Journal of Quantitative Criminology (2023) 39:519–581
https://doi.org/10.1007/s10940-022-09545-w

ORIGINAL PAPER

In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction

Caroline Wang[1] · Bin Han[2] · Bhrij Patel[3] · Cynthia Rudin[3,4]

Accepted: 6 March 2022 / Published online: 28 March 2022

Abstract

Objectives We study interpretable recidivism prediction using machine learning (ML) models and analyze performance in terms of prediction ability, sparsity, and fairness. Unlike previous works, this study trains interpretable models that output probabilities rather than binary predictions, and uses quantitative fairness definitions to assess the models. This study also examines whether models can generalize across geographic locations.

Methods We generated black-box and interpretable ML models on two different criminal recidivism datasets from Florida and Kentucky. We compared predictive performance and fairness of these models against two methods that are currently used in the justice system to predict pretrial recidivism: the Arnold PSA and COMPAS. We evaluated predictive performance of all models on predicting six different types of crime over two time spans.

Results Several interpretable ML models can predict recidivism as well as black-box ML models and are more accurate than COMPAS or the Arnold PSA. These models are potentially useful in practice. Similar to the Arnold PSA, some of these interpretable models can be written down as a simple table. Others can be displayed using a set of visualizations. Our geographic analysis indicates that ML models should be trained separately for separate locations and updated over time. We also present a fairness analysis for the interpretable models.

Conclusions Interpretable ML models can perform just as well as non-interpretable methods and currently-used risk assessment scales, in terms of both prediction accuracy and fairness. ML models might be more accurate when trained separately for distinct locations and kept up-to-date.

Keywords Criminal recidivism · Interpretability · Fairness · COMPAS · Machine learning

C. Wang, B. Han: These authors contributed equally to this work.

✉ Bin Han
bh193@uw.edu

[1] Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA

[2] Department of Information Science, The University of Washington, Seattle, WA 98195, USA

[3] Department of Computer Science, Duke University, Durham, NC 27708, USA

[4] Department of Statistical Science, Duke University, LSRC D342, Research Drive, Durham, NC 27708, USA

Abstract of Main Results for Criminal Justice Practitioners

Our goal is to study the predictive performance, interpretability, and fairness of machine learning (ML) models for pretrial recidivism prediction. ML methods are known for their ability to automatically generate high-performance models that sometimes even surpass human performance from data alone. However, many of the most common ML approaches produce "black-box" models—models that perform well, but are too complicated for humans to understand. "Interpretable" ML techniques seek to produce the best of both worlds: models that perform as well as black-box approaches, but also are understandable to humans. In this study, we generate multiple black-box and interpretable ML models. We compare the predictive performance and fairness of the ML models we generate, against two models that are currently used in the justice system to predict pretrial recidivism—namely, the Risk of General Recidivism and Risk of Violent Recidivism scores from the COMPAS suite, and the New Criminal Activity and New Violent Criminal Activity scores from the Arnold Public Safety Assessment.

We first evaluate the predictive performance of all models, based on their ability to predict recidivism for six different types of crime: `general, violent, drug, property, felony, and misdemeanor`. Recidivism is defined as a new charge for which an individual is *convicted* within a specified time frame, which we specify as 6 months or 2 years. We consider each type of recidivism over the two time periods to control for time, rather than to consider predictions over an arbitrarily long or short pretrial period. Next, we examine whether a model constructed using data from one region suffers in predictive performance when applied to predict recidivism in another region. Finally, we consider the latest fairness definitions created by the ML community. Using these definitions, we examine the behavior of the interpretable models, COMPAS, and the Arnold Public Safety Assessment, on race and gender subgroups.

Our findings and contributions can be summarized as follows:

- We contribute a set of interpretable ML models that can predict recidivism as well as black-box ML methods and better than COMPAS or the Arnold Public Safety Assessment for the location they were designed for. These models are potentially useful in practice. Similar to the Arnold Public Safety Assessment, some of these interpretable models can be written down as a simple table that fits on one page of paper. Others can be displayed using a set of visualizations.
- We find that recidivism prediction models that are constructed using data from one location do not tend to perform as well when they are used to predict recidivism in another location, leading us to conclude that models should be constructed on data from the location where they are meant to be used, and updated periodically over time.
- We reviewed the recent literature on algorithmic fairness, but most of the fairness criteria don't pertain to risk scores, they pertain only to yes/no classification decisions. Since we are interested in criminal justice risk scores in this work, the vast majority of the algorithmic fairness criteria are not relevant. We chose to focus on the evaluation criteria that were relevant, namely calibration and balanced group AUC (BG-AUC). We present an analysis of these fairness measures for two of the interpretable models (Risk-SLIM and Explainable Boosting Machine) and the Arnold Public Safety Assessment (New Criminal Activity score) on the two-year general recidivism outcome in Kentucky. We found that the fairness criteria were approximately met for both interpretable models for blacks/whites and males/females—that is, the models were fair according

to these criteria. The Arnold Public Safety Assessment's New Criminal Activity score failed to satisfy calibration for higher values of the score. The results on fairness were not as consistent for the "Other" race category. It is difficult to interpret the fairness result for the "Other" race category, due to low-resolution race data.

Introduction

Predicting criminal recidivism using statistics has been the subject of almost a hundred years of research in criminal justice, psychology, and law. Today, actuarial risk assessments are in wide use across many countries, helping judges make life-changing decisions in pretrial release, sentencing, and probation. Risk assessments can help reduce costs, racial disparity, and incarceration rates—and these benefits have already been realized in some jurisdictions (Berk 2017). However, some of the most widely used algorithms are secret, black-box models created by corporations. As a result, individuals affected by these algorithms cannot know how these decisions were made, or whether they were made in error. These problems resulted in various lawsuits over the last decade, and came to the fore in 2016, when investigative journalists from the nonprofit organization ProPublica claimed that the COMPAS black-box recidivism prediction model — standing for **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions — was rife with racial bias (Larson et al. 2016; Freeman 2016).

Though ProPublica's findings were not validated (Rudin et al. 2020; Flores et al. 2016; Dieterich et al. 2016), the COMPAS scandal demonstrated the issues with for-profit, secret algorithms making decisions in the justice system—namely, possible violations of defendants' due process rights, difficulty in ensuring that the scores were calculated based on correct inputs, and the lack of independent fairness or performance guarantees. It highlighted the ways that systemic bias in data can be propagated into the future, and was symptomatic of growing public distrust in the algorithms that impact our daily lives (Barabas et al. 2019; O'Neil 2016; Wexler 2017).

To prevent errors, prevent due process violations, allow independent validation of models, and gain public trust, we must create interpretable and fair models. Fortunately, techniques for interpretable ML and theories of fairness have advanced considerably over the last few years. Multiple works have demonstrated that publicly available interpretable ML algorithms can perform as well as black-box ML algorithms (Zeng et al. 2017; Angelino et al. 2018; Lou et al. 2013). Moreover, high-dimensional data sets on criminal recidivism have become increasingly available. However, most ML papers are focused on algorithm construction and do not consider factors such as data quality or ease of computating model predictions, which are paramount for creating models that would be useful in practice. To our knowledge, there is only one prior work (Soares and Angelov 2019) that jointly considers interpretability, fairness, and predictive performance; however, it does not do so in a comprehensive way and focuses primarily on the design of a new algorithm.

Beyond the problem of model optimization, various methodological questions remain with existing risk assessment systems. First, existing systems—such as COMPAS (Correctional Offender Management Profiling System for Alternative Sanctions) and LSI-R (Level of Service Inventory Revised)—are often used across states, or even countries, with only minor normalization (MHS Assessments 2017; Northpointe 2013). However, populations in different states can significantly differ because the data generation process is not the same, so applying the same model across states may not lead to the best possible performance. Second, empirical evidence indicates that the underlying probability distribution of recidivism has changed over time in multiple locations (Gelb et al. 2018). For instance, a significant shift in the age distribution—a key predictor in many recidivism prediction models—has been observed in New York (Kim et al. 2016). Thus, rather than using a static model with uneven performance across districts, a better solution might be to algorithmically generate models, so that they can be trained for specific locations and retrained if recidivism distributions shift over time.

Using modern tools of both interpretable and black-box ML, we revisit the recidivism prediction problem. We define recidivism as a new charge that an individual is convicted for within a certain time frame: 6 months or 2 years. We find that (1) black-box models do not perform significantly better than interpretable models for any of the twelve recidivism problems we consider. (2) Interpretable models generally perform better than existing actuarial risk assessments. (3) Models do not generalize well across regions. (4) Only a small subset of the many proposed fairness definitions can be applied to regression problems and they vary across different models. We also note that existing techniques to enforce fairness generally require non-interpretable transformations, and therefore do not work well with interpretable models.

This paper is structured as follows. The "Background" section discusses the evolution of risk assessment in America, the current debate over risk assessments, and briefly reviews the ML literature on risk assessment. The "Data" section describes the study's data sources. The "Methodology" section discusses aspects of our methodology, including the prediction problems, problem setup, and the existing risk assessments we compare against. The "Baseline Machine Learning Methods" section presents the performance of baseline, non-interpretable ML methods, while the "Interpretable Machine Learning Methods" section presents the performance of interpretable ML methods. The "Recidivism Prediction Models Do Not Generalize Well Across Regions" section examines the generalization of recidivism prediction models across states. In the "Fairness" section, we describe the selection of fairness metrics and assess the fairness of the interpretable models. Finally, in the "Discussion and Future Work" section, we discuss broader impacts and future lines of inquiry.

Contribution

Our main contribution is a set of interpretable, risk-calibrated linear models that perform approximately as well as— sometimes better than—existing actuarial risk assessments, and predict specific crime types. Other important aspects of our contribution are as follows:

- We consider multiple types of recidivism (`general`, `violent`, `drug`, `property`, `felony`, and `misdemeanor`) at two time scales (6-month, two-year) for a total of 12 prediction problems.
- Our analysis was conducted on two criminal history data sets (one from Broward County, Florida, and the other from the state of Kentucky), which allowed us to understand variability in model performance across locations. We found that models do not generalize well between locations, and conclude that models should be trained on data from the location where they are meant to be used.

- The risk models trained as part of this study are interpretable, and could potentially be useful in practice after a careful, location-specific evaluation of their accuracy and fairness.
- We provide an understanding of how to evaluate both interpretability and fairness in an important real application. The same type of analysis could be ported to financial lending decisions, hiring decisions, or any other type of high-stakes decisions that require an assessment of both interpretability and fairness.

Similar to Zeng et al. (2017), we use ML techniques optimized for interpretability, and address multiple prediction problems. This work is an improvement in the following ways. We use interpretable ML techniques to create risk scores representing probabilities of recidivism rather than making binary predictions—techniques which were not available at the time of publication for Zeng et al. (2017). We compare with COMPAS and the Arnold Public Safety Assessment (PSA), two models currently used in the justice system, whereas Zeng et al. (2017) compared only with other ML methods. We use data obtained at the pretrial stage rather than at prison-release. Since many jurisdictions utilize prediction instruments to determine pretrial release, this better aligns with the use cases of risk scores. Our data come from two locations, and include more detailed information than in Zeng et al. (2017), and are more recent than 1994. Finally, models are assessed for multiple definitions of fairness in addition to performance.

Background

Algorithmic risk assessment dates back to the early 1900s (Bureau of Justice Assistance 2020), and is used today at various stages of the criminal justice system, such as at pretrial, parole, probation, or even sentencing. In this work, we focus on forecasting recidivism at the pretrial stage. Though some states have implemented their own tools (Virginia, Pennsylvania, Kentucky), many utilize systems produced by companies, non-profits and other organizations (Kehl et al. 2017). These externally-produced risk assessments and some of the jurisdictions that utilize them include COMPAS (Florida, Michigan, Wisconsin, Wyoming, New Mexico), the Public Safety Assessment (New Jersey, Arizona, Kentucky,[1] Phoenix, Chicago, Houston), LSI-R (Delaware, Colorado, Hawaii), and the Ohio Risk Assessment System (Public Safety Assessment 2019; Latessa et al. 2009; Electronic Privacy Information Center 2016). The United States is not alone in using actuarial risk assessments. Canada uses the Static-2002 to assess risk of violent and sexual recidivism (Hanson and Thornton 2003); the Netherlands uses the Quickscan to assess static and dynamic risks of recidivism (Tollenaar and van der Heijden 2013); the U.K. uses the Offender Group Reconviction Scale to predict reoffense while on probation (Howard et al. 2009).

The Debate over Risk Assessments

Since the inception of actuarial risk assessments, there has been debate over whether they should be used in the criminal justice system at all. Proponents claim that statistical models reduce overall violence levels and ensure the most efficient use of treatment and

[1] Kentucky created and implemented their own tool in 2006 but transitioned to the Arnold PSA in 2013.

rehabilitative resources by helping judges identify the individuals that are truly dangerous. A large body of evidence appears to support this claim. Various studies have shown that statistical models are more accurate than human experts (Dawes et al. 1989; Grove and Meehl 1996). Others have shown that a small percentage of individuals commit the majority of crimes (Wolfgang 1987; Sherman 2007; Milgram 2014), indicating that correctly identifying dangerous individuals could lead to substantial decreases in violence levels. Proponents also claim that risk assessments are instrumental to reducing racial/economic disparity, allocating social services, and reducing mass incarceration (James 2018). In particular, some jurisdictions have adopted risk assessments at the pretrial stage to replace cash bail, since cash bail is widely viewed as biased against poor defendants (Zweig 2010; Desmarais et al. 2019).

In practice, reducing overall violence levels, mass incarceration, and racial/economic disparity through actuarial risk assessment is complex (Ludwig and Mullainathan 2021). Critics have argued that as recidivism prediction models always rely on racially-biased features such as arrest records, actuarial risk assessment will only exacerbate racial and socioeconomic disparity, and should therefore be abolished (The Leadership Conference on Civil and Human Rights 2018; Pretrial Justice Institute 2020). In a well-known incident, ProPublica claimed that COMPAS was biased against African-Americans because there was a disparity in false positive rates and false negative rates between African-Americans and Caucasians (Angwin et al. 2016). Follow-up research showed that this bias was likely a property of the data generation process rather than the COMPAS model, and that even a model that only relied on age showed a similar disparity in false positives and false negatives (Rudin et al. 2020). Actuarial risk assessment might be vulnerable to feature bias, but it is important to remember that other parts of the court system (such as bail and sentencing guidelines for judges) are not immune to feature bias either—they also use criminal history and arrest records. Similarly, in one of the first large-scale empirical studies, Stevenson (2018) showed that in Kentucky, the use of the Arnold PSA seemingly increased disparity between whites and blacks at pretrial release. Because the risk scores were applied differently by judges in different counties, it seemed that white people benefited more than black people in terms of pretrial release numbers—but within the same county, white and black defendants saw similar *increases* in release. Thus, rather than eliminating the use of risk scores, using them uniformly across counties may have made risk assessments more fair across the state, and could have reduced overall incarceration.

Others have argued that a fundamental flaw with risk assessments is that their simple labels obscure the true uncertainty behind their predictions (Barabas et al. 2019). This may be true for currently used risk assessments, but merely underscores the necessity for researchers to develop models that *do* quantify uncertainty. While actuarial risk assessments are not perfect, we must remember that in the absence of risk assessments, judges can only rely on their intuition—and human intuition has been shown to be less reliable than statistical models (Dawes et al. 1989; Grove and Meehl 1996; Desmarais et al. 2019; Skeem et al. 2020).

Another problem is that some of the most widely used risk prediction algorithms are for-profit and secret. For instance, while COMPAS's guidelines are published and validation studies have been performed, the full forms of the models are not available and some of the validation studies do not conform to standards of open science because they do not publish the validation data (Garrett and Stevenson 2020), thus yielding concerns over due process rights. In the 2017 Wisconsin Supreme Court case, Loomis v. Wisconsin, Loomis challenged the use of the proprietary risk prediction software, COMPAS, on the grounds that this violated his due process and equal protection rights (Freeman 2016). Yet today, there are plenty of equally accurate, transparent risk prediction tools that publish their guidelines and full models. See Table 7 in the Appendix for examples. In this article, we compare against the Arnold PSA, an interpretable and publicly available tool which is used in multiple jurisdictions.

There is also a general fear that the use of risk assessments could lead to situations similar to those depicted in the movie, “Minority Report.” In Minority Report, individuals were punished *before* they committed a crime based on oracles’ visions of the future. However, one of the major principles common to American criminal justice texts (Roberts and von Hirsch 2010; Frase et al. 2015) is that individuals should be punished based on the crimes they committed in the past. This illustrates why risk assessments have played only a minor role in sentencing. In reality, risk prediction tools are most heavily used in bail, parole, and social services decisions.

Risk scores will not solve everything, but abolishing risk assessment without a useful alternative plan will not solve the problems above either. Reducing feature bias requires generations of community investment; jurisdictions must train judges on how to use risk scores; and communities must provide treatment resources for those deemed high risk. Risk assessments and other evidence-driven practices can be an important part of this solution. In the most recent revision of the Model Penal Code, the American Law Institute has supported giving people shorter prison terms or sending them to the community through the use of risk assessment tools (Starr 2015; American Law Institute 2017). By providing simple and interpretable risk scores, we hope to mitigate the possibility that risk assessments are miscomputed, and enable judges and defendants to fully understand their scores.

Black-Box and Interpretable ML for Predicting Criminal Recidivism

There is an abundance of past research on using ML methods to predict criminal recidivism. However, many of these studies utilize black-box, non-interpretable models, and only optimize for predictive performance. For instance, Neuilly et al. (2011) used random forests to predict homicide offender recidivism. Other black-box models applied to this problem include stochastic gradient boosting (Friedman 2002), neural networks (Palocsay et al. 2000), and ensemble methods (Singh and Mohapatra 2021).

In comparison, there is relatively little work using interpretable ML techniques to forecast recidivism, and there is not a consensus on how interpretability should be defined in this domain. Berk et al. (2005) used classical decision trees to build a simple screener for forecasting domestic violence for the Los Angeles Sheriff’s Department. Goel et al. (2016) created a simple scoring system by rounding logistic regression coefficients, which helped address stop-and-frisk for the New York Police Department. Zeng et al. (2017) was the first work using modern ML methods that globally optimized over the space of sparse linear integer models to predict criminal recidivism. Despite the range of interpretable models that have been applied to the criminal recidivism problem, a common thread among these works is that simple, interpretable models can do just as well as black-box models, and better than humans. For instance, Angelino et al. (2018) found that COMPAS shows no benefit in accuracy over very simple ML models involving age and criminal history. Skeem et al. (2020) showed that algorithms outperformed humans on predicting criminal recidivism in three data sets, and demonstrated that the performance gap was especially large when abundant risk factors were considered for risk prediction.

The approaches outlined above achieved interpretability through training models with interpretable forms. Another major approach is *post-hoc explainability*, in which a simpler model provides insights into a black-box model. However, post-hoc explanations are notoriously unreliable, or are not thorough enough to fully explain the black-box model (Rudin 2019). Additionally, there seems to be no clear benefit of black-box models over inherently interpretable models in terms of prediction accuracy on the criminal recidivism problem (Zeng et al. 2017; Tollenaar and van der Heijden 2013). Thus, for a high-stakes problem such as predicting criminal recidivism, we choose not to utilize these methods.

