Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions

Article *in* Journal of Empirical Legal Studies · March 2016

DOI: 10.1111/jels.12098

CITATIONS
70

READS
2,057

3 authors:

Richard A. Berk
University of Pennsylvania
325 PUBLICATIONS **13,415** CITATIONS

SEE PROFILE

Susan Sorenson
University of Pennsylvania
134 PUBLICATIONS **7,000** CITATIONS

SEE PROFILE

Geoffrey C. Barnes
University of Cambridge
22 PUBLICATIONS **1,184** CITATIONS

SEE PROFILE

JOURNAL OF
EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 13, Issue 1, 94–115, March 2016

Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions

*Richard A. Berk, Susan B. Sorenson, and Geoffrey Barnes**

Arguably the most important decision at an arraignment is whether to release an offender until the date of his or her next scheduled court appearance. Under the Bail Reform Act of 1984, threats to public safety can be a key factor in that decision. Implicitly, a forecast of "future dangerousness" is required. In this article, we consider in particular whether usefully accurate forecasts of domestic violence can be obtained. We apply machine learning to data on over 28,000 arraignment cases from a major metropolitan area in which an offender faces domestic violence charges. One of three possible post-arraignment outcomes is forecasted within two years: (1) a domestic violence arrest associated with a physical injury, (2) a domestic violence arrest not associated with a physical injury, and (3) no arrests for domestic violence. We incorporate asymmetric costs for different kinds of forecasting errors so that very strong statistical evidence is required before an offender is forecasted to be a good risk. When an out-of-sample forecast of no post-arraignment domestic violence arrests within two years is made, it is correct about 90 percent of the time. Under current practice within the jurisdiction studied, approximately 20 percent of those released after an arraignment for domestic violence are arrested within two years for a new domestic violence offense. If magistrates used the methods we have developed and released only offenders forecasted not to be arrested for domestic violence within two years after an arraignment, as few as 10 percent might be arrested. The failure rate could be cut nearly in half. Over a typical 24-month period in the jurisdiction studied, well over 2,000 post-arraignment arrests for domestic violence perhaps could be averted.

I. Introduction

In this article, we address empirically the potential role of domestic violence forecasts when, at an arraignment, a judge, commissioner, or magistrate decides whether an

*Address correspondence to Richard A. Berk, Department of Criminology, University of Pennsylvania, 19119; email: berkr@sas.upenn.edu. Sorenson is Professor in the School of Social Policy and Practice at the University of Pennsylvania; Barnes is Research Assistant Professor of Criminology in the Department of Criminology at the University of Pennsylvania.

We appreciate the help of Michael Gallagher, a retired police officer, who currently works with a domestic violence agency in the jurisdiction studied. He provided important information on the meaning of some of our variables and on related law enforcement procedures. We also received important assistance on the prosecutorial context of domestic violence from Marian G. Braccia, Deputy District Attorney in the District Attorney's Office of the jurisdiction. Thanks also go to the anonymous reviewers of this article.

offender can be released awaiting a formal hearing on the charges. In the past, forecasts of "future dangerousness" have sometimes been used at arraignments (Goldkamp & Gottfredson 1985; VanNostran & Keebler 2009; Arnold Foundation 2013), but there have been to our knowledge no regularized applications of domestic violence risk assessment procedures in this setting. Our forecasting efforts are part of a larger pretrial reform initiative in a major metropolitan area.[1]

A combination of machine learning and routine electronic information normally available at arraignment might be able to provide timely and useful domestic violence forecasts of risk. There are examples of successful forecasting in other criminal justice settings and for other kinds of crimes (Berk 2012). Moreover, machine learning forecasts can be delivered within a real time of several seconds. However, a major question is whether the information routinely available electronically prior to an arraignment is sufficiently rich to produce usefully accurate forecasts. We address this question using data on over 28,000 domestic violence arraignments. The performance of the forecasts is the empirical focus of this article.

II. Background

Although the details vary across jurisdictions, shortly after an arrest or an apprehension through a summons, there is an arraignment at which the offender receives a written copy of the charges alleged by the police. These charges typically have been reviewed by a representative from the district attorney's office and revised as needed.[2] After the charges are read, a date for a subsequent hearing is specified during which a judge will determine whether there is sufficient evidence to require a trial. If the judge rules that there is sufficient evidence, the offender must enter a plea of guilty or not guilty.

At the arraignment, a court official, variously called a judge, commissioner, or magistrate, decides whether to detain the offender in jail until the hearing or to release the offender, sometimes on bond or subject to certain conditions, with the requirement that the offender return to court on the hearing date.[3] That decision is usually shaped by two factors specified in the Bail Reform Act of 1984 (Adair 2006): the risk of flight and the threat to public safety. By law and by court decisions from the Seventh and Ninth Circuits, both are effectively forecasts that are to be made after carefully considering a range of factors beyond the immediate charges (Federal Judicial Center 1993:8–10). In practice, the forecasts are commonly made on a magistrate's subjective

[1]It is not fully clear what to call an arrested individual at a preliminary arraignment. At some point, that individual can become a defendant, but in this jurisdiction, the individual does not have to answer to formal charges until a subsequent hearing. Therefore, we will use the term "offender" rather than "defendant." There should be no confusion in context.

[2]The arraignment is sometimes called a "preliminary arraignment," in part because prosecutors have the option to unilaterally revise the charges subsequently.

[3]In this article, we will use the title "magistrate" rather than judge or commissioner. Regardless of title, the tasks performed at a preliminary arraignment are essentially the same.

and nearly instantaneous judgments (Reitler et al. 2013). Forecasts of flight risks are being addressed in other work. Here, we consider forecasts of domestic violence (DV).

Domestic violence is among the more common charges heard at arraignments. Definitions vary, but in the jurisdiction for which our analysis is undertaken, "domestic" means an intimate relationship including dating, or a familial or blood relationship. Some equate domestic violence solely with intimate partner violence.[4] The jurisdiction's definition incorporates a much wider range of relationships. For example, an altercation between a man and his adult son is considered domestic violence. "Violence" means attempts to cause or actually causing bodily injury or serious bodily injury, and committing rape or attempted rape. Placing an individual in reasonable fear of those acts is included as well. But domestic violence also can include actions that some may not see as violent, such as burglary, destruction of property, threats of economic retribution, threats against pets, witness intimidation, and arguments over child custody, many of which can also be charged as other kinds of crimes. Consistent with the jurisdiction's practice and the state statutes under which magistrates operate, we will use the term "domestic violence" in this inclusive fashion. Later, we will make an important empirical distinction between domestic violence associated with a physical injury and other forms of domestic violence.[5]

Because domestic violence often is serial, concerns about victim safety can be very salient and forecasting accuracy really matters. An offender may be held unnecessarily or an offender may be released only to reoffend. Even a few days in jail when incapacitation is not needed can cost an offender his or her job and more generally disrupt a variety of important household functions (e.g., childcare). There are also concerns that a lengthy period of detention can jeopardize an offender's due process rights (Federal Judicial Center 1993:29–30; LaFrentz & Spohn 2006; Spohn 2009), and some reason to worry that it is at the pretrial stage where a substantial fraction of false convictions and coerced guilty pleas originate (Cohen 2012). Yet, when incapacitation is needed but not used, a new domestic violence offense can result, sometimes escalating to homicide.

