AI for Predicting Recidivism

Monica STEBBINS

2025

Today nearly two million adults are incarcerated in the United States, across various correctional facilities.[[1]](#footnote-1) This number does not include the addition 3.7 million adults under community supervision, probation and parole.[[2]](#footnote-2) The justice system by its very nature is a system put into place to respond to a committed crime. The diagram below describes the sequence of events that occur in the United States Criminal Justice System.[[3]](#footnote-3) While the proactive/preventable aspect of crime can and will be debated, the purpose of this paper is to discuss recidivism and the ability to predict recidivism, which is the tendency of a convicted criminal to reoffend. In addition, the introduction of a machine learning model is presented to predict the likelihood of recidivism among offenders. By analyzing historical criminal justice data, the model identifies key risk factors and provides actionable insights for decision-makers. The model achieved a predictive accuracy of about **80%**, highlighting the most influential factors in reoffending risk and supporting data-driven interventions.

A diagram of a system

AI-generated content may be incorrect.

**Introduction**

Recidivism—the tendency of a convicted individual to reoffend—is a persistent challenge for criminal justice systems worldwide. Traditional risk assessment methods, while useful, often rely on subjective judgment and limited data, leading to inconsistent outcomes. Advances in artificial intelligence (AI) offer the potential to improve predictive accuracy, optimize resource allocation, and support more effective rehabilitation strategies.

However, the application of AI in this context carries significant ethical, legal, and social considerations. Models can unintentionally perpetuate systemic biases, raise privacy concerns, and lack transparency in decision-making. This whitepaper explores the use of AI to predict recidivism, examines existing implementations, identifies key challenges, and presents a prototype that can be leveraged for further study. By combining predictive technology with human oversight and robust governance, AI can be leveraged to enhance justice outcomes while minimizing unintended harm.

**Current Tools**

Widely used tools in industry today are summarized in the following table. Existing tools reveal a trade-off between accuracy, interpretability, and fairness. Proprietary models often achieve operational efficiency at the cost of transparency, while open or interview-based systems enhance interpretability but remain vulnerable to human bias and calibration drift. Emerging ML approaches offer flexibility but must incorporate explainability mechanisms and bias auditing to be ethically and practically viable in justice contexts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tool** | **Focus Area** | **Key Features** | **Strengths** | **Limitations** |
| **COMPAS** | General recidivism | Combines demographic, criminal history, and attitudinal data into a composite score | Widely used; can integrate multiple risk factors | Proprietary; opaque; potential racial bias; limited external audit |
| **PSA** | Pretrial general recidivism | Uses nine static predictors (e.g., age, prior arrests) for pretrial risk | Transparent; easy to deploy; validated in multiple jurisdictions | Limited to static factors; may underfit local contexts; needs recalibration |
| **LSI-R** | General recidivism & criminogenic needs | Combines static and dynamic risk factors; supports treatment planning | Links risk to interventions; validated across populations | Requires trained raters; inter-rater variability; subjectivity in scoring |
| **ORAS** | General recidivism | Multi-phase risk assessment with static and dynamic factors; pretrial to reentry | Integrates multiple justice phases; supports supervision planning | Jurisdiction-specific calibration; needs ongoing training and recalibration |
| **Static-99 / Static-99R** | Sexual recidivism | Static, historical factors (age, prior sex offenses); produces categorical risk strata | Simple, validated for sexual offenders; reproducible | Narrow scope; excludes dynamic factors; limited to adult males |
| **START** | Violence, mental health, suicide, victimization | Assesses dynamic risk and protective factors in clinical settings | Useful for treatment planning; dynamic assessment | Requires professional judgment; may lack standardization |
| **SORAG** | Sexual recidivism | 14 static items including age, prior offenses, psychopathy | Brief; effective for static sexual risk assessment | Cannot target treatment; ignores dynamic changes |
| **RRASOR** | Sexual recidivism | 4-item screening tool (prior sex offenses, victim characteristics, age) | Quick to administer; evidence-based | Limited to static factors; does not capture dynamic risk |
| **STABLE-2007** | Sexual recidivism | 13 stable dynamic risk factors | Focuses on changeable traits; aids treatment planning | Requires professional judgment; may lack standardization |
| **ACUTE-2007** | Sexual recidivism | Acute dynamic factors; used alongside STABLE-2007 | Captures short-term changes; monitors offender behavior | Requires professional judgment; not standardized |
| **SARATSO** | Sexual recidivism (California) | Combines STABLE-2007 and ACUTE-2007; used by certified treatment providers | Evidence-based; integrates dynamic risk | Limited to California; requires certified professionals |
| **PATTERN** | General recidivism (federal) | Developed under First Step Act; dynamic reassessment opportunities | Incentivizes risk reduction; validated for federal prisoners | Limited to federal system; jurisdiction-specific |

