California Center for Population Research
University of California - Los Angeles

Forecasting Dangerous Inmate Misconduct: An Applications of Ensemble Statistical Procedures

Richard A. Berk, Brian Kreigler, and Jong-Ho Baek

CCPR-029-05

May 2005

California Center for Population Research
On-Line Working Paper Series

Forecasting Dangerous Inmate Misconduct: An Applications of Ensemble Statistical Procedures[*]

Richard A. Berk
Brian Kriegler
Jong-Ho Baek

Department of Statistics
UCLA

May 27, 2005

[*] 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. Support for work on this paper was provided by the National Science Foundation: (SES - 0437169) "Ensemble Methods for Data Analysis in the Behavioral, Social and Economic Sciences." The support is gratefully acknowledged.

Abstract

In this paper, we attempt to forecast which prison inmates are likely to engage in very serious misconduct while incarcerated. Such misconduct would usually be a major felony if committed outside of prison: drug trafficking, assault, rape, attempted murder and other crimes. The binary response variable is problematic because it is highly unbalanced. Using data from nearly 10,000 inmates held in facilities operated by the California Department of Corrections, we show that several popular classification procedures do no better than the marginal distribution unless the data are weighted in a fashion that compensates for the lack of balance. Then, random forests performs reasonably well, and better than CART or logistic regression. Although less than 3% of the inmates studied over 24 months were reported for very serious misconduct, we are able to correctly forecast such behavior about half the time.

Keywords: prison, incarceration, misconduct, classification

1 Introduction

Inmate misconduct is a universal problem in state prisons across the United States. A key response is to design and employ classification procedures that identify inmates most likely to offend. It is then sometimes possible to implement a variety of prevention strategies. Among the most popular are housing systems that place higher risk inmates in more restrictive surroundings.

We argue below that although some of the existing classification systems have demonstrable effectiveness, they are also rather blunt instruments. In particular, they typically fail to identify with useful forecasting skill the very few inmates who are likely to commit the most serious offenses. In this paper, therefore, we use data from the California Department of Corrections (CDC) to examine how ensemble statistical procedures can be applied to find these especially problematic inmates. We also examine which attributes of such inmates are useful predictors. Because the goal is to forecast well, effective predictors may or may not have a causal interpretation. Nevertheless, identifying predictors that contribute substantially to forecasting skill necessarily raise questions about why they work and what they convey about the sources of serious misconduct in prison. And insofar as useful forecasting skill is demonstrated, the importance of true forecasting for criminology is underscored.

2 Background

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 to homicide.

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 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 $30,000 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 special inmate 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). However, CDC's classification system will find the inmates at high risk for engaging in *any* form of misconduct. Minor infractions are treated the same as major infractions. No special effort is made to identify inmates who are likely to commit very serious offenses while in prison.

The one-size-fits-all approach to classification 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. Moreover, 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.

It would seem to make good policy sense in general to identify inmates likely to engage in such behavior. But with the recent resurrection of rehabilitation as a important goal, there is added reason for finding the small fraction of inmates likely to put themselves and others in harm's way. There are few things that disrupt prison life more than violence, and even the threat of violence can have a chilling effect on a wide range of constructive prison initiatives. Identifying the small fraction of inmates likely to be responsible for the vast majority of very serious misconduct may help to free the rest of the inmates to participate in more constructive activities, and perhaps even facilitate the development of special rehabilitative programs for the "hard cases."

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). The population of interest is essentially male adults, convicted of one or more serious felonies, who are in close contact with other felons and housed in a special kind of very restrictive environment. The usual outcome is "misconduct," which is a mix of relatively minor rule infractions and incidents that would be serious crimes if committed outside of prison.

Not surprisingly, little use is made of conventional theory in criminology. Theories of criminal behavior are meant primarily to apply to crimes committed within the population at large in everyday settings. Although many of the concepts carry over to prison environments, there is a substantial disconnect in the details. For example, there is little apparent correspondence between a garden-variety robbery of a neighborhood convenience store and an inmate not reporting to a job assignment or failing to come to the front of his cell during a head count. Likewise, male-on-male sexual assault seems to be a uniquely salient feature of prison life.

