Interpretable Classification Models for Recidivism Prediction

May 26, 2015

Abstract

We investigate a long-debated question of how to create predictive recidivism models that are sufficiently accurate, transparent, and interpretable to use for decision-making. These models are used to support decisions from sentencing, determining release on probation, to allocating preventative social services. 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, Ridge Regression) produce equally accurate models along the full ROC curve. However, methods 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 new method known as SLIM (Supersparse Linear Integer Models) to produce transparent and interpretable models along the full ROC curve. These models can be used for decision-making for many different cases, since they rival the most powerful black-box machine learning models in accuracy, but are more transparent and highly interpretable.

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 instance, 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. Clearly, we want to ensure that public resources are directed towards the right individuals, whether these resources are 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 (Milgram 2014). 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 tools (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) such as those being developed recently by (Turner et al. 2009).

There has been some controversy as to whether sophisticated machine learning methods (such as random forests, 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 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 a "black-box," where it is difficult to understand how the input variables are combined to produce a predicted outcome. If a simpler, transparent, but equally accurate predictive model could be developed, it would be more usable and defensible for all different decision-making applications. There is a clear precedent for using such models in criminology (Steinhart 2006, Andrade 2009) where some have argued that a "decent transparent model that is actually used will outperform a sophisticated system that predicts better but sits on a shelf" (Ridgeway 2013). This discussion is captured nicely by Bushway (2013), who described a discrepancy between papers 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 these methods are not necessary and that logistic regression is sufficient. In this work, we show 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 are going to 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 the 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 will show in this paper, for a wide range of recidivism prediction problems, many machine learning methods (support vector machines, random forests) produce equally accurate predictive models along the full 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 full 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) and, in this case, there may exist interpretable models that also attain the same level of accuracy. Finding these 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 powerful new machine learning method known as a Supersparse Linear Integer Model (SLIM, see Ustun et al. 2013, Ustun and Rudin 2014, 2015). SLIM is designed to produce models that are highly accurate but simple enough to make predictions by hand, without the use of a calculator or computer. We use SLIM to produce accurate transparent models along the full ROC curve. These models can be used for decision-making for many different cases; they are just as accurate as the most powerful black-box machine learning models, but completely transparent and highly interpretable. Black box models are indefensible. One may not agree with a particular transparent model, but one can at least have a clear idea of what it is doing.

The remainder of our paper is structured as follows. In Section 2, we discuss related work. In Section 3, we describe how we derived eight recidivism prediction problems and provide simple insights into each problem. In Section 4, we compare the accuracy and interpretability of models produced by the eight machine learning methods on the eight recidivism prediction problems, and include examples of accurate and interpretable SLIM models for each problem.

With this manuscript, all of our code will be published for the purpose of reproducibility and clarity to show how data were processed and how models were trained. We invite others to build on our work and adapt our methodology to produce interpretable models for recidivism prediction on future applications.

2 Related Work

We discuss related work in both criminal justice and in machine learning.

2.1 Related Work in Criminal Justice

Since the 1920's (Borden 1928, Burgess 1928, Tibbitts 1931), predictive models for recidivism have been in widespread use in different countries and areas of the criminal justice system, 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). Countries that currently use risk assessment tools include: Canada (Hanson and Thornton 2003); the Netherlands (Tollenaar and van der Heijden 2013); the U.K. (Howard et al. 2009); and the U.S (Turner et al. 2009). Applications of these tools can be seen in evidence-based sentencing (Hoffman 1994), correction / 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, Ritter 2013), determining appropriate sanctions for parole violations (Turner et al. 2009), and targeted policy interventions (Lowenkamp and Latessa 2004).

In this work, we consider predictive models for general recidivism (recidivism of any crime type) as well as crime-specific recidivism. Risk assessment tools for general recidivism risk prediction 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 majority of recidivism risk assessment tools were produced using logistic regression and substantially modified for the purposes of interpretability (see, e.g., the recommendations of Gottfredson and Snyder 2005). These approaches have led to serious issues in practice (see, e.g., Gottfredson and Moriarty 2006, for a detailed overview). In particular, risk assessments are not well suited for decision-making because they output risk estimates as opposed to prediction. Risk estimates can be converted to predicted outcomes by imposing a threshold (i.e. classify a prisoner as "high-risk" if the predicted probability of arrest $> 70\%$). Many tools use the risk estimate to produce several outcomes (e.g.,"low risk," "medium risk" and "high risk) with thresholds

that are decided arbitrarily (Hannah-Moffat 2013). This makes it difficult, if not impossible, to assess predictive accuracy: 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." This problem is further exacerbated by the fact that the performance of each tool is reported using statistics that do not immediately relate to predictive accuracy. In many cases, performance is measured using in-sample error (i.e., 'training' error) or in-sample AUC, despite the fact that the out-of-sample TPR/FPR at the decision-making point is far more relevant. There has been continued interest in more principled evaluation methods for these programs in recent years (see for instance the review papers of Skeem and Monahan 2011, Hanson and Morton-Bourgon 2009).

Predictive models for recidivism have been used extensively, especially for parole decisions and sentencing. In the 1970s, for instance, the United States Parole Commission began using an actuarial measurement built from 2,497 prisoners to inform parole decisions, called the Salient Factor Score (SFS) (Hoffman and Adelberg 1980). A follow-up study showed that the SFS had been fairly accurate in the first twenty years of its implementation (Hoffman 1994). Since 1987, the United States Sentencing Commission's Federal Sentencing Guidelines has mandated the use of a predictive recidivism measure for sentencing, in particular, the Criminal History Category (CHC) (U.S. Sentencing Commission 1987). A series of reports compared the CHC and SFS along various dimensions (U.S. Sentencing Commission 2004, 2005). We remark that the form of *all* of these models are linear models with integer coefficients like the ones we develop in this paper; of course, the CHC and SFS were not built using the sophisticated optimization techniques we use. Unlike the CHC and SFS, our models are created in a completely automated way, for each crime type, which implies that our tools can be used for population-specific models, or other datasets. The studies found that the AUC was 0.70 for the CHC and 0.73 for the SFS (U.S. Sentencing Commission 2005), which are within the range of the values we report in Section 4.4 (0.66 – 0.72 depending on the method). Note that our dataset is slightly different; it is over five times the size of that used for the USSC's recidivism study, and we predict for 3 years rather than 2 years.

2.2 Related Work in Statistics and Machine Learning

Many current criminologists and statisticians still depend heavily on traditional statistical tools such as logistic regression, e.g., the work of Penner et al. (2013). They consider in-sample performance, rather than the machine learning perspective of considering out-of-sample performance on a held-out test set; the machine learning perspective handles problems with overfitting and multiple testing, and it is well-known that only reporting in-sample accuracy can be very misleading. There are some works that consider machine learning approaches, and in particular, classification trees. For instance the work of Steadman et al. (2000) favors classification trees over logistic regression due to its similarity to the clinical method of decision making. Stalans et al. (2004) also found that decision trees are better than logistic regression for prediction of violent recidivism, and Silver and Chow-Martin (2002) used trees as a meta-learner over multiple existing recidivism models. Berk et al. (2005) conducted a study that helped the Los Angeles Sheriff's Department develop a simple and practical screener for forecasting domestic violence, also showing that decision trees were more accurate than logistic regression. Berk and Bleich (2014) used classification trees CART to provide a prototype of risk forecasting for sentencing. We extensively consider decision trees in this study, but we find there are flaws in the standard decision tree algorithms and implementations that other methods do not have. In particular, these decision tree methods are greedy and do not produce optimal solutions; they also cannot easily be tuned to different decision points.

Random forests (Breiman 2001b) is another popular method that produces very complex black-box models: Neuilly et al. (2011) found that random forests were better for prediction of recidivism for homicide offenders; Berk et al. (2006) used the method to forecast a prisoner's likelihood to commit a serious misconduct while incarcerated; Berk et al. (2009) used it to forecast potential murders for criminals on probation or parole; and Berk and Sorenson used random forests to forecast domestic violence and help inform court decisions at arraignment.