In fact, there have been cases in criminal justice where post-hoc explanations led to incorrect conclusions and pervasive misconceptions about what information some of the most common recidivism models use. The 2016 COMPAS scandal—where ProPublica reporters accused the proprietary COMPAS risk scores of an explicit dependence on race (Angwin et al. 2016)—was caused by a flawed, post-hoc explanation of a black-box model. In particular, ProPublica reasoned that if a post-hoc explanation of COMPAS depended linearly on race, then COMPAS depended on race, even after controlling for age and criminal history. However, as Rudin et al. (2020) demonstrated, just because an explanation model depends on a variable *does not* mean that the black box model depends on that variable. Thus, ProPublica's reasoning was incorrect. In particular, this analysis found that COMPAS does not seem to depend linearly on some of its input variables (age), and does not seem to depend on race after conditioning on age and criminal history variables. Criminologists have also criticized the ProPublica work for other reasons (Flores et al. 2016). Despite the flaws in the ProPublica article, it is widely viewed as being a landmark paper on fairness in ML.

A notable advantage of interpretable modelling for criminal justice is that some interpretable models allow a decision-maker to incorporate factors not in the database in a way that black-box models cannot. For instance, scoring systems—linear models with integer coefficients—place all of the model inputs onto the same scale: every input receives a number of points. The points of each factor in the model provide clarity on how important each input is relative to the others.

Fair Machine Learning

Fairness is a crucial property of risk scores. As such, the recidivism prediction problem is a key motivator for many of these works. However, recidivism prediction is rarely the primary focus of fairness papers. Many of these papers seek to make theoretical contributions by proposing definitions of fairness and creating algorithms to achieve these definitions, using recidivism prediction as a case study (Hardt et al. 2016; Agarwal et al. 2018). Others have proven fairness impossibility theorems, showing when different fairness constraints cannot be achieved simultaneously. For instance, the two fairness definitions at the heart of the debate over COMPAS' fairness (calibration and balance for positive/negative class) cannot be achieved simultaneously in nontrivial cases. However, by placing relaxations on the conditions, the fairness definitions can be *approximately* satisfied simultaneously. (Kleinberg et al. 2017; Pleiss et al. 2017). These theorems show that many fairness definitions directly conflict, so there cannot be a single universal definition of fairness (Kleinberg et al. 2017; Binns 2018; Verma and Rubin 2018). Moreover, there is often a trade-off between performance and fairness (Berk et al. 2017b; Richard 2019; Corbett-Davies et al. 2017). The emerging consensus is that any decision about the "best" definition of fairness must rely heavily on model characteristics and domain-specific expertise.

The question of what should count as fair in criminal recidivism prediction can be answered by discussion among ethicists, judges, legislators, and stakeholders in the criminal justice system. Existing American anti-discrimination law provides a general legal framework for addressing this question. Under Title VII of the Civil Rights Act of 1964, there are two theories of liability: disparate impact and disparate treatment (Barocas and Selbst 2016). In this article, we use the definitions of fairness from the field of fair ML, as they apply directly to ML models and are more specific than the general legal guidelines of disparate impact and treatment. Moreover, some of the definitions of fairness proposed by the field of fair ML community are inspired by these guidelines. See Corbett-Davies and Goel (2018) for a detailed discussion of the relationship between algorithmic definitions of fairness and economic/legal definitions of discrimination.

Data

In this study, we used criminal history data sets from Broward County, Florida, and the state of Kentucky, allowing us to analyze how models perform across regions. The Broward County data set consists of publicly available criminal history and court data from Broward County, Florida. This data set consists of the full criminal history, probational history, and demographic data for the 11,757 individuals who received COMPAS scores at the pretrial stage from 2013-2014 (as released by ProPublica Angwin et al. 2016). The probational history was computed from public criminal records released by the Broward Clerk's Office. Though the full data set includes 11,757 individuals, this analysis includes only the 1954 for which we could also compute the PSA. We processed the Broward data using the same methods as Rudin et al. (2020). From the processed data, we computed various features such as number of prior arrests, prior charges, prior felonies, prior misdemeanors, etc.

The Kentucky pretrial and criminal court data was provided by the Department of Shared Services, Research and Statistics in Kentucky. The data came from two systems: the Pretrial Services Information Management System (PRIM) and CourtNet. PRIM data contain records regarding defendants, interviews, PRIM cases, bonds, etc., that were connected with the pretrial service interviews conducted between July 1, 2009 and June 30, 2018. CourtNet data provide further information about cases, charges, sentences, dispositions, etc. In total, the Kentucky data set consists of over 25 million tuples. When constructing features from the Kentucky data set, we computed features that were as similar as possible to the Broward features (e.g., prior arrests, prior charges with different types of crimes, age at current charge) in order to compare models between the two regions. There are several features from Broward data which could not be computed from the Kentucky data, such as "age at first offense" and "prior juvenile charges." A limitation of the Kentucky data set is that the policies governing risk assessments changed over the period when the data was gathered, possibly impacting the consistency of the data collection.

A difference in the data processing between the two data sets is that when constructing prediction features and predictive labels, we considered non-convicted charges in the Broward data, but considered convicted charges in the Kentucky data. The reason for this choice is sample size. The processed Broward data contains only 1954 records, and limiting the scope to convicted charges would yield only 1297 records. The use of convicted versus non-convicted charges between the two regions might explain some discrepancies in the results in the "Recidivism Prediction Models Do Not Generalize Well Across Regions" section, where we discuss the generalization of recidivism prediction models between states. Note that many models currently implemented within the justice system rely on non-convicted charges such as counts of prior arrests, but for the applications such as bail and parole, the use of non-convicted charges could be problematic—it holds individuals accountable for crimes that they may not have committed.

Please refer to the "Broward Data Processing" and "Kentucky Data Processing" sections in the Appendix for more details on data processing and a full list of features.

Methodology

Throughout our analysis, we compare with two tools that are currently used to predict recidivism in the U.S. justice system: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) and the Arnold PSA (Public Safety Assessment, created by Arnold Ventures, which was previously named the Laura and John Arnold Foundation). Although we would have liked to compare against more assessment tools, many of them use data that are not publicly available, or are owned by for-profit companies that do not release their models. For a detailed discussion of the other risk assessments we considered and the features we were missing, please consult the "Kentucky Data Processing" section in the Appendix.

More specifically, we compared our models against the Arnold PSA's New Criminal Activity (NCA) and New Violent Criminal Activity (NVCA) scores on the `general` and `violent` recidivism problems, respectively. Note that the time-frames and labels for prediction are important here, and our choices distinguish this study from past works on recidivism prediction. Let us explain the time-frames next.

It is important that we chose *fixed* time-frames for prediction, in our case, 2 years or 6 months past the current charge dates. In reality, the scores are used to assess risks during the pretrial period. However, there is a huge amount of variation in pretrial periods, which can span a few days to a few years: the average pretrial time span in Kentucky is 109 days, and could last upwards of 3-4 years. Since the pretrial period depends on the jurisdiction, we chose to fix time spans so that the models do not depend on the policy used for determining how long the pretrial period would be. That way, the risk calculations we produce depend mainly on the inherent characteristics of the individual, rather than the length of the pretrial period, which is potentially a characteristic of the jurisdiction. Also, this way, individuals with the same propensity to commit a new crime within 6 months (or 2 years) are given identical risk scores, even if they have different expected time periods until their respective trials. The 6-month time span represents an approximate length of pretrial period. The two-year time span provides more balanced labels, since 2 years provides more time to commit crimes than 6 months. Additionally, our evaluation metric is AUC, which is a rank statistic, and considers relative risk rather than absolute risk; that is, an individual who actually commits crimes within 2 years of their current charge date should be ranked higher than an individual who does not. The relative risk within the two-year time-frame is related to the relative risk for other shorter or longer time-frames, allowing these models to potentially generalize to varying pretrial time-frames.

Another important aspect of our prediction problems is the *definition of recidivism* we chose. We predict the occurrence of a *convicted* charge within 6 months/2 years for Kentucky. In other words, we would like to predict whether someone will be arrested, within 6 month or 2 years from their current charge, for another crime that they were later convicted for. This definition potentially alleviates a due process concern: if we instead include non-convicted charges, our models might be more likely to predict who will be arrested, which is tied to policing practices. For Broward, where we did not have conviction information for later charges, we predicted *any charge* within 6 months/2 years, which is the typical approach to recidivism prediction.

In Broward, we directly computed Arnold PSA scores, as the Arnold PSA is publicly available. The features used by the Arnold PSA are provided in Tables 12 & 13 in the Appendix. For Kentucky, we used the unscaled Arnold PSA scores that came with the data set, because those are what are reported to the judges in Kentucky. We compared against COMPAS' Risk of General Recidivism and Risk of Violent Recidivism risk scores on the two-year `general` and two-year `violent` prediction problems, respectively. Note that both models are designed to predict recidivism within 2 years. The COMPAS suite is proprietary, but COMPAS General and Violent scores were provided with the Broward County data set. We do not compare against COMPAS on the Kentucky data set. The COMPAS General and COMPAS Violent scores appear to have been developed for a parole population (Northpointe 2013), but have been applied for pretrial decisions in Broward. In this study, we consider the COMPAS scores for the outcomes they were actually applied for (pretrial decisions), rather than the outcomes they were developed for (parole decisions).

In the "Baseline Machine Learning Methods" and "Recidivism Prediction Models Do Not Generalize Well Across Regions" sections, we compare the performance of black-box and interpretable algorithms on the Broward and Kentucky data sets. We caution readers against comparing an algorithm's performance in Broward with its performance in Kentucky. An algorithm's differences in performance between the data sets could be attributed to the many differences between the two regions. For instance, the Broward data set is at the county level while the Kentucky data set is at the state level. As the Kentucky data is at the state level, it embeds diverse information about 120 counties (e.g., demographics, legislation, culture, local policing practices). Thus, in the "Baseline Machine Learning Methods" and "Recidivism Prediction Models Do Not Generalize Well Across Regions" sections, the comparisons between baseline models and interpretable models are conducted *within* each data set. In the "Recidivism Prediction Models Do Not Generalize Well Across Regions" section, we discus in detail the regional differences between Broward County and Kentucky, and present a set of experiments that illustrate model performance gaps resulting from these regional differences.

Prediction Labels

In addition to two-year `general` recidivism and two-year `violent` recidivism—the two types of criminal recidivism considered by COMPAS and the PSA—we computed recidivism prediction labels specific to various crime types, such as `property`, `drug` related recidivism and recidivism with `felony` or `misdemeanor` level charges. For clarity, we apply the `typewrite` font to indicate prediction tasks. Note that an individual could have multiple positive labels, indicating that the newly committed crime involves multiple charge types. We defined recidivism as a recorded charge within a certain time frame. Out of all the possible recidivism prediction tasks we considered, we selected the six most balanced: `general`, `violent`, `drug`, `property`, `felony`, and `misdemeanor`. To investigate the effect of temporal scale on predictive performance, we generated these six tasks using the time windows `two-years` and `six-months` after the current charge date (or release date, if the individual went to prison for their current charge), for a total of

Table 1 Label distributions for Broward and Kentucky

Labels	Kentucky		Broward	
	Two year $P(y_i = 1)$ (%)	Six month $P(y_i = 1)$ (%)	Two year $P(y_i = 1)$ (%)	Six month $P(y_i = 1)$ (%)
`General`	20.4	5.7	45.5	21.8
`Violent`	3.4	0.7	21.0	8.4
`Drug`	8.7	2.0	9.3	4.0
`Property`	3.9	0.9	9.0	5.0
`Felony`	9.6	2.4	17.6	8.9
`Misdemeanor`	15.6	3.9	27.2	12.5

$y_i = 1$ if the defendant had the corresponding type (`general, violent, drug etc.`) of charge within 2 years (resp. 6 months) from current charge/release date

twelve tasks. The summary of prediction tasks and the base rate of recidivism for each task is provided in Table 1.

Problem Setup

Due to the binary nature of recidivism tasks, we approached these prediction problems as binary classification problems, but do not binarize the final predicted probabilities/scores of the ML models for the following reasons. First, existing risk scores are usually nonbinary. For instance, the Arnold PSA's unscaled New Criminal Activity (NCA) score takes integer values from 0 through 13, while the COMPAS Risk of Recidivism and Risk of Violent Recidivism scores take on integer values from 1 through 10 (Northpointe 2013; Public Safety Assessment 2019). Second, we want to create more nuanced risk scores both by predicting highly specific types of recidivism, in addition to coarser categories like general recidivism, and by presenting nonbinary scores which reflect a range of risk values. Since the predictions are nonbinary, we use Area Under the Curve (AUC) as our evaluation metric. This decision also impacts the fairness metrics we assess, which we discuss in the "Fairness" section. We applied nested cross validation process to train the models. Please refer to the "Nested Cross Validation Procedure" section in the Appendix for the details.

Baseline Machine Learning Methods

To provide a basis of comparison for the interpretable models presented in the "Interpretable Machine Learning Methods" section, we evaluated the performance of six common, non-interpretable ML methods in this section. Baseline models and descriptions are provided below. The tuned hyperparameters and packages used for each problem are provided in "Hyperparameters" in the Appendix.

- ℓ_2 Penalized Logistic Regression: To prevent overfitting, there is an ℓ_2 penalty term on the sum of squared coefficients in the loss function for logistic regression. Although

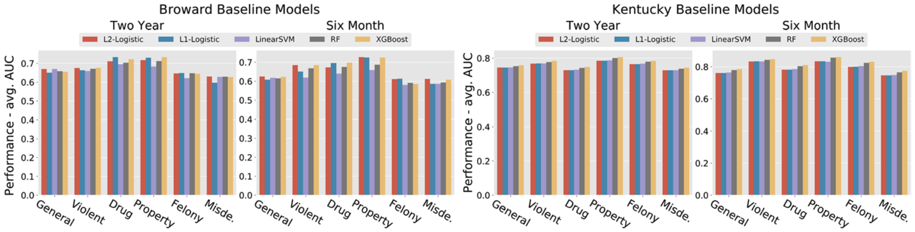


Fig. 1 Visualizations of Broward and Kentucky baseline results

this method produces linear models, we consider ℓ_2-penalized logistic regression to be non-interpretable because if the number of input features is large, there could be a large number of nonzero terms in the model.

- ℓ_1 Penalized Logistic Regression: To prevent overfitting, there is an ℓ_1 penalty term on the sum of absolute values of coefficients in the loss function for logistic regression. This algorithm creates sparser models than ℓ_2 penalized logistic regression. Notice that the sparsity of the model depends on the magnitude of the penalty and must be balanced with consideration of prediction performance. In our experiments, ℓ_1 models with Broward data were sparse yet maintained good predictive performances. However, the best ℓ_1 models with Kentucky data still had too many features, which made it difficult to interpret the results. Therefore, we classified ℓ_1-penalized logistic regression as a non-interpretable algorithm.
- SVM with a Linear Kernel (Vapnik and Chervonenkis 1964): An algorithm that outputs a hyperplane that separates two classes by maximizing the sum of margins between the hyperplane and all points. Incorrectly classified points are penalized. Although SVM with linear kernel yields a linear model, the concerns with ℓ_1 and ℓ_2 penalized logistic regressions apply here as well: the number of nonzero terms could be large, making it difficult to interpret the model.
- Random Forest (Breiman et al. 1984): An ensemble method that combines the predictions of multiple decision trees, each of which is trained on a bootstrap sample of the data. The implementation we use combines individual trees by averaging the probabilistic prediction of each tree. Random Forest is usually considered a black-box classifier because it is difficult to understand the individual contribution of each feature, which can be found in many trees, and the joint relationship between features.
- Boosted Decision Trees (Freund and Schapire 1997): An ensemble method where a sequence of weak classifiers (decision trees) are fit to weighted versions of the data. Similar to random forest, boosted decision trees produce black-box models because it is difficult to understand the joint relationships of the features. We use the XGBoost implementation (Chen and Guestrin 2016).

Broward Baseline Results

The performances of baseline algorithms on the Broward data are visualized in Fig. 1; details are presented in Table 14 in the Appendix. We noticed that in the two-year prediction problems, no algorithm consistently performs better than the others. Simple linear models can even outperform black-box models in some prediction problems. For instance, in the two-year prediction problems, ℓ_2-penalized logistic regression and LinearSVM tie in performance for the `general` recidivism prediction. XGBoost performs the best in `violent` and `property` prediction problems. ℓ_1-penalized logistic regression has the best performance in `drug` and `felony` prediction tasks, while ℓ_2-penalized logistic regression has the best performance in `misdemeanor` recidivism prediction. The largest performance gap is 5.1%, from `property` recidivism prediction. In the 6-month prediction problems, we see the same phenomenon that no single model dominates the others in performance. Overall, the performance gaps across baseline models for the `general`, `felony`, and `misdemeanor` prediction tasks are small, while other prediction problems have larger gaps.

Kentucky Baseline Results

The performances of baseline algorithms on the Kentucky data are visualized in Fig. 1; details are presented in Table 15. We noticed that complex and nonlinear baselines perform slightly better than linear models, potentially due to the larger size of the Kentucky data set (1956 records in Broward versus about 250K records in Kentucky). In particular, Random Forest and XGBoost uniformly perform slightly better than all the other algorithms on all prediction tasks, over both time periods we examined. XGBoost performs the best on all tasks. However, performance gaps, across all prediction problems and in both time frames, are very small. Thus, we conclude that all the baseline algorithms perform similarly over the Kentucky data set. On Kentucky, all algorithms perform slightly better on the 6-month recidivism period than on the two-year period.

Summary of Baseline Models' Results: We found that all baseline ML algorithms performed similarly across recidivism problems for the Kentucky data set. We also found that models performed better on the 6-month prediction problems than on the two-year problems on Kentucky data, but not on Broward data. These findings will be discussed throughout the following subsections.

Interpretable Machine Learning Methods

For recidivism prediction, we considered several different types of interpretable ML methods with different levels of interpretability, ranging from scoring systems to decision trees, to additive models. Since the Burgess model in 1928 (Burgess 1928), recidivism risk assessments have traditionally been scoring systems, which are sparse linear models with positive integer coefficients. A scoring system can be visualized as a simple scoring table or set of figures. There have only recently been algorithms designed to optimally learn scoring systems directly from data, without manual feature selection or rounding. Scoring systems have several advantages: they allow an understanding of how variables act *jointly* to form the prediction; they are understandable by non-experts; risks can be computed without a calculator; and they are consistent with the form of model that criminologists have built over the last century, where "points" are given to the individual, and the total points are transformed into a risk of recidivism. External information, such as risk factors that are not in any database, can be more easily incorporated into the risk score: it is much easier to determine how many points to assign to a new factor if the points are integer-valued for the known risk factors. For instance, we could choose to subtract three points for drug treatment, to counteract four points of past drug-related arrests.

While scoring systems appear to be the accepted standard for interpretability in the domain of criminal justice, imposing the constraints of linearity, sparsity, and integrality of coefficients could potentially be strong enough to reduce accuracy. Thus, we also consider modern algorithms that satisfy a subset of the conditions of interpretability: sparsity in features, ability to visualize/explain any variable interactions, linearity, integer coefficients. Specifically, we tested four interpretable ML algorithms: Classification and Regression Trees (CART), Explainable Boosting Machine (EBM), Additive Stumps, and RiskSLIM (Risk-Calibrated Supersparse Linear Integer Models). Algorithm specifics are articulated below and the tested hyperparameters are provided in the Appendix. We also tested two existing risk assessments—the Arnold PSA and COMPAS—and compared their performances to both baseline and interpretable ML models.

