*), he is highly likely to be rearrested. On the other hand, if he is relatively older (*age_at_release*$\geq$*40*) or has no history of arrests (*no_prior_arrests*), he is unlikely to commit another crime.

- The CART model also allows users to make predictions without a calculator. In comparison to 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*$\geq$*40* and *age_at_release_18_to_24*. However, the effect is difficult to comprehend since it depends on prior arrests for misdemeanor: if *prior_arrests*$\geq$*5* = 1 and *age_at_release_18_to_24* = 1 then the model predicts $\hat{y} = +1$; if *prior_arrests*$\geq$*5* = 0 and *age_at_release*$\geq$*40* = 0 then $\hat{y} = +1$; however, if *prior_arrests*$\geq$*5* = 0 and *age_at_release*$\geq$*40* = 1 then $\hat{y} = +1$ only if *prior_arrest_for_misdemeanor* = 1. Such issues do not affect linear models such as SLIM and Lasso, where users can immediately gauge the direction and strength of the relationship between a 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).

PREDICT ARREST FOR ANY OFFENSE 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**	=	······

Fig. 3. SLIM scoring system for `arrest`. This model has a test TPR/FPR of 76.6%/44.5%, and a mean 5-CV validation TPR/FPR of 78.3%/46.5%.

PREDICT ARREST FOR ANY OFFENSE 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**	=	······

Fig. 4. Lasso model for `arrest`, with coefficients rounded to two significant digits. This model has a test TPR/FPR of 70.9%/43.8%, and a mean 5-CV validation TPR/FPR of 72.2%/44.0%.

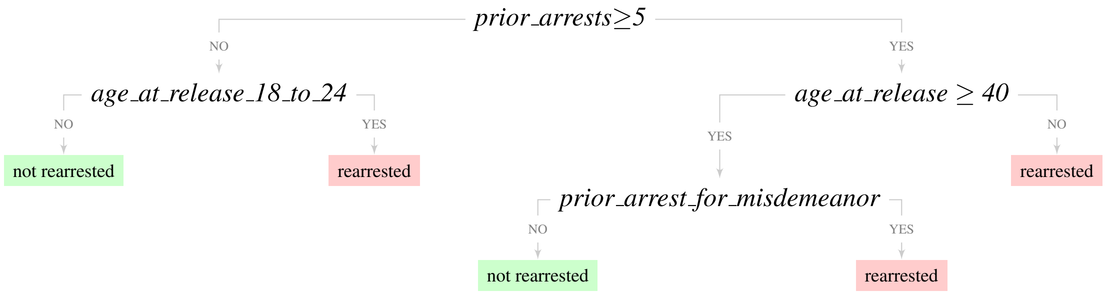

Fig. 5. CART model for `arrest`. This model has a test TPR/FPR of 79.1%/47.9%, and a mean 5-CV validation TPR/FPR of 79.9%/48.5%.

4.5. Scoring Systems for Recidivism Prediction

We show a SLIM scoring system for each of the prediction problems that we consider in Figures 6–10. The models are chosen at specific decision points, with the constraint that 5-CV FPR$\leq$ 50% except for `sexual_violence`, which is chosen at 5-CV FPR$\leq$ 20%. The models presented here may be suitable for screening tasks. To obtain a model 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 dataset, evident by the close match between test TPR/FPR (Table 5) and training TPR/FPR (Table 6).

Many of these models exhibit the same "rule-like" tendencies discussed in Section 4.4. For example, the model for `drug` in Figure 6 predicts that a person will be arrested for a drug-related offense if he/she has ever had any prior drug offenses. Similarly, model for `sexual_violence` in Figure 9 effectively states that a person will be rearrested for a sexual offense if and only if he/she has prior history of sexual crimes. For completeness, we include comparisons with other models in Appendix B. Additional risk calibration plots for models with constrained model size are included in Appendix D.

PREDICT ARREST FOR DRUG OFFENSE IF SCORE > 7

1.	*prior_arrest_for_drugs*	9 points	$\cdots\cdots$
2.	*age_at_release_18_to_24*	5 points	$+\ \cdots\cdots$
3.	*age_at_release_25_to_29*	3 points	$+\ \cdots\cdots$
4.	*prior_arrest_for_multiple_types_of_crime*	2 points	$+\ \cdots\cdots$
5.	*prior_arrest_for_property*	1 points	$+\ \cdots\cdots$
6.	*age_at_release_30_to_39*	-1 point	$+\ \cdots\cdots$
7.	*no_prior_arrests*	-6 points	$+\ \cdots\cdots$
	ADD POINTS FROM ROWS 1-7	**SCORE**	$=\ \cdots\cdots$

Fig. 6. SLIM scoring system for `drug`. This model has a test TPR/FPR of 85.7%/51.1%, and a mean 5-CV validation TPR/FPR of 82.3%/49.7%.

PREDICT ARREST FOR GENERAL VIOLENCE OFFENSE IF SCORE > 7

1.	*prior_arrest_for_general_violence*	8 points	$\cdots\cdots$
2.	*prior_arrest_for_misdemeanor*	5 points	$+\ \cdots\cdots$
3.	*infraction_in_prison*	3 points	$+\ \cdots\cdots$
4.	*prior_arrest_for_local_ord*	3 points	$+\ \cdots\cdots$
5.	*prior_arrest_for_property*	2 points	$+\ \cdots\cdots$
6.	*prior_arrest_for_fatal_violence*	2 points	$+\ \cdots\cdots$
7.	*prior_arrest_with_firearms_involved*	1 point	$+\ \cdots\cdots$
8.	*age_at_release$\geq$40*	-7 points	$+\ \cdots\cdots$
	ADD POINTS FROM ROWS 1-8	**SCORE**	$=\ \cdots\cdots$

Fig. 7. SLIM scoring system for `general_violence`. This model has a test TPR/FPR of 76.7%/45.4%, and a mean 5-CV validation TPR/FPR of 76.8%/47.6%.

PREDICT ARREST FOR DOMESTIC VIOLENCE OFFENSE IF SCORE > 3

1.	*prior_arrest_for_misdemeanor*	4 points		······
2.	*prior_arrest_for_felony*	3 points	+	······
3.	*prior_arrest_for_domestic_violence*	2 points	+	······
4.	*age_1st_confinement_18_to_24*	1 point	+	······
5.	*infraction_in_prison*	-5 points	+	······
	ADD POINTS FROM ROWS 1-5	**SCORE**	=	······

Fig. 8. SLIM scoring system for `domestic_violence`. This model has a test TPR/FPR of 85.5%/46.0%, and a mean 5-CV validation TPR/FPR of 81.4%/48.0%.

PREDICT ARREST FOR SEXUAL VIOLENCE OFFENSE IF SCORE > 2

1.	*prior_arrest_for_sexual*	3 points		······
2.	*prior_arrests≥5*	1 point	+	······
3.	*multiple_prior_jail_time*	1 point	+	······
4.	*prior_arrest_for_multiple_types_of_crime*	-1 point	+	······
5.	*no_prior_arrests*	-2 points	+	······
	ADD POINTS FROM ROWS 1-5	**SCORE**	=	······

Fig. 9. SLIM scoring system for `sexual_violence`. This model has a test TPR/FPR of 44.3%/17.7%, and a mean 5-CV validation TPR/FPR of 43.7%/19.9%.

PREDICT ARREST FOR FATAL VIOLENCE OFFENSE 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**	=	······

Fig. 10. SLIM scoring system for `fatal_violence`. This model has a test TPR/FPR of 55.4%/35.5%, and a mean 5-CV validation TPR/FPR of 64.2%/42.4%.

5. Discussion

Our paper merges two perspectives on recidivism modeling: the first is to obtain accurate predictive models using the most powerful machine learning tools 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 different 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 Stochastic Gradient Boosting – 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 e.g., Dawes, 1979; Holte, 1993, 2006).

Our study shows that there may be major advantages of using SLIM for recidivism prediction, 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 to current machine learning methods, the main drawback of running SLIM is increased computation involved in solving an integer programming problem. To this end, we proposed two new techniques to reduce computation involved in training high quality SLIM scoring systems: (i) a polishing procedure that improves the quality of feasible solutions found by an IP solver; and (ii) an IP formulation that makes it easier for an IP solver to provide a certificate of optimality. In our experiments, the time required to train SLIM was ultimately comparable to the time required to train random forests or stochastic gradient boosting. However, it was still significant compared to the time required for other methods such as CART, C5.0 and penalized logistic regression. In theory, the computation required to find an optimal solution to the SLIM integer program is NP-hard, meaning that the runtime increases exponentially with the number of features. In practice, the runtime 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 datasets (since each observation is a person), and since computer speed of solving MIPs is also increasing exponentially, it is possible that mathematical programming techniques like SLIM are well-suited for criminological problems that are substantially larger and more complex than the one discussed in this work.

References

Andrade, Joel T. *Handbook of violence risk assessment and treatment: New approaches for mental health professionals*. Springer Publishing Company, 2009.

Andrews, Donald A and James Bonta. *The level of service inventory-revised.* Multi-Health Systems, 2000.

Baradaran, Shima. Race, prediction, and discretion. *Geo. Wash. L. Rev.*, 81:157, 2013.