The work of Tollenaar and van der Heijden (2013) provides somewhat of an exception to the works on machine learning for recidivism, by claiming that machine learning approaches do not outperform classical statistical modeling approaches, and that they both perform similarly. Our findings are similar in some ways

to those of Tollenaar and van der Heijden (2013) in that for some decision points, we find that all methods - machine learning methods and classical statistical methods - have essentially the same performance. This "Rashomon" effect (Breiman 2001a) can be exploited to find models that are accurate but also are beneficial in other ways, such as interpretability. We agree with the commentary of Yang et al. (2010) who noted that for many problems, many prediction methods perform approximately the same, and in that case, one should use another measure to gauge how useful a model is. (Yang et al. 2010, was not a machine learning study however.) Our findings disagreed with Tollenaar and van der Heijden (2013) in an important way. Specifically, we find that for some decision points, the choice of algorithm is extremely important.

There are some works on the topic of handling issues such as race in predictive modeling (e.g., Gottfredson and Jarjoura 1996, Berk 2009), and broadly, guidelines on constructing and understanding models. Gottfredson and Moriarty (2006) provide a set of warnings for those using statistical tools to create predictions (though the evaluation of performance that they recommend differs from our approach). For instance, it is clearly true that all of our results are conditioned on our dataset. It is easily possible to change the conditioning by re-running our code on a subset of our data, or on recidivism data from another source.

3 Data and Preliminary Analysis

The recidivism prediction problems in our paper were all derived from the "Recidivism of Prisoners Released in 1994" database, which was put together by the U.S. Department of Justice, Bureau of Justice Statistics (2014). Each problem is a binary classification problem with $N = 33796$ prisoners and $P = 49$ input variables, where the goal is to predict if a prisoner will be arrested for a certain type of crime within 3 years of being released from prison. In what follows, we describe the original database, explain how we created each prediction problem, and provide insights. The "Recidivism of Prisoners Released in 1994" database (U.S. Department of Justice, Bureau of Justice Statistics 2014) is the largest publicly available database on prisoner recidivism in the United States.[1] It tracks a sample of 38,624 prisoners for 3 years following their release from prison in 1994. These prisoners are randomly sampled from the population of all prisoners released from 15 major states[2] and account for two-thirds of prisoners that were released from prison in the U.S. in 1994.

The database is composed of 38,624 rows and 6,427 columns, where each column represents a prisoner and each row represents 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 points in the 3 year follow-up period (e.g., if a prisoner was arrested until that point). 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.[3] To mirror the approach of Langan and Levin (2002), we also restricted our attention to the same subset of prisoners.

This dataset also has serious flaws which we point out below (almost every data set has serious flaws of some kind). For our dataset, many important factors that could be used to predict recidivism are missing, and many included factors are noisy enough that they were not useful for prediction in 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, and 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.,

[1] Other studies that use this database include Bhati and Piquero (2007), Bhati (2007), Zhang et al. (2009).

[2] The states in the database include: Arizona, California, Delaware, Florida, Illinois, Maryland, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Texas, and Virginia.

[3] 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.

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. For instance, we have only an indicator that someone was once a drug or alcohol abuser, but we have little details about the drug treatment or the extent of drug abuse. While some of these factors, such as drug/alcohol treatment and in prison vocational programs, exist, 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." In order for the study to include as many prisoners as possible, we chose to exclude factors with extremely sparse information.

Furthermore, there are more detailed categories of the crimes reported (e.g. fatal violence can be broken than into 6 categories), but we did not find those to be useful in preliminary analysis; an avenue of further investigation would be to consider these features further; however, without the education and family information, it is not clear that this would be worthwhile. The major benefits of this dataset over others are: (i) it is publicly available, benefiting reproducibility (ii) it is large (iii) its criminal history records are fairly complete. Indeed, the BJS study by Langan and Levin (2002) also split the crimes into exactly the same major categories that we did.

3.1 Prediction Problems

Input Variables: We derived a total of $P = 49$ input variables. We encoded each input variable as a binary rule of the form $x_{ij} \in \{0, 1\}$ where $x_{ij} = 1$ if condition j holds true about prisoner i, allowing us to encode highly nonlinear functions of the original variables. For clarity, we refer to input variables in the text using italicized font (e.g., *female*). We provide a summary of all input variables in Table 1. The final set of input variables represents known risk factors (Bushway and Piehl 2007, Crow 2008) and have been used in risk assessment tools since 1928 (see, e.g., Borden 1928, U.S. Sentencing Commission 2005, Berk et al. 2006, Baradaran 2013). Specifically, the variables are based on: 1) information about prison release in 1994 (e.g., *time_served*, *age_at_release*, *infraction*); 2) information from past arrests, sentencing, and convictions (e.g., *prior_arrests≥1*, *any_prior_jail_time*);[4] 3) history of substance abuse (e.g., *alcohol_abuse*) 4) gender (e.g., *female*). Thus our actuarial tools assess *static* recidivism risk[5] in the sense that a) the information is easily accessible to law enforcement officials (all above information can be found in state RAP sheets); and b) the input variables do not include socioeconomic factors such as race, which would directly eliminate the potential to use these tools in applications such as sentencing.

Outcome Variables: We created eight recidivism prediction problems by encoding a binary outcome variable $y_i \in \{-1, +1\}$, where $y_i = +1$ if a prisoner is arrested for a particular type of crime within 3 years after being released from prison; and $y_i = -1$ otherwise. For clarity, we refer to each prediction problem in the text using typewriter font (e.g., `arrest`). We provide details on the recidivism prediction problems that we consider in Table 2. We note that the percentages $P(y_i = +1)$ in Table 2 do not add up to 100% because a prisoner could be arrested for multiple types of crime within a single incident (e.g., both drug and public order offenses), and could also be arrested multiple times over the 3 year follow-up period.

The final problems that we consider include: an arrest for any crime (`arrest`); an arrest for a drug-related offense (`drug`); an arrest for a property-related offense (`property`); an arrest for a public order-related offense (`public_order`); or an arrest for a certain type of violent offense (`general_violence`, `domestic_violence`, `sexual_violence`, `fatal_violence`).[6]

[4]Note that the *prior_arrests* variable does not count the crime for which they were released from prison in 1994; thus, about 12% of the prisoners in the dataset have *no_prior_arrests* =1 even though they had to have been arrested at least once the crime for which they were released from prison in 1994.

[5]Static recidivism risk assessment tools use risk factors that do not change over time (see, e.g., Tollenaar and van der Heijden 2013, Hannah-Moffat 2013, for a discussion).

