Predicting Recidivism in the U.S. Prison Population

Seth Taylor, Viola Hilbert, Shashank Shekhr Rai

May 8, 2017

Introduction

Risk assessment has a long history in the United States judicial system, with the first attempts at structured, qualitative risk assessment being employed as early as 1923. The history of quantitative risk assessment is much shorter. The first quantitative risk assessment tool was constructed in 1991, based upon multivariate regression. This tool has been superseded in the federal system by the Post-Conviction Risk Assessment (PCRA), which uses a combination of factors fom criminal history, education/employment, substance abuse, social networks, and cognitions in a multivariate analsyis to generate a risk measurement of low, low/moderate, moderate, or high risk. The AUC for the PCRA has been shown to range between .709 and .783.

This project analyzes correlations between recidivism and risk factors such as socio-economic characteristics, mental health, family background, drug use, and prior offenses to predict the probability of recidivism. We compare several classification methods against each other, including binomial GLM, K-nearest neighbors, decision trees, and random forests to determine the best predictive method. We also ues the results from our GLM model to examine what factors are most importan in predicting recidivism, and to draw policy conclusions as to how the risk of recidivism might be lowered.

Data

The data used in this project is retrieved from the 2004 Survey of Inmates in State Correctional Facilities (SISCF) and the 2004 Survey of Inmates in Federal Correctional Facilities (SIFCF). These surveys, collectively referred to as the 2004 Survey of Inmates in State and Federal Correctional Facilities (SISFCF), provide nationally representative data on inmates held in state prisons and federal prisons for the year 2004. Collected through personal interviews conducted from October 2003 through May 2004, the data captures information about prisoners' current offense and sentence, criminal history, family background, socio-economic characteristics, prior drug and alcohol use and treatment programs, gun possession and use, as well as prison activities, programs, and services.

We combine the data from the federal analysis and state analysis datasets for our analysis, which provides information on 14,499 prisoners in the state correctional system and 3686 prisoners in the federal correctional system.

Since this data is survey data, it requires further cleaning and recoding to ensure it is appropriate for analysis and to handle missigness (coded in the survey as values from 999997 to 999999). We use the package "memisc", which is designed to handle survey data.

Descriptive Statistics

Our descriptive statistics show that about 68 percent of our sample are recidivists, 37 percent committed a violent offense, 27 percent committed a drug offense, and 20 percent committed a propery crime. The average sentence for inmates is 132 months or 11 years, and roughly 80% of those inmates reside in state prisons.The average number of prior arrests for prisoners in our sample is 5. The most common crime committed is drug trafficking, followed by robbery, drug possession, and assault.

Prisoners are most commonly between 25 and 44 years old, with relatively few prisoners above the age of 55. Prisoners are most commonly highschool educated to some degree. About 71 percent of prisoners

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
CH_CRIMHIST_COLLAPSED	13,260	0.70	0.46	0	1
OFFENSE_VIOLENT	13,260	0.37	0.48	0	1
OFFENSE_DRUG	13,260	0.27	0.45	0	1
OFFENSE_PROPERTY	13,260	0.20	0.40	0	1
SES_PHYSABUSED_EVER	13,260	0.19	0.39	0	1
CS_SENTENCEMTH	13,260	132.29	174.47	0	4,440
SES_PARENTS_INCARCERATED	13,260	0.17	0.38	0	1
SES_FAMILY_INCARCERATED	13,260	0.38	0.49	0	1
SES_HASCHILDREN	13,260	0.71	0.45	0	1
SES_SEXABUSED_EVER	13,260	0.12	0.33	0	1
DRUG_ANYREG	13,260	0.73	0.44	0	1
DRUG_ANYTME	13,260	0.34	0.47	0	1
black.nh	13,260	0.38	0.49	0	1
hispanic	13,260	0.19	0.39	0	1
asian	13,260	0.01	0.10	0	1
state	13,260	0.79	0.41	0	1
SES_FATHER_INCARCERATED	13,260	0.17	0.38	0	1
DRUG_COCRKTME	13,260	0.13	0.33	0	1
DRUG_HROPTME	13,260	0.05	0.22	0	1
DRUG_METHATME	13,260	0.07	0.26	0	1
GENDER	13,260	1.22	0.41	1	2
DRUG_MARIJTME	13,260	0.15	0.36	0	1
CH_PRIORARREST_CAT	13,260	4.98	7.92	0	99
SES_LIVE_CHILD_ARREST	13,260	0.27	0.44	0	1
DRUG_ABUSE_ONLY	13,260	0.17	0.38	0	1
DRUG_TRT	13,260	0.62	0.49	0	1

