ORIGINAL ARTICLE

Journal Section

Simple rules for complex decisions

Jongbin Jung[1] | **Connor Concannon**[2] | **Ravi Shroff**[3] | **Sharad Goel**[1] | **Daniel G. Goldstein**[4]

[1]Stanford University

[2]John Jay College of Criminal Justice

[3]New York University

[4]Microsoft Research

Correspondence
Sharad Goel, Stanford University
Email: scgoel@stanford.edu

Funding information
N/A

Judges, doctors, and managers are among those decision makers who must often choose a course of action under limited time, with limited knowledge, and without the aid of a computer. Because data-driven methods typically outperform unaided judgments, resource-constrained practitioners can benefit from simple, statistically derived rules that can be applied mentally. In this work, we formalize longstanding observations about the efficacy of improper linear models to construct accurate yet easily memorized rules. To test the performance of this approach, we conduct a large-scale evaluation in 23 domains and focus in depth on one: judicial decisions to release or detain defendants while they await trial. In these domains, we find that simple rules rival the accuracy of complex prediction models that base decisions on considerably more information. Further, comparing to unaided judicial decisions, we find that simple rules substantially outperform the human experts. To conclude, we present an analytical framework that sheds light on why simple rules perform as well as they do.

KEYWORDS
Heuristics, judgment and decision making, policy evaluation, sensitivity analysis

Abbreviations: FTA, failure to appear; RoR, release on recognizance.

All authors contributed equally.

1 | INTRODUCTION

In field settings, decision makers often choose a course of action based on experience and intuition rather than on statistical analysis (Klein, 2017). This includes doctors classifying patients based on their symptoms (McDonald, 1996), judges setting bail amounts (Dhami, 2003) or making parole decisions (Danziger et al., 2011), and managers determining which ventures will succeed (Åstebro and Elhedhli, 2006) or which customers to target (Wübben and Wangenheim, 2008). Despite the prevalence of this approach, a large body of work shows that in many domains intuitive inferences are inferior to those based on statistical models (Meehl, 1954; Dawes, 1979; Dawes et al., 1989; Camerer and Johnson, 1997; Tetlock, 2005; Kleinberg et al., 2015, 2017).

In this work, we generalize from research on improper linear models (Einhorn and Hogarth, 1975; Green, 1977; Dawes, 1979; Gigerenzer and Goldstein, 1996; Waller and Jones, 2011) to suggest a straightforward method for constructing simple yet accurate decision rules. This *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. Fast rules that can be applied mentally reduce transaction costs, encouraging persistent use. In medicine, frugal rules require fewer tests, which saves time, money, and, in the case of triage situations, lives (Marewski and Gigerenzer, 2012). The clarity of simple rules engenders trust from users, providing insight into how systems work and exposing where models may be improved (Gleicher, 2016; Sull and Eisenhardt, 2015). Clarity can even become a legal requirement when society demands to know how algorithmic decisions are being made (Goodman and Flaxman, 2016; Corbett-Davies et al., 2017).

We evaluate the efficacy of this approach on 22 datasets from the UCI Machine Learning repository, and show that in many cases simple rules are competitive with state-of-the-art machine learning algorithms. To illustrate in detail the value of simple rules, we present a case study of judicial decisions for pretrial release. Based on an analysis of over 100,000 cases, we show that simple rules substantially improve upon the efficiency and equity of unaided judicial decisions. In particular, we estimate that judges can detain one-third fewer defendants while simultaneously increasing the number that appear at their court dates.[1] In the judicial context, as in many policy settings, it is statistically challenging to evaluate decision rules based solely on historical data. The key difficulty is that one cannot observe what would have happened under an alternative course of action. What would have happened, for example, if one released a defendant who in reality was detained? We address this issue by first estimating the relevant counterfactual outcomes, and then assessing the sensitivity of our estimates to unobserved confounding, generalizing the technique of Rosenbaum and Rubin (1983a).

