The Pitfalls of Prediction

by Greg Ridgeway

The criminal justice system should take advantage of the latest scientific developments to make reliable predictions.

Prediction is common in everyday life. We make predictions about the length of our morning commute, the direction of the stock market, and the outcomes of sporting events. Most of these common-sense predictions rely on cognitive shortcuts — or heuristics — that shape our expectations of what is likely to occur in the future. But these heuristics are not necessarily true; they rely on cognition, memory and sensory impressions rather than a balanced analysis of facts. Consequently, they can result in biased predictions.

The challenge of predicting the future has always been at the heart of the criminal justice system. Judges weigh the risks of releasing offenders to probation, police agencies try to anticipate where officers should be deployed to prevent future crime, and victims wrestle with the uncertain odds of being revictimized.

There is a long history of research on prediction in criminology and criminal justice, and two developments are helping the criminal justice system improve its ability to make reliable, scientific predictions. First, more and more jurisdictions are accumulating rich data and are getting better at linking across their data sources. Second, a growing set of sophisticated analytic prediction tools is available to help agencies make decisions about future events, unknown risks and likely outcomes.

Practitioners can now combine expert assessment with data-driven prediction models to discern how much risk a probationer poses,

Editor's Note: *This article was presented to seven law enforcement agencies that were developing predictive policing programs.*

determine whether a pair of illicit drug transactions signals the emergence of a drug market, or project whether crime will increase or decrease during the next month. More and more, police departments are using forecasting tools as a basis for formal predictive policing efforts; these statistical prediction methods inform their prevention strategies so they can anticipate rather than react to crime.[1] (See sidebar, "NIJ's Role in Predictive Policing," on this page.)

Although the science of prediction continues to improve, the work of making predictions in criminal justice is plagued by persistent shortcomings. Some stem from unfamiliarity with scientific strategies or an over-reliance on timeworn — but unreliable — prediction habits. If prediction in criminal justice is to take full advantage of the strength of these new tools, practitioners, analysts, researchers and others must avoid some commonplace mistakes and pitfalls in how they make predictions.

Pitfall #1: Trusting Expert Prediction Too Much

Using data and computers to predict or help experts predict shows promise, but the pace of adoption has not matched that promise. Why? Perhaps we trust ourselves more than we trust machines.

For example, more than 30 years ago, Stanford scientists developed a pathbreaking, computer-based medical expert system that could synthesize patient features and therapeutic options.[2] The system, called MYCIN, outperformed practitioners in selecting the right antibiotic treatments. Despite MYCIN's demonstrated success and similar kinds of computer-based prediction successes, we still do not see these systems being used in our doctors' offices. Some researchers have found that physicians have "a high regard for their own decision-making ability and are afraid of any competition from computers."[3]

So how do experts and machines compare in their ability to predict in the justice system?

Consider this example: A panel of 83 experts — law professors, deans of law schools and others who had practiced before or clerked at the U.S. Supreme Court — set out to predict how the U.S. Supreme Court would vote on the 68 upcoming cases on the 2002 docket. Based on their knowledge of the justices and the ins and outs of the court, they correctly predicted how the Supreme Court would vote on 59 percent of the cases.

Researchers used a computer program to make the same prediction. The computer analyzed 628 previous Supreme Court cases and generated data-derived rules.[4] The researchers created a decision-tree prediction model based on a simple set of these rules.

NIJ's Role in Predictive Policing

Law enforcement work is frequently reactive: Officers respond to calls for service, control disturbances and make arrests. But law enforcement work is becoming increasingly proactive: Departments combine data with street intelligence and crime analysis to understand why a problem arises and predict what might happen next if they take certain actions.