Equally important for our purposes, there appear to be no studies focused on the most serious forms of prison misconduct, whether informed by theory or not. The existing research is directed toward the broad-gauged classification systems that are either in use or under development. How best to forecast the very most serious and rare kinds of prison misconduct is largely unexplored in the research literature.

3 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 system. Details are discussed in 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. Some predictors that turned out to be very important were not used for placement in the old system. Consequently, cases placed under the old system are not included in this analysis.

The data come from two sources. The predictors are taken from the new CDC intake form. Soon after an inmate arrives at the CDC reception center, the intake process begins. That process can last several weeks, depending on the information needed and the work load of prison staff. One of the key forms filled out during that period is an "839" (which is the form number), that is used to determine an inmate's classification score. That score is a key factor in the security level to which the inmate is assigned. All but the last variable in Table 1 are predictors taken from the "839."

Given how CDC records intake information at reception, all of predictors are binary with the following exceptions. Age at first arrest is a quantitative variable grouped into five cateogores. Age at arrival at the CDC reception center is a quantitative variable grouped into four categories. For both age variables, the categories are not all the same size so that the resulting scales are ordinal only. For example, the youngest category for age at first arrest is 0 to 17 years old, and the oldest is 36 years of age or older. The modal category is 18 to 21 years old. The youngest category for the age at CDC intake is 16 to 20 and the oldest is 36 years of age or older. The modal category is 36 years of age or older.[1]

Sentence length is recorded in years, capped at 50. That is, a death sentence and life without the possibility of parole are recorded at the upper

[1] For both age variables, the categories represent how CDC records such data. We could not get other age breakdowns.

Variable	Description
term	Current sentence length (median = 4.00 years)
gang	Associated with gang activity (proportion = 0.190)
agearr	Age at first arrest (mode = 18-21 years old)
agerec	Age at CDC intake (mode = 36 years old or older)
cya	Served time with California Youth Authority (proportion = 0.080)
mill	Diagnosed with mental illness (proportion = 0.08)
cdc	Served time previously with the CDC (proportion = 0.30)
jail	At least 31 days in jail or county youth facility (proportion = 0.66)
bgood	Good behavior in previous CDC incarceration(s) (proportion = 0.56)
bbad	Bad behavior in previous CDC incarceration(s) (proportion = 0.03)
infrac	Committed serious infraction in prison (proportion = 0.027)

Table 1: List of Variables

bound at 50. The number of years recorded is the nominal sentence given by the courts. The actual sentence served in most cases is somewhat shorter.[2]

Shown in the last row in Table 1, the response variable for each inmate was recorded subsequent to admissions. A "Rules Violation Report" (called a "115") is completed when prison staff observe an inmate engaging 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 crimes such as assault, drug trafficking, and robbery 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. The former is coded as "1," and the latter is coded as "0."

[2] The upper bound of 50 represents how CDC records sentence length information. We could not distinguish between the several different kinds of sentences all labeled as "50."

4 Data Analysis

Overfitting is a potential problem with all data analyses that include substantial exploration of the data. That exploration can be informal "data snooping" or searches that are guided by some computer algorithm (e.g., stepwise regression). The analyses adapt to idiosyncratic features of the data so that the results do not generalize well, even to other random samples from the same population. In this context, overfitting is said to result in "generalization error" (Hastie et al., 2001: 193-196).

In response, we began by holding out a random sample of 1000 from the total of 9662 inmates. The remaining 8662 inmates constitute our "training" data set. The 1000 inmates constitute a "testing" data set to be used to evaluate the results. How well will models constructed from the training data perform when applied to the testing data? The test sample has the added benefit of illustrating how the procedures used in the analyses to follow could be used in practice to inform decisions made by prison administrators.

4.1 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 at 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.

4.2 Using Random Forests

Could one do better with ensemble methods (Berk, 2005) as a special case of what Breiman (2001b) calls algorithmic modeling? We turned primarily 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 produce consistent estimations of the generalization error (Breiman, 2001c). As a result, random forests does not overfit. Moreover, there is growing 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).[3]

Random forests, using its default parameter settings, was applied to the training data set. In particular, the prior distribution of the response was taken to be the observed marginal distribution, and the ratio of the costs of false negatives to false positives is taken to be 1.0. Here, a false negative is classifying an inmate as not engaging in serious misconduct when he actually did. A false positive is classifying an inmate as having engaged in serious misconduct when in fact he did not.

