Title
Ensemble Procedures for Finding High Risk Prison Inmates

Authors
Richard A. Berk
Jong-Ho Baek

Publication Date
2011-10-25

Ensemble Procedures for Finding High Risk Prison Inmates*

Richard A. Berk
Jong-Ho Baek

Department of Statistics
UCLA

September 30, 2003

1 Introduction

The California Department of Corrections (CDC) currently houses more inmates than any state corrections department in the country (Harrison and Karlberg, 2003). There are approximately 160,000 inmates in 33 institutions, 16 community corrections facilities, 41 camps, and 8 prisoner mother facilities. A key issue for correction officials and a range of stakeholders is how best to maintain order and safety in a cost-effective manner. Every year, approximately 15% of the inmates engage in some form of "misconduct" that can range from failing to report for a work assignment to insubordination to possession of narcotics to an assault on a guard or fellow prisoner.

Responding in part to such concerns, the CDC has for several decades employed an inmate placement system that attempts to match prisoners to prison housing so that inmates more likely to engage in misconduct are placed

*The research reported in this paper would have been impossible without the talents and efforts of our colleagues at the California Department of corrections: George Lehman, Maureen Tristan, Gloria Rea, Penny O'Daniel, Micki Mitchell, Mark Cook, Martha Pyog, and Terrence Newsome. Andy Liaw provided a number useful suggestions on the random forests analysis.

in more secure settings. Some facilities are essentially dormitories. At the other extreme are very restrictive settings characterized by celled housing, a lethal perimeter, controlled movement, and armed supervision within the housing units and dining halls. The average cost for housing an inmate in the CDC is over $25,000 a year, but the costs for more restrictive housing are significantly higher. The goal, therefore, is to house each inmate in the least restrictive setting that can insure the inmate's safety and the safety of others.

Each inmate's supervision needs are quantified by a classification score based on the inmate's background (e.g., age) and current offense (e.g., sentence length). The score is computed soon after arrival at the CDC Reception Center, and is essentially a linear combination of about a dozen items. For about 75% of the inmates, placement in one of four security levels is determined by whether a score falls within certain ranges. For example, a score of less than 18 typically leads to placement in one of the lowest security level facilities (i.e., a "Level I" facility). A score greater than 52 typically leads to placement in one of the highest security level facilities (i.e., a "Level IV" facility).

About 25% of the inmates are placed through a set of "mandatory minimums" that respond to special features of offense (e.g., violent sex offenders) or specialinmate needs (e.g., targeted by a rival gang). Earlier research has shown that overall, the CDC classification score is a useful way to determine the allocation scarce prison space (Berk and de Leeuw, 1998, Berk et al., 2003).

Much has been written on quantitative inmate classification systems, including their development and evaluation (Austin, 1986; Austin, Baird and Neuenfeldt, 1993; Baird, 1993; Brennan, 1987; ; 1993; Kane, 1986; Alexander and Austin, 1992; Harer and Langan, 2001; Hardyman, Austin, and Tulloch, 2000; Hardyman and Adams-Fuller, 2001). But we know of no studies addressing procedures to find the inmates at highest risk to the most serious forms of prison misconduct. Thus, the CDC's classification system will find the inmates as highest risk for engaging in any form of misconduct. But minor infractions are treated the same as major infractions.

On the one hand, this is a conscious feature of the CDC system; the goal is to have operational procedures for the vast majority of inmates and the vast majority of infractions they may commit. On the other hand, there are daunting technical challenges to developing a set of practical procedures to find "the worst of the worst." In particular, very serious infractions are quite

rare. These include infractions that would be major felonies if committed outside of prison, such as drug trafficking, assault, sexual assault, robbery, attempted murder, and homicide.

In this paper, we consider how ensemble procedures can be used to find better the small fraction of inmates who are likely to commit the most serious violations. We also consider whether it is then possible to develop practical tools that prison staff might use to forecast who such inmates might be.

2 Study Design

The data are taken from a completed randomized trial testing the CDC inmate classification system. 21,734 male and female felon commitments arriving at the CDC Reception Center between November 1, 1998 and April 30, 1999 were included in the study. Approximately, half were assigned at random to be placed under an existing CDC inmate classification system, and half were randomly assigned to be placed under a revised system. The revised system included an updated list of risk factors and new set of weights for the items that were to be combined into each inmate's classification score. The results of the experiment showed the new system to be a significant improvement over the old. Details are discussed a recent paper by Berk and his colleagues (2003).