Table 2: Descriptive Statistics: Factor Variables

model.data.EDUCATION	model.data.AGE_CAT	model.data.TYPEOFFENSE
1	model.data.EDUCATION	0000012:3407
2	model.data.EDUCATION	0000011:2448
3	model.data.EDUCATION	0000010:2101
4	model.data.EDUCATION	0000009:1544
5	model.data.EDUCATION	0000014: 826
6	model.data.EDUCATION	0000008: 742
7	model.data.EDUCATION	(Other):2192
8	model.data.AGE_CAT	(1) <25 yrs:2106
9	model.data.AGE_CAT	(2) 25-34 :4486
10	model.data.AGE_CAT	(3) 35-44 :4174
11	model.data.AGE_CAT	(4) 45-54 :1915
12	model.data.AGE_CAT	(5) 55-64 : 485
13	model.data.AGE_CAT	(6) 65-96 : 94
14	model.data.AGE_CAT	(7) Unknown : 0
15	model.data.TYPEOFFENSE	(0000017) Drug traffic :2217
16	model.data.TYPEOFFENSE	(0000006) Robbery :1629
17	model.data.TYPEOFFENSE	(0000016) Drug possession :1250
18	model.data.TYPEOFFENSE	(0000007) Assault :1120
19	model.data.TYPEOFFENSE	(0000021) Other public order:1064
20	model.data.TYPEOFFENSE	(0000009) Burglary : 869
21	model.data.TYPEOFFENSE	(Other) :5111

have children. 38 percent of our sample is black, about 19 percent is hispanic, and about 1% is asian. The remaining 42 percent is categorized as white or other race. Roughly 22 percent of our sample is female.

About 19 percent of our sample reported ever being physically abused, and 12 percent reported ever being sexually abused. About 17.5 percent of prisoners have at least one parent who has been incarcerated (this is equivalent to the father incarcerated variable suggesting this is commonly the father who is incarcerated and not the mother), and 38 percent have family members who have been incarcerated. Drug use is common among prisoners, with 73 percent having regularly used any drug, and 34 percent of prisoners using drugs at the time of arrest. The most common drug used at the time of arrest was marijuana, followed by crack coccaine, meth, and heroin. 61.5 percent of our sample has been in a drug treatment program at least once, with 17 percent classifying themselves as a drug abuser.

Methodology

To predict recidivism, we compare binomial GLM, KNN, decision tree, and random forest using mean-F1 as our measure of accuracy. Our GLM approach partitions the data as a 70/15/15 train/validate/test while the rest of our methods use k-folds cross validation (k=5) as an alternative to the 70/15/15 partition.

Due to the fact that this data comes from only a single year survey, it is not possible to predict whether or not an individual will become a recidivist in the future. To do this you would hae to follow the same individuals over time and observe which become recidivists and which do not. Instead, what we are predicting is whether an individual already in the system is a redividist or someone new to the correctional system.

GLM

We test two cutoffs for prediction in our GLM. First, we default to 0.5 as our cutoff, which produces high sensitivity, but very low specificity. While high sensitivity is good, we feel that erroneously predicting non-recidivists as recidvists could be very harmful. Therefore, we move to a cutoff of .6, which provides a better balance between sensitivity and specificity as shown in the confusion matrixes below.

	0.5 Cutoff	0.6 Cutoff
Train	0.8017235	0.7963942
Validate	0.7943777	0.7942008
Test	0.7947043	0.7924353

	GLM Accuracy Measures
Accuracy	0.1630670
True Positive Rate	0.8275862
True Negative Rate	0.7024221
Positive Predictive Value	0.8652038

The mean F1 for our test set at the 0.6 cutoff is shown to be .794, with a sensitivity of .828 and a specificity of .702. However, accuracy for this model is very low at .16.

GLM: Marginal Effects

We also run our GLM model on the full dataset and calculate the marginal effects at means for each variable.

Our model shows that the primary factors that influence recidivism are drug use, the type of offense committed, prior criminal history, and family criminal history. Education and length of setnence are both not signfiicant. Prisoners who committed a violent crime are 7 percentage points less likey to be a recidivist. If the offense was a property crime, then the probability rises by about 9.7 percent. The type of offense variable is in comparison to murders, and shows that individuals who commit kidnapping, sexual assault, robbery, assualt, or other violent crimes are between 3 and 8 percentage points more likely to be recidivists when compared against murderers. Individuals sentenced for drug possession, weapon crimes, DWIs, and other public order offenses are also between 7 and 8.7 percentage points more likley to be recidivists when compared against murderers. Finally, for every prior arrest an individual has, their probability of being a recidivist increases by 4.6 percentage points.

