UNIVERSITY of PENNSYLVANIA

Department of Criminology

Working Paper No. 2016-5.0

A Primer on Fairness in Criminal Justice Risk Assessments

Richard Berk

A Primer On Fairness in Criminal Justice Risk Assessments[1]

Richard Berk
Department of Criminology
Department of Statistics
University of Pennsylvania

Introduction

There are widespread concerns about fairness when actuarial risk assessments are used to inform criminal justice decisions (Harcourt, 2008; Tonrey, 2014; Starr, 2014; Berk and Hyatt, 2015, Crawford, 2016). Some such concerns are driven by ideology in which facts do not matter. The only response may be to point out alternative and legitimate ideological positions leading to different conclusions. Some concerns result from invidious comparisons to ideal risk assessments whereas the proper benchmark is current practice, typically informal judgments from criminal justice decision-makers. Some concerns derive from principled objections to actuarial methods, although risks determined by decision-maker judgment are implicitly, but no less, actuarial. There also can be jurisprudential issues, although these too seem to overlook that informal judgment can be questioned on the very same grounds. Finally, some concerns fail to consider the tradeoffs between different features of risk assessments. In particular, there can be an inevitable need for risk assessment tools to balance different kinds of fairness as well as fairness against forecasting accuracy.

The goal of this primer is to help clarify the meaning of fairness when risk assessment tools are evaluated. Even if the concerns just listed are effectively addressed, there may still be disputes because of misunderstandings about what kind of fairness is at stake. Confusion tables will be used as a didactic device.

Confusion Table Measures of Performance

Confusions tables are a common output from machine learning classifiers and an excellent way to represent how any classifier performs (e.g., random forests, logistic regression, discriminant function analysis). A confusion table is nothing more than a cross-tabulation of actual response classes against response classes predicted when a fitting procedure is applied to data. For example, the response classes might be failing on parole or not. A confusion table would show the numerical results when the actual parolee outcomes are cross-tabulated against the predicted parolee outcomes. There can be more than two response classes such as an arrest for a violent crime, and arrest for a nonviolent crime, or no arrest of any kind. This often is very desirable. But, for simplicity, only two response classes will be discussed. The conceptual issues are much the same regardless of the number of response classes.

[1] Very helpful comments were provided by Geoffrey Barnes, Aaron Chaflin, John MacDonald, Sandy Mayson, and Emily Owens.

Table 1 : An Idealized Confusion

	Failure Predicted	*Success Predicted*	*Model Error*
Failure - A Positive	*a (true positives)*	*b (false negatives)*	*b/(a+c)*
Success - A Negative	*c (false positives)*	*d (true negatives)*	*c/(c+d)*
Use Error	*c/(a+c)*	*b/(b+d)*	*Overall Error =* *(a+b)/(a+b+c+d)*

Table 1 shows an idealized confusion table. “Success” and “Failure” are the two classes for the response variable. The observed response class is shown on the left margin of the table. The predicted response class is shown on the top margin of the table. Each letter in an internal cell of the table is a cell count. The letter *a* is the number of observations in the upper-left cell. The letters in the other three internal cells have the same meaning. All of the observations in a particular cell are characterized by an observed class and a predicted class. For example, *a* is the number of observations for which the observed response class is a failure, and the predicted response class is a failure.

When the observations are from training data, “predicted” means “assigned,” much as for fitted classes in logistic regression. Training data contain the observations used in the fitting process. When the observations are from test data, “predicted” means “forecasted.” Test data are not used in the fitting process, but are employed to obtain an honest, out-of-sample, assessment of fitting performance.

There are generally five kinds of performance assessments that legitimately can be made from confusion tables.

1. The proportion of cases *incorrectly* classified overall is a popular way to assess performance quality. It is nothing more than the number of observations in the off-diagonal cells divided by the total number of observations (i.e., *(b+c)/(a+b+c+d)*). Should all of the observations fall along the main diagonal its value is 0.0. Should no cases fall along the main diagonal its value is 1.0. Ideally, the overall proportion misclassified should be the same for each suspect group (e.g., black parolees v. white parolees). By this performance measure, the suspect groups are treated identically.