[6]We formulated mutually exclusive outcome variables for violent offenses 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`)

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≤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≥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
released_other	0.02	prisoner i released by other means
time_served≤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≥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≤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≥40	0.04	prisoner i was $\geq$ least 40 years at 1st arrest
age_1st_confinement≤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≥40	0.12	prisoner i was $\geq$ 40 years at 1st confinement
prior_arrests_for_drug	0.47	prisoner i was once arrested for drug offense
prior_arrests_for_property	0.67	prisoner i was once arrested for property offense
prior_arrests_for_public_order	0.62	prisoner i was once arrested for public order offense
prior_arrests_for_general_violence	0.52	prisoner i was once arrested for general violence
prior_arrests_for_domestic_violence	0.04	prisoner i was once arrested for domestic violence
prior_arrests_for_sexual_violence	0.03	prisoner i was once arrested for sexual violence
prior_arrests_for_fatal_violence	0.01	prisoner i was once arrested for fatal violence
prior_arrests_for_multiple_types	0.77	prisoner i was once arrested for multiple types of crime
prior_arrests_for_felony	0.84	prisoner i was once arrested for a felony
prior_arrests_for_misdemeanor	0.49	prisoner i was once arrested for a misdemeanor
prior_arrests_for_local_ordinance	0.01	prisoner i was once arrested for local ordinance
prior_arrests_with_firearms_involved	0.09	prisoner i was once arrested or an incident involving firearms
prior_arrests_with_child_involved	0.17	prisoner i was once arrested for an incident involving children
no_prior_arrests	0.12	prisoner i has no prior arrests
prior_arrests≥1	0.88	prisoner i has at least 1 prior arrest
prior_arrests≥2	0.78	prisoner i has at least 2 prior arrests
prior_arrests≥5	0.60	prisoner i has at least 5 prior arrests
multiple_prior_prison_time	0.43	prisoner i has been to prison multiple times
any_prior_jail_time	0.47	prisoner i has been to 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 been on probation or paid a fine at least once
multiple_prior_probation_or_fine	0.22	prisoner i has been been on probation or paid a fine multiple times

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

Problem	$\mathbf{P}(y_i = +1)$	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
property	25.2%	$y_i = +1$ if prisoner i is arrested for a property-related offense (e.g., burglary, larceny, arson, fraud) within 3 years of release from prison
public_order	27.9%	$y_i = +1$ if prisoner i is arrested for a public order offense (e.g., weapons possession, DUI) 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

Table 2: Overview of recidivism prediction problems.

3.2 Conditional Probabilities for Each Outcome and Variable

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 4.5. Consider, for instance, arrest. 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 (e.g., 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_arrests*).

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 arrested 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%).

4 Prediction Methodology and Empirical Results

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

4.1 Imbalanced Problems and Cost-Sensitive Classification

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

Imbalanced classification problems necessitate changes in the way that we train classification models as well as the way that we evaluate their performance. Consider, for instance, a heavily imbalanced problem such as fatal_violence where only $\mathrm{P}(y_i = +1) = 0.4\%$ 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 model that predicts no one will be arrested for fatal offenses – a result that is not surprising given that the trivial model

Input Variable	Prediction Problem							
	arrest	drug	property	public order	general violence	domestic violence	sexual violence	fatal violence
female	0.54	0.21	0.27	0.23	0.11	0.02	0.01	0.0005
prior_alcohol_abuse	0.58	0.18	0.29	0.28	0.20	0.04	0.03	0.01
prior_drug_abuse	0.61	0.23	0.32	0.27	0.21	0.03	0.03	0.004
age_at_release≤17	0.84	0.35	0.46	0.36	0.31	0.01	0.01	0.04
age_at_release_18_to_24	0.71	0.24	0.31	0.35	0.25	0.04	0.03	0.01
age_at_release_25_to_29	0.66	0.23	0.29	0.32	0.21	0.04	0.03	0.01
age_at_release_30_to_39	0.59	0.20	0.26	0.28	0.17	0.04	0.03	0.01
age_at_release≥40	0.41	0.12	0.15	0.18	0.09	0.02	0.03	0.003
released_unconditional	0.65	0.20	0.28	0.36	0.23	0.06	0.04	0.01
released_conditional	0.58	0.20	0.25	0.27	0.17	0.03	0.03	0.01
released_other	0.61	0.19	0.19	0.19	0.11	0.004	0.02	0.004
time_served≤6mo	0.67	0.27	0.30	0.32	0.19	0.04	0.03	0.01
time_served_7_to_12mo	0.63	0.22	0.28	0.28	0.19	0.04	0.03	0.01
time_served_13_to_24mo	0.59	0.20	0.24	0.28	0.17	0.04	0.03	0.01
time_served_25_to_60mo	0.53	0.16	0.22	0.26	0.17	0.03	0.03	0.01
time_served≥61mo	0.48	0.11	0.17	0.21	0.15	0.02	0.04	0.004
infraction_in_prison	0.65	0.19	0.30	0.30	0.20	0.01	0.04	0.01
age_1st_arrest≤17	0.73	0.27	0.36	0.37	0.27	0.04	0.04	0.01
age_1st_arrest_18_to_24	0.64	0.22	0.27	0.30	0.20	0.04	0.03	0.01
age_1st_arrest_25_to_29	0.47	0.14	0.17	0.21	0.10	0.02	0.02	0.005
age_1st_arrest_30_to_39	0.34	0.10	0.11	0.13	0.06	0.02	0.02	0.003
age_1st_arrest≥40	0.21	0.05	0.05	0.09	0.03	0.01	0.02	0.002
age_1st_confinement≤17	0.78	0.28	0.39	0.38	0.29	0.04	0.04	0.02
age_1st_confinement_18_to_24	0.68	0.24	0.30	0.33	0.23	0.05	0.04	0.01
age_1st_confinement_25_to_29	0.60	0.20	0.25	0.28	0.17	0.03	0.03	0.005
age_1st_confinement_30_to_39	0.50	0.16	0.20	0.23	0.12	0.03	0.02	0.003
age_1st_confinement≥40	0.34	0.09	0.12	0.16	0.07	0.01	0.02	0.002
prior_arrests_for_drug	0.68	0.32	0.31	0.33	0.21	0.04	0.02	0.01
prior_arrests_for_property	0.67	0.24	0.33	0.33	0.22	0.04	0.03	0.01
prior_arrests_for_public_order	0.65	0.24	0.30	0.35	0.22	0.04	0.03	0.01
prior_arrests_for_general_violence	0.67	0.25	0.32	0.35	0.26	0.05	0.04	0.01
prior_arrests_for_domestic_violence	0.66	0.21	0.28	0.33	0.27	0.13	0.04	0.01
prior_arrests_for_sexual_violence	0.49	0.13	0.18	0.25	0.16	0.04	0.06	0.01
prior_arrests_for_fatal_violence	0.54	0.19	0.20	0.25	0.21	0.04	0.03	0.01
prior_arrests_for_multiple_crime_types	0.64	0.23	0.29	0.32	0.21	0.04	0.03	0.01
prior_arrests_for_felony	0.60	0.21	0.27	0.30	0.19	0.04	0.03	0.01
prior_arrests_for_misdemeanor	0.69	0.26	0.33	0.38	0.24	0.06	0.03	0.01
prior_arrests_for_local_ordinance	0.91	0.29	0.45	0.84	0.43	0.15	0.05	0.02
prior_arrests_with_firearms_involved	0.70	0.30	0.32	0.35	0.27	0.06	0.03	0.01
prior_arrests_with_child_involved	0.48	0.13	0.17	0.26	0.14	0.03	0.06	0.01
no_prior_arrests	0.32	0.07	0.09	0.13	0.08	0.02	0.02	0.003
prior_arrests_≥_1	0.63	0.22	0.27	0.30	0.19	0.04	0.03	0.01
prior_arrests_≥_2	0.66	0.23	0.29	0.32	0.20	0.04	0.03	0.01
prior_arrests_≥_5	0.70	0.25	0.33	0.35	0.22	0.04	0.03	0.01
multiple_prior_prison_time	0.65	0.23	0.30	0.30	0.19	0.03	0.03	0.01
any_prior_jail_time	0.69	0.25	0.32	0.33	0.21	0.04	0.03	0.01
multiple_prior_jail_time	0.73	0.27	0.37	0.36	0.22	0.04	0.03	0.01
any_prior_probation_or_fine	0.67	0.24	0.31	0.33	0.20	0.04	0.03	0.01
multiple_prior_probation_or_fine	0.71	0.27	0.35	0.37	0.22	0.05	0.03	0.01

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.

is 99.6% accurate on the overall population. Unfortunately, this model will never be able to identify individuals that are arrested for fatal offenses, and will actually be 0% accurate on the population of interest.

In order to provide a clear measure of performance of classification model 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 metrics known as 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).[7] A detailed discussion of ROC analysis in recidivism prediction can be found in the work of Maloof (2003).

