Support over Penalty: Identifying Key Factors of and Solutions to the U.S. Recidivism Problem

Bryan Wang

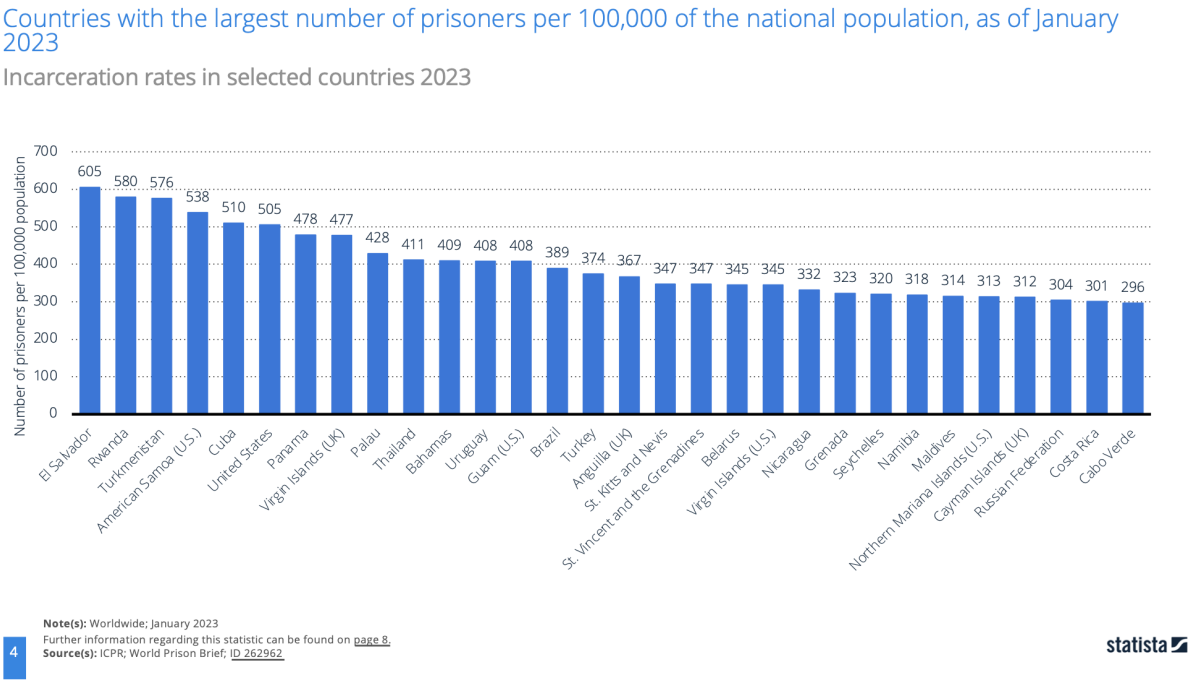
Professor Jonathan Hanke

SML 312

I pledge my honor that I have not violated the Honor Code during this assignment

Overview

The United States, although economically, culturally and martially eminent among other major world powers, bears a unique, paradoxical mark of shame: Ranked sixth seed, trailing close behind totalitarian Cuba, the United States has one of the highest incarceration rates in the world with a rate of 505 prisoners per 100,000 of the national population in the year 2023 (“Most prisoners per capita by country 2023”).



While this statistic might not shock relatively young Americans, it is a sharp increase from what those alive before the 1980s may remember, when incarceration rates were one-fifth—roughly 1 in 1000—of the current ~5-in-1000 proportion. Yet, while incarceration rates have risen, violent crime rates have followed the opposite trend, from its 1990 peak of 758 crimes per 100,000 of the national population to a 2022 level of 370 crimes per 100,000—over a 50% decrease (“Reported violent crime rate in the U.S. 2022”). And while the United States can

boast in a moderately low crime rate compared to peer countries, the incarceration expense that it pays—both in terms of penitentiary maintenance fees and the count of freedomless inmates—to support such a rate is far greater than comparable nations.

This single statistic only scratches the surface of the complex network of life-altering decision-making that is the United States justice system—of coercive plea bargains and steep bail praises and decentralized, patchwork case reporting that shrouds courts and law enforcement behind a shadowy veil. Rather than entangle ourselves in the Kafkaesque business of court reform, we look past the system and to the prisoners, highlighting factors that might put them before a judge and jury—for the first or even second and third time. In this paper, we look at recidivism.

About Recidivism

Along with its uniquely high incarceration rate, the United States's recidivism rate is equally as surprising, with 44% of released prisoners returning to penitentiaries in a single year—and even more within two and three years ("Recidivism Rate by State 2023"). On the other hand, the United States has also some of the world's longest prison sentences, ranking first, 40.6 years, for homicide-related sentences, 6 years more than Mexico's second-place 34.2 years (Warren). Indeed, a 2021 meta-analysis of 116 studies found that "custodial sanctions"—prison and other forms of confinement—"have no effect on reoffending or slightly increase it when compared with the effects of noncustodial sanctions such as probation" (Petrich et al. #).

Simply increasing penalties for crime is a well-proven strategy for failure. That is not to suggest that prisons and law enforcement should be abolished, as has become vogue in the aftermath of horrific police brutality cases, but rather to suggest that if we can make sentences

more effective, even supportive of the convicted—if we can bring down the 44% rate of prisoners reoffending in just one year—we can drastically reduce the count of Americans confined behind bars without even broaching problems of the court system.

Thus, the goal of this paper is two-fold: By building a model to predict recidivism, we hope to first identify those most likely to reoffend, informing where to focus *support* rather than *penalty*, and second, to highlight factors that contribute to recidivism, allowing policymakers to effectively target root problems that affect genuine change.

