Algorithmic approach to forecasting rare violent events
An illustration based in intimate partner violence perpetration

Richard A. Berk | Susan B. Sorenson

University of Pennsylvania

Research Summary: Mass violence, almost no matter how defined, is (thankfully) rare. Rare events are difficult to study in a systematic manner. Standard statistical procedures can fail badly, and usefully accurate forecasts of rare events often are little more than an aspiration. We offer an unconventional approach for the statistical analysis of rare events illustrated by an extensive case study. We report research aimed at learning about the attributes of very-high-risk intimate partner violence (IPV) perpetrators and the circumstances associated with their IPV incidents reported to the police. “Very high risk” is defined as having a high probability of committing a repeat IPV assault in which the victim is injured. Such individuals represent a very small fraction of all IPV perpetrators; these acts of violence reported to the police are rare. To learn about them nevertheless, we sequentially apply in a novel fashion three algorithms to data collected from a large metropolitan police department: stochastic gradient boosting, a genetic algorithm inspired by natural selection, and agglomerative clustering. We try to characterize not just perpetrators who on balance are predicted to reoffend but also who are very likely to reoffend in a manner that leads to victim injuries. Important lessons for forecasts of mass violence are presented.

Policy Implications: If one intends to forecast mass violence, it is probably important to consider approaches less
dependent on statistical procedures common in criminology. Given that one needs to “fatten” the right tail of the rare events distribution, a combination of supervised machine learning and genetic algorithms may be a useful approach. One can then study a synthetic population of rare events almost as if they were an empirical population of rare events. Variants on this strategy are increasingly common in machine learning and causal inference. Our overall goal is to unearth predictors that forecast well. In the absence of sufficiently accurate forecasts, scarce resources to help prevent mass violence cannot be allocated where they are most needed.

**KEYWORDS**
forecasting, genetic algorithms, intimate partner violence, machine learning, mass violence, synthetic populations

Forecasts of risk are routinely made in a wide variety of situations. What is the probability that a hurricane will strike the Gulf Coast in a particular hurricane season? What is the probability that a given high school student will be accepted by his or her college of choice? What is the probability that a particular business firm will declare bankruptcy? Coupled with each probability is the expected cost should the event of concern occur. For the bankruptcy example, repayment of debt at 10 cents on the dollar means a loss of 90 cents for every dollar invested. Risk formally is defined as the costs of a particular event multiplied by the probability that the event will occur.

Forecasts of risk can be useful if they lead to actions that are better informed. For undesirable outcomes, one hopes that prevention strategies can be implemented or that plans for remedial action after the fact can be made. This has long been well understood by criminal justice decision makers in the United States. Indeed, risk assessments have been used to inform criminal justice decisions since the 1920s (Burgess, 1928). One might wonder, therefore, whether forecasts of risk might be instructive for contemporary incidents of mass violence. Without good forecasts, scarce prevention and remedial resources easily can be misallocated. One cannot, for example, place armed guards at every church, mosque, or synagogue. Likewise, one cannot have grief counsellors at every business establishment. Currently, those resources are distributed in a haphazard manner. Might legitimate risk assessments help?

For almost any reasonable definition of mass violence, constructing sufficiently accurate forecasts is a daunting undertaking. This holds whether one is trying to forecast the likely perpetrator, location, or timing of an event. One obstacle is that mass violence is heterogeneous. It can include school shootings, homicides committed by disgruntled employees, brutal hate crimes, systematic execution of witnesses at a crime scene, fatal assaults by perpetrators of intimate partner violence (IPV), and other mass violence in which the motives are obscure (e.g., the October 2017 Las Vegas music festival mass shooting in which 58 people were killed and 851 were injured). Although understanding mass violence in general is an admirable aspiration, in the medium term, different forms of mass violence might be productively examined separately. Useful forecasts probably will require different
approaches for different kinds of mass violence because the risk factors and their importance likely will vary.

Another obstacle is achieving a consensus about what the most relevant observational unit should be. One important distinction is between the settings in which the violence occurs and the people found in these settings. Does one want a forecast for a school as a whole or a forecast for each student in that school? Likewise, should the observational units be businesses or their individual employees? What about places of worship versus individual members of their congregations? In addition, the setting may be a kind of event rather than a place. For example, the observational units may be armed robbery incidents or rock concerts. Simply put, what goes in the denominator when risk probabilities are to be estimated?

Even if clear definitions for different kinds of mass violence could be provided and, for each, sensible observational units specified, a third obstacle is very low base rates. One consequence is that the raw numbers of such events will be small, often in no more than double digits. For example, one large metropolitan area had in a recent year fewer than 10 homicides related to intimate partner violence, and for none were there more than 3 victims; most had a single victim. One would need to accumulate intimate partner homicides from across the country to arrive at a mass violence total of more than ∼10 such incidents per year (Krouse & Richardson, 2015, p. 18). Not much information can be extracted from so few observations, especially when one might hope to learn what risk factors distinguish IPV mass violence incidents from the hundreds of thousands of IPV incidents in which no one dies.

A more subtle concern is that very low base rates lead to accurate but trivial forecasts. For example, Time magazine reported that in 2018, there were a total of 17 “school shootings” in the United States (Wilson, 2018). Suppose during that year there were approximately 100,000 public schools in the United States (National Center for Education Statistics, 2018). The probability that any given public school will be victimized by a school shooting in that year is .00017. If one had forecasted that for any given school there would be no shootings in 2018, that forecast would have been correct with a probability of over .999 using no risk factors whatsoever. It is hard to imagine that any forecasting procedure with risk factors could do better. If one proceeded nevertheless with standard statistical tools, it is likely that no useful risk factors would be identified. The findings from the numerical methods used would rapidly reveal that nothing to improve forecasting accuracy could be found. So why bother?

The answer lies in the costs of mass violence. Although mass violence is rare, it can have devastating consequences. In addition to the tragic loss of life and the grieving of family members and friends, mass violence can undermine trust in government institutions to guarantee public safety. Mass violence also can weaken confidence in appointed and elected public officials and elicit racial, ethnic, and religious scapegoating. For these reasons and others, efforts to reduce mass violence can be terribly important. Risk forecasting can help, at least in principle. In this article, we consider ways to estimate the probabilities of mass violence assuming that the costs will be large by almost any metric. Hence, the challenge.

Effective forecasts of mass violence may prove to be useful, but the statistical obstacles are formidable. Herein, we will illustrate the potential of a novel approach to forecasting rare events. In part because of our access to unique data, the test bed is comprised of incidents of IPV in which the victim sustains injuries. Such incidents are usually not crimes of mass violence, but as a form of intentional violence, they raise many of the same statistical difficulties. In particular, for a typical set of IPV incidents, cases in which the victim is injured are rare.
1 | IPV RISK ASSESSMENT WITH LOW BASE RATES

Like most criminal justice risk assessments, risk assessments for IPV typically use very broad definitions of the forecasting target. Often the forecasting target is simply the presence or absence of any actions that qualify under existing statutes; for example, a loud argument can suffice. At the other extreme is a lethal assault. Consequently, the usual search for risk factors can be compromised by heterogeneous outcomes. An important risk factor for an argument may be an unimportant risk factor for an assault causing injuries.

