Breaking the Cycle: Predictive Modeling of Repeat Offenses in Broward County, Florida

David Lubbers, Jay Mangrulkar, Rikesh Patel,
Ross Lohrisch, Shelli Hatcher

November 25, 2025

Abstract

Recidivism remains a central challenge in criminal justice settings, reflecting the combined influence of social, structural, and behavioral factors that shape an individual's likelihood of reoffending. Using Broward County COMPAS records merged with county-level arrest data and local weather information, this project evaluated several statistical approaches, including logistic regression with LASSO selection, ridge-regularized logistic regression, Poisson regression using prior counts, logistic regression guided by AIC/BIC feature selection, and logistic regression enhanced with interaction terms, to estimate two-year reoffending risk. Across methods, age and prior offenses consistently emerged as the strongest predictors, while environmental and county-level variables showed smaller but directionally consistent effects. Multiple models demonstrated reliable discrimination between recidivists and non-recidivists, with visuals supporting a clear separation in risk profiles. Overall, the findings reinforce that recidivism is best understood through a multivariate lens that combines demographic, behavioral, and contextual factors, providing a clear and evidence-based foundation for analyzing recidivism patterns.

Introduction

Problem Origin

All throughout human history, criminal justice systems around the world have adopted the practice of using prisons, jails, and other forms of rehabilitation for people who commit crimes. The practice of re-arresting individuals with a criminal history has been around since colonial times; initially, there was no specific field of study or name given to the act of re-arresting individuals who have previously committed crimes. Throughout history, in the United States, logs of incarcerated individuals and formal tracking of the prison systems was not always kept in a very detailed manner (Johnson, 2017). However, over time, and specifically in the 19th century, tracking became more formal and, in many places, required. Within the field of criminal justice, one specific topic that is studied is recidivism. Recidivism is one of the most fundamental concepts in criminal justice and is described as a person's relapse into criminal behavior after the release from prison, or the completion of a prison sentence (National Institute of Justice, n.d.).

The study of recidivism and the tracking of recidivism statistics began after enactment of the Sentencing Reform Act of 1984, and originally began as the criminal justice reform systems around the United States were instructed by the Act to provide systems of reform and help for those that were exiting the prison system (Breyer, 2022). When considering various ideas for prison reform, and the governmental budget that the prison reform programs were going to get, the individual states and counties as well as the federal government began to track recidivism rates to determine which areas of the United States (counties, cities, and states) were going to get budgets for prison reform programs.

At a very high level, it is recorded that around 49% of offenders were rearrested within 8 years, making the national average recidivism rate at around 49% over the course of 8 years (United States Sentencing Commission, 2022). Overall, recidivism—the study and tracking of re-incarceration rates—is crucial in determining crime statistics of an area. This report serves as a detailed quantitative study on recidivism, the factors that affect it, and general trends of recidivism within the United States.

Problem Evolution

Although the behavior of criminals reoffending has been consistently observed throughout human history, the systematic tracking of recidivism in the U.S. was only made feasible with the formalization of prison records and centralized criminal registries in the 19th and early 20th centuries. As more data was collected on inmate demographics, sentence lengths, release dates, and subsequent arrests, researchers gained the ability to link individual "before" and "after" outcomes (Shoemaker & Ward, 2017). This expansion of data transformed recidivism from a more speculative topic into a tangible social indicator. Early 20th century criminologists quickly began using these emerging datasets to estimate baseline rates of reoffending and to identify demographic

or institutional factors, such as age at release, prior convictions, and sentence length (Glueck & Glueck, 1930). Key response variables in these early recidivism studies included rates of rearrest, reconviction, and reincarceration over time. Decades later, federal efforts such as the U.S. Department of Justice's Historical Corrections Statistics in the United States, 1850–1984 report sought to consolidate and standardize scattered historical records. These initiatives created a comprehensive dataset that researchers could use to analyze long-term trends in incarceration and recidivism.

By the middle of the 20th century, U.S. correctional policy began to pivot from a purely correctional approach toward one that emphasized rehabilitation and structured re-entry to society. Parole and probation systems expanded, reflecting a growing sentiment that supervised release, education, and job training could reduce the likelihood of reoffending (Cullen & Gilbert, 2012). Criminal theory naturally evolved alongside these policy changes. For example, labeling theory began to argue that the "ex-convict" stigma could enforce criminal identity and serve as a barrier to successful reintegration (Becker, 1963) while strain theory and social disorganization theory highlighted the roles of poverty, unemployment, and neighborhood instability in shaping post-release outcomes (Sampson & Groves, 1989). The evidence during this time frame linked stable employment, family support, and educational access with lower recidivism rates. Therefore, reoffending was framed as a process influenced by social and economic opportunity rather than individual moral failure. These developments in the mid 20th century laid important groundwork for the evidence-based interventions and advanced statistical modeling techniques that make up modern recidivism research.

Beginning in the late 1970s, U.S. criminal justice policy shifted sharply towards a "tough-on-crime" orientation. Legislation reflected this shift with initiatives such as mandatory minimum sentencing, "three-strike" laws, and the intensification of the war on drugs. Together, policies like these expanded both the scale and duration of incarceration (Tonry, 1996; Zimring, 2010). Between 1980 and 2000, the U.S. state and federal prison population more than tripled, creating historically high incarceration rates (Western, 2006).

This abrupt expansion of incarceration reversed many of the rehabilitation gains of the mid 20th century. Resources for programs such as education, vocational training, and substance-use treatment failed to keep pace with the rising prison population, sharply limiting the reentry supports that were once viewed as critical to reducing recidivism (Immarigeon & Petersilia, 2009). Studies during this period documented these effects, showing that high rates of imprisonment disrupted family structure and labor-market activity, both shown to increase the likelihood of post-release offending (Clear et al., 2001). The resulting prison environment showed overcrowding and limited programming, associated with elevated risk of misconduct and rearrest (Steiner & Wooldredge, 2009). Large cross-state studies by the Bureau of Justice Statistics reported that roughly two-thirds of individuals released from state prisons in the 1980s and 1990s were rearrested within three years, demonstrating that recidivism was not effectively discouraged by harsher sentences (Beck & Shipley, 1989; Langan & Levin, 2002).

With the limits of harsh sentencing becoming increasingly clear, the early 2000s brought along an era emphasizing evidence-based strategies and more data-driven decision making. Federal initiatives such as the Second Chance Act of 2007 and First Step Act of 2018 signaled a shift back towards programs designed to improve reentry outcomes through education, employment assistance, and behavioral health service (Visher & Travis, 2003). At the same time, advances in computing and more efficient dataset management fostered the development of risk-assessment tools such as the Level of Service Inventory-Revised (LSI-R) and COMPAS to estimate an individual's likelihood of reoffending based on demographic, criminal history, and psychological factors (Andrews et al., 2006). This period also brought more attention to fairness and transparency, as researchers documented how predictive tools can reflect or amplify existing socioeconomic disparities (Skeem & Lowenkamp, 2006). Overall, the 21st century has been characterized by a shift from broad disciplinary measures towards data-driven strategies that aim to reduce recidivism while reducing equity concerns.

