

# A Comparative Multi-State Analysis of Criminal Recidivism Risk Factors

Solomon Flax, Trenor Hamilton, Stella Kim, Seung Min Oh and Aldila Yunus

July 27, 2025

Georgia Institute of Technology  
ISYE 6414 - Regression Analysis

## Abstract

*Recidivism remains a persistent challenge for correctional systems worldwide, with a reported 82% of people released from U.S. state prisons rearrested within ten years. This study employed a multi-state regression modeling approach to identify the significant risk factors associated with recidivism and develop predictive models using data from Georgia and Wisconsin between 2013-2015, supplemented with U.S. census microdata to introduce socioeconomic proxy variables. Our methodology compared individual state models with combined multi-state models using various regression and variable selection techniques. The Georgia model achieved strong predictive performance (AUC = 0.777), while Wisconsin performed more modestly (AUC = 0.684). Key findings revealed that employment stability (80-100% days employed) served as a protective factor, while job instability increased recidivism risk. Community-level variables, particularly county gender composition, emerged as significant predictors, highlighting the importance of contextual factors. The differential performance between states demonstrates the critical importance of multi-jurisdictional validation, as single-state studies may produce misleading conclusions about model effectiveness and generalizability. Our findings suggest that recidivism reduction strategies should emphasize employment stability over solely job placement and invest in evidence-based programming during incarceration. This research contributes to literature advocating for diverse, multi-dimensional datasets in criminal justice prediction modeling.*

## Introduction

In 2021, the Bureau of Justice Statistics of the US Department of Justice published a report indicating that 82% of people released from state prisons in 2008 were arrested at least once in the 10 years following their release (Antenangeli & Durose, 2021). The tendency for people to commit crime even after experiencing imprisonment isn't unique to the United States; societies have long dealt with the challenge of re-offense by former convicted persons, as recidivism remains a persistent problem for correctional systems around the world and throughout history. Derived from the Latin *recidivus* (from *recidere*, meaning 'to fall back'), recidivism is generally understood as the reoccurrence of criminal behavior by a previously incarcerated individual (Getoš Kalac & Feuerbach, 2023; Eaglin, 2017).

### Early Research

Before the late 19th century, criminal justice systems lacked methodologies for understanding repeat offending patterns (Garland, 1991; Beirne, 1987). The field of criminology – largely developed due to Cesare Beccaria's foundational work critiquing the effectiveness of capital punishment – studied rational choice and deterrence of aberrant behavior but provided no empirical approaches for measuring the effectiveness of sanctions or other methods for

preventing future crime (Beirne, 1987; Beccaria & Voltaire, 1872). During this period, criminal justice operated under what Garland (1985) describes as a "classical framework" where judicial punishment was viewed primarily through moral and legal lenses rather than as a sociological phenomenon subject to scientific investigation. The absence of systematic record-keeping of criminal activities meant that any understanding of re-offense remained anecdotal and localized, staving off the development of evidence-based approaches to crime prevention (Rafter, 1992).

### Evolution & Prior Art

Over the next century, research on recidivism has followed three distinct methodological trajectories, each developed based on different assumptions about the nature of criminal behavior and the application of methods for its study (Grove et al., 2000).

**Clinical Assessment.** Early attempts at addressing recidivism – such is the case within Beccaria's time – relied on the professional judgment of trained clinicians, emphasizing expertise in evaluating individual risk factors (Quinsey et al., 1998). Although these approaches acknowledged the complexity of human behavior, they presented significant limitations. Meta-analyses demonstrated that clinical predictions were less accurate than quantitative methods, revealing problems with reliability, validity, and susceptibility to cognitive biases that undermined the credibility of

this approach (Grove et al., 2000).

**Actuarial Models.** In response to clinical limitations, researchers began to employ statistical methods to weight empirically-derived risk factors based on their observed relationships with recidivism outcomes (Gendreau et al., 1996; Silver & Miller, 2002). This approach was pioneered by Ernest Burgess (1931), whose work analyzed 3,000 paroled individuals in Illinois. Burgess produced one of the first instruments that converted categorical factors – such as marital status and previous criminal record – into numerical scores for quantitative assessment (Silver & Miller, 2002). Over time, more refined statistical models like LSI-R and VRAG made marginal improvements and introduced statistical metrics like AUC (Kröner et al., 2007; Fass et al., 2008; Bonta & Andrews, 2007). These tools demonstrated substantially improved accuracy, with AUC values typically ranging from 0.60 to 0.74, and provided standardized, replicable procedures (Hanson & Morton-Bourgon, 2009). Despite these significant improvements, actuarial methods showed reduced accuracy when applied across different populations, could not account for dynamic changes over time, and faced criticism for potentially perpetuating systemic biases through over-emphasizing specific demographic features without further consideration beyond statistical analyses (Brennan et al., 2009; Skeem & Lowenkamp, 2016).

**Machine Learning Algorithms.** In recent years, many researchers have experimented with machine learning (ML) techniques, leveraging computational advances to develop more sophisticated models for recidivism risk assessment (Tollenaar & van der Heijden, 2013; Berk & Bleich, 2014; Travaini et al., 2022). These models are capable of handling large sets of high-dimensional data, identifying non-linear relationships between risk factors, and have the potential to achieve improved accuracy. Some studies report ML models reaching AUC values typically ranging between 0.70 to 0.78 (Tollenaar & van der Heijden, 2013; Travaini et al., 2022; Liu et al., 2011; Karimi-Haghighi & Castillo, 2021). Despite these advancements, ML approaches still face challenges. For example, the "black box" nature of some models makes it difficult for legal practitioners and defendants to understand how sentencing decisions are made, raising concerns about due process and fair treatment of suspected persons (Doshi-Velez & Kim, 2017; Biddle, 2022).

## Recidivism in the United States

In the aftermath of the Civil War, the United States faced immense public pressure for prison reform

**Table 1:** Comparison of AUC values

Literature Source	Model AUC
Kröner et al. 2007; Fass et al. 2008; Bonta & Andrews 2010	0.60 - 0.74
Tollenaar & van der Heijden, 2012; Karimi-Haghighi & Castillo, 2021	0.70 - 0.78
COMPAS	0.65 - 0.80

(Keve, 1995). Due to converging social, economic, and political pressures that demanded accountability in the criminal justice system as whole, the systematic study of recidivism accelerated in tandem (Feeley & Simon, 1992).

**Strain on institutional resources.** Prison overcrowding, inhumane imprisonment conditions, and escalating costs created an urgent need for prison administrators to demonstrate the effectiveness of their institutions and invest in reform (Keve, 1995).

**Data-driven methods for social science use cases.** The rise of scientific management principles in the public sector and governmental activities, as documented by efficiency movement scholars, demanded data-driven approaches to evaluating institutions that extended into correctional settings (Sutton, 1987). Additionally, experts in the field of criminology wanted to establish scientific legitimacy through developing quantifiable methodologies, with recidivism providing a binary observable outcome variable for testing theories about criminal behavior (Beirne, 1987; Rafter, 1992).

**Need for standardized, codified practices.** The establishment of organizations like the National Prison Association, the National Conference of Charities and Correction, and the National Probation Association spurred research into comparing corrections approaches (e.g. medical, education, disciplinary) using data and identifying which were effective (Sutton, 1987).

## COMPAS: A Case Study In Algorithmic Risk Assessment

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system represents one of the most widely deployed and controversial algorithmic risk assessment tools in contemporary criminal justice (Brennan et al., 2009; Dressel & Farid, 2018). Developed by Northpointe (now Equivant), COMPAS uses a proprietary algorithm to assess defendants' likelihood of recidivism based on a 137-item questionnaire covering criminal history, social environment, and behavioral patterns.

**Merits.** As indicated in Table 1, COMPAS demonstrated relatively high predictive accuracy comparable to other validated instruments, with reported AUC values ranging from 0.65 to as high as 0.80 (Brennan et al., 2009). The system provided a standardized approach to risk assessment, generating risk scores that could be implemented consistently across jurisdictions and integrated into existing judicial workflows.

**Shortcomings.** In 2016, ProPublica published results from their investigation that revealed significant racial disparities in COMPAS predictions, finding that black defendants were almost twice as likely to be incorrectly flagged as high-risk compared to white defendants (Flores et al., 2017). This analysis sparked intense public debate about algorithmic fairness and highlighted how seemingly race-neutral algorithms can produce systematically different outcomes for demographic groups (Corbett-Davies et al., 2017; Chouldechova, 2017). Additionally, the proprietary nature of the algorithm limited transparency and made it difficult for defendants to understand or challenge their risk scores (Wexler, 2017).

## Ethical Considerations

As demonstrated through the existing body of research into recidivism, predictive modeling in criminal justice raises profound ethical concerns across multiple dimensions.

**Fairness and Bias.** The primary concern involves algorithmic bias and unfair treatment of demographic groups. Predictive models may perpetuate historical discrimination patterns, and legitimate predictive factors often correlate with demographic variables due to social inequality (Han et al., 2025; Scurich & Monahan, 2016). This challenge is further complicated by conflicting definitions of fairness (Foulds et al., 2020). Achieving one form of fairness may require accepting inequality along other dimensions (Chouldechova, 2017; Kleinberg et al., 2017).