There is a growing literature on ways to improve the processes and outcomes associated with arraignments (e.g., Bock & Frazier 1977; Frazier et al. 1980; Goldkamp & Gottfredson 1985; Bridges et al. 1987; Demuth 2003; Devers 2011), even some studies based on randomized experiments (Goldkamp & White 2006; McElroy 2011; Bornstein et al. 2013). Risk assessments have of late figured very significantly in these studies (VanNostrand & Keebler 2009; Arnold Foundation 2013), but the risks being forecasted are typically defined as an arrest for *any* crime.

It might seem that there is an easy fix: at the arraignment, use one of the better intimate partner risk assessment tools. However, even if one ignores troubling methodological

[4]The U.S. Department of Justice defines domestic violence as "as a pattern of abusive behavior in any relationship that is used by one partner to gain or maintain power and control over another intimate partner ..." (U.S. Department of Justice 2014).

[5]A list of the crimes included can be provided upon request. There are several hundred such crimes. For example, there are about 50 kinds of sexual crimes.

concerns (Farrington & Tarling 2003; Berk & Sorenson 2005; Gottfredson & Moriarty 2006; Berk & Bleich 2013; Ridgeway 2013), popular intimate partner risk assessment tools typically do not consider the wide range of crimes that are folded into statutory definitions of domestic violence and require data that currently cannot be routinely obtained within the very short interval between arrest and arraignment (Roehl et al. 2005). That interval is typically less than 48 hours.[6] In addition, most release decisions are made in a matter of a few minutes, almost regardless of charge. There is little time to collect and analyze information at the arraignment itself beyond what might be included in brief summaries of the charges.

Alternatively, one might consider whether machine learning procedures applied to routinely available criminal justice data lead to practical and sufficiently accurate forecasts. Data that typically are available include whatever is maintained on rap sheets, on records of past contacts with the courts, and some basic biographical information recorded when an arrest is made. These are no doubt slim pickings if the goal is to forecast domestic violence. For example, there often is no information on how the perpetrator and victim are related beyond what may be inferred from the charge of domestic violence, nothing about a household's economic circumstances, and no information on the offender's behavior and life circumstances more generally. Nevertheless, the inputs to most machine learning procedures are little more than raw material. Machine learning algorithms can transform, combine, and reconstruct a relatively small number of inputs into hundreds of predictors that may have little apparent relation to the inputs initially provided, but earn their keep by improving forecasting accuracy (Hastie et al. 2009; Berk 2012; Berk & Bleich 2013; Jordan & Mitchell 2015).

For example, an offender's age may be, in effect, reconfigured as a set of indicator variables so that an empirically determined, nonlinear relationship with domestic violence is built. For violent street crimes, the ages of highest risk are the late teens and early 20s, after which the risk of perpetration drops dramatically (Berk & Bleich 2013). For domestic violence, one might anticipate a high risk well into the 30s and 40s. In both cases, age serves as a proxy for physical aging, life course events, and evolving relationships with others. There are no direct measures of such processes but perhaps appropriate reconstructions of age can suffice. Moreover, because a very large number of predictors can be exploited, one is not limited to predictors that are strongly related to the outcome being forecasted. A large number of weak predictors, which would ordinarily be dismissed, can *in the aggregate* dramatically improve forecasting accuracy.

But all this comes at a price: machine learning procedures are "algorithmic" in nature (Breiman 2001b). There is no model in the usual statistical sense, and how the inputs are related to the outputs is not fully apparent. One is working with black-box procedures whose goal is accurate forecasts. Explanations of why the inputs are related to

---

[6]It might be possible to construct different forecasting tools for different kinds of domestic violence, but the limitations of data available at arraignment remain. For example, time spent in jail after arraignment is plausibly related to the chances of a new domestic violence arrest and could be especially important were the goal to understand *why* some offenders fail. However, post-arraignment time spent in jail is unknowable at arraignment.

outputs in any particular manner are a secondary concern and perhaps even unknowable (Berk & Bleich 2013).

Readers already familiar with machine learning applications in criminal justice might wonder what is novel about the analysis to follow. To the best of our knowledge, this is the first machine learning application to forecasts of domestic violence, defined broadly as in the governing statute, to inform release decisions at a preliminary arraignment. The forecasting exercise is part of a local pretrial reform initiative. Past machine learning forecasting applications in criminal justice have addressed intimate partner violence to support police decisions when they arrive at the scene (Berk et al. 2005), misconduct in prison to help prison administrators assign inmates to appropriate security levels (Berk et al. 2006), homicides committed by probationers or parolees to identify individuals who pose a very serious threat to public safety (Berk et al. 2009), arrests for three different categories of crime committed while on probation or parole to rationalize the intensity of supervision and the kinds of services offered (Berk et al. 2010), and future violence to help judges make post-conviction sentencing decisions (Berk & Bleich 2013, 2014; Berk & Hyatt 2015). The application reported in this article is unique and addresses offenses of great concern in many courtrooms across the country.

III. Data and Methods

The data used in this study were compiled as part of a larger effort to reform pretrial procedures and outcomes in a large metropolitan area. Here, we analyze 28,646 domestic violence arraignments leading to official charges and a release between January 1, 2007 and October 31, 2011.[7] This is the full population of such cases during the specified time interval. The data are restricted to released cases because it is only these cases that have the opportunity to reoffend.

The research design called for a two-year followup for each arraigned case through the end of October 2013. A two-year followup might seem to be a strange policy choice. One might assume that a magistrate's primary concern at a preliminary arraignment is what could happen between the time of a release and the time of the next court appearance. In most cases, that would be less than two years. However, local stakeholders made clear that a two-year followup is actually responsive to their decision-making needs because the two-year followup captures most of the cases they care about.

> It is possible for people to be detained while in pretrial posture after bail magistrates release them at arraignment. For example, if they fail to appear, or incur new charges or there is some issue, a judge can order detention at any phase of the trial. Traditionally, for measurement purposes we have looked at any pretrial misconduct from the point of release, through the adjudication of the case. Until the case is adjudicated, it is under the purview of Pretrial Services. The goal has always been to get offenders through trial successfully without any misconduct, not just

[7]These were not cases in which a release followed from all charges being dropped.

to get them to their first court appearance. It would be fair to say this sentiment is shared by Pretrial Services, the Judiciary, and all stakeholders." (personal communication)[8]

These priorities are fully consistent with 1984 Bail Reform Act, which allows for the revocations and modifications of release conditions for wide range of stated reasons (Federal Judicial Center 1993:5–6), even if the release conditions have not been violated.

Inputs used for forecasting were taken from electronic information available at arraignment. A list of inputs is provided in Table 1. All appear to comport well with the requirements of the 1984 Bail Reform Act (VanNostrand & Keebler 2009:8–9). "Charges" refers to criminal charges associated with an arrest. "Instant" refers to charges associated with the current arraignment.[9]

The observational units are cases, not individuals, because a release decision is made about each case as it is taken from the court docket. One consequence is that an individual can appear in our data more than once, but according to court officials, prosecutors, and defense attorneys intimately familiar with pretrial procedures in the study's jurisdiction, such offenders are a distinct minority. To be arraigned, an offender must be arrested, and to be arrested, the motivating domestic violence incident must be reported to the police. Given the study design, all this would have to happen before October 31, 2011. Finally, potential dependence resulting from some repeat individuals creates no special problems for the machine learning procedure we use (Hastie et al. 2009:Algorithm 15.1).