Problem Impact

The persistence of recidivism challenges the fundamental purpose of the justice system: rehabilitation. High rates of re-offending indicate that the current approach often fails to prepare individuals for successful reintegration, instead expelling them from society and trapping them in a cyclical life in and out of prison. Scholars posit recidivism as a "failure of rehabilitation," noting how the environment leads to reinforced dependence on others and leaves returning citizens unprepared to navigate socioeconomic hurdles after release (Sanyal, 2019). This cycle undermines rehabilitation, contributes to overcrowded prisons, and leaves communities vulnerable to ongoing crime. If recidivism can be predicted, then those predictions can support policy reform. Models can

guide policy and interventions, such as rehabilitation. Reducing recidivism will lead to safer neighborhoods and greater public trust in the fairness of the prison system.

Previous Analytical Efforts

Previous analytical evaluations of crime recidivism span a wide range of topics and approaches. These efforts fall primarily into two categories: factors that affect likelihood of recidivism and factors that affect time until recidivism. Research by Evans (1968) examined the relationship between success in the labor market and recidivism. His research focused on released offenders as a class of disadvantaged workers in the labor market, finding that success in the labor market and the "quality of post-prison employment is important to the problem of recidivism." Later work confirms the importance of employment on rates of recidivism and expands relevant factors to include gender, age, and nationality (Monnery, 2013).

Additional research has investigated the effects of demographic factors such as marriage and religion. Andersen et al. (2015) found that "marriage reduced recidivism compared to nonmarriage only when the spouse had no criminal record" while Stansfield et al. (2017) concluded that "religious and spiritual support does have a strong and robust effect on the likelihood of ex-offenders desisting from substance abuse." However, this finding was not found to apply to other types of crimes beyond criminal drug users.

Some research has also been conducted on factors affecting a criminal's experience during their incarceration and their effects on recidivism. Tan and Zapryanova (2021) find that prisoners in more homogeneous prisons, with peers similar in race and age, exhibit increased tendency towards future offenses and career criminality. Their research concludes that increased heterogeneity in prisons could lead to reduced recidivism. Additionally, correctional education programs are shown to have a large effect on reducing recidivism when compared to other recidivism reduction programs (Hall, 2015). The benefits of correctional education are especially pronounced when the program is provided to Black/African American men with only a high school education or GED equivalent (Duke, 2018).

Visitation during the prison sentence has also been found to affect rates of recidivism, with the effect being highest among inmates receiving more frequent visits (Casey et al., 2020). Specifically, "visits from siblings, in-laws, fathers, and clergy were the most beneficial in reducing the risk of recidivism, whereas visits from ex-spouses significantly increased the risk" (Duwe & Clark, 2011). However, there is disagreement amongst the literature on whether this effect of ex-spouse visitation is causal (Cochran et al., 2018).

Researching factors affecting time until recidivism aims to identify factors that affect time between release and next arrest. Witte and Schmidt (1977) found the factors of previous convictions, age, race, alcohol history, and drug history to be significant in predicting the time until recidivism. Models utilizing a Cox proportional hazard technique reinforce the importance of age and previous convictions factors, while additionally finding significance in gender and offense type (Bowles & Florackis, 2007). Black males also "recidivated in a shorter time frame than their white peers with the covariates age at release and length of stay in jail controlled" (Jung et al., 2010). These results have been consistent in countries outside the United States, indicating culture may play only a limited role on recidivism (Souza et al., 2024).

Analysis Approach

Three data sources were used for the analysis. The first data source for these models comes from ProPublica's compilation of two years of COMPAS data from Broward County. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a commercial algorithmic tool made by Northpoint, Inc. It is used by judicial and correctional personnel to assist in making decisions regarding sentencing and parole (Propublica, n.d.). The dataset contains several possible predictor variables for over 7,000 criminal defendants in Broward County. The second dataset comes from the National Weather Service via Open Meteo and provides the daily weather including temperature, rain, wind for Broward County between 2008 – 2025 (Zippenfenig, 2024). A third dataset records the total annual arrests in the state of Florida by county "as reported to the Uniform Crime Reports program for inclusion in the annual Crime in Florida Report" (Florida Department of Law Enforcement, n.d.).

Methods

Five approaches are utilized to examine this data.

Logistic Regression with Lasso Feature Selection Approach 1 used Logistic regression model, applying L1-penalty for feature selection. The outcome variable will be the 2-year recidivism binary variable. The model was trained by a dataset merging ProPublica's public COMPAS dataset, Broward County's weather data, and total arrests by county and year data, noted above.

Logistic Regression with Ridge Regularization Approach 2 examined the likelihood of recidivism using

logistic regression while implementing ridge regularization. Ridge regression is a helpful tool in cases where linear dependence and multicollinearity exist between the predictor variables. The method "obtain[s] biased estimates with smaller mean square error" for the regression coefficients compared to ordinary least squares optimization coefficient estimates (Hoerl & Kennard, 2000). This logistic regression model will combine the three datasets as predictor variables and examine the response variable indicating recidivism from the ProPublica dataset. This binary variable indicates whether the criminal defendant had "a new arrest within two years" of the initial conviction (Larson et al., 2016).

Poisson Regression Leveraging prior_counts Approach 3 utilized a Poisson Regression model, using the prior_counts continuous variable, and the outcome will be a recidivism binary variable to demonstrate whether an individual is likely to recidivate based on prior counts. The model will be trained by a dataset using the COMPAS data set, and specifically, the prior_counts variable.

Logistic Regression with AIC, BIC Feature Selection Approach 4 used a standard logistic regression model to predict the binary outcome of two-year recidivism while performing feature selection based on information criteria. Specifically, variables will be iteratively added or removed to minimize the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which balance model fits with model complexity. AIC and BIC penalize the inclusion of unnecessary predictors, helping to identify the most efficient set of features that explain recidivism risk without overfitting. This data-driven selection process will be applied to the combined datasets to identify the strongest predictors of re-offense.

Logistic regression with interaction variables Approach 5 utilized interaction variables (Walters & Crawford, 2014) to search for nonlinear combinations of predictors that have statistical significance. The baseline surface for a two-dimensional linear regression is the plane. Allowing second order interactions admits fits to quadratic surfaces, e.g., ellipses. The theory generalizes in higher dimensions, e.g. hyperplanes, ellipsoids, and saddles. Certainly Box-Cox will be the starting point for further investigation (Tollenaar & van der Heijden, 2019). It is expected that Box-Cox will suggest exponential transformations of incarceration duration. One interaction candidate is the ratio of education courses attempted during incarceration divided by the duration of incarceration. This ratio may measure an inmate's dedication to education. Certain products of predictors may also be useful, certainly time until re-offense times the local crime rate will be investigated.

Analysis Results

Approach 1: Logistic Regression with LASSO Feature Selection

To train this logistic regression model with L1 penalty, the Broward County COMPAS records data was joined with local weather and federal holiday data. With recidivism data from 7,214 individuals, data science modeling techniques can be applied to help improve the estimation and understanding of recidivism and its influential factors like age, sex, and number of priors.

The first step taken in prepping the data was removing several variables that would not help the model much, i.e. identifiers, open-ended text columns, and other redundant information. After doing this, the model was left with only interpretable features.

Diving into exploring the dataset, the balance of individuals between both classes—those who have recidivated, and those who have not — is moderately balanced. Priors count is very skewed to the right, but it may be a useful feature to look at because those who commit crimes are more likely to recidivate. It was interesting to note that the dataset is dominated by males, and individuals are primarily classified as African American or Caucasian. Another interesting insight is that offenders are more commonly younger individuals. Figure 1 highlights the meaningful differences in conditional recidivism rates across number of prior offenses of offenders in the dataset. This shows how those who have a higher number of prior offenses are more likely to be committing recidivism, so prior offenses can be a strong indicator for the model to predict higher likelihood for future offenses.

