Modeling Parole and Conditional Release

An Application of Predictive Analytics Techniques

Arnold Johnsen

Northwestern University, B.A. in Statistics expected June 2019

ArnoldJohnsen2019@u.northwestern.edu

Faculty adviser: Dr. Ajit Tamhane

Abstract

The list of released prisoners disseminated by the National Corrections Reporting Program (NCRP) documents fifteen reasons for release. Eight of them—after discounting special cases such as executions or escapes—pertain to the categories conditional release or non-conditional release. Prisoners in the latter category serve the entirety of their sentence in prison, whereas conditional releasees are allowed to serve the remainder of their sentences under community supervision. Accurately predicting conditional release is of great consequence to studying social justice and defendants' rights, so in this paper I aim to illustrate how and to what degree different methods can improve prediction of conditional release. By analyzing missing values, state-to-state variations in parole rate, sampling methods, and different predictive models, I arrived at a useful practical guide for dealing with the NCRP data and a methodological outline for better predictive performance, both of which can serve as a foundation for more sophisticated analysis in the future.

1. Introduction, Definitions, and Literature Review

The response variable of interest, conditional release, is a category of prison release which allows the prisoner to serve the remainder of their sentence at home subject to supervision and certain *conditions*—violation of these conditions will result in the rearrest of that prisoner. Conditions usually include meeting with a parole officer and avoiding even misdemeanor offenses—even minor alcohol charges can result in revocation of conditional release. Parole and conditional release are more or less interchangeable: although supervised release is technically distinct, for modeling purposes I will consider the terms parole and conditional release to be interchangeable. Unconditional release, intuitively, means that the prisoner has finished their sentence physically in prison.

Parole or conditional release is determined in one of two points along the criminal justice track: first by a sentencing judge, who determines whether parole is permitted at all, after how long of a prison term it might be permitted, and if at some point during the sentence mandatory parole is implemented (i.e., the convict is paroled automatically at a certain point and does not go through the parole board process). Otherwise, dispensation of conditional release is regulated by a parole board, which determines in a deliberative, trial-like process in which board members assess whether individual prisoners are deserving of parole and whether they might pose a threat to society if released early. For example, in the state of Texas, parole board guidelines include a numerical scoring process where prisoners are assigned "risk" levels based on their criminal history, behavior while incarcerated, and demographic factors like age and gender. However, parole board members also "retain the discretion to vote outside the guidelines when the circumstances of an individual case merit their doing so" (Texas Board of Pardons and Paroles).

In other words, the dispensation of conditional release, despite growing tendency towards an actuarial approach (see Glaser, 1985) still maintains an arbitrary, human element.

This is why a statistical approach to the problem is valuable. While robust statistical methodologies have been developed for modeling parole outcomes (Alumbaugh et al., 1978; Glaser; Rhodes, 1986) and recidivism in general (Berk et al., 2009; Andersen and Wildeman, 2015), the study of conditional release decisions on their own lacks a hardcore statistical approach. These articles apply a variety of parametric and nonparametric methods to parole outcomes, but this approach fails to study those who are never granted parole in the first place. Even Glaser's article, which is titled "Who Gets Probation and Parole," in the end focuses more on who "should" get parole—the criteria for his evaluation of parole assignation methods is whether prisoners granted parole via those methods later violate their parole. Again, prisoners denied parole never even appear in his dataset. Methods for predicting conditional release from prisoner data are not necessarily similar, so a review of how different approaches at each step affect predictive performance is a valuable academic exercise. Meanwhile, a well-performing predictive model is of immense practical utility. Defense lawyers can use this model to advise clients and their families about the chances of getting parole, or defendants themselves can use it to decide if hiring a quality defense lawyer is worth it. To dig into these relationships and build a predictive model, we need data.

2. Data Preparation

The most complete data on this subject is available from the Bureau of Justice Statistics (BJS), which annually sponsors the National Corrections Reporting Program (NCRP). A

collection of about 15 million prisoner term records from 2000-2015 formed the basis for my analysis. Observed variables can be roughly organized into demographic characteristics and criminal attributes (e.g. length of longest sentence, previous felony incarceration, AWOL during current sentence), and of course the type of release granted, which is the response variable I am interested in modeling.

The most pressing data cleaning issue is missing values—reporting practices are not standard across states, so for many potential useful predictors there are entire states for which the value of that predictor is either missing for every prisoner or set to some arbitrary value. Therefore, a careful analysis of what values are missing by variable and by state is necessary to proceed; the results are summarized in the tables on the next two pages. The most intuitive presentation of this information is data "completeness": what percentage of the values in a given state-variable pair are *not* missing. We can see in a general sense that certain states perform better than others, but the more striking feature is the number of variables which are completely missing in many states—in order to use them at all, we would have to excise entire states from the data, since imputing across state lines seems unreasonable. Thus, variable selection and state selection must be balanced when analyzing NCRP data or similar reporting data; we want to keep as many potentially useful variables as possible, but not lose the national picture by looking at only a small sample of states that have relatively complete data.

After a long series of experiments with different cutoff points, comparing the predictive success and significance of individual predictors in various output models, I decided to use only variables that were present in 90% of the observations. This resulted in 11 predictors after discounting technical variables such as BJS offense code. These 11 predictors are summarized in the appendix Once again, an imputation scheme for an entire U.S. state of missing observations

is likely impossible, so I deleted all observations missing values in the 11 chosen variables. This cut the number of observations down to about 6 million, which is obviously a significant amount of data lost, but we can at least be confident in the quality of that data rather than relying on imputed data of questionable accuracy.

