Simple Rules for Complex Decisions

Jongbin Jung
Stanford University
jongbin@stanford.edu

Connor Concannon
John Jay College of Criminal Justice
cconcannon@jjay.cuny.edu

Ravi Shroff
New York University
ravi.shroff@nyu.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

Daniel G. Goldstein
Microsoft Research
dgg@microsoft.com

ABSTRACT

From doctors diagnosing patients to judges setting bail, experts often base their decisions on experience and intuition rather than on statistical models. While understandable, relying on intuition over models has often been found to result in inferior outcomes. Here we present a new method—*select-regress-and-round*—for constructing simple rules that perform well for complex decisions. These rules take the form of a weighted checklist, can be applied mentally, and nonetheless rival the performance of modern machine learning algorithms. Our method for creating these rules is itself simple, and can be carried out by practitioners with basic statistics knowledge. We demonstrate this technique with a detailed case study of judicial decisions to release or detain defendants while they await trial. In this application, as in many policy settings, the effects of proposed decision rules cannot be directly observed from historical data: if a rule recommends releasing a defendant that the judge in reality detained, we do not observe what would have happened under the proposed action. We address this key counterfactual estimation problem by drawing on tools from causal inference. We find that simple rules significantly outperform judges and are on par with decisions derived from random forests trained on all available features. Generalizing to 22 varied decision-making domains, we find this basic result replicates. We conclude with an analytical framework that helps explain why these simple decision rules perform as well as they do.

KEYWORDS

Interpretable models, policy evaluation, causal inference

1 INTRODUCTION

In decision-making scenarios, experts often choose a course of action based on experience and intuition rather than on statistical analysis [10]. This includes doctors classifying patients based on their symptoms [24], judges setting bail amounts [6] and making parole decisions [3], and managers determining which customers to target [34]. A large body of work shows that intuitive judgments are generally inferior to those based on statistical models [4, 5, 19, 20, 32]. However, decision makers have consistently eschewed formal decision models in part because it has been difficult to create, understand, and apply them.

Here we present a simple method for constructing simple decision rules that often perform on par with traditional machine learning algorithms. Our *select-regress-and-round* strategy results in rules that are fast, frugal, and clear: fast in that decisions can be made quickly in one's mind, without the aid of a computing device; frugal in that they require only limited information to reach a decision; and clear in that they expose the grounds on which classifications are made. Decision rules satisfying these criteria have many benefits. For instance, rules that can be applied quickly and mentally are likely to be adopted and used persistently. In medicine, frugal rules require fewer tests, which saves time, money, and, in the case of triage situations, lives [23]. The clarity of simple rules engenders trust from users, providing insight into how systems work and exposing where models may be improved [11, 31]. Clarity can even become a legal requirement when society demands to know how algorithmic decisions are being made [2, 13].

Our results add to a growing literature on *interpretable machine learning* [17, 18, 21, 22, 33]. Several methods recently have been introduced to construct the kind of simple decision rules we discuss here, including supersparse linear integer models (SLIM) [33], Bayesian rule lists [22], and interpretable decision sets [21]. These methods all produce rules that are easy to interpret and to apply. One important difference between our approach and past techniques is that our rules are also easy to create.

To illustrate our method, we begin with a case study of judicial decisions for pretrial release. We show that simple rules improve upon the efficiency and equity of unaided decisions while rivaling the accuracy of a random forest model trained on all available data. We further evaluate the efficacy of our method on 22 datasets from the UCI ML repository and show that in many cases simple rules are competitive with state-of-the-art machine learning algorithms. We conclude with an analytical framework that helps explain why simple decision rules often perform well.

2 ILLUSTRATION: BAIL DECISIONS

As an initial example of how to create simple rules that make accurate and transparent decisions, we turn to the domain of pretrial release determinations. In the United States, a defendant is typically arraigned shortly after arrest in a court appearance where he is provided with written notice of the charges alleged by the prosecutor. At this time, a judge must decide whether the defendant, while he awaits trial, should be *released on his own recognizance* (RoR), or alternatively, subject to monetary bail. In practice, if the judge rules that bail be set, defendants often await trial in jail since many of them do not have the financial resources to post bail. Moreover, when defendants are able to post bail, they often do so by contracting with a bail bondsman and in turn incur hefty fees. The judge, however, has a legal obligation to consider taking measures

Working paper, Stanford University
2017. 978-x-xxxx-xxxx-x/YY/MM.
DOI: 10.1145/nnnnnnn.nnnnnnn

necessary to secure the defendant's appearance at required court proceedings. Pretrial release decisions must thus balance flight risk against the high burden that bail requirements place on defendants. In many jurisdictions judges may also consider a defendant's threat to public safety, but that is not a legally relevant factor for the specific jurisdiction we analyze below.

A key statistical challenge in this setting is that one cannot, with historical data alone, directly observe the effects of hypothetical decision rules. For example, if a proposed policy recommends releasing some defendants who in reality were detained by the judge, one does not observe what would have happened had the rule been followed. This counterfactual estimation problem—also known as offline policy evaluation [7]—is common in many domains. We address it here by adapting tools from causal inference to the policy setting, including the method of Rosenbaum and Rubin [28] for assessing the sensitivity of estimated causal effects to an unobserved binary covariate.

Our analysis is based on 165,000 adult cases involving nonviolent offenses charged by a large urban prosecutor's office and arraigned in criminal court between 2010 and 2015. This set was obtained by starting with a random sample of 200,000 cases provided to us by the prosecutor's office, and then restricting to those cases involving nonviolent offenses and for which the records were complete and accurate. Our initial sample of 200,000 cases does not include instances where defendants accepted a plea deal at arraignment, obviating the need for a pretrial release decision. For each case, we have a rich set of attributes: 49 features describe characteristics of the current charges (*e.g.*, theft, gun-related), and 15 describe characteristics of the defendant (*e.g.*, gender, age, prior arrests). We also observe whether the defendant was RoR'd, and whether he failed to appear (FTA) at any of his subsequent court dates. We note that even if bail is set, a defendant may still fail to appear since he could post bail and then skip his court date. Overall, 69% of defendants are RoR'd, and 15% of RoR'd defendants fail to appear. Of the remaining 31% of defendants for whom bail is set, 45% are eventually released and 9% fail to appear. As a result, the overall FTA rate is 13%.

In our analysis below, we randomly divide the full set of 165,000 cases into three approximately equal subsets; we use the first fold to construct decision rules (both simple and complex), and the second and third to evaluate these rules, as described next.