**Privacy and Data Use.** Methods involving the ingestion and analysis of sensitive personal information is inherent in recidivism research, raising questions about consent and appropriate data usage, particularly as governing bodies across the world are beginning to enact AI-related policies (van Dijck, 2022). Individuals in the criminal justice system have limited consent capacity, and data linkage across systems introduces the risk of privacy violation (Ferguson, 2017).

**Transparency and Accountability.** Complex AI and ML models create transparency concerns around judicial proceedings. When defendants – nor those

in the justice system relying on risk assessment programs – cannot understand algorithmic predictions, it becomes difficult to challenge decisions, particularly concerning fundamental liberty and supervision decisions (Wexler, 2017).

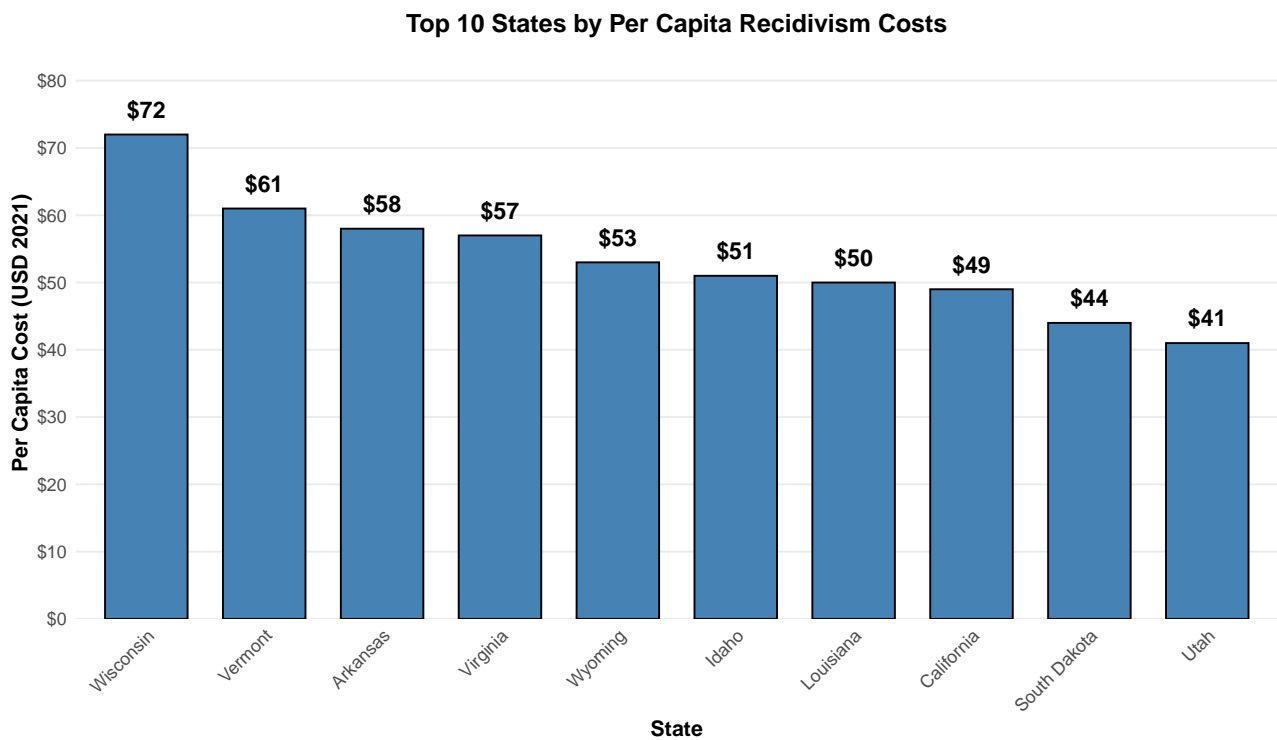
## Research Motivation

The implications of recidivism extend far beyond academic exercise, as its costs carry impacts on both the economic and societal levels. In 2015, FWD.us published a report stating that U.S. taxpayers spent \$273 billion annually on the criminal justice system, and an estimated \$87 billion in annual GDP is lost due to the effects of over-criminalization (deVunopowell et al., 2015). Moreover, the Council of State Governments Justice Center published research in 2023 estimating the costs of recidivism at the state level [Appendix B]; Figure 1 highlights the top 10 states with the highest reported cost of recidivism per capita (The Council of State Governments Justice Center, 2023). From an anthropological perspective, high recidivism rates perpetuate cycles of incarceration that disproportionately impact marginalized populations (Martensen, 2012).

Continuing research into effective recidivism prediction that addresses the shortcomings of contemporary models is imperative for securing the health and well-being of our future. Successful modeling has the potential to greatly improve administrative resource allocation, judicial decision making, rehabilitation and reintegration programs, and evidence-based policy development.

When selecting a topic for our research project, our shared attitudes around social justice combined with our desire to explore this issue in greater depth led us to choose recidivism as our subject of interest. Ultimately, our goal was to identify the significant risk factors associated with higher recidivism rates and to design a model using those predictors to predict the likelihood of recidivism using historical data across multiple jurisdictions. Based on our review of the existing literature and our initial analysis plan, we conducted our research with the following expectations in mind:

- Potential challenges sourcing data suitable for our research approach (e.g., publicly accessible, individual-level, robust, combinable).
- Small likelihood of identifying novel predictors with high impact while mitigating biases.
- Producing similar accuracy values – as opposed to exceeding – compared to existing approaches with given resources and time constraints.
- Focus on applying rigorous statistical methodology and interpreting results.



Source: The Council of State Governments Justice Center (2023)

**Figure 1:** Recidivism costs \$40+ per resident in 10 states, representing the highest per capita in the US. Source: Council of State Governments Justice Center, <https://csgjusticecenter.org/publications/the-cost-of-recidivism/>.

Rather than strive for perfection – building a high-accuracy model containing no bias – we aimed to adopt a more holistic approach to modeling recidivism that balances ethics with efficacy by applying rigorous statistical analyses to diverse, multi-faceted datasets.

## Methods

To address our research objectives, we opted to build a multi-state regression model incorporating demographic, criminal history, and socioeconomic variables. Our approach combined recidivism data from Georgia and Wisconsin, supplemented with aggregated U.S. census microdata to incorporate socioeconomic proxies that the state datasets lacked. The purpose of using a multi-jurisdiction design was to address concerns about geographic biases and limited perspective that could arise when predictive models are developed using data from one location (Silver & Miller, 2002; Berk & Bleich, 2014).

Given the nature of demographic and behavioral data comprising our datasets, we anticipated the issue of multicollinearity between predictors. Thus, we tested multiple regression techniques including logistic regression, ridge regression for regularization, and various variable selection techniques (e.g., logarithmic transformation, outlier analysis) to opti-

mize predictive performance while addressing challenges with multicollinearity. For our research, we conducted both state-specific modeling to identify predictive factors unique to the jurisdiction and combined multi-state analyses to examine generalizable patterns in an effort to balance external validity with local applicability.

Throughout our analysis, we applied statistical techniques including exploratory data analysis, diagnostics (e.g., Variance Inflation Factor analysis), and compared the performance across multiple models. This multi-dimensional approach enabled us to examine recidivism prediction from both a broad comparative perspective and a targeted jurisdictional lens, while attempting to address the challenges of working with data across different policy and demographic contexts.

## Defining Recidivism

During the early stages of the research process, we quickly determined the need to establish a definition for recidivism due to a general lack of consensus across the academic field (Fazel & Wolf, 2015; Yukhnenko et al., 2023). This became even more apparent given the differences in methodology between the Georgia and Wisconsin datasets. For the purpose of our study, we established recidivism as *any conviction within two years of release between 2013-2015*.

Applying this uniform criteria across both datasets allowed for consistency and comparability of results in addition to the merge-ability of the two state-level datasets.

## Georgia Data Analysis

We sourced the Georgia dataset from the National Institute of Justice website (U.S. Department of Justice, 2021). It contains detailed, individual-level recidivism data from in-state prisons. The raw dataset contains 25,835 observations and 55 variables and includes demographic factors (e.g., gender, race, age at release), criminal history, supervision conditions, violations, employment metrics, and program participation details.

