

Forecasting Murder Within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning *

Richard Berk^{a,b}
Lawrence Sherman^{b,c}
Geoffrey Barnes^b
Ellen Kurtz^d
Lindsay Ahlman^d

Department of Statistics, University of Pennsylvania^a
Department of Criminology, University of Pennsylvania^b
Institute of Criminology, Cambridge University^c
Philadelphia Department of Adult Probation and Parole^d

November 27, 2007

Abstract

In the United States, forecasts of future dangerousness are often used to inform the sentencing decisions of convicted offenders, and for individuals sentenced to probation or parole, the conditions under which they are to be supervised. The target for these forecasts is commonly almost any new offense, most of which are not considered

*Richard Berk's work on this paper was funded by seed money from the University of Pennsylvania, and a grant from the National Science Foundation: SES-0437169, "Ensemble methods for Data Analysis in the Behavioral, Social and Economic Sciences." Geoffrey Barnes' work and Lawrence Sherman's work were funded by the Jerry Lee Foundation. Data and helpful expertise were provided by the Philadelphia Department of Adult Probation and Parole. All of this support is gratefully acknowledged.

to be very serious. This can produce a one-size-fits-all decision that may not allocate scarce criminal justice resources effectively. In this paper, we focus on individuals sentenced to probation or parole. Using data on over 60,000 cases beginning supervision under the Philadelphia Department of Adult Probation and Parole, we forecast whether a homicide or attempted homicide will be committed. We use statistical learning procedures that take the relative costs of false negatives and false positives into account and evaluate our forecasting skill with a large test sample. Homicide and attempted homicide are relatively rare crimes, but are considered to be among the most serious. Insofar as prospective murderers can be usefully identified, there is the possibility of shifting supervisory and rehabilitation resources to a subset of offenders who may be in greatest need.

Key Words: Statistical Learning; Random Forests; Forecasting, Probation, Parole, Homicide.

1 Introduction

In the United States, forecasts can be an essential determinant of the manner in which convicted offenders are sanctioned. Following a conviction, a decision is often made about whether incarceration or probation, for all or part of a sentence, is an appropriate punishment. In what is called “probation” in the United States, the offender is placed under supervision in his or her community. For offenders who are incarcerated, there is often a subsequent decision about whether a parole release is warranted and if so, the parole conditions to be imposed. Just as with probation, a parole release entails local supervision. For both probation and parole decisions, there is often an explicit forecast made of the chances that there will be a new crime committed or a “technical” violation of the conditions under which the local supervision is implemented. Such forecasts are often used to inform the decision to impose the sentence of probation or to release on parole.

Most forecasting in this context uses repeat offending for almost any offense as the criterion variable (Gottfredson and Tonry, 1987). A substantial majority of such offenses are not especially serious (Rossi, et al, 1974). Yet, the political discourse about public safety typically is dominated by relatively rare, very serious crimes that attract widespread publicity. In particular, ever since the Willie Horton case became a national issue in the U.S. Presidential election of 1988, judges and parole boards have been extremely sensitive to

crimes committed while convicted felons under sentence are authorized to be in the community (Anderson, 1995). More recently, rising homicide rates in some American cities have drawn attention to the disproportionate contribution of offenders under current sentence compared to the total homicide count, both as offenders and victims. In 2006, for example, over 22% of the murder arrestees and 16% of the 406 murder victims in Philadelphia were among the 52,000 clients of the county probation and parole department. Forecasts of very serious offenses would seem to be a useful complement to current practice.

In this paper, we apply statistical learning procedures to construct forecasts of failures on probation or parole. We focus on the commission of homicides because homicides are the most serious form of failure and because individuals on probation or parole can account for substantial fraction of all homicides in large metropolitan areas in the United States. Our work seems to be the first of its kind.

If useful forecasts can be made, there is the prospect of preventing homicides that might otherwise occur. Unlike the general population of prospective murders, individuals on probation or parole are already under supervision and already subject to any number of supervisory conditions intended to reduce the risk of failure. Insofar as individuals more likely to commit a homicide can be identified, it is possible to intervene with greater resources to help insure their successful completion of probation or parole.

An important obstacle to forecasting which probationers or parolees will commit a homicide is the need to work with data routinely available to probation and parole case workers as initial supervisory assignments are made. Data unavailable at that time cannot be used to inform decisions at intake. Another important obstacle is that within the population of individuals on probation or parole, a relatively small fraction, perhaps 1 in 100, will commit a homicide or even an attempted homicide. Consequently, the events to be forecasted are very rare.

There is also the matter of forecasting errors and the need to take the costs of false positives and false negatives into account. Failing to identify a probationer or parolee who subsequently commits a murder is likely to be far more costly than incorrectly labeling a probationer or parolee as a prospective murderer. The convenient assumption of equal costs is likely to be unresponsive to the true preferences of stakeholders.

2 Some Background

Forecasts of behavior while under supervision in the community have a long history. Work in the 1920s (Borden, 1928; Burgess, 1928) was often clinical in style. Factors thought to affect the chances of success or failure were used to make judgments about how parolees would fare. Sometimes points were awarded for each risk factor, and the sum was used as a summary measure of risk.

Over the next several decades, decisions about whether to release felons on probation or parole, and the conditions imposed on such releases, were often informed by relatively simple data analyses. Data containing information on behavior while under supervision and on various background factors such as age, gender, and past crimes, were examined for associations that might predict future behavior. The work was often substantively and conceptually sophisticated, even by today's standards (Ohlin and Duncan, 1949; Reiss, 1951; Ohlin and Lawrence, 1952; Goodman, 1952; 1953a; 1953b). More recent work has by and large substituted regression analysis of various kinds for the earlier use of cross-tabulations and bivariate measures of association (Farrington and Tarling, 1985; Maltz, 1984; Schmidt and Witte, 1988; Gottfredson and Tonry, 1987).

In retrospect, it is difficult to know how accurate the forecasts really have been (Farrington, 1987). Perhaps the most favorable assessment is that the level of forecasting skill has been “modest” (Gottfredson, 1987). Among the many obstacles for arriving at some overall evaluation is a general failure to quantify forecasting skill with data from a test sample (Berk, 2008).

The published studies evaluating performance on probation or parole have used a wide range of behavior to define success or failure. For example, a parole failure might be any conviction for a new crime. To the best of our knowledge, homicide has never been used as the sole criterion for success or failure. Yet, that is essentially what we do here.