2.1 Rule construction

We start by constructing traditional (but complex) decision rules for balancing flight risk with the burdens of bail. These rules serve as a benchmark for evaluating the simple rules we create below. On the first fold of the data, we restrict to cases in which the judge RoR'd the defendant, and then train a random forest model to estimate the likelihood an individual fails to appear at any of his subsequent court dates. Random forests are considered to be one of the best off-the-shelf classification algorithms [8], and we fit the model on all available information about the case and the defendant.[1] The fitted model lets us compute risk scores (*i.e.*, estimated flight risk if RoR'd) for any defendant. These risk scores can in turn be converted to a binary decision rule by selecting a threshold for releasing individuals. One might, for example, RoR a defendant if and only if his flight risk is below 20%.

[1] We use the `randomForest` package in `R`, fit with 1,000 trees.

Table 1: A defendant's flight risk is obtained by adding the scores for age and prior failure to appear (FTA).

Feature	Score	Feature	Score
$18 \leq$ age < 21	8	no prior FTAs	0
$21 \leq$ age < 26	6	1 prior FTA	6
$26 \leq$ age < 31	4	2 prior FTAs	8
$31 \leq$ age < 51	2	3 prior FTAs	9
$51 \leq$ age	0	4 or more prior FTAs	10

We now construct a family of simple rules for making release decisions. We begin by fitting a logistic regression model that estimates a defendant's flight risk as a function of his age and prior history of failing to appear. These two factors are well understood to be highly predictive in this context, but we later show how such features can be selected in a principled fashion without domain expertise. Specifically, we fit the following model:

$$\Pr(Y_i = 1) = \text{logit}^{-1}\Big(\beta_0 + \beta_1^{\text{priors}} H_i^1 + \beta_2^{\text{priors}} H_i^2 + \beta_3^{\text{priors}} H_i^3 + \beta_{4+}^{\text{priors}} H_i^{4+} + \beta_{18-20}^{\text{age}} A_i^{18-20} + \beta_{21-25}^{\text{age}} A_i^{21-25} + \cdots + \beta_{46-50}^{\text{age}} A_i^{46-50}\Big),$$

where $Y_i \in \{0, 1\}$ indicates whether the i-th defendant failed to appear; $H_i^* \in \{0, 1\}$ indicates the defendant's number of past failures to appear (exactly one, two, three, or at least four); and $A_i^* \in \{0, 1\}$ indicates the binned age of the defendant (18–20, 21–25, 26–30, 31–35, 36–40, 41–45, or 46–50). For identifiability, indicator variables for zero past FTAs and age 51-and-older are omitted. As before, this model is fit on the subset of cases in the first fold of data for which the judge released the defendant. Next, we rescale the age and prior FTA coefficients so that they lie in the interval $[-10, 10]$; specifically we multiply each coefficient by the constant

$$\frac{10}{\max\left(|\beta_1^{\text{priors}}|, \ldots, |\beta_{4+}^{\text{priors}}|, |\beta_{18-20}^{\text{age}}|, \ldots, |\beta_{46-50}^{\text{age}}|\right)}.$$

Finally, we round the rescaled coefficients to the nearest integer.

Table 1 shows the result of this procedure. For any defendant, a risk score can be computed by summing the relevant terms in the table. Unsurprisingly, past FTAs are indeed strong predictors of future failure to appear; an individual's risk also declines with age, in line with conventional wisdom. These risk scores can be converted to a binary decision rule by selecting a threshold for releasing individuals. For example, one might RoR a defendant if and only if his risk score is below 9.5. A graphical representation of that rule is shown in Figure 1.

2.2 Policy evaluation

There are two key considerations in evaluating a decision rule for pretrial release: (1) the proportion of defendants who are released under the rule; and (2) the resulting proportion who fail to appear at their court proceedings. It is straightforward to estimate the former, since one need only apply the rule to historical data to

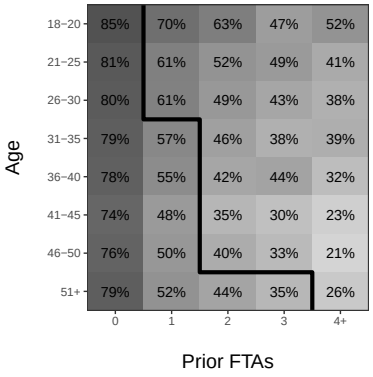

Figure 1: Graphical representation of a simple rule for release decisions, based on setting a release threshold of 9.5 on the risk scores described in Table 1. Groups to the left of the black line are those that would be released under the rule; for comparison, the shading and numbers show the proportion of defendants that are currently RoR'd in each group.

see what actions would have been recommended.[2] For example, if defendants are released if and only if their risk score is below 9.5, 77% would be RoR'd. Forecasting the proportion who would fail to appear, however, is generally much more difficult. The key problem is that for any particular defendant, we only observe the outcome (*i.e.*, whether or not he failed to appear) conditional on the action the judge ultimately decided to take (*i.e.*, RoR or bail). Since the action taken by the judge may differ from that prescribed by the decision rule, we do not always observe what would have happened under the rule. This problem of *offline policy evaluation* [7] is a specific instance of the fundamental problem of causal inference.

To rigorously describe the estimation problem and our approach, we first introduce some notation. We denote the observed set of cases by $\Omega = \{(x_i, a_i, r_i)\}$, where x_i is a case, $a_i \in \{\text{RoR}, \text{bail}\}$ is the action taken by the judge, and $r_i \in \{0, 1\}$ indicates whether the defendant failed to appear at his scheduled court date. We write $r_i(\text{RoR})$ and $r_i(\text{bail})$ to mean the *potential outcomes*, what would have happened under the two possible judicial actions. For any policy π, our goal is to estimate the FTA rate under the policy:

$$V^\pi = \frac{1}{|\Omega|} \sum_i r_i(\pi(x_i))$$

where $\pi(x)$ denotes the action prescribed under the rule. The key statistical challenge is that only one of the two potential outcomes, $r_i = r_i(a_i)$, is observed. We note that policy evaluation is a generalization of estimating average treatment effects. Namely, the average treatment effect can be expressed as $V^{\pi_{\text{RoR}}} - V^{\pi_{\text{bail}}}$, where π_{RoR} is the policy under which everyone is released and π_{bail} is defined analogously.