- Classification and Regression Trees (CART) (Breiman et al. 1984): A method to create decision trees by continuously splitting input features on certain values until a stopping criterion is satisfied. CART constructs binary trees using the feature and threshold that yields the largest information gain at each node. We constrain the maximum depth of the tree to ensure that it does not use too many features. CART models are nonlinear. They cannot be written as scoring systems, but can be written as logical models.
- Explainable Boosting Machine (EBM) (Lou et al. 2013): An algorithm that uses boosting to train Generalized Additive Models with a few interaction terms (GA^{2}Ms). The contribution by each feature and feature interaction pair can be visualized. The models are interpretable and modular, thus editable by experts. The models are generally not sparse, and cannot be written as scoring systems.
- RiskSLIM (Ustun and Rudin 2017, 2019): An algorithm that generates sparse linear models with integer coefficients that have risk-calibrated probabilities. The models generated by RiskSLIM have form similar to that of models used in criminal justice over the last century.
- Additive Stumps: An interpretable variation on ℓ_1-penalized logistic regression: for each feature, we generate multiple binary stumps (defined in the "Preprocessing Features into Binary Stumps" section), and apply ℓ_1-penalized logistic regression to these stumps. Ideally, the features will have monotonically increasing (or decreasing) contributions to the estimated probability of recidivism. Models constructed using this method generally use fewer features than those constructed with vanilla ℓ_1-penalized logistic regression. These models are flexible and nonlinear. These models also cannot be written as scoring systems because they are not sparse in the number of nonlinearities.
- Arnold PSA (Public Safety Assessment 2019): A widely used, publicly available, interpretable risk assessment system that consists of three scores: New Criminal Activity (NCA), New Violent Criminal Activity (NVCA), and Failure to Appear (FTA). We compare against the NCA for the `general` recidivism problem, and against the NVCA for the `violent` recidivism problem, on both two-year and 6-month time scales. The NCA has 7 factors, while the NVCA has 5 factors.
- COMPAS (Northpointe 2013): A widely used risk assessment system that consists of several scores, including the three that we study: Risk of General Recidivism (COMPAS General), Risk of Violent Recidivism (COMPAS Violent), and Risk of Failure to Appear. We compare against the COMPAS General score for the two-year `general` recidivism problem, and compare against the COMPAS Violent score for the two-year `violent` problem.

Preprocessing Features into Binary Stumps

We performed a data preprocessing technique for two of the interpretable ML algorithms: RiskSLIM and Additive Stumps. This technique consists of transforming all original features into binary stumps using Equation 1. Preprocessing the features into stumps allows us to include nonlinear interactions between the features (e.g. age, criminal history) and labels. It also allows us to visualize each Additive Stumps model as a set of monotonically increasing (or decreasing) curves. Formally, stumps are binary indicators, which are created by splitting features at pre-specified thresholds. For a feature $X^{(j)}$, and a set of threshold values $K \in \mathbb{R}$, we generate decreasing stumps $S_k^{(j)}$ for all $k \in K$ as follows:

$$S_k^{(j)} = \begin{cases} 1, & \text{for } X^{(j)} \leq k \\ 0, & \text{else} \end{cases} \tag{1}$$

We can generate increasing stumps analogously by substituting $\geq$ for $\leq$ in the definition above. The rationale behind the naming convention is as follows. Linear models constructed from increasing (respectively, decreasing) stumps have the nice property that if one sums the contribution from all stumps corresponding to a fixed original feature, i.e., $f(X^{(j)}) = \sum_{k \in K} c_k S_k^{(j)}$ for the feature $X^{(j)}$, and the coefficients c_k are mostly non-negative,[2] the resulting function $f(X^{(j)})$ is monotonically increasing (respectively decreasing), which is desirable for interpretability.

More concretely, the "age_at_current_charge" feature ranges from 18 to 70 in our data. For all age-related features, we construct decreasing stumps for $k = \{18, 19, ..., , 60\}$. We chose decreasing stumps for age features because based on past studies (Rudin et al. 2020; Stevenson and Slobogin 2018) and criminological theory (Gelb et al. 2018; Bindler and Hjalmarsson 2018; Bushway and Piehl 2007), the probability of recidivism decreases with age. On the other hand, intuitively, the probability of recidivism should increase as criminal history increases. Thus, we construct increasing stumps for the remaining features, which relate to criminal history.

To select a collection of stumps for the RiskSLIM and Additive Stumps model, we selected threshold values for all features by examining each feature visualization from EBM and choosing the threshold values that correspond to sharp drops in the predicted scores.

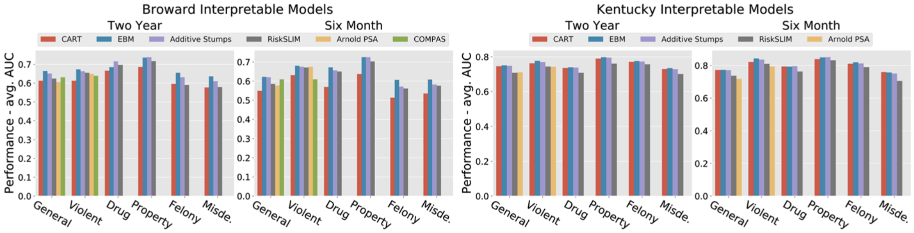


Fig. 2 Visualizations of Broward and Kentucky interpretable results

[2] For decreasing (respectively increasing) stumps, if the coefficient for the largest (respectively smallest) stump is negative, the function *f* will still be monotonic because the negative value will be subtracted from all values of the remaining stumps

Broward Prediction Results for Interpretable Models

Figure 2 show the results of interpretable models on the Broward data set; details are presented in Table 16 from the Appendix. For all prediction problems in both two-year and six-month prediction periods that we examined, we observed that CART consistently performed worse than other algorithms. Additive Stumps and EBM performed similarly on all the prediction tasks and outperformed other models, including the Arnold PSA and COMPAS, on most of the prediction tasks. The performances of the best interpretable models are very similar to that of the best baseline models—this is true for each of the prediction problems we considered. The AUC gaps between the best interpretable models and best baseline models for all two-year prediction tasks range from 0.3% to 1.7% in absolute value, and range from 0.2% to 2.6% for six-month prediction tasks. Prediction gaps from all other problems are smaller than 1%.

Kentucky Prediction Results for Interpretable Models

The Kentucky prediction results are visualized in Fig. 2; details are provided in Table 17. For all prediction problems in both time frames, CART, EBM, and Additive Stumps all had similar performances. RiskSLIM had relatively lower results compared to other interpretable models. All interpretable models performed better than the Arnold PSA, with the exception that the Arnold PSA performed slightly better (0.3%) than RiskSLIM on two-year `general` recidivism. Once more, we observed that the best interpretable models can perform approximately as well as the best black-box models (XGBoost). For the two-year prediction tasks, the differences in performance between the best interpretable and the best black-box models ranged from 0.7% to 0.9% in absolute value; for six-month problems, the difference ranged from 0.4% to 1.5%.

Summary of Interpretable Models' Results: We found that the best interpretable models performed approximately as well as the best black-box models, on both data sets and both time periods we considered, which is consistent with previous studies on other data sets (Zeng et al. 2017). The best interpretable models allow judges and defendants a better understanding of the predictions that the model outputs.

Tables and Visualizations of Interpretable Models

Each of the interpretable ML methods produces models that can be visualized, either as a decision tree (CART), scoring table (RiskSLIM), or as a set of visualizations (EBM, Additive Stumps). In this section, we present these tables and visualizations for EBM, Additive Stumps and RiskSLIM, to give a clearer understanding of each model's interpretability. Here we used the two-year `general` recidivism prediction problem on Kentucky data as an example.

EBM Models

The EBM package provides visualizations for each feature in the data set along with a bar chart of feature importance, both of which are displayed in an interactive dashboard. The dashboard allows users such as judges to see the scores corresponding to each bar or line by hovering the mouse over it. EBM models are not sparse in the number of features, so there could be visualizations for all features. Here, we show screenshots of the bar chart and visualizations for the three most important features.

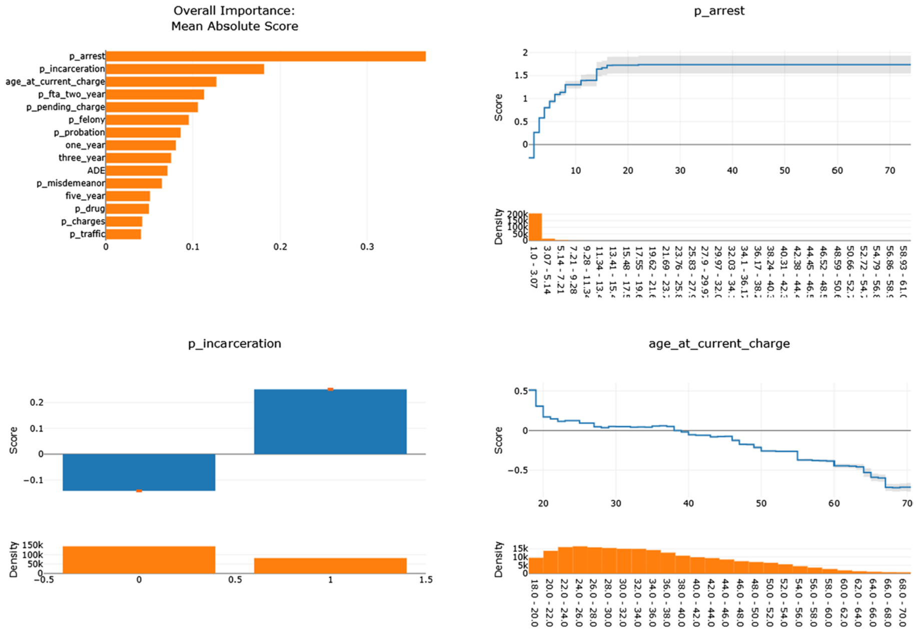


Fig. 3 Visualizations from EBM on two-year `general` recidivism. Top left: overall importance of each feature, ranked from the most important variable to least important. Remaining three: visualization for the contribution of the feature to the overall score (top) and histograms of feature values to show the distribution (bottom). Features contributions are visualized as bar charts if the feature takes binary value. The shaded grey area represents the confidence region. We see that as values get larger, there is more uncertainty in the predictions, which may be because we have fewer data points for such large feature values

EBM visualizations are similar to those from Additive Stumps, in that each feature's contribution to the score can be displayed separately. However, EBM scores do not tend to be monotonically increasing or decreasing in each feature. The visualization is provided as Fig. 3.

Additive Stumps

Additive Stumps models are constructed by thresholding the original features, such as age or criminal history, into binary stumps, followed by running ℓ_1-penalized logistic regression on the stumps. Choosing an appropriate regularization value for ℓ_1-penalized logistic regression can give us a model that is sparse in the number of original features—despite the fact that the regularization is directly on the stumps, not on the original features. For the Kentucky two-year `general` recidivism problem, the final model contains 28 stumps plus an intercept. These stumps are rooted under only 14 original features. Visualizations of the contributions for these 14 features are presented in Fig. 4. Table 10 in the Appendix, containing a scoring table that includes all 28 stumps plus an intercept.

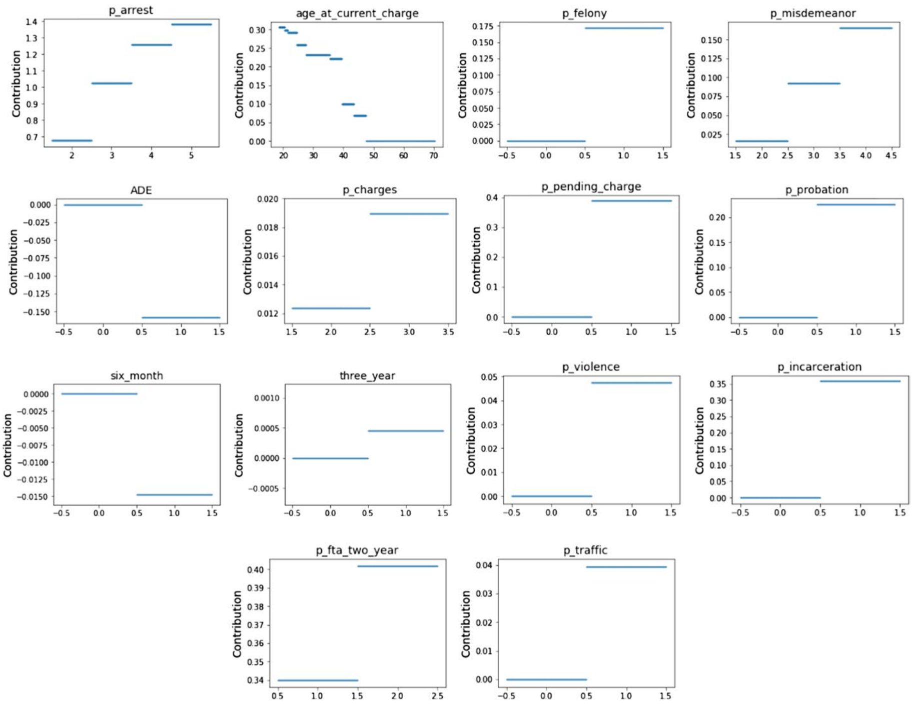


Fig. 4 Visualizations of the total contribution for each of the original features in the Additive Stumps model on two-year `general` recidivism. The contribution from each stump feature is the estimated coefficient from ℓ_1-penalized logistic regression

Table 2 Two-year `general` recidivism RiskSLIM models for Broward (left) and Kentucky (right)

Broward			Kentucky		
Pr(Y = +1) = 1 / (1 + exp(-(-2 + score)))			Pr(Y = +1) = 1 / (1 + exp(-(-2 + score)))		
Age at current charge ≤ 31	1 points	+.	Number of prior arrest≥ 2	1 points	+.
Number of prior misdemeanor charges ≥ 4	1 points	+.	Number of prior arrest≥ 3	1 points	+.
Had charge(s) within last three years = Yes	1 points	+.	Number of prior arrest≥ 5	1 points	+.
Add points from rows 1 to 3	Score	= .	Add points from rows 1 to 3	Score	= .

Each feature is given an integer point. The final predicted probability is calculated by inputting the total score to the logistic function provided on the top of the tables

RiskSLIM

RiskSLIM produces scoring tables with coefficients optimized to be integers (“points”), which makes the predictions easier to calculate and interpret for users, such as judges. The total points are translated into probabilities using the logistic function provided at the top of the table. By examining a RiskSLIM model, users can easily identify which features contribute to the final score and by how much. We provide scoring tables in Table 2 for two-year `general` recidivism prediction on both Broward and Kentucky data sets. More tables are provided in the Appendix.

We noticed that for each prediction problem, almost all five of the cross validation folds for the RiskSLIM algorithm yielded the same model on the larger Kentucky data set. In more detail, for Kentucky two-year `drug` and `violent` recidivism prediction problems, all five RiskSLIM models produced during cross validation were identical. For the rest of the prediction labels, four out of five cross validation models were the same. For the six-month recidivism prediction problems, the `misdemeanor` prediction problem resulted in five identical RiskSLIM models, and the `violent` recidivism prediction problem had four models that were the same. Kentucky RiskSLIM models are often the same, despite being trained on different—albeit overlapping—subsets of data, suggesting that they are robust to the exact subsample used for training.

Recidivism Prediction Models Do Not Generalize Well Across Regions

It is common practice for recidivism prediction systems to be applied across states, or even countries, with only minor tuning on local populations. Implicit in this practice is the assumption that models trained on data from one collection of locations will perform well when used in another collection of locations—i.e., that models *generalize* across locations. For instance, the Arnold PSA, which was developed on 1.5 million cases from approximately 300 U.S. jurisdictions, has been adopted in the states of Arizona, Kentucky, New Jersey, and many large cities including Chicago, Houston, Phoenix, etc. (Public Safety Assessment 2019). These systems have remained in place for years without any updates.

However, based on our experimental results, we conjecture that different locations would benefit from specialized models that conform to the specific aspects of each location. For instance, let us briefly compare the state of Kentucky and Broward County in Florida. The demographics are completely different: Kentucky is not a diverse state (87.8% white, 7.8% black, and 4.4% other groups in 2019 (United States Census Bureau 2019)), whereas Broward County is more racially diverse (62.3%, white; 17.1% Hispanic or Latino; 12.2% black or African American; 5.07% Asian and other groups (United States Census Bureau 2015). The geographies of the locations are drastically different as well: Kentucky is an interior state located in the Upland South with a humid subtropical climate, whereas Broward County is at the eastern edge of Florida with a tropical climate. Several studies have indicated an association between climate (temperature, humidity, and precipitation) and crime (Mishra 2014; Ranson 2014; Defronzo 1984). There are many other factors that differ between the locations that might affect the generalization of the recidivism prediction models, such as different local prosecution practices, laws and the way they are administered, social service programs, local cultures, educational systems, and judges' views.

Because models tend to be used broadly across locations, in this section we aim to investigate how well predictive models generalize between the two locations for which we have data. We trained models on Kentucky and tested on Broward, and vice versa. We looked more closely at *age*, and examined how the joint probability distribution of age and recidivism differs between Broward and Kentucky. We focused on age because of its important relationship to recidivism (Stevenson and Slobogin 2018; Bushway and Piehl 2007; Kleiman et al. 2007).

Major Findings: Our analysis shows that models do not generalize well across regions, and the joint probability distribution of age and recidivism varies across states. Therefore,

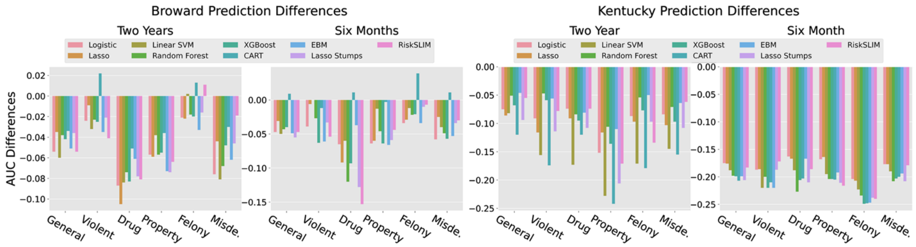


Fig. 5 Visualizations of prediction differences on Broward and Kentucky data. Broward prediction differences are the AUCs of models trained on Kentucky and tested on Broward minus AUCs of models trained and tested on Broward. Kentucky prediction differences are the AUCs of models trained on Broward and tested on Kentucky minus AUCs of models trained and tested on Kentucky. These results are also presented in Tables 18, 19, 20, 21 in the Appendix

we suggest that when possible, recidivism prediction models should be more location-specific, and be updated periodically.

Training on One Region and Testing on the Other

In order to construct models on one region and test them on the other, we only used the shared features from both data sets. Nested cross validation was used to train both the models that were trained in one region and tested in the other, and the models that were trained and tested in the same region. More details about this procedure can be found in the "Nested Cross Validation Procedure" section.

Figure 5 shows the difference between the performances of the models trained and tested on different regions and the performances of the models trained and tested on the same region. For Broward prediction results, we observed that there is an overall decrease in model performance when models were trained in Kentucky and tested on Broward. For instance, for the two-year `general` recidivism problem, the performances drop between 3.5% to 6.0% on the baseline models. A similar pattern can be observed for the interpretable models. Conversely, when we trained models on Broward and tested on Kentucky, we observed even larger performance decreases from the models trained and tested on only Kentucky. For the two-year `general` prediction task, performance gaps from baseline models range between 5.1% and 8.6%, while the gaps range from 4.6% to 12.0% for interpretable models.

Through this experimentation, we concluded that for at least the twelve prediction problems in our setup, models do not generalize across states. This could be attributable to differences in the joint probability distribution of features and outcomes between locations. To understand the difference in these distributions more closely, we examine the age feature.

Age-Recidivism Probability Distributions by Region

Age has traditionally been a highly predictive factor for recidivism (Stevenson and Slobogin 2018; Bushway and Piehl 2007; Kleiman et al. 2007). Therefore, differences in the age distributions between two regions could significantly impact a model's ability to generalize between regions.