In a few studies, researchers have narrowed their focus to very serious forms of IPV in which the victim is injured or even killed. Such outcomes make the research extremely important. But to be effective, the researchers must overcome very low base rates, making identification of risk factors immensely difficult.

In the pages ahead, we address and try to circumvent the problems caused by low base rates for IPV in which the victim is injured. Using a unique data set, we focus on the attributes of very-high-risk IPV perpetrators and the circumstances associated with their IPV incidents that are reported to the police. “Very high risk” is characterized as having a high probability of committing a repeat IPV assault in which the victim is injured.

Rather than rely solely on a conventional statistical analysis of IPV incidents, we apply three algorithms sequentially to data from a large metropolitan police department: stochastic gradient boosting, a genetic algorithm inspired by natural selection, and agglomerative clustering. The first is used to define a fitness function, the second is used to construct a population of very-high-risk IPV offenders, and the third is used to help visualize the results. The constructed population does not have a problem with low base rates, and instructive results are obtained.

2 | PAST RESEARCH

The vast amount of literature on risk factors for IPV can be organized into three categories. In some studies, scholars have tried to construct a causal account in which risk factors are treated as causes. Pathbreaking work by Straus and Gelles (1990) is an excellent example. Abramsky and colleagues (2011) provided an international example. Weitzman (2018) recently continued in this tradition, asserting that greater educational achievement for women could reduce victimization. For our purposes, such work is peripheral because intimate partner violence typically is too broadly defined. For example, IPV can comprise threats, yet can include severe injuries requiring medical care. These are treated as different manifestations of the same underlying phenomenon.

In a second tradition, scholars have used risk factors to characterize the ongoing dangers faced by victims of intimate partner violence. This approach can be traced back to work by Campbell (1995) and has led to several important follow-up studies (Campbell, Glass, Sharps, Laughon, & Bloom, 2007; Campbell, Webster, & Glass, 2009; Campbell, Webster, Koziol-McLain, et al., 2003). Storey and Hart (2014) provided a recent example of the strengths and weakness of this approach. Although the attention to very serious IPV, often homicide, fits within our goals, the concern with explanation rather than with prediction does not. A risk factor that may help explain why a homicide is more likely may have little forecasting power. Perhaps the major hurdle for such research, however, is the very low base rate. Lethal intimate partner violence, although certainly tragic, is rare.

A final approach by scholars has centered on forecasting, typically to help inform criminal justice actions (Berk, 2018; Berk, Sorenson, & Barnes, 2016; Berk, Sorenson, & He, 2005; Cunha & Goncalves, 2016). There usually is no causal account because risk factors are evaluated primarily by
how much they improve forecasting accuracy. The research cited can include intimate partner violence in which there are injuries or even fatalities, but it, too, is challenged by very low base rates.

3 | DATA

For all domestic violence dispatches confirmed as domestic violence cases by responding officers, a special offense form was filled out. We had worked with the local police department to design the form, in which a much wider range of information was elicited than what previously had been collected. (The form is still in use.) We were provided with a total of more than 54,000 forms for the calendar year 2013. Each form characterized a domestic violence incident.

Domestic violence was broadly defined, as is customary in law enforcement, to include disputes between parents and children, between siblings, and other variants on “domestic,” including intimate partners. We reorganized the data to include only incidents of IPV with the perpetrator as the analysis unit. Using the perpetrators as the unit of analysis is more than a technical fix. Because police are the first responders and because of their law enforcement mission, the most directly relevant individual in an incident of intimate partner violence is the perpetrator. Focus on the perpetrator is a unifying theme in this article.

Once irrelevant incidents were removed (e.g., a request for information only), there were 22,449 IPV perpetrators. For each perpetrator, we used the information in the earliest recorded incident in 2013 as our platform from which to forecast whether the victim in any subsequent incidents by the same perpetrator in that year were recorded by police as having physical injuries. Approximately 20% of the perpetrators in the initial incident had at least one subsequent IPV incident in 2013, and approximately 5% had a subsequent IPV incident in 2013 in which the victim was injured. Repeat IPV incidents with reported victim injuries are rare, which presents a substantial data analysis challenge. Nevertheless, our response variable is whether a given perpetrator has a subsequent IPV incident during 2013 in which the victim is injured. By “subsequent” we mean chronologically later than the IPV incident in 2013 that constitutes the baseline data. Further details about the data are provided in a paper by Sorenson (2017).

We selected all predictors from the collected offense forms for each perpetrator’s initial IPV incident in 2013. The far left column of Table 2 lists the predictors used. All are indicator variables coded so that a “1” represents the presence of the attribute and a “0” represents the absence of the attribute. It will later help conceptually if one thinks of a “1” as switching a predictor on and a “0” as switching a predictor off. We say more about the predictors shortly.

4 | METHOD

As an initial benchmark and to motivate our statistical approach, a conventional logistic regression was applied to the data. Poor performance was expected. Because 95% of the perpetrators did not commit a new, reported IPV incident in which the victim was injured, one can predict using no predictors that such an incident will not occur and automatically be right 95% of the time. One cannot expect logistic regression to do any better.

Figure 1 illustrates the problem. The mass of fitted probabilities fall around .04, and none exceed .27. The highest risk perpetrator had less than a 30% chance of reoffending. With no fitted probabilities larger than .50, no perpetrators would be forecasted to be reported for a new IPV incident in which the victim was injured.
Even with such difficult data, machine learning procedures can do better (Berk & Bleich, 2013). We applied stochastic gradient boosting (\textit{gbm} in R).\footnote{Building on past forecasting studies of domestic violence in which machine learning was used (Berk et al., 2016), our target cost ratio treated false negatives as 10 times more costly than false positives. In other words, failing to classify correctly an IPV incident with injuries was 10 times worse than failing to classify correctly an IPV incident with no injuries.}

We use the cost ratio as a place holder. When working with stakeholders on real applications, the cost ratio becomes a policy preference they would need to specify. But the 10-to-1 ratio is plausible. In any case, for the analyses that follow, the target cost ratio is peripheral to our main concerns.

The data were randomly divided into training data having 20,000 observations and into test data having 2449 observations. For reasons that will be apparent later, we will lean far more heavily on the training than on the test data, which justifies the substantially larger number of training observations. There is no other formal rationale for our setting (cf. Faraway, 2016).