Here we take a straightforward and popular statistical approach to estimating V^π: response surface modeling [15]. With response surface modeling, the idea is to use a standard prediction model (*e.g.*, logistic regression or random forest) to estimate the effect on each defendant of each potential judicial action. The model estimates of these potential outcomes are denoted by $\hat{r}_i(t)$, for $t \in \{\text{RoR}, \text{bail}\}$. Our estimate of V^π is then given by

$$\hat{V}^\pi = \frac{1}{|\Omega|} \sum_i \left[r_i \mathbf{I}(\pi(x_i) = a_i) + \hat{r}_i(\pi(x_i)) \mathbf{I}(\pi(x_i) \neq a_i) \right]$$

where $\mathbf{I}(\cdot)$ is an indicator function evaluating to 1 if its argument is true and to 0 otherwise. If the prescribed action is in fact taken by the judge, then $r_i = r_i(\pi(x_i))$ is directly observed and can be used; otherwise we approximate the potential outcome with $\hat{r}_i(\pi(x_i))$. Table 2 illustrates this method for a hypothetical example.

Table 2: For each defendant, $\hat{Y}_{\text{RoR}}$ and $\hat{Y}_{\text{bail}}$ are model-based estimates of the likelihood of FTA under each potential action. In cases where the observed action equals the proposed action, the observed outcome (FTA or not) is used to estimate the policy's effect; otherwise, the model-based estimates are used. The gray shading indicates which values are used in each instance. The overall FTA rate under the policy is estimated by averaging the shaded values over all cases.

Proposed action	Observed action	Observed outcome	$\hat{Y}_{\text{RoR}}$	$\hat{Y}_{\text{bail}}$
RoR	RoR	0	20%	10%
Bail	Bail	1	80%	30%
Bail	RoR	1	90%	70%
RoR	Bail	0	30%	25%
RoR	RoR	0	20%	15%

Response surface modeling implicitly assumes that a judge's action is *ignorable* given the observed covariates (*i.e.*, that conditional on the observed covariates, those who are RoR'd are similar to those who are not). Formally, ignorability means that

$$(r(\text{RoR}), r(\text{bail})) \perp\!\!\!\perp a \,\big|\, x.$$

This ignorability assumption is unavoidable, and is similarly required for methods based on propensity scores [1, 7, 16, 26, 27, 29, 30]. We examine this assumption in detail in Section 2.3, and find that our conclusions are robust to unobserved heterogeneity.

To carry out this approach, we derive estimates $\hat{r}_i(t)$ via an L^1-regularized logistic regression (lasso) model trained on the second fold of our data. For each individual, the model estimates his likelihood of FTA given all the observed features and the action taken by the judge. In contrast to the rule construction described above, this time we train the model on all cases (not just those for which the judge RoR'd the defendant) and include as a predictor the judge's action (RoR or bail).[3] Then, on the third fold of the data, we use the observed and model-estimated outcomes to approximate the overall FTA rate for any decision rule.

Figure 2 shows estimated RoR and FTA rates for a variety of pretrial release rules. Points on the solid line correspond to rules

[2] In theory, implementing a decision rule could alter the equilibrium distribution of defendants. We do not consider such possible effects, and assume the distribution of defendants is not affected by the rule itself.

[3] The model was fit with the `glmnet` package in R. The `cv.glmnet` method was used to determine the best value for the regularization parameter λ with 10-fold CV and 1,000 values of λ. The model includes all pairwise interactions between the judge's decision and defendant's features. We opt for lasso instead of random forest for this prediction task because the latter, while very good for classification, is known to suffer from poor calibration [25], which can in turn yield biased estimates of a policy's effects.

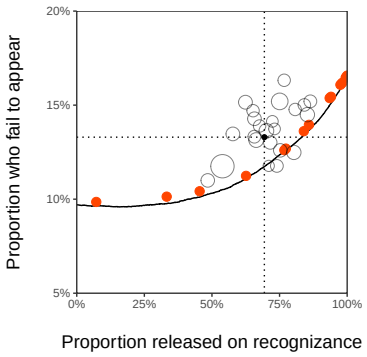

Figure 2: Each point on the solid line corresponds to decision rules derived from a random forest risk model with varying thresholds for release. The red points correspond to the simple risk score in Table 1 for all possible release thresholds. The simple rules perform nearly identically to the complex models. The open circles show the observed RoR and FTA rates for each judge in our data who presided over at least 1,000 cases, sized in proportion to their case load. In nearly every instance, the statistical decision rules outperform the human decision-maker.

constructed via the random forest model described above for various decision thresholds. The red points correspond to rules based on the simple scoring procedure in Table 1, again corresponding to various decision thresholds. For each rule, the horizontal axis shows the estimated proportion of defendants ROR'd under the rule, and the vertical axis shows the estimated proportion of defendants who would fail to appear at their court dates. The solid black dot shows the status quo: 69% of defendants RoR'd and a 13% FTA rate. Finally, the open circles show the observed RoR and FTA rates for each of the 23 judges in our data who have presided over at least 1,000 cases, sized in proportion to their case load.

The plot illustrates three key points. First, simple rules that consider only two features—age and prior FTAs—perform nearly identically to a random forest that incorporates 64 features. Second, the statistically informed policies in the lower right quadrant all achieve higher rates of RoR and, simultaneously, lower rates of FTA than the status quo. In particular, by releasing defendants if and only if their risk score is below 9.5, we expect to release 77% of defendants while achieving an FTA rate of 13%. Relative to the existing policy, this rule results in 8 percentage points more people being released while still maintaining the overall FTA rate. Finally, for nearly every judge, there is a statistical decision rule that simultaneously yields both a higher rate of release and a lower rate of FTA than the judge currently achieves. The statistical decision rules consistently outperform the human decision-makers.

Why do these statistical decision rules outperform the experts? Figure 1 sheds light on this phenomenon. Each cell in the plot corresponds to defendants binned by their age and prior number of FTAs. Under a rule that releases defendants if and only if their risk score is below 9.5, one would release everyone to the left of the solid black line, and set bail for everyone to the right of the line. The number in each cell shows the proportion of defendants in each bin who are currently released, and the cell shading graphically indicates this proportion. Aside from the lowest risk defendants, who have no prior FTAs, the likelihood of being released does not correlate strongly with estimated flight risk. For example, the high risk group of young defendants with four or more prior FTAs is released at about the same 50% rate as the low risk group of older defendants with one prior FTA. This low correlation between flight risk and release decision is in part attributable to extreme differences in release rates across judges, with some releasing more than 90% of defendants and others releasing just 50%.[4] Whereas defendants experience dramatically different outcomes based on the judge they happened to appear in front of, statistical decision rules improve efficiency in part by ensuring consistency.

