Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment*

Richard Berk
Justin Bleich
Department of Statistics
Department of Criminology
University of Pennsylvania

August 22, 2013

Research Summary

There is a substantial and powerful literature in statistics and computer science clearly demonstrating that modern machine learning procedures can forecast more accurately than conventional parametric statistical models such as logistic regression. Yet, several recent studies have claimed that for criminal justice applications, forecasting accuracy is about the same. In this paper, we address the apparent contradiction. Forecasting accuracy will depend on the complexity of the decision boundary. When that boundary is simple, most forecasting tools will have similar accuracy. When that boundary is complex, procedures such as machine learning, that proceed adaptively from the data will improve forecasting accuracy, sometimes dramatically. Machine learning has other benefits as well, and effective software is readily available.

Policy Implications

The complexity of the decision boundary will in practice be unknown, and there can be substantial risks to gambling on simplicity. Criminal justice

*Thanks go to Bill Rhodes and three anynomous reviewers for many helpful comments on this paper.

decision makers and other stakeholders can be seriously misled with rippling effects going well beyond the immediate offender. There seems to be no reason for continuing to rely on traditional forecasting tools such as logistic regression.

1 Introduction

Forecasts of recidivism have been widely used in the United States to inform parole decisions since the 1920s (Burgess, 1928; Borden, 1928). Of late, such forecasts are being proposed for a much wider range of criminal justice decisions. One important example is recent calls for predictions of "future dangerousness" to help shape sentencing (Pew Center of the States, 2011; Casey, 2011). The recommendations build on related risk assessment tools already operational in many jurisdictions, some mandated by legislation (Kleinman et al., 2007; Turner et al., 2009; Hyatt et al., 2011; Skeem and Monahan, 2011; Oregon Youth Authority, 2011). In Pennsylvania, for instance, a key section of a recent statute reads as follows.

> **42 Pa.C.S.A.§2154.7. Adoption of risk assessment instrument.**
> (a) General rule. – The commission shall adopt a sentence risk assessment instrument for the sentencing court to use to help determine the appropriate sentence within the limits established by law for defendants who plead guilty or nolo contendere to, or who were found guilty of, felonies and misdemeanors. The risk assessment instrument may be used as an aide in evaluating the relative risk that an offender will reoffend and be a threat to public safety.
> (b) Sentencing guidelines. – The risk assessment instrument may be incorporated into the sentencing guidelines under section 2154 (relating to adoption of guidelines for sentencing).
> (c) Pre-sentencing investigation report. – Subject to the provisions of the Pennsylvania Rules of Criminal Procedure, the sentencing court may use the risk assessment instrument to determine whether a more thorough assessment is necessary and to order a pre-sentence investigation report.
> (d) Alternative sentencing. – Subject to the eligibility requirements of each program, the risk assessment instrument may be an

> aide to help determine appropriate candidates for alternative sentencing, including the recidivism risk reduction incentive, State and county intermediate punishment programs and State motivational boot camps.
> (e) Definition. – As used in this section, the term risk assessment instrument means an empirically based worksheet which uses factors that are relevant in predicting recidivism.

With such widespread enthusiasm and very high stakes, one might assume forecasting accuracy has been properly evaluated and determined to be good. In fact, competent evaluations can be difficult to find for a wide variety of criminal justice decisions. Some of the problems have a long history (Ohlin and Duncan, 1949; Reiss, 1951; Ohlin and Lawrence, 1952). For example, it is relatively rare for evaluations to be based on "test data" that were not used to construct the forecasting procedures. The danger is grossly overoptimistic assessments. More recent commentaries have documented a number of other problems, sometimes including no evaluation at all (Farrington and Tarling, 2003; Gottfredson and Moriarty, 2006; Berk, 2012).

The need for thorough and thoughtful evaluations has become even more important over the past decade because in addition to calls for a more routine use of crime forecasts, new forecasting tools from computer science and statistics have been developed. Often supported by formal proofs, simulations, and comparative applications across many different data sets, these tools promise improved accuracy in principle (Breiman, et al., 1984; Breiman, 1996; 2001a; Vapnick, 1998; Friedman, 2002; Chipman et al., 2010).[1] For example, Breiman (2001a) provides a formal treatment of random forests and its comparative performance across 20 different datasets. There now several instructive criminal justice applications in print as well (Berk, 2012).

Yet, there are also several recent articles claiming that for criminal justice applications, the new tools perform no better than the old tools (Yang, 2010; Liu et al., 2011; Tollenaar and van der Heijden, 2013). Logistic regression (Berkson, 1951) is a favorite conventional approach. The conclusion seems to be "why bother?" For criminal justice forecasting applications, the new procedures are mostly hype.

> "The conclusion is that using selected modern statistical, data mining and machine learning models provides no real advantage

[1] Very accessible treatments can be found in a number of textbooks (Bishop, 2006; Berk, 2008; Hastie et al., 2009).

> over logistic regression and LDA.[2] If variables are suitably transformed and included in the model, there seems to be no additional predictive performance by searching for intricate interactions and/or non-linear relationships" (Tollenaar and van der Heijden, 2013).

How can the proofs, simulations and many applications provided by statisticians and computer scientists be so wrong? How can it be that statistical procedures being rapidly adopted by private firms such as Google and Microsoft and by government agencies such as the Department of Homeland security and the Federal Bureau of Investigation are no better than regression methods readily available for over fifty years? Why would the kinds of new analysis procedures being developed for analyzing a variety of datasets with hundreds of thousands of cases (Dumbill, 2013; National Research Council, 2013: Chapter 7) not be especially effective for a criminal justice dataset of similar size?

A careful reading of the technical literature and recent criminal justice applications suggests that there can be a substantial disconnect between that technical literature and the applications favored by many criminal justice researchers. Statisticians and computer scientists sometimes do not distinguish between forecasting performance in principle and forecasting performance in practice. Criminal justice researchers too often proceed as if the new procedures are just minor revisions of the generalized linear model. In fact, the conceptual framework and actual procedures can be very different and require a substantial change in data analysis craft lore. Without a proper appreciation of how the new methods differ from the old, there can be serious operational and interpretative mistakes.

In this paper, we try to improve the scientific discourse by providing an accessible discussion of some especially visible, modern forecasting tools that can usefully inform criminal justice decision-making. The discussion is an introduction to material addressed far more deeply in *Criminal Justice Forecasts of Risk: A Machine Learning Approach* (2012), written by the senior author. We also try to provide honest, apples-to-apples performance comparisons between the newer forecasting methods and more traditional approaches.

For some readers, it may be useful to make clear what this paper is not about. As one would expect, there have been jurisprudential concerns

[2] "LDA" stands for linear discriminant analysis.

about "actuarial methods" dating from at least the time when sentencing guidelines first became popular (Messinger and Berk, 1987; Feely and Simon, 1994), and more recent discussions about the role of race have introduced an important overlay (Harcourt, 2007; Berk, 2009). The issues are difficult and real. They are also not addressed in this paper. Our concerns are more immediate. Forecasts of future dangerousness are being developed and used. Real decisions are being made affecting real people. At the very least, those decision should be informed by the best information available. And that information depends significantly on the forecasting procedures deployed.

2 Proper Criminal Justice Forecasting Comparisons

The conceptual foundation for criminal justice forecasting can easily be misconstrued (Ridgeway, 2013). We begin, therefore, with a fundamental conceptual point that some readers may at first find counterintuitive. As a formal matter, one does not have to understand the future to forecast it with useful accuracy. Accurate forecasting requires that the future be substantially like the past. If this holds, and one has an accurate *description* of the past, one has an accurate forecast of the future. That description does not have to explain why the future takes a particular form and certainly does not require a causal interpretation. Readers comfortable with traditional time series analysis (Box iand Jenkins, 1970), should have no problem with this reasoning.

It follows that there is a key distinction between forecasting and explanation that has been badly conflated in some accounts (Andrews et al., 2006). Understanding a phenomena may lead to improved forecasting accuracy, or it may not, but forecasting and explanation are different enterprises that can work at cross-purposes. For example, explanatory models should be relatively simple and provide instructive interpretations. Such models can leave out a large number of weak predictors that one-by-one do not enlighten but *in the aggregate* dramatically improve forecasting accuracy. Common practice implicitly folds such variables into the disturbance term. Alternatively, such predictors, often called "nuisance variables" in limited information structural models, are associated "nuisance parameters" and given "minimal attention" (Cameron and Trivedi, 2005: 36). Similar issues arise if simple, easily inter-

pretable functional forms (e.g., linear) are used when complex functional forms might fit the data somewhat better.[3]

The approach we take is to maximize forecasting accuracy, and that is the premise on which the underlying mathematics depend. We take this approach because it leads to clear performance criteria and various proofs of optimal forecasting accuracy for a given dataset. Such clarity is an undeniable virtue about which more will be said shortly.