Table 2 shows that when the test data are used to evaluate the results, random forests did no better than the marginal distribution of the response variable. Not a single inmate was correctly classified as engaging in serious misconduct. Indeed, all inmates were classified as not engaging in serious misconduct although 39 of 1000 actually did. Because of the highly unbalanced response variable, the default cost ratio of 1.0 could not be achieved.[4]

Moreover, conversations with officials from the California Department of Corrections suggested that the 1 to 1 cost ratio of false negatives to false positives was inappropriate. Rather, 1 false negative was worth about 10 false positives. That is, it was about 10 times more costly to fail to identify an inmate who would in fact commit an act of serious misconduct than to identify an inmate as one who would commit such an act when in fact he would not. For purposes of this analysis, we accepted the 10 to 1 cost ratio

[3] For a very accessible discussion of boosting, see Berk, 2005. For an in depth treatment, see Friedman, 2002; Manor et al., 2002; Buehlmann and Yu, 2002.

[4] With no false positives, the cost ratio was infinite.

	Forecast No Misconduct	Forecast Misconduct	Model Error
Observed No Misconduct	961	0	0.00
Observed Misconduct	39	0	1.00
Use Error	0.039	0.00	Overall Error = 0.039

Table 2: Random Forest Confusion Table with Default Costs

	Forecast No Misconduct	Forecast Misconduct	Model Error
Observed No Misconduct	753	208	0.216
Observed Misconduct	19	20	0.487
Use Error	0.024	0.912	Overall Error = 0.227

Table 3: Random Forest Confusion Table with 10 to 1 Costs

as appropriate.

There are several ways within random forests to take the cost ratio of false negatives to false positives into account. They typically lead to effectively the same results, except in cases such as this in which the response variable is highly unbalanced. Then, the best approach is to use stratified random sampling when for each classification tree a new data set is constructed. In effect, one over-samples for the rare cases so that the ratio of false negatives to false positives comes out about right.[5]

Weighting misconduct cases to no misconduct cases by a cost ratio of 10 to 1 produced the "confusion table" in Table 3. Note that there are 208 false positives and 19 false negatives for a ratio of a little more than 10 to 1 and forecasting accuracy into the test data set has improved substantially. Looking at the model error for the second row in the table, inmates who engage in serious misconduct are correctly forecasted a little more than half the time (1 - .487). Given that such misconduct occurs only about 2.5% of the time, this represents considerable forecasting skill. Looking at the model error for the first row in the table, forecasts of no serious misconduct are correct about 80% of the time (1 - .216). Combining the two, random forests with a cost ratio of 10 to 1 makes correct forecasts about 75% of the time

[5] Without the oversampling, one risks getting a number of bootstrap samples with none of the rare cases. The response variable is then a constant.

	Forecast No Misconduct	Forecast Misconduct	Model Error
Observed No Misconduct	837	124	0.129
Observed Misconduct	24	15	0.615
Use Error	0.028	0.892	Overall Error = 0.148

Table 4: Random Forest Confusion Table with 5 to 1 Costs

(i.e., 1 - .227).

There are a related set of conclusions for how accurate the forecasts would be in use. From the first column, if a forecast of no misconduct is actually made, it would be correct over 97% of the time (1-.024). Were a forecast of misconduct made, it would be correct about 9% of the time (1 - .912). The latter figure might seem disappointing, but accurately reflects the policy decision to accept a relatively large number of false positives.

To help put these results in context, Table 4 shows the confusion table when the cost ratio of false negatives to false positives is 5 to 1. While this ratio is substantially smaller than the cost ratio preferred by CDC officials, it could be be preferred by other stakeholders. The tradeoffs are readily apparent. Forecasting error for inmates engaging in serious misconduct has increased from .49 to .62. On the other hand, Forecasting error for inmates not engaging in serious misconduct has decreased from .22 to .13. Other figures in the table change in an analogous fashion.