For a non-linear model, it is not appropriate to calculate Cook's distance to detect outliers, so Cook's distance was substituted with interquartile range. Interquartile range (IQR) was applied to count potential outliers within the numeric variables for each feature. Looking at the summary per feature, no observations were dropped as outliers because the minimum and maximum of each feature was not too far away from the mean. Some features did have a higher percentage of data points outside of the interquartile range, but that is understandable given the high skewness for select features.

Because age_cat is largely redundant with age, the model will exclude age_cat to avoid duplicate information. Missing values in numeric values were imputed with their means while missing values in

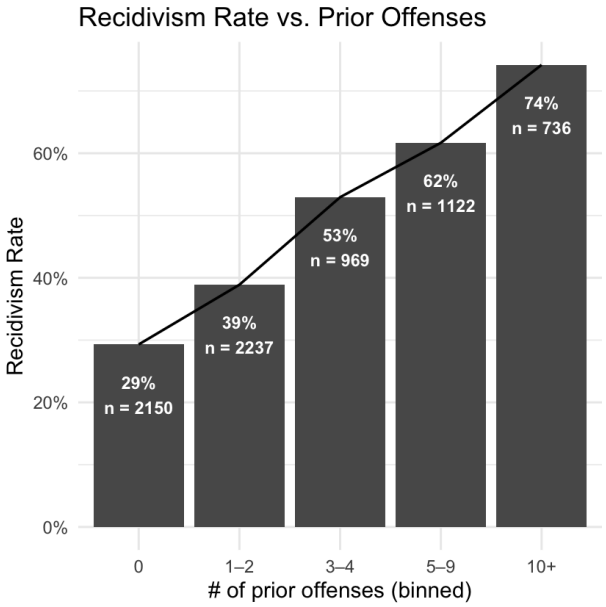

Figure 1: *Bar chart for relationship between recidivism and number of prior offenses.*

categorical predictors were explicitly replaced with a "-1" category to preserve those unknown observations from legitimate categories. To prevent multicollinearity, variance inflation factors were computed after fitting a baseline logistic regression without any penalties. Several weather variables had extremely high VIFs, indicating severe multicollinearity, so to stabilize the model and improve interpretability, most collinear variables were removed. After removing those variables, the VIFs all fell below the threshold for alarm. Using L1-penalty logistic regression (LASSO) is an excellent choice to also help with variable selection and preventing overfitting because it automatically performs feature selection by shrinking unsignificant coefficients to zero, which produces a simpler model to use.

To see how the model fits the data, a simpler model was created from a dataset with replications using categorical variables sex, race, and charge degree. To calculate residuals properly with logistic regression, the dataset must have replications, where there are repeated responses for each unique set of predictors. The 3 categorical variables were chosen to apply for this simple model to conduct a residual analysis as a check before moving on to add more variables and developing a more complicated model. The simple model fails to reject the hypothesis test that logistic regression model is a good fit for the data, and the simple model appears to pass the model's assumptions for linearity and independence. Along with multicollinearity and no strong outliers, the model appears to be a good fit for this data. After including the L1 penalty to the model, the model did not automatically shrink any variables to zero, likely because many other irrelevant features were already removed in previous steps.

Looking at the final model results, demographics notably appear to be associated with recidivism rates. Male sex carries a significant, positive coefficient (approximately 0.30), corresponding to an odds ratio of about 1.34; holding other variables constant, males have about 34% higher odds of recidivism than females according to this model. Meanwhile, age is strongly negatively associated with recidivism: the coefficient is roughly -0.044, implying that each additional year of age reduces the odds of recidivism by about 4.5%. This confirms that younger individuals are at a systematically higher risk. Criminal history variables are also strong indicators. The coefficient for priors_count is around 0.145 and highly significant, yielding an odds ratio of roughly 1.16 per additional prior offense. This means that each extra prior offense is associated with about a 16% increase in the odds of recidivism, holding other factors constant. The weather variables like maximum temperature and precipitation show small but statistically detectable effects; however, their odds ratios are extremely close to 1, suggesting that their actual impact on recidivism probabilities is small compared to demographic and criminal history variables.

To best evaluate the true performance of the model for predicting recidivism, the dataset can be split into training and test datasets, so there is data that the model has not seen before to simulate how accurate the model would be with new observations. In the training set, the model achieved an accuracy of approximately 71%, and on the held-out test set, accuracy was about 68%. Receiver operating characteristic (ROC) curves were created to show the trade-off between true positive rate and true negative rate. The area under the curve (AUC) for training was approximately 0.77, and the test AUC was approximately 0.73. The ROC curve for the test set lies notably above the 45-degree line of no discrimination, indicating that the model predicts higher likelihood to reoffenders than non-reoffenders with reasonably high probability beyond random guessing. These results suggest moderate predictive performance.

Approach 2: Logistic Regression with Ridge Regularization

In examining the potential predictor variables available in the three datasets, the potential for multicollinearity is evident. In the weather dataset, it is expected that daily_temperature_2_max and daily_apparent_temperature_max would be correlated variables. Therefore, ridge regularization was examined in this approach as a method for modeling

the data and accounting for multicollinearity.

Prior to implementing this approach, the three datasets were cleaned and joined together. It is important to understand the context of these joins prior to analyzing the results. In the weather dataset, an initial examination of the weather_code variable showed a potentially uneven distribution of the number of days corresponding to sunny-day codes and those corresponding to rainy-day codes. After aggregating the sunny and rainy weather codes, we find that 71.4% of the days in the dataset are described as rainy while only 28.6% are described as sunny. This value is lower than expected, as the initial assumption was that a weather code was assigned at the same time each day. The updated assumption is that a rainy weather code was assigned if it rained at all during the day, which aligns with the expected rainy-day proportion of the tropical monsoon climate of Broward County, Florida. The weather data was joined with the COMPAS data based on the date of the criminal offense. Potential trends could be that crimes committed on a rainy day or a colder day are more correlated with recidivism.

The arrests dataset containing the number of arrests in each Forida county by year is filtered to only those records from Broward County to align with the COMPAS data. However, the COMPAS data primarily contains offenses during 2013 and 2014. If the arrests data was joined on the year of offense, there would be almost no variability across records, and this data would be less useful in modeling. Instead, the arrest data was appended based on birth year. Potential trends are that the number of murder or drug offenses in Broward County during the year of the offender's birth correlates to the likelihood of recidivism.

Finally, to address the NA values in the combined dataset, the use of imputation was considered. These data points were instead removed because missing years of weather or arrest data cannot be reasonably estimated from averaging future years. Future analysis could utilize time series techniques for estimation and imputation. The final data frame used in this approach contains 2,149 records, which were split into training and testing sets to improve the evaluation of the models.

An initial basic logistic regression model is evaluated to confirm the suspected multicollinearity. The training set includes 55 predictor variables that are expanded to 75 coefficients due to categorical factors. Of these coefficients, 25 could not be defined in the model due to high collinearity. Eleven coefficients in this model are statistically significant at confidence level $\alpha = 0.1$ and the overall regression is found to have explanatory power at significance level $\alpha = 0.001$.