Because the outcome was binary, we used the conventional Bernoulli distribution to define the boosting residuals. We retained all of the \textit{gbm} default settings for the tuning parameters except that interaction depth was set at 10 to help capture the rare outcome events we were seeking. Reasonable variation in the tuning parameters (e.g., an interaction depth of 6) made little difference. The number of iterations was determined by fivefold cross-validation. There are difficult technical problems with cross-validation, but it seems to perform well in practice (Hastie, Tibshirani, & Friedman, 2009, Section 7.10.). Such performance is important for boosting, which can badly overfit outcome probabilities (Mease, Wyner, & Buja, 2007).

The primary boosting output of interest was the learned fitting function, which can be used to construct fitted values, conventionally treated as probabilities.\footnote{Here, they convey the chances that a perpetrator will commit a new IPV incident in which the victim is injured. As discussed shortly, the fitted probabilities ranged from a little more than .30 to a little more than .70. Approximately one quarter of the perpetrators were predicted to reoffend in a manner leading to victim injuries (i.e., fitted probability > .50).} But these results also were unsatisfactory. Even with the 10-to-1 cost ratio, the results were dominated by the perpetrators who did not reoffend because it was so easy to fit those cases accurately. Moreover, many predictors that might have been useful for identifying the repeat perpetrators could...
not be discerned because the unbalanced response variable precluded them from having much impact on the boosting loss function. In addition, the predictor values for the repeat, violent offenders were limited to those in the data set. It follows that there could be other kinds of very-high-risk perpetrators with the same predictors but with different configurations of predictor values. For example, there might be no perpetrators in these data who were reported to strangle the victim when children were present, although this confluence of predictor values might be found in other data sets.

For these reasons, we extended the analysis strategy. We applied a genetic algorithm (Luke, 2013, ch. 3; Mitchell, 1998) with the goal of constructing predictor profiles that a variety of very-high-risk offenders might have. Use of this approach allowed for us to not be limited to the actual predictor values for such offenders who were in our data. Insofar as a substantial number of very-high-risk, hypothetical perpetrators “survived,” it would be possible to determine which predictors and predictor values were responsible. Put in other terms, using the genetic algorithm, we sought to construct a new population of very-high-risk, violence-prone perpetrators that could be studied in the same way one would study an observed, empirical population. In this manner, we hoped to circumvent, at least in part, the statistical problems caused by low base rates. Genetic algorithms have been applied in economics, population genetics, ecology, immunology, and biology (Mitchell, 1998, Section 1.8); our application is novel. Possible concerns about treating “synthetic” data as real data will be addressed later.

We applied the genetic algorithm GA available in R. The earlier gbm prediction module developed from training data served as the fitness function. Perpetrators with larger predicted probabilities to violently reoffend were defined as more fit. This meant that if predictors for members of the hypothetical population of very-high-risk offenders were used as input data for the learned boosting results, the predicted probabilities of repeat violence would be much greater than .50.

One hundred populations, each with 500 perpetrators, were constructed sequentially by the genetic algorithm, although there was little improvement after about the twentieth population. Just as with gbm, we found that the default values for the tuning parameters worked well. There was no meaningful improvement when the tuning parameters were varied within reasonable values. In the end, we had a single population of 500 particularly unsavory perpetrators.

Finally, we sought to characterize the most important predictors of repeat IPV when the victim is injured and how those predictors were related to one another. In addition to some simple calculations for the population of 500, we applied an agglomerative clustering algorithm (Kaufman & Rousseeuw, 2005, ch. 5), a form of unsupervised learning (agnes in the cluster library in R). The clusters produced confirmed our earlier conclusions in an easily understood visualization. We also devised a new way to estimate predictor importance.

## 5 | BOOSTING RESULTS

Table 1 shows a conventional, machine learning classification table for the test data constructed from the output of the stochastic gradient boosting application. Even with an achieved 10-to-1 target cost ratio (i.e., 763 / 77 = 9.9), correctly classifying the rare cases was difficult. The row labeled “Actual

<table>
<thead>
<tr>
<th>Variable</th>
<th>Forecast No Injuries</th>
<th>Forecast Injuries</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No Injuries</td>
<td>1,542</td>
<td>763</td>
<td>0.33</td>
</tr>
<tr>
<td>Actual Injuries</td>
<td>77</td>
<td>67</td>
<td>0.53</td>
</tr>
<tr>
<td>Forecasting Error</td>
<td>0.04</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>
Injuries” shows that 53% of the cases in which there was a repeat IPV incident with injuries were incorrectly classified. Although a dramatic improvement over the logistic regression results in which all such incidents were misclassified, the classification error rate for the rare events is hardly inspiring. The misclassification rate when there are no injuries is smaller (33%), but because injury-free IPV is by far the most common outcome, the classification task is much easier.

Turning from classification to forecasting, as shown in the columns of Table 1, the target 10-to-1 cost ratio led to more than 750 false positives that, in turn, resulted in a forecasting error for repeat violence of 92%; forecasts of repeat violence would be wrong 92% of the time. We do far better with forecasts of the absence of repeat violence, but an error rate of 4% is only slightly better than the 5% error rate one would obtain by simply applying a Bayes classifier to the marginal distribution of the response variable. In short, compared with using only the marginal distribution of the response variable, the predictors do not help much if the goal is more accurate forecasts. The improvement for incidents in which there were no injuries could be somewhat greater if the relevant cost ratio were increased.

For our purposes, far more instructive are the fitted values. Figure 2 shows that the fitted probabilities range from a little more than .30 to a little more than .70. Clearly, there is a dramatic improvement over the fitted values from the logistic regression. A substantial number of fitted probabilities are to the right of .50, leading to forecasts of victim injuries. From Table 1, however, we know that most of these forecasts are false positives. Moreover, probabilities larger than .50 vary widely with a small minority having the among the highest risk probabilities. Finally, because the probabilities range widely, perhaps there is a range of etiologies leading to IPV injuries. Perpetrators forecasted to injure their intimate partners can differ substantially from one another. For example, some may on occasion lose control, perhaps, under the influence of drugs or alcohol. Some may use violence on a routine basis systematically to enforce domination (Johnson, 2008).

For this article, more important than the fitted values is how the predictors contribute to the fitted probabilities. The far left column of Table 2 shows all of the predictors in order of their fitting importance according to gbm. Fitting importance for a given predictor is defined as the average reduction in the deviance over the boosted regression trees used by gbm. That is, fitting importance is the in-sample, average contribution to the fit of the data. The second column shows the importance of each predictor as a proportion of the total deviance accounted for over all predictors. We call this “in-sample importance.”