	AWOL	HIGHEST_GRADE	HISPANIC	PRIOR_FELONY	PRISON_ADMISSION_TYPE	PRISON_RELEASE_FROM	PRISON_RELEASE_TO_1	RACE	SENT_DETERMINATE	SENT_INDETERMINATE	SENT_MANDATORY_MINIMUM	SEX	TOTAL_SENTENCE	TOTAL
AL	0.0%	0.0%	48.2%	0.0%	93.5%	0.0%	0.0%	99.8%	0.0%	0.0%	0.0%	100.0%	100.0%	47.5%
AK	0.0%	31.9%	89.9%	0.0%	35.7%	99.9%	0.0%	97.3%	0.0%	0.0%	0.0%	100.0%	99.4%	44.5%
AZ	100.0%	99.4%	100.0%	95.5%	100.0%	100.0%	99.9%	82.5%	100.0%	100.0%	42.9%	100.0%	99.9%	79.2%
CA	97.9%	0.0%	97.9%	98.3%	100.0%	100.0%	97.9%	97.8%	99.1%	7.5%	0.0%	100.0%	100.0%	71.1%
CO	100.0%	90.6%	100.0%	100.0%	99.8%	100.0%	98.6%	68.9%	46.7%	46.7%	0.0%	100.0%	100.0%	83.2%
DC	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	43.0%
FL	100.0%	83.7%	97.1%	100.0%	99.9%	97.8%	67.9%	100.0%	100.0%	100.0%	99.2%	100.0%	100.0%	81.3%
GA	100.0%	99.7%	100.0%	100.0%	99.3%	100.0%	81.7%	97.8%	100.0%	100.0%	100.0%	100.0%	100.0%	83.0%
ID	0.0%	27.8%	91.5%	0.0%	52.2%	0.0%	0.0%	78.9%	70.0%	70.0%	0.0%	92.7%	70.3%	54.0%
IL	73.1%	88.9%	100.0%	100.0%	100.0%	99.9%	100.0%	96.8%	100.0%	100.0%	0.0%	100.0%	100.0%	74.9%
IN	0.0%	90.3%	99.4%	0.0%	90.6%	100.0%	0.0%	96.1%	0.0%	0.0%	0.0%	100.0%	100.0%	61.7%
IA	73.6%	95.2%	99.3%	0.0%	91.9%	81.9%	77.5%	99.5%	0.0%	0.0%	0.0%	99.8%	87.2%	55.1%
KS	0.0%	97.8%	99.8%	90.0%	100.0%	100.0%	0.0%	100.0%	99.1%	99.1%	0.0%	100.0%	98.7%	64.5%
KT	100.0%	93.6%	49.0%	100.0%	99.8%	97.9%	77.1%	99.9%	63.4%	63.4%	63.4%	100.0%	99.9%	82.2%
ME	0.0%	83.5%	99.3%	0.0%	100.0%	0.0%	0.0%	99.2%	0.0%	0.0%	0.0%	100.0%	99.4%	49.4%
MD	100.0%	0.0%	0.0%	0.0%	100.0%	100.0%	0.0%	98.3%	100.0%	100.0%	100.0%	100.0%	99.8%	65.5%
MA	100.0%	58.3%	96.4%	0.0%	100.0%	100.0%	86.8%	100.0%	0.0%	0.0%	99.5%	100.0%	99.5%	74.7%
MI	99.2%	97.6%	0.0%	99.7%	95.8%	100.0%	98.3%	99.8%	100.0%	100.0%	100.0%	100.0%	100.0%	81.3%
MN	0.0%	95.5%	100.0%	0.0%	100.0%	100.0%	99.9%	100.0%	95.7%	95.7%	42.5%	100.0%	100.0%	64.7%
MS	0.0%	94.8%	100.0%	0.0%	100.0%	0.0%	0.0%	99.2%	0.0%	0.0%	0.0%	100.0%	99.7%	49.1%
MO	95.0%	65.5%	95.0%	99.6%	100.0%	100.0%	100.0%	99.9%	99.2%	97.3%	99.2%	100.0%	97.9%	87.1%
MT	91.9%	0.0%	2.7%	0.0%	100.0%	91.7%	100.0%	98.6%	100.0%	100.0%	100.0%	100.0%	99.3%	71.9%
NE	65.8%	34.3%	99.4%	65.6%	96.1%	65.8%	63.6%	97.9%	30.4%	30.4%	30.4%	100.0%	100.0%	68.5%
NV	0.0%	96.7%	99.7%	100.0%	76.6%	100.0%	100.0%	98.8%	87.9%	87.9%	0.0%	100.0%	100.0%	76.0%
NH	100.0%	0.0%	100.0%	0.0%	100.0%	21.9%	57.7%	93.1%	100.0%	100.0%	100.0%	100.0%	97.7%	64.4%
NJ	0.0%	0.0%	96.2%	0.0%	100.0%	99.9%	92.5%	89.4%	95.8%	14.2%	14.2%	100.0%	99.2%	58.0%
NM	0.0%	0.0%	99.7%	0.0%	87.0%	0.0%	0.0%	99.7%	0.0%	0.0%	0.0%	100.0%	100.0%	43.6%
NY	0.0%	87.9%	99.1%	0.0%	100.0%	100.0%	0.0%	98.5%	100.0%	100.0%	0.0%	100.0%	100.0%	64.4%
NC	10.9%	99.8%	95.3%	100.0%	100.0%	89.1%	14.6%	99.7%	87.9%	87.9%	0.0%	100.0%	100.0%	71.5%
ND	0.0%	76.6%	99.8%	0.0%	99.6%	9.5%	0.0%	95.4%	0.0%	0.0%	9.6%	100.0%	96.5%	54.1%
OH	0.0%	0.0%	64.5%	100.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	100.0%	48.0%
OK	59.2%	89.0%	81.8%	59.2%	99.8%	100.0%	61.4%	99.8%	88.2%	82.4%	68.1%	100.0%	99.5%	80.0%
OR	100.0%	4.6%	100.0%	99.4%	46.9%	70.0%	73.6%	89.7%	72.4%	72.4%	73.0%	100.0%	99.9%	76.1%
PA	0.0%	85.7%	100.0%	0.0%	96.5%	0.0%	0.0%	88.3%	84.6%	87.2%	0.0%	100.0%	100.0%	61.7%
RI	0.0%	99.8%	100.0%	0.0%	100.0%	0.0%	0.0%	82.3%	0.0%	0.0%	0.0%	100.0%	100.0%	56.4%
SC	100.0%	97.9%	98.5%	97.3%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.8%	90.3%
SD	0.0%	88.7%	0.0%	0.0%	100.0%	0.0%	0.0%	95.8%	0.0%	0.0%	0.0%	100.0%	88.5%	43.8%
TN	100.0%	63.1%	99.9%	100.0%	99.7%	100.0%	66.5%	99.9%	75.1%	75.1%	75.1%	100.0%	100.0%	84.1%
TX	37.5%	85.8%	100.0%	81.6%	100.0%	100.0%	100.0%	81.8%	100.0%	100.0%	100.0%	100.0%	100.0%	82.0%
UT	100.0%	88.3%	91.2%	100.0%	100.0%	99.6%	100.0%	99.7%	0.0%	0.0%	0.0%	100.0%	99.9%	66.5%
WA	64.9%	59.4%	98.5%	64.9%	99.9%	100.0%	100.0%	99.7%	94.1%	94.1%	46.8%	100.0%	99.9%	78.0%
WV	0.0%	87.4%	91.5%	0.0%	100.0%	0.0%	0.0%	91.1%	0.0%	0.0%	0.0%	100.0%	98.3%	49.2%
WI	0.0%	95.3%	100.0%	0.0%	100.0%	99.9%	0.0%	99.9%	0.0%	0.0%	0.0%	100.0%	99.7%	60.7%
WY	100.0%	99.1%	100.0%	100.0%	100.0%	98.0%	96.2%	87.9%	98.9%	98.9%	100.0%	100.0%	99.9%	84.9%