Table 1: *Variable pairs with correlation greater than 0.95. Variable names have been shortened.*

Predictor 1	Predictor 2	Correlation
precipitation_sum	rain_sum	1
appar_temp_max	temp_2m_max	0.9609
appar_temp_min	temp_2m_min	0.9680
liquor	non_forcible	0.9599
counterfeit	prostitution	-0.9665
forcible	robbery	0.9711
rate_per_100k	adult_arrests	0.9709
prostitution	adult_arrests	0.9507
misc	adult_arrests	0.9594
rate_per_100k	total_arrests	0.9748
adult_arrests	total_arrests	0.9725

To further explore the presence of multicollinearity in the data, the correlation coefficient between all pairs of numeric variables is calculated. Table 1 contains all variable pairs with correlation greater than 0.95. Correlation above this threshold has near-perfect correlation and is examined prior to the ridge regularization modeling approach.

The weather-related variables related to rain and temperature are likely near-duplicates in the original data as there are minor differences in measuring the apparent maximum temperature and the actual maximum temperature. Therefore, one of the predictors in these pairs is removed from the dataset. Some of the arrest-related variables are linear combinations of the others. For example, total_arrests is a summation of all individual arrest categories. Therefore, these aggregation categories are removed as this approach does not aim to examine the effects of combinations of predictors. After removing these variables, a few pairs with correlations above 0.95 remain in the dataset. Further analysis should utilize variable selection techniques to determine if these or other variables should be removed from the model.

A second basic logistic regression model is developed as a baseline to compare the ridge regularized models. This model could only define 45 out of 65 coefficients due to the remaining high collinearity. Seven coefficients in this model are statistically significant at confidence level $\alpha = 0.1$ and the overall regression is found to have explanatory power at significance level $\alpha = 0.001$. Overdispersion is estimated to be 0.947. Based on a threshold of 2, this model is not overdispersed.

Ridge regularization is implemented using the glmnet library. Alpha is set equal to 0 to utilize an L2 penalty. This penalty accounts for the multicollinearity discovered in the base logistic regression model, but it does not perform variable selection. Cross

validation with 10 folds is used to determine the optimal value of lambda. Standardization of the predictor variables is performed as a part of the implementation, and the function is set to minimize misclassification error in the optimized model. The optimal lambda for this ridge regularized model is 0.0133. Standard errors and statistical significance of coefficients are not calculated for ridge regularization models as they "are not very meaningful for strongly biased estimates such as arise from penalized estimation methods" (Goeman, 2010).

Evaluation of overall regression and model performance can still be completed. For this model using the test dataset, the misclassification error rate is equal to .4, the mean squared error is equal to 0.468, and the mean absolute error is equal to 0.93. Per the calculated AUC, the model has a 64.3% chance of estimating a random yes point higher than a random no point. Accuracy for the model is 0.594, precision is 0.749, sensitivity is 0.599, and specificity is 0.586. The overall regression is found to have explanatory power at significance level $\alpha = 0.001$. Overdispersion is estimated to be 0.948. Based on a threshold of 2, this model is not over-dispersed. Because the dataset is binomial data without replications, goodness of fit tests cannot be reliably performed on this model. Only the categorical variables in the data could be aggregated to create data with replications, and the resulting model could not be directly equated to this model containing both categorical and numeric data.

A second ridge regularized model is developed to compare the differences between cross validation tuning parameters. For this second ridge regularization model, squared loss is minimized instead of classification error. The optimal lambda found via cross validation is 0.164.

The three models - basic logistic regression (model2), ridge regression minimizing classification error (model3), and ridge regression minimizing squared loss (model4) - are compared via their performance metrics. All three models have similar accuracy measures, but model2 is the least precise, and model4 is the most. Model2 has the highest sensitivity measure and the lowest specificity measure. Both model3 and model4 have similar sensitivity and specificity measures. All four models have similar AUC. Based on these measures, ridge regularization methods improve the logistic regression model and its ability to predict recidivism due to the improvements to precision performance. Model3 is a slightly better model than model due to its higher accuracy, sensitivity, and specificity measures.

Model3 estimates coefficients for 72 predictor variables. The glmnet library returns estimated coefficients from standardized predictors in the original scale. Based on the signs of the coefficients, there is positive relationship between recidivism and daily_rain_sum. Specifically, model3 indicates that the log odds of recidivism increase by $8.075e^{-3}$ for a one unit increase in daily_rain_sum. This indicates that if there is more total rain on the date of offense, the offender is more likely to recede. The estimated coefficients also indicate a negative relationship between recidivism and drug_arrest. Specifically, model3 indicates that the log odds of recidivism decrease by $9.472e^{-6}$ for a one unit increase in the drug_arrest. This indicates that if there are more drug arrests in Broward County during the birth year, the offender is less likely to recede. However, the ridge regression has forced all coefficients close to zero due to the high multicollinearity. Therefore, these effects are relatively small.

Approach 3: Poisson and Logistic Regression Leveraging the prior_counts Variable

As mentioned in the title of this section, approach 3 leverages the prior_counts.1 variable to help determine if the number of prior counts that an individual has, influences how likely they are to commit recidivism. In other words, the question this approach aims to answer is if someone who more or less prior counts has is more likely or less likely to commit recidivism.

To complete this approach, the original data set, titled “final_merged_dataset.csv” was used. Essentially, this data set is a compilation of 2 years of recidivism data from Broward County, Florida. The original data set is also merged with a COMPAS data as well as weather data as mentioned in the Analysis Approach section above, in this paper. To begin, the data set was cleaned and only certain fields and variables in the data set were kept – while the rest were removed. This was because not all variables in the large data set were required for this approach. Variables such as the weather, and characteristics of the ID number (which represents the individual person) were not required. Therefore, the first step was to filter and clean the data so that the only 4 variables that were left to perform this approach on were the following:

- ID
- Priors_count.1
- Event
- Two_years_recid

The ID variable, as mentioned in the paragraph above represents the numbering system leveraged by the data set to showcase the individual person.

The priors_count.1 variable represents the amount of prior criminal counts that an individual has. Lastly, the Event and Two_years_recid variables are actually identical in the sense that they are both binary (0/1) variables that represent if a person has committed recidivism (1), or not (0).

After the data was cleaned and only the relevant variables were selected, as part of some data exploration a summary of the data was gathered and presented. The full summary can be found by reproducing the code, however, below are some key statistics that were helpful to understand when exploring the data.

- The average prior counts of the individuals in-scope of this data was 3.4 (can round down to 3 prior counts)
- The maximum prior counts of an individual was 38. Which seemed to be an outlier, given that it is so much larger than the mean.

The hypothesis ahead of running the analysis was that if an individual has a greater number of prior counts, then that person will have committed recidivism, and is more likely to commit recidivism. Throughout this approach, and to test the hypothesis, not only was Poisson regression run, but another method was also tested. Since this approach included a binary variable, as well as a numerical variable, the approach involved fitting a Poisson regression model, and then to confirm and check results also involved fitting a logistic regression model (particularly due to the binary variable that was included).

A Poisson regression model was fit, and a summarized output of the model was generated. The coefficient of the intercept of the Poisson model was -1.118 while the coefficient of the priors_count.1 variable was 0.039.