By far, the most important predictor is whether the initial IPV incident occurred after the first 90 days of 2013. If so, the follow-up period was between 9 and 12 months. A partial dependence plot showed
Table 2  Predictor importance forecasting victim injuries

<table>
<thead>
<tr>
<th>Candidate Predictors</th>
<th>In-Sample Importance</th>
<th>Commonality Importance</th>
<th>Switched On or Off or In Between</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up &gt; 3 months</td>
<td>27.12</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Prior DV Reports</td>
<td>4.77</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Victim &lt; 30</td>
<td>3.64</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Contact Information Given to Victim</td>
<td>3.48</td>
<td>0.00</td>
<td>Always Off</td>
</tr>
<tr>
<td>Offender Arrested</td>
<td>3.23</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Offender &lt; 30</td>
<td>3.18</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Victim Frightened</td>
<td>3.13</td>
<td>0.43</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Polite</td>
<td>3.08</td>
<td>0.00</td>
<td>Always Off</td>
</tr>
<tr>
<td>Currently Married</td>
<td>2.70</td>
<td>0.73</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Cooperative</td>
<td>2.67</td>
<td>0.62</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender White</td>
<td>2.66</td>
<td>0.20</td>
<td>In Between</td>
</tr>
<tr>
<td>Victim Shaking</td>
<td>2.66</td>
<td>0.44</td>
<td>In Between</td>
</tr>
<tr>
<td>Visible Injuries</td>
<td>2.61</td>
<td>0.49</td>
<td>In Between</td>
</tr>
<tr>
<td>Victim Strangled</td>
<td>2.50</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Victim Latina</td>
<td>2.44</td>
<td>0.68</td>
<td>In Between</td>
</tr>
<tr>
<td>Victim Crying</td>
<td>2.34</td>
<td>0.49</td>
<td>In Between</td>
</tr>
<tr>
<td>Children Present</td>
<td>2.33</td>
<td>0.49</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Threatened</td>
<td>2.22</td>
<td>0.26</td>
<td>In Between</td>
</tr>
<tr>
<td>Weapon Used</td>
<td>2.17</td>
<td>0.26</td>
<td>In Between</td>
</tr>
<tr>
<td>Former Relationship</td>
<td>2.10</td>
<td>0.31</td>
<td>In Between</td>
</tr>
<tr>
<td>PFA Ever</td>
<td>2.02</td>
<td>1.00</td>
<td>Always On</td>
</tr>
<tr>
<td>Offender Angry</td>
<td>1.95</td>
<td>0.42</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Apologetic</td>
<td>1.82</td>
<td>0.12</td>
<td>In Between</td>
</tr>
<tr>
<td>Furniture in Disarray</td>
<td>1.71</td>
<td>0.00</td>
<td>Always Off</td>
</tr>
<tr>
<td>Relationship Breaking Up</td>
<td>1.54</td>
<td>0.32</td>
<td>In Between</td>
</tr>
<tr>
<td>Victim’s Clothes in Disarray</td>
<td>1.47</td>
<td>0.76</td>
<td>In Between</td>
</tr>
<tr>
<td>Evidence Collected</td>
<td>1.38</td>
<td>0.76</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Stalked</td>
<td>1.23</td>
<td>0.47</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Black</td>
<td>1.17</td>
<td>0.63</td>
<td>In Between</td>
</tr>
<tr>
<td>Taken to Hospital</td>
<td>1.13</td>
<td>0.36</td>
<td>In Between</td>
</tr>
<tr>
<td>Formerly Married</td>
<td>1.12</td>
<td>0.35</td>
<td>In Between</td>
</tr>
<tr>
<td>Statements Taken from Kids</td>
<td>1.07</td>
<td>0.22</td>
<td>In Between</td>
</tr>
<tr>
<td>Offender Broke In</td>
<td>0.74</td>
<td>0.22</td>
<td>In Between</td>
</tr>
<tr>
<td>PFA Expired</td>
<td>0.57</td>
<td>0.51</td>
<td>In Between</td>
</tr>
</tbody>
</table>

Notes. DV = domestic violence; PFA = protection from abuse order.

The relationship to be positive. The likely explanation is that perpetrators who entered the study early in the year had more time to reoffend. This is of little substantive interest, but it serves as a sanity check on the boosting results. It confirms the well-known importance of time at risk.

The second most important predictor is whether there had been prior domestic violence reported to the police. The relationship also is positive and not a surprise. The relationship is weak, however.
Even weaker are other predictors including whether the offender is younger than 30 years of age. The relationship is positive. When the offender is younger than 30, the chances of a repeat incident with injuries are increased.

One can certainly proceed farther in this fashion, but it is not clear what of practical use would be learned. Beyond the most important predictor, the fraction of the fitted deviance attributed to each predictor is very small and often does not materially differ from one predictor to another. One does not even have a meaningful rank ordering. The predictor importance measure by used by *gbm* (i.e., the average standardized fraction of deviance “explained”) also results in no insight into how forecasts or their accuracy are affected. Better measures are associated with other machine learning procedures (e.g., Brieman, 2001). Finally, because boosting is not a model, causal inferences are unjustified whatever the particular importance measures computed (Berk, 2018).

For this analysis, an additional problem is that the importance measures in Table 2 are produced by a fitting exercise in which it is challenging to increase classification accuracy beyond applying a Bayes classifier to the marginal distribution. Many potentially important predictors for identifying high-risk offenders may not surface. Moreover, imposing an outcome class depending on which side of .50 a risk probability falls obscures that there is range of values above .50 that could represent a variety of perpetrator types and true IPV risk. Finally, the measures of importance are not immediately responsive to one of our motivating questions: What do very-high-risk offenders have in common, and how are such attributes related?

### 6 | GENETIC ALGORITHM RESULTS

For those questions, we turn to the results from the genetic algorithm whose output must be understood in the context of genetic algorithmic machinery. For this application, the values of indicator variables are randomly altered with no regard for whether certain *combinations* of such values make subject-matter sense. The algorithmic fitness function is not designed to weed out unlikely or even impossible combinations of predictor values automatically because all that matters is the probability of violent reoffending. It has no inkling about predictor combinations of perpetrator values that are not possible in reality. The result can be “unicorns,” interesting perhaps but ultimately not real. For example, there could be IPV outcomes in which statements are taken from children in households where there are no children reported. Such potential problems are exacerbated by errors in the offense reports. For these data, however, none of the most important predictors identified seem to suffer from such problems. In principle, such potential problems could be anticipated by a genetic algorithm and precluded.

Figure 3 shows that we now have a hypothetical population of 500, *composed almost entirely of very-high-risk perpetrators*. Almost all have a risk probability of .70 or larger. We now ask the following: What do these perpetrators have in common?

We computed for each predictor the proportion of the simulated population for which that predictor was switched on (i.e., that the predictor value was equal to 1). For example, if that proportion is 1.0 for a given predictor, all members of the very-high-risk population would have that predictor switched on. Should that proportion be, say, 0.35, 35% members of the very-high-risk population would have that predictor switched on. Should that proportion be 0.0, all members of the very-high-risk population have that predictor switched off. We call these proportions “commonality importance.”