3 The Setting

For each of the past several years, the Philadelphia Department of Adult Probation and Parole has had approximately 50,000 individuals under supervision. At intake, administrative personnel assign each individual to a “unit” within the Department. For “garden variety” cases, the unit is ge-

ographical; individuals under supervision are generally assigned to the geographical unit in which they live. There are also a large number of different “special” units for individuals such as sex offenders, individuals with alcohol or drug dependencies, and individuals with mental health problems. The special assignments are usually determined by information routinely available at intake, although a judge may also directly assign a convicted offender to one of the special units.

In addition, a probation/parole officer (P/PO) is assigned to each case. For garden variety cases, P/POs typically supervise 150 to 200 individuals, far too large a number to deliver intensive, targeted services. In effect, most probationers and parolees receive perfunctory supervision of, for example, a requirement to report to their P/PO once a month.

There is substantial heterogeneity even within the pool of garden variety cases. Most are unlikely to commit serious offenses while under supervision, but an important minority can be a genuine threat to public safety. Among the individuals assigned to special units, there is also considerable variability in the likelihood that a serious crime will be committed. One possible response is to forecast which individuals are likely to commit a serious crime while under supervision and provide those individuals with more intensive oversight and specialized services. In effect, a new kind of “special” unit would be defined.

A decision was made to focus on homicides. From the etiological perspective of a P/PO, however, a homicide is little different from an attempted homicide. Most homicides in the United States are committed with handguns. Whether a shooting results in a death depends on many factors such as whether a bullet happens to hit a vital organ, the speed with which paramedics arrive at the scene, the length of time it takes to get to a hospital, the quality of the medical care in that hospital’s emergency room, and the presence of a trauma care unit. None of these have much to do with the intent of the perpetrator. Consequently, either a homicide or an attempted homicide was selected as the outcome to be forecast. From a behavioral point of view, the two crimes are effectively the same.

There are other definitions that one might use for very serious crime. Homicide and attempted homicide were chosen because of their practical implications for the Philadelphia Department of Adult Probation and Parole and the local politics of crime. Homicides and attempted homicides dominate the public discourse and as such, are key determinants of how the Department is evaluated, how the media characterizes crime, and how the public

feels about its safety. In the recent mayoral election in Philadelphia, for instance, promises to do something about crime meant doing something about homicide.

One of the convenient consequence of using homicide and attempted homicide as an outcome is that compared to most other crimes, these are reported with greater accuracy. They are taken very seriously by the public and by law enforcement, and when there is a body or gun shots, there is physical evidence of a crime that is difficult to ignore. At the same time, insofar as there are homicides or attempted homicides that go unreported, it matters little for how the Department of Adult Probation and Parole does its job. It is reported crime by which the Department and city officials keep score.

Over the past several years, a little over 1% of the individuals in Philadelphia on probation or parole were charged with a homicide or an attempted homicide while under supervision and within 24 months of intake. By selecting as the key outcome a charge of a homicide or an attempted homicide, a very challenging forecasting task was being proposed. On the one hand, the forecasting exercise was trivial. If for each case, a forecast of neither a homicide nor an attempted homicide were made, that forecast would be correct about 99% of the time. On the other hand, the concern about forecasting such events implied that the costs of failing to usefully identify dangerous individuals were much higher than the costs of falsely labeling offenders as such.

Several conversations with officials within the Department of Adult Probation and Parole were undertaken in an effort to pin down the likely costs of forecasting errors. False negatives — failing to identify individuals likely to commit a homicide or attempted homicide — were seen as very costly. In addition to the loss of life and costs of arresting, trying and punishing the perpetrator, there were concerns about criticisms that would be made of the Department's operating procedures. In contrast, false positives — falsely identifying individuals as prospective murderers — were not seen as especially troubling. The primary loss would be the costs of delivering more intensive and specialized services, which if resources permitted, might be a good idea regardless. Individuals identified as likely murderers are also likely to commit other serious crimes. When pushed to quantify such judgments, there seemed to be general agreement that the costs of false negatives were approximately 10 times greater than the costs of false positives.

4 The Data

Data from the Philadelphia Department of Adult Probation and Parole were obtained on all cases with intake in between January 1, 2002 and June 30, 2004. A two-year follow-up period was used to help insure all cases could be followed for the same amount of time. A longer follow-up period would have led to some cases being right censored. A shorter follow-up period would have made the rare event of a homicide or attempted homicide even more rare. A difficult forecasting task would have been made more difficult still. Fortunately, officials at the Department of Adult Probation and Parole felt that the use of a two-year follow-up would meet their needs. The data set consisted of over 66,000 cases.

A failure was defined as a charge with homicide or attempted homicide within 2 years after the beginning of supervision. Charge can be an imperfect reflection of the crime actually committed. For example, in Philadelphia less than half of all homicides are cleared by an arrest. In addition, prosecutors sometimes charge with a more serious crime than they think can be proved in order to facilitate a plea bargain. However, the alternative of a conviction offense is also a flawed measure of the crime committed and would have required an impractically long follow-up period. Perhaps most important, the key stakeholders were comfortable with using a charge of homicide or attempted homicide as a measure of failure on probation or parole. Once a homicide or attempted homicide has been charged, the parolee or probationer is almost always taken off the streets and placed in jail awaiting trial. It is very rare for them to get bail. For all practical purposes, there is no longer any need for supervision in the community.

Predictors included all of the variables available to administrative staff at intake that had been identified by previous research as potentially useful: age, gender, race, prior record, the nature of the conviction offenses, features of the individual's neighborhood, the age at which there was an initial contact with the adult court system, and many others.

30,000 observations were selected at random to serve as the training sample, and the rest were to be used as the test sample. Having so large a test sample means that one does not have to resort to second-best approximations to evaluate how well forecasting procedures perform. The cross-validation statistic (Hastie et al., 2001: 214-217), for example, is an estimate of the generalization error that one would obtain had a large test sample been available. But we have such a sample. Within the training sample, there

were 322 true positives or about 1.1% of the total. This was fully consistent with expectations.

5 Data Analysis

The goal of the analysis was to produce usefully accurate forecasts from the information that a P/PO would have readily and routinely available when a case first arrived from the courts. Our mandate was to help improve the initial decision about the supervision and services to be provided to each new parolee or probationer. Information that might be available later was formally irrelevant for our forecasts.

Although the forecasting results would be more easily accepted if sensible predictors were found, the data analysis was not motivated by the need to better understand the causes of serious crime, let alone to construct a credible causal model. For example, had female parolees been forecasted to be at higher risk than male parolees, questions quite properly would have been raised about the forecasts. But the role of gender in the genesis of homicide was not a primary concern.