* The **COMPAS** system, developed by Equivant, is among the most widely used recidivism risk assessment tools in the United States. It combines demographic, criminal history, and attitudinal data to produce composite risk scores. However, its proprietary design obscures the underlying algorithms and feature weights, raising persistent concerns about transparency and accountability. ProPublica’s 2016 analysis found that Black defendants were more likely than White defendants to be incorrectly labeled as high risk (Angwin et al., 2016), though later critiques questioned the statistical basis of these claims (Rudin et al., 2020). Nonetheless, the debate illustrates the difficulty of assessing fairness in opaque models. By embedding historical inequities within its inputs and lacking routine recalibration or bias audits, COMPAS risks perpetuating structural discrimination. Moreover, risk scores may be misinterpreted as objective measures of “dangerousness,” influencing judicial discretion in punitive ways (Dressel & Farid, 2018).
* The **Public Safety Assessment (PSA)** is a transparent pretrial tool that uses nine publicly documented factors to predict failure-to-appear and new criminal activity risks (Stevenson, 2018). Its open methodology improves accountability compared to proprietary systems. However, PSA’s fixed factor set may underfit local contexts and requires periodic recalibration to maintain validity across jurisdictions (Lowenkamp et al., 2019). Despite reasonable accuracy, subgroup disparities can persist, and overreliance on static predictors limits the tool’s ability to capture dynamic behavioral change (Austin, 2017).
* The **LSI-R** assesses both recidivism risk and criminogenic needs using structured interviews that include dynamic factors such as employment and peer associations (Andrews & Bonta, 2010). Its dual focus supports individualized rehabilitation planning. However, it relies on human raters which introduces subjectivity and reduces objectivity (Kroner & Mills, 2018). The instrument’s validity may also decline across different populations without contextual recalibration. (Skeem & Lowenkamp, 2016).
* The **Ohio Risk Assessment System (ORAS)** was developed to evaluate offender risk across multiple criminal justice phases, integrating both static and dynamic factors (Latessa et al., 2010). Its design supports consistent assessment from pretrial through reentry. However, because ORAS is calibrated to Ohio’s offender population, performance often declines when applied in other jurisdictions without revalidation (Desmarais & Lowder, 2019). The tool also depends on consistent training and scoring fidelity; without ongoing recalibration, predictive accuracy may degrade over time.
* The **START** instrument targets dynamic risk and protective factors related to violence, mental health, suicide, and victimization, primarily in clinical and community settings. Its professional judgment framework supports individualized treatment planning but limits standardization across assessors.
* The **Static-99** and **Static-99R** are actuarial instruments that estimate sexual recidivism risk among adult male offenders using only static historical factors such as age and prior sex offenses (Hanson & Thornton, 2000). Their simplicity and strong domain-specific validation are key strengths. Yet, their limited scope excludes dynamic change and renders them unsuitable for general recidivism prediction or diverse populations (Phenix et al., 2016). The categorical risk outputs may also obscure nuanced differences in probability or treatment progress.
* Specialized tools for sexual recidivism include **SORAG**, **RRASOR**, **STABLE-2007**, **ACUTE-2007**, and **SARATSO**. **Static-99** and **RRASOR** rely solely on static historical factors such as age and prior offenses, providing simplicity and reproducibility but limited adaptability to dynamic changes or interventions (Hanson & Thornton, 2000; Phenix et al., 2016). **SORAG** offers a brief actuarial assessment of sexual recidivism risk using 14 static items, while **STABLE-2007** and **ACUTE-2007** focus on dynamic stable and acute factors, respectively, supporting treatment monitoring and behavioral change tracking (Hanson et al., 2007). **SARATSO**, implemented in California, integrates STABLE-2007 and ACUTE-2007 within certified treatment programs, emphasizing evidence-based, dynamic assessment but remaining jurisdiction-specific.
* **PATTERN**, developed under the First Step Act for federal prisoners, evaluates general recidivism risk while offering inmates opportunities to reduce risk scores through interventions. Although PATTERN incorporates dynamic reassessment, its applicability is limited to the federal system.