Table 1: Data completeness of selected variables by state

3. Model Fitting

With data issues addressed, we can return to the original goal of finding a well-performing predictive model. While machine-learning techniques are attractive due to their predictive power, it is ultimately wise to avoid these for a couple of reasons. Firstly, the NCRP data is heavily protected because of the personal identifiers it contains, and disabling all network capabilities while the data is removed from its encrypted vault is a condition of its use. Thus, the cloud computing services needed to apply neural networks or similar models to data of this quantity are not an option for a project of this scale. Secondly, the opaqueness of these black box methods is well-known, and specifically within criminal justice contexts these machine-learning algorithms have acquired a bad reputation for unseen biases that make a well-defined parametric model attractive. With this in mind, I used logistic regression to classify conditional release versus unconditional release, and later employed random forests as a compromise model that increased performance but still maintains some semblance of interpretability.

Variable	% Present
BJS_OFFENSE_3	21.10%
OFFENSE_COUNT_2	21.50%
PAROLE_ELIG_DATE	28.10%
PRIOR_PRISON_TIME	34.00%
MAND_PRISON_RELEASE_DATE	36.60%
SENT_MANDATORY_MINIMUM	38.40%
BJS_OFFENSE_2	42.00%
OFFENSE_COUNT_1	50.60%
PRIOR_COMMUNITY_RELEASE	60.80%
HIGHEST_GRADE	61.90%
SENT_INDETERMINATE	62.00%
AWOL	62.10%
PROJ_PRISON_RELEASE_DATE	68.50%
PRISON_RELEASE_TO_1	68.70%
PRIOR_FELONY	72.50%
PRIOR_JAIL_TIME	80.90%
SENT_DETERMINATE	82.10%
PRISON_RELEASE_FROM	90.00%
HISPANIC	91.70%
OFFENSE_1_SENTENCE	93.70%
RACE	95.20%
PRISON_ADMISSION_TYPE	97.80%
TOTAL_SENTENCE	99.30%
BJS_OFFENSE_1	99.60%
STATE_CODE	100.00%
PRISON_RELEASE_TYPE	100.00%
SEX	100.00%
DATE_OF_BIRTH	100.00%
PRISON_ADMISSION_DATE	100.00%
PRISON_RELEASE_DATE	100.00%
AGE_AT_RLS	100.00%

Table 2: Data completeness by variable

The metric I chose for analyzing predictive performance is the simplest and most intuitive: correct classification rate (CCR). After training the model on half of the data, called the training

set, I had the model make predictions for the other half of the data, called the test set, and the percentage of test set observations that the model classified correctly is the CCR. This is how I compared different model families, combinations of variables, and sampling methods.

State	No. Obs.	Parole rate
CA	1583521	97.0%
TX	690745	62.5%
IL	467312	85.5%
FL	452275	32.0%
NY	387211	86.2%
NC	300719	15.0%
GA	286733	58.7%
MO	262275	80.5%
TN	216143	67.8%
AZ	213577	82.6%
IN	209299	88.9%
SC	180932	49.7%
WI	140683	94.5%
NJ	135758	61.9%
WA	117031	69.7%
OK	104185	55.0%
MN	91860	84.7%
KT	91464	68.3%
CO	90105	86.7%
UT	45183	69.0%
AK	34248	21.3%
IA	33303	57.9%
NV	31810	60.8%
MA	23300	21.0%
NE	20715	36.1%
KS	16940	64.5%
OR	15185	99.7%
WY	6936	63.1%

Table 3: State-by-state breakdown of data size and parole rate

The first, simplest model was a logistic regression applied to all six million observations. I fit a model for all eleven candidate variables and applied a stepwise variable selection algorithm which minimized AIC and in the end removed SEX and PRISON_RELEASE_FROM from the model. Of note in the output for this model (see Appendix A.2) is the size of the coefficients for each level of the STATE_CODE indicator relative to the other predictors; in determining the odds of parole versus non-parole, what state a prisoner is being held in has a bigger impact than more common-sense predictors like the length of their sentence or the type of facility they are being held in (PRISON_RELEASE_FROM). As far as predictive performance, this model correctly classified 79% of the observations in the test set, which might seem a flawless victory for the power of the big data approach, but upon closer inspection this naïve approach that used all the available data actually bankrupted the results, as Table 3 demonstrates. Notice the extreme disparity in the number of observations each state contributes to the dataset, and how widely the rates of parole vary. This explains why the logistic regression was so heavily influenced by STATE_CODE, but also reveals part of why the model

performed so well; California accounted for more than a quarter of the available observations, and 97% of its prisoners are getting some form of conditional release. This "big data" national model focuses in on which states give an easy guess, and then makes that easy guess, but ultimately it is not producing a coherent national picture.

The most obvious way to get around this is to abandon national summarizing and fit individual models for each state, recording CCR and parole rate for each one. This is a bit of a chore computationally, but allows us to continue to use all of the available data. However, a national summary is still quite useful because of criminal justice's increasing profile in national politics, as well as the ability to apply this national model to states other than these 28 which happen to have enough data to model. If we wanted to apply the predictive model in New Hampshire with its 400-odd usable observations, using a national model which captures a broad signal rather than deciding which state is most likely to capture the same signal as New Hampshire is much more viable.

This is what naturally leads us to sampling; although more data is better in a general sense, California swamping our training set is not helpful to the goals of the study. I applied simple stratified sampling and probability proportional to size (PPS) sampling in order to better balance the dataset, and observed predictive performance. The stratified sample had about 7000 observations from each state for a total of about 200,000 observations, which presents a more balanced picture of the country and slashes computing time for model-fitting and predictions. For PPS sampling I fixed the sample size to be the same as the stratified sample to maintain comparability. I applied the same model-fitting procedure to these samples: fit a logistic regression with all predictors and then apply a stepwise AIC algorithm to the initial model. Interestingly, both the stratified-sample-based model kept all candidate predictors, unlike the

whole-dataset model which threw out SEX and PRISON_RELEASE_FROM. The PPS-based model was different still; minimal AIC was achieved when SEX and HISPANIC were removed from the model, but PRISON_RELEASE_FROM was kept.