Barnes, Geoffrey C and Jordan M Hyatt. Classifying adult probationers by forecasting future offending. Technical report, National Institute of Justice, U.S. Department of Justice, 2012.

Belfrage, Henrik, Ran Fransson, and Susanne Strand. Prediction of violence using the hcr-20: A prospective study in two maximum-security correctional institutions. *The Journal of Forensic Psychiatry*, 11(1):167–175, 2000.

Berk, Richard. The role of race in forecasts of violent crime. *Race and social problems*, 1(4):231–242, 2009.

Berk, Richard. Balancing the costs of forecasting errors in parole decisions. *Alb. L. Rev.*, 74:1071, 2010.

Berk, Richard. Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, 27(1):107–123, 2011.

Berk, Richard and Justin Bleich. Forecasts of violence to inform sentencing decisions. *Journal of Quantitative Criminology*, 30(1):79–96, 2014.

Berk, Richard, Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay Ahlman. Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):191–211, 2009.

Berk, Richard A and Justin Bleich. Statistical procedures for forecasting criminal behavior. *Criminology & Public Policy*, 12(3):513–544, 2013.

Berk, Richard A. and Susan D. Sorenson. Machine learning forecasts of domestic violence to help inform release decisions at arraignment. Technical report, University of Pennsylvania, 2014.

Berk, Richard A, Yan He, and Susan B Sorenson. Developing a practical forecasting screener for domestic violence incidents. *Evaluation Review*, 29(4):358–383, 2005.

Berk, Richard A, Brian Kriegler, and Jong-Ho Baek. Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology*, 22(2):131–145, 2006.

Bhati, Avinash Singh. Estimating the number of crimes averted by incapacitation: an information theoretic approach. *Journal of Quantitative Criminology*, 23(4):355–375, 2007.

Bhati, Avinash Singh and Alex R Piquero. Estimating the impact of incarceration on subsequent offending trajectories: Deterrent, criminogenic, or null effect? *The Journal of Criminal Law and Criminology*, pages 207–253, 2007.

Borden, Howard G. Factors for predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, pages 328–336, 1928.

Borum, Randy. *Manual for the structured assessment of violence risk in youth (SAVRY)*. Odessa, Florida: Psychological Assessment Resources, 2006.

Breiman, Leo. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001a.

Breiman, Leo. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001b.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

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

Bushway, Shawn D. Is there any logic to using logit. *Criminology & Public Policy*, 12(3):563–567, 2013.

Bushway, Shawn D and Anne Morrison Piehl. The inextricable link between age and criminal history in sentencing. *Crime & Delinquency*, 53(1):156–183, 2007.

Clements, Carl B. Offender classification two decades of progress. *Criminal Justice and Behavior*, 23(1): 121–143, 1996.

Copas, John and Peter Marshall. The offender group reconviction scale: a statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):159–171, 1998.

Cristianini, Nello and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

Crow, Matthew S. The complexities of prior record, race, ethnicity, and policy: Interactive effects in sentencing. *Criminal Justice Review*, 2008.

Dawes, Robyn M. The robust beauty of improper linear models in decision making. *American psychologist*, 34(7):571–582, 1979.

Dawes, Robyn M, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899): 1668–1674, 1989.

Freitas, Alex A. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, March 2014.

Freund, Yoav and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Friedman, Jerome H. Stochastic gradient boosting. *Computational Statistics &amp; Data Analysis*, 38(4): 367–378, 2002.

Goel, Sharad, Justin M Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *Understanding Racial Disparities in New York City's Stop-and-Frisk Policy (March 2, 2015)*, 2015.

Goh, Siong Thye and Cynthia Rudin. Box drawings for learning with imbalanced data. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

Gottfredson, Don M and Howard N Snyder. The mathematics of risk classification: Changing data into valid instruments for juvenile courts. ncj 209158. *Office of Juvenile Justice and Delinquency Prevention*, 2005.

Grove, William M and Paul E Meehl. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2):293, 1996.

Hahsler, Michael, Christian Buchta, Bettina Gruen, Kurt Hornik, and Christian Borgelt. *Package 'arules': Mining Association Rules and Frequent Itemsets*, December 2014. URL `http://cran.r-project.org/web/packages/arules/arules.pdf`.

Hannah-Moffat, Kelly. Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, 30(2):270–296, 2013.

Hanson, RK and D Thornton. Notes on the development of static-2002. *Ottawa, Ontario: Department of the Solicitor General of Canada*, 2003.

Hesterberg, Tim, Nam Hee Choi, Lukas Meier, and Chris Fraley. Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

Hoffman, Peter B. Twenty years of operational use of a risk prediction instrument: The United States parole commission's salient factor score. *Journal of Criminal Justice*, 22(6):477–494, 1994.

Hoffman, Peter B and Sheldon Adelberg. The salient factor score: A nontechnical overview. *Fed. Probation*, 44:44, 1980.

Holte, Robert C. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.

Holte, Robert C. Elaboration on Two Points Raised in "Classifier Technology and the Illusion of Progress". *Statistical Science*, 21(1):24–26, February 2006.

Howard, Philip, Brian Francis, Keith Soothill, and Leslie Humphreys. OGRS 3: The revised offender group reconviction scale. Technical report, Ministry of Justice, 2009.

Kropp, P Randall and Stephen D Hart. The spousal assault risk assessment (sara) guide: reliability and validity in adult male offenders. *Law and human behavior*, 24(1):101, 2000.

Kuhn, Max and Kjell Johnson. *Applied predictive modeling*. Springer, 2013.

Kuhn, Max, Steve Weston, Nathan Coulter, and R. Quinlan. C50: C5.0 decision trees and rule-based models, 2012. URL `http://CRAN.R-project.org/package=C50`. R package version 0.1.0-013.

Langan, Patrick A and David J Levin. Recidivism of prisoners released in 1994. *Federal Sentencing Reporter*, 15(1):58–65, 2002.

Langton, Calvin M, Howard E Barbaree, Michael C Seto, Edward J Peacock, Leigh Harkins, and Kevin T Hansen. Actuarial assessment of risk for reoffense among adult sex offenders evaluating the predictive accuracy of the static-2002 and five other instruments. *Criminal Justice and Behavior*, 34(1):37–59, 2007.

Liaw, Andy and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, 40(3):203–228, 2000.

Lowenkamp, Christopher T and Edward J Latessa. Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in community corrections*, 2004:3–8, 2004.

Maden, Anthony, Paul Rogers, Andrew Watt, Glyn Lewis, Tim Amos, Kevin Gournay, and P Skapinakis. 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*, 2006.

Maloof, Marcus A. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1, 2003.

McCord, Joan. A thirty-year follow-up of treatment effects. *American psychologist*, 33(3):284, 1978.

McCord, Joan. Cures that harm: Unanticipated outcomes of crime prevention programs. *The Annals of the American Academy of Political and Social Science*, 587(1):16–30, 2003.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-1.

Milgram, Anne. Why smart statistics are the key to fighting crime. Ted Talk, January 2014.

Miller, George. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

Nafekh, Mark and Laurence Louis Motiuk. *The Statistical Information on Recidivism, Revised 1 (SIR-R1) Scale: A Psychometric Examination*. Correctional Service of Canada. Research Branch, 2002.

Netter, Brian. Using group statistics to sentence individual criminals: an ethical and statistical critique of the virginia risk assessment program. *The Journal of Criminal Law and Criminology*, pages 699–729, 2007.

Neuilly, Melanie-Angela, Kristen M Zgoba, George E Tita, and Stephen S Lee. Predicting recidivism in homicide offenders using classification tree analysis. *Homicide studies*, 15(2):154–176, 2011.

Pennsylvania Commission on Sentencing. Risk Assessment Project Interim Report 4: Development of Risk Assessment Scale. 2012.

Petersilia, Joan and Susan Turner. Guideline-based justice: Prediction and racial minorities. *Crime & Justice*, 9:151, 1987.

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

Quinlan, J Ross. *C4. 5: programs for machine learning*. Elsevier, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org`.

Ricardo H. Hinojosa et al. A comparison of the federal sentencing guidelines criminal history category and the U.S. parole commission salient factor score. Technical report, U.S. Sentencing Commission, January 2005.

Ridgeway, Greg. gbm: Generalized boosted regression models. *R package version*, 1(3), 2006.

Ridgeway, Greg. The pitfalls of prediction. *NIJ Journal,* National Institute of Justice, 271:34–40, 2013.

Ritter, Nancy. Predicting recidivism risk: New tool in philadelphia shows great promise. *NIJ Journal*, 271: 4–13, 2013.

Rubin, Paul A. Mixed integer classification problems. In *Encyclopedia of Optimization*, pages 2210–2214. Springer, 2009.

Sherman, Lawrence W. The power few: experimental criminology and the reduction of harm. *Journal of Experimental Criminology*, 3(4):299–321, 2007.

Simon, Jonathan. Reversal of fortune: the resurgence of individual risk assessment in criminal justice. *Annu. Rev. Law Soc. Sci.*, 1:397–421, 2005.

Steinhart, David. Juvenile detention risk assessment: A practice guide to juvenile detention reform. Annie E. Casey Foundation, 2006.

Therneau, Terry, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning*, 2012. URL `http://CRAN.R-project.org/package=rpart`. R package version 4.1-0.