NIJ is supporting predictive policing efforts in a number of ways:

- **Predictive policing symposiums.** NIJ convened two symposiums at which researchers, practitioners and law enforcement leaders developed and discussed the concept of predictive policing and its impact on crime and justice. Read summaries of both sessions at http://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/symposium/welcome.htm.
- **Predictive policing grants.** The Chicago and Shreveport police departments are using grants to explore data-driven policing strategies. In Phase 1, they received funding to identify a problem and develop predictive policing strategies to solve it. In Phase 2, they were awarded additional funding to implement and evaluate the strategies. For more on these grants, see http://www.nij.gov/nij/topics/law-enforcement/strategies/predictive-policing/symposium/discussion-demonstrations.htm.

For more information:

- To learn more about predictive policing in general, read the *NIJ Journal* article "Predictive Policing: The Future of Law Enforcement?" at http://www.nij.gov/journals/266/predictive.htm.

Figure 1. Decision Tree for Supreme Court Justice Sandra Day O'Connor

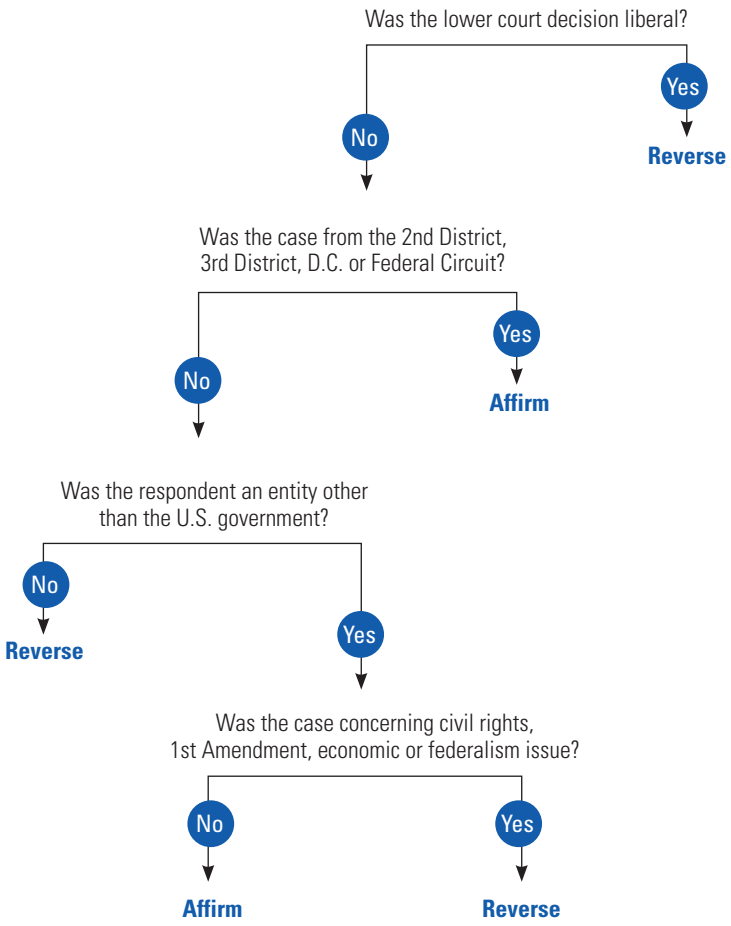


Figure 1 shows the decision tree for Justice Sandra Day O'Connor. Based on a simple set of rules — such as whether the lower court decision was liberal — the model was able to predict how Justice O'Connor would decide 70 percent of the cases in 2002. Using similar decision trees for the other eight justices,[5] the model correctly predicted the majority opinion in 75 percent of the cases, substantially outperforming the experts' 59 percent. The experts lost out to a machine that had a few basic facts about the cases.

So what can we take away from this example? It should lead us to question — but not necessarily dismiss — the predictions of experts, including ourselves. Of course, not all cases afford us the data to build predictive models. But if we have data that we can use to construct predictive models, then we should build the models and test them even if our expert detectives, probation officers and others in the field indicate that they already know how to predict. They may be as surprised as the expert panel was in the Supreme Court example.