Equally important, there are a wide variety of decisions made by criminal justice officials in which a *necessary condition* is the best possible forecasting accuracy. Consider a judge's decision to sentence an offender to either incarceration or probation. Pennsylvania's statute states that a "risk assessment instrument may be used as an aide in evaluating the relative risk that an offender will reoffend and be a threat to public safety." Presumably, accuracy really matters. Imagine the ethical and legal implications of using a particular risk tool to justify a long incarceration when there exist more accurate risk tools from which a sentence of probation could be more appropriate. There is also no requirement in the legislation that a judge understand why an individual is high or low risk. Indeed, it is not even clear what a judge would do with such information.[4] Other examples, include pre-trial decisions to release defendants on bail or decisions by parole boards to release under supervision inmates who have not served their full terms. One could also imagine forecasts of future dangerousness helping to determine charging decisions by prosecutors.

Thus, there is no formal concern in this paper with why certain predictors improve forecasting accuracy and no attempt to interpret them as explanations for the forecasted behavior. For example, if other things equal, shoe size is a useful predictor of recidivism, it can be included as a predictor. Why shoe size matters is immaterial. In short, we are not seeking to identify risk factors that may or may not make any subject-matter sense. That can be a useful enterprise, but it is a different enterprise.

Indeed, if the enterprise really is explanation, than some form of structural equation modeling may be called for. There is an extensive and largely

[3]Some differences in jargon can be instructive. In machine learning a "predictor" is often called an "input," and a response or dependent variable is often called a "target."

[4]In the special case when there are clear indications of substance dependency or psychological problems, a judge might order treatment along with the sentence. But such conditions are not necessarily risk factors for many kinds of crime, and indications of need can be sufficient.

unrebutted literature highly critical of structural equation modeling in general. An excellent, accessible, and technically sound treatment can be found in David Freedman's textbook *Statistical Models* (2005). We cannot rehash the issues here except to stress that machine learning is not a form of structural equation modeling and should never be interpreted as such.[5] Moreover, if the goal is to use one or more risk factors to design and test interventions, many would argue that the only sound approach is randomized experiments or very strong quasi-experiments.

2.1 Some Common-Sense Requirements for Fair Forecasting Comparisons

If one intends to compare the forecasting performance of different forecasting tools, there are several basic, common-sense requirements. These provide the ground rules.

1. One must be clear on what features of forecasting procedures are being compared. As we explain below, "black box" forecasting methods may forecast with remarkable accuracy and provide decision makers with tools that can be enormously helpful (Breiman, 2001b). But black box forecasting methods may have little to say about which risk factors matter most. If the goal is to compare different procedures by their forecasting accuracy, forecasting accuracy should be the benchmark.

2. Forecasting comparisons must be based on data not used to construct the competing forecasting procedures. Such data are often called "test data," and accuracy is often called "out-of-sample performance." Data used to build the forecasting procedures can be called "training data." If training data are also used as test data, all comparisons risk contamination through overfitting (Hastie et al., 2009: 219-226). As already noted, this point has been appreciated for well over 50 years, but is often ignored.

[5] A structural equation model is an algebraic theory of how nature generated the data and as such, can be right or wrong. Machine learning employs algorithms that seek some well defined empirical goal, such as maximizing forecasting accuracy. There is no structural model. Concerns about whether the model is correct are irrelevant. What matters is how well the algorithm performs.

3. Proper performance criteria must be used that are the same across competing methods. For example, measures of fit are not appropriate if the competition claims to be testing forecasting accuracy. In addition, there are many different measures of forecasting performance (Hastie et al., 2009, chapter 7), and the same measure should be used for all of the competitors. For example, the area under a receiver operating characteristic curve (ROC) provides very different information from that available through direct estimates of generalization (forecasting) error (Hastie et al., 2009: 314-317).

4. All of the forecasting competitors should be accurately characterized if comparisons are to be properly understood. For example, there are a number of forecasting procedures represented as state-of-the-art that actually are not. There are also forecasting procedures characterized as machine learning that actually are not. Classification trees, for instance, (Brieman, et al., 1984) is neither state-of-the-art nor a machine learning technique. AdaBoost (Freund and Schapire, 1995) is a machine learning procedure, but was state-of-the-art 15 years ago. Bayesian additive regression trees (Chipman et al., 2010) can be considered state-of-the-art, but is not formally within machine learning traditions. Random Forests (Breiman, 2001a) is state-of-the-art and a machine learning procedure.[6]

5. Many of the popular forecasting procedures have tuning parameters that researchers can use to improve forecasting accuracy.[7] In addition,

[6]What qualifies as state-of-the-art can certainly be debated, but within sensible boundaries, there can be remarkable consensus. For example, random forests is certainly not the newest machine learning procedure, but for a wide range applications nothing else seems to consistently perform better. Likewise, sharp distinctions between machine learning, statistical learning and a variety of other related procedures are increasingly difficult to defend and probably not worth quarreling over (National Research Council, 2013: 61). Nevertheless, within somewhat fuzzy boundaries, there can be widespread agreement.

[7]Tuning parameters can be set at particular values to improve the performance of a given statistical procedure (National Research Council, 2013: 70-73). In the estimation of a logistic regression, for instance, the convergence threshold of the iteratively reweighted least squares algorithm is a tuning parameter. It needs to small enough to produce a close approximation to a maximum likelihood estimate, but not so small that unnecessary iterations are performed. Another example is a decision in stepwise regression to fix the number of predictors that can be included in the final model. In forecasting settings, tuning parameters usually are chosen in service of forecasting accuracy.

sometimes researchers do not understand that in their effort to maximize forecasting accuracy they are implicitly tuning their procedure. Fair comparisons require that all competitors are tuned in a comparable fashion. This can be difficult because the tuning is often based on principles that can depend on the particular forecasting procedure being used.

6. All forecasting competitions are necessarily data dependent and can vary across different applications. Forecasting competitions do not reveal fundamental and invariant forecasting truths. To take a simple example, a procedure that performs poorly in small samples may be a star in large samples because its best properties only materialize asymptotically. Appropriate caveats should be attached to the results of all forecasting comparisons.

7. Performance differences across different forecasting procedures must be thoughtfully evaluated. This will often mean a careful consideration of *how a forecasting procedure will be used*. A small difference in forecasting accuracy can translate into a difference of hundreds of crimes. Academic researchers may not care. But stakeholders surely do. There is also the equally important matter of taking uncertainty into account. Some apparent differences wash out in new realizations of the data. They are just chance artifacts.

8. It should go without saying, but all of the forecasting procedures must be implemented correctly. There is ample evidence that too often this is not the case (Berk, 2012).

3 Some Conceptual Fundamentals

We turn now to a conceptual overview of classification and forecasting. The intent is to provide a very accessible, didactic overview that can apply to a very broad range of forecasting procedures used previously in criminal justice applications. Readers interested in a technical discussion should consult the references cited.

Consider the decision of whether or not to release an individual on parole. Since the 1920's, such decisions have often been informed by forecasts of whether a given inmate will be arrested for a new crime soon after release.

The forecasts are shaped by actuarial procedures applied to information from inmates who had been released in the past. In effect, profiles are developed that can classify inmates by whether they succeeded or failed on parole. These profiles are used to forecast parole outcomes when they are not yet known. In the next few pages, we provide a basic, nontechnical overview of how this can be done. We build on a prior treatment written for criminal justice researchers (Berk, 2012) and on more formal textbook discussions as needed (Bishop, 2006; Hastie et al., 2009).

3.1 The Basic Account

Figure 1 is a very simplified and initial plot illustrating how classification and forecasting can be undertaken. The red circles represent individuals who have failed on parole in the past. The blue circles represent individuals who have succeeded on parole in the past. There are two predictors in this illustration. One predictor is the number of prior arrests. The other predictor is the number of rule infractions during the most recent incarceration. Both can be seen as "dynamic" predictors, but "static" predictors would have not materially changed the discussion. Figure 1 can be seen as a 3-dimensional scatterplot.[8]

The statistical task is to impose a "decision boundary" on the 2-dimensional predictor space that can be used to define two classes: those who fail and those who do not. The term "decision boundary" is used because the intent is to directly inform actual decisions.[9] Statistical procedures that partition the data into different grouping are often called "classifiers." In this instance, the partitioning should result in the fewest classification errors possible. For Figure 1, there will necessarily be two regions defined, one for failures and one for successes. Ideally, the failure region has no successes, and the success region has no failures. Usually, one has to settle for less.

[8]The meanings of "dynamic predictors" and "static predictors" can depend on the context and the decision to be informed by the forecast. For example, the difference between statistic and dynamic predictors plays a key role in the fairness of parole decisions. Is it appropriate to use static predictors already employed at sentencing when later parole decisions are made? Is there a risk of unfair "double counting"? Thus, the crime that sent an individual to prison is static. Should it be also used to help inform parole decisions? In contrast, time in a prison secure housing unit (SHU) is in this context dynamic. There would be no concerns about double counting if it were empoyed by a parole board.

[9]The underlying mathematics is shaped by the same goal.