5.1 Random Forests

Random forests (Breiman, 2001; Lin and Jeon; 2006) was chosen as the forecasting method. Random forests is a statistical learning procedure that arrives at forecasts by aggregating the results from many hundreds of classification or regression trees. We used the implementation based on Fortran code written by Leo Breiman and Adele Cutler, ported to R (www.r-project.org/) by Andy Liaw and Matthew Weiner. (See the Appendix A1 for a summary of the algorithm.) We had had some success with this implementation of random forests in previous work (Berk et al., 2005; 2006; Lennert-Cody and Berk, 2007).¹

There are no classifiers to date that will consistently classify and forecast more accurately than random forests (Breiman, 2001; Berk 2006). Most will

¹Stochastic gradient boosting (Friedman, 2002), would probably do about as well finding instructive nonlinear functional forms, but at least within its current implementations in R (in the procedure *gbm* written by Greg Ridgeway), it does not provide a way for the a priori relative costs of false positives and false negatives to affect the fitted values. This is an especially important issue insofar as output such as partial dependence plots are to be examined.

not do as well, especially when highly nonlinear and noisy relationships are salient. Random forests uses an ensemble of classification or regression trees, each of which can inductively capture substantial nonlinearities (including interaction effects) when the trees are large. In addition, because random forests samples predictors when the trees are built, highly specialized predictors, that would not ordinarily be selected, are given a chance to contribute to the forecasts. Thus, random forests can inductively arrive at forecasts that are low in bias.

However, it is well known that classification and regression trees can be very unstable. They tend to overfit, which can inflate forecasting error. In response, random forests averages over trees to help stabilize the forecast. Stability is further enhanced when predictors are sampled as trees are grown. The combination of sampling predictors and averaging over trees serves the same function as more conventional shrinkage estimators such as the Lasso (Tibshirani, 2006). In addition, the forecasts are constructed from data *not* used to when a tree is grown. That is, the data used to grow each tree are not used when forecasts are made. For all of these reasons, random forests does not overfit. As the number of trees increases without limit, the estimate of population generalization error is consistent. (Breiman, 2001).

Finally, random forests allows the costs of forecasted false negatives and false positives to be built directly into the algorithm. As a result, all of the output responds to costs of forecasting errors that decision makers will need to consider (See Appendix A2 and A3). For applications such as ours, this is vital.

5.2 Forecasting Skill

A variety of parametric regression procedures have been used to analyze re-offending in various populations. Survival analysis, discriminant function analysis, probit regression and logistic regression are common examples when the key outcome is categorical. To help provide a benchmark for the performance of random forests, Table 1 shows the classification skill that results when logistic regression is applied to the training data. More detail will be provided shortly about the predictors and how to read the table, but it is obvious that logistic regression does only a little bit better than ignoring the predictors altogether. Using the training sample of 30,000 cases, two cases are classified as committing a homicide or attempted homicide. Of those two, one is a false positive. There are in fact 322 individuals in the training

data who committed a homicide or attempted homicide. The other standard procedures performed as poorly. It is very difficult for these analysis methods to overcome the low base rate of a little over 1%. Insofar as the predictors are related to the response in a highly nonlinear manner, there is reason to anticipate that random forests can do substantially better.

	Classified No Homicide	Classified Homicide	Model Error
No Homicide	29677	1	0.00001
Homicide	321	1	0.99709
Use Error	0.01	0.50	Overall Error = 0.01

Table 1: Logistic Regression Classification Table for Forecasts of Homicide or Attempted Homicide Using the Training Sample

Table 2 shows a conventional confusion table from a random forest analysis. The observed outcome is cross-tabulated against the forecasted outcome. Unlike for Table 1, this is a true forecasting exercise because the table uses the class forecasted when an observation is not being used to grow a given tree. The table is constructed from the out-of-bag observations as defined in Appendix A1. Other things equal, therefore, random forests should perform worse. It is usually much easier to classify accurately than to forecast accurately. For ease of exposition, the term “homicide” is used for both homicide and attempted homicide.

In Table 2, the row summaries on the far right hand side of the table are the proportions of cases incorrectly forecasted by the model, conditional on the truth. The column summaries at the bottom of the table are the proportions of cases incorrectly forecasted, conditional on the forecast. The single cell at the lower right hand side of the table contains the overall proportion of cases incorrectly forecasted.

From the off-diagonal cells in Table 2, one can see that there are 185 false negatives and 1764 false positives. The ratio of the latter to the former is about 9.5, very close to the target cost ratio 10 to 1. The cost ratio was introduced into the analysis by the stratified bootstrap sampling option available in the R implementation of random forests. The strata sizes chosen for the two response variable categories determined approximately the balance of false positives to false negatives. Because there is no exact correspondence,

	Forecast No Homicide	Forecast Homicide	Model Error
No Homicide	27914	1764	0.06
Homicide	185	137	0.57
Use Error	0.007	0.93	Overall Error = 0.07

Table 2: Confusion Table for Forecasts of Homicide or Attempted Homicide Using the Training Sample

some trial and error is necessary. In this instance, cost ratios ranging between 7 to 1 and 12 to 1 did not alter the results sufficiently for stakeholders to express any misgivings. Note that if there is about 1 false negative for every 9 false positives, false *negatives* are about 9 times more costly.

Overall, the forecasting skill looks promising. Only about 7% of the cases are forecasted incorrectly. However, this is not surprising given the highly unbalanced distribution of the response variable. Also, overall forecasting error is somewhat misleading as an summary measure of performance. The overall proportion of cases forecasted incorrectly does not by itself take the relative costs of false negatives to false positives into account.

About 43% of the probationers and parolees (i.e., 100% – 57%) are correctly forecasted as being charged with a homicide or attempted homicide, given the cost ratio used. Perhaps more important for practice, there are a little less than 13 false positives for every true positive (1764/137). It follows that whereas about 1 in 100 of the overall population of probationers and parolees will be charged with a homicide or attempted homicide within two years while under supervision, a little less than 8 in 100 within the identified subgroup are charged with such crimes. The various stakeholders have to date found this 8-fold improvement to be very promising.