Tibbitts, Clark. Success or failure on parole can be predicted: A study of the records of 3,000 youths paroled from the illinois state reformatory. *Journal of Criminal Law and Criminology (1931-1951)*, pages 11–50, 1931.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Tollenaar, Nikolaj and P.G.M. van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

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

U.S. Department of Justice, Bureau of Justice Statistics. Recidivism of prisoners released in 1994. http://doi.org/10.3886/ICPSR03355.v8, 2014.

U.S. Sentencing Commission. 2012 guidelines manual: Chapter four - criminal history and criminal livelihood, November 1987. URL `http://www.ussc.gov/guidelines-manual/2012/2012-4a11`.

Ustun, Berk. slim_for_matlab v0.1, March 2016. URL `http://dx.doi.org/10.5281/zenodo.47964`.

Ustun, Berk and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, pages 1–43, 2015. ISSN 0885-6125. doi: 10.1007/s10994-015-5528-6. URL `http://dx.doi.org/10.1007/s10994-015-5528-6`.

Wang, Fulton and Cynthia Rudin. Falling rule lists. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2015.

Webster, Christopher D et al. HCR-20: Assessing risk for violence. Technical report, Mental Health, Law, and Policy Institute, Simon Fraser University, in cooperation with the British Columbia Forensic Psychiatric Services Commission, 1997.

Wolfgang, Marvin E. *Delinquency in a birth cohort*. University of Chicago Press, 1987.

Wolsey, Laurence A. *Integer programming*, volume 42. Wiley New York, 1998.

Wroblewski, Jonathan J. Annual letter, U.S. department of justice: Criminal division, July 2014.

Wu, Xindong, Vipin Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey Mclachlan, Angus Ng, Bing Liu, Philip Yu, Zhi-Hua Zhou, Michael Steinbach, David Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2008.

Yang, Min, Stephen CP Wong, and Jeremy Coid. The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological bulletin*, 136(5):740, 2010.

Zadrozny, Bianca and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM, 2002.

Zhang, Yan, Lening Zhang, and Michael S Vaughn. Indeterminate and determinate sentencing models: A state-specific analysis of their effects on recidivism. *Crime & Delinquency*, 2009.

A. Additional Results on Predictive Accuracy

To supplement the experimental results in Section 4.2, we include the training and 5-CV validation results. Table 6 shows the training AUC performance for all methods on all prediction problems, and Table 7 shows the 5-CV validation AUC performance for all methods. A table of test AUC for all methods on all prediction problems can be found in Table 5.

Table 6. Training AUC for all methods on all prediction problems.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	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 7. 5-CV validation AUC for all methods on all prediction problems. We report the 5-CV mean validation AUC. The ranges underneath each cell represent the 5-CV minimum and maximum.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SGB	SLIM
arrest	0.72 0.72 - 0.74	0.73 0.72 - 0.74	0.71 0.71 - 0.73	0.71 0.70 - 0.72	0.67 0.66 - 0.69	0.73 0.72 - 0.74	0.71 0.70 - 0.72	0.73 0.72 - 0.74	0.72 0.71 - 0.73
drug	0.73 0.72 - 0.74	0.73 0.71 - 0.74	0.62 0.61 - 0.64	0.62 0.61 - 0.64	0.59 0.58 - 0.60	0.73 0.72 - 0.74	0.72 0.71 - 0.73	0.74 0.72 - 0.74	0.72 0.71 - 0.73
general_violence	0.71 0.70 - 0.71	0.71 0.70 - 0.71	0.56 0.55 - 0.57	0.57 0.55 - 0.59	0.56 0.55 - 0.58	0.70 0.69 - 0.71	0.69 0.69 - 0.70	0.71 0.70 - 0.71	0.70 0.69 - 0.71
domestic_violence	0.76 0.75 - 0.79	0.76 0.75 - 0.78	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.53 0.51 - 0.54	0.63 0.59 - 0.66	0.76 0.74 - 0.78	0.77 0.75 - 0.79	0.75 0.72 - 0.78
sexual_violence	0.70 0.68 - 0.74	0.69 0.66 - 0.74	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.51 0.50 - 0.51	0.54 0.53 - 0.55	0.67 0.63 - 0.70	0.68 0.65 - 0.72	0.68 0.66 - 0.72
fatal_violence	0.66 0.59 - 0.74	0.67 0.62 - 0.75	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.50 0.50 - 0.52	0.51 0.50 - 0.53	0.67 0.63 - 0.73	0.67 0.61 - 0.74	0.65 0.61 - 0.69

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 shown here correspond to the *best* models we produced using Lasso and Ridge (i.e., the ones that were plotted as points in Figure 1). We omit CART and C5.0 models are shown because all models that were produced were either trivial or 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.

Note that 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 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 5-CV validation TPR/FPR of 82.3%/49.7%.

$$
\begin{array}{rlrlrl}
 & 9.00\,\textit{prior\_arrest\_for\_drugs} & + & 5.00\,\textit{age\_at\_release\_18\_to\_24} & + & 4.00\,\textit{age\_at\_release\_25\_to\_29} \\
+ & 3.00\,\textit{prior\_arrest\_for\_multiple\_types\_of\_crime} & + & 1.00\,\textit{prior\_arrest\_for\_property} & - & 6.00\,\textit{no\_prior\_arrests} \\
- & 1.00\,\textit{age\_at\_release\_30\_to\_39} & - & 7.00 & &
\end{array}
$$

This is the best Lasso model for `drug`. This model has a test TPR/FPR of 82.0%/45.9%, and a mean 5-CV validation TPR/FPR of 81.2%/45.9%.

$$
\begin{array}{rlrlrl}
 & 1.14\,\textit{prior\_arrest\_for\_drugs} & + & 0.27\,\textit{prior\_arrest\_for\_property} & + & 0.26\,\textit{time\_served}{\leq}\textit{6mo} \\
+ & 0.19\,\textit{prior\_arrest\_for\_other\_violence} & + & 0.18\,\textit{prior\_arrest\_for\_multiple\_types\_of\_crime} & + & 0.17\,\textit{prior\_arrest\_for\_misdemeanor} \\
+ & 0.16\,\textit{age\_at\_release\_18\_to\_24} & + & 0.14\,\textit{prior\_arrests}{\geq}\textit{5} & + & 0.13\,\textit{age\_1st\_confinement\_18\_to\_24} \\
+ & 0.12\,\textit{prior\_arrest\_for\_public\_order} & + & 0.10\,\textit{prior\_arrest\_with\_firearms\_involved} & + & 0.08\,\textit{any\_prior\_jail\_time} \\
+ & 0.06\,\textit{age\_1st\_arrest}{\leq}\textit{17} & + & 0.04\,\textit{multiple\_prior\_jail\_time} & + & 0.04\,\textit{drug\_abuse} \\
+ & 0.03\,\textit{multiple\_prior\_prison\_time} & + & 0.03\,\textit{any\_prior\_prb\_or\_fine} & - & 0.62\,\textit{age\_at\_release}{\geq}\textit{40} \\
- & 0.25\,\textit{prior\_arrest\_for\_sexual} & - & 0.23\,\textit{age\_at\_release\_30\_to\_39} & - & 0.12\,\textit{time\_served\_25\_to\_60mo} \\
- & 0.11\,\textit{prior\_arrest\_with\_child\_involved} & - & 0.08\,\textit{alcohol\_abuse} & - & 0.07\,\textit{age\_1st\_confinement}{\geq}\textit{40} \\
- & 1.11 \times 10^{-03}\,\textit{time\_served}{\geq}\textit{61mo} & - & 1.01 & &
\end{array}
$$

This is the best Ridge model for `drug`. This model has a test TPR/FPR of 84.0%/48.2%, and a mean 5-CV validation TPR/FPR of 83.1%/48.4%.