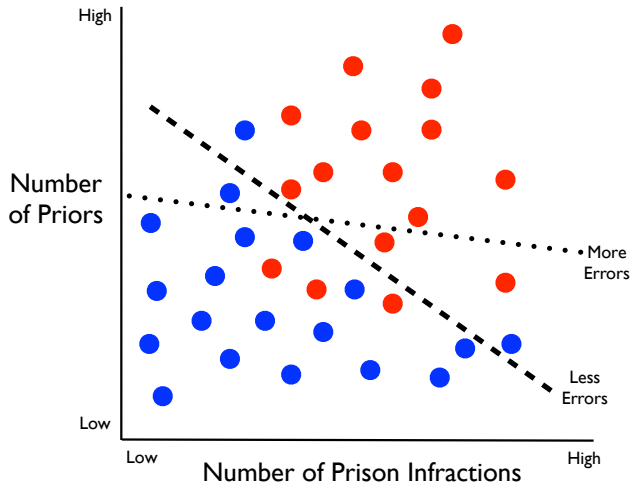

Figure 1: Two Linear Decision Boundaries in 2-Dimensional Predictor Space

The dotted line is one possible linear decision boundary. In the region above the dotted line, failures predominate by a count of 13 to 2. So, that region is assigned the class of "failure." In the region below the dotted line, successes predominate by a count of 17 to 5. So, that region is assigned the class of "success."

The assigned classes can be used for forecasting. When a new case is found for which a forecast is needed, that case is placed in one region or the other depending on its values for the two predictors. For example, a case with a very large number of priors and a very large number of prison infractions would be placed in the "failure" region to the upper right, and a forecast of failure would be made. A decision to impose a stiff prison sentence could follow.

The dotted decision boundary results in several classification errors. There are 2 (blue) successes classified as failures, and 5 five (red) failures classified as successes. Overall, there are 7 errors for 35 cases, which means that the classification procedure is right about 80% of the time. In real applications, this would be considered very good performance.

The dashed line is another attempt to accurately separate the successes

from the failures. Above this alternative linear decision boundary, the majority of cases once again are failures. Therefore, the class of "failure" is assigned to that region of the figure. Below the alternative linear decision boundary, the majority of cases are successes. Therefore, the class of "success" is assigned to that region of the figure. Now there are only five misclassified cases: 2 blue circles are in the red region and 3 red circles are in the blue region. The new boundary produces correct classifications about 85% of the time, and on those grounds is likely to be preferred to the old boundary.

As before, any cases with predictor values that place them above the decision boundary, but whose outcomes are not yet known, are forecasted to be failures. Similarly, any cases with predictor values that place them below the decision boundary, but whose outcomes are not known, are forecasted to successes. From a classification exercise comes a forecasting procedure. The forecasts, in turn, are used to inform parole decisions.

How might one arrive at the best linear decision boundary? If the two outcomes are coded as 1 or 0, and conventional linear regression is applied using the two predictors as regressors, one important kind of optimal linear decision boundary can be imposed on the predictor space. That line is defined by fitted values of .50. Cases with regression fitted values greater than .50 are assigned one class and cases with regression fitted values equal to or less than .50 are assigned the other class. By minimizing the sum of squared residuals and imposing a fitted value threshold at .5, one is also minimizing the sum of the classification errors (Hastie et al., 2009: 20-22).

Alternatively, one can apply logistic regression. The same basic reasoning works. When the response is represented as the log of the odds of the category coded as 1, there is again a linear decision boundary in "logit" units. The threshold is a logit of 0.0 (Hastie et al., 2009: 102), which in a probability metric is .50. Forecasting accuracy may be better or worse than for linear regression. Linear regression assumes that in the metric of the 1/0 outcome, relationships with the predictors are linear. Logistic regression assumes that in the metric of the 1/0 outcome, relationships with the predictors are S-shaped (i.e., the cumulative logistic function). Which of these leads to better forecasts in a given setting will usually be an empirical matter. Both functions are typically arbitrary because there will rarely be compelling subject-matter theory requiring one or the other.[10]

[10]Linear and quadratic discriminant function analysis has much in common with logistic regression and has been used in criminal justice risk assessments. We do not consider linear

3.2 Building in Differential Forecasting Error Costs

To this point, all classification errors are given equal weight. A success classified as a failure counts the same as a failure classified as a success. This is why the least squares regression minimizes the number of forecasting errors. In many criminal justice settings, the assumption of equal weights is not responsive to the preferences of stakeholders. For example, the consequences of forecasting a parole success for an individual who will fail can be far more serious than forecasting a parole failure for an individual who will be a success. The parole failure may entail a heinous crime. Failing to release an individual who would be crime-free leads to increased time behind bars. Both forecasting errors are costly, but for many stakeholders, the costs to victims of a heinous crime are far greater than the costs of extra prison time. Whether or not these relative costs generally hold, an assumption that all forecasting errors have equal costs is likely to be unrealistic.[11]

And costs matter for forecasts meant to inform real decisions. Figure 2 shows why. Using the broken line as the decision boundary, there are two successes that are incorrectly classified as failures. For this illustration, suppose that stakeholders think that the costs of "over-incarceration" are greater than the costs of crimes committed while on parole. There are reasons, therefore, to upweight the blue mistakes relative to the red mistakes. We show this in Figure 2 by making the two blue mistakes much larger. A new linear decision boundary results. Least squares regression can be used as before. But the decision boundary shifts toward the upper right with perhaps also a change in the slope.

The two blue mistakes are now accurately classified as successes. They no longer count as errors. But in trade, there are now five rather than three misclassified red circles. It looks like a wash — there are two fewer successes classified as failures, and two more failures classified as successes. But it is not a wash. The new decision boundary is to be preferred because the original two blue mistakes were much more costly than the two new red mistakes.

If the new decision boundary is preferred, many of the forecasts can

---

or quadratic discriminant function analysis because one must assume that the predictors have a multivariate normal distribution (Hastie et al., 2009: section 4.3). This is unrealistic for most predictors in criminal justice settings, especially when any of the predictors are categorical.

[11] A more complete discussion about the role of asymmetric costs is beyond the scope of this paper. An excellent treatment can be found in a special issue of the Albany Law Review, edited by Shawn Bushway (2011).

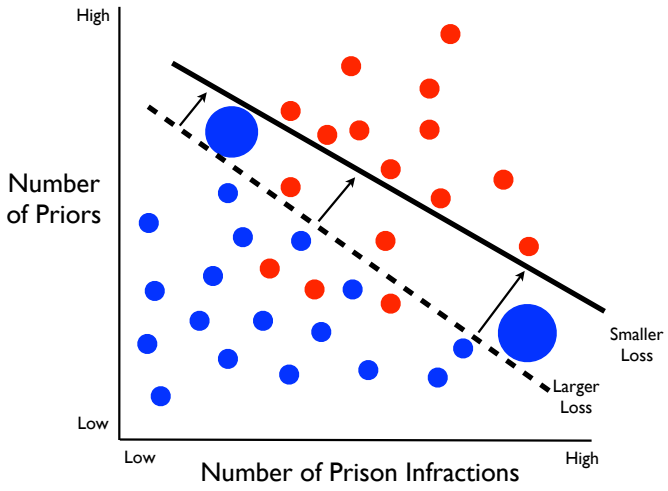

Figure 2: Impact of Asymmetric Costs in 2-Dimensional Predictor Space

change. In this example, cases to be forecasted as failures will need a greater number of priors and a greater number of prison infractions than previously. The increase will be larger for the number of prison infractions because the new decision boundary was shifted outward more for the infractions predictor.

The point is that not all forecasting errors are created equal, and the relative costs of different kinds of forecasting errors should be built into any classification/forecasting procedure. To ignore this issue is to assume equal costs. And if equal costs are not consistent with stakeholder preferences, the forecasts will not be properly responsive. Misleading forecasts can result.

3.3 Nonlinear Decision Boundaries

Why be limited to linear decision boundaries? Nonlinear boundaries can in principle perform better. In Figure 3, we reproduce much of Figure 1, but now with a nonlinear decision boundary shown by the dotted line. There are no red circles falling below the nonlinear decision boundary, and no blue circles falling above the nonlinear decision boundary. Classification is perfect. The prospects for forecasting accuracy look very promising indeed.

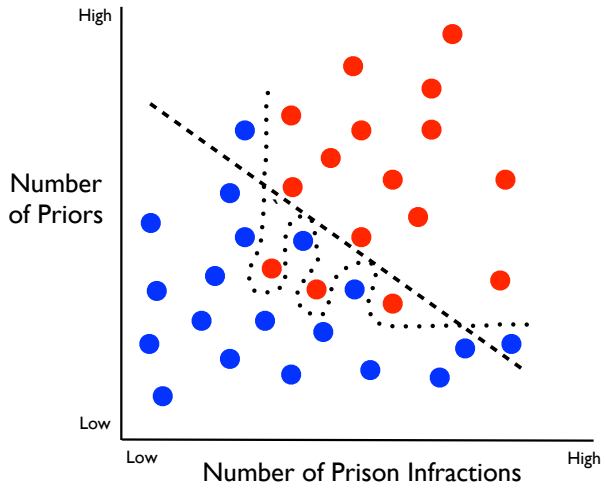

Figure 3: A Linear and Nonlinear Decision Boundary in 2-Dimensional Predictor Space

The linear decision boundary is far less complex than the nonlinear decision boundary.[12] The price for greater simplicity is more classification errors. Clearly, one's ability to classify accurately is enhanced when the decision boundary can be more complex. It is easier for the nonlinear decision boundary to respond to complicated data structures.