The post-arraignment outcomes to be forecasted were determined by stakeholders, which for this initiative included individuals on the oversight committee of the local Pre-Trial Reform Project. There were three domestic violence outcome classes defined to be consistent with our discussion above:

1. DV0—no arrest for a domestic violence offense;
2. DV1—a domestic violence arrest *not* involving physical injury, an attempt to cause physical injury, or the threat of physical injury; and
3. DV2—a domestic violence arrest involving physical injury (including rape), an attempt to cause physical injury, or the threat of physical injury.

The most pressing goal was to find a subset of offenders who could be released with no conditions and who were good bets not to be rearrested for domestic violence. Should a substantial number of such individuals be found, attention would then turn to the remaining offenders, who might be detained or released under a variety of imposed conditions, depending on the seriousness of domestic violence predicted. The mix of possible interventions is described briefly later.

---

[8]Email from members of the oversight committee of the local Pre-Trial Reform Project.

[9]Offenders who spent more than 18 months of the follow-up period in jail were excluded from the analysis because of a much reduced opportunity to reoffend. The available data were not detailed enough to allow for a more subtle approach to "time at risk." Moreover, had we used data management methods that court IT staff could not easily reproduce, our forecasting procedures could not be put into practice.

Table 1: Inputs for the Forecasting Exercise

Age
Gender
Living in a high crime zip code (an indicator variable for each)
Number of prior charges for murder
Number of prior charges for DUI
Number of prior charges for domestic violence
Number of prior charges overall
Number of prior charges for animal mistreatment
Number of prior charges for property crimes
Number of prior charges for serious crimes
Number of prior charges for violent crimes
Number of prior charges for sex crimes
Number of prior charges for firearm crimes
Number of prior charges for weapons offenses
Number of prior charges for drug offenses
Number of prior arrest warrants
Number of prior prison/jail sentences
Age at first adult charge
Number of prior failures to appear
Number of prior probation sentences
Number of prior abscondings
Number of prior probation violations
Number of prior days in jail
Number of prior jail terms
Currently on probation
Number of instant charges overall
Number of instant murder charges
Number of instant domestic violence charges
Number of instant weapons counts
Number of instant property counts
Number of instant drug distribution counts
Number of instant violent crime counts
Number of instant serious crime counts
Number of instant sex crime counts
Number of instant firearm crime counts
Number of instant drug crime counts

NOTE: These are the inputs to the machine learning algorithm that could be routinely downloaded from existing electronic administrative records.

Stakeholders readily understood that the absence of a new arrest for domestic violence did not mean that no such crimes had occurred. Such crimes are underreported and, even when reported, often do not lead to an arrest. It is common, for instance, for police to arrive at the scene after the perpetrator has fled. The difference between a domestic violence incident and a domestic violence arrest has important policy implications that we address below.

No effort was made to pare down the list of inputs, and there is no doubt substantial overlap in what is being measured. The machine learning procedure we favor has none of the problems conventional regression analysis would have with so many

correlated variables and can even work with more forecasting inputs than observations. Consistent with practice in machine learning, the primary goal is not to determine which inputs are important and which are not, but to use all the inputs as a group to arrive at accurate forecasts.

Our machine learning method of choice is random forests (Breiman 2001a), which is essentially a large ensemble of classification or regression trees. An outline of the algorithm is provided in the Appendix. There are several published examples of very successful criminal justice forecasting exercises using random forests, although some other machine learning procedures can forecast about as well (Berk 2012).[10] Random forests is preferred here because there are easy ways to weight forecasting errors by their relative costs and because it provides visualizations of forecasting performance that can be very instructive. Both assets are substantial and are discussed at some length elsewhere (Berk 2008:Ch. 5). Nevertheless, stakeholders understood that because of the limited range of inputs available, achieving good forecasting accuracy would be difficult, and the results might not be of much use.

IV. Results

The mix of offenders at arraignment can be very different from the mix of offenders at other criminal justice decision points. Offenders at arraignment have been arrested, but have yet to be officially charged. For many offenders, prosecutors will choose not to proceed. If a decision is made to prosecute, some will be found not guilty at trial, and a much larger number will plead guilty to a much less serious offense. In short, offenders at arraignment can look somewhat different from those who commit crimes but who are not arrested and from those convicted of a crime who are then sentenced.

At the same time, individuals at arraignment charged with domestic violence on the average look much like individuals at arraignment charged with other crimes who are released. We have information on all offenders arraigned, charged, and released between January 1, 2007 and October 31, 2011, not just those charged with domestic violence. Table 2 shows that with the exception of priors for domestic violence, those charged with domestic violence are less "hard core." However, the differences are very small and probably of no importance. In contrast, those charged with domestic violence are about three times more likely to have prior domestic violence arrests (15 percent compared to 5 percent).

The 15 percent figure for domestic violence priors might seem surprisingly low. However, as already noted, a domestic violence prior requires a domestic vio-

---

[10]All these methods will forecast as least as well as traditional regression approaches, and usually substantially better (Berk & Bleich 2013). Like all forecasting procedures, however, they assume that there is reasonable stability over the medium term in the processes that lead to domestic violence arrests. In this case, there were no major changes in statutes, administrative practices, or domestic violence interventions that would undermine the forecasting procedures.

Table 2: Domestic Violence Arraignment Cases Compared to All Other Arraignment Cases

Summary Statistic	*DV Cases* *N = 28,646*	*All Others* *N = 197,770*
Proportion male	0.81	0.83
Mean age	34.3	32.4
Mean age at first adult charge	25.5	24.1
Mean number of priors overall	22.3	23.0
Mean number of drug priors	2.8	4.0
Mean number of firearm priors	1.4	1.5
Mean number of times incarcerated	1.8	2.0
Proportion on probation	0.17	0.21
Proportion with DV priors	0.15	0.05

NOTE: The table compares several important summary statistics of domestic violence cases at arraignment to the same summary statistics for all other cases at arraignment. The statistics are very similar except for the proportion of cases with domestic violence priors. The domestic violence cases are more likely to have domestic violence priors, but are overall much like other arraignment cases.

lence arrest. Someone must report a crime consistent with the statutory requirements for domestic violence, the police must respond and find that a crime has been committed consistent with that definition, and then the police must make an arrest.

Consistent with the figure of 15 percent, a little less than 19 percent of the offenders are arrested for a new domestic violence offense during the two-year follow-up period. Only 1.7 percent are arrested for incidents in which there was no physical injury, no attempts to cause physical injury, and no threat of physical injury (DV1) compared to about 17 percent who are arrested for incidents involving physical injury, an attempt to cause physical injury, or a threat of physical injury (DV2). The relative absence of DV1 incidents is perhaps counterintuitive, but police officers may be disinclined to make an arrest unless there is visible evidence of simple assault, aggravated assault, or rape. We have anecdotal information to this effect.[11]

Table 3 shows the random forests results in what many call a confusion table. It is nothing more than a cross-tabulation of actual outcomes against forecasted outcomes. The rows of the table are for the actual outcomes. The far-right entry in each row is the proportion of actual outcomes that is correctly identified. The rows condition on what really happened. The columns of the table are for the forecasted outcomes. The bottom

[11]Within the UCR program, police departments can clear offenses in one of two ways: by arrest or by "exceptional" means. Clearance by exception requires that there is: (1) an identified perpetrator, (2) sufficient evidence to make an arrest and support a charge, (3) an exact location of the perpetrator so that an arrest can be quickly made, and (4) circumstances beyond the control of the police that prohibit making an arrest, filing charges, or later prosecution. This leaves lots of room for judgment calls so that some domestic violence crimes can be cleared without an arrest. In addition, there is no need to clear a domestic violence crime if it is not officially defined as such. This gives police another avenue when the assault is thought to be insufficiently "serious."