**Preprocessing.** We applied the following methods to prepare the dataset for analysis:

1. **Variable Type Conversion.** All categorical variables were converted into factors; this is necessary to be compatible with logistic regression models. There were some binary indicators that were originally stored as character strings (like “true” or “false”), so we standardized and recoded these variables as factors with levels 1 and 0, respectively.
2. **Standardization.** Race categories were renamed to align with the Wisconsin dataset. For instance, *WHITE* was recoded as *Caucasian*, and *BLACK* was recoded as *African\_American* for the combined model.
3. **Response Variable.** Since we defined recidivism as any arrest occurring within two years of release, we created the binary response variable *Recidivism\_Within\_2years* using available variables: if either *Recidivism\_Arrest\_Year1* (indicating a prisoner recidivated within the first year after release) or *Recidivism\_Arrest\_Year2* (indicating a prisoner recidivated in their second year after release) were indicated as “1”, then *Recidivism\_Within\_2years* would be “1”. We then removed *Recidivism\_Arrest\_Year1*, *Recidivism\_Arrest\_Year2*, and other similar response variables from the dataset.
3. **Feature Selection with Lasso.** We fit a Lasso regression model using the *glmnet* package. This method applies a penalty to shrink coefficients and remove weaker predictors by setting their coefficients to zero, thus performing variable selection. Our model was trained using cross-validation to determine the best penalty parameter  $\lambda$ . This step reduced the number of variables to 47 predictors correlated with recidivism.
4. **Variable Transformation.** Several predictors underwent transformations to improve model fit. Examples include:
  - (a) *Jobs\_Per\_Year*: A log transformation was applied to address heavy right-skewedness, resulting in the *log\_jobs* predictor.
  - (b) *Percent\_Days\_Employed*: Due to heavy skewing, we binned this variable into categories (low, medium, high) to represent employment stability more effectively.
5. **Model Evaluation.** We assessed predictive performance using the following metrics:
  - (a) AUC-ROC curves to measure discriminative ability.
  - (b) Confusion matrices and classification accuracy at a 0.5 probability threshold.

**Modeling.** We began the modeling process for the Georgia dataset with exploratory analysis and preprocessing, followed by logistic regression and variable selection.

1. **Training/Test Split.** The cleaned dataset was randomly split into training (80%) and testing (20%) sets.
2. **Baseline Logistic Regression.** After preprocessing the data, creating the binary outcome

*Recidivism\_Within\_2years*, and performing initial variable selection to focus on key variables, an initial logistic regression model was fit (named *reduced\_logit\_model*). This model served as a baseline for comparison. Goodness-of-fit diagnostics, including Pearson and Deviance residual checks, indicated some lack of fit and evidence of multicollinearity among predictors, most likely due to the large number of categorical variables and interactions. The multicollinearity issue was particularly notable: we discovered that all *Female* prisoners were coded as *Unknown* for the variable *Gang\_Affiliated*, while all *Male* prisoners were coded as either 1 or 0 (indicating positive or negative gang affiliation). Two models were run: one without *Gender* and one without *Gang\_Affiliated*. Since the model without *Gender* reported a lower AIC value, we ultimately removed *Gender* from the dataset.

## Wisconsin Data Analysis

We sourced the Wisconsin dataset from the Wisconsin Circuit Courts (Ash et al., 2023). Similar to the Georgia dataset, it contains detailed, individual-level recidivism data from prisons within the state. The raw dataset contains 1,048,575 observations and 54 variables, and includes demographic variables (e.g., gender, race, age at time of offense), criminal history (e.g.,

prior felony, misdemeanor, traffic charges) and socioeconomic information (e.g., percentage of Black or Hispanic populations, median household income, population density). The dataset also contains multiple recidivism response variables, specifically whether an individual re-offended within 180 days (*recid\_180d*) or two years (*is\_recid\_new*). For the purpose of our analysis, we retained the *is\_recid\_new* variable (renamed *Recidivism\_Response*) and removed the others.

**Preprocessing.** We applied the following methods to prepare the dataset for analysis:

1. **Variable Type Conversion.** All categorical variables were converted into factors to ensure that they were compatible with logistic regression models. Our response variable, along with other binary indicators, was converted from integers to factors with levels 0 and 1. Age at judgment (*age\_judge*) was binned into eight categories (e.g., "17 or younger," "18-22," "23-27") to mirror the Georgia dataset structure and allow for more direct comparisons between the two.
2. **Handling Missing Values.** The variable *jail* (jail sentence length in days) had 56.7% missing values. Observations with missing jail time were filtered out, resulting in a smaller subset of 189,263 observations.
3. **Sample Selection.** Since we were focused on predicting whether an individual recidivated or not, we removed entries where the individual's ID (*new\_id*) was repeated, indicating that they returned to court. As the observations were arranged chronologically, we retained only the first offense per individual. After removing additional rows with missing data, the remaining dataset was randomly sampled to produce a comparably-sized dataset (25,000 observations) to Georgia's. This would also improve computational efficiency, particularly when applying variable selection techniques.

**Modeling.** We began the modeling process for the Wisconsin dataset with exploratory analysis and preprocessing, followed by logistic regression and variable selection.

1. **Training/Test Split.** The cleaned dataset was randomly split into training (80%) and testing (20%) sets.
2. **Baseline Logistic Regression.** We fit an initial logistic regression model using all of the predictors. The deviance and Pearson residuals showed large deviations from normality, suggesting non-linear relationships.
3. **Variable Transformation.** After identifying skewedness and other potential model fit issues,

we applied transformations to several predictors to improve fit. Examples include:

- (a) **Log Transformations:** Skewed variables (*jail*, *median\_hist\_jail*, *med\_hhinc*, *pop\_dens*) were log-transformed to potentially improve normality.
- (b) **Categorical Recoding:** Rare categories in *wcisclass* (type of offense) were grouped into an "Other" category, retaining only the top five most frequent offenses.

4. **Feature Selection with Lasso.** We fit a Lasso regression model using the *glmnet* package to shrink coefficients and remove weaker predictors. We cross-validated with the identified optimal penalty parameter  $\lambda$ , and retained 47 variables. Several important retained predictors included demographic factors (*gender*, *race*, age bins), criminal history (*prior\_felony*, *prior\_misdemeanor*), and transformed variables (*log\_jail*, *log\_median\_hist\_jail*).
5. **Model Evaluation.** We assessed predictive performance using the following metrics:

- (a) AUC-ROC curves to measure discriminative ability.
- (b) Confusion matrices and classification accuracy at a 0.5 probability threshold.

## IPUMS Data

IPUMS USA – part of the broader IPUMS Center for Data Integration consortium – collects, preserves and harmonizes U.S. census microdata (IPUMS). We incorporated this dataset into our analysis to address the limited socioeconomic variables present in the Georgia and Wisconsin datasets, introducing continuous proxy measures that could enhance variable selection and improve explanatory power and interpretability for our results. IPUMS USA offers extensive socioeconomic indicators through an accessible data extraction interface, with samples that are stratified and designed to ensure geographic, demographic, and economic representativeness despite not constituting full-census enumeration.

Our data extraction focused on census years 2013-2015 to normalize between Georgia and Wisconsin data, selecting variables including census year, state FIPS code, age, race, Hispanic origin, health insurance coverage, employment status, total personal income, and official poverty status. We filtered the sample to include only respondents from Georgia and Wisconsin, then harmonized the race and Hispanic origin variables to create a unified race classification consistent with the Georgia dataset syntax. To create meaningful aggregated measures, we calculated the median values for each socioeco-

**Table 2:** Variable Availability Across Datasets

Variables	Georgia	Wisconsin	IPUMS USA	Combined Dataset
Recidivism within 2 years of release*	X	X	X	X
Age at Release (group)	X	X	X	X
Race	X	X	X	X
Sex	X	X	X	X
Median Income			X	X
Mean Unemployment Rate			X	X
Mean Health Coverage Rate			X	X
Mean Poverty Rate			X	X
Mean Violent Crime Rate			X	X
Education Level	X	X		
Type of Offense	X	X		
Severity of Offenses	X	X		
Gender	X			
Primary Metropolitan Area (PMA)	X			
Gang Affiliation	X			
Number of Delinquency Reports	X			
# of Days employed	X			
Number of Program Attendances	X			
Age at Intake (group)		X		
Food Stamps		X		
Other Benefits		X		

Note: X indicates variable is available in the dataset.

\*Response variable

nomical variable by grouping observations into categories of year, age, state, and race. Subsequently, we binned respondents' ages into seven groups matching Georgia's age categorization and computed averages across the three-year period, age groups, state, and race combinations. This process enabled us to merge our three datasets – Georgia, Wisconsin, and the census microdata – and introduce four continuous predictive variables to supplement our analysis: *proxy\_median\_income*, *employment\_rate*, *health\_coverage\_rate*, and *poverty\_rate*. These socioeconomic proxies provided essential context for understanding the broader community conditions that may influence recidivism patterns beyond individual demographic and criminal history factors.

## Combined Analysis

**Preprocessing.** The combined dataset refers to the aggregation of individual prisoner observations from the Georgia and Wisconsin datasets for the years 2013 through 2015 as well as calculated demographic statistics from the IPUMS census sample. There were many other years of observations in the Wisconsin dataset, but our guiding philosophy was to transform the two states' datasets to as close as possible to parity. All preprocessing steps such as standardization, new Response Vari-

able definition, and factor variable setting were inherited from the preprocessing described for each individual state dataset. Furthermore, we renamed the Georgia variable *Prior\_Arrest\_Episodes\_Felony* to *Prior\_Felony* to conform to the associated Wisconsin variable. The Wisconsin dataset originally contained *Age\_at\_Intake* as opposed to *Age\_at\_Release*. We calculated *Age\_at\_Release* for all Wisconsin observations by adding their sentence length in days to their existing age information and then rounding to the nearest year. Our group recognized that the estimated *Age\_at\_Release* contained some error, but this random error is partially mitigated by converting each inmate's age to match the factor levels of Georgia inmates.