A sensible statistical aim, therefore, can be to use predictors in a manner that allows for nonlinear decision boundaries as needed. There can be two related approaches (National Research Council, 2013: 63). For parametric procedures such as logistic regression, greater complexity can in principle be addressed by including a larger number of predictors. Transformations of predictors can help. For instance, one might include not just the age of an inmate, but some polynomial function of age. One might even break up age

[12]There seems to be no consensus on how best to define the amount of complexity. One popular approach is the degrees of freedom used to construct the decision boundary. In this example, the nonlinear decision boundary would use many more degrees of freedom than the linear decision boundary. A closely related approach is link complexity to the "effective dimension" of the statistical procedure or in some cases, the data itself (National Reserch Council, 2013: 70).

into a set of binary dummy variables. Statistical interactions might also be captured with products of variables. The point is that the capacity to address greater complexity needs to be built in from the beginning or determined later in a set of very effective exploratory procedures. Also required is that the requisite predictors are included in the dataset. Many would argue that these requirements cannot be met in practice.

For nonparametric procedures such as smoothing splines (Hastie et al., 2009: section 5.4), one may include as many predictors as possible, along with promising transformations, but the *procedure* attempts to determine the decision boundary complexity needed. At one extreme, the fitted values are a hyperplane (just as in conventional linear regression). At the other extreme, the fitted values are an interpolation between all data points. The former is much less complex than the latter. In practice, some result between these extremes is typical. In contrast to parametric methods like logistic regression, an *adaptive* process is used to arrive at a decision boundary — the procedure exploits information in the data to determine both the shape and location of a decision boundary.[13] Unless a researcher is close to prescient and has the data rich enough to constructively respond, adaptive procedures start with a substantial forecasting advantage.[14]

But there is a downside to adaptively determined decision boundaries. As a greater number of degrees of freedom is used up for a given sample size, there is the real risk of increased instability in the results. There is less information available per procedure parameter. In addition, there can be overfitting in which the procedure responds to idiosyncratic features of the

[13]Stepwise regression is an example of a very simple adaptive procedure within a conventional regression framework. But again, distinctions may not be sharp. When researchers respecify their models after looking at the results, the final model is shaped by data-informed induction. Some would say that the difference is that the model selection process is not built into the data analysis algorithm itself.

[14]If resources allow, a parametric brute force approach may help to level the playing field. With thousands of observations and hundreds of predictors, one can in addition construct *a priori* many nonlinear transformations and interaction variables. In effect, the researcher tries to anticipate how a effective adaptive procedure could respond. All of the original predictors and new transformations can then be included in a single "kitchen sink" regression. The regression will likely be uninterpretable. The complexity and multicollinearity alone could be toxic. If model selection procedures are applied to simplify, one is doing a seat-of-the-pants adaptive modeling with all of its attendant problems (Berk et al., 2010). Why settle for a brute force approximation to the desired procedure? An example can be found in the recent paper by Tollenaar and van der Heijden (2013).

data. Because forecasting involves new data, not the data used to develop the decision boundary, forecasting accuracy can be disappointing. The procedure does not generalize well to new data, which is precisely what forecasting entails.

For example, an individual with a large number of priors and a large number of prison misconducts may have a high probability of failure on parole. But a high probability is not a certainty. If that individual does not fail, a complex decision boundary would try to accurately classify that individual as a success. As a result, an anomalous case inconsistent with most of the data would help shape the decision boundary. When that decision boundary is then used for forecasting with data in which such anomalous cases were absent, the decision boundary would not perform as well. It would be unnecessarily complex and risk an increase in forecasting errors. Looking back at Figure 3, if any one of the 3 red circles had as little as one or two more prison infractions or priors, the red circle would have fallen above the linear decision boundary, and one of the fingers in the nonlinear decision boundary would not have been constructed.

There are useful responses to overfitting, often called "shrinkage" or "regularization." The intent is to reduce the instability. With smoothing splines, for instance, the fitting function is penalized for increases in complexity (Hastie et al., 2009: section 5.4). In a least squares context, the residual sum of squares is increased so that what might be the smallest sum of squared residuals no longer is the smallest. A residual sum of squares that starts out being larger, but has a smaller penalty because of less complexity, can be the preferred minimizer. In other words, a price is put on complexity that does not substantially improve the fit.

Another approach, called bagging (Breiman, 1996), capitalizes on a large number of random samples with replacement from the data on hand. A classification procedure is applied to each sample, and the results are averaged across samples. One important consequence is that idiosyncratic results tend to cancel out.

Finally, in this illustration, the two predictors have substantive interpretations. In general, parolees with a great number of prior arrests and a greater number of prison infractions are more likely to fail on parole. However, any substantive insights are a bonus. The primary goal is to classify accurately because that can lead to the most accurate forecasts. With respect to that goal, the two predictors could as well be longitude and latitude. This allows for the possibility of using "black box" classification procedures,

for which no apologies need be made. One does not have to rely a "structural model" when forecasting is the primary motive. Indeed, the requirement of a structural model can undercut forecasting accuracy (Breiman, 2001b). Two different masters are being served.

In summary, when forecasting accuracy is the primary goal, parametric approaches such as logistic regression can in principle perform as well as nonparametric approaches when the best decision boundary is relatively simple, and when the predictors required by the correct model are available in their proper form. When the best decision boundary is complex and/or the requisite predictors are not all available, nonparametric procedures will forecast more accurately, often substantially more accurately.

3.4 Enter Machine Learning

Where does machine learning come in? Machine learning, sometimes when called "statistical learning," can be viewed as a special form of nonparametric regression. The goal can be to find the "right model." But when machine learning is used strictly as a forecasting procedure, the connections to conventional regression models become very distant indeed. As will soon be discussed in more detail, there is no structural model even in principle.

The transition to machine learning can confer a number of important benefits, some of which are not readily available otherwise.

1. One is not limited to classifiers able to forecast one of two outcome categories. In some recent applications, for instance, parole outcomes are forecasted for three classes: an arrest for a violent crime, an arrest for a crime that is not violent, and no arrest (Berk et al., 2010). Increasingly, criminal justice agencies want to forecast more than the binary outcome of any arrest versus no arrest (Berk, 2012). The kind of arrest really matters. In particular, arrests for crimes of violence are distinguished from other kinds of arrests.

2. Forecasting errors that do not have equal costs can be introduced into the procedure at the beginning so that all of the results properly represent the preferences of stakeholders (Berk, 2011).

3. Regularization is often built directly into the procedure to increase forecasting accuracy (Hastie et al., 2009: chapter 5, section 8.7).

4. Highly unbalanced distributions for the classes to be forecasted create no special problems as long as the rare outcomes are important enough to be given extra weight in the analysis. For example, in some recent work for individuals primarily on probation, the outcome classes to be forecasted included a class for homicide or attempted homicide, which represented only about 2% of the outcomes (Berk et al., 2009a; 2009b).

5. Some procedures work well and in a principled manner with an enormous number of predictors and even when there are more predictors than cases (Hastie et al., 2009: chapter 15).

4 The Forecasting Contestants

We will compare the forecasting performance of three different classifiers: logistic regression, random forests, and stochastic gradient boosting. Logistic regression represents business as usual over the past 50 years. It is a special case of the generalized linear model, and very familiar to criminal justice researchers. Random forests (Breiman, 2001a) and stochastic gradient boosting (Friedman, 2002) represent true machine learning procedures based on ensembles of classification trees. Both are nonparametric, rest on solid mathematical foundations, and both have been widely battle tested. All of the evidence to date indicates that they can perform well in criminal justice applications (Berk, 2013). All three are worthy competitors.[15]

4.1 Forecasting Class Membership with Logistic Regression

Logistic regression, sometimes called binomial regression, is a special case of the generalized linear model. As such, it is meant to represent how nature generated the data — it is an algebraic translation of subject-matter theory. In that sense it is a "structural model," and forecasting can be little more

[15]There are other worthy competitors such as Bayesian neural nets (Hastie et al., 2009: Section 11.9) and support vector machines (Hastie et al., 2009 Chapter 12). They are not considered here for lack of space and the need to introduce a substantial amount of new technical material. Suffice it to say that they too are well equipped to address complex decision boundaries and should have foresting skill roughly comparable to random forests and stochastic gradient boosting. But comparisons are difficult because a new suite of tuning parameters is introduced.

than an afterthought. Nevertheless, if the theory is correct and its algebraic representation is consistent with the theory, accurate forecasting can result.

Forecasting is undertaken through the regression's fitted values. These can either be in logit (i.e., log odds) units or probability units. Researchers typically use the probabilities when forecasting. To get from the probabilities to a forecasted class, a single threshold must be applied. For example, it is common to use a threshold of .50. Probabilities greater than .50 are assigned one outcome class (e.g., failed on parole). Probabilities less than or equal to .50 are assigned the other outcome class (e.g., succeeded on parole).