Our results add to a growing literature on *interpretable machine learning*. In addition to methods for better understanding complex machine learning models and data structures (Kim et al., 2015; Ribeiro et al., 2016), several methods have been introduced to construct interpretable decision rules, similar to the simple decision rules we discuss here. For example, Van Belle et al. (2012) use convex optimization to build interval coded scoring models for survival analysis. More general methods for constructing interpretable decision rules have been recently proposed, including supersparse linear integer models (SLIM) (Ustun and Rudin, 2016), Bayesian rule lists (Wang and Rudin, 2015), and interpretable decision sets (Lakkaraju et al., 2016). These methods all produce rules that are easy to interpret and to apply but the methods differ considerably on the ease of rule creation. As an important practical consideration, the method we investigate here can be carried out by a practitioner with basic knowledge of statistics using popular open-source software.

[1] Kleinberg et al. (2017) recently and independently proposed using machine learning models to assist judicial decisions, but they do not consider simple rules.

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

We begin by presenting a simple method—which we call *select-regress-and-round*—for constructing simple decision rules. This procedure generalizes ideas that appear throughout the judgment and decision-making literature on improper linear scoring rules, and formalizes heuristics used by practitioners in creating decision aids.

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 pending trial), 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 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. 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 cross-validation and 1,000 values of λ.

The select-regress-and-round method for rule construction extends research on unit-weighted linear models by incorporating feature selection and by adopting more general integer weights to generate a richer family of rules, the accuracy of which we examine in the next section.

3 | EVALUATING THE EFFICACY OF SIMPLE RULES

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

TABLE 1 Summary of UCI datasets. For each domain, we report the name of the dataset, the number of rows and inputs (columns excluding the class label) of the original dataset, number of complete rows with no missing data, number of features after discretizing continuous variables and expanding categorical variables to binary variables, and the proportion of the plurality class.

Domain	Instances	Features	Complete instances	Binarized features	Prop. majority
1. adult	32,561	14	30,162	93	25
2. annealing	798	38	798	56	76
3. audiology-std	200	70	190	55	24
4. bank	41,188	20	41,188	59	11
5. bankruptcy	250	6	250	13	43
6. car	1,728	6	1,728	16	70
7. chess-krvk	28,056	6	28,056	38	10
8. chess-krvkp	3,196	36	3,196	37	52
9. congress-voting	435	16	232	17	53
10. contrac	1,473	9	1,473	25	43
11. credit-approval	690	15	653	38	45
12. ctg	2,126	38	2,126	54	78
13. cylinder-bands	541	39	277	51	64
14. dermatology	366	34	358	38	31
15. german_credit	1,000	20	1,000	59	70
16. heart-cleveland	303	13	303	26	46
17. ilpd	583	10	579	11	72
18. mammo	961	5	830	21	49
19. mushroom	8,124	22	5,644	76	38
20. aus_credit	690	14	690	23	44
21. wine	178	13	178	14	40
22. wine_qual	6,497	12	6,497	13	63

criteria: (1) the dataset involves a binary classification (as opposed to a regression) problem;[2] (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 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. Table 1 summarizes the domains. On each of the 22 datasets we analyze here, we construct simple rules for a range of the number of features $k \in \{1, \ldots, 10\}$ and the magnitude of the weights $M \in \{1, 2, 3\}$.

[2] 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.

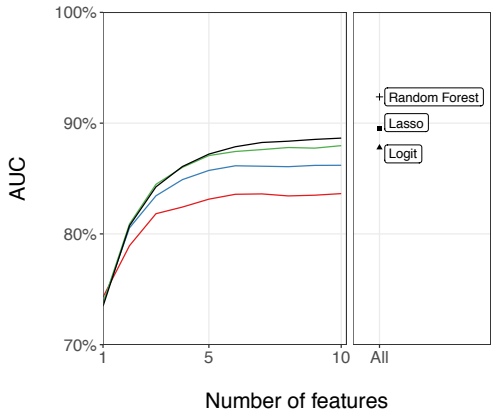

FIGURE 1 Performance of simple and complex rules. Performance is measured in terms of mean cross-validated AUC over all 22 datasets. The black line represents simple models with no rounding, and the green, blue, and red lines represent simple models rounding coefficients to $[-3, 3]$, $[-2, 2]$, and $[-1, 1]$, respectively. The simple models can predict with up to 10 features. The number of "all" features used by random forest, lasso, and logistic regression varied by domain, with an average of 38.