Upon joining the three datasets, we only kept the common columns from the two data sets, namely *Gender*, *Race*, *Age\_at\_Release*, *Violent\_Crime*, *Prior\_Felony*, *median\_income*, *mean\_employment\_rate*, *mean\_health\_coverage\_rate*, *mean\_poverty\_rate*, and *Prison\_Years* as indicated in Table 2. Each Georgia and Wisconsin dataset was split into 80-20 training and testing data sets, respectively. Then, the two individual training and testing datasets were combined for the joint model training and evaluation. We followed this process to ensure the same proportion of Georgia and Wisconsin data in both the training and test datasets.

The overall training and test datasets contained 39,257 and 9,814 observations, respectively.

**Modeling.** Once the datasets were combined, we grouped the training dataset by all variables into repetitions so that we could perform goodness of fit tests for the following models: *Full Logistic Regression*, *Full Probit Regression*, and *Stepwise Forward variable selection*. We also assessed the normality of residuals by plotting residuals into a QQ plot and a histogram. The initial *Full Logistic Regression* model's performance was not satisfactory, so we performed variable selection and regularization to improve the model performance. The *Full Probit Regression* model was trained in the hopes of improving goodness of fit, but any difference in residual normality was negligible. Goodness of fit tests were discontinued for further models using Regularization techniques due to difficulties implementing model design matrices on grouped data.

As shown in Table 2, the combined dataset had far fewer predictor variables to select from compared to the individual state datasets. Therefore, advanced variable selection methods such as *Lasso Regression* and *Group Lasso Regression* were abandoned in favor of *Ridge Regression* and an *Elastic Net Model* to address multicollinearity. We calculated AUC-ROC curves, and confusion matrices with a 0.5 probability threshold for *Full Logistic Regression*, *Full Probit Regression*, *Stepwise Forward variable selection*, *Ridge Regression*, and *Elastic Net*.

## Results

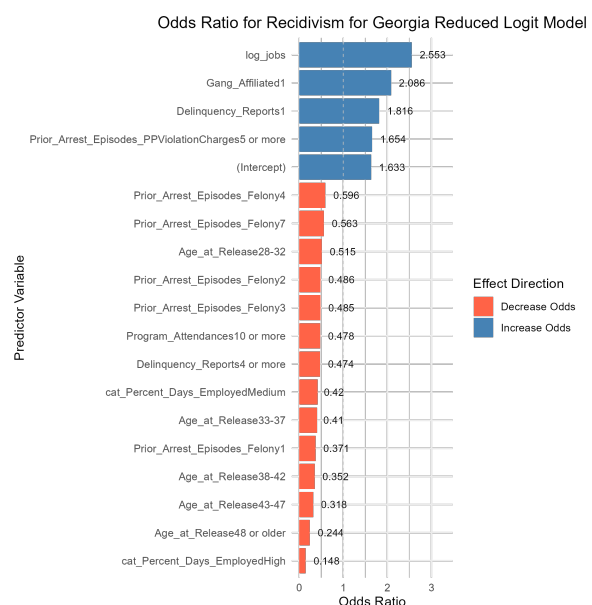
### Estimated Effects

Given that the majority of predictor variables were categorical across all datasets, each model had numerous estimated coefficients to account for the contribution to recidivism at each possible level of the factor variables. The most impactful variables on recidivism are modeled in Figure 2, Figure 3 and Figure 4. The full list of estimated coefficients for the best-performing models on each dataset are available in Appendix C (Georgia) and Appendix D (Wisconsin).

The initial list of variables from the Georgia model is filtered to interpret factors with larger impacts (absolute log-odds coefficient values greater than 0.45) which focused on 19 variables (reduced from 99 variables initially). Looking at Figure 2, we observe that the predictor variables associated with the log of number of jobs, gang affiliation, and delinquent reports and 5 or more prior arrests with probation/parole violation increase the odds for recidivism. The highest impact variable is *log\_jobs* which underwent log transformation of the predictor jobs

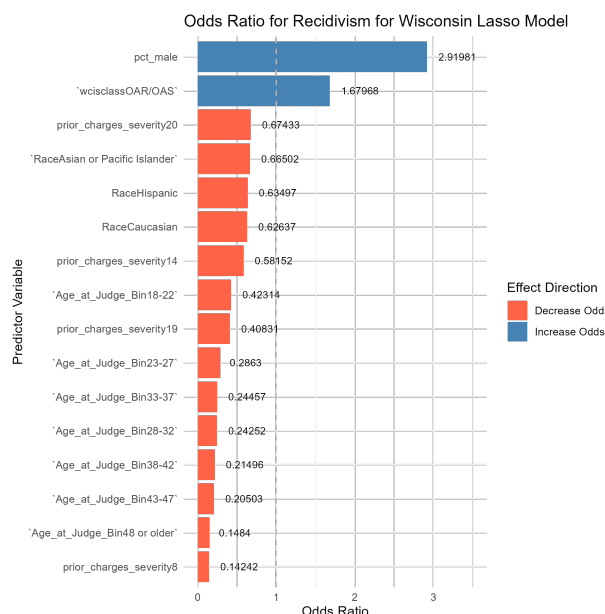
(number of jobs while on parole). Thus, one unit increase in *log\_jobs* is equivalent to an approximate increase of  $e^1 \approx 3$  jobs per year. A Georgia individual who has an increase of every 3 jobs (or one additional log (jobs)) per year while on parole increases the odds of recidivism by factor of 2.55 or 155%, holding all other variables constant. Additional gang affiliation increases the odds of recidivism by factor of 2.09 or 109% compared to a non-gang-affiliated individual, holding all other variables constant.

However, there are protective factors of recidivism shown in Figure 2. Increase in age of release strengthens the protective factor for recidivism. Program attendance of 10 or more has an odds ratio of 0.478, translating to an odds reduction of 52% for a Georgia individual having 10 or more program attendances to commit recidivism than those who have not attended a program, holding all other variables constant. The protective variable with the lowest odds ratio is the high category of percent of days employed. Variable *Percent\_Days\_Employed* was transformed/categorized into various levels to optimize model fitting. Thus, the odds of recidivism is  $(1-0.148 =)$  85% lower for individuals in the high category of percentage (80-100%) of employment days while on parole (U.S. Department of Justice, 2021) compared to an individual in the low category percentage (0 to 20%), holding all other variables constant.



**Figure 2:** Above are the strongest expected changes in odds of recidivism in either direction for the best performing Logistic Regression model using only Georgia prisoner data. Values are the expected change in odds of recidivism associated with the variable, not the change in probability.

The number of predictor variables for Wisconsin is also filtered and reduced to highlight variables

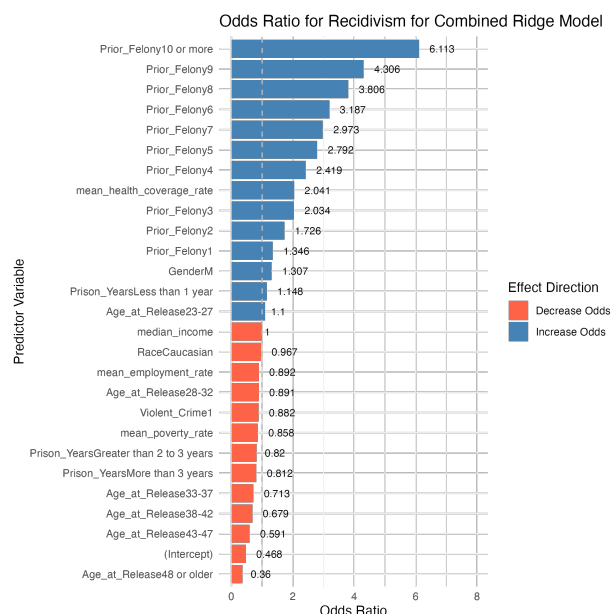


**Figure 3:** Above are the strongest expected changes in odds of recidivism in either direction for the best performing Logistic Regression model using only Wisconsin prisoner data. Values are the expected change in odds of recidivism associated with the variable, not the change in probability.

with larger impacts on recidivism (absolute log odds coefficient values greater than 0.3). We focused on the 17 most impactful variables on recidivism of 106 initial variables. The intercept was removed from the visualization due to skewing the image and thus making it difficult to interpret the other factors. Per Figure 3, we observe that the neighborhood factor *pct\_male*, is associated with the largest increases in recidivism. Wisconsin defendants who were arrested in counties with an imbalanced population with respect to gender were far more likely to recidivate as the proportion of the male population increased. Additionally, the odds of recidivism increased by 1.35 when an individual from Wisconsin was convicted of OAR/OAS. Both OAR and OAS are classified as non-violent traffic violations within the state of Wisconsin (Wisconsin Court System, 2012).

Protective factors shown on the Figure 3 for recidivism are ages 18 years and up. The odds of recidivism decreases as the age of a Wisconsin individual increases. The highest protective impact was the variable *prior\_charges\_severity8* which is described as the least severe misdemeanor offense (Misdemeanor C). For an individual in Wisconsin with a prior charge of Misdemeanor C (Wisconsin Legislative Reference Bureau, 2019), the odds of recidivism is reduced by a multiplicative factor of  $2.365926 \times 10^{-5} \approx 0.002\%$  compared to an individual who did not have a prior Misdemeanor C charge.