2.3 Sensitivity to unobserved heterogeneity

As noted above, our estimation strategy assumes that the judicial action taken is ignorable given the observed covariates. Under this ignorability assumption, one can accurately estimate the potential outcomes. Judges, however, might base their decisions in part on information that is not recorded in the data, which could in turn bias our estimates. For example, a judge, upon meeting a defendant, might surmise that his flight risk is higher than one would expect based on the recorded covariates alone, and may accordingly require the defendant to post bail. In this case, since our estimates are based only on the recorded data, we may underestimate the defendant's counterfactual likelihood of failing to appear if released.

We take two approaches to gauge the robustness of our results to such hidden heterogeneity. First, on each subset of cases handled by a single judge, we use response surface modeling to estimate V^{π}. Each judge has idiosyncratic criteria for releasing defendants, as evidenced by the dramatically different release rates across judges; accordingly, the types and proportion of cases for which the policy π coincides with the observed action differ from judge to judge. This variation allows us to assess the sensitivity of our estimates to the observed actions $\{a_i\}$. In particular, if unobserved heterogeneity were significant, we would expect our estimates to systematically vary depending on the proportion of observed judicial actions that agree with the policy π. Figure 3 shows the results of this analysis for the simple decision rule described in Figure 1, where each point corresponds to a judge. We find that the FTA rate of the decision rule is consistently estimated to be approximately 12–14%. Moreover, some judges act in concordance with the decision rule in nearly 80% of cases; for this subset of judges, where our estimates are largely based on directly observed outcomes, we again find FTA is estimated at around 12–14%.

As a second robustness check, we adapt the method of Rosenbaum and Rubin [28] for assessing the sensitivity of estimated causal effects to an unobserved binary covariate. We specifically tailor their approach to offline policy evaluation. At a high level, we assume there is an unobserved covariate $u \in \{0, 1\}$ that affects both a judge's decision (RoR or bail) and also the outcome conditional on that action. For example, u might indicate that a defendant is sympathetic, and sympathetic defendants may be more likely to be RoR'd and also more likely to appear at their court proceedings.

[4] Defendants are not perfectly randomly assigned to judges for arraignment, but in practice judges see a similar distribution of defendants.

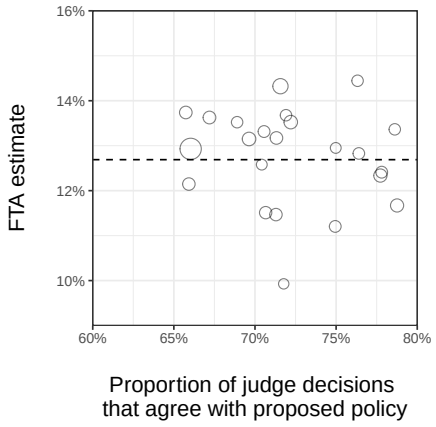

Figure 3: For the simple decision rule illustrated in Figure 1, FTA rate is estimated by separately applying response surface modeling to each judge's cases, where each point corresponds to a judge; the dashed horizontal line indicates the FTA rate of the decision rule estimated on the full set of cases. Though judges have different criteria for releasing defendants—and the corresponding response models may thus differ—the FTA rate of the decision rule is consistently estimated to be approximately 12–14%.

Our key assumption is that a judge's action is ignorable given the observed covariates x and the unobserved covariate u:

$$(r(\text{RoR}), r(\text{bail})) \perp\!\!\!\perp a \,\big|\, x, u. \tag{1}$$

There are four key parameters in this framework: (1) the probability that $u = 1$; (2) the effect of u on the judge's decision; (3) the effect of u on the defendant's likelihood of FTA if RoR'd; and (4) the effect of u on the defendant's likelihood of FTA if bail is set. Our goal is to quantify the extent to which our estimate of V^π changes as a function of these parameters.

Without loss of generality, we can write

$$\Pr(a = \text{RoR}|u, x) = \text{logit}^{-1}(\gamma_x + u\alpha_x) \tag{2}$$

for appropriately chosen parameters γ_x and α_x that depend on the observed covariates x. We note that randomness in judicial decisions may arise from a multitude of factors, including idiosyncrasies in how judges are assigned to cases. Here α_x is the change in log-odds of being RoR'd when $u = 0$ versus when $u = 1$. For $t \in \{\text{RoR}, \text{bail}\}$, we can similarly write

$$\Pr(r(t)|u, x) = \text{logit}^{-1}\left(\beta_x^t + u\delta_x^t\right) \tag{3}$$

for parameters β_x^t and δ_x^t. In this case, δ_x^{RoR} is the change in log-odds of failing to appear if RoR'd when $u = 0$ versus when $u = 1$, and δ_x^{bail} is the corresponding change if bail is set.

Now, for any posited values of $\Pr(u = 1|x)$, α_x, δ_x^{RoR} and δ_x^{bail}, we use the observed data to estimate γ_x, β_x^{RoR} and β_x^{bail}. We do this in three steps. By (2),

$$\Pr(a = \text{RoR}|x) = \Pr(u = 0|x) \cdot \text{logit}^{-1}(\gamma_x) + \Pr(u = 1|x) \cdot \text{logit}^{-1}(\gamma_x + \alpha_x).$$

The left-hand side of the equation can be estimated with a regression model fit to the data. For fixed values of $\Pr(u = 1|x)$ and α_x, the right-hand side is an increasing function of γ_x that takes on values from 0 to 1 as γ_x goes from $-\infty$ to $+\infty$. There is thus a unique value $\hat{\gamma}_x$ such that the right-hand side equals $\hat{\Pr}(a = \text{RoR}|x)$. Rosenbaum and Rubin [28] derive a simple closed form solution for $\hat{\gamma}_x$, facilitating fast computation on large datasets, which we omit for space.

Second, we use the fitted values of γ_x to estimate the distribution of u given the observed covariates and judicial action. By Bayes' rule,

$$\Pr(u = 1|a = t, x) = \frac{\Pr(a = t|u = 1, x)\Pr(u = 1|x)}{\Pr(a = t|x)} = \frac{\Pr(a = t|u = 1, x)\Pr(u = 1|x)}{\Pr(a = t|u = 1, x)\Pr(u = 1|x) + \Pr(a = t|u = 0, x)\Pr(u = 0|x)}.$$

With $\hat{\gamma}_x$, the $\Pr(a = t|u, x)$ terms on the right-hand side can be estimated from (2), and we can thus approximate the left-hand side.