3.1 | Benchmarking to complex prediction models

We benchmark the performance of our simple rules against three standard statistical models: logistic regression, L^1-regularized logistic regression, and random forest. Random forest, in particular, is considered to be one of the best off-the-shelf classification algorithms in machine learning (Fernández-Delgado et al., 2014; Kleinberg et al., 2017). 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 nested 10-fold cross-validation 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 the features in a domain, which number between 11 and 93 with a mean of 38.

Figure 1 shows model performance—measured in terms of mean cross-validated 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 cross-validation. 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 non-linear 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.

3.2 | 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 method, and supersparse linear integer models, abbreviated SLIM (Ustun and Rudin, 2016), is the leading instantiation of that approach.

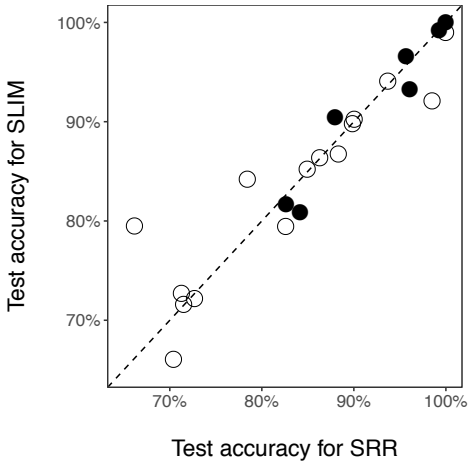

FIGURE 2 Comparing binary classification accuracy for select-regress-and-round (SRR) and SLIM on 22 UCI datasets. Solid dots are cases in which SLIM successfully found an optimal integer solution, while the open circles are cases in which the time limit of 6 hours was exceeded.

We compare SLIM to select-regress-and-round. SLIM is known to work best when the features are discrete. 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 (2016) we set a time limit for SLIM; a 10-minute limit is set in the original paper, but we allow up to 6 hours of computation per model. For 7 of the 22 datasets, SLIM found an integer-optimal solution within the time limit, and returned approximate solutions in the remaining 15 cases.

Figure 2 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 hold for AUC; accuracy is computed out-of-sample via 10-fold cross-validation. 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 7 cases where SLIM found integer-optimal solutions, performance is nearly identical to the simple select-regress-and-round method.

In terms of classification accuracy, select-regress-and-round generates rules that are 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, accordingly easing adoption for practitioners.

4 | CASE STUDY: PRETRIAL RELEASE DECISIONS

To illustrate the value—and challenges—of applying simple decision rules in practice, we now turn to the domain of pretrial release determinations and present an extended case study. 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 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 directly observe the effects of hypothetical decision rules. Unlike the class of prediction problems discussed in Section 3, outcomes in this domain are affected by a judge's decisions, and one only observes the outcomes that result from those decisions. 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 (Dudík et al., 2011)—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 (1983a) for assessing the sensitivity of estimated causal effects to unobserved confounding.

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.

4.1 | 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 fit an L^1-regularized logistic (lasso) regression and random forest, using the procedures described in Section 3.1, to estimate the likelihood an individual fails to appear at any of his subsequent court dates. We fit these models on all available information about the case and the defendant, excluding race.[3] The fitted models let 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 estimated flight risk is below 20%.

We now construct a family of simple rules for making release decisions. We begin by applying select-regress-and-

[3]We exclude race from the presented results due to legal and policy concerns with basing decisions on protected attributes (Corbett-Davies et al., 2017). We note, however, that including race does not significantly affect performance.

TABLE 2 A simple rule for estimating flight risk with 10 features. A defendant's flight risk is obtained by summing the corresponding scores for the features that apply to the case.

Feature	Score	Feature	Score
$18 \leq$ age < 21	2	1 prior FTA	2
$21 \leq$ age < 26	2	2 prior FTAs	3
$26 \leq$ age < 31	1	3 or more prior FTAs	3
Unstable housing	3	Top charge is DUI	-3
Misdemeanor	1	Charged with assault	-1

round on all available features, as described in Section 2. A resulting rule using 10 features with integer coefficients between -3 and 3 is presented in Table 2. Unsurprisingly, missing court appearances in the past is a strong indicator of flight risk, and an individual's risk also declines with age, in line with conventional wisdom. The rule in Table 2, however, may be inappropriate for implementation given that some features and their associated scores could be challenged as undesirable. For example, defendants with unstable housing are rated higher risk, which may by statistically true but which could lead to adverse outcomes for poorer defendants.