As can be observed in the Poisson model output summary, the coefficient of the prior_counts.1 variable displays the log ratio of recidivism for each additional prior crime. In this case, the coefficient is 0.039412. It can also be determined that since the coefficient is less than 1, that more prior crimes can be associated with a lower recidivism probability. This is actually very surprising in this case because the hypothesis was that more prior counts would lead to a higher likelihood for recidivism. It also would make more sense because for example, a person who commits crimes frequently, is likely to commit crimes more often and go back to jail / be arrested.

In order to do some model validation and more clearly understand and confirm that the Poisson model, which inherently did not support the hypothesis was correct, since this data set included a binary variable, a logistic regression model was also

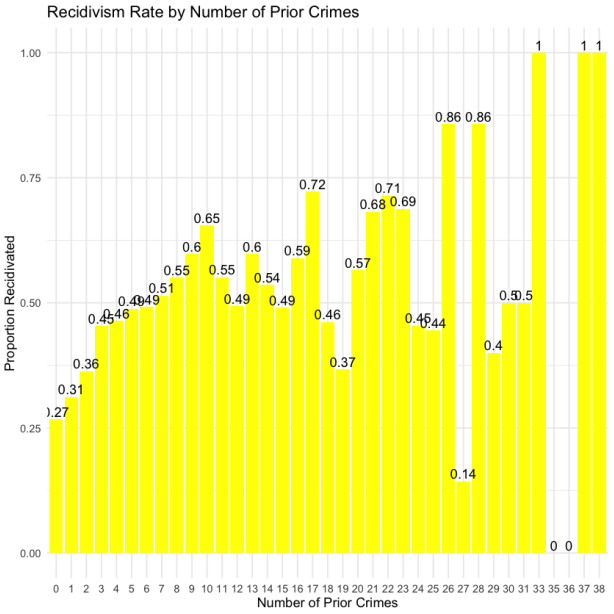

Figure 2: *Bar plot of the number of crimes vs. the percent of the population at that crime rate that recidivated.*

run. A summary of the fitted logistic regression model was displayed and the coefficient of the intercept was -0.7706 while the coefficient estimate of the priors_count.1 variable was 0.082.

It can be observed in the logistic regression model summary output that similarly to the Poisson regression, the prior count variable coefficient is 0.08, which is less than 1. This coefficient represents the change in log odds of recidivism for each additional prior count. This helps confirm the same conclusion that the Poisson model outputted in saying that since the coefficient is less than 1, prior counts do not actually have a direct effect on a higher recidivism likelihood rate. Once again, the original hypothesis was proved incorrect through the running of a logistic regression model. To help understand the trend in a view of a chart, Figure 2 outlines the proportion of individuals that committed recidivism based on the individuals prior counts amount.

As can be observed from Figure 2, there are a couple outliers of very high prior counts. For example, individuals with 31, 37 and 38 prior counts. However, according to the plot, in general, there is not a direct and consistent upwards trend in term of the number of prior crimes and the proportion of individuals that recidivated. Once again, this bar plot is a way to confirm the same findings that were output by the Poisson regression model summary output as well as the logistic regression model summary output.

While there is a short linear upwards trend from around 0 – 10 number of prior crimes, there is once again a drop and non-consistent linear growth of proportion of recidivism. Overall, this approach de-

termined that the initial hypothesis that a greater number of prior counts leads to a greater proportion or likelihood of number of individuals that recidivated was rejected, and there is no clear linear correlation between the two variables

Approach 4: Logistic Regression with AIC/BIC Feature Selection

The logistic regression model developed for this approach was estimated using the cleaned and merged recidivism dataset and reflects the final set of predictors selected after evaluating multiple model specifications. Before fitting the model, the dataset went through pre-processing in which 34 predictors were removed to ensure validity, interpretability, and independence from future information. This included several categories of variables. First, a number of original fields directly encoded by the target outcome or captured information that only becomes available after the recidivism event. Variables such as recidivism_charge fell into this category and were removed to prevent the model from accessing information it could not possibly observe at prediction time. Their inclusion would have artificially inflated predictive accuracy and undermined the integrity of the analysis. Next, raw dates including dob, c_offense_date, c_jail_in, c_jail_out, and c_arrest_date were excluded as logistic regression does not meaningfully interpret date strings and also because equivalent timing information was already captured in derivative predictors such as days_in_jail or days_b_screening_arrest. Next, several derived or duplicate COMPAS features (e.g. age_cat, decile_score.1, etc.) were removed to ensure the model remained independent of COMPAS's proprietary scoring procedures and to prevent the inclusion of information that had already been processed by the COMPAS system. These variables provided no new information beyond what already existed in the underlying numerical predictors and would compromise transparency. Finally, identifiers such as X and id were removed because they encode row-level metadata rather than valuable behavioral or contextual information. The resulting analytical dataset therefore consisted solely of predictors available at assessment time that were conceptually interpretable and free of redundancy.

A 70/30 train–test split was used, and the outcome proportions were kept the same in both sets, so the model was trained and tested under similar conditions. The model, fitted using a binomial family with a logit link, converged without numerical instability, indicating that the refined dataset supported a reliable estimation process. The resulting coefficients revealed a complex relationship of demographic, behavioral, contextual, and environmental influences shaping the probability of recidivism within two years.

Age exhibited a strong inverse relationship with recidivism risk, which reinforces a well-known phenomenon in criminology where individuals tend to age out of crime as their social roles, life responsibilities, and opportunities evolve. By contrast, variables capturing criminal history, including prior counts, prior charges, and indicators of offense severity, displayed strong positive correlations with recidivism. These findings reflect long standing evidence that repeated justice system involvement is more than a snapshot of past behavior, but a reflection of an accumulation of social disadvantage, institutional entanglement, and behavioral trajectories that strongly predict continued contact with the system.

Environmental and contextual predictors, included to explore whether situational features surrounding an offense might shape longer-term tendencies, displayed more modest but still directionally consistent effects. Weather related variables, although not among the most influential predictors, aligned with expectations of how weather might affect crime. For example, fewer crimes occurred when it was raining. County-level arrest variables provided additional context by reflecting local enforcement intensity and broader criminal justice conditions in the counties where the arrests occurred. Their presence acknowledges that recidivism risk is shaped not only by individual behaviors but also by the structural conditions surrounding the offender, including neighborhood level influences, historical patterns of enforcement, and community level social dynamics. Collectively, the estimated relationship represented by the logistic regression model portrays recidivism as a phenomenon arising from the joint influence of individual characteristics, accumulated behavioral patterns, and broader contextual conditions. Together, the chosen predictors form a coherent risk pattern that cannot be attributed to chance or to any single factor alone. Instead, the model's multivariate framework reflects the idea echoed through decades of criminological research that reoffending behavior is rooted simultaneously in personal histories and external environments, and that predictive accuracy requires attention to both.