Figure 4 once again highlights the odds ratios as-



**Figure 4:** Above are expected changes in odds of recidivism in either direction for the best performing Logistic Regression model using the combined dataset. Unlike the prior two figures, associated change in odds is shown for all estimated coefficients.

sociated with each model coefficient for the best performing model using the combined dataset. An intercept odds ratio of 0.468 shows that the odds for recidivism at baseline: an African American female between 18-22 years of age with a non-violent crime, no prior felonies and 1-2 years of prison years has lower odds of recidivism with an odds ratio <1.

The strongest risk factor for recidivism was the number of prior felonies. The odds of an individual with 10 or more prior felony counts being re-convicted/recidivating increases by a factor of 6.11, holding all other variables constant. Moreover, the odds of recidivism increases with each additional prior felony. Prison years of less than one year had a slight increase with an odds ratio of 1.1 (10% increase) for recidivism compared to the baseline of 1-2 years of prison years, keeping all other variables constant.

A protective factor from recidivism was an increase in the age of release. For example, if the baseline individual's age increases from 18-22 years to 48 or older, the odds of recidivism decreases by 64% (1-0.36), holding all variables constant. Longer prison terms of 2 years or more is also associated with a reduction of recidivism odds of 18% (1-0.82) compared to an individual with just 1-2 years of prison. Overall, being released at ages of 27 years or more, longer prison years and violent crime classification lowers the odds of recidivism. Meanwhile, the number of felonies, prison years less than one year, and health-care coverage rate increase the odds of recidivism.

Model	Dataset	Accuracy	Sensitivity	Specificity	Precision	AUC
reduced_logit_model	Georgia	0.711	0.688	0.731	0.697	0.784
<b>reduced_logit_model_new</b>	<b>Georgia</b>	0.705	0.692	0.717	0.687	0.777
final_model	Georgia	0.659	0.627	0.688	0.644	0.719
initial model	Wisconsin	0.685	0.313	0.889	0.609	0.683
<b>after transform and var selection</b>	<b>Wisconsin</b>	0.681	0.301	0.890	0.600	0.684
Elastic Net	Combined	0.657	0.447	0.801	0.606	0.691
Full Logistic	Combined	0.657	0.449	0.799	0.605	0.691
Full Probit	Combined	0.657	0.444	0.803	0.606	0.690
<b>Ridge</b>	<b>Combined</b>	0.657	0.435	0.809	0.609	0.691
Step Forward	Combined	0.657	0.455	0.795	0.603	0.690

**Table 3:** Calculated prediction metrics for each model and dataset after rounding to the third decimal point. Model and Dataset names that are jointly in bold indicate the chosen best model for a given dataset.

## Prediction Performance

To evaluate each model's predictive power, we calculated the following prediction metrics using entries of a Confusion Matrix: Accuracy, Sensitivity, Specificity, and Precision. For the following formulas let the acronyms TP, TN, FP, and FN take their typical meanings True Positives, True Negatives, False Positives, and False Negatives, respectively. With this notation defined, we calculate model Accuracy as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensitivity is calculated as

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \quad (2)$$

Model Specificity is calculated with the following formula

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (3)$$

Finally, Precision is given by

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (4)$$

These metrics as calculated with a 0.5 probability threshold for predictions are reported in Table 3 along with AUC. AUC, short for area under curve, refers to the shape of the receiver operator characteristic, a common tool to capture the tradeoff between model Sensitivity and Specificity while altering probability thresholds for predictions (Robin et al., 2011). To choose the best model for each dataset, as indicated in Table 3 by bolding of the model and dataset names, we prioritized decreased model complexity followed by higher Precision scores.

## Explanation of Changes

**Addition of IPUMS USA data.** During the development of the analysis plan, our team identified several

state-level datasets consisting primarily of individual-level demographic and criminal history data across multiple US jurisdictions. Our original approach intended to combine three datasets from different states using common predictive variables present across all datasets. However, after submitting the analysis plan and examining the data further, we determined that the shared variables between the Georgia and Wisconsin datasets were limited to only five measures: gender, race, age at release group, violent crime indicator, and prior felony history. This constraint presented two significant methodological challenges: first, the available variables were exclusively demographic and crime-related categorical indicators, leaving us without continuous predictive variables; second, the limited number of variables was insufficient for robust variable selection procedures.

To address these challenges, we incorporated aggregated census microdata to introduce continuous socioeconomic proxy variables, including median income by age group and poverty rates. This approach was further motivated by the Georgia dataset's age group structure, which necessitated creating aggregate socioeconomic variables by age cohorts even if individual-level data had been available, particularly for harmonization with the Wisconsin dataset. By integrating IPUMS USA census microdata as our third dataset and combining it with the state-level recidivism data, we successfully introduced several continuous predictive variables suitable for variable selection algorithms and model inclusion, thereby enhancing our ability to apply comparative analytics.

**Individual state analysis.** Given our team's shared interest in recidivism as both an intellectually compelling and socially relevant research topic, we elected to conduct jurisdiction-specific analyses in addition to our combined modeling approach. This decision enabled us to identify state-specific predic-

tive variables and examine their unique contributions to recidivism prediction within each unique context, while also allowing for comparative analysis of predictive factors between Georgia and Wisconsin.

**Multicollinearity and ridge regression.** We employed the *model.matrix* function in R for data transformation and conducted comprehensive multicollinearity diagnostics using Variance Inflation Factor (VIF) analysis, correlation matrices, and heatmap visualizations across all three datasets. Our analysis revealed significant multicollinearity among certain variable pairs, including gender and gang affiliation, as well as race and age at offense. These findings led us to implement ridge regression as a regularization technique to address multicollinearity through coefficient shrinkage. For the combined dataset, we incorporated multiple variable selection methods and regularization approaches, subsequently comparing their predictive performance to identify optimal modeling strategies.

## Discussion

The primary objectives for this research were to identify significant predictors associated with recidivism and subsequently use those variables to build an accurate predictive model. By combining data from two states – Wisconsin and Georgia – we attempted to reduce biases that could arise from population heterogeneity, limited representativeness, policy variations, or other factors related to studying a single geography (Silver & Miller, 2002; Berk & Bleich, 2014). The addition of U.S. census microdata addressed a critical limitation in state datasets by introducing socioeconomic proxy variables — including median income, employment rates, health coverage, and poverty rates — that are theoretically important predictors of recidivism but typically absent from state correctional records (Cullen & Gendreau, 2000).

Through applying variable selection and regularization techniques, we sought to balance model complexity with interpretability while addressing multicollinearity concerns identified in our exploratory analysis. Our analytical framework compared individual state models with combined multi-state models to assess whether cross-jurisdictional data integration enhances predictive performance and produces more robust, generalizable findings. We believe this approach reflects current best practices in criminology research that emphasize external validity and model transportability across different institutional and geographic contexts and recommend future recidivism research incorporates these considerations (Myers & Spraitz, 2011; Farrington & Welsh, 2005).

## Successes

Findings across Georgia, Wisconsin and combined dataset models uncovered possible insights for reducing recidivism. In Georgia, employment stability was a key protective factor. While a higher number of jobs held was associated with increased odds of recidivism, a large percentage of days employed (80-100%) was associated with lower recidivism. These results suggest that job placement programs which align individuals with their interests may be more impactful in maintaining continuous employment and reducing recidivism than simply navigating access to job openings. This finding is particularly salient given how well logistic models trained on Georgia prisoner data performed on test data. Not only did our best Georgia model perform all others in our study, but with an AUC of 0.777 it competes with the upper echelon of predictive models in the known literature as indicated in Table 1.

The differential performance between the two state-level datasets provides compelling evidence to support our hypothesis that diverse, multi-dimensional datasets are essential for recidivism research. Had we conducted our study using only Wisconsin data, we might have concluded that our approach was ineffective, potentially resulting in misleading implications for policy development. Conversely, relying solely on the Georgia data might have led to overconfident conclusions about model performance that wouldn't generalize effectively. A key difference between the Wisconsin dataset and the Georgia data is the inclusion of community-level data for the county an offense was committed. These community variables inspired our inclusion of IPUMS for the joint dataset. Moreover, the county statistic *pct\_male* proved to have the strongest influence on recidivism rates. This could indicate the need to revise general recidivism data collection methodologies to include more granular collection activity on socioeconomic and demographic factors that may influence repeat criminal behavior.

In the combined model, longer jail stays were associated with reduced recidivism. This could be possibly due to effective rehabilitation or education programs within the jail systems. There may be a potential value in investing and scaling these programs to reduce recidivism and serve as rationale for budget/resource allocations and system reforms for future policies. Across all three models, individuals of older ages were less likely to recidivate. Although this is a demographic variable that is not intervene-able, it helps us understand underlying risk factors. Another key success of the combined model was the reduction of racial disparity in false positive outcomes after using Ridge Regression. At a

prediction probability threshold of 0.5, the Precision measure for Caucasian and African-American subsets of test data was within 0.01. We were particularly motivated to close this racial gap given prior controversy on this front surrounding the groundbreaking COMPAS system.