Third, we have

$$\begin{aligned}
&\Pr(r(t) = 1|a = t, x) \\
&\quad = \Pr(u = 0|a = t, x)\Pr(r(t) = 1|a = t, x, u = 0) \\
&\qquad + \Pr(u = 1|a = t, x)\Pr(r(t) = 1|a = t, x, u = 1) \\
&\quad = \Pr(u = 0|a = t, x)\Pr(r(t) = 1|x, u = 0) \\
&\qquad + \Pr(u = 1|a = t, x)\Pr(r(t) = 1|x, u = 1) \\
&\quad = \Pr(u = 0|a = t, x) \cdot \text{logit}^{-1}\left(\beta_x^t\right) \\
&\qquad + \Pr(u = 1|a = t, x) \cdot \text{logit}^{-1}\left(\beta_x^t + \delta_x^t\right).
\end{aligned}$$

The second equality above follows from the ignorability assumption stated in (1), and the third equality follows from (3). The left-hand side can be approximated by the quantity $\hat{r}_x(t)$ that we obtain via response surface modeling. Importantly, $\hat{r}_x(t)$ is a reasonable estimate of $\Pr(r(t) = 1|a = t, x)$ even though it may not be a good estimate of $r_x(t)$. This distinction is indeed the rationale of our sensitivity analysis. Given our above estimate of $\Pr(u = 1|a = t, x)$ and our assumed value of δ_x^t, the only unknown on the right-hand side is β_x^t. As before, there is a unique value $\hat{\beta}_x^t$ that satisfies the constraint.

With $\hat{\beta}_x^t$ in hand, we can now approximate the potential outcome for the action *not* taken:

$$\Pr(r(\bar{t}) = 1|a = t, x)$$

where $\bar{t} = \text{RoR}$ if $t = \text{bail}$, and vice versa. Specifically, we have

$$\hat{\Pr}(r(\bar{t}) = 1|a = t, x) = \hat{\Pr}(u = 0|a = t, x) \cdot \text{logit}^{-1}\left(\hat{\beta}_x^{\bar{t}}\right) + \hat{\Pr}(u = 1|a = t, x) \cdot \text{logit}^{-1}\left(\hat{\beta}_x^{\bar{t}} + \delta_x^{\bar{t}}\right). \tag{4}$$

Finally, the Rosenbaum and Rubin estimator adapted to policy evaluation is

$$\hat{V}_{\text{RR}}^\pi = \frac{1}{|\Omega|}\sum_i \left[r_i \mathbf{I}(\pi(x_i) = a_i) + \hat{r}_i(\bar{a}_i)\mathbf{I}(\pi(x_i) \neq a_i)\right],$$

where $\hat{r}_i(\bar{a}_i) = \hat{\Pr}(r(\bar{a}_i) = 1|a_i, x_i)$ is computed via (4).

Figure 4 shows the results of computing $\hat{V}_{\text{RR}}^\pi$ on our data in two parameter regimes. In the first (left-hand plot), we assume $\alpha = \log 2$ and consider all combinations of $p(u = 1) \in \{0.1, 0.2, \ldots, 0.9\}$,

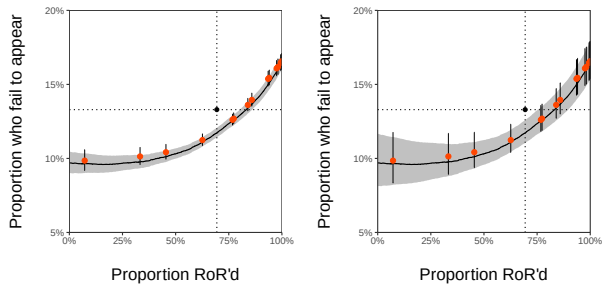

Figure 4: The grey band (for the complex rules) and the error bars (for the simple rules) indicate minimum and maximum FTA estimates for a variety of parameter settings. In the left-hand plot, we assume $\alpha = \log 2$ and consider all combinations of $p(u = 1) \in \{0.1, 0.2, \ldots, 0.9\}$, $\delta^{\text{RoR}} \in \{-\log 2, 0, \log 2\}$, and $\delta^{\text{bail}} \in \{-\log 2, 0, \log 2\}$, where all parameters are constant independent of x. In the right-hand plot, we consider a more extreme situation, with $\alpha = \log 3$, $\delta^{\text{RoR}} \in \{-\log 3, 0, \log 3\}$, and $\delta^{\text{bail}} \in \{-\log 3, 0, \log 3\}$. The results are relatively stable in these parameter regimes.

$\delta^{\text{RoR}} \in \{-\log 2, 0, \log 2\}$, and $\delta^{\text{bail}} \in \{-\log 2, 0, \log 2\}$. All parameters are constant independent of x. We thus assume that holding the observed covariates fixed, a defendant with $u = 1$ has twice the odds of being RoR'd as one with $u = 0$, and that u can double or half the odds a defendant fails to appear. For each complex policy (*i.e.*, one based on a random forest), the grey band shows the minimum and maximum value of $\hat{V}^{\pi}_{\text{RR}}$ across all parameters in this set; the error bars on the red points show the analogous quantity for the simple rules. In the right-hand plot, we consider a more extreme situation, with $\alpha = \log 3$, $\delta^{\text{RoR}} \in \{-\log 3, 0, \log 3\}$, and $\delta^{\text{bail}} \in \{-\log 3, 0, \log 3\}$. We find that our estimates are relatively stable in these parameter regimes. In the first case ($\alpha = \log 2$) the estimated FTA rate for a given policy typically varies by only half a percentage point. Even in the more extreme setting ($\alpha = \log 3$), policies are typically stable to about one percentage point. It thus seems our conclusions are robust to unobserved heterogeneity across defendants.

3 SELECT-REGRESS-AND-ROUND: A SIMPLE METHOD FOR CREATING SIMPLE RULES

We now introduce and evaluate a simple method—which we call *select-regress-and-round*—that formalizes and generalizes the rule construction procedure we applied for pretrial release decisions. In particular, we dispense with ad hoc feature selection and adopt a standard statistical routine.

3.1 Rule construction

The rules we construct are designed to aid classification or ranking decisions by assigning each item in consideration a score z, computed as a linear combination of a subset S of the item features:

$$z = \sum_{j \in S} w_j x_j,$$

where the weights w_j are integers. In the cases we consider, the features themselves are typically 0-1 indicator variables (indicating, for example, whether a person is male, or whether an individual is 26–30 years old), and so the rule reduces to a weighted checklist, in which one simply sums up the (integer) weights of the applicable attributes. Often, one seeks to make binary decisions (*e.g.*, whether to detain or to release an individual), which amounts to setting a threshold and then taking a particular course of action if and only if the score is above that threshold.