Related Work

Inspiration for this research came from a variety of sources, namely two: *Custodial Sanctions and Reoffending: A Meta-Analytic Review* by Damon M. Petrich, Travis C. Pratt, Cheryl Lero Jonson, and Francis T. Cullen, researchers at Xavier University and the University of Cincinnati; and *National Institute of Justice's Forecasting Recidivism Challenge: Team "DEAP" (Final Report)* by David B. Wilson, Evan M. Lowder, Peter Phalen, Ashley Rodriguez, one of the top-performing teams who entered the NIJ Recidivism Forecasting Challenge.

Custodial Sanctions and Reoffending: A Meta-Analytic Review provided the theoretical framework that convinced me to consider non-punitive measures to deter crime. Not only does the meta-analysis highlight the null effect of incarceration on lowering recidivism, it further argues that America's preference for incarceration is in fact *criminiogenic*, catalyzing *more* crime rather than less. For one, prison can function as a "school of crime" where criminals effectively teach other criminals how to be better, more evasive criminals, and are themselves likewise instructed. Moreover, prison is also a highly depressive, harsh environment, inducing heavy strains on physical and mental health that can precipitously engender criminal behavior. Nonetheless, the paper insists there must be a better way than barbed wire and metal bars.

Team DEAP (an acronym made up of the first letter of each team member's name) also gave me considerable insight into the data as well as direction for my analysis. In particular, I got the idea for my model ensembling method from one of their solutions (although the exact models and hyperparameters they chose were different/unspecified). Moreover, it helped to refer to their findings as a sanity check for my own results, ensuring that nothing was critically, obviously wrong about the numbers I was reading off from my screen.

One final, informal inspiration for this paper was Princeton University's Professor Thomas Leonard and his ECO 324 class, Law and Economics. Many of the central ideas and several statistics offered in this paper came from his lectures which I found incredibly interesting and useful.

Relevant Data

The National Institute of Justice (NIJ) serves as the research, development, and evaluation agency within the U.S. Department of Justice, with a primary commitment to advancing scientific knowledge and understanding of crime and justice issues. To this end, between January 1, 2013 and December 31, 2015, the NIJ hosted the NIJ Recidivism Forecasting Challenge, a competition designed to create innovative strategies to reduce violent crime and safeguard public safety personnel by mitigating recidivism.

The challenge called for participants to develop forecasting models using 53 variables—49 feature variables and 4 target variables: Recidivism_Arrest_Year1 (recidivism occurring within 1 year), Recidivism_Arrest_Year2, Recidivism_Arrest_Year3, and Recidivism_Within_3years—aiming to improve the prediction of recidivism outcomes for individuals under community supervision. The data for my research project came from this very challenge.

The dataset in its original form contained 25,835 observations, each of which represents a released Georgian prisoner on discretionary parole, supervised between the dates of the challenge. Many of the observations, however, have missing values, and after filtering for "complete" observations, my dataset shrunk down to a smaller, yet still considerable and useful 6,352 observations.

Modeling

Overview

My solution to this recidivism prediction problem was, in the first stage, to implement and fine-tune five different models: Logistic regression (lgr), k-nearest neighbors (knn), decision trees (dtc), random forests (rfc), and neural networks (nn). With each of these models, i was looking at three different performance metrics: Accuracy, specificity, and AUC score. While accuracy gives us a general idea of model goodness, for my particular dataset and research topic, specificity was especially important.

First, the dataset: My dataset contained more positive cases (recidivism occurred within three years) than negative cases, and so a naive model that simply predicted every inmate to recidivate within three years would perform at nearly a 60% (57.60%) accuracy rate. To protect my model against this error, I also measure specificity, the true negative rate, which would yield a 0% rate under the naive positive model. As a final measurement of goodness, I looked at AUC score which is informed by both true and false positive rates, providing a more holistic evaluation of model performance.

Second, for the research topic: While the purpose of this recidivism prediction task is not to more heavily penalize those more likely to recidivate, others interested in purely punitive solutions might take the findings of this and other papers to pursue such an end. As a matter of principle, I agree with the English jurist William Blackstone who posits “the law holds that it is better that 10 guilty persons escape, than that 1 innocent suffer.” While there may be further discussion on the exact ratio of guitly-free to innocent-confined, it should nevertheless be right that the declaration of true innocence be preeminent to the censure of true culpability.

However, as my evaluation criteria favored specificity, there was a natural trade-off with lowered accuracy. Thus, as a compromise between the two metrics, my first modeling stage, training the 5 starting models, prioritized accuracy, whereas my second modeling stage, combining the first-stage models in a meta-model ensemble, prioritized specificity. Ultimately, this produced a final model that performed relatively well in both accuracy and specificity, surpassing the five other models in AUC score.

Data Cleaning

The initial difficulty posed by this dataset was data type conversion. Out of the starting 53 variables, 21 had values of numerical ranges formatted as strings (Ex: Age_at_Release: ['43-47' '33-37' '48 or older' '38-42' '18-22' '23-27' '28-32']). Thus, when I first created subplot histograms of each feature variable, the x axis lacked an ordinal ordering, unable to tell which string of numbers came before or after the other strings of numbers. My first challenge then was to create a unique mapping for each of these 21 variables, converting their string-represented ranges to integers. Using the same example as Age_at Release: “18-22” became 0, “23-27” became 1, “28-32” became 2 and so on. Any value that was of the form “X or more,” simply took on the subsequent value in the sequence—and so “48 or older” became 7.

I then had to convert the nominal variables, namely “Prison_Offense,” into numerical features and used one-hot encoding to split the variable into five different dummy variables, one for each value of the original variable.

Finally, I had a variable named Residence_PUMA that gave the PUMA code (a geographical classifier) of each inmate upon release. The codes were both nominal and many, and since most geographical data I found pertained only to the county-level rather than PUMA,

given the scope of and time I had to complete this paper, I decided to remove this variable from the list of features. Had I more time, I would have sought out further data about the PUMA districts (average income, crime rates, education rates, etc.) and converted those as well into feature variables.