At this point, a reasonable concern is whether the results are substantially procedure specific. Would another procedure that has many of the same desirable performance characteristics produce similar results? For this, we turned to boosted trees (Friedman, 2002), a statistical procedure that also makes many passes through the data. Here, there is neither random sampling of the training data nor random sampling of predictors. Rather, after each tree is grown, a function of the residuals is constructed that provides weights for the the next pass through the data. Observations that are incorrectly fit are given more weight and observations that are correctly fit are given less weights. Then, classification is determined by averaging over all trees so that trees that fit the data better overall are weighted more heavily in the averaging process.[6]

[6] For our software, there was no way to directly introduce costs into the algorithm. But there were fitted values one could interpret as estimates of the probability of serious misconduct. We set the threshold not at the usual .50, but at a value slightly below the

	Forecast	Forecast	
	No Misconduct	Misconduct	Model Error
Observed No Misconduct	785	176	0.183
Observed Misconduct	21	18	0.538
Use Error	0.0261	0.907	Overall Error = 0.197

Table 5: Boosting Confusion Table with 10 to 1 Costs

From Table 5, one can see that in this case boosting and random forests give very similar results. Random forests does a bit better forecasting the presence of serious misconduct and a bit worse forecasting the absence of serious misconduct. But both differences are well within random sampling error.

4.3 Which Predictors Matter for Forecasting?

If the goal is to forecast accurately, a key indicator of a variable's importance is its contribution to forecasting skill. Random forests provides this information by randomly shuffling each predictor in turn and then computing how much forecasting error increases. The shuffling makes the predictor unrelated to the response variable (and all other variables) on the average. The greater the increase, the more important the predictor. Figures 1 and 2 show the results for these data.

Figure 1 shows that term length makes the most important contribution to forecasting accuracy when serious misconduct is being forecasted.[7] If its values are shuffled, forecast error for inmates who engage in serious misconduct increases by over .06 (i.e. from about .49 to .55, based on Table 3). Exactly how term length is related to serious misconduct will be explored shortly.

observed proportion of cases for which serious misconduct was reported. This value was chosen to approximate the desired 10 to 1 balance of false positives to false negatives (implying that the costs of false negatives to false positives was 10 to 1). This is analogous to one of the methods for handling costs in random forest where the voting threshold would not be set at 50%, but at the marginal percentage for the response category that needed to be given more weight.

[7] The small negative values represent sampling error and are properly interpreted as effectively zero.