Pitfall #2: Clinging to What You Learned in Statistics 101

If your knowledge of prediction is limited to what gets covered in introductory statistics courses, you are probably unfamiliar with the prediction model used above. Instead, you most likely learned how to check model assumptions and carefully test hypotheses. But when it comes to prediction, the rules are different and rather simple: Are the predictions accurate, and can you get them when you need them? You can judge the quality of a specific prediction model by considering the following:

Performance criteria. Do the model's goals and constraints match the intended use? Methods that are good at predicting, for example, whether an injury will result from a mission are not necessarily the same as those that are good at predicting the number of days an officer will be out with that injury. If you are planning a tactical unit's staffing, it is important for you to know the expected person-hours that will be lost to injuries. Thus, using a model that can accurately predict only whether an injury will occur — and not how long an officer will be out — would be insufficient.

Accuracy. Can the model make accurate forecasts? More specifically, the implemented model should be better at prediction than the agency's current practice. For example, if cops are allocating resources to neighborhoods where they think crime will spike, then going forward we should test whether the prediction model is better at selecting those neighborhoods. If a probation officer is assigning remote monitoring anklets to DUI probationers, then we should test whether the prediction model is better at picking which DUI probationers will reoffend in the next six

months. For a prediction model to be useful, it must outperform practice as usual.

Computation time. Can you apply the prediction model in a reasonable amount of time? Some models can be computationally intensive to run and use. There is little point in using a model that cannot produce predictions in time for them to be useful.

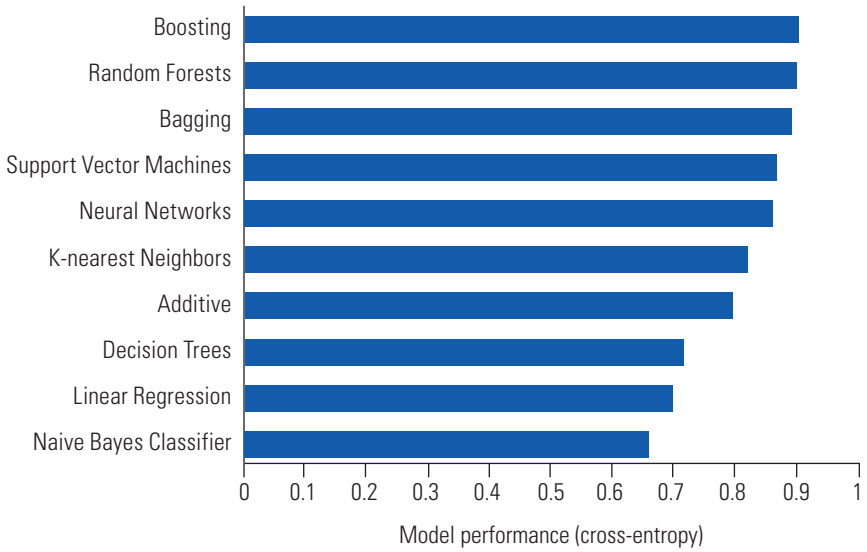
Handling mixed data types. Can the prediction model manage and properly interpret numbers, dates and times, geography, text, and missing values — which datasets almost always have?

Interpretability. Can a person understand why the prediction model makes the predictions it does? We would prefer to be able to understand the reasoning behind a prediction. However, if getting transparency requires using a model that is *less* accurate in predicting, say, when and where a gang retaliation shooting will take place, then a more transparent model might not be worth the cost. This issue will be discussed further under Pitfall #5.

Pitfall #3: Assuming One Method Works Best for All Problems

In 2006, researchers examined how the most commonly used prediction methods performed head-to-head.[6] They looked at 11 datasets covering a variety of prediction tasks and measured each method's accuracy. The researchers found that the more modern methods of boosting and random forests consistently performed best, whereas linear regression — well over 70 years old and by far the most widely used method — did not fare well. (See Figure 2.) Note that decision trees, the method used in the Supreme Court example, is also near the bottom of the list, suggesting that even better accuracy in predicting case outcomes is possible. The University of Pennsylvania team working with Philadelphia's Adult Probation and Parole Department to predict probationers at high risk of violent crime opted for random forests. (See "Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise" on page 4.)