The threshold of .50 implies that the costs of false negatives and false positives are the same. As already noted, they are usually not the same. Suppose a "positive" is a person who commits a violent crime. Suppose a "negative" is a person who does not commit a violent crime. It follows that if false negatives are three times more costly than false positives, one should use a threshold of .25. Cases with predicted probabilities greater than .25 are forecasted to be violent offenders. Cases with predicted probabilities equal to or less than .25 are forecasted to not be violent offenders. It is three times easier for a person to be forecasted a violent offender than a nonviolent offender or no offender at all ($.75/.25 = 3$).

Altering the threshold *only affects the step from probabilities to classes.* All of the other logistic regression output is computed under the assumption that false negatives have the same costs as false positives. In particular, the logistic regression coefficients would almost surely be different had the actual relative costs of false negatives and false positives been properly taken into account. It can be a serious error, for instance, to use the regression coefficients as weights for constructing risk assessment instruments.[16]

Based on the logistic regression model results the risk factors were assigned weights or points. The points included 1 point for all factors, with the exception of Two or More Failure to Appear Convictions, which was assigned

[16]They are in logits units, not probability units. If one follows the common practice of exponentiating the regression coefficients and intercept, one is now at working in odds units. In addition, the regression coefficients and intercept are then multipliers and do not represent additive weights. If the intent is to obtain risk factor weights in probability units, one must go back to the original nonlinear logistic model. But because of the nonlinear functional form, there is not one weight for each risk factor — there is a limitless number. So that strategy fails too. It is also possible to ignore the regression coefficients and weight risk factors by simply "assigning weights or 'points' " (VanNostrand and Rose, 2009: 9). The statistical foundation for that approach is obscure.

2 points due to the predictive strength of the risk factor. The points were totaled to create a score from 0 to 10. The scores were then used to create risk levels. As a result, the VPRAI consists of five risk levels including low, below average, average, above average, and high as shown in the following figure.

Finally, logistic regression can only be used for binary outcomes. These days, criminal justice stakeholders often want much more — they want to forecast different kinds of crimes. As already noted, in some applications the intent is to work with three crime categories: arrests for violent crimes, arrests for crimes that are not violent, and no arrest at all (Berk et al., 2010). In the context of probation supervision, one motivation is to move supervisory resources from individuals who do not threaten public safety to individuals who do, a strategy that has been shown to work well (Berk et al., 2010). When there are more than two outcome classes, multinomial logistic regression may be an option, but there are a number of unresolved issues about how best to go from predicted probabilities for each class to the classes themselves.

4.2 Random Forests

A random forest is an ensemble of classification trees. The classification trees are an intermediate product used because they fit the data adaptively. They have no stand-alone role, and in the end are effectively invisible. They disappear into a machine learning black box through the follow algorithm.

1. A random sample of size N is drawn with replacement from a "training" dataset. Observations not selected are retained as the "out-of-bag" (OOB) data to later serve as "test data." On average about a third of the data will be OOB. The growing process for the first classification tree then begins.

2. A small sample of predictors is randomly drawn (e.g., 3 predictors).

3. After selecting the best split as usual from among the randomly selected predictors, the first partition is determined. There are then two subsets of the data that together maximize the improvement in the Gini index.

4. Steps 2 and 3 are repeated for all later partitions until the fit does not improve or the observations are spread too thinly over terminal nodes.

5. The Bayes' classifier is applied to each terminal node to assign a class — the class for each terminal node is determined by the class in the node that has the largest number of cases.

6. The OOB data are "dropped down" the classification tree. Each observation is labeled with the class assigned to the terminal node in which it lands. The result is the predicted class for each observation in the OOB data for that tree.

7. Steps 1 through 6 are repeated many times to produce a large number of classification trees. There are often 500 trees or more.

8. For each observation, the class assigned is determined by "vote" over all trees in which that observation is OOB. The class with the most votes is chosen. That class can be used for forecasting when the predictor values are known but the outcome class is not.

The adaptive nature of classification trees helps to reduce bias. In addition to the predictors used as inputs, there are "derived" predictors constructed as needed. The sampling of training data and predictors serves as a form of regularization that can improve the stability of class assignments and help make those assignments more independent over trees. Averaging over trees enhances both results. Finally, the use of OOB data helps to prevent overtfitting. In the end, random forests does not overfit as the number of trees in the random forest increases. A formal proof can be found in Breiman's seminal paper on random forests (2001a).

There are several ways to introduce asymmetric costs. Perhaps the best way is to employ stratified sampling in step 1. There is one stratum for each outcome class. Sample sizes for each stratum are determined so that some outcome classes are oversampled and some are undersampled. In effect, the oversampled classes are given more weight as each tree is grown, which in turn will affect the balance of false negatives to false positives. That balance captures relative costs. For example, if there are 10 false positives for every false negative, false negatives are necessarily 10 times more costly than false positives.

In addition to "confusion tables" in which forecasted outcomes from OOB data are cross-tabulated with the observed outcomes, there are measures of the contribution to forecasting accuracy for each predictor, and plots that show the way in which each predictor is related to the response, holding all

other predictors constant. The details are beyond the scope of this paper but some examples are provided later. (See, for example, Berk, 2008 for the details.)

4.3 Stochastic Gradient Boosting

Stochastic gradient boosting proceeds by applying a "weak learner" repeatedly to the data. After each pass through the data, all observations are reweighted, giving more weight to observations that were more difficult to accurately classify. The fitted values from each pass are used to update earlier fitted values. The weak learner is "boosted" to perform as a strong learner. Here is an outline of the algorithm for a binary outcome coded numerically as "1" for failure on parole or "0" for success on parole.

1. The algorithm is initialized with fitted values for the binary outcome. The overall proportion of cases that fail is a popular choice.

2. A random sample without replacement is drawn from the training data with a sample size of about half the sample size of the training data.[17]

3. The "negative gradient" (sometimes called the "pseudo-residuals") is computed. Just like with usual residuals, each fitted value is subtracted from its corresponding observed value of 1 or 0. The residual is a *quantitive* outcome variable within the algorithm: $(1-p)$ or $-p$, where p is the overall proportion coded as "1."

4. Using the randomly-selected observations, a *regression* tree is grown to fit the pseudo-residuals.[18]

5. The conditional mean in each terminal node is the estimate of the probability of failure.

6. The fitted values are updated by adding to the existing fitted values the new fitted values weighted to get the best fit.

[17]The goal is much the same as the sampling with replacement used in random forests. A smaller sample is adequate because when sampling without replacement, no case is selected more than once; there are no "duplicates."

[18]The procedure is much the same as for classification trees, but the fitting criterion is the error sum of squares or a closely related measure of fit.

7. Steps 2 through 7 are repeated until the fitted values no longer improve by a meaningful amount. The number of passes can in practice be quite large (e.g., 10,000), but unlike random forests, stochastic gradient boosting can overfit. Some care is needed because there is formally no convergence.

8. The fitted probability estimates can be transformed into outcome classes just as they were for logistic regression.

9. When forecasts are needed for new cases, they are constructed from the aggregated fitted values and their relationships with the predictors.

Like random forests, stochastic gradient boosting capitalizes on random samples of the training data, adaptive fitting tree by tree, and aggregation over trees. However, asymmetric costs can only be introduced at the end when probabilities are transformed into classes. Experience to date suggests that it can forecast about as well as random forests.

4.4 A Simulation

Logistic regression can forecast well when it is able to capture the data structure. However, logistic regression is not adaptive and depends on the researcher to specify an effective model. Important nonlinearities and interaction effects must be anticipated and included using the available predictors. If the researcher lacks the requisite insight or data, logistic regression will necessarily stumble. In contrast, adaptive procedures such as random forests or stochastic gradient boosting can shine because both algorithms are designed to search for structure with each pass through the data.

Figure 4 shows a fictitious dataset constructed to illustrate when logistic regression will perform poorly and random forests or stochastic gradient boosting will perform well.[19] It is by intent a worst case scenario for logistic regression and is not meant to represent in general the relative merits of the forecasting competitors. We are trying to address *why* nonparametric methods can forecast better than parametric methods. The exercise is didactic.

[19]The lessons learned can be applied far beyond logistic regression to any parametric regression approach. The lessons also apply to a wide range functions that have clear structures, but are very difficult for parametric regression models to capture.