$$
\begin{array}{rlrlrl}
 & 0.91\,\textit{prior\_arrest\_for\_drugs} & + & 0.25\,\textit{time\_served}{\leq}\textit{6mo} & + & 0.24\,\textit{age\_at\_release\_18\_to\_24} \\
+ & 0.21\,\textit{prior\_arrest\_for\_multiple\_types\_of\_crime} & + & 0.20\,\textit{prior\_arrest\_for\_property} & + & 0.17\,\textit{prior\_arrest\_for\_misdemeanor} \\
+ & 0.17\,\textit{prior\_arrest\_for\_other\_violence} & + & 0.17\,\textit{age\_1st\_confinement\_18\_to\_24} & + & 0.14\,\textit{prior\_arrests}{\geq}\textit{5} \\
+ & 0.13\,\textit{prior\_arrest\_with\_firearms\_involved} & + & 0.12\,\textit{age\_at\_release\_25\_to\_29} & + & 0.11\,\textit{drug\_abuse} \\
+ & 0.11\,\textit{prior\_arrest\_for\_public\_order} & + & 0.09\,\textit{age\_1st\_arrest}{\leq}\textit{17} & + & 0.08\,\textit{age\_1st\_confinement}{\leq}\textit{17} \\
+ & 0.08\,\textit{any\_prior\_jail\_time} & + & 0.07\,\textit{multiple\_prior\_jail\_time} & + & 0.07\,\textit{age\_at\_release}{\leq}\textit{17} \\
+ & 0.06\,\textit{multiple\_prior\_prison\_time} & + & 0.06\,\textit{released\_unconditonal} & + & 0.05\,\textit{any\_prior\_prb\_or\_fine} \\
+ & 0.05\,\textit{prior\_arrests}{\geq}\textit{2} & + & 0.04\,\textit{time\_served\_7\_to\_12mo} & + & 0.04\,\textit{multiple\_prior\_prb\_or\_fine} \\
+ & 0.02\,\textit{prior\_arrests}{\geq}\textit{1} & + & 0.01\,\textit{age\_1st\_confinement\_25\_to\_29} & + & 0.01\,\textit{released\_conditonal} \\
+ & 2.52 \times 10^{-03}\,\textit{prior\_arrest\_for\_felony} & + & 1.76 \times 10^{-03}\,\textit{age\_1st\_arrest\_18\_to\_24} & + & 9.58 \times 10^{-04}\,\textit{prior\_arrest\_for\_fatal\_violence} \\
- & 0.33\,\textit{age\_at\_release}{\geq}\textit{40} & - & 0.25\,\textit{prior\_arrest\_for\_sexual} & - & 0.19\,\textit{age\_1st\_confinement}{\geq}\textit{40} \\
- & 0.16\,\textit{prior\_arrest\_with\_child\_involved} & - & 0.15\,\textit{time\_served\_25\_to\_60mo} & - & 0.14\,\textit{alcohol\_abuse} \\
- & 0.13\,\textit{time\_served}{\geq}\textit{61mo} & - & 0.10\,\textit{prior\_arrest\_for\_domestic\_violence} & - & 0.09\,\textit{age\_at\_release\_30\_to\_39} \\
- & 0.05\,\textit{age\_1st\_arrest}{\geq}\textit{40} & - & 0.04\,\textit{female} & - & 0.04\,\textit{infraction\_in\_prison} \\
- & 0.03\,\textit{age\_1st\_arrest\_30\_to\_39} & - & 0.02\,\textit{age\_1st\_confinement\_30\_to\_39} & - & 0.02\,\textit{no\_prior\_arrests} \\
- & 4.71 \times 10^{-03}\,\textit{prior\_arrest\_for\_local\_ord} & - & 4.45 \times 10^{-03}\,\textit{time\_served\_13\_to\_24mo} & - & 2.23 \times 10^{-03}\,\textit{age\_1st\_arrest\_25\_to\_29} \\
- & 1.09 & & & &
\end{array}
$$

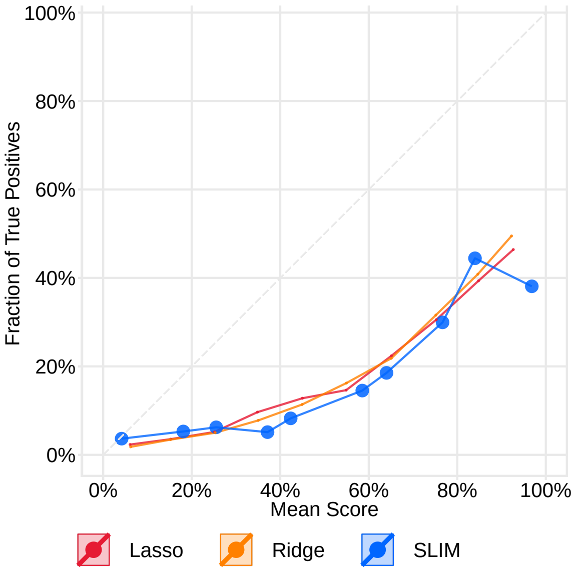

Fig. 11. Risk calibration plot for `drug`.

B.2. `general_violence`

SLIM model for `general_violence`. This model has a test TPR/FPR of 76.7%/45.4%, and a mean 5-CV validation 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 5-CV validation TPR/FPR of 77.3%/45.7%.

	0.90 *prior_arrest_for_other_violence*	+	0.35 *prior_arrest_for_property*	+	0.28 *prior_arrest_for_misdemeanor*
+	0.28 *age_at_release_18_to_24*	+	0.24 *prior_arrest_for_public_order*	+	0.20 *age_1st_arrest≤17*
+	0.20 *released_unconditonal*	+	0.17 *age_1st_confinement_18_to_24*	+	0.16 *alcohol_abuse*
+	0.14 *prior_arrest_for_fatal_violence*	+	0.14 *age_1st_confinement≤17*	+	0.10 *prior_arrest_for_felony*
+	0.10 *prior_arrests≥5*	+	0.10 *prior_arrest_with_firearms_involved*	+	0.10 *age_1st_arrest_18_to_24*
+	0.09 *infraction_in_prison*	+	0.04 *time_served≤6mo*	+	0.03 *time_served_7_to_12mo*
+	2.89×10^{-03} *prior_arrest_for_drugs*	−	0.72 *age_at_release≥40*	−	0.41 *female*
−	0.27 *age_at_release_30_to_39*	−	0.15 *prior_arrest_with_child_involved*	−	0.07 *age_1st_confinement≥40*
−	0.05 *age_1st_arrest≥40*	−	0.01 *time_served_25_to_60mo*	−	1.84×10^{-03} *age_1st_confinement_30_to_39*
−	1.19				

This is the best Ridge model for `general_violence`. This model has a test TPR/FPR of 81.4%/48.1%, and a mean 5-CV validation TPR/FPR of 80.0%/48.5%.

	0.62 *prior_arrest_for_other_violence*	+	0.27 *age_at_release_18_to_24*	+	0.24 *prior_arrest_for_property*
+	0.23 *prior_arrest_for_misdemeanor*	+	0.19 *age_1st_confinement_18_to_24*	+	0.18 *prior_arrest_for_public_order*
+	0.17 *age_1st_arrest≤17*	+	0.14 *prior_arrest_for_multiple_types_of_crime*	+	0.13 *released_unconditonal*
+	0.13 *prior_arrests≥5*	+	0.13 *prior_arrest_for_felony*	+	0.12 *prior_arrest_with_firearms_involved*
+	0.11 *age_1st_confinement≤17*	+	0.11 *alcohol_abuse*	+	0.10 *age_at_release_25_to_29*
+	0.10 *prior_arrest_for_fatal_violence*	+	0.09 *infraction_in_prison*	+	0.08 *age_1st_arrest_18_to_24*
+	0.07 *prior_arrest_for_domestic_violence*	+	0.05 *drug_abuse*	+	0.05 *time_served≤6mo*
+	0.05 *prior_arrest_for_local_ord*	+	0.04 *time_served_7_to_12mo*	+	0.04 *age_at_release≤17*
+	0.03 *prior_arrests≥2*	+	0.03 *multiple_prior_prb_or_fine*	+	0.02 *multiple_prior_jail_time*
+	0.01 *prior_arrest_for_drugs*	+	3.41×10^{-03} *no_prior_arrests*	−	0.32 *age_at_release≥40*
−	0.20 *female*	−	0.18 *age_1st_confinement≥40*	−	0.12 *prior_arrest_with_child_involved*
−	0.12 *age_1st_arrest≥40*	−	0.11 *age_1st_arrest_30_to_39*	−	0.09 *age_1st_confinement_30_to_39*
−	0.08 *age_at_release_30_to_39*	−	0.05 *age_1st_arrest_25_to_29*	−	0.04 *prior_arrest_for_sexual*
−	0.04 *time_served_25_to_60mo*	−	0.03 *time_served≥61mo*	−	0.03 *released_conditonal*
−	0.03 *age_1st_confinement_25_to_29*	−	0.02 *any_prior_prb_or_fine*	−	0.02 *time_served_13_to_24mo*
−	5.89×10^{-03} *multiple_prior_prison_time*	−	3.60×10^{-03} *any_prior_jail_time*	−	3.47×10^{-03} *prior_arrests≥1*
−	1.13				

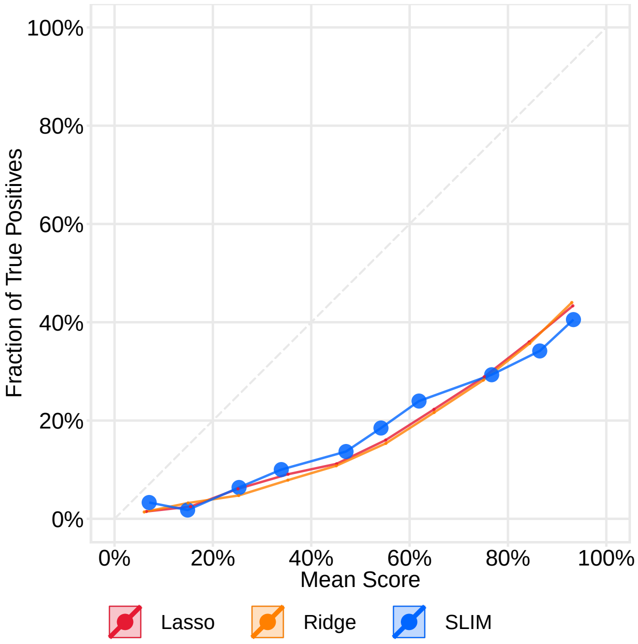

Fig. 12. Risk calibration plot for `general_violence`.