In this paper, we use the data from the 9662 male inmates assigned to placement under the revised system. As a routine matter, an intake form called an "839" is filled out at reception on all new inmates. Under the experiment, the 839 was revised to remove some background variables that were no longer seen as useful for placement (e.g., served in the military) and to add some new predictors thought to be associated with the risk of misconduct in prison. The revised form, therefore, provides a better set of predictors. These include the following variables (with the percentages of inmates in each category or key summary statistics in parentheses).

1. Age at arrival at the Reception Center in years (AgeRec) — 16-20 (12%); 21-26 (23%); 27-35 (30%); 36 or older (35%)

2. Age at the time of the earliest arrest in years (AgeArr) — 0-17 (30%); 18-21 (41%); 22-29 (19%) 30-35 (6%); 36 or older (4%)

3. Associated with gang activity (Gang) — (19%)

4. Mental Illness (Psych) — (8%)

5. Previously served time in a county jail (Jail)— (66%)

6. Previously served time under the California Youth Authority (CYA) — (8%)

7. Previously served time under the California Department of Corrections (CDC) — (30%)

8. Sentence length in years (Term) — (median = 2; IQR = 2)[1]

A "Rules Violation Report" (called a "115") is completed when prison staff observe an inmate engaged in some form of prison misconduct. These reports are the source of our response variable. For this study there was a 24 month follow-up for each inmate starting with admission into the CDC Reception Center.

The particular response of interest is a serious type of misconduct recorded under "Division Levels" A1, A2, B and C. As noted earlier, these offenses include such crimes as assault, drug trafficking, and robbery and are the types of misconduct that can automatically send an inmate to a Level IV CDC facility. "Administrative" violations, such as failing to report for a work assignment, are not included. Division levels A1,A2, B and C represent about 2.5% of all 115s, and there were virtually no inmates who committed more than one such offense in the 24 month follow-up period. Therefore, our response variable will be treated as binary: committed a serious 115 or not.

3 Data Analysis

We began by holding out a random sample of 500 from the total of 9662 inmates. The remaining 9162 inmates are our "training" data set. The 500 inmates constitute a data set that will be used for forecasting.

[1]Sentence length is recorded on the CDC intake form 839 so that it is capped at 25 years. Thus, for example, a life sentence without the possibility of parole is recorded as 25 years. The median and IQR are here unaffected by the truncation at 25.

3.1 Using Logistic Regression and CART

In one sense, the classification problem is trivial. If one were to ignore all predictors and always classify an inmate as uninvolved in serious misconduct, one would be right about 97.5% of the time. Assuming that the inmates included in this study were representative of inmates coming to the CDC Reception Center over medium term (a reasonable assumption), forecasts would also be accurate the vast majority of the time.

Not surprisingly, adding predictors in a conventional manner does not help much. When logistic regression was applied, there was no meaningful improvement in fit with the predicted probabilities of serious misconduct never higher than .03. When Classification and Regression Trees (Breiman et al., 1984) was applied, it was difficult to get any tree built at all. There were no splits that could perform substantially better than the root node unless priors for misconduct were employed that compensated at least in part for the lack of balance. And then the resulting tree varied substantially depending on the priors chosen.

3.2 Using Random Forests

Could one do better with ensemble methods (Hastie et al., 2001) as a special case of what Breiman (2001b) calls algorithmic modeling? We turned to random forests as one promising ensemble approach (Breiman, 2001a; 2001b; 2001c). For binary outcomes, random forests constructs an ensemble of classification trees. Each tree is built from a bootstrap sample of the data and at each split, a random sample of predictors is examined. In the end, classification is determined by a majority vote for each case over the ensemble of classification trees.

Random forests will converge and produce consistent estimations of key population parameters (Breiman, 2001c). There is some evidence that random forests will usually classify more accurately than CART and at least as well as the the most effective existing statistical learning alternatives, such as ADAboost (Breiman,2001a).[2] Random forests has the added benefit of no apparent problems with overfitting. Recent implementations (e.g., Random Forests V4.0) have a rich set of outputs that can help enormously in interpreting the results.