Given our earlier discussion of varying parole rates, it is unsurprising that these more balanced samples produced different predictive results. As before, it is prudent to compare CCR to the true parole rate in the test set, which I will display graphically below alongside the state-by-state data. The stratified-sample model yielded a CCR of 74% against a parole rate of 65%, and the PPS model yielded a CCR of 78% against a parole rate of 73%. These two points appear in red and blue in the plot on the next page, respectively.

The parabolic shape of this plot is key: as I said before, the model performs better in states where the classification is easy; where the true parole rate is either very low or very high. A perfect 50% parole rate yields the lowest CCR, and the CCR increases as the parole rate moves left or right. From a certain point of view this is discouraging; our predictive success is definitively handcuffed to the parole rate in whatever setting we apply our model to, regardless of technique. On the other hand, this gives us a very rigorous way to quantify how well a particular model does given the underlying parole rate.

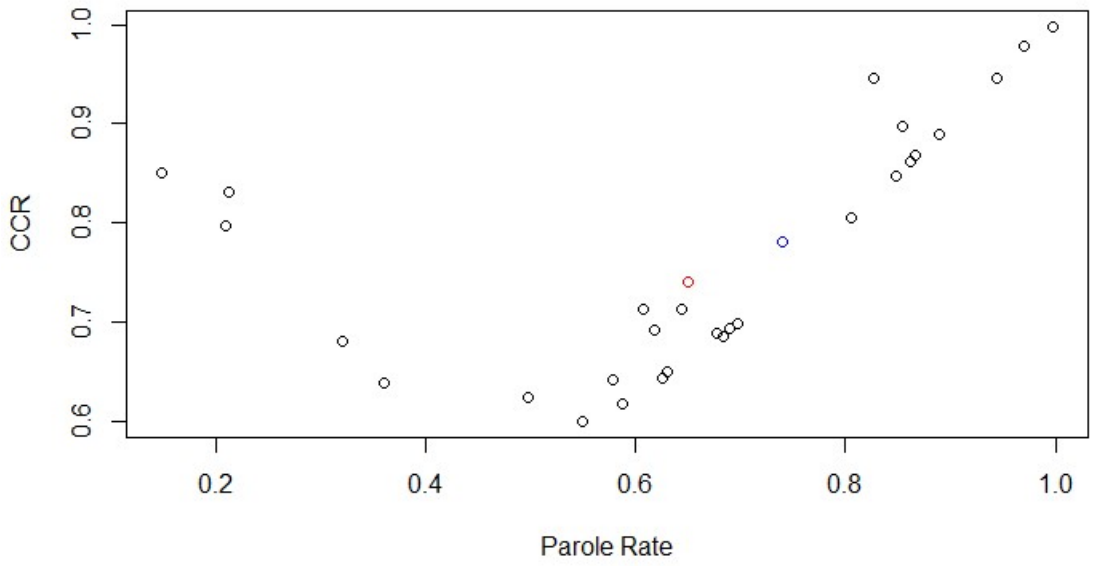


Figure 1: Correct Classification Rate vs. parole rate for states and samples

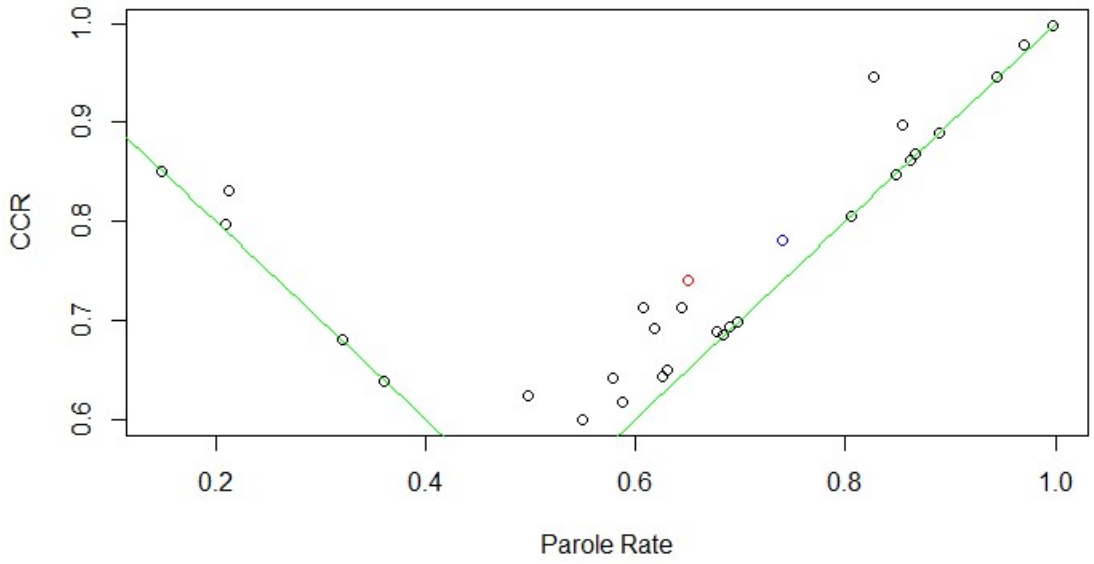


Figure 2: CCR vs. parole rate, plotted with f(x) = |x-.5|+.5

Consider the modified plot on the previous page; the plotted function represents a worst-case scenario where the predictive model fares no better than simply guessing parole for everyone in a parole-friendly state and guessing no parole for everyone in a less-parole-friendly state. Many of the state-level model points fall on this line, particularly as the parole rates tend to the extremes, which reinforces that the naïve big data model is inadequate in many states. However, sampling gave us some positive results—notice that both the stratified sample and PPS models buck this trend somewhat. To make this more concrete, I linearized the state-by-state data, ran a linear regression, and then compared the sampling CCRs to a 95% confidence interval for that line.