Table 3: Random Forests Confusion Table: Actual Outcome Classes Tabulated Against Forecasted Outcome Classes Using Out-of-Bag Data

	Forecasts			
Actual	*DV0*	*DV1*	*DV2*	*Accuracy*
DV0	10,087	272	12,936	0.43
DV1	85	37	358	0.08
DV2	1,206	77	3,587	0.74
Accuracy	0.89	0.10	0.21	

NOTE: DV0 = no arrest for domestic violence; DV1 = a domestic violence arrest in which there is no physical injury, no attempt to cause physical injury, and no threat of physical injury; DV2 = a domestic violence arrest in which there is physical injury, an attempt to cause physical injury, or threat of physical injury. Accuracy measures are shown on the margins of the table. Cases that pose little risk can be forecasted accurately nearly 90 percent of the time.

entry in each column is the proportion of actual outcomes forecasted correctly. The columns condition on the forecast.

The rows and the columns address different questions. The rows provide information on how well a random forests application *classifies* known outcomes. Classification performance is used diagnostically to help determine the values of tuning parameters. The columns provide estimates of how well a random forests application will *forecast* when in practice the outcomes are not known. In the discussion to follow, the distinction between classification accuracy and forecasting accuracy is important and sometimes unappreciated.

The performance assessments in Table 3 are derived from "out-of-bag" test data not used to grow the random forest. This is a default feature of the random forest algorithm available in R.[12] The assessments are, therefore, "honest" and not subject to overfitting. Breiman (2001a) provides a formal proof.

A key factor affecting the entries in Table 3 is the costs assigned to different kinds of forecasting errors. These costs affect the forecasts made. As a mathematical matter, it is only the *relative* costs that count (Berk 2008:Sec. 5.5).[13] For example, all one needs to consider is that a given kind of false negative is, say, three times more costly than a given kind of false positive. As a policy matter, such cost ratios should be elicited from stakeholders but no claims about the monetized accuracy of those cost ratios need be made. One can think of the cost ratios as just the relative preferences of stakeholders. The main diagonal of the table contains the counts for the cases in which the actual outcome class corresponds to the forecasted outcome class. All the other cells contain

[12]The procedure in R, called "randomForest," was coded in fortran by Leo Breiman and Adele Culter. That code was ported to R by Andy Liaw and Matthew Weiner, who added a number of excellent enhancements. Because each tree is grown from a new random sample of the existing data, *with replacement*, about a third of the data on the average are available to each tree as test data (Breiman 2001a).

[13]This is a property of classification trees (Therneau & Atkinson 2015:Sec. 3.2) that carries over to random forests because a random forest is an ensemble of classification trees.

counts for different kinds of forecasting errors. One tries to fix the ratio of each possible pair of off-diagonal cells to correspond to the tradeoffs stakeholders prefer.

In this application, stakeholders believed that forecasting a DV0 when there was subsequently a DV2 was the worst possible false negative. They also believed that forecasting a DV2 when subsequently there was a DV0 was the worse possible false positive. The cost ratio of these kinds of false negatives to these kinds of false positives was set provisionally at 10 to 1.

It is usually not possible with real data and an outcome with three classes to hit desired cost ratios exactly. The problem is partly mathematical. The most effective way to introduce asymmetric relative costs allows only for three cost-weighting tuning parameters. Yet, the cost ratios depend on the six off-diagonal cells. A complicating factor is that the data used to construct a confusion table are based on the randomly selected out-of-bag data, which necessarily introduces some noise. Nevertheless, with some trial and error, one can often come quite close. Consider the 2 × 2 subtable that includes only DV0 and DV2 cases. For this analysis, we approximate the 10 to 1 cost ratio reasonably well (12,936/1,206 = 10.7 to 1).[14]

There are several cascading consequences for classification and forecasting accuracy that are shaped by the 10.7 to 1 cost ratio. First, the random forests algorithm will accept statistical evidence that is over 10 times *weaker* for a DV2 forecast than for a case forecasted as DV0. As a result, a large fraction of the actual DV2 cases are correctly classified. The bottom row of Table 3 shows that 74 percent of them are correctly classified. Second, it follows that, in trade, there will be a large number of DV2 false positives. Consequently, when a DV2 forecast is made, the last column shows that it is correct only 21 percent of the time. Third, it also follows that the relatively low costs assigned to DV2 false positives lead to a modest level of classification accuracy for DV0 cases: only 43 percent are correctly classified. Finally, because the 10.7 to 1 cost ratio means that very *strong* statistical evidence is required for a *forecast* of DV0, forecasting accuracy for DV0 is very high. In the first column, one can see that forecasts of no domestic violence arrests are accurate nearly 90 percent of the time. This is perhaps the most important policy result for reasons that will soon be addressed. It cannot be overemphasized that these tradeoffs result from stakeholder policy preferences built into the forecasts. A lower cost ratio than 10.7 to 1 would have led to fewer DV2 cases being correctly classified but also fewer false positives.

Similar reasoning for relative costs can be applied to comparisons between any two outcome categories. Consider the 2 × 2 subtable that includes only DV0 and DV1 cases. Forecasting a case as DV0 when it is subsequently a DV1 is 3.2 times worse than forecasting that a case is a DV1 when it is subsequently a DV0 (272/85). Consider the 2 × 2 subtable that includes only DV1 and DV2 cases. Forecasting a case as a DV1 when it is subsequently a DV2 is 4.6 times worse than forecasting a case as DV2 when it is subsequently a DV1 (358/77). All cost ratios can affect the forecasts made because they shape

[14]The difference between an actual cost ratio of 10 to 1 compared to an actual cost ratio of 10.7 to 1 makes no practical difference for this analysis.

the various tradeoffs built into the table. In this instance, the cost ratios in Table 3 are roughly consistent with the provisional preferences of stakeholders.[15] Those preferences could well change as the content of the pretrial reforms becomes more clear over time. Then, the forecasts would change as well.

In short, the empirical cost ratios computed from confusion tables are not findings. They are empirical realizations of the cost ratios determined by stakeholders. But as such, they affect forecasting accuracy. Findings are interpretations of that forecasting accuracy.

A closely related cost issue is that with more individuals released awaiting trial, more money is saved. From a fiscal perspective, the number of cases forecasted as DV0 ideally would be large. In principle, this is built into the cost ratios just discussed. For example, more cases will be forecasted as DV0 if the target cost ratio of 10 to 1 was decreased to, say, 5 to 1. That would also mean there will be fewer DV2 false positives and forecasts of DV2 would be correct more often. However, that would also lead to a smaller fraction of the DV2 cases being correctly classified. These matters will be revisited as the proposed pretrial reforms are clarified over the next year to 18 months.