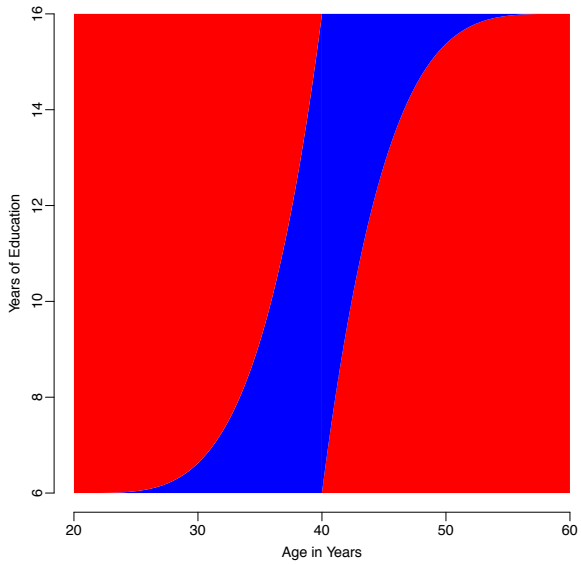

Figure 4: A Very Challenging Classification Example

There are 100,000 observations. The outcome is binary. Red is coded 1 and blue is coded 0. There are two predictors. The 2-dimensional predictor space contains a blue area that is homogeneously successes and two red areas that are homogeneously failures. The graphical conventions are no different from those used for the earlier figures except that the colored circles for individual observations are replaced by solid colors for different regions. It is as if we have printed a very large number of overlapping red circles and a very large number of overlapping blue circles. However, the data structure is far more complex because the blue region has red regions to its left and its right. Complex data structures of this sort are routinely analyzed in the classification literature (Hastie et al., 2009), but usually with many more than two predictors so that visualizations such as Figure 4 are unavailable. Any researcher trying to arrive at the correct parametric model from an examination of a scatter plot would necessarily be flying blind.

The surface was built by first drawing one predictor from a uniform distribution. The second predictor was constructed as a power function of the first. Then the predictor space was partitioned to show an interaction effect: both predictors had be high or low for the area to be red. That is, there are

nonlinear effects and an interaction effect. Because each of the three regions is perfectly homogeneous, the data provide a clear and compelling signal that a good classifier should be able to accurately detect.

After the fact, one might overlay the following subject-matter account. The outcome is whether or not a parolee finds employment. The blue area contains successes and the red area contains failures. On the horizontal axis is age in years. The young and the old do not do well. The vertical axis is years of education. The association is not strong but parolees with a lot of education or very little do slightly better. In addition, when the educational level is higher, the best ages for finding work are older.[20]

Why might such patterns occur? The kinds of positions for which parolees apply and the kinds of employers who would hire them represent a very limited subset of all jobs. By and large, the positions will involve physical labor for which not much experience or skill is required. Pay will be low and the work will be hard. Younger parolees may not be inclined to seek such positions, and older parolees may be incapable doing the work. Education may be largely irrelevant for most of the jobs a parolee will seek. But, those who have very little education may correctly target their job search only for menial positions. Those with more education may correctly understand that they have a wider range of employment options. Finally, having more education may give some older workers, who would have difficulty working at demanding menial jobs, the chance to take entry level white collar positions (e.g., taking orders and making change at fast food restaurants).[21]

This *post hoc* account may well be wrong, perhaps very wrong. The intent is to provide a less abstract setting in which to think about each contestant's performance. By itself, the story has no impact whatsoever on how well a given classifier performs. Any good classifier should forecast with near perfect accuracy. Unlike in real data, there is no noise.

When logistic regression is used, both regression coefficients are virtually zero.[22] Logistic regression is unable to extract any useful information from

[20]Plots of this sort may be unfamiliar and at first difficult to interpret. For the main effects, one has to do an eyeball integration over the variable whose role is not being described. For example, to gauge the marginal association between age and employment, one must consider vertical slices of the data and what fraction of each area is blue. Similar reasoning applies to years of education, but now the slices are horizontal.

[21]Some preliminary analyses we are doing for the program "Ready, Willing & Able" supported by the Doe Fund, are consistent with this account.

[22]The two regression coefficients are -.03 and -.01. Even with 100,000 observations, one

the two predictors. All that remains is the intercept, which is effectively the logit of the outcome variable's proportion of reds (i.e., .80). The distribution of the predicted probabilities ranges from .7958 to .8022. The predicted probabilities have almost no variability.

For didactic purposes and with no important loss of generality, we assume that the costs of false negatives are the same as the costs of false positives. The corresponding threshold of .50 is applied. It follows that forecasting error is minimized by always predicting red. 20% of the time the forecast would be wrong. The true reds would be forecasted with 100% accuracy, and the true blues would be forecasted with 0% accuracy. Table 1 shows the results.[23]

	Predict Blue	Predict Red	Model Error
Actual Blue	0	20078	1.0
Actual Red	0	79922	0.0

Table 1: Logistic Regression Confusion Table Using Simulated Test Data

Suppose a researcher is astute enough to include in advance the product of the two predictors to capture an interaction effect. Our reading of criminal justice forecasting applications is that such interactions are rarely used, but it is useful to see how logistic regression performs when given an especially good opportunity to deliver.

Table 2 shows the results. Although there are now nonzero regression coefficients for all three regressors, there are still no predicted probabilities smaller than .5. As before, forecasting error is minimized by always forecasting red. Nevertheless, there is some meaningful information in the predicted probabilities, and with cost ratios that weight forecasting errors for blue cases more heavily than for red cases, some blue cases will be correctly predicted.[24] For example, if a cost ratio of 4 to 1 is used, actual blues and actual reds are

cannot reject the null hypothesis of 0.0 for either. In an odds multiplier metric, both coefficients are very close to 1.0.

[23]Other thresholds would not change the performance of logistic regression. A threshold a very little bit below .80 would allow some blues to be correctly forecasted. The price would be a commensurate increase in reds forecasted incorrectly. Virtually no predictive information from the predictors is being used. The predictors might as well be ignored.

[24]The predicted probabilities now range from .5333 to .9384.

both correctly forecasting about 2/3rds of the time. That may seem quite good, but for these data the appropriate target is perfection.

	Predict Blue	Predict Red	Model Error
Actual Blue	0	20078	1.0
Actual Red	0	79922	0.0

Table 2: Logistic Regression with Interaction Confusion Table Using Simulated Test Data

How does an adaptive machine learning procedure perform? For illustrative purposes, we take random forests as our machine learning champion.[25] Table 3 shows the results for random forests assuming equal costs. With respect to the cost ratio, we are comparing apples to apples. The same two predictors are used, but there is *no* product variable for an interaction effect. The researcher using random forests is not allowed to be as clever as the researcher using logistic regression — random forests begins with a model specification disadvantage. Still, random forests is just about perfect. Given either outcome, random forests forecasts correctly more than 99% of the time. The failure to be literally perfect results from randomness in the random forest algorithm itself.

	Predict Blue	Predict Red	Model Error
Actual Blue	19975	102	0.005
Actual Red	92	79830	0.001

Table 3: Random Forests Confusion Table Using Simulated Test Data

The implications of this forecasting contest are clear. When the data structure is complex, machine learning procedures can perform very well. An adaptive process that "learns" from data can be very effective. This is

[25]We used the procedure *randomForest* in R, originally written by Leo Breiman and Adele Cutler and ported to R by Andy Liaw and Matthew Wiener. To the best of our knowledge, there is no implementation of random forests in any of the popular statistical packages such as SPSS, STATA, or SAS. Salford Systems has a procedure they call random forests, but the source code is proprietary, and it is difficult to know exactly what is being done. Also, according to the current Salford Systems website, the available version of random forests will not run on a Mac computer.

precisely what the large literature in statistics and computer science says. Logistic regression and other parametric forecasting procedures will not perform as well unless the researcher is able to construct a parametric model that captures all of the significant features of the data structure. As already noted, this can be a daunting task.

5 An Empirical Example

We turn now to analyses of real data. The dataset was selected to be typical of those recently used in parole or probation settings. Recall, however, that it is very difficult with real data to arrive at results that are broadly generalizable.

5.1 Forecasting Arrests for Serious Crimes

The data address how well parolees manage under supervision. There are 20,000 observations in the training data and 5,000 observations in the test data. We consider whether an individual is arrested for a serious crime within 2 years of release on probation. Serious crimes include murder, attempted murder, rape, aggravated assault, and arson. About 13% fail by this definition. Such crimes are of widespread concern. Static and dynamic predictors include:

1. Date of Birth;
2. Number of Violent Priors as an Adult;
3. Earliest Age for a Charge as an Adult;
4. Total Number of Priors as an Adult;
5. Earliest Age for a Charge as a Juvenile;
6. Total Number of Priors as a Juvenile;
7. Number of Charges for Drug Crimes as an Adult; and
8. Number of Sex Crime Priors as an Adult.

There is nothing special about these predictors. They represent the usual kinds of information that is routinely available on parolees when they begin their supervision. From past experience, they can make important contributions to forecasting accuracy (Berk, 2012).

We first apply logistic regression to the training data. A threshold of .135 is imposed on the predicted probabilities in order to arrive empirically at a 5 to 1 cost ratio of false negatives to false positives. Table 4 is the confusion table that results when the model is applied to test data. From the column on the far right, about 44% of the true failures are misclassified and about 32% of the true successes are misclassified. The forecasting accuracy is within the range of recent studies with similar data (Berk, 2012) and could well be useful for decision-makers.