What this plot shows us is that stratified sampling (in red) outperforms its underlying parole rate better than both the PPS approach (in blue) and, generally speaking, the big data approach, although there are outliers among the former. Although the linear regression indicates that the linear relationship between transformed parole rate and CCR is highly significant (see

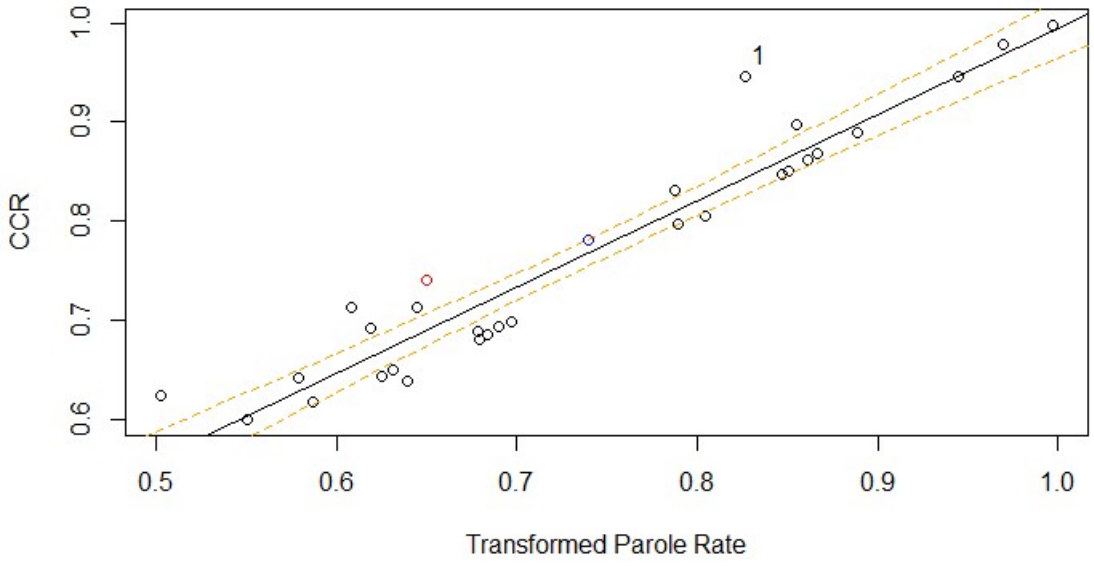


Figure 3: Regression of CCR vs. |Parole Rate - .5| + .5, with 95% confidence interval for the regression line

Appendix A.3), the point labeled 1 corresponds to Arizona, which boasted a 94% CCR against a parole rate of 83%, making it an interesting candidate for further study.

Finally, I dipped into the wide world of machine learning in search of even further predictive improvements. As mentioned before, I chose random forests because, through their variable importance plots and Gini values, they can provide some insight into how different variables affect the model's output. The focus of this paper is not interpretation, but in predictive models that *can* be interpreted, so the curious reader is referred to Chapter 17 of Efron and Hastie for more information on random forest models.

For our purposes, random forests work much the same as logistic regression; they are trained on half the data and then make predictions on the other half, which yields a CCR just like our logistic models. The computational resources required to fit random forests are much higher than those for logistic regression, so I did not fit forests for all 6 million observations; this means no big data model, and no state-level models. Rather, I applied the technique to our two samples from before (which took about twenty minutes each to compute), and used our magic plot to compare them to all our logistic CCRs:

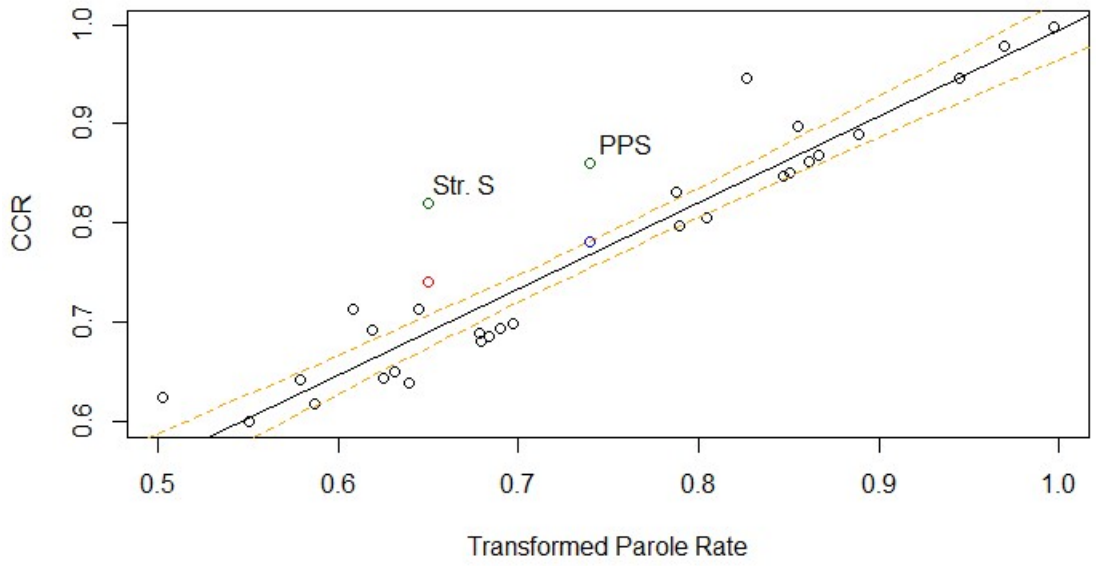


Figure 4: Regression of CCR against parole rate, with random forests

The improvement is obvious; both of the green points corresponding to the forest predictive models far outstrip the corresponding logistic points that were applied to the same samples. The drawbacks to random forests are that they are computationally expensive and there is some loss of interpretation—see Appendix A.4 for the variable importance output mentioned before. However, the sampling methods we have explored here make them feasible even without cloud computing and ensure that the results are capturing true national trends without being swamped by certain outlier states. And I have also shown that logistic models can perform quite well on the samples, so those can be used for more explicit interpretations if the application is more academic than practical.

4. Conclusions

While the above study was neither exhaustive or all that sophisticated, it provides a valuable launching point for further study and underscores some valuable philosophical approaches to working with NCRP data and similar prisoner records. The balancing act among missing variables and missing states, the importance of accounting for variation between states, and the need for context in analyzing predictive performance must all factor in to future studies based on this data.