The third column in Table 2 shows the commonality importance of each predictor. Values range widely. Some have a value of 1.00, which means that for this population, they are always switched on.
These are as follows:

1. Follow-up > 3 months
2. Prior domestic violence (DV) reports
3. Offender < 30 years old
4. Victim < 30 years old
5. Offender arrested
6. Victim strangled
7. Protection from abuse order (PFA) ever

The predictors may serve as “aggravators.” When switched on, each of these variables, except arrest, increases the probability of a subsequent repeat incident with injuries reported to the police. The gbm partial dependence plot shows that after an arrest, there is a slight reduction of .04 in the probability of a repeat IPV incident with injuries. Yet all of the perpetrators in the constructed, very-high-risk population had their arrest predictor assigned a value of 1; all were characterized by an arrest at the initial IPV incident. There seems to be a contradiction.

Police in this jurisdiction are required to make an arrest for IPV incidents involving injuries (Pennsylvania Statute, Crimes and Offenses, chapter 27, section 2711, paragraph a). Thus, it follows that all of our very-high-risk population would have been arrested at the initial, reported incident. Perhaps an arrest reduced their estimated risk but not enough to remove them from the subset of very-high-risk subset offenders. We return to the role of arrest later; there is more to the story.12

Some predictors have a commonality importance value of 0.0, which means that for this population, they are always switched off. When switched on, they are potential “mitigators.” These are as follows:

1. Victim assistance contact information given to the victim13
2. Offender polite14
3. Furniture in disarray15

In short, in Table 2, potential aggravators and mitigators for the risk of very serious intimate partner violence are identified. The archetypal perpetrator who injures his victim can be characterized by these aggravators and mitigators that often were not apparent in the boosting measures of importance.
Other variables in the table can matter as well, but they are not consistently switched on or off for this synthetic population of 500.

Although the simulated population of very-high-risk offenders was generated from real data, there are no perfect matches with perpetrators in those data. Across each of the 34 predictors, a maximum of 27 predictors had values that were the same for any perpetrator in the actual data and any perpetrator from the simulated population. That should be no surprise because very few perpetrators in the data had risk probabilities as high as those for the population of 500. The synthetic population, however, is not composed of outliers. There are several members of the empirical data set with almost identical risk probabilities. We have, in effect, just fattened the right tail of the risk probability distribution.

One might wonder how many of these very-high-risk offenders would be false positives. There is no way to know because they are not included in the empirical data; there is, for the simulated population, no “ground truth.” But from those data, one can see that the number of false positives declines as the fitted risk probabilities increases beyond .50. For these very-high-risk offenders, most of the forecasted positives probably would be true positives.

7 | CLUSTERING RESULTS

In principle, a cluster analysis of the predictors should replicate the story extracted informally from Table 2 and add instructive visualizations. Predictors with proportions of 1.0 should constitute one strong cluster, and predictors with proportions of 0.0 should constitute another strong cluster. Both extreme proportions imply maximum similarity, a bit like a correlation of 1.0 and −1.0. We are addressing which predictors are strongly related to one another for the synthetic population.

Figure 4 shows the dendrogram that results when an agglomerative clustering algorithm (Kaufman & Rousseeuw, 1990, ch. 5) is applied to the predictors in the very-high-risk population. "Height" on the left margin of the figure is a standardized measure of dissimilarity within a cluster that ranges from 0 to 1, and necessarily gets larger when as one moves toward the top of the figure; the clustering begins at the bottom of the figure for the predictors that are the least dissimilar. The agglomerative coefficient at the bottom of Figure 4, which also ranges from 0.0 to 1.0, is a standardized measure of cluster distinctness. The moderate coefficient of .59 indicates that on the average, the clusters
are distinct; the between-cluster dissimilarities are usually substantially larger than the within-cluster dissimilarities.

The cluster on the lower left reproduces the “always on” predictors from the far right column in Table 2, this time using the Gower measure, not the importance commonality. Because the next predictor to join (Victim’s Clothes in Disarray) is substantially higher (i.e., \(\sim 0.25\) units), its distance from the cluster is meaningful.

The initial cluster to the right identifies predictors that are “always off.” (i.e., have a proportion value of 0.0). These variables include Contacts to Victim, Polite Offender, and Furniture in Disarray. According to their partial dependence plots, each would decrease the chances of injury when switched on. Therefore, one would not expect them to be switched on for very-high-risk offenders. Because this cluster forms at about the same height as the cluster with predictors always switched on, both clusters have about the same, small within-cluster dissimilarity. If one wanted to enlarge the cluster, the next predictor to join would be whether the offender was apologetic.

Perhaps the most important conclusion from the clustering results is that selecting as important only those predictors that are universally switched on or off is discarding other potentially relevant information. The proportions in Table 2 provide information that also could be useful. It may be worth noting that the boosting importance measures and the commonality importance measures sometimes order predictors in the same way, especially for predictors toward the very top of the table. But the predictors highlighted in red or blue can differ substantially as well.

8 | NEW APPROACH TO PREDICTOR IMPORTANCE

A routine task in the development of criminal justice risk assessment tools is to document how each risk factor affects forecasts of risk. Typically, some risk factors will alter risk forecasts more than others. Put another way, the credibility of risk forecasts depends substantially on the weight given to each predictor as risk scores are determined. We need to know how much each predictor “moves the needle.” None of our earlier results provide such information.\(^\text{18}\)

Working broadly from machine learning traditions formalized by Leo Breiman (2001), we developed new predictor importance measures implemented in the following steps:

1. Use the very-high-risk population of 500 as new data from which predicted probabilities are desired. Figure 3 provides that information visually.
2. Compute the mean risk probability to serve as a benchmark. This number is the average risk probability when all 34 predictors are set to the values determined by the genetic algorithm.
3. Construct new data sets, one for each predictor with a universal 1 or a universal 0. There are 10 such data sets, each containing the full set of 34 predictors.
4. For one of the 10 new data sets, select a predictor that is universally 1 or 0. Apply reverse coding for that predictor.\(^\text{19}\) For example, with the new data set for the predictor Prior DV Reports, recode the 1s to 0s. All other predictor values are unchanged. Repeat the reverse coding within each data set, so that each predictor has its own data set in which it alone has been reverse coded.
5. Using the boosting algorithmic structure found earlier, obtain the 500 predicted probabilities of a repeat IPV incident in which the victim is injured for each of the 10 data sets. There will be 10 sets of predicted probabilities.
6. Compute the mean of the predicted risk probabilities for each data set.
7. Compare each of these means with the mean probability when no predictors are recoded.\(^\text{20}\)
Table 3 shows the results. The mean risk probability in the population of 500 is 0.718; this is the value of the benchmark. When each predictor is reverse coded (one at a time), the mean probability will drop in value. The larger the drop, the greater the impact that predictor has on the average risk.