This class of rules has two natural dimensions of complexity: the number of features and the magnitude of the weights. Given integers $k \geq 1$ and $M \geq 1$, we apply the following three-step procedure to construct rules with at most k features and integer weights bounded by M (*i.e.*, $|S| \leq k$ and $-M \leq w_j \leq M$).

(1) **Select.** From the full set of features, select k features via forward stepwise regression. For fixed k, we note that standard selection metrics (*e.g.*, AIC or BIC) are theoretically guaranteed to yield the same set of features.
(2) **Regress.** Using only these k selected features, train an L^1-regularized (lasso) logistic regression model to the data, which yields (real-valued) fitted coefficients $\beta_1, \ldots, \beta_k$.
(3) **Round.** Rescale the fitted coefficients to be in the range $[-M, M]$, and then round the rescaled coefficients to the nearest integer. Specifically, set

$$w_j = \text{Round}\left(\frac{M\beta_j}{\max_i |\beta_i|}\right).$$

We note that rules constructed in this way may have fewer than k features, since the lasso regression in Step 2 may result in coefficients that are identically zero, and rescaling and rounding coefficients in Step 3 may zero-out additional terms.[5] This select-regress-and-round strategy for rule construction builds upon findings that "improper" weighting schemes for linear models (e.g, unit weighting) lead to accurate predictions [4, 9, 12, 14]; in particular, our strategy incorporates feature selection and more general integer weights to generate a richer family of simple rules. We next examine the accuracy of these rules.

3.2 Rule evaluation

We apply the select-regress-and-round procedure to 22 publicly available datasets to examine the tradeoff between complexity and performance. These datasets all come from the UCI ML repository, and were selected according to four criteria: (1) the dataset involves a binary classification (as opposed to a regression) problem;[6] (2) the dataset is provided in a standard and complete form; (3) the dataset involves more than 10 features; and (4) the classification problem is one that a human could plausibly learn to solve with the given features. For example, we included a dataset in which the task was to determine whether cells were malignant or benign based on various biological attributes of the cells, but we excluded image recognition tasks in which the features were represented as pixel

[5] We select features in Step 1 with the `R` package `leaps`. The models in Step 2 are fit with the `R` package `glmnet`. The `cv.glmnet` method is used to determine the best value of the regularization parameter λ with 10-fold CV and 1,000 values of λ.
[6] For those datasets whose outcome variable takes more than two values, we set the majority class as the target variable, so that all the tasks we consider involve binary classification.

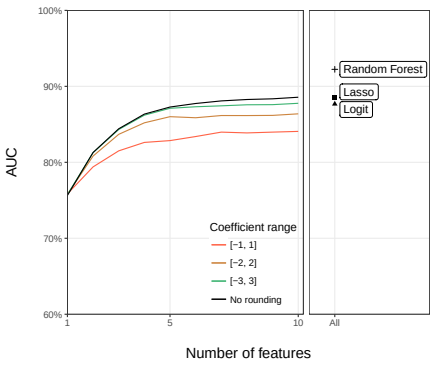

Figure 5: Mean test AUC of decision rules over 22 datasets. The simple rules use up to 10 features, with integer coefficients in the specified ranges. The black line shows performance of lasso with feature selection but without rounding the coefficients. "All" features — used by random forest, lasso, and logistic regression — varies by domain, with an average of 38.

values. This fourth requirement limits the scope of our analysis and conclusions to domains in which human decision makers typically act without the aid of a computer.[7]

Unlike the judicial decisions discussed in Section 2, outcomes in the domains we consider here are unaffected by a decision maker's actions. For example, assessing the likelihood a cell is malignant—and then acting on that knowledge—does not change the fact that the cell was either malignant or not at the time of the measurement. Similarly, when identifying potential tax fraud, assessing risk and conducting an audit does not affect whether an individual indeed cheated on his taxes. In contrast, a judge's decision to release or detain an individual necessarily alters the defendant's likelihood of appearing at trial. In both types of scenarios, decision rules are constructed similarly. Evaluating the resulting rules, however, is significantly easier when outcomes are independent of actions, as one need not consider subtle issues of causal inference.

On each of the 22 datasets we analyze here, we construct simple rules for a range of values of $k \in \{1, \ldots, 10\}$ and $M \in \{1, 2, 3\}$, the number of features and the magnitude of the weights. We benchmark the performance of these rules against three standard statistical models: logistic regression, L^1-regularized logistic regression, and random forest. These models were fit in `R` with the `glm`, `glmnet`, and `randomForest` packages, respectively. For the L^1-regularized logistic regression models, the `cv.glmnet` method was used to determine the best value of the regularization parameter λ with 10-fold CV and 1,000 values of λ. We used 1,000 trees for the random forest models. This head-to-head comparison is a difficult test for the simple rules in part because they can only base their predictions on 1 to 10 features. The complex models, in contrast, can train and predict with all features, which number between 11 and 93 with a mean of 38.

[7]The 22 UCI datasets we consider are: `adult`, `annealing`, `audiology-std`, `bank`, `bankruptcy`, `car`, `chess-krvk`, `chess-krvkp`, `congress-voting`, `contrac`, `credit-approval`, `ctg`, `cylinder-bands`, `dermatology`, `german_credit`, `heart-cleveland`, `ilpd`, `mammo`, `mushroom`, `aus_credit`, `wine`, and `wine_qual`.

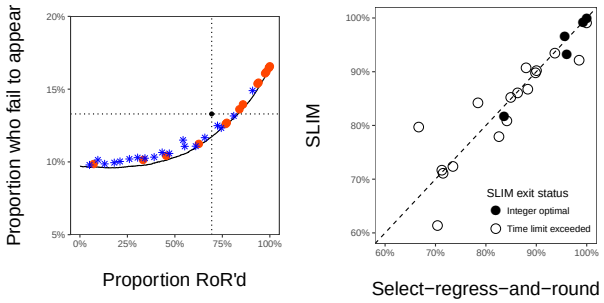

Figure 6: Left panel: comparison of rules for pretrial release decisions produced by select-regress-and-round (red), SLIM (blue), and random forest (black line). Right panel: binary classification accuracy for select-regress-and-round and SLIM on 22 UCI datasets.

Figure 5 shows model performance—measured in terms of mean AUC across the 22 datasets—as a function of model size and coefficient range. The AUC for each model on each dataset is computed via 10-fold CV. We find that simple rules with only five features and integer coefficients between -3 and 3 perform on par with logistic regression and L^1-regularized logistic regression trained on the full set of features. For 1 to 10 features, the [-3, 3] model (green line) differs from the unrounded lasso model (black line) by less than 1 percentage point. The performance of the random forest model is somewhat better: trained on all features, random forest achieves mean AUC of 92%; the mean AUC is 87% for simple rules with at most five features and integer coefficients between -3 and 3. Complex prediction methods certainly have their advantages, but the gap in performance between simple rules and fully optimized prediction methods is not as large as one might have thought.