Specifically, an interpretative analysis of the models produced, in concert with a deep-dive into legal, political science, and social justice literature on the subject of parole and conditional release, would be a tremendously valuable accompaniment to this methodological experiment. In addition, policymakers and activists could explore how best a well-performing, interpretable predictive model for parole could be applied responsibly in the field. For statisticians at much higher level than mine I leave the difficult theoretical work of mathematically justifying the conclusions I reached—that stratified sampling finds the best balance of summarizing responsibly and optimizing predictive performance, and that even large differences in correct classification rates can be mostly explained by variations in underlying parole rate.

Appendix

A.1: Variable summaries

Variable name	Description
STATE_CODE	State in which prisoner was being held; for codes, see attached pages from NCRP codebook.
SEX	Gender, coded as 1 for male and 2 for female.
HISPANIC	Whether the prisoner is Hispanic or assumed Hispanic, coded as 1 for Hispanic and 2 for non-Hispanic.
RACE	The prisoner's race, coded as follows: 1 for white, 2 for black, 3 for American Indian or Alaskan Native, 4 for Asian, 5 for Hawaiian or Pacific Islander, 6 for other.
PRISON_ADMISSION_TYPE	Type of admission to prison, e.g. court commitment, parole revocation, etc. See attached codebook pages for more details.
PRISON_RELEASE_FROM	The type of facility that the prisoner was released from, e.g. local jail, state prison, federal prison, halfway house, etc. See attached codebook pages for more details
TOTAL_SENTENCE	The total prison sentence in days for all counts of all crimes.
AGE_AT_RLS	The prisoner's age on their data of release, calculated from their given birth month and year.
BINARY_RELEASE	Whether the given prisoner's release was conditional or unconditional, coded as 0 for unconditional and 1 for conditional.

A.2: Logistic regression output, full data national model

```
Call:
glm(formula = BINARY_RELEASE ~ as.factor(STATE_CODE) + as.factor(HISPANIC) +
    as.factor(PRISON_ADMISSION_TYPE) + as.factor(RACE) + TOTAL_SENTENCE +
    AGE_AT_RLS, family = "binomial", data = train.1)

Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                   -1.407e+00  2.588e-02  -54.382  < 2e-16 ***
as.factor(STATE_CODE)4         3.040e+00  2.578e-02  117.933  < 2e-16 ***
as.factor(STATE_CODE)6         5.709e+00  2.782e-02  205.188  < 2e-16 ***
as.factor(STATE_CODE)8         3.426e+00  2.935e-02  116.731  < 2e-16 ***
as.factor(STATE_CODE)12        5.724e-01  2.442e-02   23.439  < 2e-16 ***
as.factor(STATE_CODE)13        1.640e+00  2.470e-02   66.395  < 2e-16 ***
as.factor(STATE_CODE)17        4.200e+00  2.619e-02  160.384  < 2e-16 ***
as.factor(STATE_CODE)18        3.541e+00  2.828e-02  125.182  < 2e-16 ***
as.factor(STATE_CODE)19        1.310e+00  3.109e-02   42.148  < 2e-16 ***
as.factor(STATE_CODE)20        2.054e+00  3.713e-02   55.332  < 2e-16 ***
as.factor(STATE_CODE)21        2.162e+00  2.683e-02   80.561  < 2e-16 ***
```

```
as.factor(STATE_CODE)25            -9.058e-02  3.717e-02   -2.437 0.014797 *
as.factor(STATE_CODE)27             4.462e+00  3.043e-02  146.625  < 2e-16 ***
as.factor(STATE_CODE)29             2.938e+00  2.613e-02  112.445  < 2e-16 ***
as.factor(STATE_CODE)30             2.650e+00  4.959e-01    5.343 9.16e-08 ***
as.factor(STATE_CODE)31             6.665e-01  3.458e-02   19.275  < 2e-16 ***
as.factor(STATE_CODE)32             1.800e+00  3.134e-02   57.427  < 2e-16 ***