Figure 2. Comparison of 10 Widely Used Prediction Methods



For more information on these prediction models, see Caruana, Rich, and Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics," in *Proceedings of the 23rd International Conference on Machine Learning,* New York: Association for Computing Machinery, 2006: 161-168.

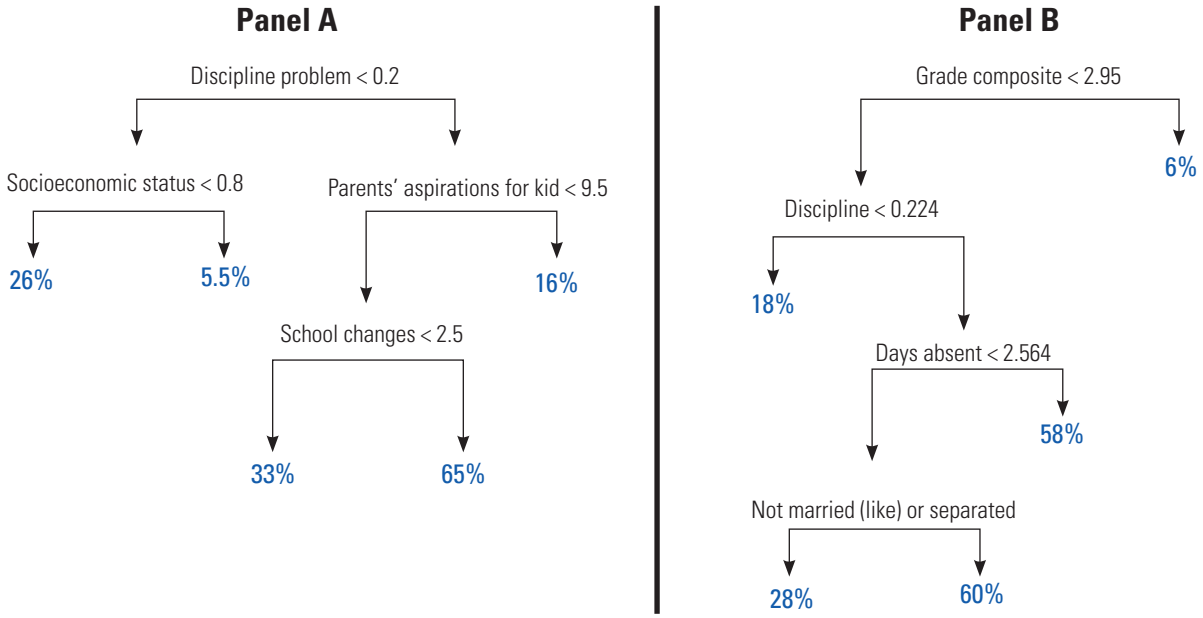
Prediction can play a major role in the criminal justice system. Even small improvements in where police are assigned, which cold cases receive more attention, or which probationers receive more intense supervision can result in performance and efficiency gains.

However, the researchers who compared these prediction methods also found that the best-performing method for any particular dataset varied. This means that analysts cannot fall in love with a single model — depending on the particular prediction problem, their preferred method might not be the best fit.

Pitfall #4: Trying to Interpret Too Much

Practitioners tend to favor decision-tree models like the one used in the Supreme Court example because

Figure 3. Example of Decision-Tree Models Fit to Two Samples from NELS88 Dataset



they offer transparency. One can, after all, trace the pathways through the tree. And the Justice O'Connor tree, based on a set of simple rules, provides a compact, easy-to-follow story.

But not all trees are as straightforward — they can have many branches, the path may not be easy to follow, and the rules can be quite sensitive to small changes in the dataset.