[2]Also, ADAboost does not converge and does not produce consistent estimates (Mannor et al., 2002)

Random forests with its default parameters was applied to the training data set. In particular, the prior distribution of the response was taken to be the observed marginal distribution. With these settings, random forests did virtually no better than the marginal distribution of the response variable. Only one inmate was classified (correctly, as it turned out) as engaging in serious misconduct. All of the other inmates were classified as not engaging in serious misconduct although 253 of the 9162 actually did.

Taking seriously the idea that random forests is only an algorithm, we weighted the data in a fashion that to varying degrees compensated for the highly skewed response variable distribution. These weights can be in some circumstances viewed as reflecting the prior distribution of the response. For example, weighting misconduct cases to no misconduct cases by a ratio 10 to 1 implies a prior distribution with about 91% of the inmates guilty of serious misconduct and about 9% not. For our analysis, however, the weights are just a tuning parameter. The weights are given no statistical or substantive interpretation.

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	1676	7233	0.81
Misconduct	36	217	0.14

Table 1: Confusion Table for a 10 to 1 Weighting for Misconduct to No Misconduct

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	4348	4561	0.51
Misconduct	81	172	0.32

Table 2: Confusion Table for a 2 to 1 Weighting of Misconduct to No Misconduct

Weighting misconduct cases to no misconduct cases by a ratio 10 to 1 produced the "confusion table" in Table 1. The 10 to 1 weighting does a pretty good job in picking out the true positives, but the price is very high. For every true positive there are over 33 false positives. One implication is that about 80% of all new inmates are classified as committing a serious

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	7100	1809	0.20
Misconduct	150	103	0.59

Table 3: Confusion Table for a 1 to 1 Weighting of Misconduct to No Misconduct

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	8390	519	0.06
Misconduct	211	42	0.83

Table 4: Confusion Table for a 1 to 2 Weighting of Misconduct to No Misconduct

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	8814	95	0.01
Misconduct	233	20	0.92

Table 5: Confusion Table for a 1 to 4 Weighting of Misconduct to No Misconduct

violation. The real figure is about 2.5%. It appears that we have grossly overcompensated for the lack of balance. These results are clearly of no help to the CDC.

Table 2 through Table 5 provide results for different sets of weights. Two conclusions follow. First, the weighting can make an enormous difference in how the inmates are classified. For example, while in Table 1 about 80% of the inmates are classified as engaging in serious misconduct. In Table 5, using a 1 to 4 ratio, approximately 1% are classified as engaging in serious misconduct.

Second, the numbers of false positives and false negatives also vary dramatically. In Table 1, over 80% of the no misconduct cases are misclassified and only about 14% of the misconduct cases are misclassified. In Table 5 the results are dominated by false negatives.

3.3 Selecting the More Useful Results

It is extremely difficult to know which of the results in Tables 1-5 (or some other table) are preferable without introducing additional information. To begin, it can be useful to know which predictors are driving the results. Consider first the random forest fit responsible for Table 4.

Figure 1 shows the reduction in the Gini index fitting criterion that may attributed to each predictor if that predictor's influence on the fit were removed. These reductions are calculated for a given predictor by randomly shuffling the values of that predictor and re-running random forests. If that variable is important to the fit, the value of the Gini criterion will decline substantially. The predictor is then returned to its original state, and the process repeated for each of the other predictors in turn. The values reported in Figure 1 are the average drop in the Gini criterion over the ensemble of trees.

Figure 1 shows that term length makes the most important contribution to the fit. Age at first arrest, age at reception, and gang activity are also important. From past research on prison inmate, these results are to be expected. (Berk et al., 2003). The age effects are also consistent with extensive work in criminology (Gottfredson and Hirschi, 1990, Sampson and Laub, 1990, Hamil-Luker et al., 2003).