To mitigate such practical challenges, we fit the model in the second step of select-regress-and-round using only age and prior history of failing to appear. These two factors represent the majority of features included in the original rule, and are generally viewed as legitimate risk factors to consider. In this case, we can think of the "select" step in the select-regress-and-round strategy as incorporating both human and machine judgment. 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 no prior 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 $[-3, 3]$; specifically, we multiply each coefficient by the constant

$$\frac{3}{\max\left(|\beta_1^{\text{priors}}|, \ldots, |\beta_{4+}^{\text{priors}}|, |\beta_{18-20}^{\text{age}}|, \ldots, |\beta_{46-50}^{\text{age}}|\right)}.$$

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

Figure 3 shows the result of this procedure. For any defendant, a risk score can be computed by summing the relevant terms in the table. 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 3.5; a graphical representation of such a binary decision rule is also shown in the figure.

Feature	Score	Feature	Score
$18 \leq$ age < 21	3	no prior FTAs	0
$21 \leq$ age < 31	2	1 prior FTA	2
$31 \leq$ age < 51	1	2 or more prior FTAs	3
$51 \leq$ age	0		

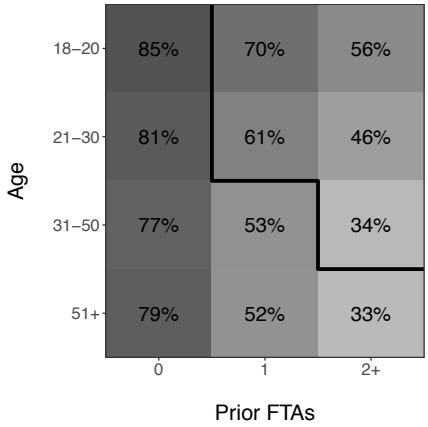

FIGURE 3 A simple rule for estimating flight risk, where a defendant's risk is obtained by summing the appropriate scores for age and prior history of failing to appear (FTA). Also shown is a graphical representation of this rule, based on setting a release threshold of 3.5. Groups to the left of the solid black line in the grid are those that would be released under the rule. For comparison, the shading and numbers in the grid show the proportion of defendants that are currently RoR'd in each group.

4.2 | Policy evaluation

The AUC is a useful general measure of performance, and hence the metric we consider when evaluating the 22 UCI datasets in Section 3. But in applied settings it is often necessary to directly measure the costs and benefits of any given rule. We do that here by assessing decision rules for pretrial release on two key dimensions: (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 see what actions would have been recommended.[4] For example, if defendants are released if and only if their risk score is below 3.5, 79% would be RoR'd; under this rule, bail would be required of only two-thirds as many defendants relative to the status quo. 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* (Dudík et al., 2011) is a specific instance of the fundamental problem of causal inference.

To rigorously describe the estimation problem and our approach, we introduce some notation. For concreteness, we frame our methodology in terms of the pretrial release example, but the ideas presented here are common to many

[4] 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.

TABLE 3 An illustrative example of response surface modeling for offline policy evaluation. 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%

policy decisions. 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 (Hill, 2012).[5] 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 3 illustrates this method for a hypothetical example.

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

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

This ignorability assumption is typically unavoidable, and is similarly required for methods based on propensity

[5] We investigated two alternative approaches—inverse propensity weighting (Rosenbaum and Rubin, 1983b, 1984) and doubly robust estimation (Cassel et al., 1976; Robins et al., 1994; Robins and Rotnitzky, 1995; Kang and Schafer, 2007; Dudík et al., 2011)—and found qualitatively similar results.