If an individuals parents were incarcerated, then they are about 1.7 percentage points more likely to be a recidivist, and if anyone in their family was incarcerated they are about 2.4 percentage points more likely to be a recidivist.

Age is consistently a positive factor for recidivism, except for the highest age category which is only significant at the 10 percent level. This is likely due to the fact that younger prisoners have more time to re-enter the system. People between the ages of 65-96 are less likely to be recommitting offenses.

Drug use is also a prominent factor in recidivism. Individuals who regularly did drugs were about 5 percent more likely to be recidivists. While individuals who were using any drug at the time of arrest are about 2.5 percentage points less likely to be recidivists, individuals who were using heroin, crack, or methamphetamines are between 3 and 4 percentage points more likley to be recidivists. This suggests that harder drug use is more common among recidivists while drug use like marijuana is more common among first offenders. Individuals who had been in drug treatment programs are about 4 percentage points more likely to be recidivists. This is likely to be a reverse causality in that recidivists are more often hard drug users and therefore more likely to

Table 5: Logistic Regression: Marginal Effects

	dF/dx	Std. Err.	z	P>|z|
OFFENSE_VIOLENT	-0.070	0.033	-2.112	0.035
OFFENSE_DRUG	-0.001	0.034	-0.016	0.987
OFFENSE_PROPERTY	0.090	0.034	2.619	0.009
CS_SENTENCEMTH	0.00002	0.00002	1.101	0.271
SES_PARENTS_INCARCERATED	0.019	0.007	2.797	0.005
SES_FAMILY_INCARCERATED	0.027	0.006	4.599	0.00000
SES_HASCHILDREN	0.010	0.006	1.615	0.106
AGE_CAT(2) 25-34	0.060	0.008	7.355	0
AGE_CAT(3) 35-44	0.089	0.009	9.770	0
AGE_CAT(4) 45-54	0.074	0.008	8.700	0
AGE_CAT(5) 55-64	0.053	0.011	4.876	0.00000
AGE_CAT(6) 65-96	0.034	0.022	1.523	0.128
SES_SEXABUSED_EVER	-0.007	0.009	-0.759	0.448
DRUG_ANYREG	0.057	0.009	6.416	0
stateTRUE	0.069	0.010	7.199	0
GENDER	-0.064	0.010	-6.481	0
DRUG_COCRKTME	0.035	0.010	3.572	0.0004
DRUG_HROPTME	0.027	0.013	2.021	0.043
DRUG_ANYTME	-0.023	0.009	-2.563	0.010
DRUG_METHATME	0.039	0.011	3.570	0.0004
CH_PRIORARREST_CAT	0.047	0.001	46.114	0
TYPEOFFENSE(0000002) Manslaughter	-0.015	0.025	-0.614	0.539
TYPEOFFENSE(0000003) Kidnapping	0.047	0.019	2.416	0.016
TYPEOFFENSE(0000004) Rape	0.021	0.015	1.404	0.160
TYPEOFFENSE(0000005) Other sexual assualt	0.032	0.013	2.549	0.011
TYPEOFFENSE(0000006) Robbery	0.077	0.010	7.513	0
TYPEOFFENSE(0000007) Assault	0.059	0.011	5.407	0.00000
TYPEOFFENSE(0000008) Other violent	0.067	0.017	3.900	0.0001
TYPEOFFENSE(0000009) Burglary	-0.044	0.051	-0.862	0.389
TYPEOFFENSE(0000010) Larceny	-0.032	0.050	-0.633	0.527
TYPEOFFENSE(0000011) Mvt	-0.048	0.060	-0.795	0.427
TYPEOFFENSE(0000012) Arson	-0.092	0.074	-1.250	0.211
TYPEOFFENSE(0000013) Fraud	-0.070	0.055	-1.269	0.204
TYPEOFFENSE(0000014) Stolen property	0.007	0.049	0.152	0.879
TYPEOFFENSE(0000016) Drug possession	0.040	0.018	2.225	0.026
TYPEOFFENSE(0000017) Drug traffic	0.020	0.019	1.054	0.292
TYPEOFFENSE(0000019) Weapons	0.087	0.017	5.249	0.00000
TYPEOFFENSE(0000020) Dwi	0.084	0.019	4.528	0.00001
TYPEOFFENSE(0000021) Other public order	0.068	0.019	3.506	0.0005
DRUG_TRT	0.043	0.007	6.558	0

have ever gone to drug treatment programs, since it is unclear why drug treatment programs would make you more likely to commit a crime.