as.factor(STATE_CODE)33             4.718e+00  3.087e-01   15.285  < 2e-16 ***
as.factor(STATE_CODE)34             1.987e+00  2.692e-02   73.807  < 2e-16 ***
as.factor(STATE_CODE)36             3.602e+00  2.551e-02  141.209  < 2e-16 ***
as.factor(STATE_CODE)37            -3.940e-01  2.532e-02  -15.558  < 2e-16 ***
as.factor(STATE_CODE)38             2.847e+00  1.360e-01   20.939  < 2e-16 ***
as.factor(STATE_CODE)40             1.455e+00  2.588e-02   56.230  < 2e-16 ***
as.factor(STATE_CODE)41             7.098e+00  2.579e-01   27.525  < 2e-16 ***
as.factor(STATE_CODE)45             1.329e+00  2.523e-02   52.680  < 2e-16 ***
as.factor(STATE_CODE)47             1.859e+00  2.528e-02   73.554  < 2e-16 ***
as.factor(STATE_CODE)48             1.684e+00  2.428e-02   69.381  < 2e-16 ***
as.factor(STATE_CODE)49             2.222e+00  3.004e-02   73.988  < 2e-16 ***
as.factor(STATE_CODE)53             2.119e+00  2.605e-02   81.358  < 2e-16 ***
as.factor(STATE_CODE)55             5.446e+00  3.289e-02  165.573  < 2e-16 ***
as.factor(STATE_CODE)56             1.709e+00  4.986e-02   34.282  < 2e-16 ***
as.factor(HISPANIC)2                1.280e-02  8.246e-03    1.552 0.120674
as.factor(PRISON_ADMISSION_TYPE)20  1.338e-01  8.427e-02    1.588 0.112316
as.factor(PRISON_ADMISSION_TYPE)30 -5.730e-01  4.447e-02  -12.885  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)46  1.247e-01  1.241e-02   10.047  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)47 -7.159e-01  7.545e-03  -94.882  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)49  4.503e-01  2.187e-02   20.586  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)56  5.857e-01  2.476e-02   23.657  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)57 -2.558e+00  1.123e-02 -227.814  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)59  5.907e-01  2.454e-01    2.407 0.016066 *
as.factor(PRISON_ADMISSION_TYPE)65 -4.738e-01  2.647e-01   -1.790 0.073485 .
as.factor(PRISON_ADMISSION_TYPE)66  7.386e-01  2.178e-01    3.390 0.000698 ***
as.factor(PRISON_ADMISSION_TYPE)67 -7.358e-01  3.420e-02  -21.515  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)69 -3.032e-01  7.657e-02   -3.960 7.49e-05 ***
as.factor(PRISON_ADMISSION_TYPE)70 -7.345e-01  1.644e-02  -44.689  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)80 -1.260e+00  1.774e-02  -71.022  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)86  6.134e-01  1.625e-02   37.745  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)87  4.900e-02  9.844e-03    4.978 6.43e-07 ***
as.factor(PRISON_ADMISSION_TYPE)88  1.999e+00  1.655e-02  120.816  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)89  2.293e-01  1.165e-02   19.679  < 2e-16 ***
as.factor(PRISON_ADMISSION_TYPE)90 -5.137e-01  4.017e-02  -12.790  < 2e-16 ***
as.factor(RACE)2                   -8.779e-02  4.190e-03  -20.953  < 2e-16 ***
as.factor(RACE)3                   -1.107e-01  1.600e-02   -6.921 4.48e-12 ***
as.factor(RACE)4                    4.686e-02  3.778e-02    1.241 0.214790
as.factor(RACE)5                   -5.461e-02  1.839e-01   -0.297 0.766450
as.factor(RACE)6                    6.031e-02  1.169e-02    5.158 2.49e-07 ***
as.factor(RACE)7                    1.279e-01  7.755e-02    1.649 0.099083 .
TOTAL_SENTENCE                      1.584e-05  3.845e-07   41.183  < 2e-16 ***
AGE_AT_RLS                          2.968e-03  1.852e-04   16.029  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A.3 Output for linear regression of CCR against a function of parole rate

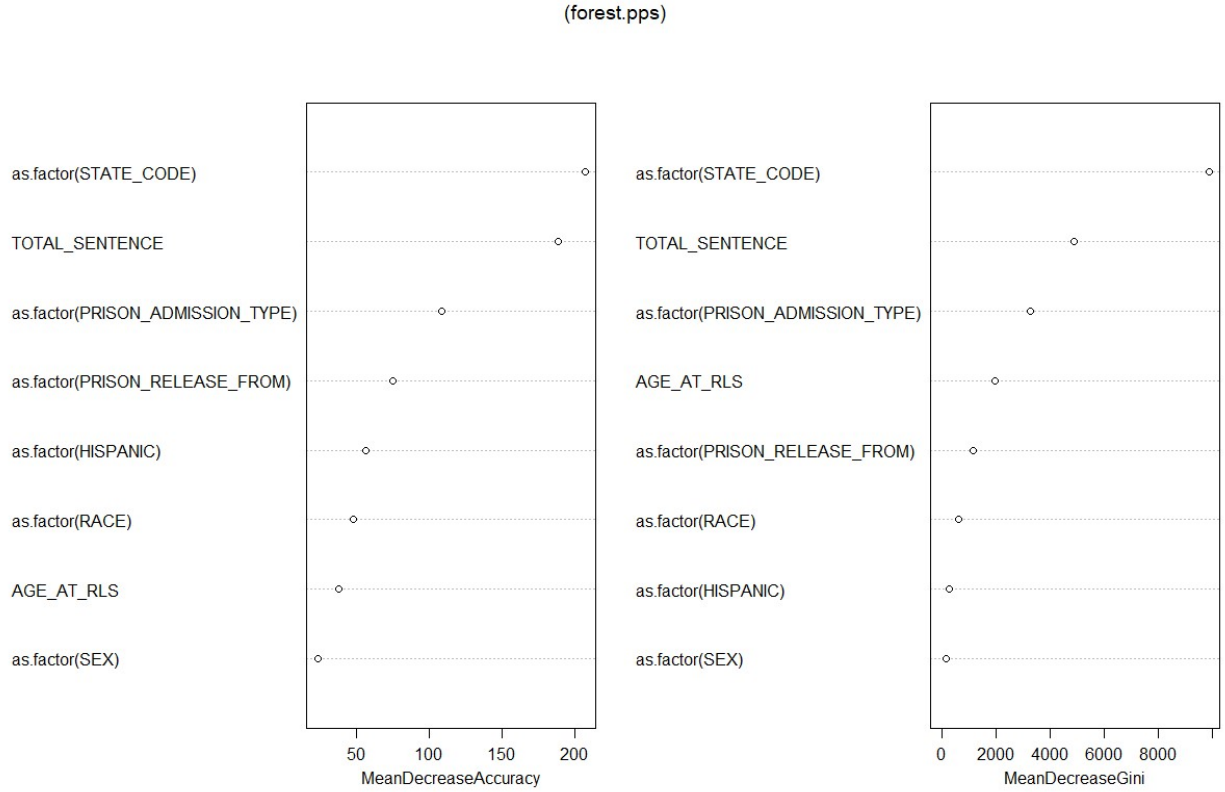
```
Call:
lm(formula = CCR ~ trans)