Models:

For all models, I used a 70-30 train-test split and k-fold (5 or 10) cross-validation to fine-tune the model (except for the neural network where I created a validation set from the training data instead of using k-fold cross-validation). Each model, however, posed unique challenges and decisions, so in this section, I will briefly describe each model that I used, choices I made for hyperparameter values, and the final model performance.

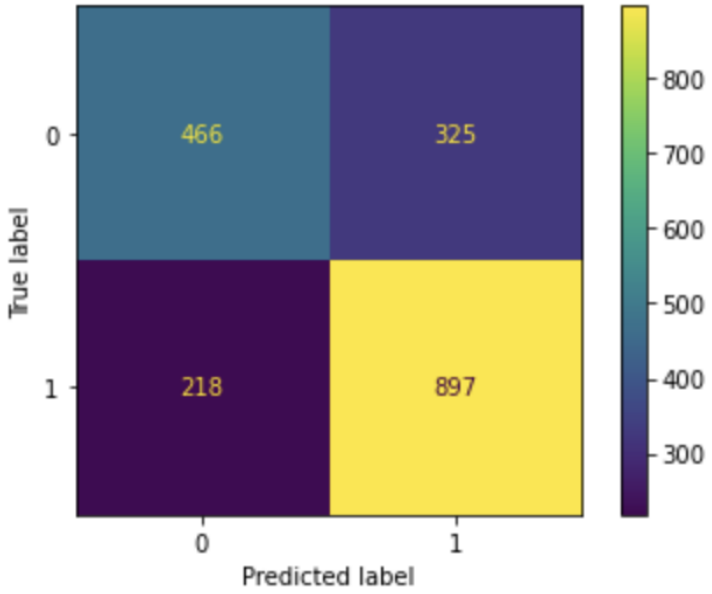
Logistic regression:

The two hyperparameters I chose for logistic regression were the type of penalty (L1, L2 or mixed) and the size of that penalty, "C." Because I believed most of my feature variables to be somewhat relevant, I chose to use the L2 penalty, ridge regression, which shrinks coefficients for each feature variable close but not exactly to zero, which the L1 penalty, lasso regression, does. Using L2 therefore yields a more complex model as opposed to the sparser L1 model.

As for the size of the penalty, "C," I used GridSearchCV(), a useful function from sklearn that iterates through different parameter options and performs cross-validation on each potential combination of parameters. I gave the function a diverse range of possible parameters, beginning with C=0.001 and ending at C=1000, and found that C=1 yielded the most accurate model.

As for performance, despite its simplicity, the logistic regression model did fairly well, as the following results demonstrate:

	Accuracy	Specificity	AUC Score
Training	0.7107512370670266	0.5940282870612886	
Test	0.7072402938090241	0.5918367346938775	0.6898577612863327



While the model had an acceptable accuracy rate, its specificity rate was lower than I preferred.

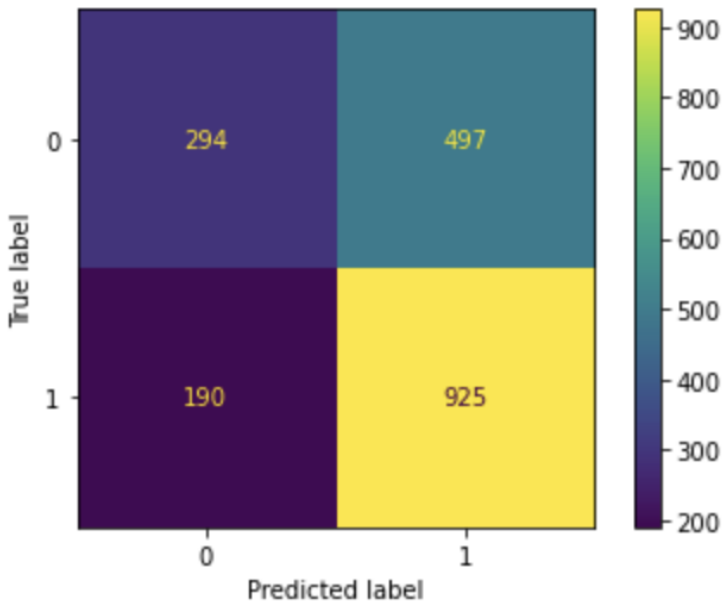
K-Nearest Neighbors

For k-nearest neighbors, the one hyperparameter I chose was the number of neighbors, "n_neighbors." Unlike with logistic regression where I iterated through every hyperparameter

value in a small list of options using GridSearchCV, for knn, I used another sklearn function, RandomizedSearchCV, which randomly chooses a specified number of hyperparameter options from a list of hyperparameters. This allowed me to create a larger list of options while reducing the function runtime. More specifically, my list contained all integer values from 1 to 100, and from this list, the RandomizedSearchCV function chose 50 to test with 10-fold cross-validation. N_neighbors = 39 proved to yield the most accurate model.

For performance, while the accuracy rate was close to the logistic regression, the specificity and AUC score were far lower, leading me to leave out the model in the final model ensembling method.