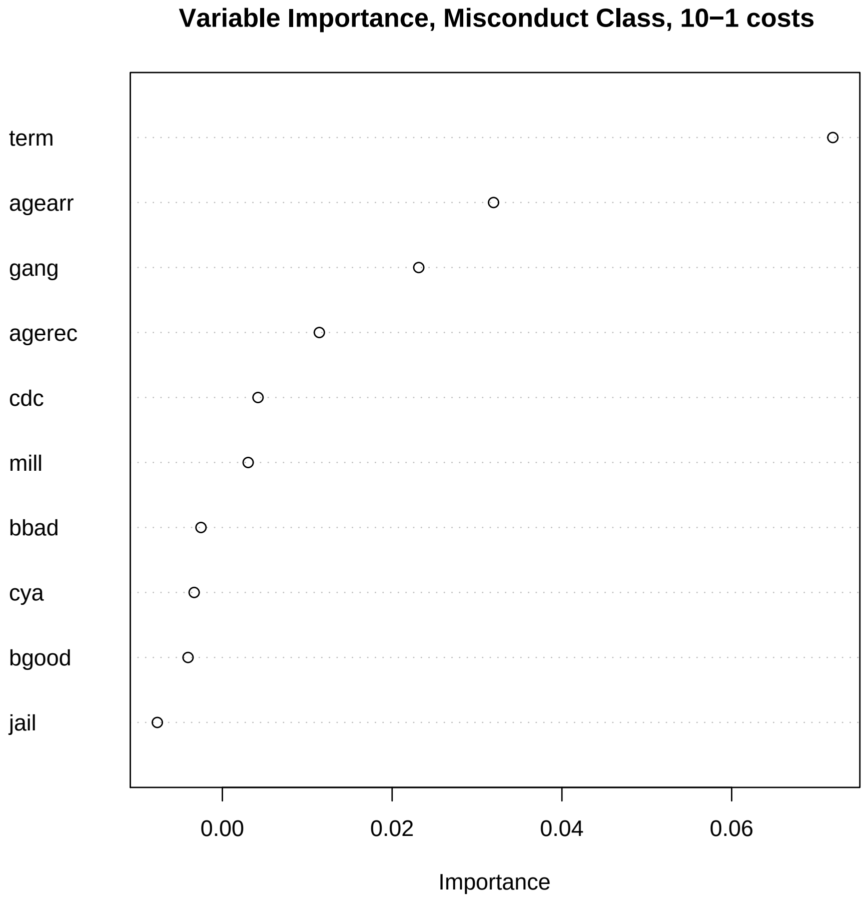

Figure 1: Average Reduction in Forecasting Skill for Serious Misconduct

Age at first arrest, age at reception, and gang activity are also important for forecasting accuracy. Shuffling each of their values increases forecasting error by more than .02. From past research on prison inmates, these contributions to forecasting skill are to be expected. (Berk et al., 2003). Age effects are also consistent with extensive work in criminology (Gottfredson and Hirschi, 1990, Sampson and Laub, 1990, Hamil-Luker et al., 2003). More details from these analysis are provided below.

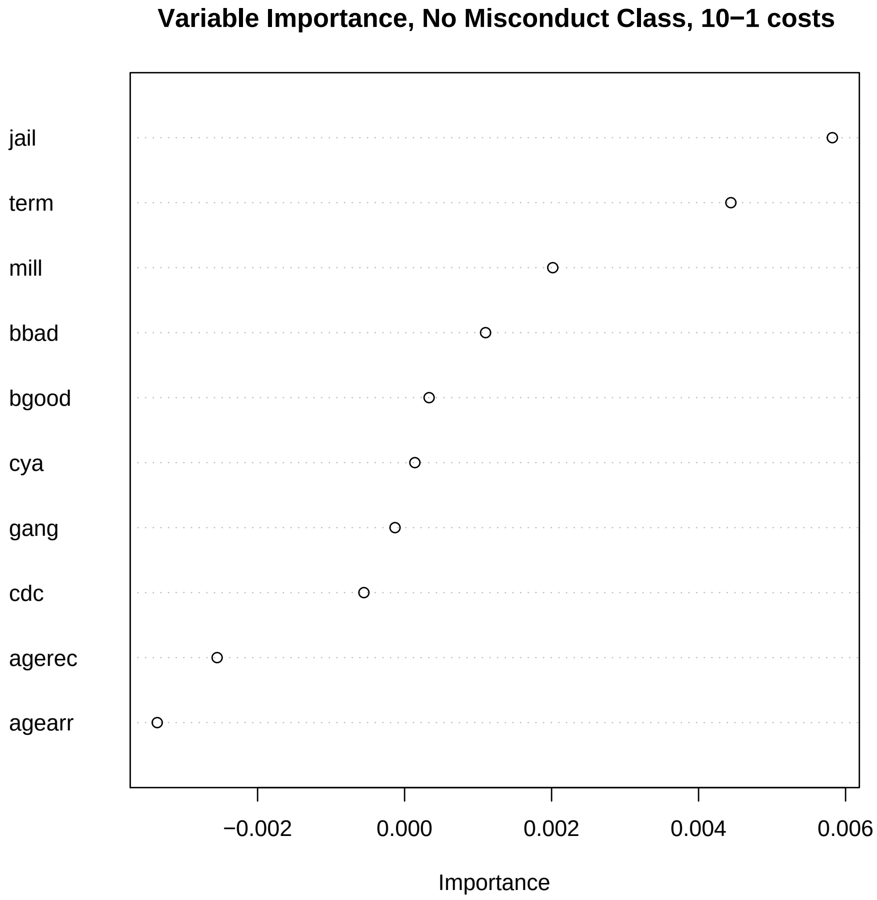

Figure 2: Average Reduction in Forecating Skill for No Serious Misconduct

Figure 2 shows that the pattern is some somewhat different for forecasting the absence of misconduct.[8] To begin, because there are so many more observations compared with inmates engaging in misconduct, one has to make many more forecasting errors to increase comparably the proportion of cases forecasted incorrectly. For example, suppose the number of incorrect forecasts is increased by 10 for both outcomes: serious misconduct and no serious misconduct. That increase will make a large difference in the proportion of cases inaccurately forecasted for acts of serious misconduct. That same increase will make a small difference in the proportion of cases inaccurately forecasted for no acts of serious misconduct. Therefore, importance for forecasting skill will be far smaller.