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 5-CV validation TPR/FPR of 81.4%/48.0%.

4 prior_arrest_for_misdemeanor + 3 prior_arrest_for_felony + 2 prior_arrest_for_domestic_violence

+ age_1st_confinement_18_to_24 − 5 infraction_in_prison − 3

This is the best Lasso model for domestic_violence. This model has a test TPR/FPR of 87.0%/45.8%, and a mean 5-CV validation TPR/FPR of 84.5%/45.8%.

0.88 prior_arrest_for_misdemeanor + 0.73 prior_arrest_for_domestic_violence + 0.73 prior_arrest_for_felony

+ 0.66 prior_arrest_for_other_violence + 0.54 released_unconditonal + 0.32 age_1st_confinement_18_to_24

+ 0.24 multiple_prior_prb_or_fine + 0.21 alcohol_abuse + 0.17 prior_arrest_for_sexual

+ 0.16 prior_arrests≥5 + 0.16 prior_arrest_with_firearms_involved + 0.08 age_at_release_18_to_24

+ 0.06 no_prior_arrests + 0.05 time_served_7_to_12mo + 0.03 prior_arrest_for_property

+ 0.01 age_1st_arrest_18_to_24 + 0.01 prior_arrest_for_public_order − 1.09 infraction_in_prison

− 0.54 age_at_release≥40 − 0.47 drug_abuse − 0.40 multiple_prior_prison_time

− 0.31 prior_arrest_with_child_involved − 0.28 multiple_prior_jail_time − 0.26 female

− 0.20 age_1st_confinement≥40 − 0.16 any_prior_jail_time − 0.07 age_1st_arrest_30_to_39

− 0.07 any_prior_prb_or_fine − 0.06 prior_arrest_for_drugs − 0.06 time_served≥61mo

− 4.48×10^{-04} time_served_25_to_60mo − 1.04

This is the best Ridge model for domestic_violence. This model has a test TPR/FPR of 87.0%/47.7%, and a mean 5-CV validation TPR/FPR of 85.2%/47.5%.

0.76 prior_arrest_for_misdemeanor + 0.59 prior_arrest_for_other_violence + 0.57 prior_arrest_for_domestic_violence

+ 0.54 prior_arrest_for_felony + 0.40 released_unconditonal + 0.27 age_1st_confinement_18_to_24

+ 0.27 multiple_prior_prb_or_fine + 0.21 prior_arrest_for_sexual + 0.19 prior_arrest_with_firearms_involved

+ 0.18 alcohol_abuse + 0.18 prior_arrests≥5 + 0.17 age_at_release_18_to_24

+ 0.15 prior_arrest_for_local_ord + 0.12 age_at_release_25_to_29 + 0.11 time_served_7_to_12mo

+ 0.10 prior_arrest_for_property + 0.10 prior_arrest_for_fatal_violence + 0.10 no_prior_arrests

+ 0.08 age_at_release_30_to_39 + 0.07 prior_arrest_for_multiple_types_of_crime + 0.07 age_1st_arrest≤17

+ 0.07 age_1st_arrest_18_to_24 + 0.07 prior_arrest_for_public_order + 0.05 age_1st_arrest_25_to_29

+ 0.05 time_served≤6mo + 0.05 time_served_13_to_24mo + 0.05 prior_arrests≥2

+ 3.08×10^{-03} age_1st_confinement_30_to_39 − 0.86 infraction_in_prison − 0.40 drug_abuse

− 0.39 multiple_prior_prison_time − 0.36 age_at_release≥40 − 0.26 prior_arrest_with_child_involved

− 0.25 multiple_prior_jail_time − 0.25 female − 0.24 age_1st_confinement≥40

− 0.19 any_prior_jail_time − 0.14 time_served≥61mo − 0.12 age_1st_arrest_30_to_39

− 0.10 any_prior_prb_or_fine − 0.10 age_1st_arrest≥40 − 0.10 prior_arrests≥1

− 0.08 prior_arrest_for_drugs − 0.06 age_1st_confinement_25_to_29 − 0.05 time_served_25_to_60mo

− 0.04 released_conditonal − 0.04 age_at_release≤17 − 0.02 age_1st_confinement≤17

− 1.01

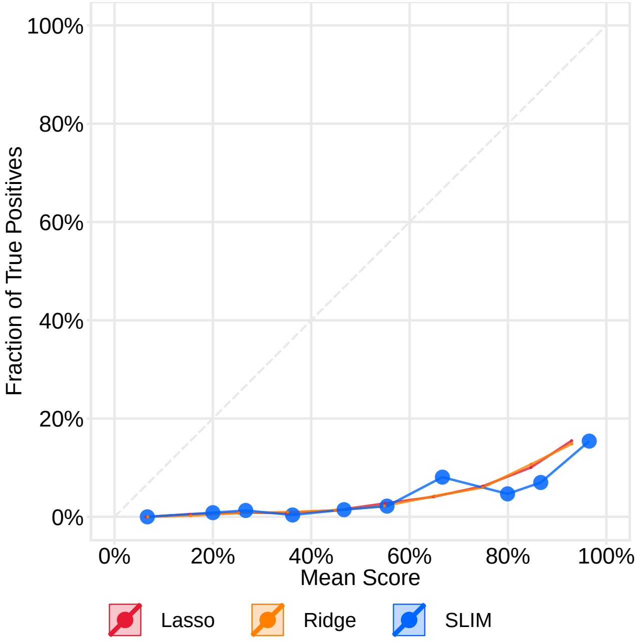

Fig. 13. Risk calibration plot for domestic_violence.

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 5-CV validation TPR/FPR of 43.7%/19.9%.

	3 *prior_arrest_for_sexual*	+	*prior_arrests*$\geq$*5*	+	*multiple_prior_jail_time*
−	2 *no_prior_arrests*	−	*prior_arrest_for_multiple_types_of_crime*	−	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 5-CV validation TPR/FPR of 43.7%/17.9%.

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

This is the best Ridge model for `sexual_violence`. This model has a test TPR/FPR of 48.6%/19.3%, and a mean 5-CV validation TPR/FPR of 44.9%/19.4%.

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

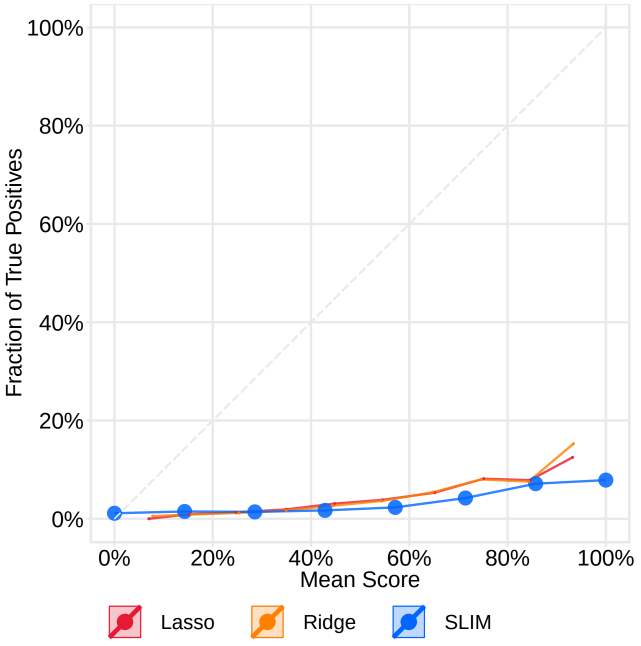

Fig. 14. Risk calibration plot for `sexual_violence`.

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 5-CV validation TPR/FPR of 64.2%/42.4%.

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

This is the best Lasso model for `fatal_violence`. This model has a test TPR/FPR of 68.9%/44.5%, and a mean 5-CV validation TPR/FPR of 67.6%/42.4%.

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

This is the best Ridge model for `fatal_violence`. This model has a test TPR/FPR of 62.2%/34.0%, and a mean 5-CV validation TPR/FPR of 60.1%/33.0%.

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

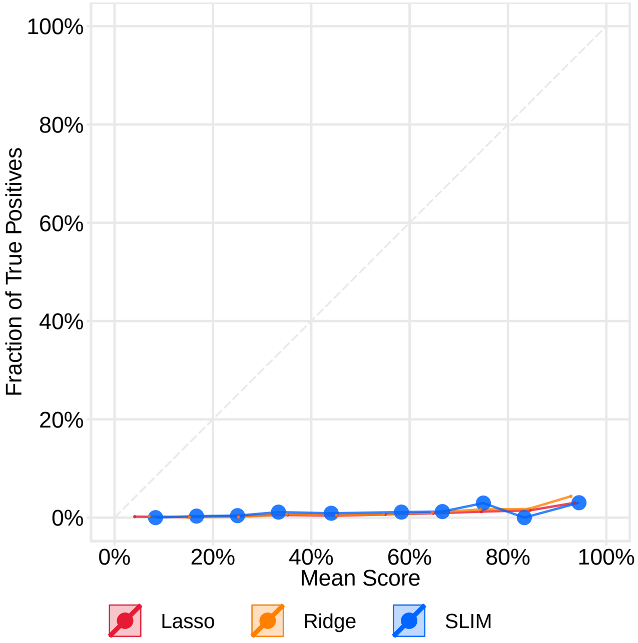

Fig. 15. Risk calibration plot for `fatal_violence`.