	Accuracy	Specificity	AUC Score
Training	0.6617183985605039	0.4043112513144059	
Test	0.6395592864637986	0.3716814159292035	0.6006389142426287



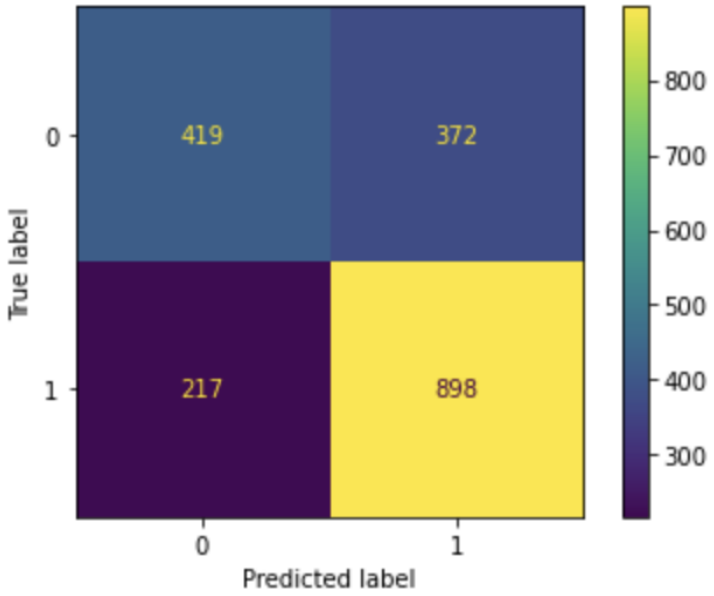
It appears that the model was biased to predicting positive labels, athwart to the philosophy underpinning this study.

Decision Trees

For my decision trees classifier, I chose two hyperparameters: min_samples_leaf_values and min_samples_split_values, the former specifying the minimum number of samples required for a node to qualify as a leaf node, and the latter, the minimum number of samples required for a split in that node. The leaf parameter takes precedence over the split parameter, meaning that if a split would result in one node having fewer samples than min_samples_leaf_values, then the split is not performed even if it meets the conditions of min_samples_split_values.

For this model, I chose to use GridSearchCV once again because I had some idea of what values I could use for the hyperparameter options, keeping my options list as small as possible. This was important because, since I was running the function on two sets of hyperparameter options simultaneously, I would be making len(leaf_values) x len(split_values) iterations, one for each leaf-split pairing. In order to reduce the runtime further, I lowered the number of folds in my k-fold cross-validation from 10 to 5. A combination of min_samples_leaf_values = 20 and min_samples_split_values = 300 produced the most accurate model. The results are as follows:

	Accuracy	Specificity	AUC Score
Training	0.7141250562303194	0.5662460567823344	
Test	0.6909758656873033	0.5297092288242731	0.6675451973717778

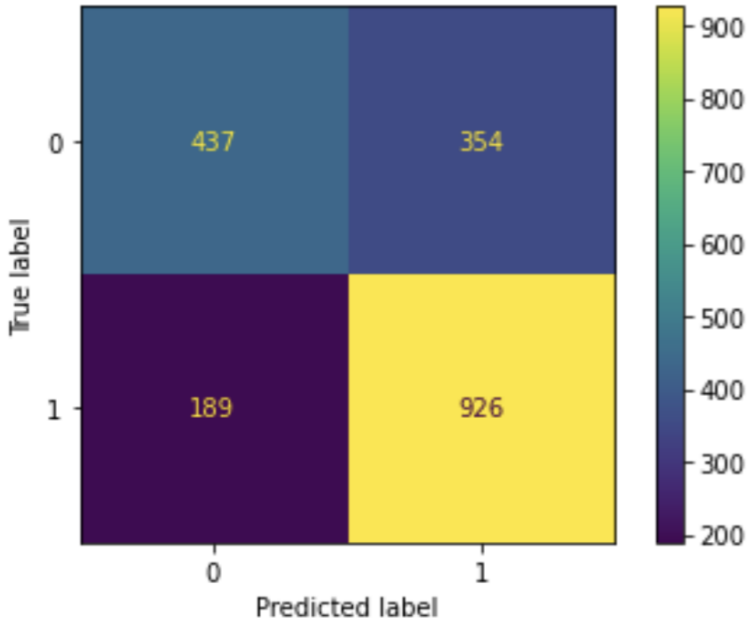


While the decision trees model did fairly well with accuracy, like the previous two models, it did far worse with specificity, resulting in a lower AUC score as well.

Random Forest

For my random forest classifier, I found that my hyperparameter choices for the decision trees model, when coupled with hyperparameters unique to random forest, did not produce the most accurate results. Moreover, rather than the minimum number of samples for leaves and splits, through trial and error, "n_estimators" (the number of decision trees in the random forest) and "max_depth" (the maximum depth of the random forest) proved to have a more significant influence on model performance. Thus, in order to minimize runtime and maximize model performance, I used RandomizedSearchCV to find the optimal hyperparameter values. Training a random forest model, because it contains multiple decision trees, took far longer than the single decision tree model, which is why I used RandomizedSearchCV instead of GridSearchCV. Doing so, I found that n_estimators=200 and max_depth=60 yielded the most accurate model. Below are the results:

	Accuracy	Specificity	AUC Score
Training	1.0	1.0	
Test	0.7151101783840503	0.5524652338811631	0.691479253711882



While the model was certainly overfitting to the training data, it still performed moderately well for the test data. Again, we see that the performance for specificity is not very high, yet the high accuracy performance raises the model's AUC score to second best, only 0.005 behind logistic regression.