3.3 Benchmarking to integer programming

The simple rules we construct take the form of a linear scoring rule with integer weights. To produce such rules, mixed-integer programming is a natural alternative to our select-regress-and-round strategy, and supersparse linear integer models (SLIM) [33] is the leading instantiation of that approach. Given constraints on the number of features and the magnitude of the integer weights, SLIM produces rules that optimize for binary classification accuracy (*i.e.*, 0-1 loss).

We compare SLIM to select-regress-and-round on the judicial decision-making problem and on the 22 UCI datasets. Figure 6 (left panel) shows estimated FTA and release rates for the random forest model (black line), our simple rules derived in Section 2 (red points), and the simple rules produced by SLIM (blue points). As with our own simple rules, we constrain SLIM to produce rules based on age and number of past FTAs, with integer weights ranging from -10 to 10. As before, decision rules are constructed from the random forest and select-regress-and-round risk scores by varying the decision threshold; in contrast, multiple rules for SLIM are computed by varying a parameter that specifies the maximum acceptable false positive rate [33]. Both methods for producing simple rules perform nearly the same as the random forest model trained on the full set of 64 features.

We next consider the 22 UCI datasets. SLIM is known to work best when the features are discrete [35]. We thus pre-process the datasets by discretizing all continuous features into three bins containing an approximately equal number of examples, representing low, medium, and high values of the feature. Integer programming is an NP-hard problem, and so following Ustun and Rudin [33] we set a time limit for SLIM; they set a 10-minute limit, but we allow up to 1 hour of computation per model. For 5 of the 22 datasets, SLIM found an integer-optimal solution within the time limit, returning approximate solutions in the remaining 17 cases. Figure 6 (right panel) compares binary classification accuracy of SLIM and select-regress-and-round on the 22 UCI datasets, where each point corresponds to a dataset. Both methods are constrained to produce rules with at most five features and integer coefficients between -3 and 3. We show 0-1 accuracy since SLIM optimizes for this metric, but similar results also hold for AUC; accuracy is computed out-of-sample via 10-fold CV. Both methods for producing simple rules yield comparable results. Averaged across all 22 datasets, SLIM and select-regress-and-round both achieve mean accuracy of 86%. Even in the 5 cases where SLIM found integer-optimal solutions, performance is nearly identical to our simple select-regress-and-round strategy.

In terms of classification accuracy, select-regress-and-round generates rules on par with those obtained by solving mixed-integer programs. We note, however, two advantages of our approach. First, whereas select-regress-and-round yields results almost instantaneously, integer programs can be computationally expensive to solve. Second, our approach is both conceptually and technically simple, requiring little statistical or computational expertise, and accordingly easing adoption for practitioners.

4 THE ROBUSTNESS OF CLASSIFICATION

Why is it that simple rules often perform as well as the most sophisticated statistical methods? In part it is because binary classification is robust to error in the underlying predictive model, an observation that we formalize in Theorem 4.1 below.

To establish this result, we start by considering the prediction scores generated via a standard statistical method—such as logistic regression trained on the full set of available features—which we call the "true" scores. As in linear discriminant analysis, we assume that the true scores for positive and negative instances are normally distributed with equal variance: $\mathrm{N}(\mu_p, \sigma^2)$ and $\mathrm{N}(\mu_n, \sigma^2)$, respectively. The homoscedasticity assumption guarantees the Bayes optimal classifier is a threshold rule on the scores. For scores estimated via logistic regression, the normality assumption is reasonable if we consider the scores on the logit scale rather than on the probability scale. Figure 7 (left panel) shows such scores for one of the UCI datasets. We further assume that the process of generating simple rules—both limiting the number of features and also restricting the possible values of the weights—can be viewed as adding normal, mean-zero noise $\mathrm{N}(0, \sigma_\epsilon^2)$ to the true scores; Figure 7 (center panel) plots the distribution of this noise for one of the datasets.[8] Thus, with simple rules, instead of making classification decisions based on the true scores, we assume decisions are made in terms of a noisy approximation. Under this analytic framework, Theorem 4.1 shows that the drop in classification performance (as measured by AUC) can be expressed in terms of the "true AUC" (*i.e.*, the AUC under the true scores) and $\gamma = \sigma_\epsilon^2/\sigma^2$, the ratio of the noise to the within-class variance of the true scores. In particular, we find that when the magnitude of the noise is on par with (or smaller than) the score variance (*i.e.*, $\gamma \lesssim 1$), then the AUC of the noisy approximation is comparable to the true AUC.

Theorem 4.1. *For a binary classification task, let Y be a continuous random variable that denotes the prediction score of a random instance, and let Y_p and Y_n denote the conditional distributions of Y for positive and negative instances, respectively. Suppose $Y_p \sim N(\mu_p, \sigma^2)$ and $Y_n \sim N(\mu_n, \sigma^2)$. Then, for $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\hat{Y} = Y + \epsilon$,*

$$\mathrm{AUC}_{\hat{Y}} = \Phi\left(\frac{\Phi^{-1}(\mathrm{AUC}_Y)}{\sqrt{1+\gamma}}\right), \tag{5}$$

where $\gamma = \sigma_\epsilon^2/\sigma^2$, and Φ is the CDF for the standard normal.

Proof. In general, AUC is equal to the probability that a randomly selected positive instance has a higher prediction score than a randomly selected negative instance, and so $\mathrm{AUC}_Y = \Pr(Y_p - Y_n > 0)$. Since $Y_p - Y_n$ is normally distributed with mean $\mu_p - \mu_n$ and variance $2\sigma^2$,

$$\frac{Y_p - Y_n - (\mu_p - \mu_n)}{\sqrt{2}\sigma} \sim \mathrm{N}(0,1).$$

Hence,

$$\begin{aligned}\mathrm{AUC}_Y &= \Pr\left(\frac{Y_p - Y_n - (\mu_p - \mu_n)}{\sqrt{2}\sigma} > -\frac{\mu_p - \mu_n}{\sqrt{2}\sigma}\right) \\ &= \Phi\left(\frac{\mu_p - \mu_n}{\sqrt{2}\sigma}\right),\end{aligned}$$

where the last equality follows from symmetry of the normal distribution.