Different applications will require 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 this paper, 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.

4.2 Overview of Classification Methods

We compared the performance of models from eight different classification methods, including those previously used for recidivism prediction (see Section 2.2) or that ranked among the "top 10 algorithms in data mining" (Wu et al. 2008). We restricted our attention to methods with publicly-available software packages that allowed us to specify misclassification costs for positive and negative classes.

- **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.
- **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), which is a predecessor to C5.0.
- **L_1 and L_2-Penalized Logistic Regression**: State-of-the-art 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).

[7] 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.

- **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 allows 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).
- **Supersparse Linear Integer Models**: A new method to create scoring systems that are optimized for accuracy and sparsity (Ustun and Rudin 2015). We provide a short overview in the following section.

4.2.1 Supersparse Linear Integer Models

A Supersparse Linear Integer Model (SLIM) is a new optimization-based method for creating *scoring systems* – 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).

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 (see, e.g., Webster et al. 1997, Webster 2013, for applications in criminology). These models are often more interpretable than traditional linear models because they are highly sparse and use a small number of integer coefficients. Such characteristics allow users to easily gauge the influence of one input variable with respect to the others – by catering to the fact that most humans are seriously limited in the number of cognitive entities they can handle at once (7 ± 2 according to Miller 1984), and seriously limited in estimating the association between three or more variables (Jennings et al. 1982).

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 $1, \ldots, P$), 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 fitted from data by solving a discrete optimization problem of the form:

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

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 small set of bounded integers and may include additional conditions to tailor the accuracy and interpretability of the final scoring system. We note that the objective also includes a *tiny* penalty on the absolute value of the coefficients to restrict coefficients to coprime values without affecting accuracy or sparsity.[8]

[8] 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, 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, $\hat{y} = \text{sign}\,(x_1 + x_2)$.

SLIM differs from state-of-the-art machine learning methods because it directly optimizes accuracy and sparsity, without making approximations that other methods make for scalability (e.g., the use of convex loss functions). By avoiding these approximations, SLIM sacrifices the ability to fit a model within seconds in a way that scales to extremely large datasets. In return, however, it gains the ability to produce accurate models that are far more practical and interpretable than the state-of-the-art.

SLIM also has the unique ability to address operational constraints related to accuracy and interpretability, without the need for parameter tuning (e.g., it can directly produce scoring systems with explicit limits on the false positive rate, or the number of input variables in the final model). In our experiments, we trained the following version of SLIM:

$$\begin{aligned}
\min_{\boldsymbol{\lambda}} \frac{W^+}{N} \sum_{i \in \mathcal{I}^+} \mathbb{1}\left[y_i \neq \hat{y}_i\right] + \frac{W^-}{N} \sum_{i \in \mathcal{I}^-} \mathbb{1}\left[y_i \neq \hat{y}_i\right] + C_0 \sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right] + \epsilon \sum_{j=1}^{P} |\lambda_j| \\
\text{s.t.} \quad \sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right] \leq 8 \\
\lambda_j \in \{-10, \ldots, 10\} \text{ for } j = 1,...,P \\
\lambda_0 \in \{-100, \ldots, 100\}.
\end{aligned}$$

Here, we constrained each coefficient λ_j to an integer between -10 and 10, we constrained the threshold λ_0 to an integer between -100 and 100, and we restricted the number of non-zero coefficients to at most 8, similar to the number of cognitive entities humans can handle (Miller 1956). These choices were intended to create a model that would allow users to make predictions without a computer or calculator, and to be able to easily assess how joint values of input variables affected the predicted outcome. The C_0 parameter was set to a sufficiently small value so that SLIM would not sacrifice accuracy for sparsity; given W^+, W^- we can set C_0 to any value $0 < C_0 < \min\{W^-, W^+\}/NP$ to ensure this condition. The ϵ parameter was set to a sufficiently small value so that SLIM would produce a model with coprime coefficients without affecting accuracy or sparsity; given W^+, W^- and C_0, we can set ϵ to any value $0 < \epsilon < C_0 / \max \sum_{j=1}^{P} |\lambda_j|$ to ensure this condition.

One cannot expect the same level of accuracy or constraint satisfaction by manually rounding or adjusting logistic regression coefficients, manually searching over integer points in a high dimensional space can be highly suboptimal.

4.3 Experimental Design

We provide an overview of the methods, software, and settings that we used to produce prediction models for all problems in Table 4.

We ran each method on each problem for 19 values of W^+ as well as method-specific free parameters. By default, we chose values of $W^+ \in \{0.050, \ldots, 1.950, 2.000\}$ (i.e. cost ratios between 1:39 and 39:1). This range was inappropriate for problems with a significant class imbalance as all methods produced trivial models. For significantly imbalanced problems, such as `domestic_violence`, `sexual_violence`, we instead used values of $W^+ \in \{1.905 \ldots, 1.995\}$ (i.e. cost ratios between 19:1 and 199:1). In the case of `fatal_violence`, which was extremely imbalanced, we used $W^+ \in \{1.9905, \ldots, 1.9995\}$ (i.e. cost ratios between 209:1 and 3999:1).

For each problem, each method, each value of W^+, and each instance of the method-specific free parameters, we trained a total of 11 models: 1 model using all of the data to assess interpretability, and 10 models using subsets of the data to assess the predictive accuracy through 10-fold cross-validation (10-CV). We generated the folds once, and used the same folds for each problem so as to allow for comparisons across algorithms and problems.

We constructed an ROC curve for each method by plotting the 10-CV mean TPR and 10-CV mean FPR of the produced model for each distinct value of W^+. For methods such as Lasso, Ridge and SVM RBF, where we trained multiple models for each value of W^+ with different settings of the free parameters, we selected a

Method	Acronym	Software	Free Parameters and Settings
CART Decision Trees	CART	**rpart** (Therneau et al. 2012)	19 values of W^+
C5.0 Decision Trees	C5.0T	**c50** (Kuhn et al. 2012)	19 values of W^+
C5.0 Decision Rules	C5.0R	**c50** (Kuhn et al. 2012)	19 values of W^+
Logistic Regression (L_1-Penalty)	Lasso	**glmnet** (Friedman et al. 2010)	19 values of $W^+ \times$ 100 values of L_1-penalty chosen by **glmnet**
Logistic Regression (L_2-Penalty)	Ridge	**glmnet** (Friedman et al. 2010)	19 values of $W^+ \times$ 100 values of L_2-penalty chosen by **glmnet**
Random Forests	RF	**randomForest** (Liaw and Wiener 2002)	19 values of W^+
Support Vector Machines (Radial Basis Kernel)	SVM RBF	**e1071** (Meyer et al. 2012)	19 values of W^+ $\times$ 7 values of $C \in (10^{-3}, \ldots, 10^3)$
SLIM Scoring Systems	SLIM	CPLEX 12.6	19 values of W^+; $C_0, \epsilon, \mathcal{L}$ set to find most accurate model with ≤ 8 coefficients where $\lambda_0 \in \{-100, \ldots, 100\}$ and $\lambda_j \in \{-10, \ldots, 10\}$

Table 4: Methods, software and free parameters used to create models for each problem. The values of W^+ are problem-specific.

single model for each of the W^+ values by choosing the instance that minimized the weighted mean 10-CV test error. This resulted in optimistic performance for these methods, as we discuss shortly.

We trained all models on a 12-core 2.7GHZ Intel Nehalem CPU with 48GB RAM. We trained all methods other than SLIM using publicly available packages in R 3.1.1 (R Core Team 2014) without imposing any time constraints. We trained SLIM by solving mixed-integer programming problems (MIP) with the CPLEX 12.6 API in MATLAB 2014a. We set a time limit of 1 hour for each MIP and solved 11 MIPs in parallel. Thus, it took 19 hours to train all of the SLIM models required to produce an ROC curve for a single problem. This was comparable to the time required to train models with RF and SVM RBF: these methods required 40–60 minutes to train models at each value of W^+, and 12–20 hours to produce all models required to create an ROC curve.