Neural Networks

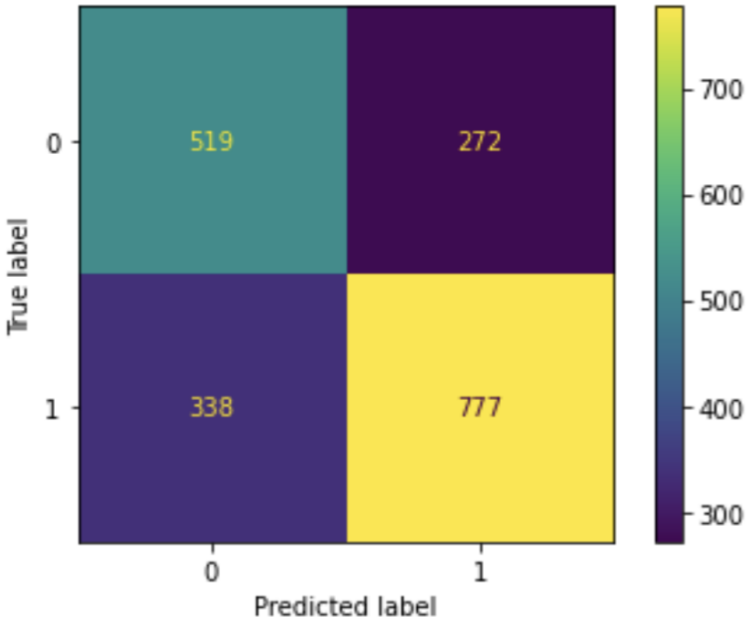
Finally, the neural network, the most complex model of the five, began with a different approach: I first had to preprocess the training data, making sure that all variables were the right datatype, and then further split my training data into a smaller training dataset and a validation

set (using the same 70-30 ratio used in the original train-test split). Keras, the package I used to build the neural network, is not compatible with sklearn's GridSearchCV or RandomSearchCV, so rather than automate the fine-tuning process, I manually changed the number of hidden dense layers and nodes per layer, trying out different combinations until the model produced satisfactory results. Because my task was binary classification (whether or not someone recidivates within three years), I used BinaryCrossentropy as my loss function and set the activation function to sigmoid, a standard choice for binary classification.

Unlike my other models that output binary labels, the neural network produced an array of probabilities of a positive label (how likely is each person to recidivate within three years). My first approach was to relabel the output by setting values < 0.5 to 0 and values >= 0.5 to 1. However, I discovered that by raising the probability threshold, I could make a positive classification more difficult and raise the model's specificity score while only minimally impacting accuracy. Ultimately, I decided to raise the threshold to 0.6, meaning that the model would only predict a person would recidivate within three years if there was at least a 60% chance of it happening.

Training my model with these parameters over 20 epochs and using this conversion approach produced the following results:

	Accuracy	Specificity	AUC Score
Training	0.7095115681233933	0.6907294832826748	
Test	0.6799580272822665	0.6561314791403287	0.676496232843707



While its accuracy and AUC score weren't the highest, this neural network's specificity score was the best out of all five models, outperforming the second-best score by nearly 7 percentage points and the median score by ~12.5 percentage points.

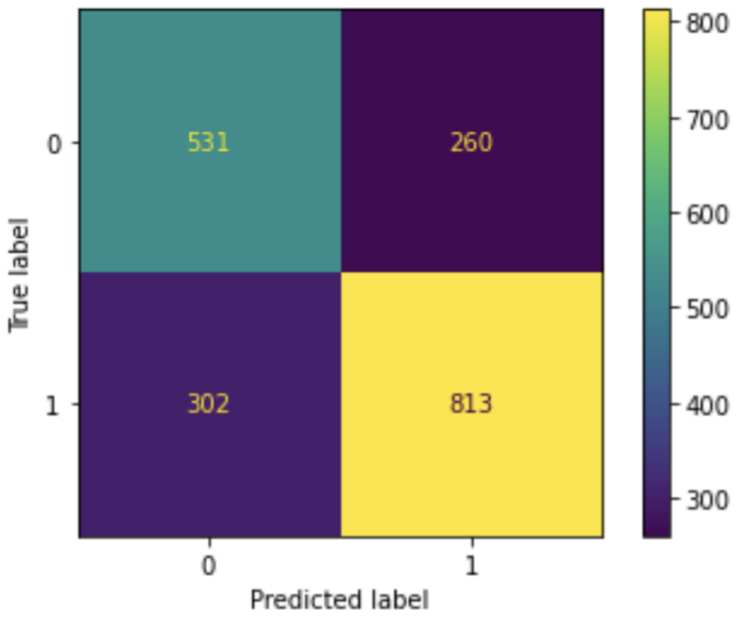
Stage 1: Final results (Test data)

Model	**Accuracy**	**Specificity**	**AUC**
lgr	0.715110	0.589128	0.696806
knn	0.639559	0.371681	0.600639
dtc	0.690976	0.529709	0.667545
rfc	0.715110	0.552465	0.691479
nn	0.679958	0.656131	0.676496

Ensembling (lgr, dtc, rfc, nn)

My final model is a simple vote-based system by the four most successful models: logarithmic regression, decision trees, random forests, and neural networks. In this second ensembling stage, since I wanted to prioritize specificity, I set up the voting system to require at least three models making a positive prediction for the ensemble model to also make a positive prediction, making the model more likely to catch true negatives. The following results are for the test data:

	Accuracy	Specificity	AUC Score
Test	0.7051416579223505	0.6713021491782554	0.7002250656205178



As stated previously, the ensemble method scored the highest in both AUC and specificity and performed third-best in accuracy, trailing just 1 percentage point behind the best-performing model.

Purpose

With these predictive models, we can reasonably identify which inmates might be at higher risk of recidivism, allowing us to focus greater attention and post-release support for these prisoners. While it is unjust to scale punishment with this recidivism likelihood—as punishment should be proportional to the crime committed and not the probability that it happens again—it is both economically efficient and proper to adjust aid in this way. We thus fulfill the first goal of this study.