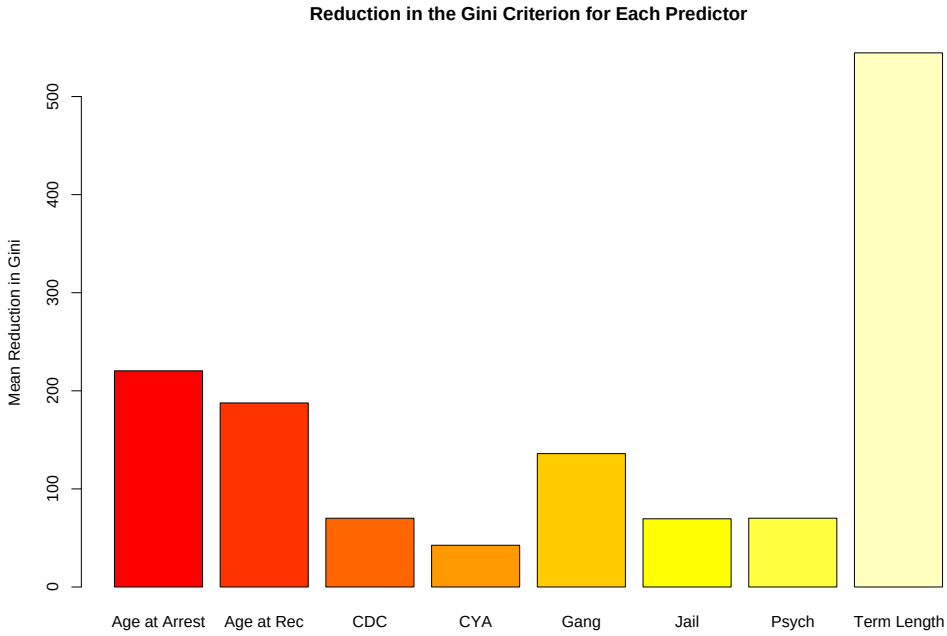

Figure 1: Average Reduction in Gini Criterion for Each Predictor for Serious Misconduct

Partial plots can be used to show for random forests the marginal relationship between a given predictor and the probability of misconduct. Figure 2 shows the partial plot for term length. The vertical axis is in logits, calculated for each value of the predictor as

$$\frac{1}{n}\sum_{i=1}^{n} log(p_k) - log(p_j), \tag{1}$$

where p_k is the proportion of votes over trees for class k, and p_j is the proportion of votes over trees for class J Here, k is for serious misconduct and j is for no serious misconduct.

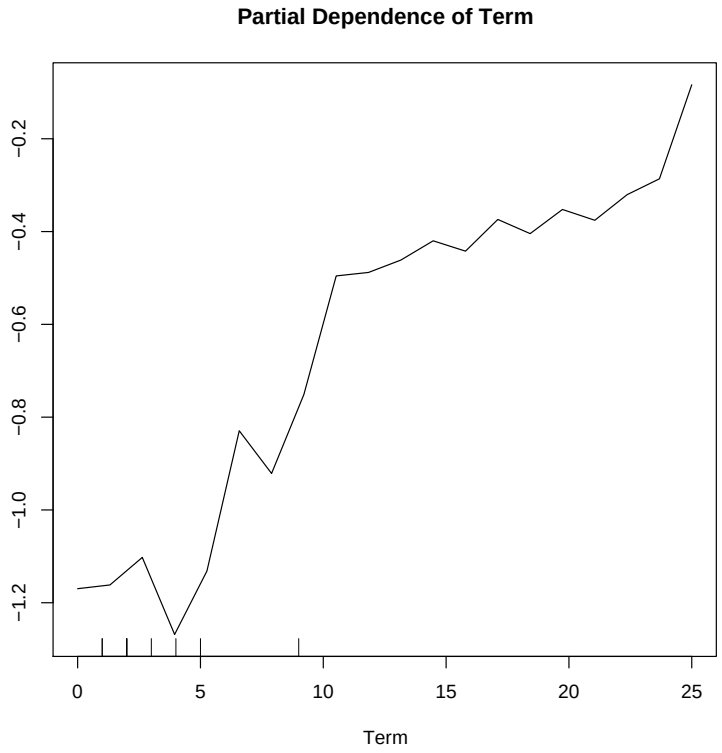

Figure 2: Partial Dependence Plot for the Predictor Term Length — in Logits of Serious Misconduct

One can see that the relationship is generally positive. But it is somewhat less steep between sentences of 10 and 24 years and is very steep for the few the very longest sentences. Similar analyses confirmed that each of the important predictors had their anticipated effects (Berk et al., 2003): serious misconduct is more common among younger inmates with longer records, inmates involved in gang activity and inmates sentenced to long terms.