Consider the `general` recidivism problem as an example. In Kentucky, the probability of general recidivism for both six-month and two-year prediction periods peaks for individuals aged around the early to mid 30s and then decreases as age increases. In Broward County, the age distribution for the corresponding `general` recidivism problem is substantially different. From Fig. 6, the probabilities seem to peak around ages 18–29, and then decrease after age 29. There are less data for higher ages, causing greater variance in the probabilities. For the `violent` recidivism problem, please refer to Fig. 8 in the Appendix. Additionally, there is a large gap in the probability magnitudes between the two regions. For instance, the probabilities of `general` recidivism from the Broward data set can exceed 0.5, while the probabilities of general recidivism from Kentucky data are all less than 0.4. Thus, the populations of individuals from Broward and Kentucky who recidivate are different with respect to age.

This difference is directly manifested in the interpretable models presented in the Appendix. We found that the selection of features differs between interpretable models trained on Broward and Kentucky data. For instance, referring to the simple Risk-SLIM models listed in the Appendix, which show the most important features in each prediction problem, we noticed that with Broward data, almost all prediction problems contain at least one age feature, either "age at current charge" or "age at first offense." This suggests that age is important in predicting recidivism across different problems trained on the Broward data. However, none of the RiskSLIM models trained on the Kentucky data set use age features. Almost all the models use "prior arrest" features,

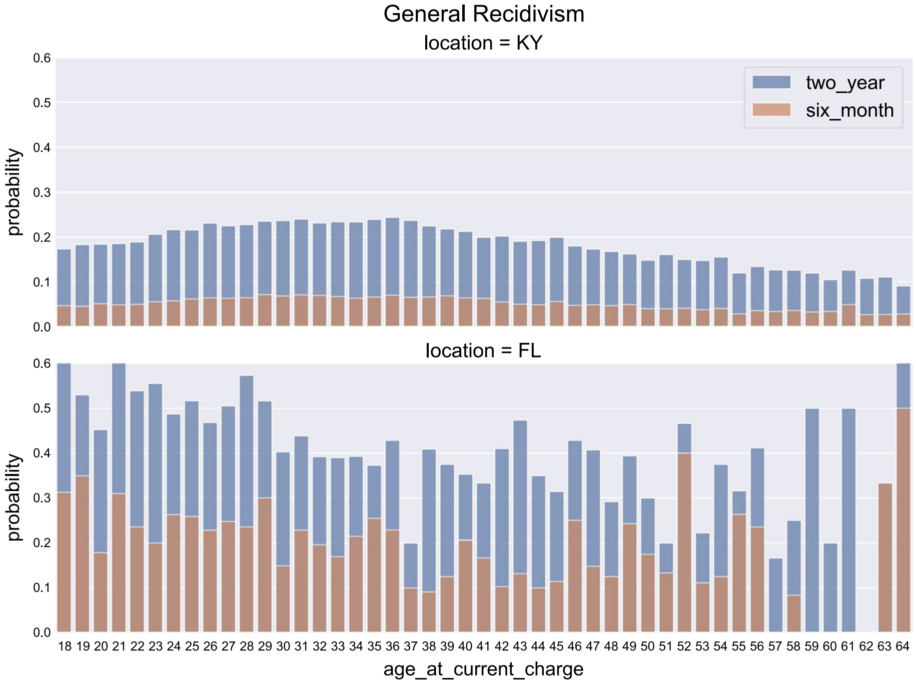


Fig. 6 Probability of recidivism vs. age at current charge—`general` recidivism

reflecting the fact that Kentucky recidivism prediction problems rely more on prior criminal history information than on age.

Fairness

In this section, we conduct a technical discussion of a small fraction of the various fairness definitions that have emerged recently, and an evaluation of how well the interpretable models satisfy them on the Kentucky data set. We first describe our rationale for selecting fairness definitions (calibration and balanced group AUC). Next, we evaluate how well the Arnold PSA, COMPAS, EBM (the best-performing interpretable models) and RiskSLIM (the most interpretable and most constrained models) satisfy these definitions on the two-year `general` recidivism and two-year `violent` recidivism problems in Kentucky. Finally, we discuss how current fairness enforcement procedures interact with interpretability.

Major Findings: Empirically, we found no violations of the fairness definitions (calibration and balanced group AUC) for both interpretable ML models we assessed (EBM and RiskSLIM) for the `two-year` general recidivism problem on the Kentucky data set. We found that the Arnold NCA raw score violated calibration at higher values of the score. Overall, we observed a larger gap in fairness for both fairness measures we examined between the largest and smallest sensitive groups, than between black and white sensitive groups. We also note that existing techniques to enforce fairness generally require non-interpretable transformations, and therefore do not work well with interpretable models.

Selection of Fairness Metrics: Calibration, Balance for Positive/Negative Class, Balanced Group AUC

As discussed in the "Problem Setup" section, we do not wish to consider binary risk scores in this study. This decision limits us to a much smaller class of fairness definitions, e.g., statistical parity would not be relevant. Below, we summarize the definitions that apply to regression that we do not consider and the reasons why:

- Fairness through unawareness states that a model should not use any sensitive features (Verma and Rubin 2018). However, if there are proxies for sensitive features present in the data set, the model can still learn an association between a sensitive group and the outcome. Fairness through unawareness could be used if one decides that a proxy feature is permissible to use—for instance, if one decided that age could be used, despite its correlation with race—but we do not presume that this is what is desired for this application. Of course, if fairness through unawareness *is* desired, it is easy to construct models that satisfy this definition.
- Individual fairness intuitively requires that "similar" individuals are treated "similarly" by the model—individuals with similar features should be given similar model scores. This type of fairness requires manually and thus subjectively defining a notion of similarity between individuals (Dwork et al. 2012). This type of subjective choice goes beyond the scope of this paper.
- Balance for Positive/Negative Class (BPC/BNC) states that it is permissible to give consistently higher (respectively lower) scores to individuals who truly belong to the

positive (respectively negative) class. However, BPC/BNC limits the set of attributes where it is permissible to "discriminate" between individuals, to the label Y. Suppose the count of prior offenses is an important feature for a recidivism prediction model—higher prior counts lead to higher scores. This is a reasonable model assumption because a higher prior count is correlated with higher recidivism rates. If on average, African-Americans have higher prior counts than Caucasians, the model will not satisfy BPC/BNC. For a model to satisfy BPC/BNC, it must give the same average score to individuals from a certain race and with a certain recidivism label, regardless of distributional differences in prior counts. Those who believe that prior counts and arrests are not racially biased against African-Americans might find this a desirable property of a fairness definition. On the other hand, those who find this undesirable can fix this by conditioning on the prior counts attribute as well. Thus, this fairness definition requires deciding which features are group-biased, a subjective conditioning that also goes beyond the scope of this work.

Once we limited ourselves to real-valued outcomes and eliminated the above definitions, only a few definitions remained. In a literature search for nonbinary fairness definitions, we found the fairness definitions of calibration and balanced group AUC (BG-AUC).

Below, G denotes a (categorical) sensitive attribute such as race, and g_i denotes one of the *sensitive groups* in G (e.g. African-American, Caucasian, and Hispanic, for the sensitive attribute of race). $Y \in \{0, 1\}$ denotes the ground-truth label (recidivism status) and S denotes the predicted score from a model.

- Calibration: We consider two notions of calibration. The first, group calibration, requires that for all predicted scores, the fraction of positive labels is approximately the same across all groups. Mathematically, group calibration over the sensitive attribute G requires:

  $$P(Y = 1|S = s, G = g_i) \approx P(Y = 1|S = s, G = g_j), \forall i, j$$

  where s is the given value of a risk score. Note that in the case where scores are binary, group calibration is equivalent to requiring *conditional use accuracy equality*. In practice, it is common to bin the score S if there are many possible values. The second, monotonic calibration, requires that if $s_1 < s_2$, then $P(Y = 1|S = s_1) < P(Y = 1|S = s_2)$.[3]

  These types of calibration are of particular concern to designers of current recidivism risk models. Group calibration means that a risk score holds the same "meaning" for each race. Monotonic calibration means that if the score increases, the risk also increases. These notions are important because human decision-makers expect risk scores to have these intuitive properties (but not all algorithms produce calibrated models) (Chouldechova 2017).
- Balanced Group AUC (BG-AUC) requires that the AUC of the risk score is approximately the same for each sensitive group. This definition is our adaptation of overall

[3] We note that a real-valued score S between 0 and 1 is *well-calibrated* if $P(Y = 1|S = s) = s$. Well-calibration says that the predicted probability of recidivism should be the same as the true probability of recidivism (Verma and Rubin 2018). Although well-calibration is the definition of calibration that is standard in the statistics community, we consider monotonic-calibration here because any score that is monotonically-calibrated can be transformed to be well-calibrated.

accuracy equality (Berk et al. 2017b), which asks that the score's accuracy is the same for each sensitive group. Our risk scores are not binary so we do not assess accuracy in this work, but assessing the AUC for each group is the natural analog.

Sensitive attributes: The two sensitive attributes that are available in the Kentucky data sets are race and gender. In the Kentucky data set, all individuals are partitioned into Caucasian, African-American, Indian, Asian, and Other, but we group the Indian and Asian attributes into other because there are very few individuals with these attributes. See Table 11 for the distribution of sensitive attributes in Kentucky. The Kentucky data set also partitions individuals into the genders female and male. To summarize: races in Kentucky = {Caucasian, African-American, other}; sexes in Kentucky = {female, male}.

Fairness Results

We assessed model fairness only on the Kentucky data because the Broward data has a limited sample size, potentially making the fairness results unreliable. We attempted the evaluation on Broward data, but conditioning on race/gender and the true label/score in the Broward data led to subgroups that were too small, and therefore noisy results. We compared the interpretable models, EBM and RiskSLIM, to the Arnold PSA on Kentucky. EBM has the best performance on most of the prediction problems on the Kentucky data set. RiskSLIM performs relatively worse, but is considerably simpler as there are no more than five features in each model, coefficients are integers, and the model is linear.

We evaluated the two-year general and two-year violent problems, as they are the primary problems that the Arnold PSA is used for. For the two-year general problem, we evaluated the unscaled Arnold New Criminal Activity (NCA) score; for the two-year violent problem, we assessed the unscaled Arnold New Violent Criminal Activity (NVCA) score. Although Arnold Ventures provides a table to scale the Arnold scores, in Kentucky, judges are presented with the unscaled scores along with a categorization of the scores as low, medium, and high risk. Results for both two-year general and violent recidivism can be found in the Appendix.

Calibration

Figure 7 shows that the Arnold NCA raw score approximately satisfies monotonic and group calibration for race and gender on lower values of the risk score (i.e., scores less than 10) and except for the "Other" group. There are fewer individuals with high Arnold

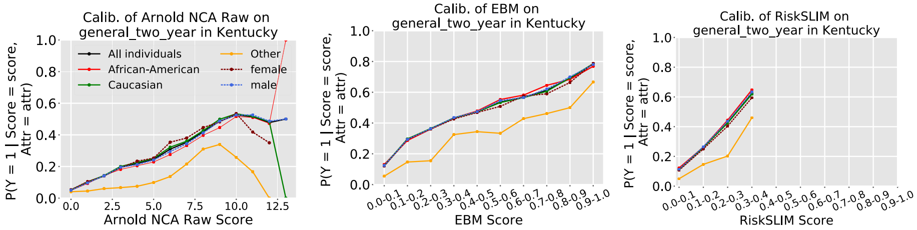


Fig. 7 Calibration results for the Arnold NCA raw, EBM and RiskSLIM for two-year general recidivism on Kentucky

NCA scores (e.g., 12 or 13) in the dataset, leading to higher variance in predictions, which may be why higher Arnold NCA raw scores fail the calibration definitions. Interestingly, we found that the scaled version of Arnold NCA fully satisfied monotonic and group calibration, but had slightly worse predictive performance. EBM and RiskSLIM both satisfy monotonic calibration and group calibration for all gender and race groups (excluding the "Other" group).

Balanced Group AUC (BG-AUC)

In Kentucky, AUC values are stable across sensitive attributes for all models. The discrepancies in AUC between African-Americans and Caucasians range from 0.3% (RiskSLIM) to 2.1% (Arnold NCA raw). The range gets smaller for gender groups, lying between 0.5% (Arnold NCA) to 1.3% (RiskSLIM). Hence, we found that RiskSLIM has the least AUC difference between race groups (excluding the "Other" group), whereas Arnold NCA has the least AUC difference between gender groups. We note that the differences in AUCs between models are small, with the largest differences manifesting with the "Other" group (Table 3).

Summary of Fairness Results: For the two-year general recidivism problem on the Kentucky data set, we found no egregious violations of calibration and BG-AUC for the models we assessed (Arnold NCA, EBM and RiskSLIM). However, we did find small violations. For example, we found the Arnold NCA raw score violated calibration for higher scores.

A caveat is that we limited the discussion of the race groups to Caucasians and African-Americans, else the "Other" group would have caused all models to fail all definitions of fairness: calibration curves for the "Other" group are significantly beneath curves for other groups and prediction AUC is significantly lower for the "Other" group. This may be because we have the least data for the "Other" race group, which is only 2.49% of the total sample. To ensure fairness, it is important that comparable amounts of data are gathered for each sensitive group when possible. However, in non-diverse states such as Kentucky, there may not be enough individuals in minority groups to create a large enough statistical sample.

A Discussion on the Interaction Between Fairness and Interpretability

There are significant hurdles to using current fairness techniques with interpretable models. Moreover, the vast majority of the work on fairness has focused on the binary classification case. Thus, few definitions of fairness (let alone algorithms) work for problems where predictions are nonbinary.

We did not attempt to use fairness enforcement techniques because many fairness techniques require a non-interpretable transformation. Once these transformations are made, there is no way to correct them to produce an interpretable model afterwards. There are generally three approaches to fairness algorithms: preprocessing of features (Zemel et al. 2013), altering the training loss function (Berk et al. 2017a; Agarwal et al. 2019), and post-processing of predictions (Hardt et al. 2016; Agarwal et al. 2018; Pleiss et al. 2017). The pre-processing steps are generally complicated transformations of the input features, which shreds the data's natural meaning. Similarly, post-processing approaches either transform the predictions in some way, performing "fairness corrections" (Pleiss et al. 2017) (which are non-interpretable), or require threshold selection, which is contrary to our goals of

Table 3 AUCs of the Arnold NCA Raw, EBM and RiskSLIM on Kentucky, conditioned on sensitive attributes

Kentucky								
		Race				Sex		
Model	Label	Afr-Am.	Cauc.	Other Race	race_range	Female	Male	sex_range
Arnold NCA Raw	General_two_year	0.692	0.713	0.653	0.059	0.714	0.709	0.005
EBM	General_two_year	0.742	0.751	0.696	0.055	0.745	0.753	0.008
RiskSLIM	General_two_year	0.705	0.708	0.620	0.088	0.699	0.712	0.013

AUC ranges are given for each sensitive attribute providing nonbinary risk assessments (Hardt et al. 2016). The approaches to modify training loss functions are the most promising, but model optimization for both fairness and interpretability constraints would require new algorithms and is beyond the scope of this work.

In problems where fairness is a significant concern, machine learning outputs are likely to be used as decision tools rather than decision-makers, so it is surprising that so little work has thoroughly examined fairness for regression or probability estimation.

Discussion and Future Work

From this analysis, we conclude that the interpretable models can indeed perform approximately as well as the black-box models in various recidivism prediction problems. On the Broward data set, we found that RiskSLIM, EBM, and Additive Stumps perform as well or better than the best black-box models. On the Kentucky data set, we observed that EBM and Additive Stumps have extremely close performance to the best black-box models—Random Forest and XGBoost—with average AUC differences around 1%, which is less than the uncertainty gap. For the purposes of criminal recidivism prediction, our work indicates that there is no practical loss in accuracy by using interpretable models and much to be gained in interpretability.

We observed that ML models for six-month outcomes generally outperform those for two-year outcomes, conditioning on the recidivism type. This may be because treatment/rehabilitation programs have a greater chance of taking effect over a two-year time span, as compared to the six-month time span, altering the probability of recidivism. Future work could investigate this hypothesis, or pose other hypotheses to explain this observation.

We observed significant differences in the age distributions for Kentucky and Broward County, and hypothesized that this difference may be why ML models do not generalize well between regions. One might easily imagine regional feature distributions shifting over *time* as well, which is supported by several studies (Kim et al. 2016; Cook and Laub 2002; Alfred 2006; Matthews and Minton 2017). Even though these studies focused on disparate crime types, they consistently observed a drop in the rate of offending among younger people since the 1990s. Studies have explicitly shown that the distributions of age versus arrest rate has changed over time as well. For instance, Kim et al. (2016) has reported that in the state of New York, the mean age of the total arrested population increased by 2 years between 1990 and 2010. They hypothesized that a decrease in arrests in younger people and an increase in arrests in older people together contributed to the increase in mean age. There are many reasons why data would change over time and over jurisdictions. Changing policies (e.g., the NYC stop and frisk program) could potentially alter who would be arrested and for what types of crime. New cultural phenomena, for instance originating from media, could also influence people's behavior at a large scale.

The above observations lead us to conclude that different recidivism prediction models could be constructed for different locations and should be periodically updated. Machine learning models are well-suited for efficient creation and updating of these kinds of models. A possible future line of work is to separate the Kentucky data at the jurisdiction level, and perform a causal analysis of the effects of different judicial and policing practices on the recidivism distribution. While local jurisdictions, e.g., at the county or district level, might not have sufficient resources to fit their own recidivism prediction models, our analysis on the Kentucky dataset shows that even considering *state-level* recidivism models produces gains over models trained across nationwide data, such as the Arnold PSA. It may be more feasible for the state-level agencies to collect enough data and hire analysts to fit the models. Future work could also investigate using small quantities of location-specific data to fine-tune more general models.

Simple, linear models have been used for criminal justice applications for almost a century (Burgess 1928; Hart 1924). They have the advantage that one can easily quantify the contributions of each feature to the predicted score. Judicial actors without much statistics background can understand these scores, and use them to help solve societal issues. Interpretable models are extremely valuable for current decision-making processes in criminal justice: they allow error-checking, help ensure due process, and allow judges to incorporate information outside the database into their decision-making process in a calibrated manner.

However, our work on interpretable risk prediction is only one step closer to what we view as the ultimate goal—placing recidivism prediction into the framework of formal decision analysis. Decision-making in the context of decision analysis involves the minimization of costs rather than risks. Towards this end, Lakkaraju and Rudin (2017) considered several costs related to pretrial release decisions; these include the societal cost of releasing an individual who might commit a crime before their trial, the cost of assigning an officer to an individual, and the cost to taxpayers of keeping an individual incarcerated. The importance of risk predictions vary between decision-making problems (release, parole, sentencing, etc.). In some cases, they play a minor role, yet in others, predictions may comprise the sole deciding factor. Because of this, it would be useful to have a cost-benefit analysis *per decision* that would help determine exactly when and where risk scores should participate.

Hence, an important and necessary direction for the future work would be to incorporate the framework of classical decision analysis into decision-making in the criminal justice system. Decision analysis tools would ideally allow practitioners to strike a balance between relevant considerations: for instance, future risk to society, costs of treatment programs, costs to families involved in the criminal justice system, costs to the individual, as well as more traditional modelling objectives such as fairness, interpretability, transparency, and predictive performance. While the full data measuring costs and risks to all stakeholders in the criminal justice process may never be available, it is important to move in this direction, as this would bring us closer to more consistent and informed decision making.

Appendix

Nested Cross Validation Procedure

We applied fivefold nested cross validation to tune parameters. We split the entire data set into five equally-sized folds for the outer cross validation step. One fold was used as the holdout test set and the other four folds were used as the training set (call it “outer training set”). The inner loop deals only with the outer training set ($\frac{4}{5}$ths of the data). On this outer training set, we conducted fivefold cross validation and grid-searched hyperparameter values. After this point, each hyperparameter value had five validation results. We selected the parameter values with the highest average validation results and then trained the model with this best set of parameters on the entire outer training set and tested it on the holdout test set.