	Predict Fail	Predict No Fail	Model Error
Actual Fail	378	302	0.444
Actual No Fail	1385	2935	0.321

Table 4: Logistic Regression Test Data Confusion Table for Serious Crime

Table 5 is the confusion table for random forests using the test data. The procedure was tuned to also arrive at a cost ratio of about 5 to 1 for false negatives versus false positives. From the column on the far right, about 37% of those who actually fail are incorrectly identified and about 28% of those who actually do not fail are incorrectly identified. Forecasting accuracy for random forests appears to be superior.

	Predict Fail	Predict No Fail	Model Error
Actual Fail	427	253	0.372
Actual No Fail	1196	3124	0.277

Table 5: Random Forests Test Data Confusion Table for Serious Crime

Table 6 is the confusion table for stochastic gradient boosting using the test data.[26] A threshold of .13 was used on the predicted probabilities from

[26]We used the R procedure *gbm*, written by Greg Ridgeway. There are several tuning parameters that can make a difference, and we are not certain that the comparisons are fully fair. To the best of our knowledge, there is no implementation of stochastic gradient boosting in any of the popular statistical packages.

the training data to empirically arrive at a cost ratio of about 5 to 1. From the column on the far right, about 42% of those who actually fail are incorrectly identified and about 32% of those who actually do not fail are incorrectly identified. Stochastic gradient boosting does appreciably better than logistic regression when forecasting failures, but only slightly better when forecasting successes.

	Predict Fail	Predict No Fail	Model Error
Actual Fail	396	284	0.418
Actual No Fail	1361	2459	0.315

Table 6: Stochastic Gradient Boosting Test Data Confusion Table for Serious Crime

It appears that across the three tables, random forests performs better than logistic regression and stochastic gradient boosting. This is consistent with published studies (Berk, 2012). But one must not overstate what is learned from the comparisons we report. It is difficult to guarantee that after tuning, one is necessarily comparing apples to apples. We have tried to insure that for all practical purposes, the false negative to false positive cost ratios are the same for all three procedures. But the cost ratios are not identical, and it is essentially impossible to make them so. The test data and training data are *different* random splits of the available dataset. Tuning done on the training data will carry over a bit differently to the test data, depending on the forecasting procedure. Moreover, each procedure was tuned with its own special set of tuning parameters. There is no guarantee that the results are fully comparable. Indeed, it is not even clear how to define such a thing.

Another important issue is whether the differences are large enough to matter. As already explained, that judgement depends on the application. For example, the agency from which these data were obtained supervises about 40,000 individuals on probation each year. About 5000 of these individuals are arrested for a serious crime within 24 months, most within less than a year. For failures, the difference of approximately 7% between the accuracy of logistic regression compared to random forests translates into about 350 serious crimes. Roughly 50 of those will be homicides or attempted homicides, the perpetrator of which could be identified in advance by random

forests, but not by logistic regression. In this instance, stakeholders found the practical difference in forecasting accuracy dramatic.

If one is looking for firm conclusions about forecasting accuracy from our results and others, it is almost certain that properly applied, random forests will always do at least as well as logistic regression and much of the time meaningfully better. Stochastic gradient boosting will do at least as well as logistic regression, but is somewhat less likely to dominate it.

There are several other reasons why random forests should be the forecasting method of choice, given currently available alternatives. For this illustration, the success category included individuals who were arrested for crimes not defined locally as "serious" and individuals not arrested at all. This is, of course, less than ideal. In fact, one of the goals of the supervising agency was to identify low risk offenders who could be supervised less intensively with no increased risk to public safety. Resources recaptured from the low risk offenders could then be allocated to the high risk offenders. To address this policy preference, random forests was applied using three outcome categories: an arrest for a serious crime, an arrest for a crime that was not serious, and no arrest at all. Three outcome classes are not an option for logistic regression. Forecasting accuracy for the low risk offenders was very good, implying that about half of the agency's case load could be minimally supervised. A reorganization of the supervisory practices followed, and a subsequent evaluation showed that re-arrest rates for the low risk individuals were not higher than under the previous, more intensive supervision regimes (Berk et al., 2010).

Random forests also provides output that can help explain how the forecasting works in practice. Recall, logistic regression coefficients, for instance, are estimated under equal costs and can be misleading if the costs of false negatives and false positives differ. In place of regression coefficients, random forests provides estimates of each predictor's contribution to forecasting accuracy. How this is done is beyond the scope of the paper, but is explained in many published papers and texts (e.g., Brieman, 2001a). Figure 5 is an example of the output that easily can be obtained.

Date of birth makes the largest contribution to forecasting accuracy for those who are arrested for a violent crime. The value of a little over .08 means that if date of birth is not allowed to contribute to forecasting accuracy, model error increases from about .37 in Table 5 to .45. The contributions of all other variables are smaller, with sexual priors contributing little or nothing.

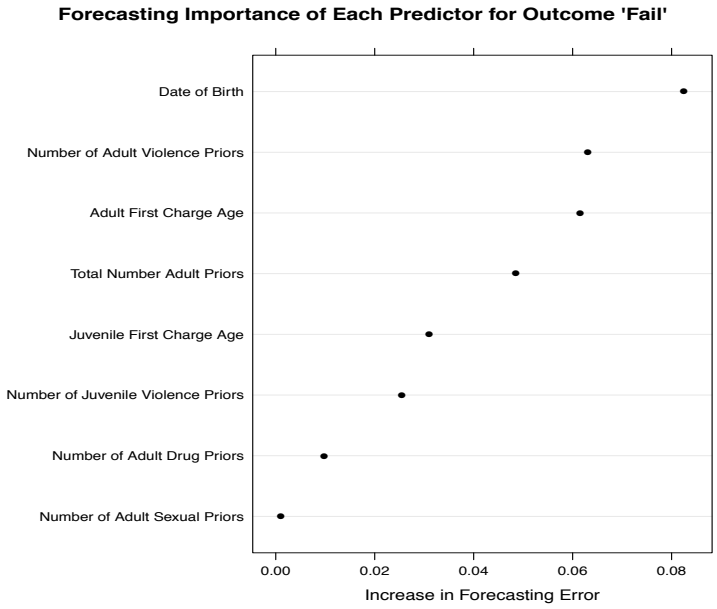

Figure 5: Random Forests Variable Importance Plot

Stakeholders have found this kind of information very useful. However, forecasting accuracy does not identify risk factors in the usual sense. A given predictor will often be transformed in many different ways, including as a component of interaction effects. All of these roles are combined when contribution to forecasting accuracy is computed. One has in test data the *net association* between a given machine learning input and the outcome being forecasted. There is currently no way to represent each role separately.[27]

There are also "partial response plots" showing how each predictor is related to each outcome class, with all other predictors held constant. Again, the details are beyond the score of this paper, but easily found elsewhere (e.g., Berk, 2012). Figure 6 is an example.

The predictor is the age at which the first arrest as an adult occurred. The response is being subsequently arrested for a serious crime while on probation. Units on the vertical axis are centered logits. The details need not concern us here — movement in the vertical direction means that the

[27]Recall that each tree in the random forest can transform each predictor differently. If there are, for instance, 500 trees, a given variable may be transformed in 500 different ways.

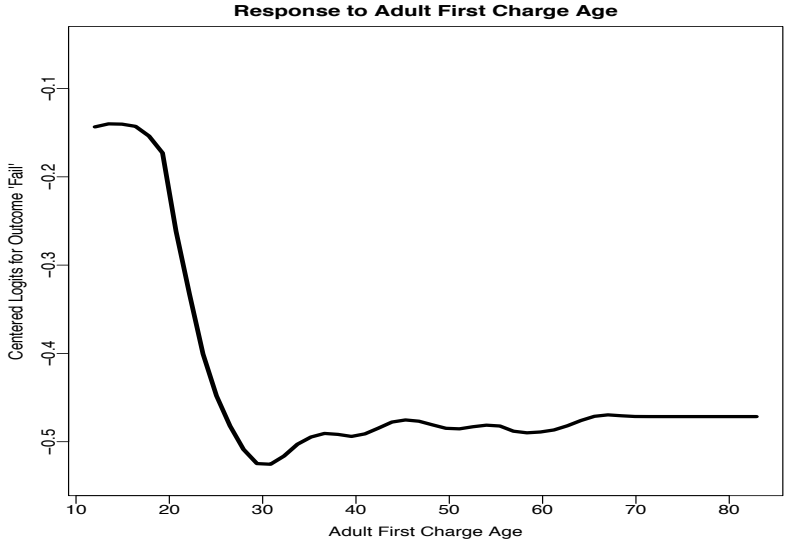

Figure 6: Random Forests Partial Response Plot for "Adult First Charge Age"

probability of failure increases.

The figure shows that the chances of an arrest for a serious crime are high for parolees whose first arrest as an adult occurred at a very young age. Starting in the late teens, those chances decline rapidly. For parolees whose first arrest occurred after age 30, increases beyond that in age of first arrest do not matter. In random forests, partial response plots are available for all predictors. For categorical predictors, the plots are bar charts.