The partial dependence plots were much the same for the other tables. But the more weight given to serious misconduct in the random forest fit, the less salient was the role of term length compared to the other predictors. That is, when the relative costs classification errors for cases of misconduct were increased relative to the costs of classification errors of cases of no misconduct, the role of term length was less pronounced. Term length generates too many false positives if false positives are given extra weight. This seems to be consistent with Figure 2 in which the relationship between term length and sentence is generally the strongest for shorter term lengths. One implication is that if false positives are an important policy concern, the role of term length in inmate classification should be reduced.

The kinds of inmates classified as engaging in very serious misconduct are broadly like the inmates likely to be placed through the regular inmate classification system in the more secure CDC facilities. The salient predictors are the very ones leading to high inmate classification scores (Berk et al., 2003). However, the inmates identified by random forests are considerable more difficult than regular level IV inmates, most of whom never get into serious trouble, and would ordinarily not be identified until after an incident of serious misconduct had occurred.

What could the CDC do with that information? One option would be to make sure that such inmates were only given job assignments that were well staffed and that would provide no access to materials that could be used to make weapons. But that alone would likely be seen by the CDC as inadequate.

Another option would be to place all such inmates in Level IV housing. The CDC's current housing arrangements allow for about 5% of all inmates to be placed in Level IV housing. Without a major and costly reorganization of existing space, this implies that one upper bound for inmates classified in our data as very high risk is around 450 (about 5% of the sample). This rules out the random forests analyses that are the basis for Tables 1 through 3.

Table 4 shows that 561 inmates are classified as engaging in serious misconduct, with 42 true positives. Although the the 561 is the ballpark, it is

a bit too high. Moreover, there is more that CDC can do within level IV placements for a difficult subset of level IV inmates.

Our definition of serious misconduct is the same definition used by the CDC to place Level IV inmates in even more restrictive settings. Level IV facilities can be differentiated by whether they have a "180 degree design" or a "270 degree design." These designations refer to the size of the viewing angle from a prison watchtowers. The 270 degree design allows more inmates to kept under surveillance at one time and is therefore used for the very most dangerous inmates. Such facilities might be used to house the inmates our methods identified. However the capacity of these prisons is about 1% of all inmates, or about 100 inmates in our sample. The results shown in Table 4 are, therefore, even less satisfactory.

Table 5 shows that 115 inmates are classified as engaging in serious misconduct. While the number of inmates classified as problematic is acceptable, only 20 of the 115 are true positives; the cost in false positives is high. We will try to do better shortly.

In summary, random forests analyses bounded at the top by the results in Table 4 and at the bottom by the results in Table 5 may be viable options for the CDC. Then, depending on the particular model chosen, it would be a simple matter to drop new cases down the trees to make predictions for serious misconduct. Inmates could then be housed accordingly.

For this application, using priors to compensate for the highly unbalanced distribution of the response variable clearly improves the classification performance for random forests. Such priors also improve the performance of CART, but it still performs less well than forests when proper adjustments for overfitting are applied (i.e. on a test sample). Analogous weighting also improves the performance of logistic regression, but it fares the worst of the three. These differences in effectiveness are no surprise (Breiman, 2001a), and random forests has some additional features that for this application can be very useful. We turn to those now.

3.3.1 Using Random Forest Margins

Recall that random forests produces an ensemble of trees through bootstrap samples of the data when each new tree is to be constructed, and then random samples of the predictors at each split. Also recall that classification is determined by a vote over the full ensemble of trees. From this, one can compute the "margin" for each case classified. Here, the margin is defined

as the proportion of votes cast over an ensemble of 500 random forest trees in favor of the correct classification minus the proportion of votes cast over the ensemble of 500 random forests in favor of the incorrect classification. For example, if for a given case actually characterized by serious misconduct 400 trees classified the case as serious misconduct and 100 trees classified the case as not serious misconduct, the margin for that case would be .60. Had 400 trees classified the case as not serious misconduct and 100 trees had, the margin would be -.60.