Our design differs from split-sample designs where a substantial portion of the data is reserved for used as a test set (see, e.g., Tollenaar and van der Heijden 2013, Berk et al. 2014). We chose to forgo allocating a test set because it would have substantially reduced the number of examples in the minority class for heavily imbalanced problems such as `fatal_violence`.[9] This design may have resulted in optimistic performance for Lasso, Ridge and SVM RBF as we set the free parameters for these methods using 10-CV statistics. However, it had the benefit of producing 10 separate estimates of predictive accuracy, and final predictive models that were trained with as much data as possible, and provided us with a measure of uncertainty on unseen data.

4.4 Observations on Predictive Accuracy

We show the ROC curves for all methods on all prediction problems in Figures 1 and 2 and summarize the 10-CV test AUC of each method in Table 5. We make the following important observations, which we believe carries 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 0.70; the highest test 10-CV test AUC was achieved by SVM RBF (0.74). This multiplicity of good models reflects the *Rashomon effect* of Breiman (2001b).

[9] There is little guidance on how much data to allocate for a test set (see e.g Faraway 2014, for a discussion.)

- Major differences between methods appeared in terms of their performance on imbalanced prediction problems. We expected different methods 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 – especially in problems with significant class imbalance (e.g., `fatal_violence`).

- SVM RBF, LARS Lasso and LARS Ridge were able to produce accurate models at different points on the ROC curve on most problems. SVM RBF usually achieved the highest AUC on most problems (e.g., `arrest`, `drug`, `property`, `public_order`, `general_violence`), though Lasso and Ridge often produce comparable AUCs. We find that these methods do respond well to cost-sensitive tuning, but that it is difficult to find appropriate misclassification costs for highly imbalanced problems such as `sexual_violence`, `domestic_violence` and `fatal_violence`.

- 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, and that this issue becomes markedly more severe as problems become more imbalanced. For `drug`, `property`, and `general_violence`, for instance, these methods could not produce models with high TPR. For `sexual_violence` and `domestic_violence`, these methods almost always produced trivial models that always predict $y = 0$ (resulting in AUCs of 0.5). This result may be attributed to the greedy nature of the algorithms that are being 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).

- Random Forests – while being able to produce accurate models at certain points on the ROC curve – tend to overfit on all problems. To see this, we can compare the difference in performance between the testing accuracy in Table 5 (usually near 0.7) and the training AUC in Table **??** (usually near 1.0). The overfitting could be due to settings in the **randomForest** package in R, or the use of a cost-sensitive approach (Berk et al. 2006, mentions that random forests may be better-suited to tackle imbalanced problems using sampling-based approaches that oversample the minority class.) The settings in the R package cannot easily be fixed without tuning in a huge parameter space, which itself grows exponentially with the number of parameters.

- SLIM performs well despite being restricted to a relatively small class of simple linear models (e.g., models with 8 non-zero integer coefficients) and without cross-validated parameters. In general, SLIM produces models that are not significantly less accurate than other methods and close to or on on the efficient frontier of the ROC curve (see, e.g., `general_violence`, `property`). Further, SLIM achieves the highest AUC in highly imbalanced problems, such as `fatal_violence` and `sexual_violence`, as it responds well to changes in misclassification costs.

4.5 Observations on Interpretability

Interpretable predictive models provide "a *qualitative understanding* of the relationship between *joint* values of the input variables and the resulting predicted response value," (Hastie et al. 2009). In assessing the interpretability of various models, we distinguish between *transparent* models, which provide a textual or visual representation of the relationship between input variables and the predicted outcome (CART, C5.0R, C5.0, Lasso, Ridge), and *black-box* models, which do not (RF, SVM RBF).

Trade-offs Between Accuracy and Interpretability

Most of the methods that we tested for are unable to produce a prediction model that is both accurate and interpretable along the full ROC curve. 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 Lasso. Among the remaining methods:

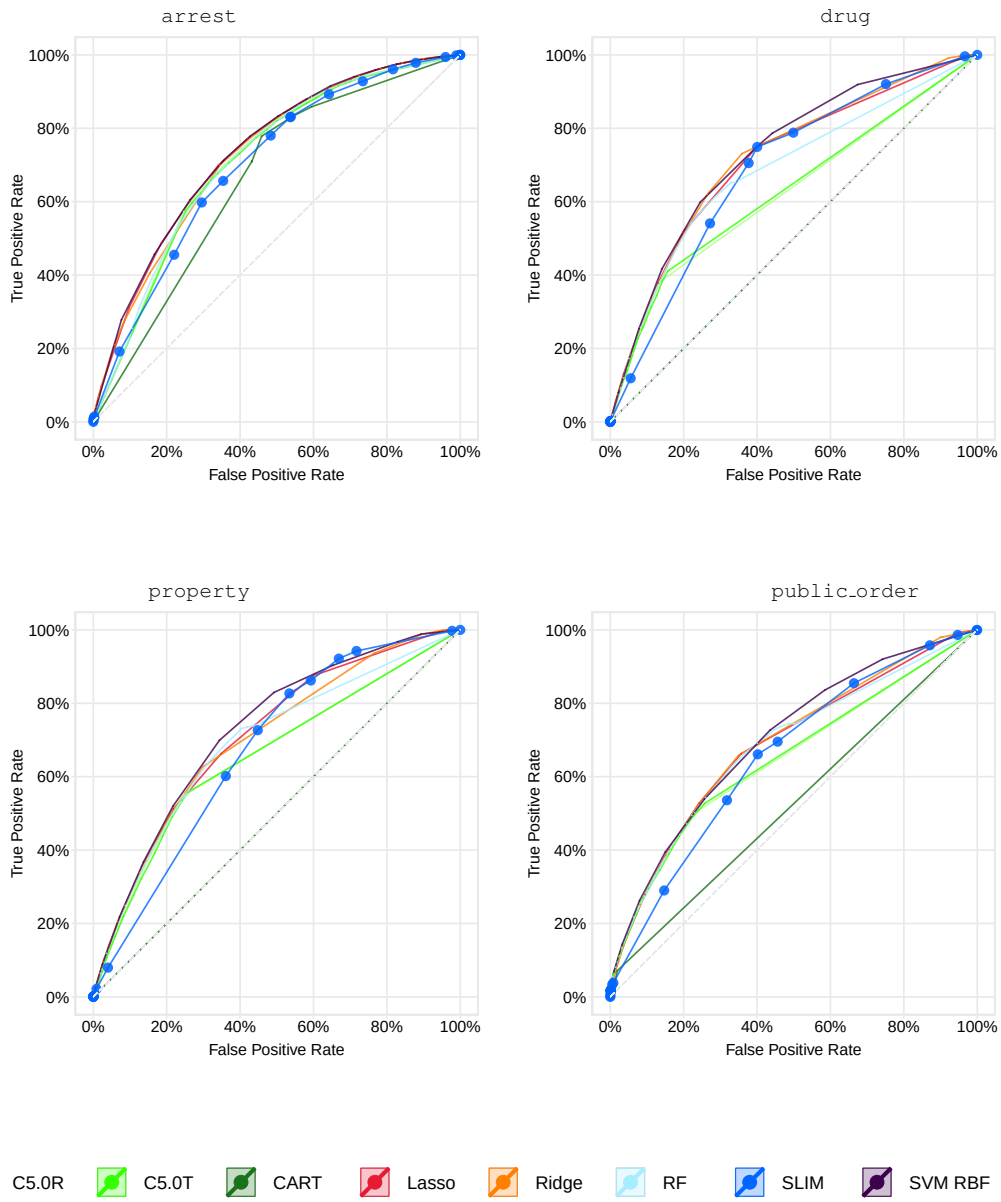

Figure 1: ROC curves for general recidivism-related prediction problems. We plot SLIM models using large blue dots. All models perform similarly except for C5.0R, C5.0T, and CART.