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 8 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, model with at most 8 non-zero coefficients, 8 leaves or 8 rules).

Table 8 shows the test AUC of each method when they are used to fit a models with a model size of 8 or less. Trivial models of size 1 are also omitted. Table 9 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 are unable to produce a suitably sparse model for some of the problems since their implementation does not provide control over model sparsity. Note that we have omitted results for Ridge because it could not produce a model with fewer than 8 coefficients for all prediction problems (see Section 4.4 for explanation).

Table 8. Test AUC on all prediction problems when transparent methods are restricted to models with at most 8 coefficients, 8 leaves or 8 rules.

Prediction Problem	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 9. Percentage in test AUC with respect to SLIM's model on all prediction problems when transparent methods are restricted to models with at most 8 coefficients, 8 leaves or 8 rules.

Prediction Problem	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%

D. Trade-off between Risk Calibration and Interpretability

Figure 16 shows the risk calibration plots of Lasso, Ridge, and SLIM for transparent models with model size constrained to 8 or less, chosen under the same decision criteria as Appendix B. Ridge is not included because no such models are achievable, as discussed also in Appendix C. For Lasso, the risk calibration performance is worse in comparison to Figures 11–15. For fatal_violence, there was no Lasso model available at the desired decision point.

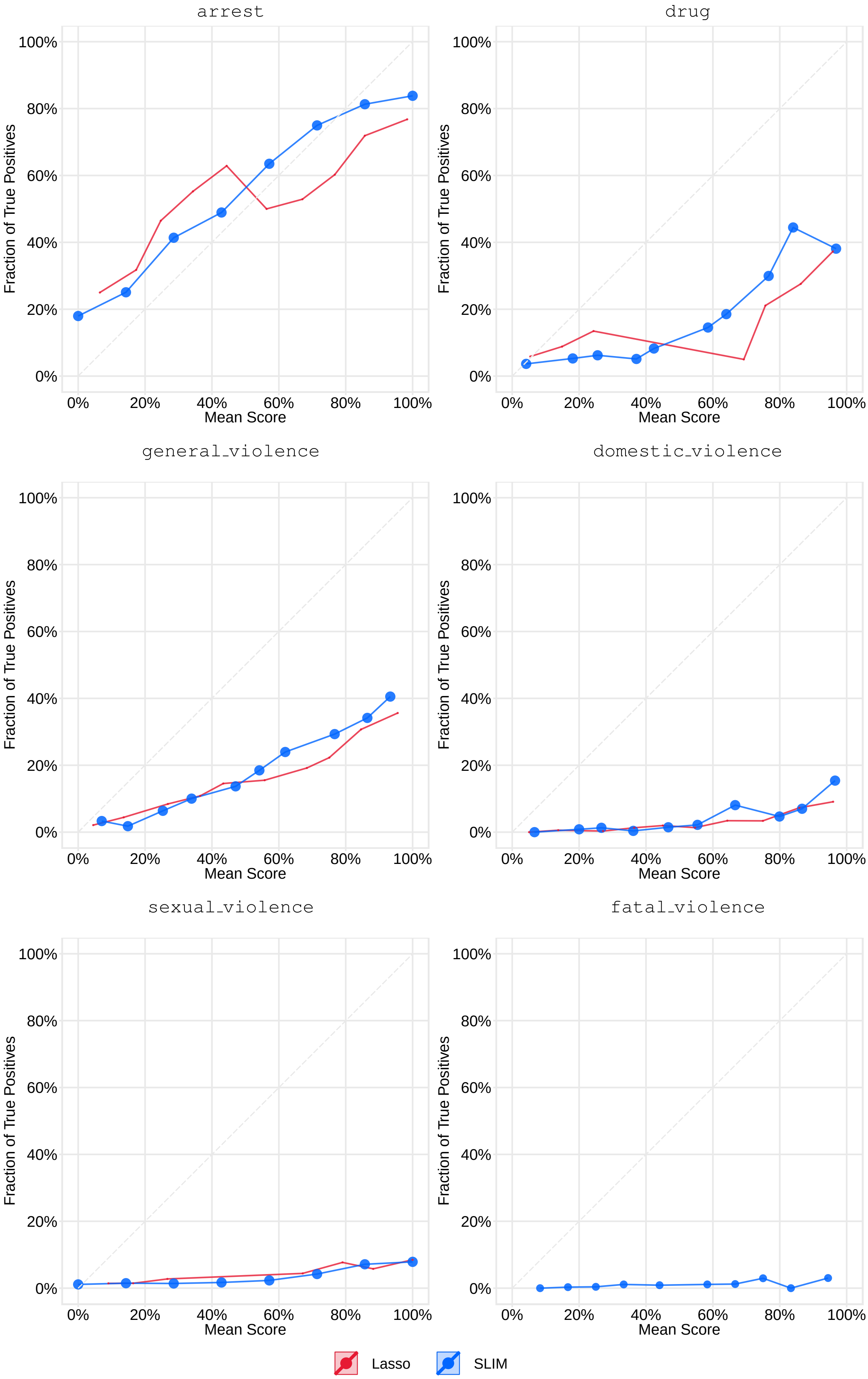

Fig. 16. Risk calibration plots for transparent models with model size constrained to 8 or less.

E. On the Predictive Accuracy of Baseline Methods with Continuous Input Variables

In our experiments in Section 4, we ran all methods with a dataset composed exclusively of binary input variables. That is, for each feature in the original database (e.g., *prior_arrests*), we derived binary variables (e.g., *no_prior_arrests*, *prior_arrests* $\geq$ *1* and so on) and trained each method using these binary variables. It is possible that machine learning methods could potentially be hindered by this removal of information. Here, we investigate how the predictive accuracy of the baseline method would have been affected had we run these methods using continuous input variables (Appendix E.1) or both binary and continuous input variables (Appendix E.2). In both cases, we find that the change in variable encoding results in a minor difference in performance.

E.1. Change in Predictive Accuracy using Only Continuous Input Variables

Instead of using 48 input variables, we now have 25 continuous variables. Table 10 summarizes the test AUC for all methods on all prediction problems when we use only continuous input variables. Table 11 shows the percentage change in test AUC due to this change in encoding (i.e. from binary input variables to continuous input variables). The largest increases in predictive accuracy are 4.6% for CART and 7.7% for SVM RBF, while the biggest decrease in accuracy is -19.6% for RF.

Our results suggest that there is no uniform gain/loss in performance for most of the methods: for any given method, the test AUC increased slightly for at least one problem, and decreased slightly for at least another. Among the methods, CART saw the most uniform improvement, performing slightly better on 5 out of the 6 problems when continuous variables are used (though CART still performs poorly compared to other methods).

Table 10. Test AUC for all methods on all datasets when features are encoded as continuous variables.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SGB
arrest	0.74	0.73	0.72	0.72	0.70	0.75	0.74	0.75
drug	0.74	0.74	0.65	0.66	0.62	0.75	0.74	0.76
general_violence	0.71	0.70	0.54	0.58	0.55	0.69	0.69	0.71
domestic_violence	0.74	0.70	0.50	0.50	0.54	0.51	0.75	0.77
sexual_violence	0.70	0.68	0.50	0.50	0.52	0.51	0.68	0.71
fatal_violence	0.69	0.68	0.50	0.50	0.51	0.50	0.74	0.72

Table 11. Percentage change in test AUC for all methods on all datasets when features are encoded as continuous variables instead of binary variables.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SGB
arrest	1.7%	0.1%	0.0%	0.1%	2.4%	2.7%	3.0%	1.7%
drug	-0.5%	-0.3%	2.5%	4.2%	4.6%	0.6%	0.7%	1.9%
general_violence	-1.5%	-2.6%	-4.1%	0.7%	-1.3%	-2.7%	-2.2%	-1.0%
domestic_violence	-3.9%	-8.7%	-0.1%	-0.1%	1.6%	-19.6%	-3.1%	-0.8%
sexual_violence	-1.5%	-5.1%	0.0%	0.0%	2.0%	-5.3%	-2.7%	0.9%
fatal_violence	2.8%	0.1%	0.0%	0.0%	1.0%	0.3%	7.7%	2.9%

E.2. Change in Predictive Accuracy using Both Binary and Continuous Input Variables

Instead of the original 48 variables, we now use a combination of 66 binary and continuous variables. Table 12 summarizes the test AUC for all methods on all prediction problems when we used both binary and continuous input variables. Table 13 shows the percentage change in test AUC due to this change in encoding (i.e., from binary input variables to both binary and continuous input variables). Most methods saw a slight AUC increase due to the addition of continuous variables, ranging from 0.2–6.3%. The most significant increases are 3.3% for CART and 6.3% for C5.0T, while the largest decrease is -16.0% for RF. In addition to RF, Ridge and SVM RBF all saw slight decreases with the inclusion. Similar to Appendix E.1, no uniform gain/loss in performance is seen.