Given the bootstrap samples of data, the margin for each cases reflects stability "under data selection" (Gifi, 1996: 36). How stable is the classification under different random samples of the data? Given, the random samples of predictors, the margin for each case also reflects stability "under model selection" (Gifi, 1996: 37). How stable is the classification under different sets of predictors? Presumably, the CDC would prefer to concentrate its scarce resources on inmates whose classifications were relatively stable: cases that could be classified with substantial confidence.

The margins will vary depending on how the inmates are classified, and we have just seen how that can change depending on the manner in which the cases are weighted. Figure 3 shows a plot of the margins for results reported in Table 4. Inmates actually engaging in serious misconduct are represented in blue and inmates not actually engaging in serious misconduct are represented in red. Correct classifications fall above 0, and incorrect classifications fall at or below 0.

With the very high density of points near the top of Figure 3, it is clear that the vast majority are correctly classified in a stable manner. And most of these are true negatives. But there are also some true positives within the sea of red dots. Insofar as the CDC might like to emphasize classifications that are both correct and stable, they would, in effect, draw a horizontal line in Figure 3 not at 0 but at, say, .50. That would guarantee that only relatively stable and correct classifications were selected.

Unfortunately, margin are only defined when the correct classification is known. Margins cannot help if the goal is to place new prisoners in appropriate housing shortly after their arrival at the Reception Center. However, the classification votes for each case over an ensemble of trees can help.

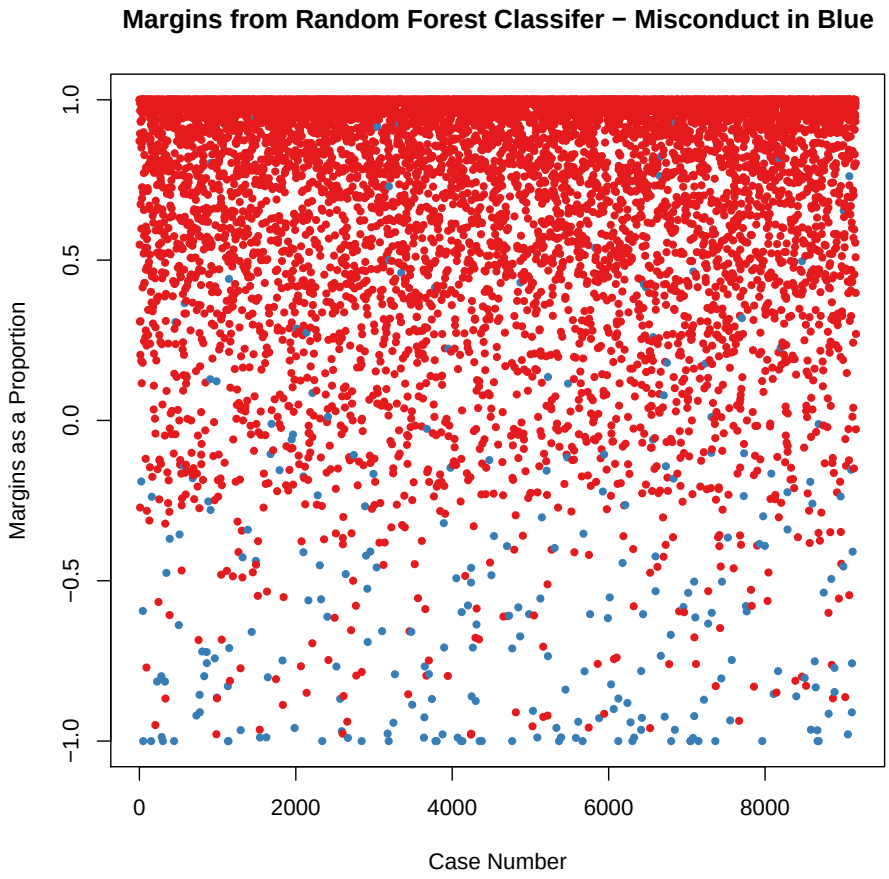

Figure 3: Random Forest Margins

3.4 Forecasting for New Cases Taking Votes into Account

Recall that we held out a random sample of 500 cases from the analyses reported above. We dropped these cases down the ensemble of trees constructed for the results shown in Table 4. The goal was to see if it were possible to usefully reduce the number of inmates predicted to engage in serious misconduct by concentrating on the more stable predictions. While in practice the true classifications would not be known, they are known here and can be used to evaluate what might happen if the true classifications were not known.

