Discover Artificial Intelligence

Case Study

Simple rules outperform machine learning for personnel selection: insights from the 3rd annual SIOP machine learning competition

Jason L. Harman[1,2] · Jaelle Scheuerman[2]

Received: 20 September 2022 / Accepted: 13 December 2022
Published online: 10 January 2023

Abstract
Machine learning (ML) algorithms are often assumed to be the most accurate way of producing predictive models despite problems with explainability and adverse impact. The 3rd annual Society for Industrial and Organizational Psychology Machine Learning Competition sought to find ML models for personnel selection that could balance the best of ML prediction with the constraint of minimizing selection bias based on race and gender. To test the possible advantages of simple rules over ML algorithms, we entered a simple and explainable rule-based model inspired by recent advances in model comparison. This simple model outperformed most ML models entered and was comparable to the top performers while retaining positive qualities such as explainability and transparency.

1 Introduction

Interest and use of artificial intelligence (AI) and machine learning (ML) applications within multiple areas including medicine, public health, and employee selection has been growing. There is good reason for this as the availability of big data and advances in both computing power and ML techniques have made highly complex prediction models with purportedly high levels of accuracy readily accessible. There are however several downsides to reliance on AI/ML models, particularly as their use expands to diverse domains. AI technologies, such as deep learning, can be a-theoretic black boxes, meaning they produce predictions based only on given data to optimize the accuracy of some output with no theoretical rationale for predictions and no explainable process. A large body of research has shown that explainability is a key component to engendering trust. Without the public's trust, AI/ML based decisions are susceptible to backlash [1]. Additionally, AI/ML algorithms have been shown to make differential decisions between individuals of different race and gender [2–4]. This adverse impact poses a problem for both the ethical and practical use of such decision tools as well as the public trust of such tools. Beyond these two major issues inherent to AI/ML, there is additional evidence that advances in accuracy of decision making from AI/ML are overstated if not absent when compared to simpler cognitively inspired decision models that are both explainable and less amenable to adverse impact [5–9]. To test this last assertion, we created a simple rule-based prediction model inspired from a recent advancement in model comparison in the field of judgment and decision making [10]. The simple and explainable model was entered into the 3rd Society for Industrial and Organizational Psychology (SIOP) Machine Learning competition [11] and outperformed 51 of 60 ML models entered with results comparable to the best performing ML models.

The third in a series of Machine Learning competitions at the Society for Industrial and Organizational Psychology (SIOP) focused on personnel selection with a special emphasis on adverse impact (under selection of protected groups,

✉ Jason L. Harman, jharman@lsu.edu | [1]Louisiana State University, Baton Rouge, LA, USA. [2]U.S. Naval Research Laboratory, Stennis Space Center, MS, USA.

e.g. ethnic minorities, women, etc.). The data for the competition consisted of around 50,000 employee selection records from Walmart. The ultimate goal for competing models was to create an algorithm that effectively predicted performance and turnover while balancing adverse impact. The organizers decided to compare models on a single quantitative score that combined prediction accuracy (top performers in addition to retained employees) and unfairness (an adverse impact ratio). The data set included over 100 possible predictors grouped into three different types: pre-employment tests (judgments of hypothetical work scenarios), personality tests, and biographical/work history data. All item content was removed from the data, leaving only numerical responses and an example of one item from each category. All personal information and demographics were removed and a contrived variable called protected group was created as a surrogate for ethnicity, gender, or other protected classes as part of the fairness/adverse impact criterion of the competition. Seven possible dependent variables were included in the data; five performance ratings and two dichotomous variables labeling employees as high performers (or not) and whether they had been retained.

Data from 44,102 individuals was provided as a training data set and two groups of 2250 were retained as test datasets for a public leaderboard and a private leaderboard. Participation was open to anyone. Following registration, participating groups were given the training data with which they built their models with the goal of selecting the top 50% of the employees on a combined measure of performance and fairness. The evaluative criteria was accuracy—unfairness where accuracy was quantified as $(\%\ of\ top\ performers\ \times 25) + (\%\ of\ retained\ employees\ \times 25) + (\%\ of\ top\ performers\ who\ were\ also\ retained\ \times 50)$. Unfairness was quantified as ABS (1-AIRatio) ×100 where the AI ratio was the ratio of the percentage of protected group members in the predicted group over the percentage of protected group members in the data set. After creating a model with the training data set, groups were allowed to enter their model into the public leaderboard with real time scores (up to five test submissions were allowed) seeing their relative performance each time, allowing for model adjustments. At the conclusion of the competition, all models were evaluated using the private leaderboard data set and results were announced at the annual SIOP conference.