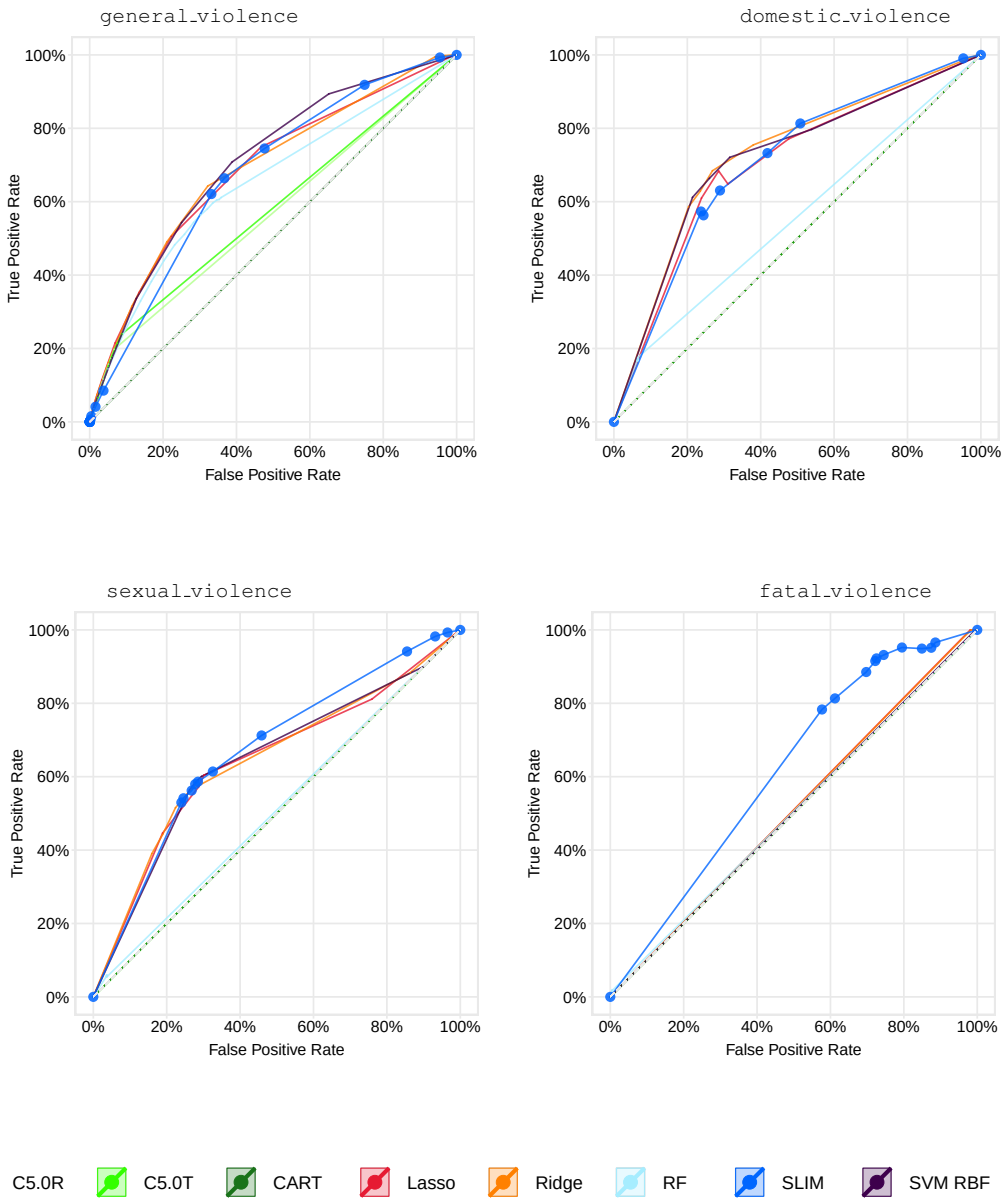

Figure 2: ROC curves for violence-related prediction problems. We plot SLIM models using large blue dots. Here C5.0R, C5.0T, and CART performed poorly, and for `fatal_violence`, all methods except SLIM performed poorly.

Problem	Lasso	Ridge	C5.0R	C5.0T	CART	RF	SVM RBF	SLIM
arrest	0.74 0.73 - 0.75	0.73 0.72 - 0.74	0.72 0.71 - 0.73	0.71 0.70 - 0.73	0.66 0.65 - 0.67	0.72 0.71 - 0.73	0.74 0.73 - 0.75	0.70 0.68 - 0.72
drug	0.71 0.68 - 0.74	0.72 0.69 - 0.74	0.63 0.59 - 0.66	0.63 0.61 - 0.65	0.50 0.50 - 0.50	0.69 0.67 - 0.70	0.73 0.71 - 0.75	0.69 0.67 - 0.72
property	0.71 0.69 - 0.73	0.70 0.68 - 0.73	0.66 0.65 - 0.67	0.66 0.63 - 0.67	0.50 0.50 - 0.50	0.69 0.67 - 0.71	0.73 0.70 - 0.74	0.67 0.65 - 0.70
public_order	0.69 0.68 - 0.70	0.69 0.68 - 0.71	0.65 0.64 - 0.67	0.65 0.64 - 0.66	0.53 0.51 - 0.54	0.68 0.66 - 0.70	0.70 0.69 - 0.73	0.66 0.64 - 0.67
general_violence	0.68 0.67 - 0.70	0.69 0.67 - 0.71	0.57 0.55 - 0.58	0.58 0.56 - 0.60	0.50 0.50 - 0.50	0.65 0.63 - 0.67	0.70 0.69 - 0.72	0.67 0.66 - 0.70
domestic_violence	0.70 0.67 - 0.72	0.73 0.68 - 0.75	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.55 0.51 - 0.58	0.72 0.69 - 0.75	0.70 0.57 - 0.74
sexual_violence	0.66 0.64 - 0.70	0.66 0.63 - 0.70	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.51 0.49 - 0.53	0.66 0.63 - 0.70	0.67 0.63 - 0.74
fatal_violence	0.51 0.50 - 0.53	0.51 0.50 - 0.52	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.50 0.50 - 0.50	0.51 0.50 - 0.53	0.50 0.50 - 0.51	0.62 0.56 - 0.68

Table 5: Test AUC for all methods on all prediction problems. Each cell contains the 10-CV mean test AUC (top), as well as the 10-CV minimum and maximum test AUC (bottom).

- Methods such as RF and SVM RBF produce "black box" models that do not provide a comprehensible representation of the relationship between input variables and the predicted outcome.
- Ridge produces models that are accurate and transparent. The models produced by ridge provide a clear representation of the relationship between input variables and the predicted outcome. However, they do not allow users to understand how *joint* values of the input variables affect the predicted outcome since they use all of the $P = 49$ input variables; these models are not sparse.
- Tree and rule-based methods such as CART, C5.0T and C5.0R are generally unable to produce to models that attain high degrees of accuracy. Even on balanced problems such as arrest where these methods are able to produce accurate models, however, we find that these 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). While it may be possible to simplify the models that we obtained through these methods through heuristic post-processing methods (e.g., pruning), it is not likely to benefit the accuracy of the model (and can drastically change the TPR or FPR).

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 example provide a nice basis for comparison as all three methods are able to produce a model at roughly the same decision point, and with the same degree of sparsity.[10] We make the following observations:

[10]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 less than or equal to that of the SLIM model (48.3%). We report the models that attained the highest test TPR among all such models. Here, neither C5.0R nor C5.0R could produce an acceptable model with 8 rules or 8 leaves.