A small proportion for overall error is desirable, but it must be compared to the baseline for fitting skill when no predictors are used. Sometimes, even a low overall error rate is larger than the overall error rate when no predictors are employed. For example, suppose that the marginal proportion of individuals on parole who are arrested is .70, and the marginal proportion of individuals on parole who are not arrested is .30. By the Bayes classifier, one should always predict an arrest. Then, the overall error is .30. Now suppose that a confusion table has an overall error proportion of .35. By this measure, the predictors don't help.

2. The overall error rate neglects that it will often be more important to accurately classify one response category than another. For example, in a medical setting, failing to diagnose a life-threatening illness will usually be seen as more costly than failing to diagnose good health. The row proportions shown in the far right-hand column are now in play. One conditions on the actual response class. For each such class, the row proportion is the number of observations incorrectly classified divided by the total number of observations of that class (i.e., *b/(a+b)* and *c/(c+d)*).

Each row proportion characterizes errors made by the fitting procedure and can be called "model error." When the true response class is known (e.g., succeeded on parole), what proportion of the time will the fitting procedure fail to correctly identify it? Ideally, misclassifications are relatively few, using as the benchmark performance with no predictors. Also ideally, the model error is the same for each suspect group. That is, the two proportions can differ from one another, but not across the suspect groups.

The two kinds of model misclassifications are commonly called false positives and false negatives. Here, failures incorrectly classified as successes are false negatives. These are individuals who failed on parole but were not correctly identified as such by the fitting procedure. Successes incorrectly classified as failures are false positives. These are individuals who succeeded on parole but were not correctly identified as such by the fitting procedure. It may seem a little odd, but this language is common in many applications where a "success" is what stakeholders are especially concerned about, whether it is a good thing or a bad thing. For example, if a diagnostic goal is to correctly detect an existing malignant tumor, finding that tumor is a true positive and failing to detect that tumor is a false negative. Still, the use of the class labels success and failure is formally arbitrary, so which off-diagonal cells contain false positives or false negatives is formally arbitrary as well. What is called a success in one study may be called a failure in another study. This is just a labeling issue, not a data analysis issue.

3. The column proportions address a different question. For each column, one conditions on the fitted class and computes the proportion of times the fitted class is incorrect (i.e., *c/(a+c)* and *b/(b+d)*). Whereas the row proportions help evaluate how well the fitting procedure performs, the column proportions capture how probative the procedure would be if used to make decisions; "use error" conveys what would happen if a practitioner uses the procedure's results to forecast. Use error will typically differ from model error, and just as for model error, error in use will typically differ depending on the response class. It will usually be possible to forecast one response class better than the other. Again, the errors should be relatively few using predictor-free performance as a benchmark. Ideally, use error should be the same for each suspect group.

4. The ratio of the number of false negatives to the number of false positives (or the inverse) shows how the fitting procedure is trading one kind of error for the other. If *c* is 5 times larger than *b*, there are five false positives for every false negative. This means that false *negatives* are taken to be five times more important than false positives; one false negative is "worth" five

false positives. Ideally, the ratio of false negatives to false positives should be the same each suspect group.

Interpretative Complications

There are factors not shown explicitly in a confusion table that can dramatically affect what a confusion table conveys. In particular, marginal distributions of key variables can cascade through a confusion table. This is important to consider when discussions of fairness are undertaken. For example, suppose a fitting procedure like logistic regression is equally accurate classifying men and women with respect to whether they fail on parole. Model error is the same for male and female parolees because for both, the fitting procedure gets failures wrong, say, 15% of the time and successes wrong, say, 20% of the time. Some would argue that, consequently, one has a fair classification procedure because it is equally accurate for male and female parolees. But, suppose there are more men than women on parole. All of the cell counts will be larger for males than for females, and there will be more false negatives (i.e., *b*) and false positives (i.e., *c*) for males than for females.