Model performance on the withheld test dataset further emphasized the strength and effectiveness of this analytical approach. The model achieved an accuracy of 72.4%, indicating strong overall classification quality. Sensitivity was 63.9%, meaning that nearly two thirds of true recidivists were correctly identified, while specificity reached 79.3%, showing that the model correctly classified a substantial majority of non-recidivists. Precision was 71.7%, reflecting

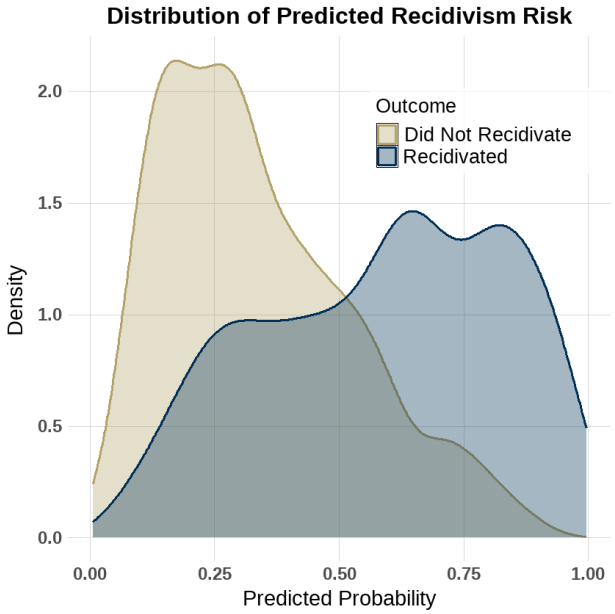

Figure 3: *Estimated density of predicted recidivism probabilities by outcome group. Individuals who recidivated exhibit substantially higher predicted probabilities, demonstrating the model's ability to differentiate higher- and lower-risk cases.*

that individuals predicted to reoffend did so at a relatively high rate. The F1 score of 0.676 balanced precision and recall, indicating a strong overall classification profile without disproportionate weight on one type of classification error. The AUC of 0.783 demonstrated strong discrimination capability, revealing that the model distinguishes meaningfully between higher and lower risk individuals across a range of probability thresholds.

To visualize how effectively the model differentiates between the two outcome groups, a density plot of predicted recidivism probabilities (Figure 3) was generated. This visualization illustrates the entire distribution of predicted risk for individuals who recidivated and those who did not. The curves reveal a clear pattern where non-recidivists cluster heavily toward the lower end of the probability spectrum (indicating confidently low predicted risk), while recidivists exhibit a pronounced shift toward higher predicted probabilities. Although some overlap exists between the two distributions, the peaks are clearly separated, mirroring the strong AUC value and demonstrating that the model discriminates effectively between higher and lower risk individuals. The density plot therefore provides compelling visual evidence that the logistic regression model translates diverse demographic, behavioral, and contextual predictors into meaningful and interpretable levels of recidivism risk.

Taken together, these findings indicate that a comprehensive multivariate logistic regression model provides an empirically grounded, transparent, and theoretically robust approach to estimating two-year recidivism risk in Broward county. The model's performance reflects the integrated nature of reoffending behavior, merging the predictive power of criminal-history variables with demographic indicators and contextual features that capture broader behavioral and structural influences. When compared with alternative modeling approaches, this strategy represents a balanced option that offers clarity and interpretability while maintaining strong predictive performance. Penalized models provide coefficient stability but limited improvements to discrimination, whereas single predictor approaches overlook the structural complexity that is intrinsic to recidivism. The results presented therefore support the conclusion that recidivism forecasting benefits from a multivariate, contextually informed structure, and that logistic regression serves as a robust and appropriate foundation for identifying patterns of reoffending in this dataset.

Approach 5: Logistic Regression with Interaction Variables

Interaction variables are investigated with respect to the COMPAS data set. The results for the particular analysis path improved model accuracy by 0.6% over the simple model of all factors. (0.679 vs .684) This percentage value is small and suggests that interaction variables are not useful for the analysis path chosen. It is noted that the AIC value for the indicator variable result is lower than the simple model by 96.9. As AIC balances model complexity and GOF (goodness of fit) the interaction model is superior. (Lower is better for AIC.)

Interaction Variables - Discussion Interaction variables add derived predictors to a data set. The strategy is that the derived predictors improve the overall regression and the goodness of fit. There are many types of interaction variables. One example considered in the analysis was the interaction of two categorical predictors. Specifically, the Sex and Race predictors, i.e. sex:race. The Sex predictor has two categories: male and female. The Race predictor has about five categories. Thus, without the sex:race interaction the modeling algorithm has seven coefficients to available. Furthermore, without the interaction sex:race there in no way for the glm algorithm to compare pairs such as (male, Native American) to (female, Asian). The sex:race interaction variable adds two (number of sex categories) times five (number of race categories) or ten new categorical variables. The glm algorithms then have much finer grained predictor variables available. For example, the algorithm can assign different coefficients to (male,

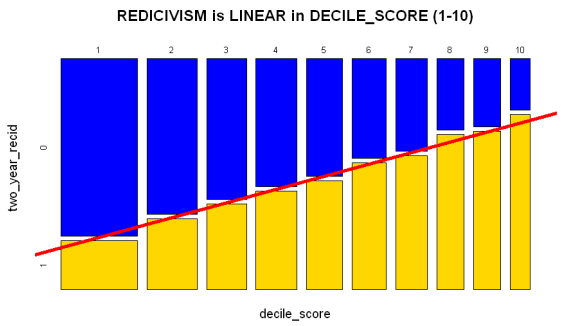

Figure 4: *Mosaic Plot: Decile Score vs Recidivism.*

Native American) and (female, Asian).

Interaction variables based on pairs of numeric factors can also be useful. Such was not the case for the analysis of the COMPAS data. A trivial example might be a gasoline mileage model with numerical predictors miles and gallons. A much more direct input to the regression algorithm would be miles divided by gallons: miles/gallons. This is an example of a numeric:numeric interaction variable.

A third type of interaction variable is the 'functional' type. A numerical predictor may be the square of some value when responses units are in terms of the square root. A 'functional' interaction variable would then be to add a predictor calculated as the square root of the squared predictor. Another example could include the arccosine of a predictor that is periodic with respect to the response.

The type of interaction used in the final analysis model was of the functional type. Using the glm algorithm, the response variable is the logit function. Two numeric predictors, priors_count and decile_score were converted mathematically to probabilities. Now the response is log-odds, so the priors and decile score probabilities were likewise converted to log-odds through cubic polynomials used to approximate the logit function. Strictly speaking, this example is a nested use of 'functional' interaction variables.

Modeling Theme – Exploratory Analysis - Conditional Probabilities A particularly novel approach based on conditional probabilities was implemented. Consider the mosaic plot in Figure 4. Each vertical bar consists of two blocks. At the far left the lower gold block corresponds to decile_score = 1. (Decile_score will be explained below.)

The lower bloc is the proportion of offenders with decile_score=1 who are recidivists. As such, the proportion of the lower bloc is the conditional probability of recidivism given the decile score is one, i.e. Prob (recidivist | decile_score). That the proportions are indeed the said probabilities is demonstrated in the analysis notebook. What is remarkable is that the conditional probabilities as a function of decile_score are strongly linear. It is not clear why such a linear relationship exists.

The linear relationship of the conditional probabilities is interpreted as follows: Any unit increase in the decile score results in a fixed increase in the conditional. The analysis endeavored to leverage the conditional probabilities and the linear relationship.

Variable Selection There are over 7,000 records each with 95 predictors. The response variable is "two_year_recid" – a binary variable where 1 indicates re-offense within two years and 0 indicates no re-offense within two years. The definition of the response variable leads to confusion. The issue is that an offender may have many offenses but happens to have a two-year gap between offenses during the period considered by the analysis. This leads to confusion in the analysis.