Analysis

Feature Importance

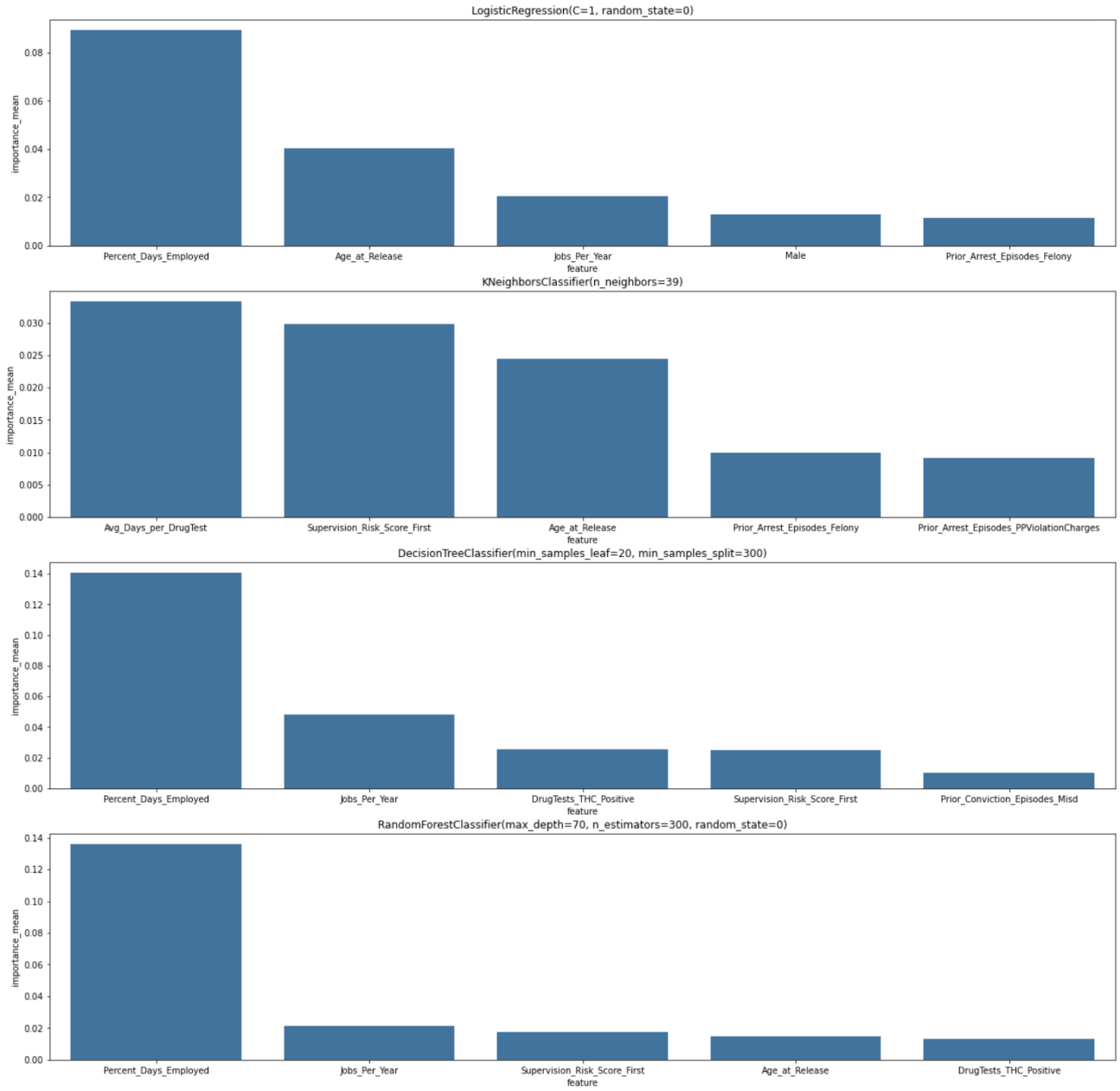
This now leads us to our second goal: Identifying the factors that engender recidivism. We do this with sklearn's permutation_importance function, calculating the importance of each feature as measured by the mean loss in the model's R-squared score when the feature is dropped from the model over a number of iterations. We perform feature selection in this way for all models except for the neural network which is incompatible with the permutation_importance function. Doing so yields the following top-five features for each model:

LogisticRegression
Percent_Days_Employed: 0.089 +/- 0.005
Age_at_Release: 0.040 +/- 0.006
Jobs_Per_Year: 0.021 +/- 0.003
Male: 0.013 +/- 0.002
Prior_Arrest_Episodes_Felony: 0.011 +/- 0.003

KNeighborsClassifier
Avg_Days_per_DrugTest: 0.033 +/- 0.005
Supervision_Risk_Score_First: 0.030 +/- 0.004
Age_at_Release: 0.024 +/- 0.004
Prior_Arrest_Episodes_Felony: 0.010 +/- 0.003
Prior_Arrest_Episodes_PPViolationCharges: 0.009 +/- 0.003

DecisionTreeClassifier
Percent_Days_Employed: 0.140 +/- 0.006
Jobs_Per_Year: 0.048 +/- 0.005
DrugTests_THC_Positive: 0.026 +/- 0.003
Supervision_Risk_Score_First: 0.025 +/- 0.003
Prior_Conviction_Episodes_Misd: 0.010 +/- 0.002

RandomForestClassifier
Percent_Days_Employed: 0.136 +/- 0.004
Jobs_Per_Year: 0.021 +/- 0.002
Supervision_Risk_Score_First: 0.017 +/- 0.001
Age_at_Release: 0.015 +/- 0.001
DrugTests_THC_Positive: 0.013 +/- 0.001