We repeated the process above until each one of the original five folds was used as the holdout test set. Ultimately, we had five holdout test results, with which we were able to calculate the average and standard deviation of the performance.

We applied a variant of the nested cross validation procedure described above to perform the analysis discussed in the "Recidivism Prediction Models Do Not Generalize Well Across Regions" section—where we trained models on one region and tested on the other region. For instance, when we trained models on Broward and tested them on Kentucky, the Kentucky data was treated as the holdout test set. We split the Broward data into five folds and used four folds to do cross validation and constructed the final model using the best parameters. We then tested the final model on the entire Kentucky data set, as well as the holdout test set from Broward. We rotated the four folds and repeated the above process five times.

Broward Data Processing

The Broward County data set consists of publicly available criminal history, court data and COMPAS scores from Broward County, Florida. The criminal history and demographic information were computed from raw data released by ProPublica (Angwin et al. 2016). The probational history was computed from public criminal records released by the Broward Clerk's Office.

The screening date is the date on which the COMPAS score was calculated. The features and labels were computed for an individual with respect to a particular screening date. For individuals who have multiple screening dates, we compute the features for each screening date, such that the set of events for calculating features for earlier screening dates is included in the set of events for later screening dates. On occasion, an individual will have multiple COMPAS scores calculated on the same date. There appears to be no information distinguishing these scores other than their identification number, so we take the scores with the larger identification number. The recidivism labels were computed for the timescales of 6 months and 2 years. Some individuals were sentenced to prison as a result of their offense(s). We used only observations for which we have 6 months/2 years of data subsequent to the individual's release date.

Below, we describe details of the feature and label generation process. The constructed features are presented in Table 4 at the end of this section.

- Degree "(0)" charges seem to be very minor offenses, so we exclude these charges. We infer whether a charge is a felony, misdemeanor, or traffic charge based off the charge degree.
- Some of our features rely on classifying the type of each offense (e.g., whether or not it is a violent offense). We infer this from the statute number, most of which correspond to statute numbers from the Florida state crime code.
- The raw Propublica data includes arrest data as well as charge data. Because the arrest data does not include the statute, which is necessary for us to determine offense type, we use the charge data to compute features that require the offense type. We use both charge and arrest data to predict recidivism.

Table 4 Features from Broward data set

Features	Explanation
person_id	Unique personal identifier
sex	Biological sex of the person
race	Race of the person
screening_date	Date that triggered the COMPAS screening
age_at_current_charge	Age at the person's current charge
age_at_first_charge	Age at the person's first charge
p_arrest	Count of prior arrests
p_charges	Count of prior charges
p_violence	Count of prior violent charges
p_felony	Count of prior felony-level charges
p_misdemeanor	Count of prior misdemeanor-level charges
p_juv_fel_count	Count of prior felony-level and juvenile charges
p_property	Count of prior property-related charges
p_murder	Count of prior murder charges
p_famviol	Count of prior family violence charges
p_sex_offenses	Count of prior sex offense charges
p_weapon	Count of prior weapon-related charges
p_felprop_viol	Count of prior felony-level, property-related, and violent charges
p_felassault	Count of prior felony-level assault charges
p_misdeassault	Count of prior misdemeanor-level assault charges
p_traffic	Count of prior traffic-related charges
p_drug	Count of prior drug-related charges
p_dui	Count of prior DUI charges
p_stalking	Count of prior stalking charges
p_voyeurism	Count of prior voyeurism charges
p_fraud	Count of prior fraud charges
p_stealing	Count of prior stealing/theft charges
p_domestic	Count of prior domestic violence charges
p_trespass	Count of prior trespass charges
p_fta_two_year	Count of prior failures to appear in court within last 2 years ($\leq$ 2 years)
p_fta_two_year_plus	Count of prior failures to appear in court beyond last 2 years (> 2 years)
p_pending_charge	Count of times charged with a new offense when there was a pending case
p_probation	Count of times charged with a new offense when the person was on probation
p_incarceration	Whether or not the person was formerly sentenced to incarceration
six_month	Whether or not the person had charges within last 6 months ($\leq$ 6 months)
one_year	Whether or not the person had charges within last year ($\leq$ 1 year)
three_year	Whether or not the person had charges within last three years ($\leq$ 3 years)
five_year	Whether or not the person had charges within last five years ($\leq$ 5 years)
current_violence	Whether or not the current charge is violent
current_violence20	Whether or not the current charge is violent and the person is $\leq$ 20 years old
total_convictions	Total count of convictions

Recall that charges can be convicted or non-convicted

- For each person on each COMPAS screening date, we identify the offense—which we call the current offense—that most likely triggered the COMPAS screening. The current offense date is the date of the most recent charge that occurred on or before the COMPAS screening date. Any charge that occurred on the current offense date is part of the current offense. In some cases, there is no prior charge that occurred near the COMPAS screening date, suggesting charges may be missing from the data set. For this reason we consider charges that occurred within 30 days of the screening date for computing the current offense. If there are no charges in this range, we say the current offense is missing. We exclude observations with missing current offenses. We used some of the COMPAS subscale items as features for our ML models. All such components of the COMPAS subscales that we compute are based on data that occurred prior to (not including) the current offense date.
- The events/documents data includes a number of events (e.g., “File Affidavit Of Defense” or “File Order Dismissing Appeal”) related to each case, and thus to each person. To determine how many prior offenses occurred while on probation, or if the current offense occurred while on probation, we define a list of event descriptions indicating that an individual was taken on or off probation. Unfortunately, there appear to be missing events, as individuals often have consecutive “On” or consecutive “Off” events (e.g., two “On” events in a row, without an “Off” in between). In these cases, or if the first event is an “Off” event or the last event is an “On” event, we define two thresholds, t_{on} and t_{off}. If an offense occurred within t_{on} days after an “On” event or t_{off} days before an “Off” event, we count the offense as occurring while on probation. We set t_{on} to 365 and t_{off} to 30. On the other hand, the “number of times on probation” feature is just the count of “On” events and the “number of times the probation was revoked” feature is just the count of “File order of Revocation of Probation” event descriptions (i.e., we do not infer missing probation events for these two features).
- Current age is defined as the age in years, rounded down to the nearest integer, on the COMPAS screening date.
- A juvenile charge is defined as an offense that occurred prior to the defendant’s 18th birthday.
- Labels and features were computed using charge data.
- The final data set contains 1954 records and 41 features.

Kentucky Data Processing

The Kentucky pretrial and criminal court data was provided by the Department of Shared Services, Research and Statistics in Kentucky. The Pretrial Services Information Management System (PRIM) data contains records regarding defendants, interviews, PRIM cases, bonds etc., that are connected with the pretrial services’ interviews conducted between July 1, 2009 and June 30, 2018. The cases were restricted to have misdemeanor, felony, and other level charges. The data from another system, CourtNet, provided further information about cases, charges, sentences, dispositions etc. for CourtNet cases matched in the PRIM system. The Kentucky data can be accessed through a special data request to the Kentucky Department of Shared Services, Research and Statistics. Please refer to Table 5 for all the raw datasets we processed, together with their sizes and general information provided.

Table 5 The table lists raw datasets obtained from the Kentucky Department of Shared Services, Research and Statistics, the number of records within each data frame, and general descriptions of the data

Data	Num. of records	Information
KY_Recidivism_Defendants	1,286,599	Contains unique identifiers for each defendant
KY_Recidivism_Interviews	1,490,545	Contains arrest information, risk level assessment, etc.
KY_Recidivism_CNet_Cases	3,179,421	Contains distinct cases in CourtNet database, including case filing, disposition, etc.
KY_Recidivism_PRIM_Cases	1,987,783	Contains distinct cases in PRIM database
KY_Recidivism_CNet_Charges	8,683,273	Contains distinct charges within cases in CourtNet database
KY_Recidivism_Sentences	2,926,446	Contains sentence information for charges disposed with a conviction
KY_Recidivism_Bonds	2,947,148	Contains pretrial records of bonds and release conditions set by court.
KY_Recidivism_Events	2,256,330	Contains information about court-mandated events, including the type of event and whether or not a defendant was compliant, etc.
KY_Recidivism_FTA	231,737	Contains specific information about court appearances for which pretrial services indicated that the defendant failed to appear (FTA).
KY_Recidivism_Supervision	42,352	Contains information specific to defendants on monitored pretrial release.

Table 6 Features from Kentucky data set

Features	Explanation
person_id	Unique personal identifier
sex	Biological sex of the person
race	Race of the person
current_date	Current charge date or the release date if there was a sentence on the current charge.
age_at_current_charge	Age at the person's current charge, or the age at current charge plus the sentence time if there was a sentence on the current charge
p_arrest	Count of prior arrests with convicted charges
p_charges	Count of prior convicted charges
p_violence	Count of prior violent charges
p_felony	Count of prior felony-level charges
p_misdemeanor	Count of prior misdemeanor-level charges
p_property	Count of prior property-related charges
p_murder	Count of prior murder charges
p_assault	Count of prior assault charges
p_sex_offenses	Count of prior sex offense charges
p_weapon	Count of prior weapon-related charges
p_felprop_viol	Count of prior felony-level, property-related, and violent charges
p_felassault	Count of prior felony-level assault charges
p_misdeassault	Count of prior misdemeanor-level assault charges
p_traffic	Count of prior traffic-related charges
p_drug	Count of prior drug-related charges
p_dui	Count of prior DUI charges
p_stalking	Count of prior stalking charges
p_voyeurism	Count of prior voyeurism charges
p_fraud	Count of prior fraud charges
p_stealing	Count of prior stealing/theft charges
p_trespass	Count of prior trespass charges
ADE	Count of times the person was assigned to alcohol/drug education classes
treatment	Count of times the person received treatment along with the sentence
p_fta_two_year	Count of prior failures to appear in court within last 2 years ($\leq$ 2 years)
p_fta_two_year_plus	Count of prior failures to appear in court beyond last 2 years ($>$ 2 years)
p_pending_charge	Count of times charged with a new offense when there was a pending case
p_probation	Count of times charged with a new offense when the person was on probation
p_incarceration	Whether or not the person was formerly sentenced to incarceration
six_month	Whether or not the person had charges within last 6 months ($\leq$ 6 months)
one_year	Whether or not the person had charges within last year ($\leq$ 1 year)
three_year	Whether or not the person had charges within last three years ($\leq$ 3 years)
five_year	Whether or not the person had charges within last five years ($\leq$ 5 years)
current_violence	Whether or not the current charge was violent
current_violence20	whether or not the current charge was violent and the person was $\leq$ 20 years old
current_pending_charge	Whether or not the person had a pending case during the current charge

The charges are convicted. ADE means assignment to alcohol and drug education classes

CourtNet and PRIM data were processed separately and then combined together. We describe the details below. The constructed features are presented in Table 6 at the end of this section.

- For the CourtNet data, we filtered out cases with filing date prior to Jan. 1st, 1996, which were claimed to be less reliable records by the Kentucky Department of Shared Services, Research and Statistics (which provided the data). To investigate what types of crimes the individuals were involved in for each charge, such as drug, property, traffic-related crime, we used the Kentucky Uniform Crime Reporting Code (UOR Code), as well as detecting keywords in the UOR description.
- From the PRIM system data, we extracted the probation, failure to appear, case pending, and violent charge information at the PRIM case level, as well as the Arnold PSA risk scores computed at the time of each pretrial services' interview. Since Kentucky did not use Arnold PSA until July 1st, 2013, we filtered out records before the this date. We omitted records without risk scores since we want to compare the performance of the PSA with other models. Only 33 records are missing PSA scores; therefore we do not worry about missing records impacting the results. Additionally, some cases in the PRIM system have "indictment" for the arrest type, along with an "original" arrest case ID, indicating that those cases were not new arrests. We matched these cases with the records that correspond to the original arrests to avoid overcounting the number of prior arrests. Then we inner-joined the data from the two systems using person-id and prim-case-id.
- For each individual, we used the date that is 2 years before the latest charge date in the Kentucky data, as a cutoff date. The data before the cutoff are used as criminal history information to compute features. The data after the cutoff are used to compute labels and check recidivism. In the data before the cutoff, the latest charge is treated as the current charge (i.e., the charge that would trigger a risk-assessment) for each individual. We compute features and construct labels using only convicted charges. However, the current charge can be either convicted or non-convicted. This ensures that our analysis includes all individuals that would receive a risk assessment, even if they were later found innocent of the current charge that triggered the risk assessment. It also ensures that criminal history features use only convicted charges, so that our risk assessments are not influenced by charges for crimes that the person may not have committed.
- In order to compute the labels, we must ensure that there are at least 2 years of data following an individual's current charge date. For individuals who are sentenced to prison due to their current charge, we consider their *release date* instead of the current charge date. We omitted individuals for whom there were less than 2 years of data between their current charge date or release date, and the last date recorded in the data set.
- To get the age at current charge information, we first calculated the date of birth (DOB) for each individual, using CourtNet case filing date and age *at the CourtNet case filing date*. Then we calculated "age at current charge" using the DOB and charge date (the charge date sometimes differs from the case filing date). Notice that there are many errors in age records in the data. For instance, some people have age recorded over 150, which is certainly wrong but there is no way to correct it. To ensure the quality of our data, we limited the final current age feature to be inclusively between 18 and 70. This is also consistent with the range from Broward analysis. If the person was not sentenced to prison, we define current age as the age at current charge date. If the person

was sentenced to prison, we compute current age by adding the sentence time to the age at the current charge date. Note that this differs from the way risk scores are computed in practice—usually risk scores are computed prior to the sentencing decision. This helps to handle distributional shift between the individuals with no prison sentence (for whom a 2-year evaluation can be handled directly) and the full population (some of whom may have been sentenced to prison and cannot commit a crime during their sentence).

- We computed features using the data before the current charge date. The CourtNet data is organized by CourtNet cases, and each CourtNet case has charge level data. The PRIM data is organized by PRIM cases. Each CourtNet case can connect to multiple PRIM cases. This occurs because a new PRIM case is logged when an update occurs in the defendant's CourtNet case — for example, if the defendant fails to appear in court. Therefore, to compute the criminal history information, we first grouped on PRIM case level to summarize the charge information. Next, we grouped on CourtNet case level to summarize PRIM case level information. Last, we grouped on the individual level to summarize the criminal histories.
- On computing the ADE feature: The ADE feature means number of times the individual was assigned to alcohol and drug education classes. Note that by Kentucky state law, any individual convicted for a DUI is assigned to ADE classes. This does not indicate whether the individual successfully completed ADE classes.
- We compute labels using the 2 years of data after the current charge date/release date. We constructed the `general` recidivism labels by checking whether a "convicted charge" occurred within 2 years or 6 months from the current charge (or release date). Then, using the charge types of the convicted charge, other recidivism prediction labels were generated, such as drug or property-related recidivism. The final data set contains 250,778 records and 40 features.

  Note: there are degrees of experimenter freedom in some of these data processing choices; exploring all the possible choices here is left for future studies.

The Arnold PSA features that were included in the Kentucky data set (e.g., prior convictions, prior felony convictions etc.) were computed by pretrial officers who had access to criminal history data from both inside and outside of Kentucky. However, the Kentucky data set we received contained criminal history information from within Kentucky only. Thus, the Arnold PSA features for Kentucky (which are included in our models as well) use both in-state and out-of-state information, but the remaining features (which we compute directly from the Kentucky criminal history data) are limited to in-state criminal history.

Additionally, we were informed by Kentucky Pretrial Services team that the data set 's sentencing information may not be reliable due to unmeasured confounding, including shock probation and early releases that would allow a prisoner to be released much earlier than the end date of the sentence. Because the sentence could be anywhere from zero days to the full length, we conducted a sensitivity analysis by excluding the sentence information in the data processing, which is equivalent to the assumption that no prison sentence was served. For that analysis, the current age of each individual was calculated to be the age at the current charge, and the prediction labels were generated from new charges within 6 months (or 2 years) from the current charge. The sensitivity analysis yielded predictive results that were almost exactly the same as the results in the main text, when the sentence information was used to determine age and prediction interval.

Table 7 Variable comparison for currently-utilized actuarial risk assessments

Models	Criminal history	Age	Finance	Residential info	Edu/emp	Peer/family	Mental health	Alc/subs abuse	Other
COMPAS Northpointe Inc. (2009)	X	X			X	X	X	X	X
Connecticut Carollo et al. (2007)	X				X	X	X	X	
CPAT of Pretrial Services (2015)	X	X		X			X	X	X
CSRA Turner et al. (2009)	X	X							X
ORAS Latessa et al. (2009)	X	X		X	X			X	
LSI-CMI MHS Assessments (2017)	X		X		X	X		X	X
PSA Public Safety Assessment (2019)	X	X							
PTRA Cadigan and Lowenkamp (2011)	X	X	X	X	X	X		X	X
Salient Factor Hoffman and Adelberg (1980)	X	X						X	
SIRS Nafekh and Motiuk (2002)	X	X			X				X
SPIn Orbis (2014)	X	X			X	X	X	X	X
VERA Lazarsfeld (1974)	X	X			X	X	X	X	X
VRAG Harris and Rice (2008)	X	X				X	X	X	X
VPRAI Virginia Department of Criminal Justice Services (2018)	X			X	X			X	X

We only have criminal history and age variables, but most models include many other variables. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS); Connecticut Risk Assessment for Pretrial Decision Making (Connecticut); Colorado Pretrial Risk Assessment Tool (CPAT); California Static Risk Assessment (CSRA); Ohio Risk Assessment System (ORAS); Level-of-Service Case Management Inventory (LSI-CMR); Public Service Assessment(PSA); (Federal) Pretrial Risk Assessment (PTRA); Statistical Information on Recidivism Score (SIRS); Service Planning Instruments (SPIn); Vera Point Scale (VERA); Violence Risk Appraisal Guide (VRAG); Virginia Pretrial Risk Assessment Instrument (VPRAI)

Why We Compare Only Against COMPAS and the PSA

The variables included in risk assessments are often categorized into *static* and *dynamic* factors. Static factors are defined as factors that cannot be reduced over time (e.g. criminal history, gender, and age-at-first-arrest). Dynamic factors are defined as variables that can change over time to decrease the risk of recidivism; they allow insight into whether a high-risk individual can lower their risk through rehabilitation, and sometimes improve prediction accuracy. Examples of dynamic factors include current age, treatment for substance abuse, and mental health status (Kehl et al. 2017). Dynamic factors are often included in *risk-and-needs-assessments* (RNAs), which in addition to identifying risk of recidivism, recommend interventions to practitioners (e.g., treatment programs, social services, diversion of individuals from jail).

With the exception of current age, our features all fall under the "static" classification. This renders us unable to compare against the risk assessment tools that use dynamic factors, whose formulas *are* public. The risk assessments that we examined are listed in Table 7. Since we have only criminal history and age variables, the only model we could compute from our data was the Arnold PSA.

However, as we demonstrated in the main body of the paper, the fact that we do not possess dynamic factors is not necessarily harmful to the predictive performance of our models. The goal behind including dynamic factors in models is to improve prediction accuracy as well as be able to recommend interventions that reduce the probability of recidivism. While an admirable goal, the inclusion of dynamic factors does not come at zero cost and may not actually produce performance gains for recidivism prediction. In the Baseline Machine Learning Methods and "Recidivism Prediction Models Do Not Generalize Well Across Regions" sections, we show that standard machine learning techniques (using only the static factors) and interpretable ML models (using only static factors) are able to outperform a criminal justice model that utilizes both static and dynamic factors (COMPAS). Furthermore, the inclusion of additional, unnecessary factors increases the risk of data entry errors, or exposes models to additional feature bias (Corbett-Davies and Goel 2018). As Rudin et al. (2020) reveals, data entry errors appear to be common in COMPAS score calculations and could lead to scores that are either too high or too low.