For the test data, 34 out of 500 cases were classified as engaging in serious misconduct. Because these were test data, we knew the truth; there were actually only 18 instances of serious misconduct. Table 6 shows the confusion table. The results are just about what one would expect given Table 4. One implication is that too large a fraction of inmates (about 7%) are predicted as very high risk. Once again, the CDC housing capacities for high risk inmates would probably be exceeded.

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	453	29	0.06
Misconduct	13	5	0.72

Table 6: Test Data: Random Forests Confusion Table for Predicting Serious Inmate Misconduct

Tables 7 - 9 show some confusion tables for which the proportion of votes over the ensemble of trees is larger than .50. As the threshold is increased, the predictions are more stable in the sense discussed earlier. Stability is desirable. In addition, with increasing thresholds, the number of inmates predicted to engage in serious misconduct decreases. For example, when the threshold is set at greater than .55 (Table 7), about 5% of the inmates are predicted to engage in serious misconduct. This is probably still too large a proportion given CDC resource and housing constraints.

With the threshold set at greater than .75 (Table 9), 1.4% of the inmates are predicted to engage in serious misconduct. Of that 1.4%, there are 5 false positives and 2 true positives for a ratio of 2.5 to 1. Put another way, the probability of correctly predicting serious misconduct is around .3

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	462	20	0.04
Misconduct	13	5	0.72

Table 7: Test Data: Random Forests Confusion Table for Predicting Serious Inmate Misconduct with Votes $> .55$

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	467	15	0.03
Misconduct	15	3	0.63

Table 8: Test Data: Random Forests Confusion Table for Predicting Serious Inmate Misconduct with Votes $> .60$

	Classified No Misconduct	Classified Misconduct	Error
No Misconduct	477	5	0.01
Misconduct	16	2	0.89

Table 9: Test Data: Random Forests Confusion Table for Predicting Serious Inmate Misconduct with Votes $> .75$

when the stability of the forecast is given substantial weight. This leads to a dramatic improvement over the marginal distribution. (Recall that the marginal probability of serious misconduct was .025.) It also leads to a meaningful improvement over the results for Tables 6 - 8 with probability of serious misconduct among those so classified was less than .20 for each. The downside is that most of the true positives are not identified, and the price for identifying more of them is, once again, increasing substantially the number of false positives.

Note, however, that there is no need for a single decision about an appropriate vote threshold. That decision could be made in the field based on legal and ethical constraints and on available resources. Hence, these decisions could in principle be altered as those constraints and resources varied. That is, CDC staff could make threshold decisions within certain boundaries on the fly, trading false positives against true positives depending upon available resources and the costs associated with various placements and staffing requirements.

To make this concrete, the CDC receives about 4000 new inmates each month and as a practical matter, forecasts could be made on a monthly basis. If results like shown in Table 9 were used, about 60 inmates a month would be predicted to engage in serious misconduct. This number is not unreasonable especially when one remembers that these are among more stable forecasts available.

4 Conclusions

Even with the highly unbalanced response variable, it is possible with these data to meaningfully improve forecasting skill. One key ingredient is the response variable's priors viewed as a tuning parameter. Another key ingredient is exploiting information on how stable the forecasts are. In the longer run, however, this is no substitute for finding better predictors that could be measured when each inmate arrives at the CDC Reception Center. For example, if the offense for which an inmate was incarcerated were "gang related," that might usefully predict gang related problems while in prison.

At the same time, just as in any real world decision-making, the forecasts made can depend on the consequences of those decisions and on the resources available. Here, if more false positives could be tolerated, the number of true positives will also rise. For example, inmates identified as false positives are

in fact very likely to commit other less serious, by still disruptive, forms of misconduct. The vote threshold could be set so that false positives would rarely be cooperative inmates who get into no trouble whatsoever.

We make no claims that the procedures we have applied are in any sense optimal. However, procedures like those we used would provide the CDC with useful forecasts of serious misconduct, substantially better than they can make now.

5 References

Alexander, Jack and James Austin. *Handbook for Evaluating Prison Classifications Systems.* San Francisco: National Council on Crime and Delinquency.