The number of false positives can be a salient fairness issue when they lead to sanctions that are inappropriate. For example, an individual forecasted to fail on parole, who would have actually succeeded (i.e., *c*), might be pointlessly denied parole at substantial cost to the state, the individual, and the individual's family. Consequently, some would argue that gender differences in the number of false positives make the procedure unfair even though the gender disparity results *solely from there being more men than women on parole to begin with.* There is no unfairness in the fitting procedure. Some of the debates in the media have been confused on this point, although the focus has been on race not gender.[2]

Now, instead suppose that men are more likely to fail on parole than women. Even if the number of men and women on parole is the same, the cell counts *a* and *b* will be larger for men than women, and the cell counts *c* and *d* will be smaller for men than for women. Consequently, even if classifications accuracy is the same for men and women, there will be more false negatives and fewer false positives for men. Moreover, the cost ratio of *c/b* will differ as well. Some debates in the media have been confused on these points too, although again, the focus is on race not gender. In short, if even a classification procedure is equally accurate for men and women, which for some defines a fair classification procedure, different marginal distributions related to the suspect classes can lead to different performance consequences.

Some Definitions are Fairness

Conceptually, there can be more to fairness than equal classification accuracy or equal forecasting accuracy. For example, one might ask a fitting procedure to *compensate* for the overrepresentation of males among those parolees who fail. This allows one to propose six

[2] A lively example is the debates over the use of the COMPAS recidivism instrument. A web search using “ProPublica risk assessment” and “Abe Gong risk assessment” will turn up lots of hits.

definitions of fairness that follow directly from the earlier discussion of confusion table performance measures.

1. “Prediction fairness” is achieved when the marginal distributions of the predicted classes are the same over two or more suspect groups (e.g., men v. women). Thus, *(a+c)/(a+b+c+d)* and *(b+d)/(a+b+c+d)*, although typically different from one another, should each be the same over suspect groups. For example, the proportion of inmates forecasted to fail on parole should be the same for male and female parolees.

2. “Overall fairness” is achieved when total classification error is the same over two or more suspect groups. That is, *(b+c)/(a+b+c+d)* should be the same. This measure assumes that a false negative and a false positive are equally costly. In many settings, the costs are unequal, and a cost-weighted approach is required.

3. “Model fairness” is achieved when model error is the same over two or more suspect groups. That is, *b/(a+b)* is the same over each suspect group, and *c/(c+d)* is the same over each suspect group. We applied this definition above.

4. “Use fairness” is achieved when use error is the same over two or more suspect groups. That is, *c/(a+c)* is the same over each suspect group, and *b/(b+d)* is the same over each suspect group

5. “Cost ratio fairness” is achieved when the cost ratios (i.e., *c/b* or equivalently, *b/c*) are the same over two or more suspect groups.

6. “Total fairness” is achieved when (1) prediction fairness, (2) overall fairness, (3) model fairness, (4) use fairness, and (5) the cost ratio fairness are all achieved.

All six definitions of fairness are in practice related to one another, which will often mean that one kind of fairness will traded off against another kind of fairness. For example, cost ratio fairness (#5) can mean that model fairness (#3) will not be achieved. Then, stakeholders will need to decide how to balance one kind of fairness against another, and different stakeholders can have different views the will need to be reconciled or compromised.

Each of the definitions of fairness apply when there are more than two response categories. However, there are more statistical summaries that need to be reviewed. For example, when there are three response classes, there are three cost ratios to be examined.

Conclusions

Until the various parties expressing strong opinions about the merits of criminal justice risk assessments clarify what they mean by fairness, no progress can possibly be made. At a more fundamental level, the possible tradeoffs between different kinds of fairness need to be explored in part to clarify which concerns are about values and which concerns are about the data and statistical methods used. Finally, there are also important tradeoffs between fairness and forecasting accuracy. The tradeoffs can be quite technical and are currently being studied. But it is likely that most definitions of fairness will require a loss of forecasting accuracy so that more mistakes will be made. These mistakes, however, will be fairly distributed over the different suspect groups. Members of both groups will be equally worse off.

References

Berk, R.A., & Hyatt, J. (2015) Machine learning forecasts of risk to inform sentencing decisions. The Federal Sentencing Reporter, 27(4): 222-228.

Crawford, K. (2016) Artificial intelligence's white guy problem. New York Times, Sunday Review, June 25.

Harcourt, B. (2008) Against prediction: profiling, policing, and punishing in an actuarial age. (Chicago: University of Chicago Press).

Starr, S.B. (2014) Evidence-Based sentencing and the scientific rationalization of discrimination. Stanford Law Review, 66: 803-872.

Tonrey, M. (2014) Legal and ethical issues in the prediction of recidivism. Federal Sentencing Reporter, 26(3): 167-176.