Although the COMPAS suite is a proprietary (and thus black-box) risk-and-needs assessment, we were still able to compare against its risk assessments thanks to the Florida's strong open-records laws. Created by Northpointe (a subsidiary company of Equivant), COMPAS is a recidivism prediction suite which is used in criminal justice systems throughout the United States. It is comprised of three scores: Risk of General Recidivism, Risk of Violent Recidivism, and Risk of Failure to Appear. In this work, we examine the two risk scores relating to violent recidivism and general recidivism. Each risk score is an integer from one to ten (Brennan et al. 2009).

As COMPAS scores are proprietary instruments, the precise forms of its models are not publicly available. However, it is known that the COMPAS scores are computed from a subset of 137 input variables that include vocational/educational status, substance abuse, and probational history, in addition to the standard criminal history variables (Brennan et al. 2009). As such, we cannot directly compute these risk scores, and instead utilize the COMPAS scores released by ProPublica in the Broward County recidivism data set. We do not compare against COMPAS on the Kentucky data set, as our data set does not include COMPAS scores.

Table 8 Hyperparameters for ℓ_1 and ℓ_2 penalized logistic regression, linear SVM, CART, random forest, XGBoost, and EBM. RiskSLIM and additive stumps are discussed separately

Models	Kentucky	Broward
ℓ_2 Logistic regression	class_weight: balanced solver: liblinear Fan et al. (2008) penalty: ℓ_2 C ∈ [1e−4, 1e−3, 1e−2, 1e−1, 1]	class_weight: balanced solver: liblinear penalty: ℓ_2 C ∈ 100 values in [1e−5, 1e−2]
ℓ_1 Logistic regression	class_weight: balanced solver: liblinear penalty: ℓ_1 C ∈ [1e−4, 1e−3, 1e−2, 1e−1, 1]	class_weight: balanced solver: liblinear penalty: ℓ_1 C ∈ 100 values in [1e−5, 1e−2]
LinearSVM	C ∈ [1e−4, 1e−3, 1e−2, 1e−1, 1]	C ∈ 100 values in [1e−5, 1e−2]
CART	max_depth ∈ [5,6,7,8,9,10]	max_depth ∈ [1,2,3,4,5] min_impurity_decrease ∈ [1e−3, 2e−3, ...5e−3]
Random forest	n_estimator ∈ [100,150,200] max_depth ∈ [7,8,9]	n_estimator ∈ [50,100,200,400,600] max_depth ∈ [1,2,3] min_impurity_decrease ∈ [1e−3, 2e−3, ..., 1e−2]
XGBoost	learning_rate ∈ [0.1] n_estimator ∈ [100,150] max_depth ∈ [4,5,6]	learning_rate ∈ [0.05] n_estimator ∈ [50,100,200,400,600] max_depth ∈ [1,2,3] gamma ∈ [6,8,10,12] min_child_weight ∈ [6,8,10,12] subsample ∈ [0.5]
EBM[a]	n_estimator ∈ [60] max_tree_splits ∈ [2] learning_rate ∈ [0.1]	n_estimator ∈ [40,60,80,100] max_tree_splits ∈ [1,2,3] learning_rate ∈ [0.01] holdout_split ∈ [0.7, 0.9]

[a] The training procedure is slow for EBM, due to the size of Kentucky data, the nested cross validation we applied, and the cross-validation within the algorithm to choose number of pairwise interactions. Therefore, we tested only one set of parameters, which gave reliable results

The PSA was created by Arnold Ventures, and is a publicly available risk assessment tool. Similar to the COMPAS suite, it is comprised of three risk scores: Failure to Appear, New Criminal Activity, and New Violent Criminal Activity. Again, we compare against latter two scores. Both are additive integer models which take nine factors as input, relating to age, current charge, and criminal history. The New Criminal Activity model outputs a score from 1 to 6, while the New Violent Criminal Activity model outputs a binary score (Public Safety Assessment 2019). The PSA is an interpretable model.

Hyperparameters

Baseline Models, CART, EBM

We applied nested cross validation to tune the hyperparameters. Please refer to Table 8 for parameter details.

Table 9 Hyperparameters for additive stumps

Models	Two year	Six month
Kentucky		
General	C ∈ [1e-3, 2e−3]	C ∈ [1e−3, 1.5e−3]
Violent	C ∈ [6e−4, 8e−4, 1e−3]	C ∈ [5e−4, 7e−4]
Drug	C ∈ [1e−3, 2e−3, 2.5e−3]	C ∈ [1e−3, 2e−3]
Property	C ∈ [1e−3, 1.5e−3]	C ∈ [1e−3, 1.5e−3]
Felony	C ∈ [1e−3, 1.5e−3]	C ∈ [5e−4, 8e−4]
Misdemeanor	C ∈ [1e−3, 1.5e−3]	C ∈ [5e−4, 1e−3]
Broward		
General	[1e−2, 2e−2...1e−1]	C ∈ [1e−2, 2e−2...1e−1]
Violent	C ∈ [1e-2, 2e−2 ...7e−2]	C ∈ [1e−2, 2e−2 ...7e−2]
Drug	C ∈ [1e−2, 2e−2 ...9e−2]	C ∈ [1e−2, 2e−2 ...6e−2]
Property	C ∈ [1e−2, 2e−2 ...8e−2]	C ∈ [1e−2, 2e−2 ...6e−2]
Felony	C ∈ [1e−2, 2e−2 ...8e−2]	C ∈ [1e−2, 2e−2 ...8e−2]
Misdemeanor	C ∈ [1e−2, 2e−2 ...7e−2]	C ∈ [1e−2, 2e−2 ...7e−2]

All models use "balanced" for the class_weight, "liblinear" for the solver, and ℓ_1 for the penalty

Additive Stumps

Stumps were created for each feature as detailed in "Preprocessing Features into Binary Stumps" section. An additive model was created from the stumps using ℓ_1-penalized logistic regression, and no more than 15 original features were involved in the additive models. But multiple stumps corresponding to each feature could be used in the models. We chose to limit the size of the model to 15 original features because then at most 15 plots would be generated to visualize the full model, which is a reasonable number of visualizations for users to digest.

We started with the smallest regularization parameter on ℓ_1 penalty that provides at most 15 original features from the model. This will be our lower bound for nested cross validation. From there, we perform nested cross validation over a grid of regularization parameters, all of which are greater than or equal to the minimum value of the regularization parameter found above. Please refer to Table 9 for more details.

RiskSLIM

RiskSLIM is challenging to train, because it uses the CPLEX optimization software, which can be difficult to install and requires a license. Moreover, since RiskSLIM solves a very difficult mixed-integer nonlinear optimization problem, it can be slow to prove optimality, which makes it difficult to perform nested cross validation as nested cross validation requires many solutions of the optimization problem. A previous study (Smith 2016) also noted similar problems with algorithms that use CPLEX (this study trained on SLIM (Ustun and Rudin 2015), which is similar to the training process of RiskSLIM in that they both require CPLEX). Here we provide details of how we trained RiskSLIM to help others use the algorithm more efficiently.

- We ran ℓ_1-penalized logistic regression on the stumps training data with a relatively large regularization parameter to obtain a small subset of features (that is, we used ℓ_1 -penalized logistic regression for feature selection). Then we trained RiskSLIM using nested cross validation with this small subset of features. The maximum run-time, maximum offset, and penalty value were set to 1000 seconds, 100, and 1e−6 respectively. The coefficient range was set to [− 5, 5], which would give us small coefficients that are easy to add/subtract.
- If the model converged to optimality (optimality gap less than 5%) within 1000 seconds, we then ran ℓ_1-penalized logistic regression again with a smaller regularization parameter to obtain a slightly larger subset of features to work with. We then trained RiskSLIM with nested cross validation again on this larger subset of features. If RiskSLIM also generated an optimality gap less than 5% within 1,000 seconds and had better validation performance, we repeated this procedure.
- Once either RiskSLIM could not converge to a 5% optimality gap within 1,000 seconds, or the validation performance did not improve by adding more stumps, we stopped there, using the previously obtained RiskSLIM model as the final model.
- This procedure generally stopped with between 12 and 20 stumps from ℓ_1-penalized logistic regression. Beyond this number of stumps, we did not observe improvements in performance in validation.

See Figs. 8, 9, 10.

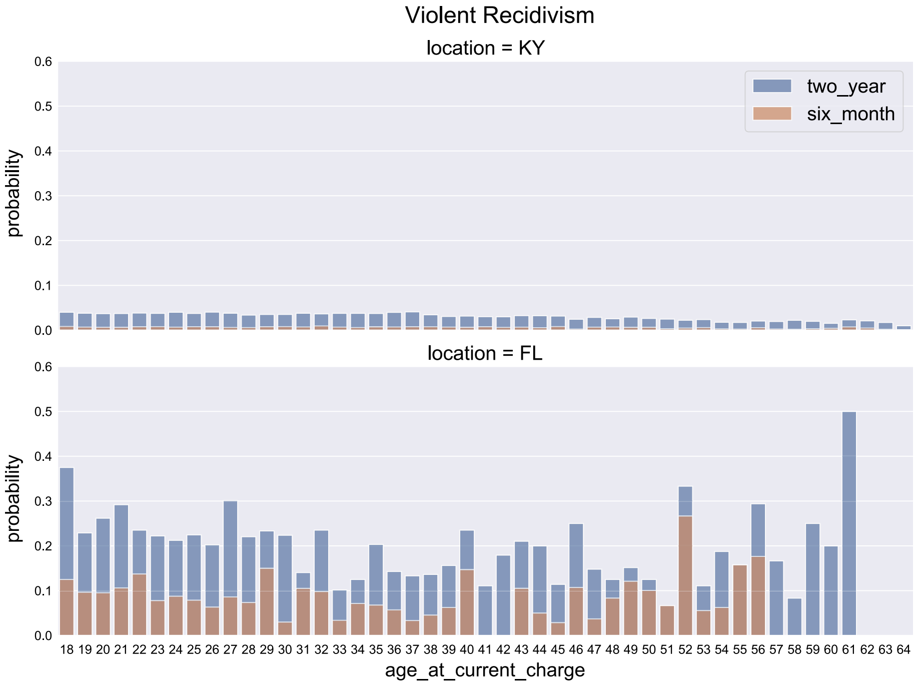


Fig. 8 Probabilities of 2-year and 6-month `violent` recidivism, given the age at current charge

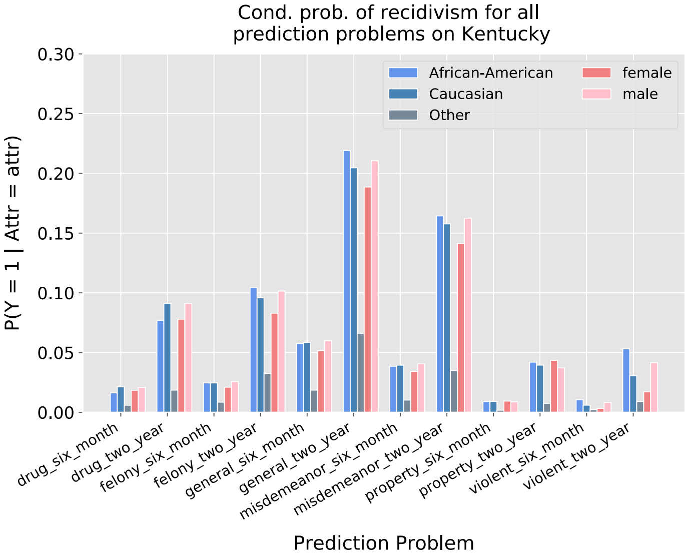
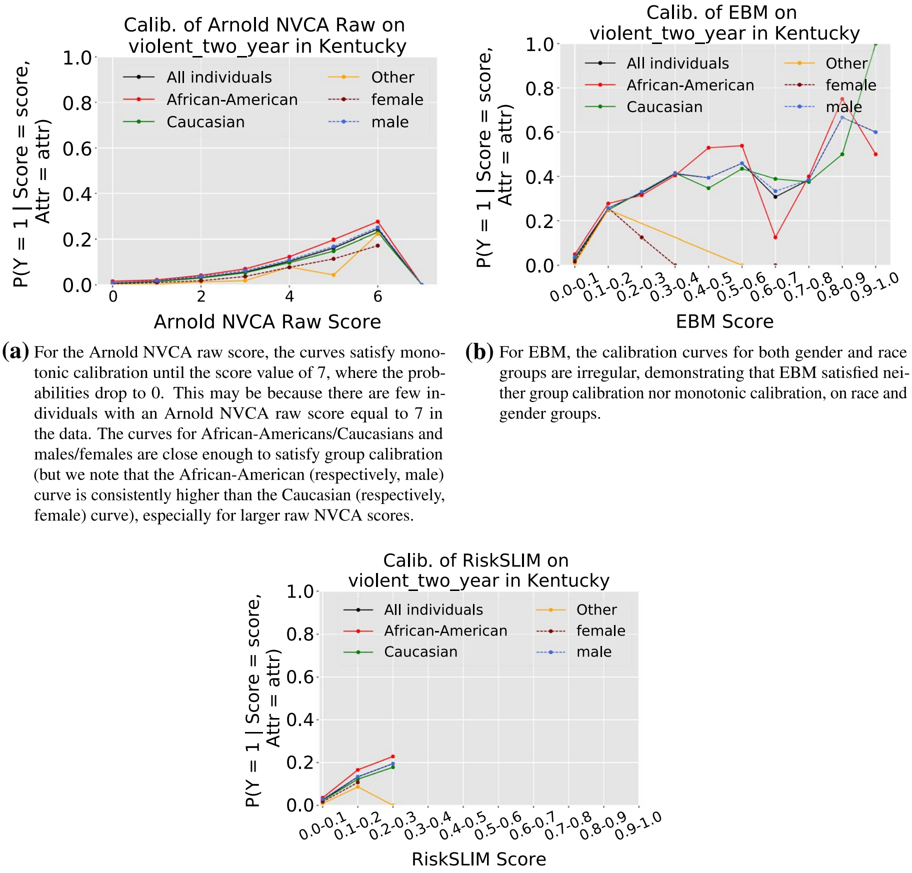


Fig. 9 Base rates of all twelve types of recidivism on Kentucky data, conditioned (separately) on race and gender



(a) For the Arnold NVCA raw score, the curves satisfy monotonic calibration until the score value of 7, where the probabilities drop to 0. This may be because there are few individuals with an Arnold NVCA raw score equal to 7 in the data. The curves for African-Americans/Caucasians and males/females are close enough to satisfy group calibration (but we note that the African-American (respectively, male) curve is consistently higher than the Caucasian (respectively, female) curve), especially for larger raw NVCA scores.

(b) For EBM, the calibration curves for both gender and race groups are irregular, demonstrating that EBM satisfied neither group calibration nor monotonic calibration, on race and gender groups.

(c) For RiskSLIM, the curves are monotonically increasing and roughly overlap with each other. The calibration curve for African-Americans is slightly higher than for the Caucasian and the "Other" race groups. For the two gender groups, the curves are close to each other. We conclude that both race and gender approximately satisfy group calibration.

Fig. 10 Calibration of the Arnold NVCA Raw, EBM and RiskSLIM for 2-year `violent` recidivism on Kentucky

See Tables 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26.

Table 10 Additive Stumps on two-year `general` recidivism

1. Age at current charge ≤ 20	0.0082	+.
2. Age at current charge ≤ 21	0.0053	+.
3. Age at current charge ≤ 24	0.0322	+.
4. Age at current charge ≤ 27	0.0270	+.
5. Age at current charge ≤ 35	0.0108	+.
6. Age at current charge ≤ 39	0.1223	+.
7. Age at current charge ≤ 43	0.0311	+.
8. Age at current charge ≤ 47	0.0686	+.
9. Prior arrest ≥ 2	0.6762	+.
10. Prior arrest ≥ 3	0.3489	+.
11. Prior arrest ≥ 4	0.2339	+.
12. Prior arrest ≥ 5	0.1226	+.
13. Prior charges ≥ 2	0.0124	+.
14. Prior charges ≥ 2 3	0.0065	+.
15. Prior violence ≥ 1	0.0474	+.
16. Prior felony ≥ 1	0.1721	+.
17. Prior misdemeanor ≥ 2	0.0162	+.
18. Prior misdemeanor ≥ 3	0.0764	+.
19. Prior misdemeanor ≥ 4	0.0733	+.
20. Prior traffic ≥ 1	0.0394	+.
21. ADE ≥ 1	0.1583	−.
22. Prior fta two year ≥ 1	0.3398	+.
23. Prior fta two year ≥ 2	0.0617	+.
24. Prior pending charge ≥ 1	0.3874	+.
25. Prior probation ≥ 1	0.2265	+.
26. Prior incarceration ≥ 1	0.3577	+.
27. 6 month ≥ 1	0.0148	−.
28. Three year ≥ 1	0.0005	+.
29. Intercept	−1.1500	+.
Add points from rows 1 to 29	Score	= .
Probability: Pr(Y = 1) = exp(score) / (1 + exp(score))		

The model consists of twenty-eight stumps with an intercept. These binary features represent fifteen original features; coefficients were rounded for display purposes only

Table 11 Race and gender distributions for Kentucky

Kentucky			
Attribute	Attribute Value	num_inds	% total
Race	African-American	42,197	16.83
Race	Asian	843	0.34
Race	Caucasian	202,341	80.69
Race	Indian	195	0.08
Race	Other	5202	2.07
Sex	Female	79,207	31.58
Sex	Male	171,571	68.42

Due to the low percentage of the Asians and Indians in Kentucky, we included them in the "Other" category in the fairness analysis

Table 12 Arnold Public Safety Assessment (PSA): New Criminal Activity (NCA)

New Criminal Activity (NCA)

Risk factor	Value	Points
Age at current arrest	23 or older	0
	22 or younger	2
Pending charge at time of offense	No	0
	Yes	3
Prior misdemeanor conviction	No	0
	Yes	1
Prior felony conviction	No	0
	Yes	1
Prior violent conviction	0	0
	1	1
	2	1
	3 or more	2
Prior FTA in past 2 years	0	0
	1	1
	2 or more	2
Prior sentence to incarceration	No	0
	Yes	2

Point scaling

Total NCA points	NCA scaled score
0	1
1	2
2	2
3	3
4	3
5	4
6	4
7	5
8	5
9	6
10	6
11	6
12	6
13	6

Table 13 Arnold Public Safety Assessment (PSA): New Violent Criminal Activity (NVCA)

New Violent Criminal Activity (NVCA)

Risk factor	Value	Points
Current violent offense	No	0
	Yes	2
Current violent offense and 20 years or younger	No	0
	Yes	1
Pending charge at time of offense	No	0
	Yes	1
Prior conviction (misdemeanor or felony)	No	0
	Yes	1
Prior violent conviction	0	0
	1	1
	2	1
	3 or more	2