As noted earlier, random forests does not overfit as the number of trees increases. But random forests was applied several times to the data with different cost ratios and different values for some of the tuning parameters. Overall, the results were much the same, but one might wonder if there was overfitting because of the number of forests grown. Earlier forests were used to help determined the tuning parameter values for later forests.²

²Random forest results are generally known to be insensitive to a range of reasonable values for the tuning parameters (Hastie et al., forthcoming: Chapter 13). For example, values for the number of predictors randomly selected at each CART split that are in the

	Forecast No Homicide	Forecast Homicide	Model Error
No Homicide	30652	2193	0.07
Homicide	198	147	0.58
Use Error	0.007	0.93	Overall Error = 0.07

Table 3: Confusion Table for Forecasts of Homicide or Attempted Homicide Using the Test Sample

Table 3 shows a confusion table constructed from the test data using the random forest fitting function responsible for Table 2. Except for the two different sample sizes, the tables are effectively identical. There is no evidence of overfitting.

5.3 Predictor Importance

The main goal of the analysis was to develop a useful forecasting device. But the degree to which that device was to be accepted by the Department of Adult Probation and Parole depended in part on whether the results broadly made sense. It was important, therefore, to examine which predictors weighed in most heavily as forecasts were constructed.

Using the algorithm summarized in the Appendix A2, Figure 1 shows the contribution of each predictor to forecasting skill. Concerns have been raised that forecasting skill computed in this manner can lead to biased estimates (Strobl et al., 2007). Nevertheless, their descriptive meaning is clear, and one would probably not be seriously misled using them solely to characterize the data on hand.

For Figure 1, a charge of homicide or attempted homicide is being forecasted. The age of the individual on probation or parole is the most important variable. If age is not allowed to contribute to forecasting skill, forecasting error increases about 12% (from 57% incorrectly forecasted to 69% incorrectly forecasted).

The age at which the individual has his or her first contact with the adult court system is the second most important predictor. Its contribution is about 8%. The number of prior convictions involving a firearm follows with a contribution of about 6%. Of somewhat less importance are in order:

neighborhood of the default work well and in about the same way.

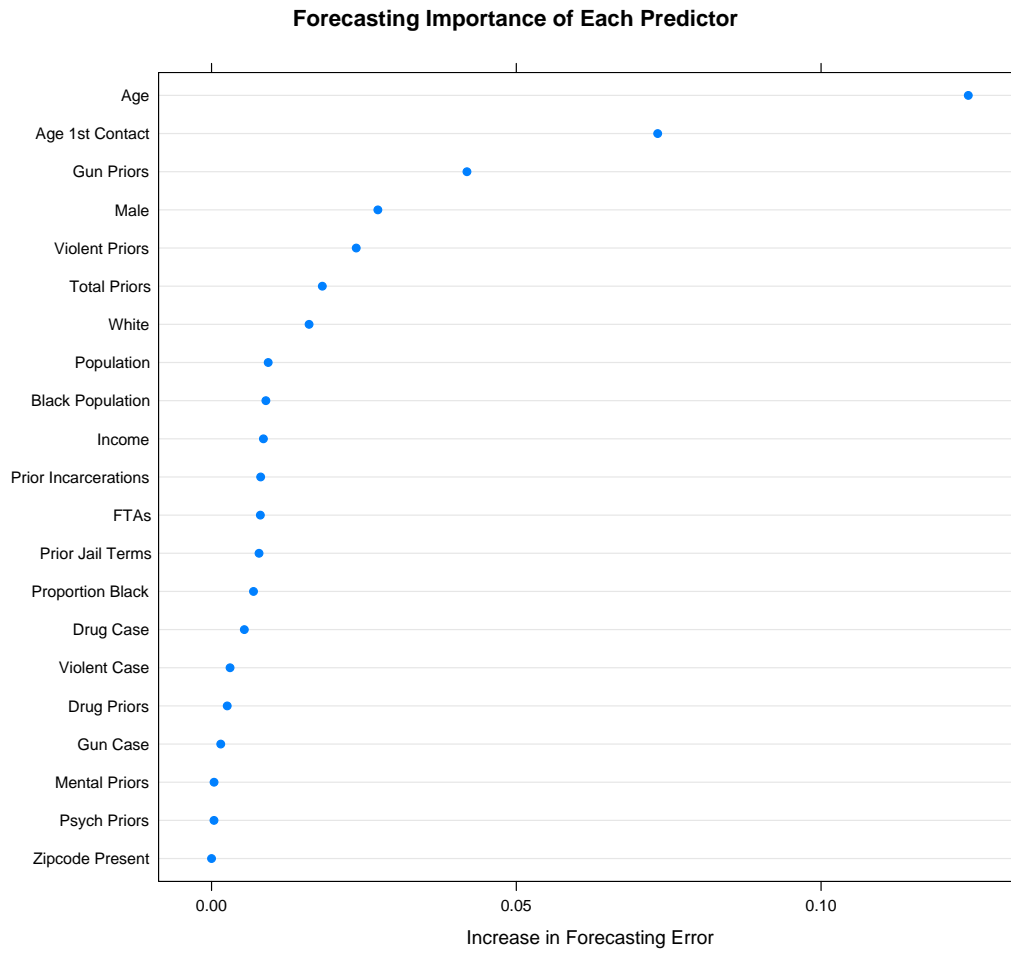


Figure 1: Predictor Importance for Forecasting Skill

gender, the number of prior convictions for violence offenses, the total number of all prior convictions, and race.³

5.4 Partial Response Functions

In addition to learning something of each predictor’s contribution to forecasting skill, it can be useful to see how each predictor is related to the response variable, other predictors held constant. In this instance, the sign and form of the response functions would need to make sense to stakeholders. Partial dependence plots can provide this information (Hastie et al., 2001: 333-334). (See the Appendix A3 for a summary of the partial dependence plot algorithm.)

Figure 2 shows the partial dependence plot for age. The predicted values are shown with small circles, and a smoother is overlaid. The vertical axis is in logit units defined as

$$f_k(X) = \log[p_k(X)] - \frac{1}{K} \sum_{k=1}^K \log[p_k(X)]. \quad (1)$$

Thus, the vertical axis is the disparity between the proportion for category k and the average of the proportions for all K categories in log units. In this case, K is 2, and $p_k(X)$ is the proportion of individuals forecasted to be charged with a homicide or an attempted homicide as a function of the set of available predictors.