Table 12. Test AUC for models created using both continuous and binary variables.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SGB
arrest	0.74	0.73	0.72	0.72	0.69	0.75	0.73	0.75
drug	0.75	0.74	0.65	0.67	0.61	0.76	0.75	0.76
general_violence	0.72	0.71	0.58	0.58	0.56	0.72	0.71	0.73
domestic_violence	0.75	0.71	0.50	0.50	0.54	0.54	0.77	0.78
sexual_violence	0.71	0.69	0.50	0.50	0.52	0.50	0.71	0.71
fatal_violence	0.69	0.68	0.50	0.50	0.50	0.50	0.70	0.72

Table 13. Percentage difference of test AUC for models created with both continuous and binary variables verses test AUC for models created with just binary variables.

Prediction Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SGB
arrest	2.4%	0.6%	0.7%	0.2%	1.9%	2.7%	1.5%	1.7%
drug	0.2%	0.2%	2.1%	6.3%	3.3%	1.9%	2.1%	2.1%
general_violence	0.5%	-1.7%	2.5%	0.7%	-0.2%	0.4%	0.7%	1.2%
domestic_violence	-2.2%	-7.4%	0.0%	0.0%	1.3%	-16.0%	-0.5%	0.4%
sexual_violence	-0.4%	-4.2%	0.0%	0.0%	1.5%	-6.4%	1.9%	0.6%
fatal_violence	2.2%	0.3%	0.0%	0.0%	-0.3%	0.3%	2.7%	2.8%

F. Association Rules

We produce insights more extensive than those in Section 2.4 by mining *association rules*. Association rules, also known as "IF-THEN" rules, are small predictive models that can be produced using search techniques or optimization techniques.

F.1. Terminology

High quality association rules are characterized by large values of *support*, *confidence*, and *lift*. To define this terminology, consider a rule such as "IF a THEN b." We denote this rule also as $a \to b$. The *support* of $a \to b$ is the empirical probability $\hat{P}(a \text{ and } b)$, that is, the proportion of observations where the conditions a and b are both satisfied. The *confidence* of $a \to b$ is the empirical probability $\hat{P}(b|a)$, that is, the proportion of observations for which condition b is satisfied given a is satisfied. The *lift* of $a \to b$ is the ratio $\frac{\hat{P}(b|a)}{\hat{P}(b)}$. Lift measures the ability of condition a to "target" the population where condition b is satisfied: if the lift of $a \to b$ is equal to 1, then outcome b could be predicted equally well if we had assumed that a and b were independent; if the lift of $a \to b$ greater than 1 then event a has some effect on predicting event b.

To illustrate these concepts, consider the following association rule:

IF *age_at_release_18_to_24* AND *prior_arrests≥5* THEN $y = +1$.

The support of this rule is 0.07, which means that 7% of prisoners were released from prison between the ages of 18 to 24, had at least 5 prior arrests, and were arrested within 3 years of being released from prison. The confidence of this rule is 0.83, which means that if a prisoner was released from prison between the ages of 18 to 24 and had at least 5 prior arrests, then there was an 83% chance that this person would be arrested within 3 years of being released from prison. Lastly, the lift of this rule is 1.41, which means that prisoners released from prison between the ages of 18 to 24 and had at least 5 prior arrests have a higher chance of being rearrested than other prisoners, i.e., the prisoners age at release and arrest history makes the conditional probability of arrest 1.41 times higher than if arrest was independent of these conditions.

F.2. Rule Mining

We list 24 interesting association rules for the `arrest` problem in Table 14. These rules were generated with the apriori method in the **arules** package in R 3.1.1 (Hahsler et al., 2014). Note that the choice of package does not matter, as mining rules through search techniques is deterministic, so all packages produce the same rules.

Here, the IF conditions are formulated using combinations of input variables (i.e. $x_j = 1$ and $x_k = 1$) and the THEN condition is that a prisoner is arrested within 3 years of being released from prison (i.e. a positive outcome $y = +1$). The rules in Table 14 have the highest levels of lift and confidence with a minimum support of 5% (i.e., the rule applied to at least 1690 of the 33796 prisoners in our dataset). This threshold value was chosen so the rules do not reflect spurious correlations. Rules A – E were produced by mining the most powerful single-variable predictors for `arrest`. Rules A – E attain the highest lift among one-variable rules with a support of at least 5% and a confidence of at least 0.70. Rules F – X were produced by mining two-variable rules that use at least one of the input variables from Rules A – E that attain the highest possible lift, as well as support at least 5% and confidence at least 0.75. Out of all these rules, Rule F performs the best with a confidence of 0.83 and a lift of 1.41. As it turns out, Rule F is often exploited by some of the best models we find for `arrest`, as we often find patterns similar to "*age_at_release_18_to_24* AND *prior_arrests≥5*" in our predictive models (see e.g., Figure 5 in Section 4.4).

Interesting observations can also be made from the discovered rules. Recall that jail is a much less severe punishment than prison. Considering Rule L and Rule M in Table 14, we can see that prisoners with multiple jail time and have *any* past probations or fines are just as likely to be arrested as those with multiple jail time and multiple prior prison records – despite *multiple_prior_prison_time* being a indicator of much more severe past actions than *any_prior_probation_or_fine*.

F.3. Falling Rule Lists for Imbalanced Problems

As we discuss in Section 4.2, it is difficult to use traditional tree and rule-based methods to create non-trivial models on imbalanced classification problems such as `sexual_violence`. This is possibly because these

Table 14. IF-THEN rules mined for `arrest`. The THEN condition for each rules is the outcome $y = +1$, which indicates that a prisoner is arrested within 3 years of being released from prison.

Rule	IF Condition	Lift	Support	Confidence
A	*multiple_prior_jail_time*	1.24	0.21	0.73
B	*age_1st_arrest≤17*	1.23	0.10	0.73
C	*multiple_prior_probation_or_fine*	1.20	0.16	0.71
D	*age_at_release_18_to_24*	1.20	0.14	0.71
E	*prior_arrests≥5*	1.19	0.42	0.70
F	*age_at_release_18_to_24* AND *prior_arrests≥5*	1.41	0.07	0.83
G	*multiple_prior_jail_time* AND *multiple_prior_probation_or_fine*	1.30	0.08	0.77
H	*age_1st_arrest≤17* AND *prior_arrests≥5*	1.28	0.08	0.76
I	*multiple_prior_jail_time* AND *time_served≤6mo*	1.34	0.06	0.79
J	*multiple_prior_jail_time* AND *age_1st_confinement_18_to_24*	1.29	0.12	0.76
K	*multiple_prior_jail_time* AND *prior_arrest_for_misdemeanor*	1.28	0.15	0.76
L	*multiple_prior_jail_time* AND *multiple_prior_prison_time*	1.28	0.13	0.75
M	*multiple_prior_jail_time* AND *any_prior_probation_or_fine*	1.27	0.13	0.75
N	*age_1st_arrest≤17* AND *prior_arrest_for_misdemeanor*	1.32	0.07	0.78
O	*age_1st_arrest≤17* AND *any_prior_jail_time*	1.28	0.06	0.76
P	*age_1st_arrest≤17* AND *age_1st_confinement_18_to_24*	1.28	0.05	0.75
Q	*multiple_prior_probation_or_fine* AND *age_1st_confinement_18_to_24*	1.31	0.08	0.77
R	*age_at_release_18_to_24* AND *prior_arrest_for_misdemeanor*	1.34	0.06	0.79
S	*age_at_release_18_to_24* AND *any_prior_jail_time*	1.34	0.06	0.79
T	*age_at_release_18_to_24* AND *prior_arrests≥2*	1.32	0.10	0.78
U	*age_at_release_18_to_24* AND *prior_arrest_for_multiple_types*	1.30	0.10	0.76
V	*prior_arrests≥5* AND *age_at_release_25_to_29*	1.31	0.10	0.77
W	*prior_arrests≥5* AND *age_1st_confinement_18_to_24*	1.28	0.21	0.76
X	*prior_arrests≥5* AND *time_served≤6mo*	1.28	0.11	0.76

algorithms employ greedy splitting and pruning procedures. Here, we aim to show that there exist rule-based models that perform well on such problems by training *Falling Rule Lists* (Wang and Rudin, 2015).

Falling Rule Lists are ordered lists of IF-THEN rules. The confidence of each rule decreases as we go down the list. In this way, the highest rule applies to the group of individuals that have the highest risk, the second highest rule applies to a group of individuals with the second highest risk, and so on. The algorithm that produces Falling Rule List globally optimizes the list, without greedy splitting and pruning.

We present a Falling Rule List for the `arrest` problem in Table 15, learned from the algorithm of Wang and Rudin (2015). This model was trained using rules with at most two input variables and a support of at least 5%. The rules listed within this model have the form "IF a THEN b" where b denotes a positive outcome $y = +1$. In Table 15, support refers to the percentage of remaining examples that satisfy the IF conditions and *probability* refers to percentage of these examples where the outcome variable is positive. This model shows that the highest risk prisoners are those who were released between ages 18 and 24, and who have at least 5 prior arrests – this is aligned with the association rule (Rule F) that we found in Section F.2. Once those individuals are removed, the second highest risk prisoners are 25–29 year olds with at least 5 prior arrests, etc. The risk of each group decreases as one moves down the rules. Rule 15 represents the default rule. If an individual does not fall under any of risk groups determined by Rules 1-14, then his/her risk of arrest is 0.21.