The predictor variables include the demographic variables sex, age, age cat, and race. A variable selection process via visual display and model fit indicated that the demographic variables were not strong predictors. Two additional predictors were obvious choices. The predictor priors_count is an integer count of the number of previous offenses and can be as high as 37 prior offenses. It seems very likely that a high number of prior offenses indicates a higher chance of recidivism. This hypothesis is verified in a mosaic plot in the analysis notebook. Likewise, the decile_score predictor was a good choice for the model given the ease of interpretation. Offenders are evaluated for recidivism risk. The evaluation result, the decile_score, ranges from 1=low risk to 10=high risk. During the variable selection, the decile_score was found to follow an easy to explain distribution for non-recidivists. However, for the recidivists, the decile_score predictor follows no simple pattern.

There were two derived predictors: priors_prob and decile_prob. The predictors held the conditional probabilities described above: P(recidivism | priors_count(decile_score)).

Models The overall regression scores were very low for all models considered. Many models that were calculated are included in the analysis notebook. Only the most significant are documented in the next sections

Baseline Model The baseline model was two_year_recid ∼ sex + age + race + decile_score + priors_count. All of the race factor coefficients were not statistically significant. The remaining predictors sex, age, decile_score and priors_count p-values were very small, essentially 0, p-values thus were statistically significant. The Accuracy was 0.679, the AUC was 0.678 and the AIC value was 8710.

Final Interaction Model The formula for the final interaction model was: two_year_recid $\sim$ sex + age + poly (decile_prob, 3)+ poly(priors_prob, 3)

Accuracy was 684, AUC was 0.682, and AIC was 8613.1. So, the interaction terms do not lead to a significant increase in accuracy.

The choice of cubic polynomials for the interactive model is researched carefully in the analysis notebook. The cubic polynomials are an attempt to fit conditional probabilities into log(odds). As shown in the notebook, a cubic polynomial can generate a decent approximation to a log(odds) function. The result then is that the poly (,3) predictors are approximate log(odds) predictors. Also note that linear approximation is also not a bad approximation to log(odds) when used as a predictor.

The model estimated a statistically significant cubic polynomial at the 0.001 level for priors_prob. The model for decile_prob was a statistically significant line at 0.05 line. The other statistically significant predictors were age and sex.

The age_cat was not significant in almost all of the models. The age_cat variable is too crudely binned, only 3 bins, to be of much use statistically. The race predictor did not contribute to almost all the models. This is not to say that race is not a recidivism component; much more likely is that the information contained in the binned race predictor is more strongly encoded in the priors_count and the decile_score.

All of the demographic predictors were absolute numbers, i.e. counts. Total population numbers would have allowed demographic rates in the models. An example is the very low number of Asian offenders. Without knowing the total number of Asian residents, it is hard to calculate what is a 'large' proportion of Asian offenders.

Performance Validation – K Fold Cross Validation A five-fold cross validation was performed for two models: For the baseline model, average performance was Accuracy = 0.677, AUC = 0.732, and F1 = 0.611. For the final version of the interaction model, average performance was Accuracy = 0.681, AUC = 0.737, and F1 = 0.611. Thus, it appears that the final interaction model has little improvement over the baseline model. As stated above, the most notable improvement was in the AIC parameter which, as explained before, balances goodness of fit against performance. In the AIC sense, the interaction model was better. The final statistically significant predictor coefficients were intercept: 0.69, age: -0.03, sexMale:0.31: decile_prob 'x' term 33.29 and priors_prob terms: x = 50.50, x2 = 5.74, x3 = 6.39.

Explanation of Changes

There were no changes to the initial analysis plan outlined prior to the study's implementation. All approaches were completed as planned.

Conclusions

Recidivism is known as a complicated issue with many factors that contribute to a person's likelihood of committing additional crimes after incarceration. This study attempted to examine traditional recidivism factors such as demographics and number of prior crimes while also integrating new factors such as weather and wider arrest trends in the area. Several approaches were utilized to examine recidivism data, and each approach found importance in different predicting factors. The variance across approaches in different selected factors and resulting performance metrics illustrates the importance of examining data from different angles when employing regression analysis.

This study found that feature selection had a large impact on the developed models. Both variable selection techniques examined, stepwise selection and LASSO selection, selected sex and age from the demographic factors. They also selected priors_count and several weather predictors. However, the LASSO selection did not keep any variables from the Florida Department of Law Enforcement while stepwise selection found statistical significance in the total_arrests, arrest_rate_per_100k, and murder predictors. These differences highlight the importance of developing and comparing multiple models.

Due to the high correlation between variables, ridge regularization and interaction variables were employed to examine the effects of transformations on the predictors. Both approaches found only modest improvements in model performance, reinforcing that variable selection is more important in model development due to the large number of predictors included in the dataset.

Future research efforts should examine elastic net and random forest approaches as these methods may lead to further improvements in model performance by combining the factor selection and regularization methods examined in this study. These approaches were not examined in this study because elastic net implementation requires additional hyperparameter selection, which can lead to increased bias without further cross-validation of the model, and random forests lead to decreased interpretability as the algorithm averages the output of many decision trees.

Successes

The method that had the most success out of all our methodologies used AIC to find the most performant model to predict recidivism. By doing an early cleaning of the dataset with imputing missing values and removing redundant features before starting stepwise AIC, the algorithm had a good baseline to start from. At each step, it adds or drops one term and checks whether AIC improves. AIC aims to keep a strong combination of predictors that improves performance while not overfitting the data. Other logistic regression methodologies like LASSO or Ridge had lower accuracy. AIC did the best at identifying the strongest predictors and leading to a great model to predict recidivism with 72% test accuracy and 0.783 AUC score.

There were a few features in the data that were very helpful in predicting recidivism, such as priors_count and the other criminal history variables. This supports the connection between those who commit a crime are more likely to do it again, and it reveals how the criminal-justice system needs to reform more toward helping those key individuals with the greatest difficulty escaping the cycle of repeating crime. If the government aims to break this cycle, then more research needs to be done on recidivism and how rehabilitation programs can be changed to improve it more.

Limitations

While the analyses conducted throughout this project led to the successes mentioned in the section above, there were also a few limitations that were faced. One limitation faced was that the group did not have access to a professional who was an expert around recidivism and criminal data analytics. If the group had access to a subject matter expert, questions could have been asked and clarifications around meanings of variables, and trends could have been better gathered. Another limitation that was faced during the course of the project was that the data sets that were gathered were limited. Recidivism data is not widely available, and locating data with many different variables was difficult.

References

Andersen, S. H., Andersen, L. H., & Skov, P. E. (2015). Effect of marriage and spousal criminality on recidivism. *Journal of Marriage and Family*, *77*(2), 496–509. https://doi.org/10.1111/jomf.12176

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*(1), 7–27. https://doi.org/10.1177/0011128705281756

Beck, A. J., & Shipley, B. E. (1989). Recidivism of prisoners released in 1983. *United States Bureau of Justice Statistics Special Report*.

Becker, H. S. (1963). Outsiders: Studies in the sociology of deviance. *American Journal of Sociology*, *69*(1), 1–18.