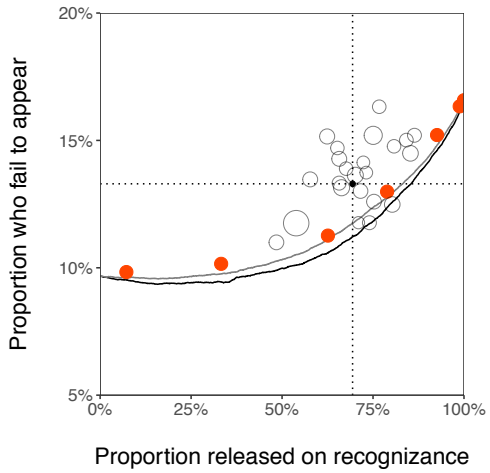

FIGURE 4 Evaluation of simple and complex decision rules. Each point on the solid lines corresponds to decision rules derived from a random forest (gray) or lasso (black) risk model with varying thresholds for release. The red points correspond to policies based on simple risk score using all possible release thresholds. The simple rules perform nearly identically to the random forest models, and comparably to the lasso 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.

scores (Rosenbaum and Rubin, 1983b, 1984; Cassel et al., 1976; Robins et al., 1994; Robins and Rotnitzky, 1995; Kang and Schafer, 2007; Dudík et al., 2011). We examine this assumption in detail in Section 4.3, and find that our conclusions are robust under a common model of 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); we also include the defendant's race.[6] 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 4 shows estimated RoR and FTA rates for a variety of pretrial release rules. Points on the solid lines correspond to rules constructed via the lasso (black line) and random forest (gray line) models described above for various decision thresholds. The red points correspond to rules based on the simple scoring procedure in Figure 3, 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 state-of-the-art machine learning models (random forest and lasso regression) that

[6] Although it is legally problematic to use race when *making* decisions, its use is acceptable—and indeed often required—when *evaluating* decisions. 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 cross-validation 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 we empirically found lasso to yield better predictions in this case.

incorporate all 64 available 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 3.5, we expect to release 79% of defendants while achieving an FTA rate of 13%. Relative to the existing policy, following this rule would result in detaining one-third fewer defendants while also slightly decreasing the overall FTA rate—from 13.3% to 13.0%. 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 3 sheds light on this phenomenon. Each cell in the grid 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 3.5, 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 two or more prior FTAs is released at about the same 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%.[7] 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.

4.3 | 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 5 shows the results of this analysis for the simple decision rule described in Figure 3, 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 (1983a) for assessing the sensitivity of estimated causal effects to an unobserved binary covariate. We specifically tailor their approach to offline policy

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

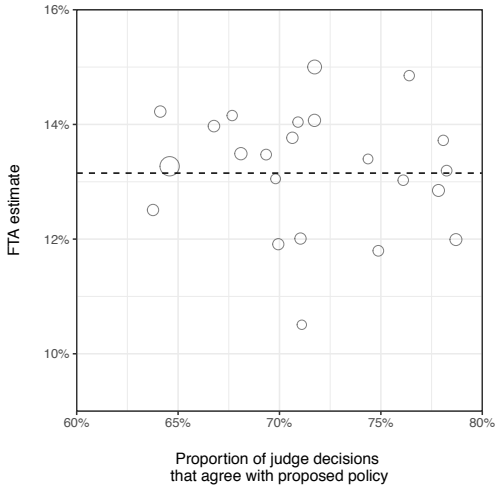

FIGURE 5 Robustness of estimated FTA rate. For the simple decision rule, FTA rate is estimated by separately applying response surface modeling to each judge's cases, where each point corresponds to a judge; the dashed horizontal line indicates the FTA rate of the decision rule estimated on the full 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%.

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. 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 (1983a) 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,