Clearly, the largest impact by far is for the length of the follow-up. The average risk probability drops from 0.718 to 0.431 when its values are recoded from 1.0 to 0.0. The next largest impact is far smaller. The average risk probability for Prior DV reports is 0.612 when its values are recoded from 1.0 to 0.0. Next, the mean probability risk for furniture in disarray falls to 0.642 when its values are recoded from 0.0 to 1.0. If one proceeds through the table, the smallest impact is for victims younger than 30 years of age when its values are recoded from 1.0 to 0.0. The mean risk probability for victims younger than 30 years of age is 0.712.

The reduction for the arrest variable is curious. From the boosting partial dependence plots, an arrest reduced the estimated probability of subsequent IPV resulting in injury. But here we see that when an arrest is not made, the probability of subsequent IPV resulting in injury declines. If one or both of the results are not a product of some statistical artifact or errors in the offense forms, the impact of an arrest differs for the actual, rather heterogeneous population of perpetrators compared with the constructed, homogeneous, very-high-risk population. Alternatively, the arrest variable may be part of a proxy for factors not measured in the data set. In the synthetic population, for example, an arrest may motivate perpetrators to retaliate against the victim.

Clearly, the universal predictors can reduce mean risk probabilities by meaningful amounts for this very-high-risk population. Less clear is whether differences in impact between the predictors should be taken seriously. We reran the genetic algorithm several times, and although, by and large, the same predictors were universally 1.0 or 0.0, the impact on predicted risk changed a bit, usually in the second or third decimal place. With the exception of the length of the follow-up period, this was sometimes enough to reorder the predictors in their importance. Modest rerankings of this sort are not surprising because the genetic algorithm has random processes built in.

The fitness function used by the genetic algorithm is no doubt highly dependent on the training data and on the cost ratio imposed. Any thoughts of generalizing the findings for variable importance are premature. Yet, although the performance of the boosting algorithm was somewhat disappointing, its fitness function when used in concert with the genetic algorithm led to results that are worth thinking about.

### Table 3

<table>
<thead>
<tr>
<th>Variable Recoded</th>
<th>Mean Risk Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.718</td>
</tr>
<tr>
<td>Follow-up &gt; 3 months: 1.0 to 0.0</td>
<td>0.431</td>
</tr>
<tr>
<td>Prior DV reports: 1.0 to 0.0</td>
<td>0.612</td>
</tr>
<tr>
<td>Furniture in disarray: 0.0 to 1.0</td>
<td>0.642</td>
</tr>
<tr>
<td>Contact information given to the victim: 0.0 to 1.0</td>
<td>0.662</td>
</tr>
<tr>
<td>Offender strangled: 1.0 to 0.0</td>
<td>0.664</td>
</tr>
<tr>
<td>Offender polite: 0.0 to 1.0</td>
<td>0.683</td>
</tr>
<tr>
<td>Offender &lt; 30: 1.0 to 0.0</td>
<td>0.700</td>
</tr>
<tr>
<td>Offender arrested: 1.0 to 0.0</td>
<td>0.701</td>
</tr>
<tr>
<td>Ever having a PFA: 1.0 to 0.0</td>
<td>0.709</td>
</tr>
<tr>
<td>Victim &lt; 30: 1.0 to 0.0</td>
<td>0.712</td>
</tr>
</tbody>
</table>

*Notes. DV = domestic violence; PFA = protection from abuse order.*
9 | DISCUSSION

The highly unbalanced IPV distribution we confronted would produce significant difficulties for any empirical analysis aimed at identifying attributes of very-high-risk, repeat offenders. Distributions for mass violence are likely to be far more unbalanced still, making a credible statistical analysis extremely difficult, and even impossible in many cases. But a lot depends on how the population of mass violence “opportunities” is formulated. As noted earlier, for example, one can easily define a population at risk as all high schools in the United States. What is the probability that there will be, in a particular year, mass violence at any given high school? One can as easily define the population at risk as all high school students. What is the probability for a particular year that a given high school student will be a fatal victim of mass violence at his or her high school? Even for these two closely related populations, the numerators and denominators that determine probability estimates differ, but for both, mass violence will be rare. Trying to produce credible analyses and forecasts for such populations may well be a fool’s errand.  

One can work, however, with more narrow populations in which mass violence is not especially rare. For example, in major metropolitan areas, SWAT teams are dispatched regularly. These dispatches are usually informed by extensive information about the setting, the perpetrator(s), and potential victims. People who might know the perpetrator(s) are contacted for background information (e.g., does he have firearms) that can be combined with information from administrative data readily available (e.g., past police mental health referrals). Information also is obtained about the physical setting: Is there a back door, are the windows barred, is there a fire escape, is there an escape route for the offender across neighboring roof tops, how are the rooms in the building arranged, and more. And for each dispatch, the manner in which the events unfolded is well documented, including any deaths and/or injuries. In short, there is a rich collection of predictors coupled with information on fatal outcomes that are not necessarily rare.

Closely related to the analyses in this article, SWAT teams often are dispatched to residences in which an IPV perpetrator is “holed up” with his partner, their children, and perhaps other family members. Many times the incident can be deescalated and there is no physical violence. But physical harm is difficult to prevent fully, and deaths and injuries can occur. From the data we have seen from one local SWAT team, multiple shootings and deaths associated with SWAT dispatches for IPV incidents are not especially rare. One could perhaps do a sensible analysis of mass violence from SWAT team dispatches—we are exploring that possibility. By gradually accumulating studies of mass violence for narrow populations in which mass violence is not nearly so rare, progress might be made.

In this article, we suggested a complementary path. Perhaps the major challenge to our results is the reliance on a synthetic population. How much can one learn from 500 hypothetical perpetrators? The use of synthetic data has a long, if somewhat narrow, history in statistics. Perhaps the most well-known example is for multiple imputation with missing data (1987). If one is able to construct the predictive distribution of the values for missing, prior crime victimizations, for instance, one can draw from that distribution, each draw being a synthetic, imputed value. From these, one can sometimes estimate the uncertainty introduced by imputation, although strong assumptions are required.

An example from a much different setting is the use of synthetic counterfactuals for causal inference in observation studies (Abadie, Diamond, & Hainmueller, 2010). Suppose one is trying to estimate the impact of a crime control program in several city precincts chosen precisely because they are high crime outliers. What should one use as the comparison precincts? One strategy is to compute a weighted average of several other city precincts so that for the period before the intervention, the weighted average is
similar to the experimental precincts in the response variable values and predictor values. One might think of this as synthetic matching; one manufactures a matched counterfactual.  

Finally, the importance and visibility of synthetic populations has become a salient feature of recent work in deep learning (Le, Günes, Baydin, & Wood, 2017; Patki, Wedge, & Veeramachaneni, 2016). Generative adversarial neural networks (GANs) is perhaps the best current example (Goodfellow et al., 2014). Its architecture is designed to force two neural network algorithms to compete. One algorithm, called the “generator,” constructs distorted images of, say, cats. The other algorithm, called the “discriminator,” tries to distinguish between the images of fake cats and images of real cats. The generator’s task is to do its best to fool the discriminator. The discriminator’s task is to distinguish between the fake cat images and the real cat images as accurately as it can. With each iteration, both algorithms improve their performance.