³Among the other predictors, “population” is the number of people living in the offender’s zipcode area; “black population” is the number of African Americans living in the offender’s zipcode area; “income” is the median household income in the offender’s zipcode area; “prior incarcerations” is the number of prior incarcerations in state prisons; “FTAs” is the number of times the offender failed to show for a court appearance; “prior jail terms” is the number of prior jail terms; “proportion black” is the proportion of the population that is African American in the offender’s zipcode area; “drug case” is whether the instant offense was drug related; “violent case” is the instant case was violent; “drug priors” is the number of prior convictions for drug offenses; “gun case” is whether the instant offense involved the use of a firearm gun; mental priors is the number of prior mental health probation or parole cases implying assignment to the special mental health unit; “psych priors” is the number of prior probation or parole cases in which a psychiatric condition was imposed; and “zipcode present” is whether there was a reported home address for an offender from which the zipcode could be determined. Some of these predictors are highly correlated, but because the algorithm samples predictors, there are usually not serious problems.

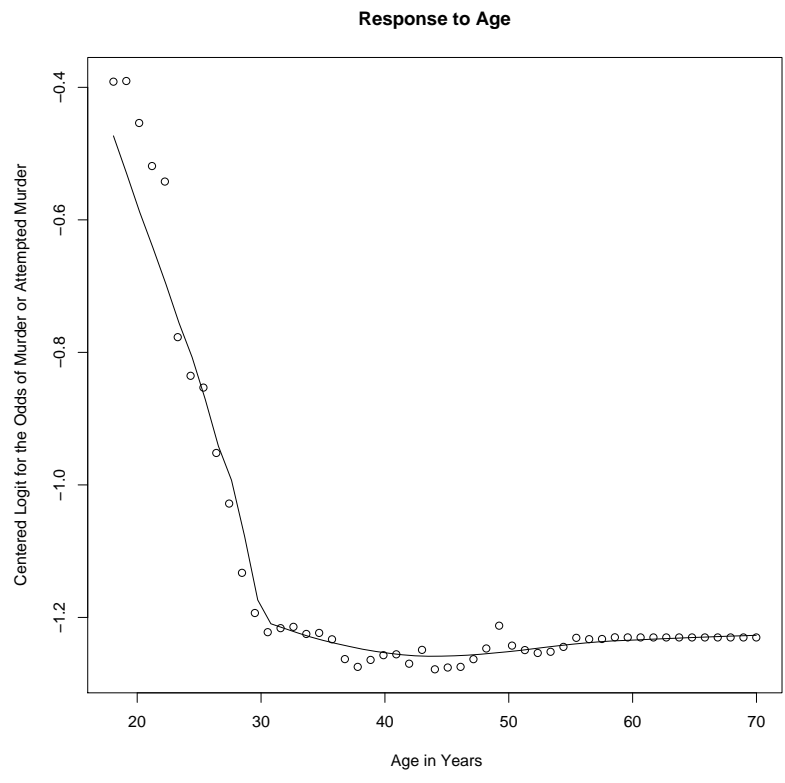


Figure 2: Partial Dependence Plot for Age

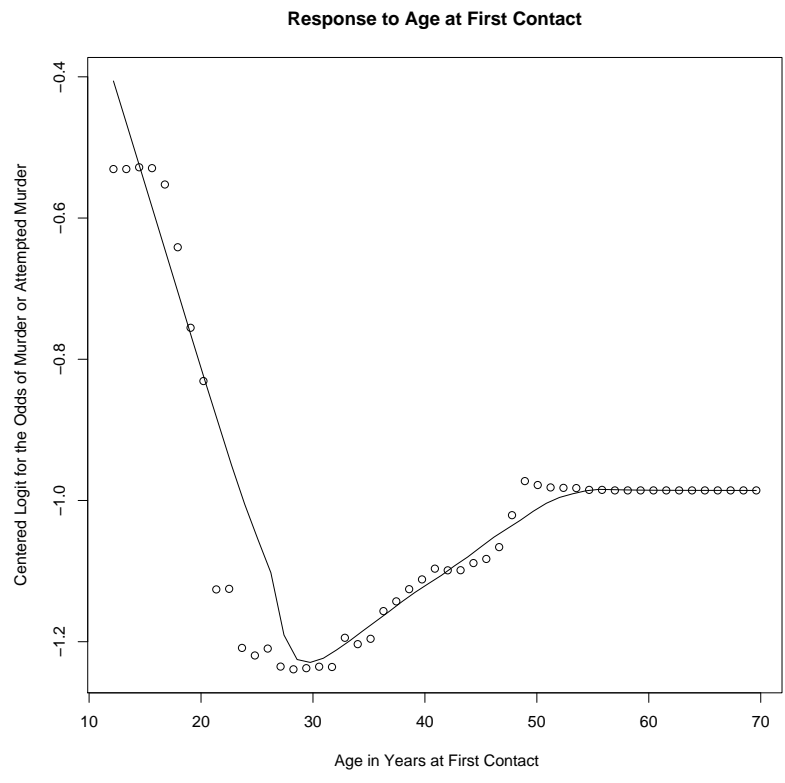


Figure 3: Partial Dependence Plot for Age at First Adult Court Contact

From Figure 2, it is apparent that the relationship between age and the log odds of being charged with a homicide or attempted homicide is highly nonlinear. The log odds decline very rapidly from about age 18 to about age 30. After age 30, the the relationship is flat. A negative relationship was anticipated. A highly nonlinear function of age was not. The message is clear: offenders in their late teens and early 20's are far more likely to be charged with a homicide or attempted homicide than offenders older than 30. Working backwards from the logit units, the odds are about twice as large.

Figure 3 shows that the age at which first contact is made with the adult court system also has a highly nonlinear relationship with the response. There is sharp decline from about age 12 to about age 30 and then the partial dependence plot begins a more gradual increase until age 50. Offenders who begin their criminal activities earlier are especially dangerous: the odds that they will charged with a homicide or attempted homicide are larger by a factor of two. One implication is that an aggravated assault, for example, committed at age 13 is a strong predictor of later violence, while that same assault committed at age 30 is not.

The increase in risk after age 30 was unexpected. It may represent offenders engaged in domestic violence, which is almost by definition not possible until dating begins, can be exacerbated with cohabitation, and often emerges after the peak years for other kinds of crime have passed. More generally, there may be a special etiology for violence committed by people who begin their criminal activities relatively late in life.

Figure 4 shows the partial dependence plot for the number of violent prior convictions. Again, the relationship is nonlinear. The risks increase very dramatically from 0 to 2 violent priors and increase strongly up to about 50 priors.⁴ Once again, the relationship with the response is strong. The difference between an offender with no violent priors and an offender with 10 violent priors is to approximately double the odds that a homicide or attempted homicide will be charged.