2 *age_at_release≤24* $+$ 2 *prior_arrests≥5* $-$ 2 *age_1st_confinement≥40* $-$ 1

Figure 3: SLIM model for `arrest`. This model has a training TPR/FPR of 78.4%/47.1%, and a mean 10-CV test TPR/FPR of 78.0%/48.3%.

	0.61 *prior_arrests≥5*	$+$	0.16 *age_1st_confinement_18_to_24*	$+$	0.07 *prior_arrests_for_property*
$+$	0.06 *prior_arrests_for_misdemeanor*	$+$	0.02 *prior_arrests≥2*	$+$	4.75×10^{-03} *prior_arrests_for_general_violence*
$-$	0.21 *age_at_release≥40*	$-$	0.31		

Figure 4: Lasso model for `arrest`. This model has a training TPR/FPR of 73.7%/45.2%, and a mean 10-CV test TPR/FPR of 73.2%/44.9%.

- All three models attain similar levels of predictive accuracy. Test TPR values ranged between 73-78% and test FPR values ranged between 45-47%. There may not exist a classification model that can attain a substantially higher accuracy. The highest test TPR attained by models with test FPR $\leq 50\%$ was produced by the SVM RBF model which had a TPR of 80%.

- The SLIM model is highly sparse and uses 4 input variables. The small integer coefficients allow users to make quick predictions without a computer or calculator (see, e.g., Figure 6). Further, they allow users to easily gauge the importance of different input variables and provide a natural rule-based interpretation. In this case, for example, the SLIM model effectively says "predict arrest for any crime if age at release is ≤ 24 or prior arrests ≥ 5, unless the age at first confinement ≥ 40."

- The CART model allows users to make hands-on predictions. 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≥40* and *age_at_release_18_to_24*. However, the effect is difficult to comprehend since it depends on prior arrests and prior jail time: if *prior_arrests≥5* = 1 and *age_at_release_18_to_24* = 1 then the model predicts $\hat{y} = +1$; if *prior_arrests≥5* = 0 and *age_at_release≥40* = 0 then $\hat{y} = +1$; however, if *prior_arrests≥5* = 0 and *age_at_release≥40* = 1 then $\hat{y} = +1$ only if *multiple_prior_jail_time* = 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.

Scoring Systems for Recidivism Prediction

We present one of the SLIM scoring systems for each of the prediction problems that we consider in Figures 6 through 13. Many of these models exhibit the same "rule-like" interpretability discussed in the previous section (see, e.g., `drug` in Figure 7, which effectively predicts that a person will be arrested for a drug-related offense if he/she has had any prior arrests *and* was at most 17 years old when released from prison).

Note that the models below are not causal, meaning that one cannot change a single feature and evaluate the change in the predictions; a person with one feature that is different may also have other correlated features that are not accounted for when changing one feature at a time. One can argue why or why not these models make sense, whereas a black box model (such as SVM) is indefensible; a black box model is often so complex that printing one on a page of this manuscript would not be possible.

5 Conclusion

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 interpretable,

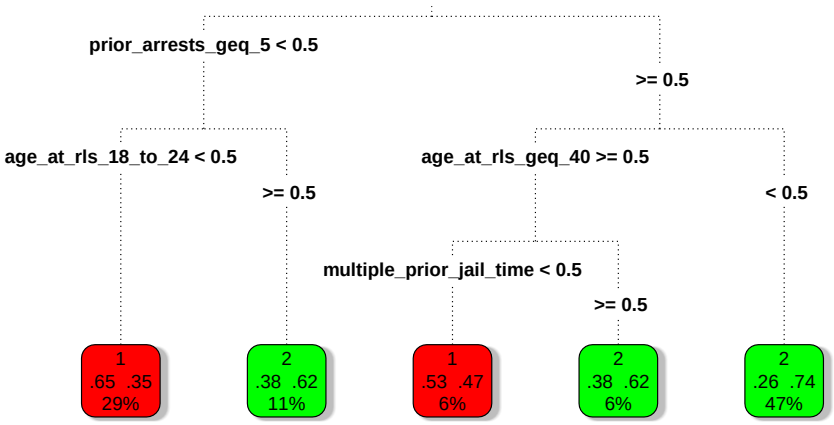

Figure 5: CART model for `arrest`. This model has a training TPR/FPR of 77.7%/45.4%, and a mean 10-CV test TPR/FPR of 77.9%/45.9%.

PREDICT ARREST FOR ANY OFFENSE IF SCORE > 1

1.	*age_at_release*$\leq$*24*	2 points		.....
2.	*prior_arrests*$\geq$*5*	2 points	+	.....
3.	*age_1st_confinement*$\geq$*40*	-2 points	+	.....
	ADD POINTS FROM ROWS 1-3	**SCORE**	=	.....

Figure 6: SLIM scoring system for `arrest`. This model has a training TPR/FPR of 78.4%/47.1%, and a mean 10-CV test TPR/FPR of 78.0%/48.3%.

PREDICT ARREST FOR DRUG OFFENSE IF SCORE > 1

1.	*prior_arrests_for_drugs*	2 points		.....
2.	*age_at_release*$\geq$*18*	-2 points	+	.....
	ADD POINTS FROM ROWS 1-2	**SCORE**	=	.....

Figure 7: SLIM scoring system for `drug`. This model has a training TPR/FPR of 75.5%/40.3%, and a mean 10-CV test TPR/FPR of 70.5%/37.7%.

PREDICT ARREST FOR PROPERTY OFFENSE IF SCORE > 1

1.	*prior_arrests_for_property*	3 points		.....
2.	*prior_arrests*$\geq$*1*	-1 point	+	.....
3.	*prior_arrests_for_sexual*	-2 points	+	.....
	ADD POINTS FROM ROWS 1-3	**SCORE**	=	.....

Figure 8: SLIM scoring system for `property`. This model has a training TPR/FPR of 69.4%/41.3%, and a mean 10-CV test TPR/FPR of 72.7%/44.8%.

PREDICT ARREST FOR PUBLIC ORDER OFFENSE IF SCORE > 1

1.	*prior_arrests_for_public_order*	2 points		.....
2.	*prior_arrests_for_general_violence*	2 points	+	.....
3.	*prior_arrests*$\geq$*1*	-2 points	+	.....
	ADD POINTS FROM ROWS 1-3	**SCORE**	=	.....

Figure 9: SLIM scoring system for `public_order`. This model has a training TPR/FPR of 55.9%/36.4%, and a mean 10-CV test TPR/FPR of 55.0%/34.5%.

PREDICT ARREST FOR VIOLENT OFFENSE IF SCORE > 1

1.	*age_at_release≤17*	2 points		.....
2.	*prior_arrests_for_fatal_violence*	2 points	+	.....
3.	*prior_arrests_for_general_violence*	2 points	+	.....
4.	*age_1st_confinement≥40*	-2 points	+	.....
5.	*age_1st_arrest≥40*	-4 points	+	.....
	ADD POINTS FROM ROWS 1-5	**SCORE**	=	.....

Figure 10: SLIM scoring system for `general_violence`. This model has a training TPR/FPR of 73.4%/44.2%, and a mean 10-CV test TPR/FPR of 73.9%/45.1%.

PREDICT ARREST FOR FATAL VIOLENCE IF SCORE > 1

1.	*prior_arrests_for_multiple_types_of_crime*	4 points		.....
2.	*released_other*	2 points	+	.....
3.	*age_1st_arrest≥40*	-2 points	+	.....
4.	*age_at_release_30_to_39*	-2 points	+	.....
5.	*age_at_release_25_to_29*	-4 points	+	.....
6.	*age_at_release≥40*	-6 points	+	.....
7.	*female*	-8 points	+	.....
	ADD POINTS FROM ROWS 1-7	**SCORE**	=	.....