Table 15. Falling rule list for `arrest`.

Conditions			Probability	Support
IF	*age_at_release_18_to_24*	AND *prior_arrests≥5*	0.83	0.08
ELSE IF	*age_at_releaser_25_to_29*	AND *prior_arrests≥5*	0.77	0.13
ELSE IF	*multiple_prior_jail_time*	AND *prior_arrests_for_drugs*	0.73	0.18
ELSE IF	*age_at_release_30_to_39*	AND *prior_arrests≥5*	0.67	0.26
ELSE IF	*age_at_release_18_to_24*	AND *prior_arrests≥1*	0.66	0.16
ELSE IF	*prior_arrests_for_drugs*	AND *prior_arrests_for_misdemeanor*	0.55	0.29
ELSE IF	*age_at_release_25_to_29*	AND *prior_arrests≥2*	0.54	0.17
ELSE IF	*multiple_prior_jail_time*	AND *prior_arrests≥5*	0.54	0.27
ELSE IF	*age_1st_arrest≤17*		0.53	0.14
ELSE IF	*age_at_release_18_to_24*		0.50	0.19
ELSE IF	*time_served≤6mo*	AND *prior_arrests_for_property*	0.48	0.17
ELSE IF	*prior_arrests≥5*	AND *prior_arrests≥1*	0.41	0.60
ELSE IF	*age_at_release_25_to_29*	AND *age_1st_arrest_18_to_24*	0.41	0.16
ELSE IF	*age_at_release_30_to_39*	AND *prior_arrests≥1*	0.37	0.35
ELSE	default		0.21	

G. The Impact of Race

As discussed earlier, we chose not to include race as an input variable in our prediction problems. Some studies have shown that race is important for accurate recidivism prediction (Petersilia and Turner, 1987; Berk, 2009).

We wanted to know the answers to two questions. First, whether including race as a feature would lead to more accurate predictions. Second, whether we could predict race from the features that we already had. If we could predict race well from our current set of features, this would show that race information could be implicitly included in any model we might construct. The results that follow show: (i) including race does not substantially increase prediction accuracy for our problems, and (ii) race can be predicted fairly well from the features we already have. These results indicate that most of the information necessary to predict recidivism is already included in the features we have, and these features also include relevant information for predicting race.

To address whether race provided an increase in accuracy for predicting recidivism, we re-ran all methods other than SLIM on all new versions of each prediction problem that included three additional race-related input variables: *white*, *black*, *hispanic*. An overview of these variables can be seen in Table 16. Table 17 presents the models' test AUC when race-related indicator variables are included. Table 18 represent the percentage increase in AUC when compared to 5. As shown, the differences for most methods are negligible. In the cases of SVM RBF and Ridge, the accuracy increased slightly. In the case of RF, including race decreases accuracy (most likely because it exacerbates the overfitting problem).

To determine whether race could be predicted from the current variables, we used three different race options (*white*, *black*, and *hispanic*) as outcomes and predicted each race as a function of our features. ROC plots are provided in Figure 17, showing that race can be predicted much better than random guessing. This is not a surprise, as we already know that blacks tend to have longer criminal histories than whites. On the other hand, we remark that we could not predict race perfectly with the features we have - in fact, our predictions (for all methods) were far from perfect. This means that not all of the information about race is contained in the features we have.

Table 16. Overview of race-related input variables, in addition to the variables in Table 1. Each variable is a binary rule of the form $x_{ij} \in \{0,1\}$.

Input Variable	$\mathbf{P}(x_{ij} = 1)$	**Definition**
white	0.53	prisoner i is white
black	0.44	prisoner i is black
hispanic	0.14	prisoner i is hispanic

Table 17. Test AUC for the baseline methods on all prediction problems using the standard set of input variables along with the race-related indicator variables *white*, *black* and *hispanic*.

Dataset	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	Boosting
arrest	0.73	0.74	0.72	0.71	0.69	0.74	0.72	0.74
drug	0.75	0.75	0.64	0.65	0.59	0.76	0.74	0.76
general_violence	0.73	0.73	0.56	0.58	0.56	0.72	0.71	0.72
domestic_violence	0.77	0.77	0.50	0.50	0.52	0.65	0.77	0.78
sexual_violence	0.72	0.72	0.50	0.50	0.51	0.55	0.70	0.70
fatal_violence	0.68	0.69	0.50	0.50	0.50	0.50	0.69	0.70

Table 18. Percentage difference of test AUC for models with the inclusion of race-related indicator variables such as *white*, *black* and *hispanic* verses test AUC for models created without.

Dataset	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	Boosting
arrest	1.2%	0.9%	-0.2%	-0.1%	1.4%	0.6%	-0.9%	0.3%
drug	0.9%	1.2%	0.5%	3.8%	0.3%	1.3%	0.9%	0.8%
general_violence	0.9%	1.1%	-0.7%	1.5%	1.1%	0.6%	1.7%	0.6%
domestic_violence	0.0%	-0.1%	0.0%	0.0%	-1.0%	1.1%	-0.1%	-0.1%
sexual_violence	0.2%	0.2%	0.0%	0.0%	-0.7%	1.2%	0.4%	0.0%
fatal_violence	1.4%	1.4%	0.0%	0.0%	-0.0%	-0.3%	-1.0%	0.8%

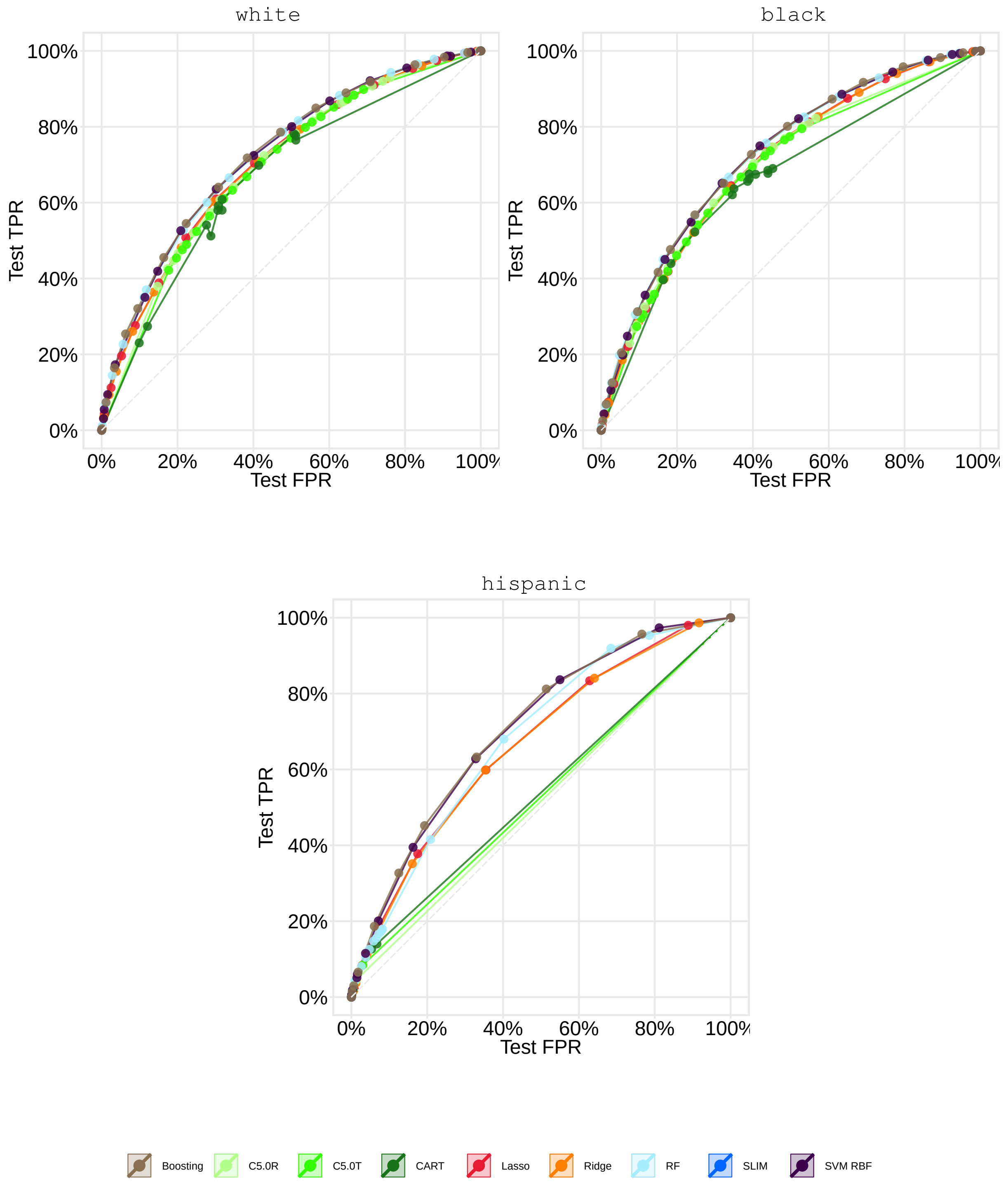

Fig. 17. ROC curves for predicting *white*, *black* and *hispanic* using the standard set of input variables.