Just as with contributions to forecasting accuracy, partial plots also do not identify risk factors in the usual sense. Each plot captures an average across trees in the forest and across each term in which that predictor is used. So age at first arrest is related to failure on parole in the manner shown in Figure 6, but all sorts of potentially important relationships involving that variables (e.g., interaction effects with gender) are masked.

6 Some Implications for Use

The forecasting output from machine learning classifiers is a forecast for given individuals and the sorts of descriptive output just discussed. Decision-makers "drop" an individual's predictor values into an algorithm, and a

forecast is computed in real time. There is no explicit use of risk factors, whether weighted or not. Thus, the algorithm must be live on some computer when forecasts are needed. Ideally, that computer is part of a network connected electronically to databases containing the predictor values. Then, a decision-maker may need only to enter an individual's unique ID number for appropriate predictor values to be properly downloaded into the machine learning algorithm. Experience to date indicates that such arrangements are well within the capabilities of IT personnel in many criminal justice settings (Berk, 2012).

7 Conclusions

Complex decision boundaries pose a significant challenge for logistic regression or any other parametric classifier. To forecast well, a researcher must understand the nature of the complexity, be able to properly translate that knowledge into an algebraic expression, and then have the data to construct an appropriate model. These are daunting requirements for criminal justice applications.

In contrast, adaptive machine learning procedures have the capacity to empirically discover patterns in the data and construct suitably complex decision boundaries. The requirements are a conventional menu of predictors and a large enough sample to exploit them. The tree-based machine learning procedures we have reviewed can then perform well and have several other important assets that logistic regression lacks: the capacity for outcome categories with more than two classes, a natural way to build in the asymmetric costs of forecasting errors, and a variety of instructive output that builds in asymmetric costs.

In practice, performance differences between logistic regression and most machine learning procedures can be small if the true decision boundary is simple. But how would one know? If logistic regression is used because a simple decision boundary is incorrectly assumed, substantial forecasting accuracy can be forfeited. In criminal justice settings where real lives can be at stake, the consequences could be significant. Why take the risk?

References

Andrews, D.A., Bonta, J., and Wormith, J. S. (2006) "The Recent Past and Near Future of Risk and/or Need Assessment." *Crime & Delinquency* January: 7–24.

Berk, R.A. (2007) "Meta-Analysis and Statistical Inference." (with commentary) *Journal of Experimental Criminology* 3(3): 247–297.

Berk, R.A. (2008) *Statistical Learning from a Regression Perspective.* New York: Springer.

Berk, R.A.(2009) "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1: 231–242.

Berk, R.A. (2011) "Asymmetric Loss Functions for Forecasting in Criminal Justice Settings." *Journal of Quantitative Criminology* 27: 107–123.

Berk, R.A. (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* New York: Springer.

Berk, R.A, (2013) "Algorithmic Criminology." *Security Informatics*, 2(5).

Berk., R.A., Brown, L., and Zhao, L. (2010) "Statistical Inference After Model Selection." *Journal of Quantitative Criminology* 26(2): 217-236, 2010.

Berk, R.A., Sherman, L., Barnes, G., Kurtz, E., and Ahlman, L. (2009a) "Forecasting Murder within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning. *Journal of the Royal Statistics Society — Series A* 172 (part I): 191–211.

Berk, R.A., Barnes, G., Ahlman, L., and Kurtz, E. (2010) "When Second Best Is Good Enough: A Comparison Between A True Experiment and a Regression Discontinuity Quasi-Experiment." *Journal of Experimental Criminology* 6(2): 191–208.

Berk, R.A., and Bleich, J. (2013) "Forecasts of Violence to Inform Sentencing Decisions." *Journal of Quantitative Criminology*, forthcoming.

Berkson, J. (1951) "Why I Prefer Logits to Probits." *Biometrics* 7: 327-339

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Borden, H.G. (1928) "Factors Predicting Parole Success." *Journal of the American Institute of Criminal Law and Criminology* 19: 328–336.

Box, G.E.P., and Jenkins, G. (1970) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Breiman, L., (1996) "Bagging Predictors." *Machine Learning* 26: 123-140.

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

Breiman, L. (2001b) "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3): 199–231.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth Press.

Burgess, E. M. (1928) "Factors Determining Success or Failure on Parole." In A. A. Bruce, A. J. Harno, E. .W Burgess, & E. W., Landesco (eds.) *The Working of the Indeterminate Sentence Law and the Parole System in Illinois* (pp. 205–249). Springfield, Illinois, State Board of Parole.

Bushway, S. (2011) *Albany Law Review* 74 (3).

Cameron, A.C. and Trivedi, P.K. (2005) *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011) "Using Offender Risk and Needs Assessment Information at Sentencing: Guidance from a National Working Group." National Center for State Courts, www.ncsconline.org/.

Chipman, H,A., George, E.I., and McCulloch, R.E. (2010) "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–298.

Dumbill, E. (2013) "Making Sense of Big Data." *Big Data* 1(1): 1-2.

Farrington, D. P. & Tarling, R. (2003) *Prediction in Criminology.* Albany: SUNY Press.

Feeley, M., & Simon, J. (1994). "Actuarial Justice: The Emerging New Criminal Law." In D. Nelken (ed.), *The Futures of Criminology* (pp. 173201). London: Sage Publications.

Friedman, J.H. (2002) "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis* 38: 367–378.

Gottfredson, S. D., & Moriarty, L. J. (2006) "Statistical Risk Assessment: Old Problems and New Applications." *Crime & Delinquency*52(1): 178–200.

Harcourt, B.W. (2007) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* Chicago, University of Chicago Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition. New York: springer.

Hyatt, J.M., Chanenson, L. & Bergstrom, M.H. (2011) "Reform in Motion: The Promise and Profiles of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing." *Duquesne Law Review* 49(4): 707–749.

Liu, Y.Y., Yang, M., Ramsay, M., Li, X.S., and Cold, J.W. (2011) "A Comparison of Logistic Regression, Classification and Regression Trees, and Neutral Networks Model in Predicting Violent Re-Offending." *Journal of Quantitative Criminology* 27: 547–573.

Kleiman, M., Ostrom, B. J., & Cheeman, F. L. (2007) "Using Risk Assessment to Inform Sentencing Decisions for Nonviolent Offenders in Virginia." *Crime & Delinquency* 53(1): 1–27.

Freedman, D.A., (2005) *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press.

Freund, Y., and Schapire, R.E. (1997) "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and system Sciences* 55(1): 119-139.

Messinger, S.L., & Berk, R.A. (1987) "Dangerous People: A Review of the NAS Report on Career Criminals." *Criminology* 25(3): 767–781

National Research Council (2013) *Frontiers for Massive Data Analysis*. Washington, D.C.: National Academies Press.

Ohlin L.E., and Duncan, O.D. (1949) "The Efficiency of Prediction in Criminology." *American Journal of Sociology* 54: 441-452.

Ohlin, L.E., and Lawrence, R.A. (1952) "A Comparison of Alternative Methods of Parole Prediction." *American Sociological Review* 17: 268–274.

Oregon Youth Authority (2011) "OYA Recidivism Risk Assessment — Violent Crime (ORRA-V): Modeling Risk to Recidivate with a Violent Crime." Oregon Youth Authority, Salem, OR.

Pew Center of the States, Public Safety Performance Project (2011) "Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders." The Pew Center of the States. www.pewcenteronthestates.org/publicsafety.

Reiss, A.J. (1951) "The Accuracy, Efficiency, and Validity of a Prediction Instrument." *American Journal of Sociology* 56: 552–561. Ridgeway, G. (2013) "The Pitfalls of Prediction." *NIJ Journal* 271.

Silver, E., & Chow-Martin, L. (2002) "A Multiple Models Approach to Assessing Recidivism Risk: Implications for Judicial Decision Making." *Criminal Justice and Behavior* 29: 538–569.

Skeem, J. .L., & Monahan, J. (2011) "Current Directions in Violence Risk Assessment." *Current Directions in Psychological Science* 21(1): 38–42.

Sorensen, J. R., & Pilgrim, R. L. (2000) "An Actuarial Risk Assessment of Violence Posed by Capital Murder Defendants." *The Journal of Criminal Law and Criminology* 90: 1251–1270.

Tollenaar, N. and van der Heijden, P.G.M. (2013) "Which Method Predicts Recidivism Best?: A Comparison of Statistical, Machine Learning and Data Mining Predictive Methods." *Journal of the Royal Statistical Society*, Series A 176 (part 2): 565–584.

Turner, S., Hess, J., & Jannetta, J. (2009) *Development of the California Risk Assessment Instrument*. Center for Evidence Based Corrections, University of California, Irvine.

VanNostrand, M., and Rose, K.J. (2009) *Pretrial Risk Assessment in Virginia*. St. Petersburg, Florida: Luminosity Inc.

Vapnick, V. (1998) *Statistical Learning Theory*. New York; Wiley.

Yang, M., Liu, Y., and Coid, J. (2010) "Applying Neural Networks and Other Statistics Models to Classification of Serious Offenders and the Prediction of Recidivism." Vol 6/10. London: Ministry of Justice.