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

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

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

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

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

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

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

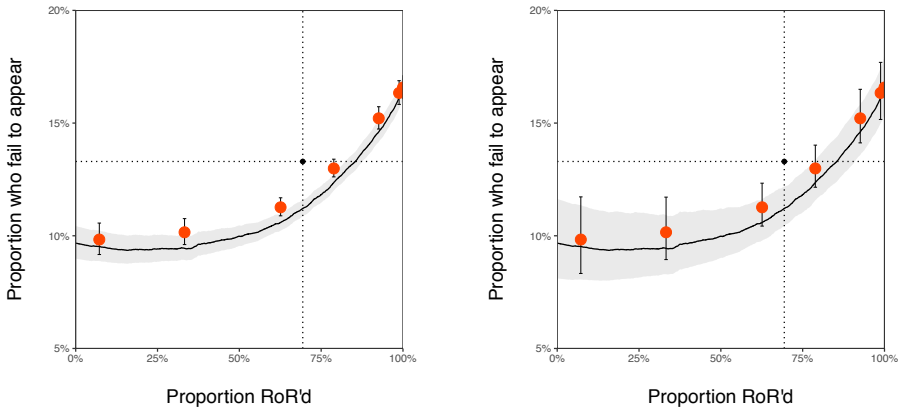

FIGURE 6 Sensitivity of FTA estimates to unobserved heterogeneity. The gray bands (for the complex rules using lasso) and the error bars (for the simple rules) indicate minimum and maximum FTA estimates for a variety of parameter settings. In the left-hand plot, we assume $\alpha = \log 2$ and consider 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.

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

Figure 6 shows the results of computing $\hat{V}^{\pi}_{\text{RR}}$ on our data in two parameter regimes. In the first (left-hand plot), we assume $\alpha = \log 2$ and consider all combinations of $p(u = 1) \in \{0.1, 0.2, \ldots, 0.9\}$, $\delta^{\text{RoR}} \in \{-\log 2, 0, \log 2\}$, 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 halve the odds a defendant fails to appear. For each complex policy (*i.e.*, one based on lasso), the gray bands show 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 potentially unobserved heterogeneity across defendants.

5 | THE ROBUSTNESS OF BINARY CLASSIFICATION

Why is it that simple rules often perform as well as the most sophisticated statistical methods? In part it is because binary classification accuracy is relatively robust to error in the underlying predictive model, an observation that we formalize in Proposition 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: $\text{N}(\mu_p, \sigma^2)$ and $\text{N}(\mu_n, \sigma^2)$, 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, Proposition 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.

Proposition 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 $Y_p \sim \mathrm{N}(\mu_p, \sigma^2)$ and $Y_n \sim \mathrm{N}(\mu_n, \sigma^2)$. Then, for $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$ 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

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

Proposition 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

[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.

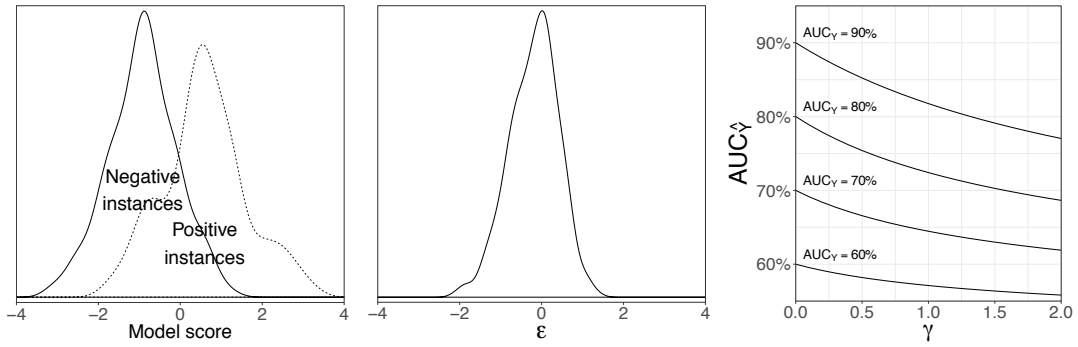

FIGURE 7 Theoretical analysis of simple rules. Left panel: empirical distribution of prediction scores, on the logit scale, for positive and negative instances of a UCI dataset (`heart-cleveland`), generated via an L^1-regularized logistic regression model. Center panel: empirical distribution of ϵ for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

expression for various parameter values. For example, the figure shows that for $\text{AUC}_\gamma = 90\%$ and $\gamma = 0.5$, we have $\text{AUC}_{\hat{\gamma}} = 85\%$. 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, Proposition 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 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.