## Limitations

This investigation was greatly challenged by the difference in data collection methodology between states. The Georgia dataset, as a result of publication by the Department of Justice, contained detailed information about each individual's activities while they were incarcerated. Unfortunately, the researchers who published Wisconsin criminal data were not privy to the same level of individual specific information. Despite our best efforts to address multicollinearity through regularization methods, we believe that the effect of some variables are still obscured by correlation of predictors. For example, the Georgia prison system collects drug test information on every type of drug individually, meaning there is no catch-all variable for the dataset for prisoners who successfully rehabilitate from drug use. Thus, the impact of drug use could be dispersed across many categorical variables. One area that regrettably remains unexplored in this study is the inclusion of interaction terms between variables in predictive modeling of recidivism.

The perceptibility of our models trained on only Wisconsin data, as indicated by AUC, and to a lesser extent the combined dataset model, was noticeably outclassed by the models with only Georgia training data. One potential explanation for this deficiency is the overabundance of traffic-related criminal records in Wisconsin's training data. 40% of all observational records for Wisconsin had a *case\_type* of Criminal Traffic. We hypothesize that the population of repeat traffic offenders and repeat offenders for other criminal activities may be fundamentally different or separable by some as of yet unknown qualities.

Finally, a two-year follow-up period, while consistent with much of the recidivism literature, may not capture longer-term patterns or delayed recidivism that could provide different insights into risk factors and their temporal relationships. Longitudinal studies with extended follow-up periods would greatly enhance our understanding of behavioral trajectories and the stability of predictive relationships over time. Additionally, future research should investigate the causal mechanisms underlying the observed associations, particularly regarding the relationship between longer jail stays and reduced recidivism, to distinguish between the effects of incarceration length per se and the impact of specific programming or treat-

ment interventions available during extended periods of confinement.

## References

- Antenangeli, L., & Durose, M. (2021, September). *Recidivism of prisoners released in 24 states in 2008: A 10-year follow-up period (2008–2018)* (NCJ 256094). U.S. Department of Justice. <https://bjs.ojp.gov/library/publications/recidivism-prisoners-released-24-states-2008-10-year-follow-period-2008-2018>
- Getoš Kalac, A.-M., & Feuerbach, L. (2023). On (measuring) recidivism, penal populism and the future of recidivism research: Neuropenology [Publisher: Akademija pravnih znanosti Hrvatske]. *Godišnjak Akademije pravnih znanosti Hrvatske*, XIV(1), 1–28. <https://doi.org/10.32984/gapzh.14.1.1>
- Eaglin, J. (2017, October 1). Constructing recidivism risk. <https://doi.org/10.2139/ssrn.2821136>
- Garland, D. (1991). Sociological perspectives on punishment [Publisher: [University of Chicago Press, University of Chicago]]. *Crime and Justice*, 14, 115–165. Retrieved July 27, 2025, from <https://www.jstor.org/stable/1147460>
- Beirne, P. (1987). Adolphe quetelet and the origins of positivist criminology [Publisher: The University of Chicago Press]. *American Journal of Sociology*, 92(5), 1140–1169. Retrieved July 27, 2025, from <https://www.jstor.org/stable/2779999>
- Beccaria, C. B. d., & Voltaire. (1872). *An essay on crimes and punishments* (A New Edition Corrected). Albany: W.C. Little & Co.
- Rafter, N. H. (1992). Criminal anthropology in the united states [\_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.1992.tb01115.x>]. *Criminology*, 30(4), 525–546. <https://doi.org/10.1111/j.1745-9125.1992.tb01115.x>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis [Place: US Publisher: American Psychological Association]. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk* [Pages: xviii, 356]. American Psychological Association. <https://doi.org/10.1037/10304-000>
- Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! [\_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.1996.tb01220.x>]. *Criminology*, 34(4), 575–608. <https://doi.org/10.1111/j.1745-9125.1996.tb01220.x>
- Silver, E., & Miller, L. L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control [Publisher: SAGE Publications Inc]. *Crime & Delinquency*, 48(1), 138–161. <https://doi.org/10.1177/0011128702048001006>
- Kröner, C., Stadtland, C., Eidt, M., & Nedopil, N. (2007). The validity of the violence risk appraisal guide (VRAG) in predicting criminal recidivism. *Criminal behaviour and mental health: CBMH*, 17(2), 89–100. <https://doi.org/10.1002/cbm.644>
- Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-r and the compas: Validation data on two risk-needs tools [Publisher: SAGE Publications Inc]. *Criminal Justice and Behavior*, 35(9), 1095–1108. <https://doi.org/10.1177/0093854808320497>

- Bonta, J., & Andrews, D. (2007). Risk-need-responsivity model for offender assessment and rehabilitation 2007-06. *Rehabilitation*, 6, 1–22. <https://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/rsk-nd-rspnsvty/index-en.aspx>
- Hanson, R., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological assessment*, 21, 1–21. <https://doi.org/10.1037/a0014421>
- Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior - CRIM JUSTICE BEHAV*, 36, 21–40. <https://doi.org/10.1177/0093854808326545>
- Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1745-9125.12123>]. *Criminology*, 54(4), 680–712. <https://doi.org/10.1111/1745-9125.12123>
- Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-985X.2012.01056.x>]. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2), 565–584. <https://doi.org/10.1111/j.1467-985X.2012.01056.x>
- Berk, R., & Bleich, J. (2014). Forecasts of violence to inform sentencing decisions. *Journal of Quantitative Criminology*, 30(1), 79–96. <https://doi.org/10.1007/s10940-013-9195-0>
- Travaini, G. V., Pacchioni, F., Bellumore, S., Bosia, M., & De Micco, F. (2022). Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction [Number: 17 Publisher: Multidisciplinary Digital Publishing Institute]. *International Journal of Environmental Research and Public Health*, 19(17), 10594. <https://doi.org/10.3390/ijerph191710594>
- Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, 27(4), 547–573. <https://doi.org/10.1007/s10940-011-9137-7>
- Karimi-Haghighi, M., & Castillo, C. (2021). Enhancing a recidivism prediction tool with machine learning: Effectiveness and algorithmic fairness. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 210–214. <https://doi.org/10.1145/3462757.3466150>
- Doshi-Velez, F., & Kim, B. (2017, March 2). Towards a rigorous science of interpretable machine learning. <https://doi.org/10.48550/arXiv.1702.08608>
- Biddle, J. B. (2022). On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning [Number: 321-341 Place: Edmonton, United Kingdom Publisher: Cambridge University Press Section: Article]. *Canadian Journal of Philosophy*, 52(3), 321–341. <https://doi.org/10.1017/can.2020.27>
- Keve, P. W. (1995). *Prisons and the american conscience: A history of u.s. federal corrections*. SIU Press. <https://books.google.com/books?id=X5-ngmwEdeQC&pg=PA24#v=onepage&q&f=false>
- Feeley, M. M., & Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.1992.tb01112.x>]. *Criminology*, 30(4), 449–474. <https://doi.org/10.1111/j.1745-9125.1992.tb01112.x>

- Sutton, J. R. (1987). Doing time: Dynamics of imprisonment in the reformist state [Publisher: [American Sociological Association, Sage Publications, Inc.]]. *American Sociological Review*, 52(5), 612–630. <https://doi.org/10.2307/2095598>
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism [Publisher: American Association for the Advancement of Science]. *Science Advances*, 4(1), eaao5580. <https://doi.org/10.1126/sciadv.aao5580>
- Flores, A. W., Holsinger, A. M., Lowenkamp, C. T., & Cohen, T. H. (2017). Time-free effects in predicting recidivism using both fixed and variable follow-up periods: Do different methods produce different results. *Criminal Justice and Behavior*, 44, 121. <https://heinonline.org/HOL/Page?handle=hein.journals/crmjusbhv44&id=117&div=10&collection=journals>
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, January 28). *Algorithmic decision making and the cost of fairness* [arXiv.org]. <https://doi.org/10.1145/3097983.309809>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments [Publisher: Mary Ann Liebert, Inc., publishers]. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Wexler, R. (2017, February 21). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. <https://doi.org/10.2139/ssrn.2920883>
- Han, J. X., Greenewald, K., & Shah, D. (2025, April 25). Fairness is more than algorithms: Racial disparities in time-to-recidivism. <https://doi.org/10.48550/arXiv.2504.18629>
- Scurich, N., & Monahan, J. (2016). Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, 40, 36. <https://heinonline.org/HOL/Page?handle=hein.journals/lwhmbv40&id=36&div=6&collection=journals>
- Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness [ISSN: 2375-026X]. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. <https://doi.org/10.1109/ICDE48307.2020.00203>
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017, February). Human decisions and machine predictions. <https://doi.org/10.3386/w23180>
- van Dijck, G. (2022). Predicting recidivism risk meets AI act [Num Pages: 407-423 Place: Amsterdam, Netherlands Publisher: Springer Nature B.V.]. *European Journal on Criminal Policy and Research*, 28(3), 407–423. <https://doi.org/10.1007/s10610-022-09516-8>
- Ferguson, A. (2017). Policing predictive policing. *Washington University Law Review*. [https://digitalcommons.wcl.american.edu/facsch\\_lawrev/749](https://digitalcommons.wcl.american.edu/facsch_lawrev/749)
- deVuono-powell, S., Schweidler, C., Walters, A., & Zohrabi, A. (2015, September). *Who pays? the trust cost of incarceration on families*. Ella Baker Center, Forward Together, Research Action Design. Oakland, CA. <https://ellabakercenter.org/who-pays-the-true-cost-of-incarceration-on-families/>
- The Council of State Governments Justice Center. (2023, April). *The cost of recidivism*. Retrieved July 27, 2025, from <https://csgjusticecenter.org/publications/the-cost-of-recidivism/>