**AI Approaches**

Artificial intelligence, particularly machine learning (ML), provides a systematic method for predicting recidivism by identifying patterns in historical data. Various AI approaches can be applied, each with its strengths and limitations.



Common Machine Learning Models

* Decision Trees / Random Forests: Non-linear models that can handle complex datasets and interactions between features. Random forests reduce overfitting by averaging multiple trees.
* Logistic Regression: A simple, interpretable model that estimates the probability of recidivism based on weighted features. Useful for transparency but may not capture complex interactions.
* Ensemble Methods: Combine multiple models (e.g., gradient boosting) to improve predictive accuracy while managing bias and variance.
* Support Vector Machines (SVM) are supervised max-margin models with associated learning algorithms that analyze data for classification, regression, and outliers detection.
* K Nearest Neighbor (KNN) The k-nearest neighbors (KNN) algorithm is a supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While the KNN algorithm can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.
* Neural Networks: Powerful for detecting intricate patterns in data but often function as “black boxes,” making explanations difficult.

**Methodology**

The purpose of this paper is to present a prototype that uses different AI models to predict recidivism in justice involved adults. The tools used for this prototype include:

* Jupyter Notebook
* Python 3 (iPyKernel)- iPyKernel enables Python code execution
* GitHub

The data sets used for this prototype:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Description** | **AI Algorithms Used** |
| ProPublica data- 2016 | compas.db - a sqlite3 database containing criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County. Contained in a GitHub repository: https://github.com/propublica/compas-analysis/blob/master/README | Random Forest    XGBoost |

**Dataset :**

The ProPublica data is from 2016 (true?). The data definition looks like the following:

id,name,first,last,compas\_screening\_date,sex,dob,age,age\_cat,race,juv\_fel\_count,decile\_score,juv\_misd\_count,juv\_other\_count,priors\_count,days\_b\_screening\_arrest,c\_jail\_in,c\_jail\_out,c\_case\_number,c\_offense\_date,c\_arrest\_date,c\_days\_from\_compas,c\_charge\_degree,c\_charge\_desc,is\_recid,num\_r\_cases,r\_case\_number,r\_charge\_degree,r\_days\_from\_arrest,r\_offense\_date,r\_charge\_desc,r\_jail\_in,r\_jail\_out,is\_violent\_recid,num\_vr\_cases,vr\_case\_number,vr\_charge\_degree,vr\_offense\_date,vr\_charge\_desc,v\_type\_of\_assessment,v\_decile\_score,v\_score\_text,v\_screening\_date,type\_of\_assessment,decile\_score,score\_text,screening\_date

However, the data is missing elements and contains many blank fields, as shown below.

**Results**

Model accuracy values:

**Precision:** The ratio of true positive predictions to all positive predictions, indicating how often the model is correct when it predicts positive.

**Recall:** The ratio of true positive predictions to the total number of actual positive cases, measuring the model's ability to find all positive instances.

**F1 Score:** A balanced measure of precision and recall.

**Accuracy:** The ratio of correct predictions to the total number of predictions.

**Model #1**

|  |  |
| --- | --- |
| **Model Name** | **Random Forest- file name: working\_plots\_top20.ipynb ( at: C:\Users\monicastebbins\working\_juptyers)** |
| **Accuracy** | Accuracy: 1.0  ROC-AUC: 1.0  precision recall f1-score support  0.0 1.00 1.00 1.00 1469  1.0 1.00 1.00 1.00 739  accuracy 1.00 2208  macro avg 1.00 1.00 1.00 2208  weighted avg 1.00 1.00 1.00 2208 |
| **High risk individuals** |  |
| importance of each feature as it pertains to the model and calculated scores | A graph with blue bars  AI-generated content may be incorrect. |
| Beeswarm | **A graph of a graph  AI-generated content may be incorrect.** |
| Comments |  |

**Model#2**

|  |  |
| --- | --- |
| **Model Name** | **XGBoost-file name: XGB.ipynb ( at: C:\Users\monicastebbins\)** |
| **Accuracy** | Accuracy: 1.0  ROC-AUC: 0.9999999999999999  precision recall f1-score support  0.0 1.00 1.00 1.00 1467  1.0 1.00 1.00 1.00 741  accuracy 1.00 2208  macro avg 1.00 1.00 1.00 2208  weighted avg 1.00 1.00 1.00 2208 |
| **High risk individuals** |  |
| importance of each feature as it pertains to the model and calculated scores |  |
| Beeswarm |  |
| Comments |  |

**Conclusions**

In summary, while traditional recidivism tools provide structured risk assessment, their opaqueness, reliance on static features, and limited generalizability raise concerns about fairness and accountability. This study demonstrates that open-source, explainable machine learning models can enhance transparency, enabling stakeholders to understand feature contributions, detect bias, and audit predictions. By combining predictive accuracy with interpretability and regular recalibration, recidivism prediction can support evidence-based decision-making while promoting ethical and equitable outcomes in the criminal justice system.