Figure 5 for the number of gun-related prior convictions has much the same pattern. The strength of the relationship is also about the same. The similarities may result in part from the fact that a large fraction of violent crimes involve the use of a handgun. In any case, nonlinear relationships

⁴There can be a prior conviction for each charge associated with a given crime incident. For example, if there is an armed robbery of a convenience store, there could well be a dozen charges. Over several crime incidents, these charges can accumulate rapidly. Offenders with more than 50 priors for violent crimes are relatively rare, but are real.

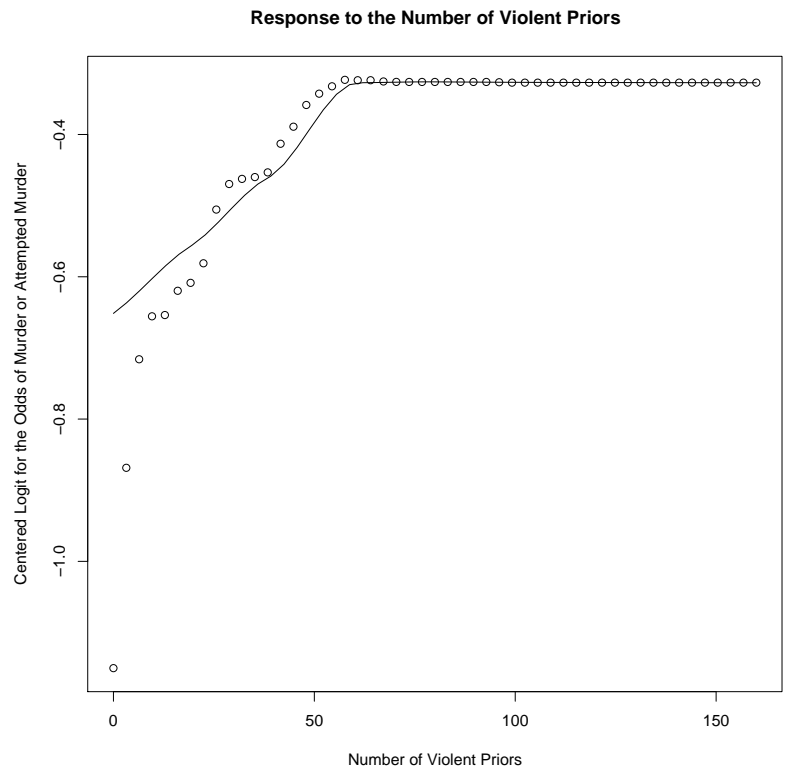


Figure 4: Partial Dependence Plot for the Number of Violent Prior Convictions

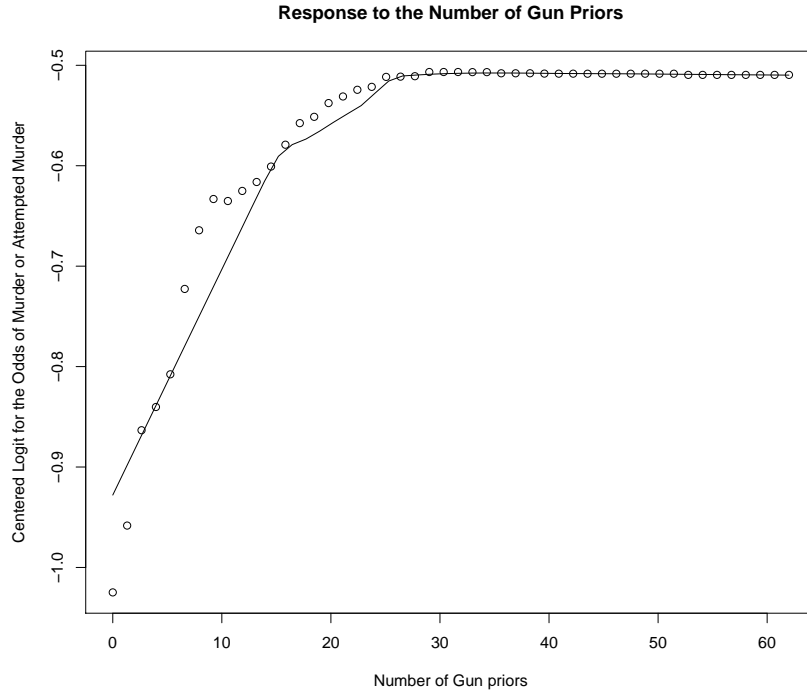


Figure 5: Partial Dependence Plot for the Number of Prior Gun Convictions

with the response are clearly very common.

Figure 6 shows the partial dependence plot for the size of the African American population in an offender’s zipcode area at time of intake. The response function is roughly S-shaped. Up to about 20,000 African American residents, there is no relationship. Between about 20,000 and 50,000 African American residents, the relationship is strongly positive. Above 50,000 African American residents, there is again no relationship. The impact on the odds of a homicide or attempted homicide charge is modest. The difference between 20,000 and 50,000 residents multiplies the odds by a factor of about 1.25.

It is not clear how the relationship should be interpreted. At face value, one might think the size of African American population is a surrogate for the number of potentially violent offenders and the number of potential victims. But that would not explain the leveling off of the relationship when the size of the African American population is very large.

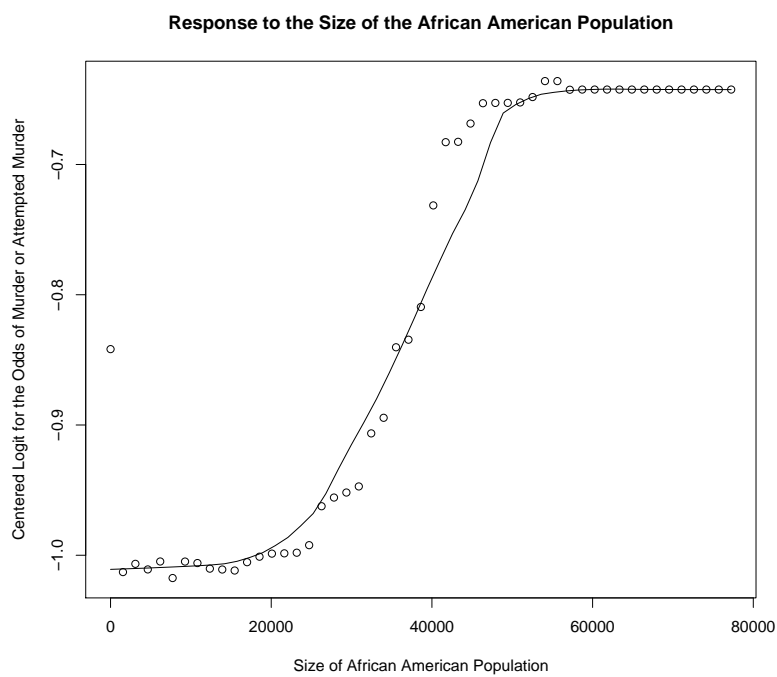


Figure 6: Partial Dependence Plot for the Size of the African American Population

Another interpretation of Figure 6 is that the proportion of a zipcode area that is African American is the real predictor. However, among the variables held constant is the overall population of the zipcode area. As a result, the role of the number of African Americans plays out with the overall population size fixed. But that too is unsatisfying because it is well known that in the United States, people tend to kill others like themselves; cross-race homicides are relatively rare. So again, there is no apparent explanation for the leveling off on the right side of the plot.