- Martensen, K. (2012). The price that US minority communities pay: Mass incarceration and the ideologies that fuel them [Publisher: Routledge \_eprint: <https://doi.org/10.1080/10282580.2012.681165>]. *Contemporary Justice Review*, 15(2), 211–222. <https://doi.org/10.1080/10282580.2012.681165>
- Fazel, S., & Wolf, A. (2015). A systematic review of criminal recidivism rates worldwide: Current difficulties and recommendations for best practice [Publisher: Public Library of Science]. *PLOS ONE*, 10(6), e0130390. <https://doi.org/10.1371/journal.pone.0130390>
- Yukhnenko, D., Farouki, L., & Fazel, S. (2023). Criminal recidivism rates globally: A 6-year systematic review update. *Journal of Criminal Justice*, 88, 102115. <https://doi.org/10.1016/j.jcrimjus.2023.102115>
- U.S. Department of Justice. (2021, July 15). NIJ's recidivism challenge full dataset. Retrieved July 27, 2025, from [https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/about\\_data](https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/about_data)
- Ash, E., Goel, N., Li, N., Marangon, C., & Sun, P. (2023). WCLD: Curated large dataset of criminal cases from wisconsin circuit courts. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/29c80c549ed67ddd7259559c1bb07c1b-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/29c80c549ed67ddd7259559c1bb07c1b-Paper-Datasets_and_Benchmarks.pdf)
- Wisconsin Court System. (2012, March 1). Statewide OAR/OWS guidelines and penalties. <https://www.wicourts.gov/publications/fees/docs/oarows.pdf>
- Wisconsin Legislative Reference Bureau. (2019). Statutory misdemeanors in wisconsin. [https://docs.legis.wisconsin.gov/misc/lrb/lrb\\_reports/lrb\\_reports\\_3\\_5.pdf](https://docs.legis.wisconsin.gov/misc/lrb/lrb_reports/lrb_reports_3_5.pdf)
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12, 77.
- Cullen, F. T., & Gendreau, P. (2000). *Assessing correctional rehabilitation: Policy, practice, and prospects*. Retrieved July 27, 2025, from <https://www.ojp.gov/ncjrs/virtual-library/abstracts/assessing-correctional-rehabilitation-policy-practice-and-prospects>
- Myers, D. L., & Spraitz, J. D. (2011). Evidence-based crime policy: Enhancing effectiveness through research and evaluation [Publisher: SAGE Publications Inc]. *Criminal Justice Policy Review*, 22(2), 135–139. <https://doi.org/10.1177/0887403410396212>
- Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology*, 1(1), 9–38. <https://doi.org/10.1007/s11292-004-6460-0>

## Appendix A. Project contributions

**Solomon Flax** led the initial outreach and organizing of the group, setting up communication channels and connecting group members. He conducted a majority of the data pre-processing and cleaning activities for the Wisconsin and Georgia datasets, making sure both datasets were ready for exploratory analysis and modeling. Solomon also heavily contributed to developing the modeling process for both the Wisconsin and Georgia datasets, which involved coding in R and utilizing many of the concepts introduced in the class, including logistic regression, variable selection, variable transformation, and residual analysis. He also contributed to discussions around interpreting results and conclusions drawn from the analysis, uncovering specific insights such as the effect of *age\_of\_release* on recidivism in the Georgia dataset.

**Trenor Hamilton** led the analysis of the combined dataset and wrote the code for additional pre-processing steps that were necessary for the joining of observational data. He was heavily involved in discussions of goodness of fit, residual analysis, and interpretation of results. He also analyzed prediction metrics on cross sections of the test data, namely by race and by state, to ensure that our selected best models had as little racial gap as possible in Precision, and authored much of the Results and Conclusion sections. Trenor was instrumental in formatting the final report in LaTeX.

**Stella Kim** initiated the discussion on appropriate data sources and mergeability. She led the exploratory data analysis for Georgia and Wisconsin and summarized key model findings, including interpretations of odds ratios. She also developed visualizations to support the model results and improve interpretability. She also authored a significant portion of the Results section. To drive the project forward, Stella managed the initial workflow, aligning the team on timelines and deliverables to maintain consistency. She also provided support in other areas as needed.

**Seung Min Oh** led initial collection of recidivism datasets and conducted exploratory analysis to discover common variables, organizing them for collective understanding. He then created a summary list table of variables across four datasets used in the analysis. During the analysis plan phase, Seung Min authored the Approaches & Methodology sections. Additionally, he sourced and introduced the IPUMS USA data and wrote data preprocessing codes in R, making the census data compatible with Georgia and Wisconsin datasets by standardizing the column names and aggregating socioeconomic variables. Once the team's inputs were gathered, Seung Min authored the Explanation of Changes section and the Analysis portion of IPUMS data and the combined data sets.

**Aldila Yunus** led the project's foundational background research and writing, collaborating with group members to source and curate academic articles supporting the research process. She ultimately conducted the comprehensive literature review and authored the Introduction sections, including Background, Prior art analysis, and research purpose. When developing the final report, Aldila established the project's organizational framework, creating the Overleaf LaTeX structure, designing the report architecture, and coordinating task delegation across team members. Aldila's individual contributions included sourcing and exploring alternative state datasets (ultimately not included in final analysis) and conducting comprehensive editing of the final report. She also contributed to collaborative aspects including conducting exploratory data analysis, variable selection, results interpretation, and conclusions development.

## Appendix B. Total cost of recidivism by state

**Table 4:** Total cost of recidivism by state (2021)

State	Cost of Recidivism (USD)
Arizona	\$226,345,172
Arkansas	\$175,858,040
California	\$1,924,810,316
Colorado	\$150,436,575
Delaware	\$17,580,794
Florida	\$313,267,794
Georgia	\$201,117,694
Hawaii	\$56,106,340
Idaho	\$94,109,337
Illinois	\$159,012,177
Indiana	\$122,065,870
Iowa	\$103,857,914
Kansas	\$89,737,440
Louisiana	\$231,943,283
Massachusetts	\$14,900,490
Michigan	\$53,462,609
Minnesota	\$91,039,760
Mississippi	\$89,046,524
Missouri	\$229,107,463
Montana	\$33,558,892
Nevada	\$60,732,073
New Hampshire	\$2,066,630
New York	\$435,556,369
North Carolina	\$323,353,719
North Dakota	\$26,133,599
Ohio	\$98,258,442
Oklahoma	\$73,414,812
Oregon	\$70,497,677
Pennsylvania	\$359,890,000
Rhode Island	\$29,827,920
South Carolina	\$70,483,143
South Dakota	\$39,338,861
Tennessee	\$39,842,360
Texas	\$585,526,120
Utah	\$134,181,891
Vermont	\$39,170,417
Virginia	\$496,861,933
Washington	\$241,855,789
West Virginia	\$25,651,470
Wisconsin	\$426,157,181
Wyoming	\$31,091,375

Source: The Council of State Governments Justice Center. (2023). The Cost of Recidivism: The high price states pay to incarcerate people for supervision violations. April 2023.