For example, we can create a decision-tree model to predict student dropout risk among 16,000 students in the 1988 National Education Longitudinal Study (NELS88). If we randomly split the data on students into two halves, each with 8,000 students, and fit decision-tree models predicting dropout risk to each half, the resulting trees will look like those in Figure 3.

We arrive at very different interpretations about the reasons behind student dropout. Looking at the first tree (Panel A), we might conclude that discipline problems are the most important factor. When we look at the second tree (Panel B), it seems that grades are most important. Incidentally, the two decision trees had identical predictive accuracy.

The lesson is this: Although it is tempting to try to interpret results, the tree's structure is actually quite unstable. Instead, users should focus on the accuracy of the predictions. In some ways, this is analogous to using a watch — you expect it to give you the time accurately even if you do not completely understand how it works.

Pitfall #5: Forsaking Model Simplicity for Predictive Strength — or Vice Versa

Earlier, we noted that we would prefer to have a more interpretable model than a less interpretable one. Unfortunately, there is often a tradeoff, with more interpretability coming at the expense of more predictive capacity. But it is crucial that predictive models are designed for those who are going to use them, and in some cases, being able to interpret results is more important than achieving greater predictive capability.

Take, for instance, the Los Angeles Police Department's (LAPD's) effort to identify new recruits.[7] The LAPD did not know why some candidates made it through the recruiting process and became officers and others did not — and thus, it did not know whether it was using its resources efficiently.

To help the LAPD predict which recruits had a better chance of becoming officers, researchers developed a priority score based on a few easily collected facts about each candidate. The score rated how likely that candidate was to join the department. Recruiters could then usher these viable candidates through the process more quickly.

Looking at LAPD data on former recruits, the researchers found that three factors were critically important:

- Whether the candidates had "issues" identified in a preliminary background questionnaire that could disqualify them from service (e.g., criminal, financial, driving and drug history)
- Level of education
- Where they lived

They developed the point system in Table 1.

Table 1. Priority Score Point System

Preliminary Background	Does not qualify because has too many issues	Potentially qualifies, has some issues	Qualifies with no issues
	+0	+13	+22
Education	Has GED	Has high school	Has some college
	+0	+4	+8
Residence	Lives outside California	Lives in California	Lives in Los Angeles County
	+0	+2	+5

Under the system, if candidates have too many background issues, they do not qualify for service and receive 0 points. Most candidates have some issues but not enough to disqualify them, and they receive 13 points. Some have no issues and receive 22 points. The other two factors, education and residence, follow a similar point structure. Most candidates have high school diplomas (4 points), and most live in Los Angeles County (5 points). A candidate's priority score is the sum of the points for their preliminary background, education and residence.

The model predicts that a candidate who has a total of 22 points — for example, a recruit with some issues, a high school degree and residency in Los Angeles County — has a 20 percent chance of joining the LAPD. Candidates who have no issues and some college and live in Los Angeles County score the maximum 35 points; according to the model, they have a 43 percent chance of joining the department. By separating these highly viable candidates from the rest of the recruits, this system allows the LAPD to prioritize candidates and more efficiently allocate its recruiting resources. And because the model is simple to understand and simple to implement, the LAPD recruiting team used it.

This simplicity gets at the important issue: A decent transparent model that is actually used will outperform a sophisticated system that predicts better but sits on a shelf. If the researchers had created a model that predicted well but was more complicated, the LAPD likely would have ignored it, thus defeating the whole purpose.

Pitfall #6: Expecting Perfect Predictions

When using prediction models, managing expectations and focusing on the big picture are critical. Predictions will not be perfect, but the ultimate goal is to improve *overall* efficiency.

In the LAPD example, a highly viable candidate has a 43 percent chance of joining the force. This means that 57 percent of highly viable recruits drop out. Invariably, a candidate given a high viability score will fail miserably in the process. Because of this, some will say that doing business the old way is a better strategy than using the prediction model. But a handful of misclassified candidates should not overshadow the gains made in recruiting efficiency.