The contest ends when the discriminator can do no better than a coin flip when deciding between the fake cat images and the real cat images because they differ only by small amounts of unstructured noise. Typically, humans cannot distinguish between the two. For us, the intriguing question is what to make of the fake images of cats. They look real to the human eye.

Some would go further and call the fake images a Platonic ideal of a cat. Just as the mathematical construction of a triangle does not empirically exist, one can, nevertheless, learn much from its Platonic ideal. Perhaps some of the same reasoning can be applied to our synthetic population of 500. They may represent, at least in part, a Platonic ideal of a certain kind of very-high-risk IPV offender from which we can learn.

Our approach is more empirically based, however, than that for much of the synthetic methods briefly reviewed. First, gradient boosting applied to real training data was used to compute the fitness function. There is nothing hypothetical about how fitness was defined. Moreover, the predictors characterizing the synthetic population were the same predictors from which the algorithm was trained.

Second, as an empirical matter, the population of 500 is not composed of outliers. There are several real perpetrators with risk probabilities as high, and there is substantial overlap between the predictor values for very-high-risk perpetrators in the training data and very-high-risk perpetrators in the synthetic population. As noted earlier, we are just fattening the right tail of the empirically produced risk probability distribution.

Third, the predictors determined to be important by and large were not surprising. Although several have not been empirically studied before, except in a much different context (Small, Sorenson, & Berk, 2019), most associations with the risk probabilities were plausible and consistent with current understandings about intimate partner violence. Researchers and practitioners likely would agree that that these associations “made sense.”

Fourth, some readers may be skeptical without more information on how the predictors we found compare with those of “garden variety” IPV offenders. Such comparisons properly cannot be made, however. Recall that the results from stochastic gradient boosting could not provide such information. Indeed, that is why we moved to a genetic algorithm. At a deeper level, stochastic gradient boosting is a black box that likely contains nonlinear transformations of the input predictors coupled with high-order interaction effects. These are carried forward into our final results. Any differences or similarities in the predictors we have identified compared with findings from past research could result simply from differences in data analysis methods. Perhaps most important, we make no claims of explanation, let alone of causal inferences. At this point, the goal is forecasting alone. The relevant question is whether our predictors improve forecasting accuracy. Why they might or might not is a different question. As noted earlier, explanation requires a much different approach based on experiments or credible models.

None of these arguments should be construed as a powerful justification for our procedures nor for proceeding quickly toward policy implementation. There are many moving parts, each of which needs
to be skeptically examined and subjected to much more empirical testing. A definitive test will come when a set of predictors from some synthetic population is used to forecast real incidents of mass violence.

10 | CONCLUSION

Perhaps it helps to restate the challenge. For incidents of mass violence, base rates will be very low. Conventional forms of analysis likely will stumble; our logistic regression results are an instructive example. The choice for research and policy, therefore, is either to abandon serious statistical science, reformulate the empirical questions, or consider newer approaches to data with low base rates. It this article, we offer very tentatively the third option.

Machine learning (i.e., stochastic gradient boosting) performed far better than logistic regression, but it still fell short of forecasting accuracy that one would ordinarily require (Berk, 2018). It also was difficult to learn much about predictor importance. When the boosting results were used to construct a fitness function, a genetic algorithm produced outcomes that seemed far more instructive. Some predictors, which to our knowledge have never been evaluated in a forecasting exercise, emerged as important. Some well-established predictors surfaced as well. From the hypothetical population of 500, we also learned which of the predictors seemed to perform as aggravators and which seemed to perform as mitigators and, for each, the size of their impact on forecasted risk.

We used perpetrators of IPV incidents reported to law enforcement as our analysis units. Although these data allow for an illustrative application, our results must been seen through the lens of IPV incidents reported to the police. Reported IPV incidents have important similarities to all IPV incidents, but they can be different enough to warrant caution when trying to generalize from reported incidents to all incidents. Whether an IPV incident is reported to authorities introduces a systematic difference, even if potentially small, between the two kinds of incidents. The associations between our predictors and subsequent IPV resulting in injury that is not reported to law enforcement may differ in strength and perhaps even in direction.

Although motivated by risk assessment prediction, in this article, we stop well short of using information from the synthetic population for IPV forecasting. In addition, because there is no model of risk, no causal claims can be made. Speculation that some of the associations found may be causal is insufficient. Our intent is to show how a sequence of machine learning algorithms can extract plausible features of a rare population that would not surface from conventional statistical analysis alone. In subsequent work, these features might dramatically improve forecasting accuracy for the subset of IPV offenders who pose the greatest risk of injury to victims. In the meantime, they may serve as an instructive checklist of warning signs for IPV that results in injury and that, at the least, has substantial face validity.

Moving on to new violence applications, and perhaps especially mass violence, is surely worth trying, but we have provided nothing like a recipe. Each new application will need to be considered from the ground up. For example, the SWAT data described earlier look promising, but there are many new challenges. Perhaps the most apparent is that SWAT incidents already have an “active shooter” or an individual who plausibly might become one. The population of interest is SWAT dispatches. Is that an instructive population?

Our methodological conclusions also are offered in a highly provisional manner. As best we can tell, our approach is novel. The challenges inherent in low base rates may require extending existing data analysis tools beyond conventional practice. In this article, we propose one way that might be done. Whether the methods will port well to forecasts of mass violence is at this point unknown.
ENDNOTES

1 It might seem that statistical modeling for extreme value distributions could be used to solve the problem (Coles, 2001). From a given type of extreme value distribution, one has the ability to extrapolate to rare events in the tails. But even before getting into several difficult details, one must know at least the form of the extreme value distribution and, to forecast, how to include the role of predictors. In other words, one has a demanding model specification problem for phenomena that currently are poorly understood. One important concern is that untestable assumptions will be introduced to justify a particular specification: what some disparagingly call “assume and proceed statistics.”

2 One might think that the results of a good meta-analysis would be a solution to the low base rate problem (Spencer & Stith, 2018). But at best, the only gain would be statistical power. As discussed in this article, a low base rate can undercut the estimated contributions of all predictors because it is difficult to fit the data better than the marginal distribution of the highly unbalanced response. A forecast for the far more common outcome will be correct most of the time with no help from the predictors whatsoever. There also are a variety of technical obstacles, especially when meta-analysis is applied to observational data (Berk, 2007).

3 Greg Ridgeway is the initial and primary author of gbm. In brief, stochastic gradient boosting sequentially fits a large number of generalized linear models (here, logistic regressions). With each iteration, the cases that were poorly fit in the previous iteration are given more weight so that the fitting procedure works harder to fit them accurately. In the end, the fitted values for the full data set from each iteration are combined as a weighted average to obtain the fitted values used for interpretation and forecasting. The weights are a measure of fitting accuracy for each iteration (Ridgeway, 1999).