The order in which predictors are ranked by importance can also vary. Recall that forecasts are based on a vote over trees. For correctly classified observations, the vote can be close (e.g., 51% v. 49%) or lopsided (e.g., 80% v. 20%). When the vote is close, switching a few votes can change the forecast from correct to incorrect. When the vote is lopsided, a large number of votes need to switch for the forecast to change from correct to incorrect. It follows that cases for which the vote was close are the cases whose forecast is most likely to change when each predictor is shuffled. The degree to which certain predictors tend differentiate these observations from the rest will determine which predictors will have the greater impact on forecasting accuracy.

For serious misconduct, sentence length is best able to distinguish the observations with close votes from the other observations. It is, therefore, the most important predictor for forecasting accuracy. For the absence of serious misconduct, Figure 2 shows that serving one or more prior terms in a county jail tends to best distinguish the observations with close votes from the rest. As a result, it rises to the top of the list of predictors. Figure 2 also shows that for inmates not engaging in serious misconduct, term is still relatively important; it comes in second. But, the predictor gang activity falls to the middle of the pack, and the two age variables are at the bottom.

However, the declines in forecasting accuracy shown in Figure 2 really do not matter much. Given the highly unbalanced distribution of the response, it is easy using no predictors at all to forecast quite well an absence of serious misconduct. Consequently, having a few more or few less predictors does not make a meaningful difference, and the variation in predictor importance is

[8] The very small negative values again represent sampling error and are properly interpreted as effectively zero.

not especially instructive.

4.4 Response Functions

Within random forests, one can explore how the response is related to each predictor with partial response plots. They depict the functional relationship of each predictor with the response, all other predictors held constant. Partial response plots are analogous to added variable plots in regression, although the "partialing" is done by a form of matching rather than by covariance adjustments. (See Berk, 2005, for an accessible discussion.) For a binary outcome, the units on the vertical axis are in logits computed 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 3 shows how serious acts of misconduct are related to term length. One can see that the relationship is generally positive. However, there is effectively no consistent relationship for sentences of 5 years or less, followed by a large jump for sentences between 6 and 10 years. After that, the slope is more moderate. The general message seems to be that term length is a very useful predictor, but it has an especially large impact for sentences of moderate length. Interestingly, there is no real increase in slope for the very longest sentences despite the fact that inmates serving them are thought to have "nothing to lose."[9]

It is tempting to provide causal interpretations for Figure 3. However, causal interpretations would require trying to determine which attributes of sentence length that we have not measured are related to both sentence length and serious misconduct. For example, taking potential "good time" credits into account, sentences less than 5 years may make parole appear to be immanent. Consequently, there may be strong incentives to stay out of trouble. But short sentences are associated with a number of factors that could well be related to misconduct.

[9] Had the partial response plot for no misconduct been shown, it would just have been the mirror image. For binary responses, only one of the two possible partial response plots need be shown. This is not true when there are more than two classification categories. Then, there needs to be one partial response plot for each response category.

Similar analyses confirmed that each of the important predictors noted earlier had their anticipated relationships with the response. Serious misconduct is more common among inmates with longer criminal records, but especially those arrested initially in their early teenage years. Inmates involved in gang activity are far more likely to get into serious trouble. And younger inmates, especially those under 20, are worse risks. [10] None of these findings are a surprise and are fully consistent with reports from prison staff. Individuals who are likely to get into serious trouble on the street are high risks in prison.

5 Discussion

The inmates predicted to engage in very serious misconduct are broadly like the inmates likely to be placed in the more secure CDC facilities by the revised inmate classification system. The salient predictors are largely the very ones leading to high inmate classification scores (Berk et al., 2003). However, the inmates identified by random forests are considerably 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.