Bowles, R. A., & Florackis, C. (2007). Duration of the time to reconviction: Evidence from uk prisoner discharge data. *Journal of Criminal Justice*, *35*(4), 365–378. https://doi.org/10.1016/j.jcrimjus.2007.05.002

Breyer, C. (2022). *Length of incarceration and recidivism united states sentencing commission*. https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2022/20220621_Recidivsm-SentLength.pdf

Casey, W. M., Copp, J. E., & Bales, W. D. (2020). Releases from a local jail: The impact of visitation on recidivism. *Criminal Justice Policy Review*, *32*(4), 427–442. https://doi.org/10.1177/0887403420919480

Clear, T. R., Rose, D. R., & Ryder, J. A. (2001). Incarceration and the community: The problem of removing and returning offenders. *Crime & Delinquency*, *47*(3), 335–351. https://doi.org/10.1177/0011128701047003003

Cochran, J. C., Barnes, J. C., Mears, D. P., & Bales, W. D. (2018). Revisiting the effect of visitation on recidivism. *Justice Quarterly*, *37*(2), 304–331. https://doi.org/10.1080/07418825.2018.1508606

Cullen, F., & Gilbert, K. (2012). *Reaffirming Rehabilitation* (2nd). Routledge.

Duke, B. (2018). A meta-analysis comparing educational attainment prior to incarceration and recidivism rates in relation to correctional education. *Journal of Correctional Education*, *69*(1), 44–59.

Duwe, G., & Clark, V. (2011). Blessed be the social tie that binds. *Criminal Justice Policy Review*, *24*(3), 271–296. https://doi.org/10.1177/0887403411429724

Evans, R. (1968). The labor market and parole success. *The Journal of Human Resources*, *3*(2), 201. https://doi.org/10.2307/145132

Florida Department of Law Enforcement, F. (n.d.). *Ucr arrest data*. https://www.fdle.state.fl.us/cjab/ucr/annual-reports/ucr-arrest-data

Glueck, S. S., & Glueck, E. T. (1930). *Five hundred criminal careers* (1st). Knopf.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, *52*(1), 70–84. https://doi.org/10.1002/bimj.200900028

Hall, L. L. (2015). Correctional education and recidivism: Toward a tool for reduction. *Journal of Correctional Education*, *66*(2), 4–29.

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80. https://doi.org/10.2307/1271436

Immarigeon, R., & Petersilia, J. (2009). *When prisoners come home: Parole and prisoner reentry* (1st). Oxford University Press.

Johnson, J. (2017). Comparison of recidivism studies: Aousc, ussc, and bjs. *Federal Probation*, *81*(1), 52–54.

Jung, H., Spjeldnes, S., & Yamatani, H. (2010). Recidivism and survival time: Racial disparity among jail ex-inmates. *Social Work Research*, *34*(3), 181–189. https://doi.org/10.1093/swr/34.3.181

Langan, P. A., & Levin, D. J. (2002). Recidivism of prisoners released in 1994. *Federal Sentencing Reporter*, *15*(1), 58–65. https://doi.org/10.1525/fsr.2002.15.1.58

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016). *How we analyzed the compas recidivism algorithm*. Retrieved May 23, 2016, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Monnery, B. (2013). The determinants of recidivism among ex-prisoners: A survival analysis on french data. *Groupe d'Analyse et de Théorie Economique (GATE) Lyon-St Étienne Working Paper No. WP 1320*. https://doi.org/10.2139/ssrn.2265349

National Institute of Justice, N. (n.d.). *Recidivism*. https://nij.ojp.gov/topics/corrections/recidivism

Propublica. (n.d.). *Propublica/compas-analysis: Data and analysis for "machine bias"*. https://github.com/propublica/compas-analysis

Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology*, *94*(4), 774–802.

Sanyal, S. (2019). *Recidivism: A failure of rehabilitation* (1st). Delhi Publishing House.

Shoemaker, R., & Ward, R. (2017). Understanding the criminal: Record-keeping, statistics and the early history of criminology in england. *The British Journal of Criminology*, *57*(6), 1442–1461. https://doi.org/10.1093/bjc/azw071

Skeem, J. L., & Lowenkamp, C. T. (2006). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology: An Interdisciplinary Journal*, *54*(4), 680–712. https://doi.org/10.1111/1745-9125.12123

Souza, R. G., Golgher, A. B., & Silva, B. F. (2024). Determinantes da reincidência prisional em santa catarina utilizando a análise de sobrevivência. *Nova Economia*, *34*(4). https://doi.org/10.1590/0103-6351/8027

Stansfield, R., Mowen, T. J., & O'Connor, T. (2017). Religious and spiritual support, reentry, and risk. *Justice Quarterly*, *35*(2), 254–279. https://doi.org/10.1080/07418825.2017.1306629

Steiner, B., & Wooldredge, J. (2009). The relevance of inmate race/ethnicity versus population composition for understanding prison rule violations. *Punishment & Society*, *11*(4), 459–489. https://doi.org/10.1177/1462474509341143

Tan, K. T., & Zapryanova, M. (2021). Peer effects and recidivism: The role of race and age. *The Journal of Law, Economics, and Organization*, *38*(3), 721–740. https://doi.org/10.1093/jleo/ewab040

Tollenaar, N., & van der Heijden, P. G. M. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS ONE*, *14*(3), e0213245. https://doi.org/10.1371/journal.pone.0213245

Tonry, M. H. (1996). *Malign neglect: Race, crime, and punishment in america* (1st). Oxford University Press.

United States Sentencing Commission, U. (2022). *Recidivism among federal offenders: A comprehensive overview*. https://www.ussc.gov/research/research-reports/recidivism-among-federal-offenders-comprehensive-overview

Visher, C., & Travis, J. (2003). Transitions from prison to community: Understanding individual pathways. *Review of Sociology*, *29*(1), 89–113. https://doi.org/10.1146/annurev.soc.29.010202.095931

Walters, G. D., & Crawford, G. (2014). Major mental illness and violence history as predictors of institutional misconduct and recidivism: Main and interaction effects. *Law and human behavior*, *38*(3), 238–247. https://doi.org/10.1037/lhb0000058

Western, B. (2006). *Punishment and inequality in america* (1st). Russell Sage Foundation.

Witte, A. D., & Schmidt, P. (1977). An analysis of recidivism, using the truncated lognormal distribution. *Applied Statistics*, *26*(2), 302. https://doi.org/10.2307/2346971

Zimring, F. E. (2010). The scale of imprisonment in the united states: Twentieth century patterns and twenty-first century prospects. *The Journal of Criminal Law and Criminology*, *100*(3), 1225–1246. https://doi.org/10.2307/25766120

Zippenfenig, P. (2024). *Open-meteo.com weather api*. https://doi.org/10.5281/zenodo.14582479

Appendix

Work was divided equally between all group members. Jay wrote the Problem Origin and Limitations sections. Ross wrote the Problem Evolution section, the abstract, and researched datasets. Rikesh wrote the Problem Impact and Successes sections, researched datasets, and kept notes during meetings. Shelli wrote the Previous Analytical Efforts and Conclusions sections and converted the final report to LaTex format. Dave edited all sections of the final report. All members wrote their own Jupyter Notebook code files, approach analysis section, and cited the sources utilized in their sections of the literature review. Members met weekly via virtual video call and communicated frequently via online chat to provide updates and suggestions to others.

Approach Assignment

1. Approach 1: Rikesh Patel
2. Approach 2: Shelli Hatcher
3. Approach 3: Jay Mangrulkar
4. Approach 4: Ross Lohrisch
5. Approach 5: David Lubbers