Point scaling

Total NVCA points	NVCA scaled score
0	No
1	No
2	No
3	No
4	Yes
5	Yes
6	Yes
7	Yes

Table 14 Broward baseline models

Baseline models						
Labels	Logistic (ℓ_2)	Logistic(ℓ_1)	Linear SVM	RF	XGBoost	Performance range
Two year						
General	0.670 (0.021)	0.650 (0.021)	0.670 (0.020)	0.658 (0.027)	0.655 (0.022)	0.020
Violent	0.675 (0.037)	0.663 (0.039)	0.659 (0.032)	0.671 (0.036)	0.676 (0.048)	0.017
Drug	0.711 (0.048)	0.733 (0.035)	0.695 (0.037)	0.703 (0.040)	0.722 (0.039)	0.038
Property	0.717 (0.052)	0.730 (0.057)	0.683 (0.048)	0.712 (0.027)	0.733 (0.034)	0.051
Felony	0.646 (0.041)	0.648 (0.050)	0.621 (0.036)	0.647 (0.046)	0.644 (0.037)	0.027
Misdemeanor	0.630 (0.019)	0.597 (0.013)	0.628 (0.018)	0.629 (0.027)	0.627 (0.024)	0.033
Six month						
General	0.625 (0.022)	0.608 (0.022)	0.618 (0.028)	0.615 (0.026)	0.623 (0.014)	0.017
Violent	0.685 (0.024)	0.651 (0.038)	0.619 (0.036)	0.668 (0.045)	0.685 (0.033)	0.066
Drug	0.673 (0.084)	0.696 (0.022)	0.640 (0.081)	0.675 (0.055)	0.698 (0.038)	0.058
Property	0.727 (0.047)	0.725 (0.053)	0.659 (0.069)	0.687 (0.047)	0.725 (0.048)	0.068
Felony	0.611 (0.050)	0.613 (0.054)	0.580 (0.086)	0.591 (0.061)	0.585 (0.066)	0.034
Misdemeanor	0.612 (0.038)	0.586 (0.040)	0.586 (0.016)	0.593 (0.039)	0.608 (0.031)	0.027

Results are the average value of test AUCs from fivefold nested cross validation, with standard deviation listed in parentheses

Table 15 Kentucky baseline models

Baseline models						
Labels	Logistic (ℓ_2)	Logistic(ℓ_1)	Linear SVM	RF	XGBoost	Performance Range
Two year						
General	0.745 (0.004)	0.745 (0.004)	0.746 (0.004)	0.753 (0.003)	0.759 (0.003)	0.014
Violent	0.768 (0.002)	0.769 (0.003)	0.769 (0.003)	0.777 (0.005)	0.784 (0.004)	0.016
Drug	0.730 (0.003)	0.730 (0.003)	0.733 (0.003)	0.743 (0.002)	0.749 (0.002)	0.019
Property	0.785 (0.005)	0.785 (0.005)	0.787 (0.005)	0.801 (0.004)	0.806 (0.004)	0.021
Felony	0.765 (0.001)	0.765 (0.001)	0.768 (0.002)	0.779 (0.002)	0.784 (0.001)	0.019
Misdemeanor	0.729 (0.005)	0.729 (0.005)	0.730 (0.006)	0.738 (0.005)	0.744 (0.005)	0.016
Six month						
General	0.761 (0.004)	0.761 (0.004)	0.764 (0.005)	0.779 (0.003)	0.785 (0.004)	0.024
Violent	0.833 (0.007)	0.834 (0.006)	0.833 (0.007)	0.843 (0.006)	0.847 (0.005)	0.014
Drug	0.782 (0.003)	0.782 (0.003)	0.785 (0.003)	0.803 (0.003)	0.811 (0.002)	0.029
Property	0.834 (0.012)	0.834 (0.013)	0.831 (0.014)	0.857 (0.011)	0.860 (0.011)	0.029
Felony	0.799 (0.002)	0.800 (0.002)	0.804 (0.003)	0.824 (0.003)	0.831 (0.002)	0.032
Misdemeanor	0.746 (0.007)	0.746 (0.007)	0.748 (0.007)	0.765 (0.006)	0.774 (0.006)	0.028

Results are the average value of test AUCs from fivefold nested cross validation, with standard deviation listed in parentheses

Table 16 AUCs of intepretable models on Broward data

Labels	Interpretable Models					Existing Risk Models	
	CART	EBM	Additive Stumps	RiskSLIM	Performance Range	Arnold PSA	COMPAS
Two year							
General	0.613 (0.025)	0.664 (0.027)	0.651 (0.020)	0.624 (0.022)	0.051	0.605 (0.022)	0.631 (0.019)
Violent	0.613 (0.045)	0.673 (0.045)	0.665 (0.034)	0.655 (0.055)	0.059	0.649 (0.028)	–
Drug	0.666 (0.026)	0.685 (0.043)	0.716 (0.037)	0.697 (0.027)	0.049	–	–
Property	0.686 (0.059)	0.736 (0.034)	0.736 (0.033)	0.717 (0.020)	0.052	–	–
Felony	0.596 (0.033)	0.655 (0.050)	0.631 (0.028)	0.590 (0.036)	0.065	–	–
Misdemeanor	0.577 (0.036)	0.636 (0.029)	0.609 (0.020)	0.579 (0.015)	0.059	–	–
Six month							
General	0.549 (0.021)	0.622 (0.022)	0.620 (0.019)	0.585 (0.021)	0.074	0.577 (0.018)	0.609 (0.019)
Violent	0.631 (0.050)	0.680 (0.040)	0.676 (0.029)	0.671 (0.039)	0.049	0.675 (0.038)	–
Drug	0.569 (0.074)	0.672 (0.043)	0.656 (0.068)	0.650 (0.068)	0.102	–	–
Property	0.637 (0.052)	0.725 (0.031)	0.725 (0.036)	0.703 (0.023)	0.089	–	–
Felony	0.513 (0.014)	0.606 (0.049)	0.574 (0.036)	0.561 (0.045)	0.093	–	–
Misdemeanor	0.535 (0.021)	0.608 (0.042)	0.582 (0.036)	0.576 (0.024)	0.073	–	–

For the violence problem, we use the Arnold New Violent Criminal Activity score. For the general problem, we use the Arnold New Criminal Activity score

Table 17 AUCs of interpretable models on Kentucky data

Labels	Interpretable Models					Existing Risk Models
	CART	EBM	Additive Stumps	RiskSLIM	Performance Range	Arnold PSA
Two year						
General	0.746 (0.003)	0.751 (0.004)	0.748 (0.004)	0.708 (0.003)	0.042	0.711 (0.004)
Violent	0.763 (0.007)	0.777 (0.004)	0.770 (0.005)	0.744 (0.008)	0.032	0.743 (0.003)
Drug	0.736 (0.002)	0.740 (0.001)	0.738 (0.002)	0.708 (0.005)	0.032	–
Property	0.790 (0.003)	0.798 (0.006)	0.796 (0.005)	0.761 (0.003)	0.037	–
Felony	0.771 (0.002)	0.776 (0.001)	0.773 (0.002)	0.757 (0.007)	0.019	–
Misdemeanor	0.730 (0.005)	0.735 (0.005)	0.729 (0.006)	0.701 (0.002)	0.033	–
Six month						
General	0.772 (0.005)	0.773 (0.004)	0.771 (0.004)	0.737 (0.002)	0.037	0.718 (0.004)
Violent	0.822 (0.011)	0.843 (0.006)	0.836 (0.004)	0.810 (0.009)	0.033	0.794 (0.011)
Drug	0.794 (0.003)	0.793 (0.004)	0.796 (0.004)	0.763 (0.004)	0.033	–
Property	0.839 (0.014)	0.850 (0.012)	0.851 (0.010)	0.832 (0.010)	0.019	–
Felony	0.811 (0.003)	0.820 (0.003)	0.813 (0.003)	0.790 (0.006)	0.030	–
Misdemeanor	0.760 (0.006)	0.757 (0.006)	0.751 (0.006)	0.705 (0.005)	0.055	–

For the violence problem, we use the Arnold New Violent Criminal Activity score. For the general problem, we use the Arnold New Criminal Activity score

Table 18 Training baseline models and interpretable models on the Kentucky data set using fivefold nested cross validation and testing the best-performing model on the Broward data set

Labels	Baseline models					Interpretable models			
	Logistic (ℓ_2)	Logistic (ℓ_1)	Linear SVM	Random Forest	XGBoost	CART	EBM	Additive Stumps	RiskSLIM
Two year									
General	0.615 (0.001)	0.614 (0.001)	0.610 (0.000)	0.619 (0.001)	0.617 (0.003)	0.595 (0.009)	0.612 (0.002)	0.608 (0.001)	0.568 (0.000)
Violent	0.655 (0.001)	0.653 (0.002)	0.630 (0.000)	0.652(0.000)	0.652 (0.004)	0.622 (0.030)	0.640 (0.010)	0.652(0.002)	0.629 (0.018)
Drug	0.629 (0.001)	0.629 (0.001)	0.618 (0.000)	0.614 (0.002)	0.637 (0.002)	0.621 (0.010)	0.629 (0.003)	0.631 (0.001)	0.625 (0.000)
Property	0.664 (0.001)	0.672 (0.001)	0.649 (0.000)	0.668 (0.002)	0.674 (0.008)	0.649 (0.017)	0.665 (0.011)	0.659 (0.001)	0.639 (0.021)
Felony	0.630 (0.001)	0.630 (0.001)	0.624 (0.000)	0.631 (0.001)	0.627 (0.005)	0.611 (0.003)	0.623 (0.005)	0.624 (0.000)	0.614 (0.000)
Misdemeanor	0.558 (0.000)	0.558 (0.000)	0.551 (0.000)	0.561 (0.001)	0.576 (0.002)	0.555 (0.004)	0.571 (0.003)	0.557 (0.000)	0.539 (0.002)
Six month									
General	0.577 (0.002)	0.576 (0.001)	0.569 (0.000)	0.577 (0.001)	0.581 (0.002)	0.562 (0.007)	0.571 (0.004)	0.562 (0.001)	0.553 (0.000)
Violent	0.641 (0.002)	0.644 (0.001)	0.614 (0.000)	0.643 (0.001)	0.626 (0.004)	0.611 (0.013)	0.622 (0.009)	0.650 (0.001)	0.637 (0.002)
Drug	0.607 (0.004)	0.604 (0.003)	0.589 (0.000)	0.567 (0.005)	0.593 ((0.007)	0.580 (0.018)	0.618 (0.006)	0.576 (0.001)	0.566 ((0.020)
Property	0.662 (0.001)	0.665 (0.002)	0.635 (0.000)	0.652 (0.002)	0.656 (0.013)	0.634 (0.016)	0.657 (0.008)	0.640 (0.004)	0.619 (0.000)
Felony	0.586 (0.001)	0.584 (0.002)	0.575 (0.000)	0.589 (0.002)	0.580 (0.002)	0.563 (0.003)	0.571 (0.005)	0.574 (0.001)	0.550 (0.001)
Misdemeanor	0.558 (0.002)	0.558 (0.000)	0.550 (0.000)	0.552 (0.002)	0.563 (0.004)	0.554 (0.012)	0.559 (0.002)	0.542 (0.001)	0.526 (0.003)

Table 19 Training baseline models and interpretable models on the Broward County data set using fivefold nested cross validation and testing the resulting best-performing model on a held out portion of the Broward data set

Labels	Baseline models					Interpretable models			
	Logistic (ℓ_2)	Logistic (ℓ_1)	Linear SVM	Random forest	XGBoost	CART	EBM	Additive stumps	RiskSLIM
Two year									
General	0.669 (0.020)	0.649 (0.021)	0.670 (0.020)	0.657 (0.034)	0.659 (0.019)	0.629 (0.028)	0.663 (0.031)	0.644 (0.027)	0.622 (0.021)
Violent	0.679 (0.038)	0.662 (0.035)	0.662 (0.034)	0.675 (0.037)	0.677 (0.050)	0.600 (0.037)	0.675 (0.049)	0.673 (0.035)	0.670 (0.032)
Drug	0.716 (0.047)	0.734 (0.034)	0.702 (0.043)	0.688 (0.044)	0.720 (0.034)	0.672 (0.041)	0.690 (0.054)	0.709 (0.044)	0.706 (0.027)
Property	0.721 (0.057)	0.731 (0.057)	0.687 (0.052)	0.725 (0.039)	0.729 (0.040)	0.685 (0.058)	0.738 (0.031)	0.733 (0.039)	0.703(0.036)
Felony	0.651 (0.040)	0.652 (0.053)	0.622 (0.036)	0.649 (0.045)	0.647 (0.039)	0.598 (0.034)	0.656 (0.050)	0.640 (0.031)	0.603 (0.042)
Misdemeanor	0.634 (0.017)	0.602 (0.012)	0.632 (0.017)	0.629 (0.022)	0.624 (0.020)	0.585 (0.041)	0.633 (0.025)	0.603 (0.016)	0.558 (0.026)
Six month									
General	0.624 (0.024)	0.607 (0.019)	0.619 (0.026)	0.620 (0.025)	0.621 (0.019)	0.553 (0.014)	0.620 (0.027)	0.617 (0.035)	0.600 (0.021)
Violent	0.680 (0.027)	0.650 (0.038)	0.614 (0.039)	0.670 (0.039)	0.689 (0.031)	0.623 (0.043)	0.683 (0.040)	0.683 (0.032)	0.691 (0.032)
Drug	0.672 (0.082)	0.696 (0.025)	0.649 (0.080)	0.687 (0.065)	0.686 (0.044)	0.569 (0.074)	0.655 (0.035)	0.704 (0.054)	0.719 (0.039)
Property	0.726 (0.049)	0.725 (0.053)	0.648 (0.058)	0.698 (0.046)	0.720 (0.052)	0.637 (0.052)	0.723 (0.030)	0.699 (0.038)	0.663 (0.048)
Felony	0.620 (0.058)	0.613 (0.054)	0.587 (0.086)	0.611 (0.076)	0.601 (0.047)	0.524 (0.015)	0.605 (0.052)	0.584 (0.034)	0.557 (0.043)
Misdemeanor	0.616 (0.030)	0.583 (0.039)	0.590 (0.022)	0.601 (0.049)	0.620 (0.044)	0.543 (0.033)	0.612 (0.050)	0.576 (0.037)	0.556 (0.040)

Table 20 Training baseline and interpretable models on the Broward County data set using fivefold nested cross validation and testing the resulting best-performing model on the Kentucky data set

Labels	Baseline Models					Interpretable Models			
	Logistic (ℓ_2)	Logistic (ℓ_1)	Linear SVM	Random Forest	XGBoost	CART	EBM	Additive Stumps	RiskSLIM
Two year									
General	0.664 (0.007)	0.653 (0.001)	0.658 (0.007)	0.701 (0.005)	0.689 (0.006)	0.626 (0.025)	0.704 (0.003)	0.653 (0.009)	0.649 (0.037)
Violent	0.674 (0.005)	0.650 (0.007)	0.611 (0.013)	0.729 (0.005)	0.724 (0.005)	0.589 (0.053)	0.720 (0.005)	0.657 (0.018)	0.663 (0.025)
Drug	0.649 (0.008)	0.632 (0.003)	0.554 (0.005)	0.655 (0.022)	0.650 (0.006)	0.613 (0.013)	0.656 (0.008)	0.626 (0.009)	0.634 (0.012)
Property	0.628 (0.022)	0.663 (0.014)	0.556 (0.017)	0.695 (0.018)	0.669 (0.023)	0.548 (0.018)	0.687 (0.011)	0.590 (0.014)	0.593 (0.052)
Felony	0.671 (0.006)	0.661 (0.002)	0.592 (0.014)	0.724 (0.003)	0.706 (0.014)	0.592 (0.042)	0.725 (0.006)	0.676 (0.023)	0.631 (0.059)
Misdemeanor	0.638 (0.007)	0.619 (0.026)	0.579 (0.010)	0.665 (0.011)	0.645 (0.014)	0.574 (0.053)	0.669 (0.007)	0.621 (0.017)	0.631 (0.025)
Six month									
General	0.676 (0.006)	0.665 (0.004)	0.601 (0.011)	0.698 (0.009)	0.685 (0.010)	0.613 (0.018)	0.709 (0.005)	0.663 (0.012)	0.602 (0.046)
Violent	0.653 (0.015)	0.662 (0.021)	0.533 (0.011)	0.762 (0.047)	0.773 (0.007)	0.625 (0.059)	0.757 (0.004)	0.728 (0.026)	0.723 (0.004)
Drug	0.663 (0.031)	0.678 (0.008)	0.521 (0.006)	0.682 (0.009)	0.658 (0.027)	0.600 (0.082)	0.609 (0.037)	0.619 (0.025)	0.635 (0.017)
Property	0.681 (0.012)	0.708 (0.009)	0.529 (0.012)	0.719 (0.053)	0.718 (0.010)	0.555 (0.007)	0.715 (0.018)	0.643 (0.022)	0.696 (0.053)
Felony	0.685 (0.008)	0.679 (0.008)	0.556 (0.011)	0.719 (0.018)	0.683 (0.025)	0.552 (0.049)	0.724 (0.010)	0.652 (0.039)	0.621 (0.036)
Misdemeanor	0.664 (0.003)	0.658 (0.008)	0.558 (0.016)	0.670 (0.004)	0.662 (0.006)	0.604 (0.019)	0.676 (0.006)	0.615 (0.019)	0.583 (0.070)

Table 21 Training baseline models and interpretable models on the Kentucky data set using fivefold nested cross validation and testing the resulting best-performing model on a held out portion of the Kentucky data set

Labels	Baseline models					Interpretable models			
	Logistic (ℓ_2)	Logistic (ℓ_1)	Linear SVM	Random Forest	XGBoost	CART	EBM	Additive stumps	RiskSLIM
Two year									
General	0.739 (0.003)	0.739 (0.003)	0.740 (0.004)	0.752 (0.004)	0.757 (0.003)	0.746 (0.003)	0.750 (0.004)	0.747 (0.004)	0.704 (0.004)
Violent	0.765 (0.001)	0.766 (0.002)	0.767 (0.002)	0.776 (0.004)	0.783 (0.004)	0.763 (0.007)	0.776 (0.004)	0.771 (0.005)	0.741 (0.010)
Drug	0.723 (0.002)	0.723 (0.002)	0.727 (0.002)	0.739 (0.002)	0.745 (0.002)	0.733 (0.002)	0.737 (0.002)	0.734 (0.003)	0.708 (0.002)
Property	0.780 (0.004)	0.779 (0.004)	0.784 (0.004)	0.801 (0.004)	0.805 (0.004)	0.790 (0.004)	0.797 (0.005)	0.796 (0.005)	0.764 (0.009)
Felony	0.758 (0.002)	0.758 (0.002)	0.763 (0.002)	0.778 (0.002)	0.783 (0.001)	0.771 (0.002)	0.775 (0.001)	0.773 (0.001)	0.765 (0.001)
Misdemeanor	0.722 (0.005)	0.722 (0.005)	0.724 (0.006)	0.736 (0.006)	0.742 (0.005)	0.729 (0.005)	0.733 (0.006)	0.729 (0.006)	0.693 (0.010)
Six month									
General	0.752 (0.004)	0.752 (0.004)	0.757 (0.004)	0.775 (0.003)	0.780 (0.003)	0.769 (0.005)	0.770 (0.004)	0.768 (0.004)	0.736(0.004)
Violent	0.828 (0.006)	0.830 (0.005)	0.834 (0.005)	0.843 (0.005)	0.846 (0.005)	0.821 (0.011)	0.842 (0.005)	0.837 (0.004)	0.809 (0.005)
Drug	0.770 (0.003)	0.771 (0.003)	0.777 (0.004)	0.794 (0.004)	0.799 (0.002)	0.783 (0.005)	0.785 (0.004)	0.786 (0.004)	0.752 (0.006)
Property	0.830 (0.010)	0.829 (0.011)	0.830 (0.013)	0.856 ((0.009)	0.860 (0.011)	0.839 (0.014)	0.849 (0.011)	0.851 (0.010)	0.835 (0.009)
Felony	0.790 (0.002)	0.791 (0.002)	0.798 (0.003)	0.823 (0.003)	0.829 (0.003)	0.811 (0.005)	0.818 (0.004)	0.812 (0.004)	0.790 (0.005)
Misdemeanor	0.735 (0.006)	0.735 (0.006)	0.740 (0.007)	0.760 (0.005)	0.766 (0.005)	0.754 (0.005)	0.753 (0.006)	0.750 (0.006)	0.705 (0.005)

Table 22 AUCs of the Arnold NVCA Raw, EBM and RiskSLIM on Kentucky for two-year `violent` recidivism, conditioned on sensitive attributes. AUC ranges are also given for each sensitive attribute class

Kentucky								
		Race				Sex		
`Model`	`Label`	`Afr-Am.`	`Cauc.`	`Other Race`	`race_range`	`Female`	`Male`	`sex_range`
Arnold NVCA Raw	violent_two_year	0.728	0.740	0.767	0.039	0.728	0.734	0.006
EBM	violent_two_year	0.775	0.770	0.766	0.009	0.744	0.766	0.022
RiskSLIM	violent_two_year	0.744	0.736	0.680	0.063	0.706	0.730	0.024

Table 23 Two year prediction problems—Kentucky

Two year general recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-2 + score)))		
Number of prior arrests≥ 2	1 points	+.
number of prior arrests≥ 3	1 points	+.
Number of prior arrests≥ 5	1 points	+.
Add points from rows 1 to 3	Score	= .
Two year violent recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-6 + score)))		
Sex = male	1 points	+.
Age at current charge ≤ 27	1 points	+.
Number of prior arrests≥ 2	1 points	+.
Number of prior violent charges≥ 1	1 points	+.
Sentenced to incarceration before = Yes	1 points	+.
Add points from rows 1 to 5	Score	= .
Two year drug recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
Number of prior arrests≥ 2	1 points	+.
Number of prior drug related charges≥ 1	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Add points from rows 1 to 3	Score	= .
Two year property recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
Number of prior property related charges≥ 1	1 points	+.
Number of prior arrests≥ 3	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Number of prior ADE ≥ 1	-1 points	+.
Add points from rows 1 to 4	Score	= .
Two year felony recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-5 + score)))		
Age at current charge ≤ 43	1 points	+.
Number of prior arrests≥ 2	1 points	+.
Number of prior felony level charges≥ 1	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Sentenced to incarceration before = Yes	1 points	+.
Add points from rows 1 to 5	Score	= .