Moreover, the proportion African Americans was included directly as a predictor. If there were to be neighborhood effects as a function of race, this was the predictor we thought would surface. It did not, and various combinations of neighborhood variables all led back to the number of African Americans.

Finally, median household income was one of the predictors included in the analysis. So, typical household income is apparently not an explanation either. Nevertheless, we suspect that the number of African Americans is a surrogate for one or more other neighborhood variables that can affect crime. How this plays out will need to be explored in future research.

Partial dependence plots for the other quantitative predictors were not especially interesting. They summarized relationships that were not strong and approximately linear. Partial dependence plots for categorical predictors are generally of little help except to establish the direction of the association. Here, African American and male offenders were more likely commit homicides or attempted homicides.

In summary, one must be clear that the partial dependence plots are meant to be solely descriptive. There is no statistical model, causal or otherwise, in random forests. However, the signs of the relationships are broadly consistent with much past research and with social science theory that would give the associations causal interpretations. A major surprise is the number of highly nonlinear associations. To the best of our knowledge, these nonlinearities are new findings that no extant theory anticipates with any real precision. The nonlinear relationships also help explain why random forests forecasts does so much better than logistic regression and other parametric regression procedures. In this instance, the parametric regression models got the functional forms very wrong because the actual functional forms were unknown when the data analysis began.

6 Discussion

In the overall pool of cases on probation or parole, the chances are about 1 in 100 that an individual will be charged with a homicide or an attempted homicide within 2 years after intake. For the subset of cases identified by random forests, the chances are a bit less than 8 in 100. An important consequence is that for every true positive identified, there will be about 12 false positives. Were the marginal distribution alone used, for every true positive there would be nearly 100 false positives. Random forests leads to an 8-fold improvement by this yardstick.

By and large, sensible predictors were influential in the construction of forecasts, and the estimated response functions seemed plausible. For these reasons, stakeholders found the forecasting enterprise credible as well as useful. It would have been unsettling, for example, had older offenders been found to be more dangerous than younger offenders.

There were also no concerns expressed about the need for an analysis of subgroups (e.g., men versus women). Separate analyses for subgroups are just interaction effects, which is precisely what classification and regression trees are designed to detect. A random forest is an ensemble of such trees. Whatever information there might be in subgroup analyses is already built into the forecasts.

With a training sample of 30,000 randomly selected cases and the nearly exact reproduction of the results with a test sample of over 30,000 randomly selected cases, there is virtually no uncertainty due to sampling error in our forecasts. How well our forecasting results would work with *new populations* is less clear. We have data on Philadelphia probationers and parolees going back eight years. There does not seem to be any dramatic changes in the mix of offenders over time. Therefore, it is reasonable to assume that our results will usefully apply at least several years into the future. At that point, a new forecasting algorithm might well need to be constructed. More problematic is how well our forecasting algorithm would work in other jurisdictions. To apply the algorithm, however, one would need local data to “drop-down” the model. And if such data were available, it would just make good sense to rebuild the model itself. Then, external validity concerns formally disappear.

There is now a special unit within the Philadelphia Department of Adult Probation and Parole to oversee the individuals our procedures identify. At intake, the background variables of new cases are “dropped down” the random forest model to generate forecasts. If these individuals fall in the high

risk group, they are eligible for the new special unit.

There is a set of special services for the eligible individuals. Case loads are under 20 so that the officer in charge can be frequently in face-to-face contact with the parolees or probationers. Among the additional services are cognitive behavioral therapy for those diagnosed to need it and access to programs that can improve health, literacy, and job skills.

As soon as the content of these services is clearly defined, and it is determined that they can be delivered with high integrity, a randomized clinical trial will be mounted. Among the high risk offenders, some will be assigned at random to the usual forms of supervision. The others will be assigned at random to various combinations of special services. The question to be answered is whether there are cost-effective interventions for offenders at high risk to commit a homicide.

In a sense, the forecasts have so far been too good. Some identified offenders have failed so quickly that there was not enough time to deliver any meaningful services. The reasons for these failures included not just homicides or attempted homicides, but a range of serious crimes. Also, two parolees were shot to death and one was shot several times at close range and lived. It will be an ongoing challenge to intervene with sufficient speed and intensity to keep these offenders out of trouble and healthy.

There is work underway to obtain better data with which to construct forecasts. Among the data sets being sought are juvenile records. Because the most dangerous felons often get into trouble at a very early age, juvenile records may contain information useful for forecasting. There are also initial steps being taken to develop similar forecasting tools for neighboring jurisdictions. The forecasting algorithms will need to be hand-tailored for each locale, and discussions have begun to assemble the data required. Finally, under serious consideration is porting the statistical tools and data analysis approach to other settings such as decisions to place children on foster homes, judges' sentencing practices, and the selection of prison inmates for early release. We have had some success with prison data in the past (Berk et al., 2006).

7 Conclusions

The Philadelphia Department of Adult Probation and Parole, like many such departments in large cities across the United States, has too few resources to

provide proper oversight and services for those who might most benefit. One policy response can be to determine which offenders are in greatest need and then devote a larger share of the resources to them. Offenders more likely to commit a homicide or attempted homicide might be one such group.

The analyses presented in this paper would seem to provide a useful way to better target scarce resources. But whether these resources can be translated into cost-effective interventions is at this point unknown. By the end of 2008, we hope that one or more clinical trials will help provide an answer.

Finally, there may be some general implications for the study of crime. Researchers have worked for decades to find the important predictors of criminal behavior. By and large, the same cluster of explanatory variables has been repeatedly rediscovered. The strong nonlinear relationships reported here suggest that progress might be made by focussing on predictors already known to be important and trying to better understand how those predictors are related to the response. The usual functional forms assumed in practice may be well off the mark.