The high risk inmates tend to be young individuals with long criminal records, active participants in street and prison gangs, and sentenced to long prison terms. As a qualitative matter, none of these predictors is surprising. However, there is a useful story in the quantitative details and particular configurations of predictor values. Sentence length makes its most important difference for sentences between 5 and 10 years. Criminal record has its greatest impact for inmates arrested at a very young age. The very youngest inmates and inmates engaged in gang activity are also trouble. These patterns imply that very young inmates who nevertheless have managed to accumulate long criminal histories, who are active in gang activities, and who are serving sentences of more than 10 years, are a “perfect storm.” Another implication is that in contrast to much prison lore, “lifers” are not especially

[10] The partial response plots are not very interesting for the two age variables and for gang activity because the age variables are measured in just a few ordinal categories and gang activity is a binary variable.

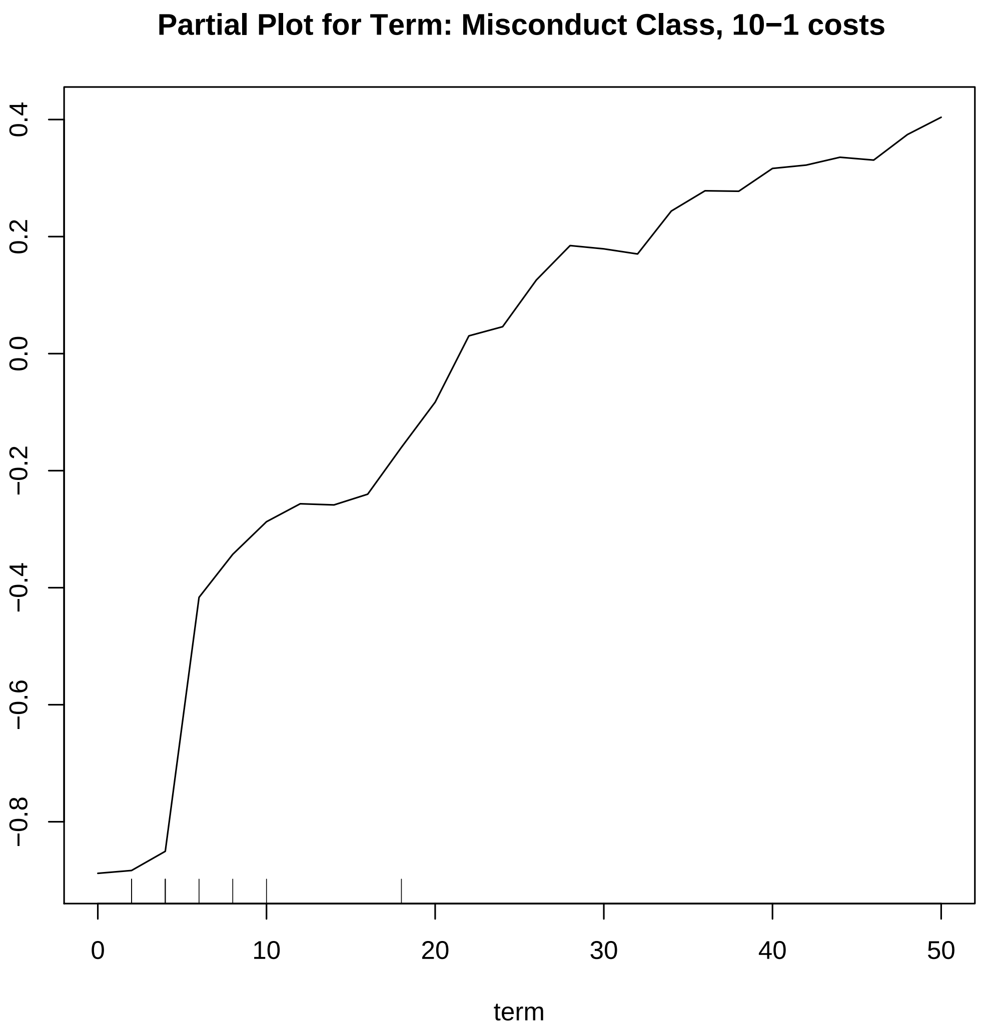

Figure 3: Serious Misconduct as a Function of Term Length

more dangerous than other inmates serving moderate to long sentences, even if it is really true that the lifers have "nothing to lose."