With these complexities, it is important to provide policymakers with some reasonable, if only tentative, bottom line. Currently, about 20 percent of the individuals arraigned in domestic violence cases are arrested for new domestic violence offenses within two year of release. It follows that if one continued current practices and forecasted no new domestic violence arrests for each offender released, that forecast would be correct approximately 80 percent of the time. The 80 percent figure sets a very high accuracy standard with no use of any predictors whatsoever. Achieving better than 80 percent accuracy is a major challenge, especially given the weak set of forecasting inputs available at arraignment.

In fact, it is possible to do substantially better. In Table 3, the first column shows that forecasts of DV0 will be correct nearly 90 percent of the time. The random forest forecasts of no new domestic violence arrest easily clear a very high bar. This means that about 40 percent of all individuals arraigned for domestic violence can be released with the expectation that only about 1 in 10 will be rearrested for domestic violence.

Stated more programmatically, under current practices, about 20 percent of those released to await a hearing fail (i.e., have a new domestic violence arrest). If magistrates were to release only those offenders forecasted to not be rearrested, the failure rate for such offenders would be only about 10 percent. Were one to introduce some criminal justice reform that legitimately cut a domestic violence rearrest rate in half, it would likely be considered a major success.

It also makes good policy sense to consider forecasts for all three outcomes, and that is precisely what Table 3 provides. The costs stakeholders assigned to forecasting

---

[15]The product of the two smaller cost ratios is 14.7 to 1, which is a little too large when compared to 10.7 to 1. For example, if Outcome A is two times more costly than Outcome B, and Outcome B four times more costly than Outcome C, A should be eight times more costly than Outcome C. But as already noted, one cannot control the values of the empirical cost ratios exactly and stakeholders did not seem especially concerned about the lack of product equivalence. They thought we had it "about right."

errors differ substantially over the three outcomes. Moreover, the kinds of policy options appropriate for each outcome might differ dramatically. At this point, however, it is not clear what viable policy alternatives there are for the 60 percent of the cases forecasted to reoffend. Detention is an expensive option that should be reserved for the very high risk cases. Other less expensive options include diversion programs, electronic monitoring, and probation-like supervision. Requiring a bond may also be an option, although forfeiture is usually attached only to a failure to appear in court when required to do so. It might make sense, in addition, to hold a subset of high-risk offenders for up to seven additional days, during which time a far more thorough investigation could be undertaken. Such information could be used in a second machine learning forecasting procedure that probably would be able to identify a substantial number of new low-risk cases, and perhaps establish finer distinctions among those forecasted to be arrested for serious domestic violence. Insofar as these or other options become viable, the cost ratios built into the forecasts will change, and the forecasts will differ.

A. Input Contributions to Forecasting Accuracy

Although there is no ability or intent to formally identify risk factors as such, it can be helpful to consider how each input is related to the outcomes being forecasted. Stakeholders will be more inclined to adopt the random forests forecasting procedure if the inputs are related to the outcomes in a sensible fashion, especially if the relationships are consistent with stakeholder preconceptions.

Figure 1 is a plot showing the forecasting importance of each input for the outcome of no new DV arrests (DV0). On the horizontal axis is the reduction in forecasting accuracy if a given input is blocked from contributing to the forecasts. An outline of the algorithm can be found in the Appendix.[16] On the vertical axis is each input in order from high to low forecasting importance.[17] Thus, for example, "Violence Priors" has a forecasting importance of a little over 0.01. This means that if "Violence Priors" is blocked from contributing to the forecasts, forecasting accuracy for DV0 drops 0.01 (i.e., from 0.43 to 0.42). In part because there are so many inputs, forecasting importance is spread around thinly, and no single input matters very much. About a third of the inputs hardly matter at all. Negative importance figures are the result of the algorithm's normal random variation—the actual contribution is effectively zero.

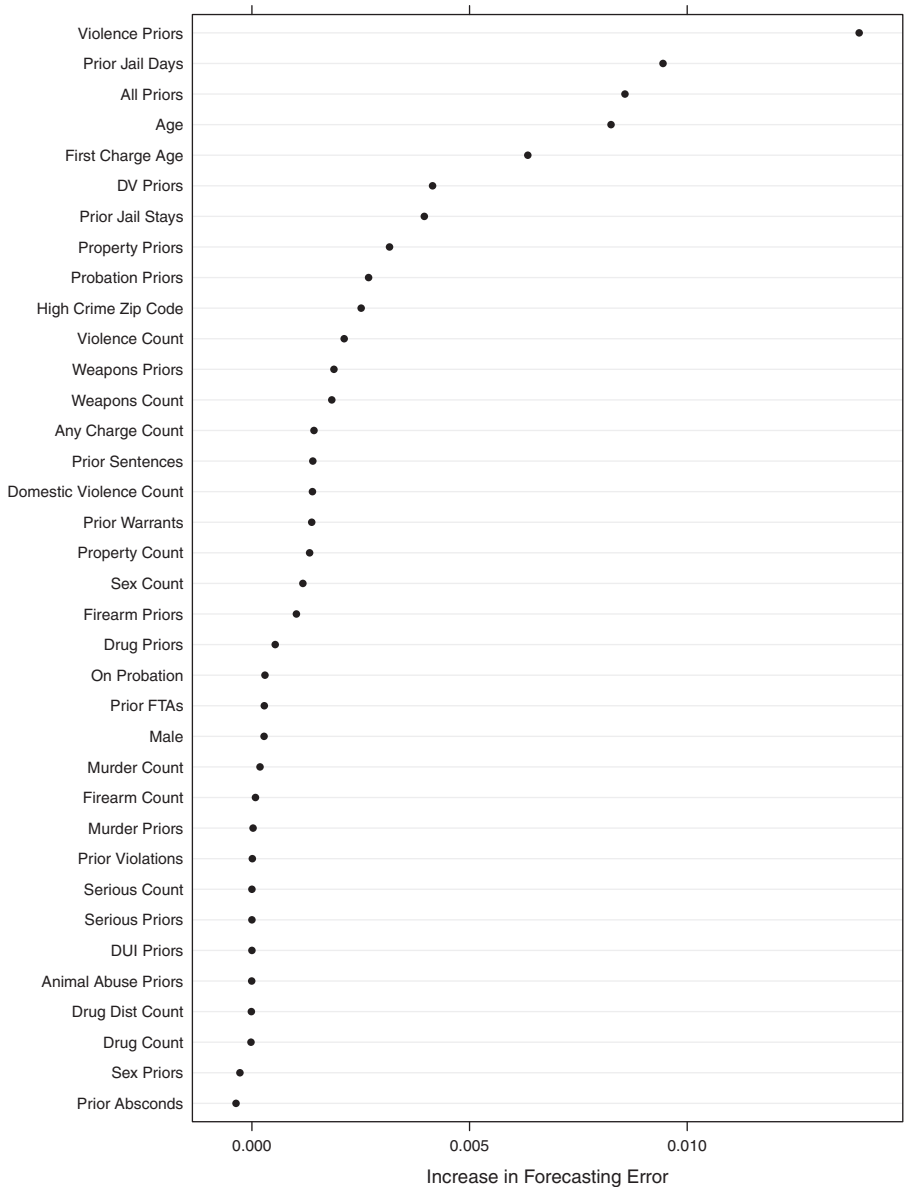
Forecasting importance plots could be reported for the two other outcome classes, and the order of the inputs would be at least somewhat different. One reason is that

---

[16]The random forest grown is unchanged. However, each predictor in turn is shuffled at random so that on the average it is unrelated to the outcome and cannot on the average contribute to forecasting accuracy. The shuffling does not change the random forest itself.