Austin, James. 1986. Evaluating How Well Your Classification System is Operating: A Practical Approach. In *Crime & Delinquency* 32, No. 3, ed. Lawrence A. Bennett. Newbury Park, Calif.: Sage Publications.

Austin, James, Christopher Baird, and Deborah Neuenfeldt. 1993. Classification for Internal Management Purposes: The Washington Experience. In *Classification: A tool for managing today's offenders.* American Correctional Association.

Baird, Christopher. 1993. Objective Classification in Tennessee: Management, Effectiveness, and Planning Issues. *Classification: A Tool for Managing Today's Offenders.* American Correctional Association.

Berk, R.A. and J. de Leeuw (1998) "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association*, Volume 94, Number 448: 1045-1052.

Berk, R.A., Ladd, H., Graziano, H., and J Baek (2003) "A Randomized Experiment Testing Inmate Classification Systems," *Journal of Criminology and Public Policy*, 2, No. 2: 215-242.

Breiman, L., Friedman, J.H., Olshen, R.A., and C.J. Stone (1984) *Classification and Regression Trees.* Monterey, Ca: Wadsworth and Brooks/Cole.

Breiman, L. (2001a) "Random Forests." *Machine Learning* 45: 5-32.

Breiman, L. (2001b) "Statistical Modeling: Two Cultures," (with discussion) *Statistical Science* 16: 199-231.

Breiman, L. (2001c) "Wald Lecture I: Machine Learning," at ftp://ftp.stat.berkeley.edu/pub/users

Breiman, L. (2001d) "Wald Lecture II: Looking Inside the Black Box," at ftp://ftp.stat.berkeley.edu/pub/users/breiman/

Breiman, L. (2003) "Manual – Setting Up, Using, and Understanding Random Forests V4.0". At ftp://ftp.stat.berkeley.edu/pub/users/breiman/

Brennan, Timothy. 1987. Classification: An Overview of Selected Methodological Issues. In *Prediction and Classification: Criminal justice decision making*. Chicago: University of Chicago Press.

Brennan, Timothy. 1993. Risk Assessment: An Evaluation of Statistical Classification Methods. In *Classification: A tool for Managing Today's Offenders*. American Correctional Association.

Gifi, A. (1996) *Nonlinear Multivariate Analysis*. New York: John Wiley.

Gottfredson, M.R., and T. Hirschi (1990) *A General Theory of Crime*. Stanford, Ca: Stanford University Press.

Hamil-Luker, J., Land K.C., and J. Blau (2003) "Diverse Trajectories of Cocaine Use through eaerly Adulthood Among Rebellious and Socially Conforming Youth." *Social Science Research*, forthcoming.

Harer, M.D., and N.P. Langan (2001) "Gender Differences in Predictors of Prison Violence: Assessing the Predictive Validity of a Risk Classification System," *Crime & Delinquency* 47: 513-536.

Hardyman, Patricia L., James Austin, and Owan C. Tulloch. 2000. *Revalidating External Classification Systems: The Experience of Seven States and Model for Classification Reform*. Report submitted to the National Institute of Corrections. Washington, D.C.: The Institute on Crime, Justice and Corrections at The George Washington University.

Hardyman, Patricia L., and Terri Adams-Fuller. 2001. National Institute of Corrections Prison Classification Peer Training and Strategy Session: What's Happening with Prison Classification Systems? September 6-7, 2000 Proceedings.

Harrison, P.M., and J.C. Karlberg. (2003) "Prison and Jail Inmates at Midyear, 2002," *Bureau of Justice Statistics Bulletin*, April, 2003, NCJ 198877.

Hastie, T., Tibshiani, R., and J. Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Kane, Thomas R. 1986. The Validity of Prison Classification: An Introduction to Practical Considerations and Research Issues. In *Crime & delinquency* 32, No. 3, ed. Lawrence A. Bennette. Newbury Park, Calif.: Sage Publications.

Mannor, S., Meir, R. and T. Zhang (2002) "The Consistency of Greedy Algorithms for Classification." In J. Kivensen and R.H. Sloan (eds.), COLT 2002, LNAI 2375: 319-333.

Sampson, R.J., and J.H. Laub (1993) "Crime and Deviance over the Life Course: The Saliance of Adult Social Bonds." *American Sociological Review* 55: 609-627.