Finally, women are about 7 percentage points less likely to be recidivist. Those individuals in state prisons are about 6 percentage points more likely to be recidivists. This is likely due to the fact that the average individual is less likely to reguarly commit federal crimes, so those in federal prison are more often first offenders.

KNN

In a next step, we compare GLM predictions to the K-Nearest-Neighbors (KNN) learning Algorithm. KNN is a non-parametric pattern recognition algorithm that measures Euclidean distances in terms of input features for any given record i to all other records in a dataset. The k records with the shortest distance to that point i build the neighborhood from which the value of y(i) can be approximated, because observations that are more similar will likely also be located in the same neighborhood.

In the following, we present our best prediction based on KNN. To see six further KNN models we ran, which include different variables for x and different parameter specifications, please see "1-KNN-Predictions.R" in the Github repository "https://github.com/GeorgetownMcCourt/Predicting-Recidivism/tree/master/Scripts".

Decision Tree

Decision trees classify by taking each input in the model, and splitting the sample based on the variable that most significantly differentiates the sample into the target categories (in our case recidivist or not recidivist). The tree is split into internal nodes, which are the features used to split the set, and leaf nodes which has the class label based on how the internal node above it split the set. The decision tree branches out through the variables creating leaf nodes for each of the internal nodes.

Decision trees tend to overfit when grown very deep, which requires pruning the tree to the optimal level. Our decision tree below shows out best result, but all decision trees can be found at https://github.com/GeorgetownMcCourt/Predicting-Recidivism/blob/master/Scripts/2-DT-Predictions.R.

Random Forest

The random forest approach constructs many decision trees and produces a final classification based on the mode of the classification from the decision trees. While decision trees that are grown very deep have a tendency to overfit, random forest corrects for this through its ensemble approach.

As with our decision tree, the results from our best random forest are shown below but all random forests can generated from the script at https://github.com/GeorgetownMcCourt/Predicting-Recidivism/blob/master/Scripts/3-RF-Predictions.R.

Model.Mean.F1	train	test	validate	full
KNN	0.8592319	0.7535503	0.7564198	0.8613525
Decision Tree	0.8403324	0.8403200	0.8388787	0.8401167
Random Forest	0.9536978	0.8404565	0.8375947	0.9413880

Model.Accuracy.Measure	KNN	Decision.Tree	Random.Forest
accuracy	0.8647813	0.8539216	0.9429864
TPR	0.9361105	0.9881286	0.9950356
TNR	0.6992989	0.5425638	0.8222334

Model.Accuracy.Measure	KNN	Decision.Tree	Random.Forest
PPV	0.8783797	0.8336520	0.9284995

As shown in our mean-F1 table, random forest and decision tree produce very similar results when run on our test set (.837 and .838 mean-F1 respectively), but random forest is far superior to decision tree when re-running the model on the full dataset after testing with a mean-F1 of .94 compared to .84. All 3 methods produce superior results in terms of predictive accuracy to our GLM model, but KNN and GLM remain close in terms of mean-F1 score.

Random forest also shows the best results for accuracy, true positive rate, true negative rate, and positive predictive value with rates of .943, .995, .822, and .928 respectively. Decisions trees are particularly bad at specificity with a true negative rate of .54.

Policy Recommendations

Our GLM analysis shows us that the most important factors influencing recidivism are type of offense, past criminal history, family criminal history, and drug use. Of these factors, drug use is the one that most lends itself towards credible policy recommendations. We see from our analysis that the use of hard drugs at the time of arrest is significantly associated with recidivism, and it also seems that drug abuse is quite prevalent among recidivists since having participated in drug treatment is significantly associated with recidivism. It seems plausible then, that more robust options for drug treatment programs, particularly ones that target individuals before they enter the correctional system, could reduce recidivism. Additionally, for prisoners released on parole who have a past history of drug abuse, providing additional support for drug treatment and training parole officers in how to appropriately manage past drug abusers could help prevent recidivism.

Conclusion

Predictive algorithms for recidivism risk have already started being used in the U.S. court system (http://nyti.ms/2qjkV6Q), and have proved controversial. Although the results from our analysis are not highly predictive, they are useful in understanding how bias might affect the results of predictive algorithms. Due to the fact that our data comes from a single year survey, missingness from non-response presents a possibly significant source of bias in our results and our predictive results lack generalizability to the overall population.

Future efforts in predicting recidivism risk would require following the same individuals over time to examine how many become recidivists. Some organizations, such as UChicago's Center for Data Science and Policy, have already begun this work studying individuals at the county level in Illinois.