What could the CDC do with forecasts about inmates predicted to commit very serious crimes while in prison? 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 (CDC's highest security level). The CDC's current housing arrangements allow for about 5% of all inmates to be placed in Level IV facilities. Without a major and costly reorganization of existing space, this implies that were our forecasts used literally, Level IV bed capacity would be substantially exceeded. Thus, another kind of tradeoff becomes salient. Are the costs of failing to act on forecasts such as ours high enough to justify increasing the number of Level IV beds or moving less dangerous Level IV inmates to Level III facilities?

Interestingly, 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 watchtower. The 270 degree design allows more inmates to be 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 identified by our methods. However, the capacity of these prisons is about 1% of all inmates. Our forecasting methods would clearly overwhelm these facilities, even with a 5 to 1 cost ratio.

It may well be that a more fundamental reorganization of the the CDC's housing practices are in order. Earlier research (Berk and de Leeuw, 1998; Berk et al., 2003) indicates that the vast majority of inmates are not a significant threat to engage in the most serious kinds of misconduct. Although this is not the place to get into the details, one could easily imagine an effective three-level security system with perhaps 90% of the inmates housed in the lowest security level (e.g., inmates convicted of drug possession or property crimes). The remaining 10% of inmates who posed a serious threat to staff and to each other would then be housed in one of two very high security settings, depending on whether they were among the "worst of the worst."

Finally, it is important to appreciate our forecasts in the test sample will overestimate forecasting skill insofar as the mix of inmates and the precursors of serious misconduct change over time.[11] Because both are likely to change slowly, our conclusions might be useful for up to a decade from the time the data were collected. But sound practice argues for a revised analysis every 5 years or so. This would not be costly because data routinely recorded could be used.

6 Conclusions

We have found a small number of predictors that can help in forecasting serious inmate misconduct. These predictors have considerable face validity. They would no doubt make sense to custodial personnel and criminologists. A fair question, therefore, is what's the news?

First, the events to be predicted are rare and as a result, pose daunting technical problems. It is very difficult to improve upon forecasts made from the marginal distribution alone when the marginal distribution is highly unbalanced. Yet, we achieved considerable improvement. Perhaps our most important result is a demonstrable and high level of forecasting skill.

Second, the forecasting skill demonstrated could enhance prison practice. The forecasts could help distinguish between those relatively few inmates who put themselves and others at great risk and the bulk of the inmates who are just trying to do their time. A number of useful program implication would follow.

Third, the predictors that proved to be important are easily measured at prisoner intake and lead to a checklist that is easily applied. Consequently, the additional overhead for prison staff is small. Moreover, the checklist is simple so that errors in use should be few.

Finally, the forecasts were constructed consciously taking the relative costs of false negatives and false positives into account. This makes good policy sense despite being rarely done in criminology. The general practice is to ignore costs, which necessarily introduces costs anyway. Typically, a 1 to 1 cost ratio is imposed, whether recognized or not, that will often make little policy sense. Then, the forecasts will make little policy sense as well.

However, an explicit introduction of relative costs can cut in several different ways. The relative costs we used, while consistent with our conversations

[11] This is not caused by overfitting, which is measured by the decline of forecasting skill into a new random sample from the *same* population. Here the issue is a changing population.

with CDC Staff, are hardly definitive, and could vary with the times. Different relative costs could alter forecasting skill and the predictors that would be most important.

7 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 T. Cooley (1979) "A Dynamic Decision Theoretic Perspective on Modeling the Performance of the Criminal Justice System," *Social Sciences Research* 8:265-286.

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.

Berk, R.A. (2005) "An Introduction to Ensemble Methods for Data Analysis," *Sociological Methods and Research*, forthcoming.

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/

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.

Buehlmann, P. and Bin Yu (2002), “Analyzing Bagging.” *The Annals of Statistics* 30: 927-961.

Friedman, J.H. (2002) “Stochastic Gradient Boosting,” *Computational Statistics and Data Analysis* 38(4):367-378.

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 Early 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.