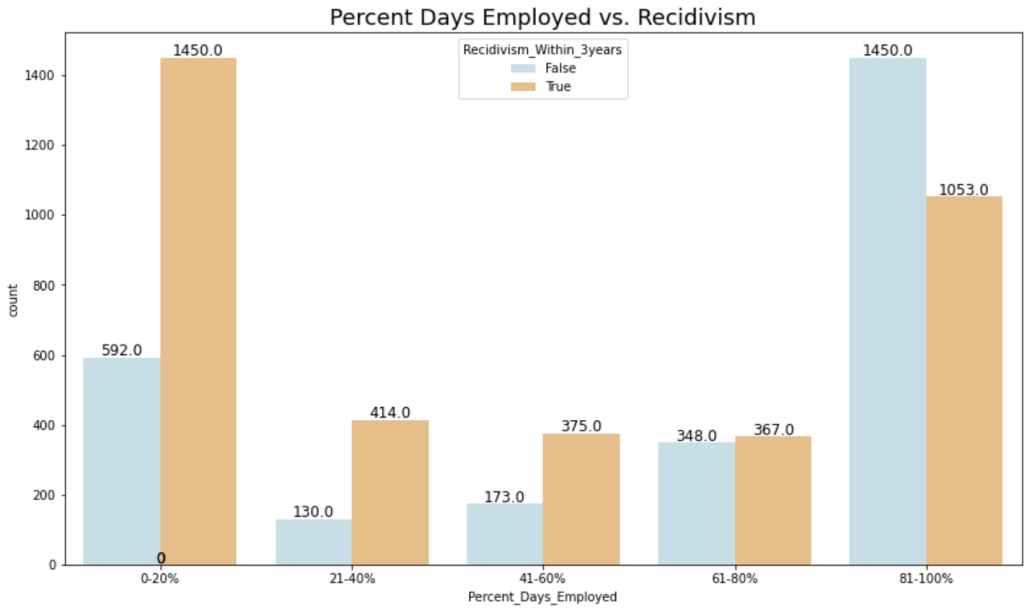


As highlighted in red-colored text, in three of the four models, Percent_Days_Employed and Jobs_Per_Year ranked in the top three most important features—often in the top two, with Percent_Days_Employed ranking first in each of the three models. Indeed, large bodies of research have linked lower employment/job opportunities to crime, with studies showing pre-prison employment rates lower than 35%, and the rate of prisoners having at least an associate's degree, four times lower than that of the average adult (Duwe and Henry-Nickie).

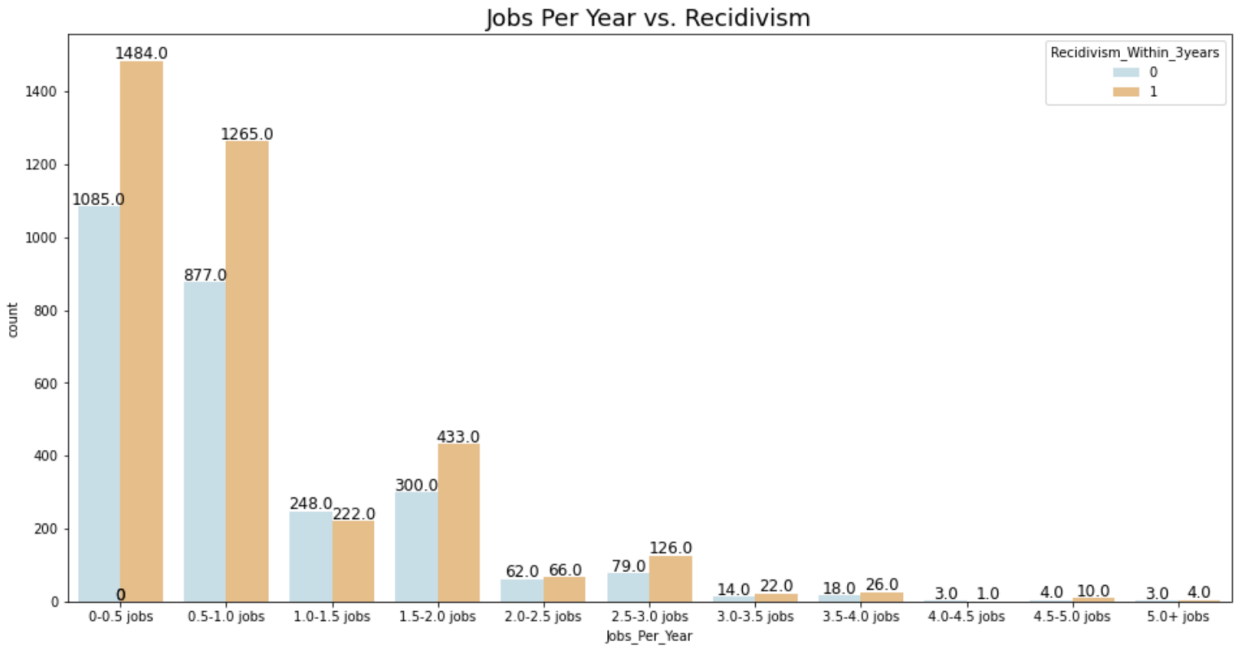
Using an economic approach, these two variables indicate the opportunity cost of committing crime: If someone is employed for 100% of the days in a year, they most likely have greater job prospects than those employed for a lower percentage of the year. Moreover, with Jobs_Per_Year: If a person is recorded to have a single job in a year, it reveals both that, one, they are able to *obtain* a job (jobs are relatively plentiful) and two, that they are able to *hold* a job (they are good workers and/or that the particular job is stable). By committing a crime and going to prison, one forfeits their current job and also makes it far more difficult to find another in the future. Thus, with these two variables, we have a measure of crime's opportunity cost.

Although Percent_Days_Employed and Jobs_Per_Year are the two most obviously relevant features to measuring crime's opportunity cost, variables such as Prior_Arrest_Episodes_Felony and other measurements of criminal history can also add to this cost. While higher scores might indicate the criminality of an inmate, they also lower the opportunity cost of crime by making it harder for the inmate to find a job upon release, making crime that much more attractive.