[17]For all inputs, "priors" following a kind of crime refers to prior charges associated with an arrest. "Count" following a kind of crime refers to the number of counts associated with the charges read at the arraignment. "High Crime Zip Code" is a categorical variable with 31 classes, one for each zip code characterized as a high crime area. "Prior Sentences," "Prior Abscondings," "Prior FTAs," "Prior Jail Days," and "Prior Jail Stays" refer to the number of times each event occurred.

Figure 1: Forecasting importance plot.



NOTE: The values on the horizontal axis show the reduction in forecasting accuracy for no new domestic violence arrests when a given variable is randomly shuffled. Various kinds of priors and age make the largest contributions to forecasting accuracy.

the reference categories would differ for each plot. For Figure 1, the reference categories are DV2 and DV1. The importance plot for DV1, for example, would have DV2 and DV0 as the reference categories. In other words, the baselines will differ. Because the policy focus is initially on forecasting DV0, Figure 1 is the relevant importance plot.[18]

Perhaps the major conclusion from Figure 1 is that the number of violence priors, prior number of days in jail, the number of all priors, age, age at which there is the first adult charge, the number of domestic violence priors, and the number of prior stays in jail make the largest contributions to forecasting accuracy. One inference is that domestic violence arrests can be seen as part of a pattern of arrests for criminal behavior in general and violent crimes in particular. Another inference is that prior record matters much more for forecasting accuracy than the charges associated with the current domestic violence incident. But, as noted earlier, with so many strong associations between the inputs, one should not make too much of the role of any single input.

B. Relationships Between Inputs of the Outcomes to be Forecasted

Forecasting importance plots do not indicate the direction of an association between an input and the outcome being forecasted. That information is contained in partial dependence plots. In effect, such plots show the nature of the association between a given input and the outcome being forecasted, all other inputs held constant. These are *not* covariance adjusted effects as one would have in a conventional regression analysis. They are more akin to matching procedures. That algorithm is outlined in the Appendix and the details can be found elsewhere (see Hastie et al. 2009:Sec. 10.13.2; Berk 2008:Ch. 5).

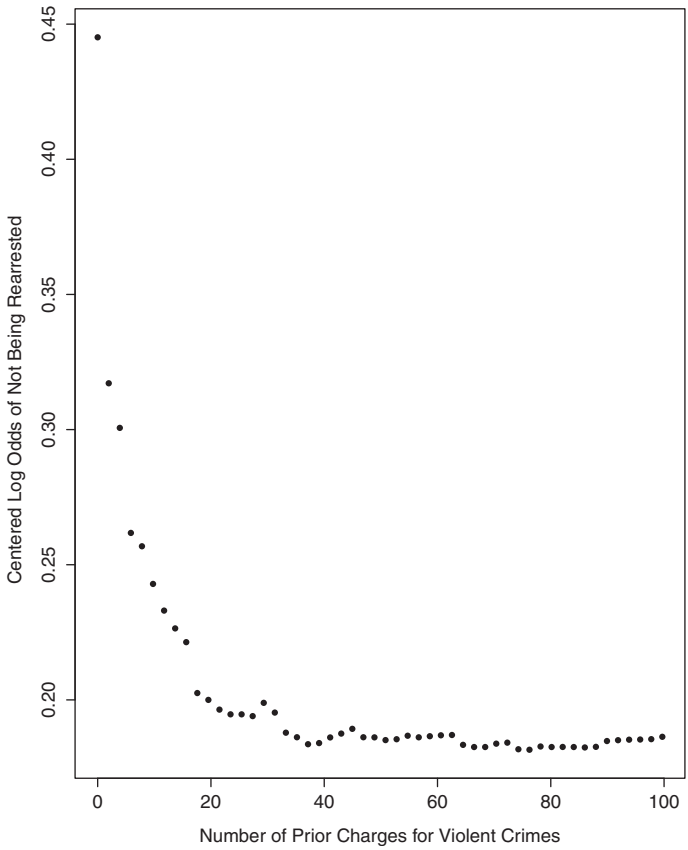
Figure 2 is a partial dependence plot for the relationship between DV0 and the number of prior charges for violent crimes in general.[19] The vertical axis is in centered logit units (Hastie et al. 2009:Sec. 10.13.2) so that larger values mean that a DV0 outcome is more likely. In other words, larger values are associated with a "success." As one would expect, Figure 2 shows that as the number of prior charges for violent crimes increases, the chances of no new DV arrests declines. But the relationship is highly nonlinear, as is often the case with partial dependence plots derived from machine learning.[20] The chances of no new DV arrests decrease dramatically with just a few priors for

[18]The order of the variables also could differ somewhat for other more technical reasons (Hastie et al. 2009:Sec. 15.3.2; Berk 2008:Ch. 5). Thus, the importance plot for DV2 will *not* be the same as the importance plot for DV0 in reverse order.

[19]The number of prior charges is generally less than 40. But, on occasion the number is very large because a substantial number of charges can be associated with a single arrest.

[20]When researchers let the data determine functional forms, rather than imposing functional (typically linear) forms with no compelling theoretical rationale, associations will often turn out to be highly nonlinear.

Figure 2: Partial dependence plot.



NOTE: On the vertical axis are the centered logits for not being rearrested for domestic violence. On the horizontal axis are the number of prior charges for violent crimes. The chances of not being rearrested for domestic violence decline very rapidly at first and then level off. The difference between no such priors and up to 20 is very important. The difference between 20 and a very large number does not appear to matter much. The largest effect is for no violent priors compared to one or more.

violent crimes and then level off. Transforming the centered logits into probabilities, the large falloff means a difference in probabilities of about 0.09.[21]

Partial dependence plots are not reported for each input because the conclusions are easily summarized. The directions of all effects are largely as one would expect. For

[21]It might seem preferable to transform the entire vertical axis in this manner, but when there are more than two outcome categories, there are significant technical complications. Among them is the reference category issue raised for importance plots (Hastie et al. 2009:370). Other than arbitrarily selecting one of the outcome categories as the reference, centering makes the mean logit the reference (analogizing to analysis of variance).

example, male offenders are more likely be arrested for new domestic violence incidents. However, the quantitative inputs generally have highly nonlinear associations with the outcome in which the strongest associations are found for smaller values of the input. Thus, individuals whose first adult charge is at an older age are less likely to be rearrested for domestic violence, but the difference between a first adult charge at 16 versus 25 matters a lot. The difference between a first adult charge at 36 versus 45 matters hardly at all.

There is perhaps one major surprise. In contrast to decades of research on street crime, the impact of the age of the offender falls off gradually. The reason seems to be, according to these data, that unlike street violence, domestic violence perpetration is nearly as common among individuals in their 30s and 40s as among individuals in their 20s.

V. Summary, Some Policy Implications, and Conclusions

Under current arraignment practices in the jurisdiction studied, arraignments are held within 48 hours of an arrest and are completed very quickly. Within a matter of several minutes, a magistrate reads the charges to the offender and decides whether the offender can be released until a subsequent formal hearing. If the offender is released, there can be release conditions, and a bond may be required that is forfeited if the offender fails to appear. The release decision is guided by two primary concerns: the risk of flight and the threat to public safety. Procedurally, domestic violence cases are handled at arraignment much like all other cases.