Residuals:
     Min       1Q   Median       3Q      Max
-0.04327 -0.02344 -0.01086  0.01566  0.10269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12560    0.03649   3.442  0.00196 **
trans        0.86781    0.04857  17.868 4.02e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.03434 on 26 degrees of freedom
Multiple R-squared:  0.9247,      Adjusted R-squared:  0.9218
F-statistic: 319.3 on 1 and 26 DF,  p-value: 4.023e-16
```

A.4 Variable importance output for a random forest



Works Cited

Alumbaugh, Richard V, et al. "Comparison of Multivariate Techniques in the Prediction of Juvenile Postparole Outcome." *Educational and Psychological Measurement*, vol. 38, no. 1, 1978, pp. 97–106.

Andersen, Lars, and H. Wildeman. "Measuring the Effect of Probation and Parole Officers on Labor Market Outcomes and Recidivism." *Journal of Quantitative Criminology*, vol. 31, no. 4, 2015, pp. 629–652.

Berk, Richard, et al. "Forecasting Murder within a Population of Probationers and Parolees: a High Stakes Application of Statistical Learning." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, 2009, pp. 191–211.

Efron, Bradley and T. Hastie. *Computer Age Statistical Inference.* New York: Cambridge University Press, 2016.

James, et al. *An Introduction to Statistical Learning : with Applications in R*. New York: Imprint: Springer, 2013.

Glaser, Daniel. "Who Gets Probation and Parole: Case Study versus Actuarial Decision Making. (Special Issue on Community Corrections)." *Crime and Delinquency*, vol. 31, no. 3, 1985, pp. 367–378.

Matloff, Norman. *Statistical Regression and Classification: From Linear Models to Machine Learning.* Boca Raton: CRC Press, 2017.

Rhodes, William. "A Survival Model with Dependent Competing Events and Right-Hand Censoring: Probation and Parole as an Illustration." *Journal of Quantitative Criminology*, vol. 2, no. 2, 1986, pp. 113–137.

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. National Corrections Reporting Program, 2000-2015. ICPSR36746-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-06-22. http://doi.org/10.3886/ICPSR36746.v1

Text: The most recent day for which the inmate was admitted into the custody of the state prison system on the current sentence.

V14: DAY OF ADMISSION TO PRISON	
Value	*Label*
95 (M)	Illegal entry
97 (M)	Blanked for confidentiality
98 (M)	Blank
99 (M)	Not known

V15 YEAR OF ADMISSION TO PRISON

Location: 36-39(width: 4; decimal: 0)

Variable Type: numeric

Question: Item 8(c): Year of admission to prison

Text: The most recent year for which the inmate was admitted into the custody of the state prison system on the current sentence.

1901-Current Study Year is valid for prison release records. Current Study Year is the only valid code for prison admission records

V15: YEAR OF ADMISSION TO PRISON	
Value	*Label*
8888 (M)	Item not applicable
9995 (M)	Illegal entry
9998 (M)	Blank
9999 (M)	Not known

V16 TYPE OF ADMISSION TO PRISON

Location: 40-41(width: 2; decimal: 0)

Variable Type: numeric

Question: Item 9: Type of admission to prison

V16: TYPE OF ADMISSION TO PRISON	
Value	*Label*
10	Court commitment
20	Returned from appeal or bond
30	Transfer
46	Parole revocation with new sentence
47	Parole revocation with no new sentence
49	Parole revocation, no information regarding new sentence
56	Mandatory parole release revocation with new sentence
57	Mandatory parole release revocation with no new sentence
59	Mandatory parole release revocation, no information regarding new sentence
65	Suspended sentence imposed
66	Escapee/AWOL returned with new sentence

V16: TYPE OF ADMISSION TO PRISON	
Value	*Label*
67	Escapee/AWOL returned with no new sentence
69	Escapee/AWOL returned, no information regarding new sentence
70	Parole status, pending revocation
80	Mandatory parole release status, pending revocation
86	Probation revocation with new sentence
87	Probation revocation with no new sentence
88	Other
89	Probation revocation no information regarding new sentence
90	Probation status, pending revocation
92	Unsentenced commitment
95 (M)	Illegal entry
98 (M)	Blank
99 (M)	Not known

V17 JURISDICTION ON DATE OF ADMISSION

Location: 42-43(width: 2; decimal: 0)

Variable Type: numeric

Question: Item 10: Jurisdiction on date of admission

Text: The state having the legal authority to enforce a prison sentence.

V17: JURISDICTION ON DATE OF ADMISSION	
Value	*Label*
01	Alabama
02	Alaska
04	Arizona
05	Arkansas
06	California
08	Colorado
09	Connecticut
10	Delaware
11	District of Columbia
12	Florida
13	Georgia
15	Hawaii
16	Idaho
17	Illinois
18	Indiana
19	Iowa
20	Kansas
21	Kentucky
22	Louisiana

V47: RELEASED FROM	
Value	*Label*
1	State prison facility
2	Local jail
3	Other
4	Halfway house
5	Work release center
6	Pre-release center
12	Federal prison
88 (M)	Item not applicable (Prison Admission records only)
95 (M)	Illegal entry
98 (M)	Blank
99 (M)	Not known

V48 1ST AGENCY THAT ASSUMED CUSTODY AT RLS

Location: 134-135(width: 2; decimal: 0)

Variable Type: numeric

Range of Missing Values (M): 88, 95, 98, 99

Question: Item 24(a): First agency that assumed custody for this person at time of release.

Text: VARIABLES V37 THROUGH V51 APPLY TO PRISON RELEASE RECORDS (RECORD TYPE "2") AND PAROLE RELEASE RECORDS (RECORD TYPE "3") ONLY.

V48: 1ST AGENCY THAT ASSUMED CUSTODY AT RLS	
Value	*Label*
0	None
1	Other prison, outside state
2	Other prison, Federal
3	Parole, within state
4	Parole, outside state
5	Parole, Federal
6	Probation, within state
7	Probation, outside state
8	Probation, Federal
9	Mental/medical facility, within state
10	Mental/medical facility, outside state
11	Mental/medical facility, Federal
12	Other, within state
13	Other, outside state
14	Other, Federal
88 (M)	Item not applicable (Prison Admission records only)
95 (M)	Illegal entry
98 (M)	Blank
99 (M)	Not known

V94: STATE IDENTIFIER	
Value	*Label*
01	Alabama
02	Alaska
04	Arizona
05	Arkansas
06	California
08	Colorado
09	Connecticut
10	Delaware
11	District of Columbia
12	Florida
13	Georgia
15	Hawaii
16	Idaho
17	Illinois
18	Indiana
19	Iowa
20	Kansas
21	Kentucky
22	Louisiana
23	Maine
24	Maryland
25	Massachusetts
26	Michigan
27	Minnesota
28	Mississippi
29	Missouri
30	Montana
31	Nebraska
32	Nevada
33	New Hampshire
34	New Jersey
35	New Mexico
36	New York
37	North Carolina
38	North Dakota
39	Ohio
40	Oklahoma
41	Oregon
42	Pennsylvania
44	Rhode Island

V94: STATE IDENTIFIER	
Value	*Label*
45	South Carolina
46	South Dakota
47	Tennessee
48	Texas
49	Utah
50	Vermont
51	Virginia
53	Washington
54	West Virginia
55	Wisconsin
56	Wyoming
57	Federal Prison System
58	California Youth Authority
60	State Not Known
52	Shared Jurisdiction
62	Northern Mariana Islands
64	Guam
66	Puerto Rico
68	Virgin Islands
70	Private Prison within State

V95 **RECORD TYPE**

Location: 246-246(width: 1; decimal: 0)

Variable Type: numeric

V95: RECORD TYPE	
Value	*Label*
1	Prison admission record
2	Prison release record
3	Parole release record

V96 **OFFENSE #1 - ORIGINAL CODE REPORTED**

Location: 247-266(width: 20; decimal: 0)

Variable Type: character

Text: Offense #1 - original code reported by state

V97 **OFFENSE #2 - ORIGINAL CODE REPORTED**

Location: 267-286(width: 20; decimal: 0)

Variable Type: character

Text: Offense #2 - original code reported by state

V98 **OFFENSE #3 - ORIGINAL CODE REPORTED**