**Recommendations**

This paper and the prototype presented here is meant to illustrate the potential benefits of having AI supplement the decision making process for recidivism. It is not an exhaustive study into AI models nor is this working prototype meant to be utilized in a production environment setting. It is meant, however, to continue the discussion on how AI can facilitate recidivism predictability in order to supplement findings and diagnosis.

Expanding this work to vary the sample data size (increase/decrease), as well as to investigate the introduction of new AI models while monitoring results, is recommended.

In addition to leveraging the work done for this study, the topic of governance should be discussed. Governance models are frameworks that define how organizations make decisions, assign roles, and ensure accountability. A governance model for using AI in recidivism predictions is discussed in the paper titled ‘*Governance for AI in Recidivism Predictions’*.

**Appendix**

The code discussed in the paper is contained within a GitHub repository, located here: [monicajurs/recidivism-ai-project](https://github.com/monicajurs/recidivism-ai-project/tree/main)

**References**

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.* ProPublica.

Rudin, C., Wang, C., & Coker, B. (2020). *The age of secrecy and unfairness in recidivism prediction.* *Harvard Data Science Review.*

Dressel, J., & Farid, H. (2018). *The accuracy, fairness, and limits of predicting recidivism.* *Science Advances, 4*(1), eaao5580.

Stevenson, M. (2018). *Assessing risk assessment in action.* *Minnesota Law Review, 103*(1), 303–384.

Austin, J. (2017). *Reevaluating pretrial risk assessment in the age of transparency.* *Federal Sentencing Reporter, 29*(2).

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Routledge.

Kroner, D. G., & Mills, J. F. (2018). *The reliability and validity of the Level of Service Inventory–Revised: A meta-analytic review.* *Assessment, 25*(1), 1–13.

Skeem, J., & Lowenkamp, C. (2016). *Risk, race, and recidivism: Predictive bias and disparate impact.* *Criminology, 54*(4), 680–712.

Latessa, E. J., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. T. (2010). *Creation and validation of the Ohio Risk Assessment System: Final report.* University of Cincinnati.  
Desmarais, S. L., & Lowder, E. M. (2019). *Pretrial risk assessment tools: A primer for judges, prosecutors, and defense attorneys.* *National Center for State Courts.*

Hanson, R. K., & Thornton, D. (2000). *Improving risk assessments for sex offenders: A comparison of three actuarial scales.* *Law and Human Behavior, 24*(1), 119–136.

Phenix, A., Helmus, L., & Hanson, R. K. (2016). *Static-99R coding rules: Revised 2016.* Public Safety Canada.

Lowenkamp, C. T., Lemke, R., & Latessa, E. (2019). *The development and validation of the Public Safety Assessment.* *Criminal Justice and Behavior, 46*(2), 227–244.

Equivant (formerly Northpointe). (2015–2023). *COMPAS practitioner documentation & product pages (COMPAS Core / COMPAS-R).* EPIC / Equivant public materials. [Epic+1](https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf?utm_source=chatgpt.com)

Laura and John Arnold Foundation / PSA project. (n.d.). *Public Safety Assessment (PSA) — factors, research & validation.* Advancing Pretrial Policy & Research.

Andrews, D. A., & Bonta, J. (1995; documentation). *Level of Service Inventory—Revised (LSI-R).* Tool description and validation summaries (multiple empirical studies aggregated). U.S. Courts / RMA summaries.

Public Safety Canada (and related gov't documents). (2022). *Static-99 / Static-99R overview and predictive validity summaries.*

1. Bureau of Justice Statistics. [Correctional Populations in the United States, 2023 – Statistical Tables | Bureau of Justice Statistics](https://bjs.ojp.gov/library/publications/correctional-populations-united-states-2023-statistical-tables), September 2025. [↑](#footnote-ref-1)
2. [Probation and Parole in the United States, 2022](https://bjs.ojp.gov/document/ppus22.pdf) [↑](#footnote-ref-2)
3. [The Justice System | Bureau of Justice Statistics](https://bjs.ojp.gov/justice-system) [↑](#footnote-ref-3)