Release decisions in domestic violence cases could perhaps be improved if sufficiently accurate forecasts of repeat domestic violence arrests could be made. However, existing threat assessment tools for domestic violence typically concentrate exclusively on intimate partner violence and the needs of victims. Forecasting accuracy, even when properly accessed, can be subverted in service of other goals. In contrast, the setting in which arraignments are held requires a focus on offenders and statutory definitions of domestic violence. Accurate forecasts are meant to help inform decisions made by magistrates. To be sure, it is important to document a victim's needs, but that is another task to be undertaken in very different settings.

We are able to provide promising forecasting procedures even with data that are far less than ideal. Under current practice, about 20 percent of the individuals released after arraignment are arrested for domestic violence within two years. If magistrates only released offenders our forecasts identified as good bets, approximately 10 percent of those offenders would be arrested for domestic violence within two years. Failures could be cut in half. One would likely be pleased with any feasible intervention that performed as well. In the jurisdiction studied, the reduction in the percentage who fail translates into well over 1,000 fewer domestic violence arrests per year.

There is, of course, the question of what many fewer domestic violence arrests means for domestic violence incidents not reported to the police. One might argue that

because domestic violence arrests are made in response to domestic violence incidents, the number of incidents is also substantially reduced. But perhaps domestic violence offenders released after arraignment are more likely to threaten their victims with retaliation if the police are called. Arrests may be reduced without a meaningful drop in domestic violence incidents. A significant reduction in domestic violence arrests can lead to important practical and fiscal benefits for police, courts, prosecutors, and public defenders, but may or may not improve public safety.

However, the forecasting inputs contributing most to forecasting accuracy suggest that we are identifying for possible release offenders who are actually less inclined to reoffend. Note that several measures of prior arrests, including priors for domestic violence, all forecast a greater likelihood of new domestic violence arrests. If arraignments simply increased offenders' incentives to thwart calls to the police, one would find that, all else equal, offenders with longer prior records would have *fewer* post-arraignment domestic violence arrests because the domestic violence would be less likely to be reported. Nevertheless, we are exploring ways to collect better data on these issues.[22]

Still to be addressed is what should be done with offenders who are forecasted to be rearrested for domestic violence, especially domestic violence in which there are physical injuries. A variety of strategies is being considered. As noted above, cost is a very significant constraint, so jail time will have to be used sparingly. One of the least costly alternatives might be to inform the offender that the court will be making regular phone calls to the victim to ask "how things are going." Much the same protocol is part of many batterer intervention programs (BIPs). Another option might be probation-like supervision at several levels of intensity, some of which would include home visits. Diversion out of the court system is yet another possibility, and a diversion program for a small number of domestic violence offenders has actually been launched in the jurisdiction. After the arraignment, there is an initial screening for possible diversion. Those selected can be referred to a BIP that includes group counseling coupled with regular monitoring to help enforce attendance and improve victim safety. A failure to "buy in" can mean a return to court, where the usual options remain. The chances of success might be greater if only those forecasted as DV1 were diverted.[23]

Almost regardless of which interventions are being considered, it could be productive to impose a hold of up to a week on offenders forecasted to fail, during which better data could be collected to make better forecasts. Those forecasts would almost certainly find another group of low-risk offenders appropriate for release. Equally important, one might use the data to better anticipate which kinds of

---

[22]There could be two competing processes. On the one hand, offenders with longer prior records might have a greater proclivity to commit acts of domestic violence that through any of several mechanisms come to the attention of the criminal justice system. On the other hand, offenders with longer prior records might be more inclined to threaten their victims should the domestic violence be reported, and the threats could work. In our data, the first process seems to dominate.

[23]The track record for batterer intervention programs is worrisome (National Institute of Justice 2011). The local program is trying to address some of the more serious problems.

interventions would be most appropriate for which high-risk offenders. For example, BIPs might be best suited for offenders who have at least a high school diploma and are employed.

Unaddressed by our forecasts is what should be done about individuals who, under current practices, are detained until the next court appearance or are otherwise incarcerated for substantial periods after an arraignment.[24] Recall that such individuals could not be included in our study because they would not have sufficient opportunity to reoffend. This presents a common counterfactual problem in recidivism studies.

We do not know whether our forecasting procedures generalize well to such individuals. Yet, one can imagine wanting a forecast of what would happen if at least some were released because there are probably many who would be predicted to not be rearrested for domestic violence. For them, alternatives to incarceration might make cost-effective sense without compromising victim safety.

We have some information about the offenders who were released but who spent a substantial amount of time behind bars. Insofar as these offenders are similar to those who were not released, some comparative information is available. Offenders who were released but who spent substantial time in jail appear to be only a little different from those who were never incarcerated during the two-year followup. Offenders who spent time in jail were more likely to be male and to have somewhat more extensive prior records in general. But the number of priors for domestic violence was about the same, the number of domestic violence counts read at the arraignment was virtually identical, and they were only a little younger. In short, differences that could matter seem quite small.

If these conclusions also apply to the offenders who were never released, generalization problems may not be serious. Moreover, the question is not whether the descriptive statistics differ when those who are released are compared to those who are not released. The question is whether those differences imply that different application of random forests is required. In more conventional terms, should there be two "models" or one? Unless, in the future, magistrates are prepared to release some offenders who they would have otherwise not released, there is no definitive way to know.[25]

---

[24]There are many ways this could happen. For example, they may be formally released on the instant charge, but that release is essentially irrelevant because they have an existing warrant, are in violation of immigration law, or are being detained because their arrest violates the terms of their probation/parole. Or they might be released and sent home at arraignment, but then be arrested a few weeks later and spend the next 18 months in jail because of some unrelated offense.

[25]Some nice research designs are possible. For example, suppose that a magistrate released all offenders predicted not to be rearrested for domestic violence within two years. Also suppose that before looking at the forecast for each, the magistrate recorded what the decision would have been absent the forecast. The offenders who were released but who would not have been released without the forecast could provide some useful information about the key counterfactual and whether our current results generalize to those who currently are not released.

In the longer term, were better data available before an arraignment, more accurate forecasts could be obtained. In particular, there likely is very useful information in offense and arrest reports. The main obstacle in many jurisdictions, including the jurisdiction in this study, is obtaining that information in appropriate electronic form. One would first require a large training data set containing the outcomes to be forecasted and information from offense and arrest reports. Next, the data would need to be prepared for analysis. Among the many problems to be solved would be anticipating errors or inconsistencies in the data. Then for real-time forecasting, one would need a way to key-enter data from those forms in the period between an arrest and an arraignment. There are a variety of technically-driven shortcuts that can help, such as direct data entry from hand-held devices or from laptops in patrol cars. All of this is feasible if, as a policy matter, there is a commitment to providing magistrates with more accurate forecasts. But even under current circumstances, we can offer some potentially useful tools.

References

Adair, D. N. (2006) *The Bail Reform Act of 1984*. Washington, DC: Federal Judicial Center.

Arnold Foundation (2013) "Developing a National Model for Pretrial Risk Assessment," research summary from the Laura and John Arnold Foundation. Available at: www.arnoldfoundation.org.

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

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

Berk, R. A., G. Barnes, L. Ahlman, & E. Kutrtz (2010) "When Second Best is Good Enough: A Comparison Between a True Experiment and a Regression Discontinuity Quasi-Experiment," 6(○) *J. of Experimental Criminology* 191.