Now define $\hat{Y}_p = Y_p + \epsilon$, so $\hat{Y}_p \sim \mathrm{N}(\mu_p, \sigma^2 + \sigma_\epsilon^2)$, with $\hat{Y}_n$ defined similarly. A short computation shows that

$$\mathrm{AUC}_{\hat{Y}} = \Pr(\hat{Y}_p > \hat{Y}_n) = \Phi\left(\frac{\mu_p - \mu_n}{\sqrt{2\sigma^2 + 2\sigma_\epsilon^2}}\right) = \Phi\left(\frac{\Phi^{-1}(\mathrm{AUC}_Y)}{\sqrt{1+\gamma}}\right).$$

□

Theorem 4.1 establishes a direct theoretical link between performance and noise in model specification. To give a better sense of how the analytic expression for $\mathrm{AUC}_{\hat{Y}}$ varies with AUC_Y and γ, Figure 7 (right panel) shows this expression for various parameter values. For example, the figure shows that for $\mathrm{AUC}_Y = 90\%$ and $\gamma = 0.5$, we have $\mathrm{AUC}_{\hat{Y}} = 85\%$. That is, if the amount of noise is equal to half the within-class variance of the true scores, then the drop in performance is relatively small.

While connecting model performance to model noise, Theorem 4.1 leaves unanswered how much noise simple rules add to the underlying scores. This question seems difficult to answer theoretically. We can, however, empirically estimate how much noise

[8] We estimate the noise distribution by taking the difference between the simple and true scores. Before taking the difference, we convert the simple scores to the scale of true scores by dividing the simple scores by M, the scaling factor used when generating the rule.

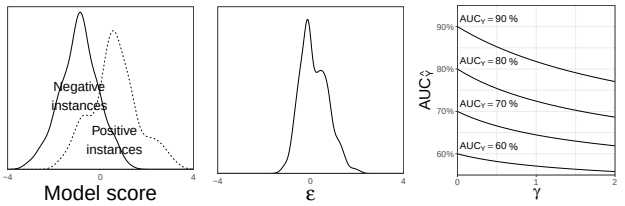

Figure 7: Left panel: the empirical distribution of prediction scores, on the logit scale, for positive and negative instances of a UCI dataset (`heart-cleveland`), generated via an L^1-regularized logistic regression model. Center panel: empirical distribution of ϵ for select-regress-and-round applied to the same dataset. Right panel: the theoretical change in AUC, under the setup of Theorem 4.1.

simple rules add in the datasets we analyze.[9] Across the 22 UCI datasets we consider, we find that rules with five features and a coefficient range of -3 to 3 have an average value of $\gamma = 0.22$. This low empirically observed noise is in line with our finding that such simple rules perform well on these datasets.

5 CONCLUSION

In this paper we introduced select-regress-and-round, a simple method for constructing decision rules that are fast, frugal, and clear. In an analysis of pretrial release decisions, simple rules outperformed human judges and matched the performance of a sophisticated statistical model. Generalizing this result, in 22 domains of varying size and complexity, the simple mental checklists produced by the select-regress-and-round method rivaled the performance of regularized regression models while using only a fraction of the information.

These results complement a growing body of work in statistics and computer science in which sophisticated algorithms are used to create interpretable scoring systems and rule sets [21, 22, 33]. Many prior rule construction methods offer great flexibility and performance, but in turn require considerable computational expertise to carry out. In contrast, the simple rules in this article can be created by practitioners with only basic statistical knowledge and generic software. For practitioners to favor statistics over intuition, we believe decision rules must not only be simple to apply but also simple to create.

ACKNOWLEDGMENTS

We thank Avi Feller, Andrew Gelman, Gerd Gigerenzer, Art Owen, and Berk Ustun for helpful conversations.

REFERENCES

[1] Claes M Cassel, Carl E Särndal, and Jan H Wretman. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, 3 (1976), 615–620.
[2] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230* (2017).
[3] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.
[4] Robyn M Dawes. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34, 7 (1979), 571.
[5] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.
[6] Mandeep K Dhami. 2003. Psychological models of professional decision making. *Psychological Science* 14, 2 (2003), 175–180.
[7] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly Robust Policy Evaluation and Learning. *ICML* (2011). DOI: http://dx.doi.org/10.1214/14-STS500
[8] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res* 15, 1 (2014), 3133–3181.
[9] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* 103, 4 (1996), 650.
[10] Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur. 2011. *Heuristics: The foundations of adaptive behavior.* Oxford University Press, Inc.
[11] Michael Gleicher. 2016. A Framework for Considering Comprehensibility in Modeling. *Big Data* 4, 2 (2016), 75–88.
[12] Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics* (2016).
[13] Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a right to explanation. *arXiv preprint arXiv:1606.08813* (2016).
[14] Joy Paul Guilford. 1942. *Fundamental statistics in psychology and education.* McGraw-Hill.
[15] Jennifer L Hill. 2012. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* (2012).
[16] Joseph DY Kang and Joseph L Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* (2007), 523–539.
[17] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In *Advances in Neural Information Processing Systems 27*. 1952–1960.
[18] Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In *Advances in Neural Information Processing Systems 28*. 2260–2268.
[19] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. (2017). Working paper.
[20] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *The American Economic Review* 105, 5 (2015).
[21] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[22] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.
[23] Julian N Marewski and Gerd Gigerenzer. 2012. Heuristic decision making in medicine. *Dialogues Clin Neurosci* 14, 1 (2012), 77–89.
[24] Clement J. McDonald. 1996. Medical Heuristics: The Silent Adjudicators of Clinical Practice. *Annals of Internal Medicine* 124, 1 Part 1 (1996), 56–62.
[25] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 625–632.
[26] James M Robins and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* 90, 429 (1995), 122–129.
[27] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.
[28] Paul R Rosenbaum and Donald B Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* (1983), 212–218.
[29] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[30] Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association* 79, 387 (1984), 516–524.
[31] Donald Sull and Kathleen M Eisenhardt. 2015. *Simple rules: How to thrive in a complex world.* Houghton Mifflin Harcourt.
[32] Philip Tetlock. 2005. *Expert political judgment: How good is it? How can we know?* Princeton University Press.
[33] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (2016), 349–391.

[9] To estimate $\gamma = \sigma_\epsilon^2/\sigma^2$ for a specific simple rule on a given dataset, we first compute the average within-class variance of the true scores, where these scores are generated via an L^1-regularized logistic regression model. We estimate σ_ϵ^2 by taking the variance of the noise, as described in Footnote 8.

[34] Markus Wübben and Florian V Wangenheim. 2008. Instant customer base analysis: Managerial heuristics often get it right. *Journal of Marketing* 72, 3 (2008), 82–93.
[35] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2016. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2016).