Predictive policing offers another example. It holds the promise of anticipating where crimes will occur so that police can prevent those crimes. However, prevention activities prompted by prediction models are poised to disappoint some. Consider a model that anticipates the time and place of the next retaliatory gang shooting almost perfectly, coupled with a model that directs officers to the right place at the right time. In such an ideal situation, the predicted shooting will never materialize. Naturally, those in the field will question why they were deployed to this place and not to another place with more pressing problems.

Pitfall #7: Failing to Consider the Unintended Consequences of Predictions

Prediction models can have unintended consequences that must be anticipated. Take, for instance, the LAPD recruiting example. The goal of the prediction model was to help the department improve its recruiting process. However, the LAPD is under a 30-year-old court order to meet diversity targets, such as having women make up 25 percent of new recruits. The priority point

system could undermine the LAPD's ability to comply with the court order if, for example, prioritizing those with some college reduced the number of minority recruits.

The researchers who developed the point system considered this unintended consequence and noted that a small change could not only avoid the problem but also actually help the LAPD achieve compliance. They determined that if female applicants received an additional 7 points, then the system would be tuned so the department would meet its goal for recruiting women and its racial diversity goals because minority candidates were more likely to be women. Although this change reduces predictive accuracy because it places priority on some candidates with a lower chance of joining, it optimizes recruiting resources subject to the department's diversity goals.

The Power of Prediction

Prediction can play a major role in the criminal justice system. Even small improvements in where police are assigned, which cold cases receive more attention, or which probationers receive more intense supervision can result in performance and efficiency gains.

However, if the criminal justice system is going to reap such gains by using prediction models, it must seek to avoid the pitfalls that are so often a part of prediction.

About the author: Greg Ridgeway is acting director of NIJ.

NCJ 240702

Watch Greg Ridgeway talk about predictive policing at the 2010 NIJ Conference: http://nij.ncjrs.gov/multimedia/video-nijconf2010-ridgeway.htm.

Notes

1. Bratton, William J., and Sean W. Malinowski, "Police Performance Management in Practice: Taking COMPSTAT to the Next Level," *Policing* 2(3) (2008): 259-265.
2. Yu, Victor L., Lawrence M. Fagan, Sharon M. Wraith, William J. Clancey, A. Carlisle Scott, John Hannigan, Robert L. Blum, Bruce G. Buchanan, and Stanley N. Cohen, "Antimicrobial Selection by a Computer: A Blinded Evaluation by Infectious Disease Experts," *Journal of the American Medical Association* 242(12) (1979): 1279-1282.
3. Engle, Jr., Ralph L., and Betty J. Flehinger, "Why Expert Systems for Medical Diagnosis Are Not Being Generally Used: A Valedictory Opinion," *Bulletin of the New York Academy of Medicine* 63(2) (1987): 193-198. Heathfield is more optimistic but echoes that "theoretical and technical limitations are not the major barriers to the successful implementation … but rather more complex professional and organizational issues are at stake." Heathfield, Heather, "The Rise and 'Fall' of Expert Systems in Medicine," *Expert Systems* 16(3) (1999): 183-188.
4. Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn, "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking," *Columbia Law Review* 104 (2002): 1150-1210.
5. The scores for all the justices were as follows: O'Connor: 70 percent; Ginsburg: 55 percent; Breyer: 57 percent; Stevens: 61 percent; Souter: 61 percent; Kennedy: 71 percent; Scalia: 73 percent; Rehnquist: 76 percent; and Thomas: 75 percent.
6. Caruana, Rich, and Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics," in *Proceedings of the 23rd International Conference on Machine Learning,* New York: Association for Computing Machinery, 2006: 161-168.
7. Lim, Nelson, Carl F. Matthies, Greg Ridgeway, and Brian Gifford, *To Protect and to Serve: Enhancing the Efficiency of LAPD Recruiting.* Santa Monica, Calif.: RAND Corporation, 2009.