Table 23 (continued)

Two year misdemeanor recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-3 + score)))		
Number of prior arrests≥ 2	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Sentenced to incarceration before = Yes	1 points	+.
Add points from rows 1 to 3	Score	= .

Here, counts of prior arrests indicate the counts of arrests with at least one convicted charge

All charges mentioned are convicted charges. ADE indicates assignment to alcohol and drug education classes

Table 24 Six month prediction problems—Kentucky

Six Month general recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
Number of prior arrests≥ 2	1 points	+.
Number of prior arrests≥ 4	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Add points from rows 1 to 3	Score	= .
Six month violent recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-7 + score)))		
Number of prior violent charges≥ 1	1 points	+.
Number of prior arrests≥ 3	1 points	+.
Number of prior felony level charges≥ 1	1 points	+.
Current violent charge = Yes	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Add points from rows 1 to 5	Score	= .
Six month drug recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-5 + score)))		
Number of prior drug related charges≥ 1	1 points	+.
Number of prior drug related charges≥ 3	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Number of prior ADE≥ 1	−1 points	+.
Add points from rows 1 to 4	Score	= .

Table 24 (continued)

Six month property recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-7 + score)))		
Number of prior property related charges≥ 1	2 points	+.
Number of prior felony level charges≥ 1	1 points	+.
Number of prior FTA within last 2 years ≥ 1	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Add points from rows 1 to 4	Score	= .
Six month felony recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-5 + score)))		
Number of prior arrests≥ 3	1 points	+.
Number of prior felony level charges≥ 1	1 points	+.
Number of times charged with a new offense when there is a pending case≥ 1	1 points	+.
Add points from rows 1 to 3	Score	= .
Six month misdemeanor recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
Number of prior arrests≥ 2	1 points	+.
Number of prior arrests≥ 4	1 points	+.
Add points from rows 1 to 2	Score	= .

Here, counts of prior arrests indicate the counts of arrests with at least one convicted charge. All charges mentioned are convicted charges. ADE means assignment to alcohol and drug education classes

Table 25 Two year prediction problems—Broward

Two Year General Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-2 + score)))		
age at current charge ≤ 31	1 points	+.
number of prior misdemeanor level charges ≥ 4	1 points	+.
had charge(s) within last three years = Yes	1 points	+.
ADD POINTS FROM ROWS 1 TO 3	Score	= .
Two Year Violent Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
age at current charge≤ 30	1 points	+.
number of prior violent charges≥ 4	1 points	+.
number of prior arrests≥ 7	1 points	+.
current violent charge=Yes	1 points	+.
had charge(s) within last three year = Yes	1 points	+.
ADD POINTS FROM ROWS 1 TO 5	Score	= .
Two Year Drug Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
age at current charge≤ 33	1 points	+.
number of prior drug related charges≥ 1	1 points	+.
number of prior drug related charges≥ 4	1 points	+.
ADD POINTS FROM ROWS 1 TO 3	Score	= .
Two Year Property Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
age at current charge ≤ 18	1 points	+.
age at current charge ≤ 23	1 points	+.
number of prior property related charges≥ 1	1 points	+.
number of prior property related charges≥ 5	1 points	+.
number of prior violent charges≥ 4	1 points	+.
ADD POINTS FROM ROWS 1 TO 5	Score	= .
Two Year Felony Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-3 + score)))		
age at current charge ≤ 33	1 points	+.
number of prior misdemeanor level charges≥ 4	1 points	+.
number of prior property related charges≥ 4	1 points	+.
ADD POINTS FROM ROWS 1 TO 3	Score	= .
Two Year Misdemeanor Recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-2 + score)))		
age at first charge≤ 30	1 points	+.
number of FTA within last 2 years≥ 1	1 points	+.
Add points from rows 1 to 29	Score	= .

Here, counts of prior arrests indicate the counts of arrests with at least one non-convicted or convicted charge. All charges mentioned are non-convicted charges

Table 26 Six Month Prediction Problems—Broward

Six month general recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-3 + score)))		
Age at first charge≤ 28	1 points	+.
Had charge(s) within last three years = Yes	1 points	+.
Add points from rows 1 to 29	Score	= .
Six month violent recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-4 + score)))		
Current violent charge = Yes	1 points	+.
Number of prior violent charges ≥ 4	1 points	+.
Had charge(s) within last three years = Yes	1 points	+.
Add points from rows 1 to 3	Score	= .
Six month drug recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-5 + score)))		
Age at first charge≤ 21	1 points	+.
number of prior drug charges≥ 2	1 points	+.
Had charge(s) within last year = Yes	1 points	+.
Add points from rows 1 to 3	Score	= .
Six month property recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-5 + score)))		
Age at current charge≤ 29	1 points	+ .
Number of prior misdemeanor level charges≥ 5	1 points	+.
Number of prior property related charges≥ 1	1 points	+.
Number of prior property related charges≥ 4	1 points	+.
Add points from rows 1 to 4	Score	= .
Six month felony recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-3 + score)))		
age at current charge≤ 29	1 points	+.
number of prior property related charges≥ 4	1 points	+.
Add points from rows 1 to 29	Score	= .
Six month misdemeanor recidivism		
Pr(Y = +1) = 1 / (1 + exp(-(-3 + score)))		
Age at current charge≤ 19	1 points	+.
Number of prior weapon related charges≥ 1	1 points	+.
Had charge(s) within last three years = Yes	1 points	+.
Add points from rows 1 to 3	Score	= .

Here, counts of prior arrests indicate the counts of arrests with at least one non-convicted or convicted charge. All charges mentioned are non-convicted charges

Acknowledgements We thank the Broward County Sheriff's office and the Kentucky Department of Shared Services, Research and Statistics for their assistance and provision of data. We would also like to thank Daniel Sturtevant from the Kentucky Department of Shared Services, Research and Statistics for providing significant insight into the Kentucky data set, and Berk Ustun for his advice on training RiskSLIM. Finally, we thank Brandon Garrett from Duke, Stuart Buck and Kristin Bechtel from Arnold Ventures, and Kathy Schiflett, Christy May, and Tara Blair from Kentucky Pretrial Services for their thoughtful comments on the article.

Funding This study was partially supported by Arnold Ventures, the Department of Computer Science at Duke University, the Department of Electrical and Computer Engineering at Duke University, and the Lord Foundation of North Carolina. This report represents the findings of the authors and does not represent the views of any of the funding agencies.

Data Availability Statement The Broward County, FL dataset generated and analyzed during the current study is available from the corresponding author on request. The Kentucky dataset is not publicly available but can be accessed through a special data request to the Kentucky Department of Shared Services, Research and Statistics.

Declarations

Conflict of interest No additional institutional conflicts.

Code Availability Our code is here: https://github.com/BeanHam/interpretable-machine-learning.

References

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. In: Proceedings of the 35th international conference on machine learning. https://proceedings.mlr.press/v80/agarwal18a.html

Agarwal A, Dudík M, Wu ZS (2019) Fair regression: quantitative definitions and reduction-based algorithms. In: Proceedings of the 36th international conference on machine learning. https://proceedings.mlr.press/v97/agarwal19d.html

Alfred B (2006) The crime drop in America: an explanation of some recent crime trends. J Scand Stud Criminol Crime Prev 7:17–35

American Law Institute (2017) Model penal code. https://www.ali.org/projects/show/sentencing/

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2018) Certifiably optimal rule lists for categorical data. J Mach Learn Res 19:1–79

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. Technical report, ProPublica

Barabas C, Dinakar K, Doyle C (2019) The problems with risk assessment tools. The New York Times. https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671–732

Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. Exp Criminol 13:193–216

Berk RA, He Y, Sorenson SB (2005) Developing a practical forecasting screener for domestic violence incidents. Eval Rev 29(4):358–383

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2017a) A convex framework for fair regression. arXiv:1706.02409

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2017b) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res

Bindler A, Hjalmarsson R (2018) How punishment severity affects jury verdicts: evidence from two natural experiments. Am Econ J 10

Binns R (2018) Fairness in machine learning: lessons from political philosophy. J Mach Learn Res 81:1–11

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, New York

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the COMPAS risk and needs assessment system. Crim Justice Behav 36(1):21–40

Bureau of Justice Assistance (2020) History of risk assessment. Bureau of Justice Assistance. https://psrac.bja.ojp.gov/basics/history

Burgess EW (1928) Factors determining success or failure on parole

Bushway SD, Piehl AM (2007) The inextricable link between age and criminal history in sentencing. Crime Delinq 53(1):156–183

Cadigan TP, Lowenkamp CT (2011) Implementing risk assessment in the federal pretrial services system. Federal Probation 75(2)

Carollo J, Hines M, Hedlund J (2007) Expanded validation of a decision aid for pretrial conditional release. Technical report, Central Connecticut State University

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163

Cook P, Laub J (2002) After the epidemic recent trends in youth violence in the United States. Crime Justice 29:1–37

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv:180800023v2

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 797–806

CPAT of Pretrial Services (2015) The colorado pretrial assessment tool (cpat): Administration, scoring, and reporting manual. https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=47e978bb-3945-9591-7a4f-77755959c5f5

Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgment. Science 243(4899):1668–1674

Defronzo J (1984) Climate and crime: tests of an FBI assumption. Environ Behav 16

Desmarais S, Garrett B, Rudin C (2019) Risk assessment tools are not a failed 'minority report'. Law360. https://www.law360.com/access-to-justice/articles/1180373/risk-assessment-tools-are-not-a-failed-minority-report-

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity: performance of the COMPAS risk scales in Broward county. Technical report, Northpointe, Inc

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, ITCS '12, pp 214–226, New York. ACM

Electronic Privacy Information Center (2016) Algorithms in the criminal justice system. Electronic Privacy Information Center. https://epic.org/algorithmic-transparency/crim-justice/

Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874

Flores AW, Lowenkamp CT, Bechtel K (2016) False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals". Federal Probation 80(2)

Frase RS, Roberts J, Hester R, Mitchell KL (2015) Robina institute of criminal law and criminal justice, criminal history enhancements sourcebook. https://robinainstitute.umn.edu/publications/criminal-history-enhancements-sourcebook

Freeman K (2016) Algorithmic injustice: How the wisconsin supreme court failed to protect due process rights in state v. loomis. N C J Law Technol 18. http://ncjolt.org/wp-content/uploads/2016/12/Freeman_Final.pdf

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378

Garrett B, Stevenson M (2020) Open risk assessments. Behav Sci Law. https://sites.law.duke.edu/justsciencelab/2019/09/15/comment-on-pattern-by-brandon-l-garrett-megan-t-stevenson/

Gelb A, Velazquez T, Trust PC, of America, U. S. (2018) The changing state of recidivism: fewer people going back to prison. The Pew Charitable Trusts

Goel S, Rao JM, Shroff R (2016) Precinct or prejudice? understanding racial disparities in New York city's stop-and-frisk policy. Inst Math Stat 10(1):365–394

Grove WM, Meehl PE (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychol Public Policy Law 2(2):293

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

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3315–3323

Harris GT, Rice ME (2008) Encyclopedia of Psychology and Law, chapter Violence Risk Appraisal Guide (VRAG), p 848. SAGE Publications, Inc.

Hart H (1924) Predicting parole success. J Crim Law Criminol 14

Hoffman PB, Adelberg S (1980) The salient factor score: a nontechnical overview. Federal Probation 44:44

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

James N (2018) Risk and needs assessment in the federal prison system. Technical report, Congressional Research Service

Kehl D, Guo P, Kessler S (2017) Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing. https://cyber.harvard.edu/publications/2017/07/Algorithms

Kim J, Bushway S, Tsao H (2016) Identifying classes of explanation for crime drop: period and cohort effects for New York state. J Quant Criminol 32:357–375

Kleiman M, Ostrom BJ, Cheesman FL (2007) Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. Crime Delinq 53(1):106–132

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the 8th conference on innovations in theoretical computer science

Lakkaraju H, Rudin C (2017) Learning cost-effective and interpretable treatment regimes. In: Singh A, Zhu J (eds) Proceedings of the 20th international conference on artificial intelligence and statistics, vol 54 of proceedings of machine learning research, pp 166–175, Fort Lauderdale. PMLR. http://proceedings.mlr.press/v54/lakkaraju17a.html

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. Technical report, ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Latessa E, Smith P, Lemke R, Makarios M, Lowenkamp C (2009) Creation and validation of the ohio risk assessment system. Technical report, University of Cincinnati School of Criminal Justice Center for Criminal Justice Research

Lazarsfeld PF (1974) An evaluation of the pretrial services agency of the Vera institute of justice. Vera Institute, New York

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

Ludwig J, Mullainathan S (2021) Fragile algorithms and fallible decision-makers: lessons from the justice system. J Econ Perspect 35(4):71–96

Matthews B, Minton J (2017) Rethinking one of the criminology's 'brute facts': the age-crime curve and the crime drop in Scotland. Eur J Criminol 15(3):296–320

MHS Assessments (2017) Level of service/case management inventory: an offender management system. *MHS Public Safety*. https://issuu.com/mhs-assessments/docs/ls-cmi.lsi-r.brochure_insequence

Milgram A (2014) Why smart statistics are the key to fighting crime

Mishra A (2014) Climate and crime. Global J Sci Front Res 14

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

Neuilly M-A, Zgoba KM, Tita GE, Lee SS (2011) Predicting recidivism in homicide offenders using classification tree analysis. Homicide Stud 15(2):154–176

Northpointe (2013) Practitioner's Guide to COMPAS Core. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Northpointe Inc. (2009) Measurement & treatment implications of COMPAS core scales. Technical report, Northpointe Inc

O'Neil C (2016) Weapons of math destruction. Crown Books, New York

Orbis (2014) Service planning instrument: an innovative assessment and case planning tool. https://orbispartners.com/wp-content/uploads/2014/07/SPIn-Brochure.pdf

Palocsay W, PingWang S, Brookshire RG (2000) Predicting criminal recidivism using neural networks. Socio-Econ Plan Sci 34:271–284

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger K (2017) On fairness and calibration. In: Advances in neural information processing systems, pp 5680–5689

Pretrial Justice Institute (2020) Updated position on pretrial risk assessment tools. Pretrial Justice Institute. https://university.pretrial.org/viewdocument/updated-statement-on-pretrial-risk

Public Safety Assessment (2019) Risk factors and formulas. Laura and John Arnold Foundation. https://www.psapretrial.org/about/

Ranson M (2014) Crime, weather, and climate change. J Environ Econ Manag 67

Richard B (2019) Accuracy and fairness for juvenile justice risk assessments. J Empir Leg Stud 16:174–194

Roberts J, von Hirsch A (2010) Previous convictions at sentening - theoretical and applied perspective. Bloomsbury Publishing, London

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215

Rudin C, Wang C, Coker B (2020) The age of secrecy and unfairness in recidivism prediction. Harvard Data Sci Rev 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269

Sherman LW (2007) The power few: experimental criminology and the reduction of harm. J Exp Criminol 3(4):299–321

Singh A, Mohapatra S (2021) Development of risk assessment framework for first time offenders using ensemble learning. IEEE Access 9:135024–135033

Skeem J, Lin Z, Jung J, Goel S (2020) The limits of human predictions of recidivism. Sci Adv 6

Smith B (2016) Auditing deep neural networks to understand recidivism predictions. PhD thesis, Haverford College

Soares E, Angelov PP (2019) Fair-by-design explainable models for prediction of recidivism. arXiv:abs/1910.02043

Starr SB (2015) The risk assessment era: an overdue debate. Federal Sentencing Reporter 27:205–206

Stevenson M (2018) Assessing risk assessment in action. Minnesota Law Review. http://www.minnesotalawreview.org/wp-content/uploads/2019/01/13Stevenson_MLR.pdf

Stevenson MT, Slobogin C (2018) Algorithmic risk assessments and the double-edged sword of youth. Washington Univ Law Rev 96(18–36)

The Leadership Conference on Civil and Human Rights (2018) The use of pretrial “risk assessment” instrument: a shared statement of civil rights concerns. http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf

Tollenaar N, van der Heijden P (2013) Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. J R Stat Soc A Stat Soc 176(2):565–584

Turner S, Hess J, Jannetta J (2009) Development of the California Static Risk Assessment Instrument (CSRA). CEBC Working Papers

United States Census Bureau (2015) Hispanic or latino origin by race 2011–2015 American community survey 5-year estimates. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_B03002&prodType=table

United States Census Bureau (2019) Quickfacts kentucy United States. https://www.census.gov/quickfacts/fact/table/KY,US/PST04521

Ustun B, Rudin C (2015) Supersparse linear integer models for optimized medical scoring systems. Mach Learn 1–43

Ustun B, Rudin C (2017) Optimized risk scores. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining

Ustun B, Rudin C (2019) Learning optimized risk scores. J Mach Learn Res 20(150):1–75

Vapnik V, Chervonenkis A (1964) A note on one class of perceptrons. Autom Remote Control 25

Verma S, Rubin J (2018) Fairness definitions explained. In: ACM/IEEE international workshop on software fairness, pp 1–7. ACM

Virginia Department of Criminal Justice Services (2018) Virginia pretrial risk assessment instrument - (vprai). https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/virginia-pretrial-risk-assessment-instrument-vprai_0.pdf

Wexler R (2017) When a computer program keeps you in jail: how computers are harming criminal justice. New York Times, p 27. Section A

Wolfgang ME (1987) Delinquency in a birth cohort. University of Chicago Press, Chicago

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International conference on machine learning, pp 325–333

Zeng J, Ustun B, Rudin C (2017) Interpretable classification models for recidivism prediction. J R Stat Soc A Stat Soc 180(3):689–722

Zweig J (2010) Extraordinary conditions of release under the bail reform act. Harvard J Legis 47:555–585