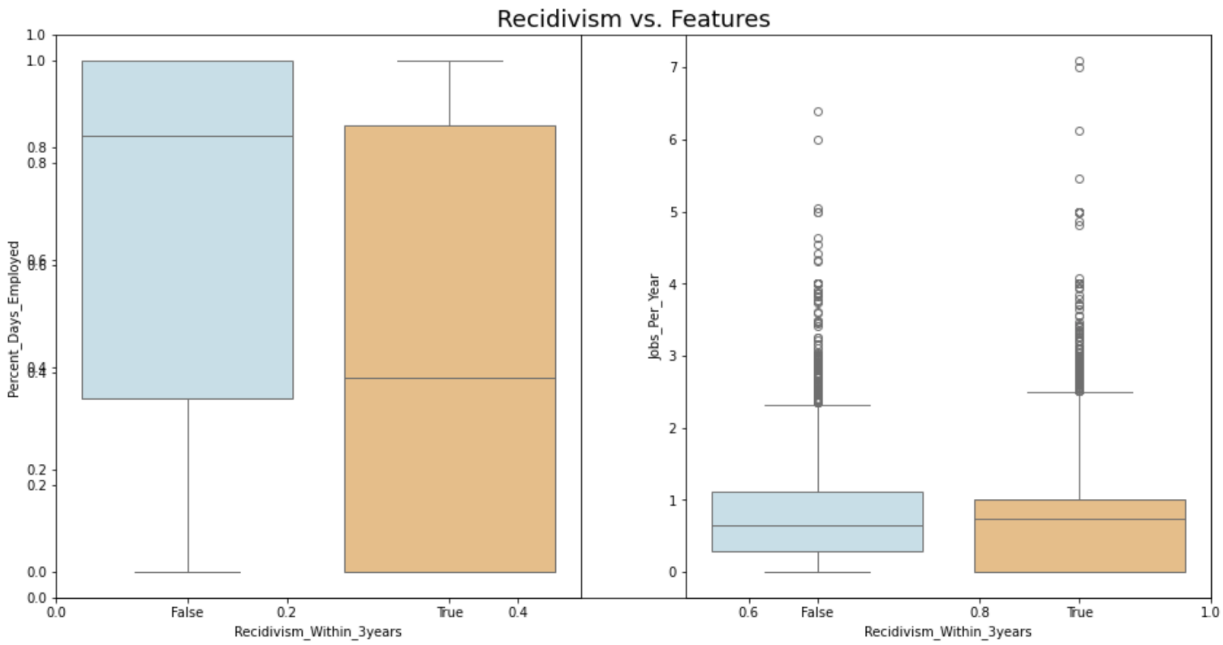
We further investigate these first two variables, Percent_Days_Employed and Jobs_Per_Year, and their relationship to recidivism.



First looking at this grouped bar chart, "Percent Days Employed vs. Recidivism," we see that those who are employed less than 60% of the year are far more likely to recidivate within three years than to stay out of prison. However, in the 61-80% range, the distribution of those who do and do not recidivate is far more evenly spread. At the 81-100% level, we see that the trend reverses, and recidivism becomes the rarer outcome. This confirms what we previously hypothesized, that higher job prospects, as reflected in higher percentages of days employed, make crime more costly and therefore deter crime.



This second graph, "Jobs Per Year vs. Recidivism," likewise tells a similar story to our hypothesis. Here, we find that the number of jobs a person has per year is not necessarily directly related to lower rates of recidivism: Those who hold less than one job per year are more likely to recidivate than not, and so is true for those who with more than 1.5 jobs. However, only those in the 1-1.5 jobs range have a higher proportion of non-reoffenders than reoffenders. Again, we posit that holding a single job for a year signals the highest job prospects and concomitantly the highest opportunity cost of committing crime. Thus, we would expect as we observe, that those in this 1-1.5 jobs category have the lowest recidivism rate.



Swapping the x and y axes of the last two graphs, we more clearly observe differences between those who do and do not recidivate in their percent of days employed and number of jobs held per year: The median reoffender is employed for far fewer days in a year than the median non-reoffender. Looking at the second graph, however, we lose sight of our original hypothesis, as the median number of jobs per year seems to be around 0.75 for both those who do and do not recidivate.

As our results show, while there is certainly complexity in the relationships between feature variables and the recidivism-rate target, it is clear that the life outside of prison also has a significant association with repeat criminal activity, expanding the arsenal of solutions legislatures and social workers have in combatting criminal activity.

Bibliography

Duwe, Grant, and Makada Henry-Nickie. “A better path forward for criminal justice: Training and employment for correctional populations | Brookings.” *Brookings Institution*, https://www.brookings.edu/articles/a-better-path-forward-for-criminal-justice-training-and-employment-for-correctional-populations/. Accessed 8 December 2023.

“Most prisoners per capita by country 2023.” *Statista*, 22 August 2023, https://www.statista.com/statistics/262962/countries-with-the-most-prisoners-per-100-000-inhabitants/. Accessed 8 December 2023.

Petrich, Damon M., et al. “Custodial Sanctions and Reoffending:A Meta-Analytic Review.” *Crime and Justice*, vol. 50, no. 1, 2021, https://www.journals.uchicago.edu/doi/epdf/10.1086/715100.

“Recidivism Rate by State 2023.” *Wisevoter*, 2023, https://wisevoter.com/state-rankings/recidivism-rates-by-state/. Accessed 8 December 2023.

“Reported violent crime rate in the U.S. 2022.” *Statista*, 20 October 2023, https://www.statista.com/statistics/191219/reported-violent-crime-rate-in-the-usa-since-1990/. Accessed 8 December 2023.

Warren, Jenifer. “New Analysis Shows U.S. Imposes Long Prison Sentences More Frequently than Other Nations.” *Council on Criminal Justice*, 20 December 2022, https://counciloncj.org/new-analysis-shows-u-s-imposes-long-prison-sentences-more-frequently-than-other-nations/. Accessed 8 December 2023.

Wilson, David B., et al. “National Institute of Justice’s Forecasting Recidivism Challenge: Team “DEAP” (Final Report).”