Figure 11: SLIM scoring system for `fatal_violence`. This model has a training TPR/FPR of 90.5%/63.7%, and a mean 10-CV test TPR/FPR of 81.3%/61.3%. The reason for the "*released_other*" term appearing in the model is unclear, though as Table 3 shows, the type of release indeed has predictive power. The codebook for the dataset has little explanation for the meaning of this variable, other than that it excludes conditional releases such as parole and probation release, unconditional releases such as expiration of sentence or commutation and pardon, and releases by death or transfer. This term actually appears in some models (see Figure 11), indicating that it is worthwhile to further investigate this variable for either data quality or more information.

PREDICT ARREST FOR SEXUAL OFFENSE IF SCORE > 9

1.	*prior_arrests_for_sexual*	6 points		.....
2.	*age_at_release_25_to_29*	6 points	+	.....
3.	*age_at_release_18_to_24*	4 points	+	.....
4.	*age_at_release≥30*	4 points	+	.....
5.	*prior_arrests_for_local_ord*	4 points	+	.....
6.	*released_unconditonal*	2 points	+	.....
7.	*age_1st_arrest≥40*	-2 points	+	.....
	ADD POINTS FROM ROWS 1-7	**SCORE**	=	.....

Figure 12: SLIM scoring system for `sexual`. This model has a training TPR/FPR of 59.9%/28.7%, and a mean 10-CV test TPR/FPR of 58.9%/28.6%.

PREDICT ARREST FOR DOMESTIC VIOLENCE IF SCORE > 17

1.	*prior_arrests_for_felony*	8 points		.....
2.	*prior_arrests_for_misdemeanor*	6 points	+	.....
3.	*prior_arrests_for_general_violence*	6 points	+	.....
4.	*age_at_release_18_to_24*	4 points	+	.....
5.	*female*	-4 points	+	.....
6.	*infraction_in_prison*	-6 points	+	.....
	ADD POINTS FROM ROWS 1-6	**SCORE**	=	.....

Figure 13: SLIM scoring system for `domestic_violence`. This model has a training TPR/FPR of 65.5%/27.6%, and a mean 10-CV test TPR/FPR of 73.1%/39.9%.

in fact, small enough to fit on an index card. We used a set of features that are commonly accessible to police officers and judges, and performed a comparison of different machine learning methods. Our findings show that there are (potentially major) advantages of using new machine learning tools like SLIM. SLIM produces models that are just as accurate as the more complicated black box algorithms; however, they are also much more interpretable, they can be customized automatically to follow criminological knowledge, and they can be directly useful in decision-making. These models can be dependably generated for any given decision point along the ROC curve.

References

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

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

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

Barnes, G. C. and Hyatt, J. M. Classifying adult probationers by forecasting future offending.

Belfrage, H., Fransson, R., and Strand, S. 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, R. The role of race in forecasts of violent crime. *Race and social problems*, 1(4):231–242, 2009.

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

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

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

Berk, R., Sherman, L., Barnes, G., Kurtz, E., and Ahlman, L. 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, R., Bleich, J., Kapelner, A., Henderson, J., Barnes, G., and Kurtz, E. Using regression kernels to forecast a failure to appear in court. *arXiv preprint arXiv:1409.1798*, 2014.

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

Berk, R. A. and Sorenson, S. D. Machine learning forecasts of domestic violence to help inform release decisions at arraignment.

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

Berk, R. A., Kriegler, B., and Baek, J.-H. Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology*, 22(2):131–145, 2006.

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

Bhati, A. S. and Piquero, A. R. 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, H. G. Factors for predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, pages 328–336, 1928.

Borum, R. Manual for the structured assessment of violence risk in youth (savry). 2006.

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

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

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

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

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

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

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

Copas, J. and Marshall, P. 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, N. and Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

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

Dawes, R. M., Faust, D., and Meehl, P. E. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

Faraway, J. J. Does data splitting improve prediction? *Statistics and Computing*, pages 1–12, 2014.

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

Goh, S. T. and Rudin, C. Box drawings for learning with imbalanced data. *arXiv preprint arXiv:1403.3378*, 2014.

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

Gottfredson, S. D. and Jarjoura, G. R. Race, gender, and guidelines-based decision making. *Journal of Research in Crime and Delinquency*, 33(1):49–69, 1996.

Gottfredson, S. D. and Moriarty, L. J. Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, 52(1):178–200, 2006.

Grove, W. M. and Meehl, P. E. 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.

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

Hanson, R. K. and Morton-Bourgon, K. E. The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychological assessment*, 21(1):1, 2009.

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

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. *The elements of statistical learning*, volume 2. Springer, 2009.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

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

Howard, P., Francis, B., Soothill, K., and Humphreys, L. Ogrs 3: The revised offender group reconviction scale. 2009.

Jennings, D., Amabile, T. M., and Ross, L. Informal covariation assessment: Data-based vs. theory-based judgments. 1982.

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

Kuhn, M. and Johnson, K. *Applied predictive modeling*. Springer, 2013.

Kuhn, M., Weston, S., and code for C5.0 by R. Quinlan, N. C. C. *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, P. A. and Levin, D. J. Recidivism of prisoners released in 1994. *Federal Sentencing Reporter*, 15(1):58–65, 2002.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., and Hansen, K. T. 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, A. and Wiener, M. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. 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, C. T. and Latessa, E. J. Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in community corrections*, 2004:3–8, 2004.

Maden, A., Rogers, P., Watt, A., Lewis, G., Amos, T., Gournay, K., and Skapinakis, P. 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, M. 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, J. A thirty-year follow-up of treatment effects. *American psychologist*, 33(3):284, 1978.

McCord, J. 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, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *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, A. Why smart statistics are the key to fighting crime. Ted Talk, January 2014.

Miller, A. J. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425, 1984.

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

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

Netter, B. 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, M.-A., Zgoba, K. M., Tita, G. E., and Lee, S. S. Predicting recidivism in homicide offenders using classification tree analysis. *Homicide studies*, 15(2):154–176, 2011.

Penner, E. K., Viljoen, J. L., Douglas, K. S., and Roesch, R. Procedural justice versus risk factors for offending: Predicting recidivism in youth. Educational Publishing Foundation, 2013.

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. R. *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, 2014. URL `http://www.R-project.org`.

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

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

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

Silver, E. and Chow-Martin, L. A multiple models approach to assessing recidivism risk implications for judicial decision making. *Criminal justice and behavior*, 29(5):538–568, 2002.

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

Skeem, J. L. and Monahan, J. Current directions in violence risk assessment. *Current Directions in Psychological Science*, 20(1):38–42, 2011.

Stalans, L. J., Yarnold, P. R., Seng, M., Olson, D. E., and Repp, M. Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law and human behavior*, 28(3): 253, 2004.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H., and Banks, S. A classification tree approach to the development of actuarial violence risk assessment tools. *Law and human behavior*, 24(1):83–100, 2000.

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

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

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

Tollenaar, N. and van der Heijden, P. 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, S., Hess, J., and Jannetta, J. 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. 2014. doi: 10.3886/ICPSR03355.v8. URL `http://doi.org/10.3886/ICPSR03355.v8`.

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`.

U.S. Sentencing Commission. Measuring recidivism: The criminal history computation of the federal sentencing guidelines. 2004.

U.S. Sentencing Commission. A comparison of the federal sentencing guidelines criminal history category and the U.S. parole commission salient factor score. January 2005.

Ustun, B. and Rudin, C. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*, 2014.

Ustun, B. and Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *arXiv preprint arXiv:1502.04269*, 2015.

Ustun, B., Traca, S., and Rudin, C. Supersparse linear integer models for predictive scoring systems. In *AAAI (Late-Breaking Developments)*, 2013.

Webster, C. Risk assessment: Actuarial instruments & structured clinical guides, 2013.

Webster, C. D. et al. Hcr-20: Assessing risk for violence. 1997.

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

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

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

Yang, M., Wong, S. C., and Coid, J. The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological bulletin*, 136(5):740, 2010.

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