Addendum

Below are broad summaries of three important algorithms used by the version of random forests we employed. These summaries are not substitutes for understanding the details of what the software actually does. For ease of exposition, some shortcuts have been taken.

A1: Random Forests Algorithm Summary

1. Take a random sample of size n with replacement from the training data. The selected observations will be used to grow a classification tree. The observations not selected are saved as the “out-of-bag” (OOB) data.
2. Take a random sample of predictors.
3. Partition the data using CART (as usual) into two subsets minimizing the Gini index.
4. Repeat steps 2-3 for all subsequent partitions.
5. Compute (as usual) the class to be assigned to each terminal node.
6. “Drop” the OOB data down the tree and assign the class associated with the terminal node in which an observation lands.
7. Repeat steps 1-6 a large number of times.
8. For each observation, classify by majority vote over all trees when that observation was OOB.

A2: Variable Importance Algorithm Summary

1. Drop OOB data down a tree.
2. Compute the proportion of cases misclassified for each response class.
3. Randomly permute a given predictor.
4. Drop the OOB with the permuted predictor data down the tree again.
5. Compute the proportion of cases misclassified for each response class.

6. Compute the increase in the proportion misclassified as a measure of that variable's importance for each response class.
7. Repeat from step 3 for each predictor.

A3: Partial Dependence Plot Algorithm Summary

1. For a given predictor with M values, construct M special data sets, setting the predictor values to m and fixing the rest at their existing values. (For example, if the predictor is years of age, M might be 20, and there would be 20 data sets. For each, age would be set to one of the 20 age values for all observations. The rest of the predictors would be fixed at their existing values.)
2. Using these data and the random forest output, predict the response in logit units.
3. Average over terminal nodes.
4. Repeat 2 and 3 for each of the M values.
5. Plot averaged responses against the M values of the predictor.
6. Repeat 1-5 for each predictor.

References

- Anderson, D. (1995) *Crime and the Politics of Hysteria: How the Willie Horton Story Changed American Justice*. New York: Crown Publishing.
- Berk, R.A. (2006) "An Introduction to Ensemble Methods for Data Analysis," *Sociological Methods and Research*, 34(3): 263-295, 2006.
- Berk, R.A. (2008) "Forecasting Methods in Crime and Justice." *Annual Review of Law and Social Science*, Palo Alto, CA: Annual Reviews Press, forthcoming
- Berk, R.A., He, Y., and S. Sorenson (2005) "Developing a Practical Forecasting Screener for Domestic Violence Incidents." *Evaluation Review* 29(4): 358-382.
- Berk, R.A., Kriegler B. and J. Baek (2006) "Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures," *Journal of Quantitative Criminology*, 22(2): 131-145.
- Borden, H.G. (1928) "Factors Predicting Parole Success." *Journal of the American Institute of Criminal Law and Criminology* 19: 328-336.
- Breiman, L. (2001) "Random Forests." *Machine Learning* 45: 5-32.
- Burgess, E.W. (1928) "Factors Determining Success or Failure on Parole," in A.A. Bruce, A.J. Harno, E.W. Burgess, and J. Landesco (eds.) *The Working of the Indeterminant Sentence Law and the Parole System in Illinois* Springfield, Illinois, State Board of Parole: 205-249.
- Dean, C.W. and T.J. Dugan (1968) "Problems in Parole Prediction: A Historical Analysis." *Social Problems* 15: 450-459.
- Farrington, D.P. and R. Tarling (1985) *Prediction in Criminology*. Albany: SUNY Press.
- Farrington, D.P. (1987) "Predicting Individual Crime Rates," in D. M. Gottfredson and M. Tonry (eds.), *Prediction and Classification*. Chicago: University of Chicago Press.

- Friedman, J.H. (2002) "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis* 38: 367-378.
- Goodman, L.A. (1952) "Generalizing the Problem of Prediction." *American Sociological Review* 17: 609-612.
- Goodman, L.A. (1953a) "The Use and Validity of a Prediction Instrument. I. A Reformulation of the Use of a Prediction Instrument." *American Journal of Sociology* 58: 503-510.
- Goodman, L.A. (1953b) "II. The Validation of Prediction." *American Journal of Sociology* 58: 510-512.
- Gottfredson, S.D. (1987) "Prediction: An Overview of Selected Methodological Issues," in D. M. Gottfredson and M. Tonry (eds.), *Prediction and Classification*. Chicago:University of Chicago Press.
- Gottfredson, D.M., and M. Tonry (eds.) (1987) *Prediction and Classification*. Chicago: University of Chicago Press.
- Hastie, T., Tibshirani, R. and J. Friedman (2001) *The Elements of Statistical Learning*. New York: Springer-Verlag.
- Hastie, T., Tibshirani, R. and J. Friedman (forthcoming) *The Elements of Statistical Learning*, Second Edition. New York: Springer-Verlag.
- Ohlin, L.E., and O.D. Duncan (1949) "The Efficiency of Prediction in Criminology." *American Journal of Sociology* 54: 441-452.
- Ohlin, L. E. and R.A. Lawrence (1952) "A Comparison of Alternative Methods of Parole Prediction." *American Sociological Review* 17: 268-274.
- Lennert-Cody, C. and R.A. Berk (2007) "Statistical Learning Procedures for Monitoring Regulatory Compliance: An Application to Fisheries Data," *JRSS, Series A*, 170, (3): 671-689, 2007.
- Lin, Y., and Y. Jeon (2006) "Random Forests and Adaptive Nearest Neighbors." *Journal of the American Statistical Association* 101: 578-590.
- Maltz, M.D. (1984) *Redidivism* New York: Academic Press.

- Reiss, A.J. (1951) "The Accuracy, Efficiency, and Validity of a Prediction Instrument." *American Journal of Sociology* 56: 552-561.
- Rossi, P., Waite, E., Bose, C.E., and R. Berk, "The Seriousness of Crimes: Normative Structure and Individual Differences," *American Sociological Review* 39(2): 224-237.
- Schmidt, P. and A.D. Witte (1988) *Predicting Recidivism Using Survival Models*. New York: Springer.
- Strobl, C., Bouestreix, A., Zeileis, A., and T. Hothorn (2007) "Bias in Random Forest Variable Importance Measures: Illustrations, Sources, and a Solution." *Bioinformatics* 8(25).
- Tibshirani, R.J. (1996) "Regression Shrinkage and Selection Via the LASSO," *Journal of the Royal Statistical Society, Series B*, 25: 267-288.