## Appendix C. Coefficients of Georgia Reduced Logit Model

Table 5: Coefficient estimates and p-values for Georgia

Variable	Estimate	Pr(>  t )
(Intercept)	0.490314	0.030069 *
RaceCaucasian	0.110681	0.003402 **
Age_at_Release23-27	-0.292647	2.12e-05 ***
Age_at_Release28-32	-0.663447	2e-16 ***
Age_at_Release33-37	-0.891576	2e-16 ***
Age_at_Release38-42	-1.043433	2e-16 ***
Age_at_Release43-47	-1.145448	2e-16 ***
Age_at_Release48 or older	-1.412167	2e-16 ***
Residence_PUMA	0.006695	0.003486 **
Gang_Affiliated1	0.735311	2e-16 ***
Gang_AffiliatedUnknown	-0.410017	7.41e-15 ***
Supervision_Risk_Score_First	0.032980	0.001196 **
Supervision_Level_FirstSpecialized	0.078463	0.103732
Supervision_Level_FirstStandard	-0.195345	8.93e-06 ***
Supervision_Level_FirstUnknown	0.036781	0.611949
Education_LevelHigh School Diploma	0.086016	0.067492 .
Education_LevelLess than HS diploma	-0.084549	0.083843 .
Prison_OffenseOther	0.176374	0.005671 **
Prison_OffenseProperty	0.158930	0.002960 **
Prison_OffenseViolent/Non-Sex	0.110872	0.078662 .
Prison_OffenseViolent/Sex	-0.140302	0.256386
Prison_OffenseUnknown	0.183107	0.001708 **
Prison_YearsGreater than 2 to 3 years	-0.301939	1.17e-09 ***
Prison_YearsLess than 1 year	0.216830	4.21e-07 ***
Prison_YearsMore than 3 years	-0.278330	4.83e-08 ***
Prior_Arrest_Episodes_Felony1	-0.990248	8.48e-08 ***
Prior_Arrest_Episodes_Felony10 or more	-0.142850	0.469511
Prior_Arrest_Episodes_Felony2	-0.720725	9.09e-05 ***
Prior_Arrest_Episodes_Felony3	-0.723701	9.88e-05 ***
Prior_Arrest_Episodes_Felony4	-0.516797	0.005914 **
Prior_Arrest_Episodes_Felony5	-0.435878	0.021774 *
Prior_Arrest_Episodes_Felony6	-0.433518	0.024215 *
Prior_Arrest_Episodes_Felony7	-0.575358	0.003211 **
Prior_Arrest_Episodes_Felony8	-0.378144	0.056990 .
Prior_Arrest_Episodes_Felony9	-0.268445	0.185447
Prior_Arrest_Episodes_Misd1	-0.003247	0.960144
Prior_Arrest_Episodes_Misd2	0.032726	0.643658
Prior_Arrest_Episodes_Misd3	0.017756	0.819263
Prior_Arrest_Episodes_Misd4	0.130516	0.114541
Prior_Arrest_Episodes_Misd5	0.035150	0.693501
Prior_Arrest_Episodes_Misd6 or more	0.237215	0.004655 **
Prior_Arrest_Episodes_Violent1	0.098580	0.019236 *
Prior_Arrest_Episodes_Violent2	0.146384	0.007859 **
Prior_Arrest_Episodes_Violent3 or more	0.105824	0.087335 .
Prior_Arrest_Episodes_Property1	0.087469	0.088065 .
Prior_Arrest_Episodes_Property2	0.092386	0.111631
Prior_Arrest_Episodes_Property3	0.151560	0.025785 *
Prior_Arrest_Episodes_Property4	0.162738	0.036285 *

Continued on next page

Table 5 – continued from previous page

Variable	Estimate	Pr(> t )
Prior_Arrest_Episodes_Property5 or more	0.384785	1.29e-07 ***
Prior_Arrest_Episodes_PPViolationCharges1	0.152482	0.006983 **
Prior_Arrest_Episodes_PPViolationCharges2	0.301542	2.88e-06 ***
Prior_Arrest_Episodes_PPViolationCharges3	0.288147	5.91e-05 ***
Prior_Arrest_Episodes_PPViolationCharges4	0.291234	0.000226 ***
Prior_Arrest_Episodes_PPViolationCharges5 or more	0.503150	
Prior_Arrest_Episodes_DVCharges1	0.101579	0.036114 *
Prior_Conviction_Episodes_Misd1	0.177242	0.000814 ***
Prior_Conviction_Episodes_Misd2	0.228182	0.000253 ***
Prior_Conviction_Episodes_Misd3	0.228496	0.001639 **
Prior_Conviction_Episodes_Misd4 or more	0.300578	3.77e-05 ***
Prior_Conviction_Episodes_Drug1	0.031266	0.468090
Prior_Conviction_Episodes_Drug2 or more	-0.006001	0.904340
Prior_Conviction_Episodes_PPViolationCharges1	-0.140057	0.001178 **
Prior_Conviction_Episodes_GunCharges1	-0.014568	0.771400
Prior_Revocations_Parole1	0.338018	3.08e-09 ***
Prior_Revocations_Probation1	-0.136981	0.003151 **
Condition_MH_SA1	0.315876	2e-16 ***
Condition_Other1	0.100987	0.012271 *
Violations_ElectronicMonitoring1	0.207086	0.002638 **
Violations_Instruction1	0.142075	0.003258 **
Violations_MoveWithoutPermission1	-0.152620	0.005071 **
Delinquency_Reports1	0.596415	1.23e-13 ***
Delinquency_Reports2	-0.079135	0.281144
Delinquency_Reports3	-0.447068	6.73e-09 ***
Delinquency_Reports4 or more	-0.747025	2e-16 ***
Program_Attendances1	0.021660	0.800136
Program_Attendances10 or more	-0.737422	2e-16 ***
Program_Attendances2	-0.093186	0.315722
Program_Attendances3	0.067641	0.515338
Program_Attendances4	-0.310060	0.001448 **
Program_Attendances5	-0.281209	0.000430 ***
Program_Attendances6	-0.416771	6.57e-14 ***
Program_Attendances7	-0.439653	3.86e-06 ***
Program_Attendances8	-0.356461	0.005232 **
Program_Attendances9	-0.326935	0.013933 *
Program_UnexcusedAbsences1	0.246985	0.000495 ***
Program_UnexcusedAbsences2	0.345126	3.47e-05 ***
Program_UnexcusedAbsences3 or more	0.251109	0.000185 ***
Residence_Changes1	0.124403	0.001602 **
Residence_Changes2	0.173498	0.000563 ***
Residence_Changes3 or more	0.335008	1.55e-09 ***
Employment_Exempt1	-0.120965	0.014812 *
THC_Positive_Flag1	-0.045786	0.294374
Cocaine_Positive_Flag1	0.013922	0.833762
Meth_Positive_Flag1	0.308140	5.71e-06 ***
Other_Positive_Flag1	-0.117309	0.108794
log_Avg_Days_per_DrugTest	-0.032250	0.116382
log_jobs	0.937267	2e-16 ***
cat_Percent_Days_EmployedMedium	-0.867843	2e-16 ***
cat_Percent_Days_EmployedHigh	-1.912608	2e-16 ***

## Appendix D. Coefficients of Wisconsin Lasso Model

**Table 6:** Coefficient estimates and p-values for Wisconsin

Variable	Estimate	Pr(>  t )
(Intercept)	8.144e+01	< 2e-16
county	-4.041e-04	0.592095
GenderM	1.485e-01	0.000285
RaceAmerican Indian or Alaskan Native	-1.489e-01	0.114991
RaceAsian or Pacific Islander	-3.880e-01	0.025945
RaceCaucasian	-3.793e-01	6.54e-12
RaceHispanic	-3.637e-01	1.50e-06
_typeFelony	-2.261e-01	0.011531
_typeMisdemeanor	3.862e-03	0.950213
wcisclassDrug Possession	-3.180e-01	0.010438
wcisclassOAR/OAS	3.037e-01	0.014656
wcisclassOperating While Intoxicated	-2.806e-01	0.013453
wcisclassOther	-1.741e-01	0.077680
wcisclassResisting Officer	-6.910e-03	0.952200
prior_felony	1.403e-02	0.654043
prior_misdemeanor	2.599e-01	< 2e-16
prior_criminal_traffic	2.133e-01	< 2e-16
highest_severity	-1.355e-02	0.289524
pct_black	-1.509e-01	2.57e-16
pct_hisp	-4.126e-02	0.044132
pct_male	1.270e+00	0.036353
pct_rural	-2.783e-02	0.025176
pct_urban	1.135e-02	0.201800
pct_college	-5.974e-02	0.106603
pct_food_stamps	1.795e-01	1.11e-06
pct_somecollege	7.939e-02	0.339814
year	-4.060e-02	< 2e-16
prior_charges_severity7	2.875e-01	0.001046
prior_charges_severity8	-1.065e+01	0.899634
prior_charges_severity9	1.438e-01	0.173225
prior_charges_severity10	2.159e-01	0.004329
prior_charges_severity11	2.991e-01	0.001235
prior_charges_severity12	-5.537e-02	0.593845
prior_charges_severity13	9.943e-02	0.319254
prior_charges_severity14	-3.655e-01	0.021554
prior_charges_severity15	-5.515e-02	0.693105
prior_charges_severity16	9.006e-02	0.377826
prior_charges_severity17	6.992e-03	0.944014
prior_charges_severity18	4.116e-02	0.686208
prior_charges_severity19	-4.960e-01	0.146282
prior_charges_severity20	-3.547e-01	0.085598
prior_charges_severity21	9.656e-02	0.855669
violent_crime1	-1.807e-01	0.012171
Age_at_Judge_Bin18-22	-8.646e-01	1.00e-11
Age_at_Judge_Bin23-27	-1.201e+00	< 2e-16
Age_at_Judge_Bin28-32	-1.363e+00	< 2e-16
Age_at_Judge_Bin33-37	-1.433e+00	< 2e-16
Age_at_Judge_Bin38-42	-1.499e+00	< 2e-16

Continued on next page

Table 6 – continued from previous page

Variable	Estimate	Pr(>  t )
Age_at_Judge_Bin43-47	-1.559e+00	< 2e-16
Age_at_Judge_Bin48 or older	-1.902e+00	< 2e-16
log_jail	5.106e-02	0.000535
log_median_hist_jail	3.403e-02	0.008050
log_pop_dens	-2.469e-02	0.371400