4 Just as for conventional logistic regression, the fitted values are proportions, which under certain assumptions are treated as probabilities. The issues can be subtle (Berk, 2018). In the interest of space, we sidestep these issues. They are peripheral to the major themes in this analysis.

5 GA was written by Luca Scrucca.

6 A machine learning classification table is often called a “confusion table.” It is a cross-tabulation of the true categorical outcome by the fitted categorical outcome.

7 With the 10-to-1 cost ratio, false positives were, as a policy matter, cheap compared with false negatives. It is then no surprise that the boosting algorithm works hard to avoid false negatives but not false positives. Indeed, it readily trades a substantial number of false positives for fewer false negatives. Should this trade-off be unacceptable to stakeholders, the cost ratio is easily changed and the algorithm reapplied.

8 The preferred cost ratio is implemented as a special form of weighting. Therefore, one might think that some form of weighting could be used to solve the low base rate problem—just give the rare events more weight in the analysis. Even if such weighting could be justified by subject matter considerations, any apparent gains could be misleading. No new information is being added. Suppose there are 50 rare events, and one gives them double the weight. This is the same as counting each rare event twice. Each rare event has an exact duplicate, and as just illustrated, one is asking for a substantial increase in false positives. In Table 1, the use of asymmetric costs was introduced to distribute the false-negative and false-positive classification errors in a way that was consistent with specific policy preferences.

9 Ideally, predictor importance would be computed in the test data, which would provide a measure of “out-of-sample importance.” In-sample measures can capitalize on overfitting (Berk, 2018).

10 In stochastic gradient boosting, additional randomness is introduced by random subsetting into training and test data. Rerunning the analysis several times from the beginning produced classification tables almost the same as Table 1, but the importance measures shuffled the order of all but the most important predictor.

11 The probability that an indicator predictor would have its value flipped (a “mutation”) was set to .10. Changing it to .05 or .25 made no important difference except for altering the number of populations needed before no further improvement was found. The probability of a crossover (“sexual” reproduction) between a random pair of perpetrators was set at .80. Dropping that value as low as .10 did not change the results in an important way, although again, the number of populations needed changed somewhat. The default crossover method was a “single point” procedure in which, for a single randomly chosen perpetrator, all predictor values for columns to the right of a randomly chosen column are swapped with the values for the same columns for another randomly chosen perpetrator (Umbarkar & Sheth, 2015, Section 2.1). By default, the fittest 5% of the cases automatically survived to the next generation with no changes. Many of the background details can be found in Scrucca (2013).
Recall that partial dependence plots are constructed for all 20,000 perpetrators in the training data, not for the 500 very-high-risk offenders.

It is difficult to know what this predictor captures. Because all of the very-high-risk population would have been arrested, the offender would have been removed from the scene. Under these circumstances, the police might have assumed that the department automatically would follow up with the victim. There was no need for other interventions that might have put the victim at least risk.

This refers to interactions with police at the scene.

When furniture is in disarray, perhaps a threat has been delivered that achieves the desired domination without physical harm to the victim.

The process begins with the construction of a dissimilarity matrix. Because the predictors are binary, the Gower method was used as a measure of similarity (Kaufman & Rousseeuw, 1990, pp. 235–236). When transformed into dissimilarities, the index becomes a measure of distance between predictors. The algorithm starts with each predictor unclustered, finds the pair that is least dissimilar, and combines them in a cluster. For each unclustered predictor in turn, a mean distance is computed between the within-cluster predictors and that predictor. In essence, the closest predictor is then taken into the cluster and the process continues. Clusters are combined using similar reasoning based on the average distance between predictors in each possible pair of clusters. To be consistent with the earlier results, we clustered variables; in principle, we could have clustered perpetrators. It would be difficult, however, to see much with 500 observations, and interpreting a predictor space of 34 dimensions with a projection on two dimensions likely would not have revealed much.

The details are beyond the scope of this discussion. See Kaufman and Rousseeuw (1990, pp. 211–212).

Our concern here is with the effects on the risk forecasts, not on actual offending. We are seeking information on what alters the fitted values.

If the predictor in question is universally a 1, recode it to 0. If the predictor in question is universally a 0, recode it to 1. There is no formal need to limit this process to predictors that are, for all members of the population, either equal 1 or 0. But these are the predictors that have the most promise of substantial importance because all 500 values will be recoded. If, for a given predictor, the proportion of 1s is, say, only .60, only 60% of the values will be recoded.

One could accomplish the same thing with single data set (not 10) by reverse coding a given “universal predictor,” obtaining the desired estimate, reversing the reverse coding, and repeating the process for each universal predictor one at a time. Here, it is conceptually and operationally more direct to work with one data set for each universal predictor.

These measures of importance isolate the impact of the given predictor with all else held constant in the sense that none of the other predictors are recoded. This is in the spirit of an experiment; only the “treatment” is varied.

The very few times one of the universal predictors had less than all 1s or all 0s, the proportion of 1s or 0s was still very large. And on the rare occasion when a new universal predictor surfaced, it earlier had a large proportion of 1s or 0s.

The same issues arise for mass violence in business establishments and government buildings. “The reality is there are millions of people who go to work every day in government buildings and millions of people who go into government buildings every day. And almost every single one of them goes home safely” (Sorenson, 2019, para. 15).

These incidents can have bizarre and even darkly humorous features. One incident of which we are aware was de-escalated once a SWAT team member complied with a request from the perpetrator to sing him a Christmas carol. A surrender followed with no physical violence.

McClelland and Gault (2017) provide an accessible discussion along with the most important assumptions.

This is the basis for current and real concerns about “deep fake” media in which “artificial intelligence” creates synthetic images that appear to be real. For a recent summary of the issues, see “New Video Editing Technology Raises Disinformation Worries” from The Washington Post (Zakrzewski & Riley, 2019). See also “Top AI Researchers Race to Detect ‘Deepfake’ Videos: ‘We are outgunned’ ” (Harwell, 2019).

The notion that the fake cat images might be seen as a Platonic ideal was suggested to us by Michael Kearns and Aaron Roth. Both are real people but close to a Platonic ideal of a computer scientist.
REFERENCES


Wilson, C. (2018, February 22). This chart shows the number of school shooting victims since Sandy Hook. Time Magazine.


AUTHOR BIOGRAPHIES

Richard A. Berk is a professor of criminology and statistics at the University of Pennsylvania. He works on foundational issues in statistics, machine learning, and on applications to criminal justice.

Susan B. Sorenson is a professor of social policy and a senior fellow in public health at the University of Pennsylvania. Her work focuses on the epidemiology and prevention of violence.

How to cite this article: Berk RA, Sorenson SB. Algorithmic approach to forecasting rare violent events: An illustration based in intimate partner violence perpetration. Criminol Public Policy. 2019; 1–21. https://doi.org/10.1111/1745-9133.12476