Berk, R. A., & J. Bleich (2013) "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment," 12(3) *J. of Criminology & Public Policy* 513.

—— (2014) "Forecasts of Violence to Inform Sentencing Decisions," 12(3) *J. of Quantitative Criminology* 515.

Berk, R. A., & J. Hyatt (2015) "Machine Learning Forecasts of Risk to Inform Sentencing Decisions," 27(4) *Federal Sentencing Reporter* 222.

Berk, R. A., B. Kriegler, & J.-H. Baek (2006) "Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures," 22(2) *J. of Quantitative Criminology* 141.

Berk, R. A., L. W. Sherman, G. Barnes, E. Kurtz, & L. Ahlman (2009) "Forecasting Murder in a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning," 172(1) *J. of the Royal Statistical Society (Series A)* 191.

Berk, R. A., S. B. Sorenson, & Y. He (2005) "Developing a Practical Forecasting Screener for Domestic Violence Incidents," 29(4) *Evaluation Rev.* 358.

Bock, W., & C. Frazier (1977) "Official Standards Versus Actual Criteria in Bond Dispositions," 5(4) *J. of Criminal Justice* 321.

Bornstein, B. H., A. J. Tomkins, E. M. Neeley, M. N. Herian, & J. A. Hamm (2013) "Reducing Courts Failure-to-Appear Rate by Written Reminders," 19(1) *Psychology, Public Policy & Law* 70.

Bridges, G., R. Crutchfield, & E. Simpson (1987) "Crime, Social Structure and Criminal Punishment: White and Nonwhite Rates of Imprisonment," 34 *Social Problems* 345.

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

—— (2001b) "Statistical Modeling: The Two Cultures," 16(3) *Statistical Science* 199.
Cohen, A. (2012) "Wrongful Convictions: A New Exoneration Registry Tests Stubborn Judges," May *Atlantic* 21.
Demuth, S. (2003) "Racial and Ethnic Differences in Pretrial Release Decisions and Outcomes: A Comparison of Hispanic, Black and White Felony Arrestees," 41 *Criminology* 873.
Devers, L. (2011) *Bail Decisions: Research Summary*. Washington, DC: Bureau of Justice Assistance, U.S. Department of Justice.
Farrington, D.P., & R. Tarling (2003) *Prediction in Criminology*. Albany, NY: SUNY Press.
Federal Judicial Center (1993) *The Bail Reform Act of 1984*, 2d ed. Washington, DC: Federal Judicial Center.
Frazier, C., E. W. Bock, & J. C. Henretta (1980) "Pretrial Release and Bail Decisions: The Effects of Legal, Community, and Personal Variables," 18(2) *Criminology* 162.
Goldkamp, J. S., & M. R. Gottfredson (1985) *Policy Guidelines for Bail—An Experiment in Court Reform*. Philadelphia, PA: Temple University Press.
Goldkamp, J. S., & M. D. White (2006) "Restoring Accountability in Pretrial Release: The Philadelphia Pretrial Release Supervision Experiments," 2 *J. of Experimental Criminology* 143.
Gottfredson, S. D., & L. J. Moriarty (2006) "Statistical Risk Assessment: Old Problems and New Applications," 52(1) *Crime & Delinquency* 178.
Hastie, T., R. Tibshirani, & J. Friedman (2009) *Elements of Statistical Learning: Data Mining. Inference, and Prediction*, 2d ed. New York: Springer.
Jordan, M. I., & T. M. Mitchell (2015) "Machine Learning: Trends, Perspectives, and Prospects," 349(6234) *Science* 255.
LaFrentz, C. D., & C. Spohn (2006) "Who is Punished More Harshly in Federal Court? The Interaction of Race/Ethnicity, Gender, Age, and Employment Status in the Sentencing of Drug Offenders," 8(2) *Justice Research & Policy* 25.
McElroy, J. E. (2011) "Introduction to the Manhattan Bail Project," 24(1) *Federal Sentencing Reporter* 8.
National Institute of Justice (2011) Batterer Intervention Programs Often Do Not Change Offender Behavior. Available at: http://www.nij.gov/topics/crime/intimate-partner-violence/interventions/Pages/batterer-intervention.aspx.
Reitler, A. K., C. Sullivan, & J. Frank (2013) "The Effects of Legal and Extra Legal Factors on Detention Decisions in US District Courts," 30(2) *Justice Q.* 340.
Ridgeway, G. (2013) "The Pitfalls of Prediction," 271 *NIJ J.* 271.
Roehl, J., C. O'Sullivan, D. Webster, & J. Campbell (2005) *Intimate Partner Violence Risk Assessment Validation Study, Final Report*. Washington, DC: U.S. Department of Justice, National Institute of Justice.
Spohn, C. (2009) "Race, Sex, and Pretrial Detention in Federal Court: Indirect Effects and Cumulative Disadvantage," 57 *Kansas Law Rev.* 879.
Therneau, T. M., & E. J. Atkinson (2015) *An Introduction to Recursive Partitioning Using RPART Routines*. Available at: http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.
U.S. Department of Justice (2014) http://www.justice.gov/ovw/domestic-violence.
VanNostrand, M., & G. Keebler (2009) *Pretrial Risk Assessment in the Federal Court*. Washington, DC: Office of the Federal Detention Trustee, U.S. Department of Justice.

Appendix: Algorithms

Outline of the Random Forests Algorithm

1. From a training data set with N observations, take a random sample of size N *with replacement*. Observations not chosen at random become the test "out-of-bag" data.
2. Take a random sample *without replacement* of the predictors (e.g., 5).
3. Construct the first classification tree partition of the data.
4. Repeat Step 2 for each subsequent split until the classification tree is as large as desired. Do not prune.
5. Drop the out-of-bag data (i.e., data not used to grow the tree) down the tree. Store the class assigned to each observation along with each observation's predictor values.
6. Repeat Steps 1–5 a large number of times (e.g., 500).
7. Using only the class assigned to each observation when that observation is not used to grow the tree, count the number of times over trees that the observation is classified in each outcome category.
8. Assign each case to an outcome category by a plurality vote over the set of trees.

Outline of the Variable Forecasting Importance Algorithm

1. Construct a measure prediction error for each classification tree as usual by dropping the out-of-bag data down the tree. Note that this is a real forecasting enterprise because data not used to grow the tree are used to evaluate its predictive skill.
2. If there are p predictors, repeat Step 1 p times, but each time with the values of the given predictor randomly shuffled. The shuffling makes that predictor on the average unrelated to the response and all other predictors. For each shuffled predictor individually compute a new measure of prediction error.
3. For each of the p predictors, average over trees the difference between the prediction error with no shuffling and the prediction error with a given predictor shuffled. This is the measure of variable forecasting importance.

Outline of the Partial Dependence Plot Algorithm

1. Grow a forest as usual.
2. Suppose x_1 is the initial predictor of interest, and it has v distinct values in the training data. Construct v data sets as follows.
   a. For each of the v values of x_1, make up a new data set where x_1 only takes on that value, leaving all other variables unchanged.
   b. For each of the v data sets, predict the response using random forests. There will be a single value averaged over all observations.
   c. Average each of these predictions over the trees.
   d. Plot the average prediction for each value for each of the v data sets against the v values of x_1
3. Go back to Step 2 and repeat for each predictor.