6 | CONCLUSION

In this paper we extended research on improper linear models to propose a simple method for constructing simple rules. The rules take the form of an easily memorized point system. In 23 domains of varying size and complexity, the rules produced by the select-regress-and-round method rivaled the accuracy of regularized regression models while using only a fraction of the information. Specifically, rules based on only five features performed on par with logistic and lasso regression models that integrated 38 features in the average domain. In a detailed analysis of pretrial release decisions, the simple rules outperformed human judges and matched machine learning models incorporating 64 features.

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 (Ustun and Rudin, 2016; Wang and Rudin, 2015; Lakkaraju et al., 2016). While many of these rule construction methods offer great flexibility, they in turn require considerable computational resources and expertise to carry out. In contrast, the method we propose can easily be carried out by ordinary practitioners using popular open-source software. It has long been noted that statistical models tend to outperform unaided human judgments. To encourage practitioners to move beyond intuition, they should be provided with models that are not only easy to apply, but also easy to construct.

[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.

ACKNOWLEDGEMENTS

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

REFERENCES

Åstebro, T. and Elhedhli, S. (2006) The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, **52**, 395–409.

Camerer, C. F. and Johnson, E. J. (1997) The process-performance paradox in expert judgment. In *Research on judgment and decision making: Currents, connections, and controversies* (eds. W. M. Goldstein and R. M. Hogarth). Cambridge University Press.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017) Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. ACM.

Danziger, S., Levav, J. and Avnaim-Pesso, L. (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, **108**, 6889–6892.

Dawes, R. M. (1979) The robust beauty of improper linear models in decision making. *American Psychologist*, **34**, 571.

Dawes, R. M., Faust, D. and Meehl, P. E. (1989) Clinical versus actuarial judgment. *Science*, **243**, 1668–1674.

Dhami, M. K. (2003) Psychological models of professional decision making. *Psychological Science*, **14**, 175–180.

Dudík, M., Langford, J. and Li, L. (2011) Doubly robust policy evaluation and learning. *ICML*. URL: `http://arxiv.org/abs/1103.4601`.

Einhorn, H. J. and Hogarth, R. M. (1975) Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, **13**, 171–192.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014) Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res*, **15**, 3133–3181.

Gigerenzer, G. and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, **103**, 650.

Gleicher, M. (2016) A framework for considering comprehensibility in modeling. *Big Data*, **4**, 75–88.

Goodman, B. and Flaxman, S. (2016) EU regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*.

Green, B. F. (1977) Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research*, **12**, 263–287.

Hill, J. L. (2012) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*.

Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 523–539.

Kim, B., Shah, J. A. and Doshi-Velez, F. (2015) Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*, 2260–2268.

Klein, G. A. (2017) *Sources of power: How people make decisions*. MIT press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017) Human decisions and machine predictions. *The Quarterly Journal of Economics*, **133**, 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015) Prediction policy problems. *The American Economic Review*, **105**.

Lakkaraju, H., Bach, S. H. and Leskovec, J. (2016) Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*.

Marewski, J. N. and Gigerenzer, G. (2012) Heuristic decision making in medicine. *Dialogues Clin Neurosci*, **14**, 77–89.

McDonald, C. J. (1996) Medical heuristics: The silent adjudicators of clinical practice. *Annals of Internal Medicine*, **124**, 56–62.

Meehl, P. E. (1954) *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, 21–27.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, **89**, 846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 212–218.

— (1983b) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

— (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, **79**, 516–524.

Sull, D. and Eisenhardt, K. M. (2015) *Simple rules: How to thrive in a complex world.* Houghton Mifflin Harcourt.

Tetlock, P. (2005) *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Ustun, B. and Rudin, C. (2016) Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, **102**, 349–391.

Van Belle, V., Van Huffel, S., Suykens, J. and Boyd, S. (2012) Interval coded scoring systems for survival analysis. *Proceedings of the European Symposium on Artificial Neural Networks*.

Waller, N. and Jones, J. (2011) Investigating the performance of alternate regression weights by studying all possible criteria in regression models with a fixed set of predictors. *Psychometrika*, **76**, 410.

Wang, F. and Rudin, C. (2015) Falling rule lists. In *Artificial Intelligence and Statistics*, 1013–1022.

Wübben, M. and Wangenheim, F. V. (2008) Instant customer base analysis: Managerial heuristics often get it right. *Journal of Marketing*, **72**, 82–93.