To explore whether simple rules could match or outperform ML models in the SIOP ML competition, we adapted a model selection procedure from Harman et al. [10]. The procedure from Harman et al. was designed to compare models in a competition across multiple evaluative criteria as opposed to the typical single criteria of prediction accuracy. This procedure involves creating multiple quantitative criteria (such as prediction error, falsifiability, and generalizability) on which each model is scored and then calculating a total score based on the relative performance of each model on each criterion. To calculate an overall score that can be decomposed into multiple evaluative criteria, Harman et al. imported the concept of Borda rule voting from the field of computational social choice. Put simply, this procedure rank orders each model on each individual criterion from best to worst (see Fig. 1 for an example). Then each model receives a score (Borda value) for each criterion based on their rank with higher ranks getting more points. Finally, the scores for each model are summed across criteria into a Borda count. This Borda count is a single quantitative score composed of the individual relative rankings for each criterion.

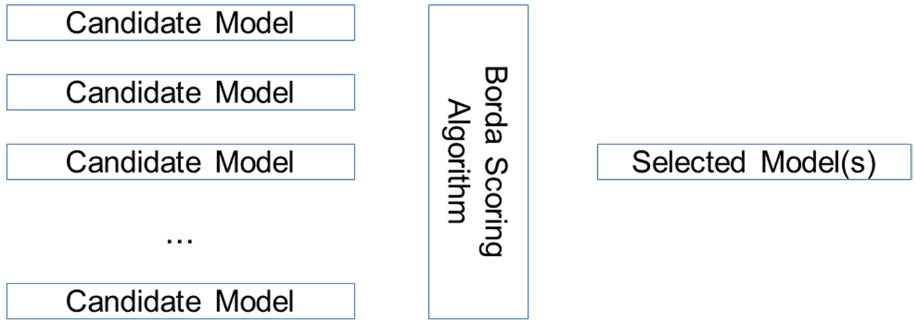
In adapting this procedure for the SIOP ML competition, we treated each person in the data set as a candidate model and treated each measure in the data as a model criterion. Because there was no content tied to the data, we had no theoretical grounding for selecting criteria or determining in which direction each measure should be ranked and had to rely on statistical analysis. We created a dependent variable by combining performance [0–1] and retention [0–1] into a pseudo continuous variable [0–2] and regressed each predictive variable onto this dependent variable (DV). With these regressions we were able to identify how each predictive variable related to the outcome and the direction of that relationship. For each prediction variable, we simply rank ordered the scores on each variable, assigned Borda values for each variable, and perform a Borda count (summing a person's Borda values across all variables), selecting the top 50% of participants based on total Borda count.

In summary, our entry "Models for All" has three steps or rules: (1) identify relevant variables and predictive direction, (2) rank order scores on each variable according to the direction identified in step 1, and (3) sum Borda counts for each candidate, choosing the top 50%.

RULE 1: Select (theoretically or statistically) predictive variables. *Because all content was removed from the data set, we ran simple regressions for each variable onto a composite DV to determine the direction of relationship between predictors and performance.*

RULE 2: Rank Order candidates' performance on each variable. *For each variable, candidates were rank ordered and given a Borda value based on their rank.*

RULE 3: Choose Y candidates based on Borda counts across X variables. *Total Borda scores were used to choose the top 50%.*



	V1	V2	V3	V4	V5
C 1	3	1	2	2	1
C 2	1	2	1	1	1
C 3	2	1	1	1	1
C 4	4	2	1	1	2
C 5	5	1	3	1	2

	V1	V2	V3	V4	V5	Borda Total
C 1	3 (3)	1 (2)	2 (2)	2 (1)	1 (2)	**10**
C 2	1 (5)	2 (1)	1 (3)	1 (2)	1 (2)	**13**
C 3	2 (4)	1 (2)	1 (3)	1 (2)	1 (2)	**13**
C 4	4 (2)	2 (1)	1 (3)	1 (2)	2 (1)	**9**
C 5	5 (1)	1 (2)	3 (1)	1 (2)	2 (1)	**7**

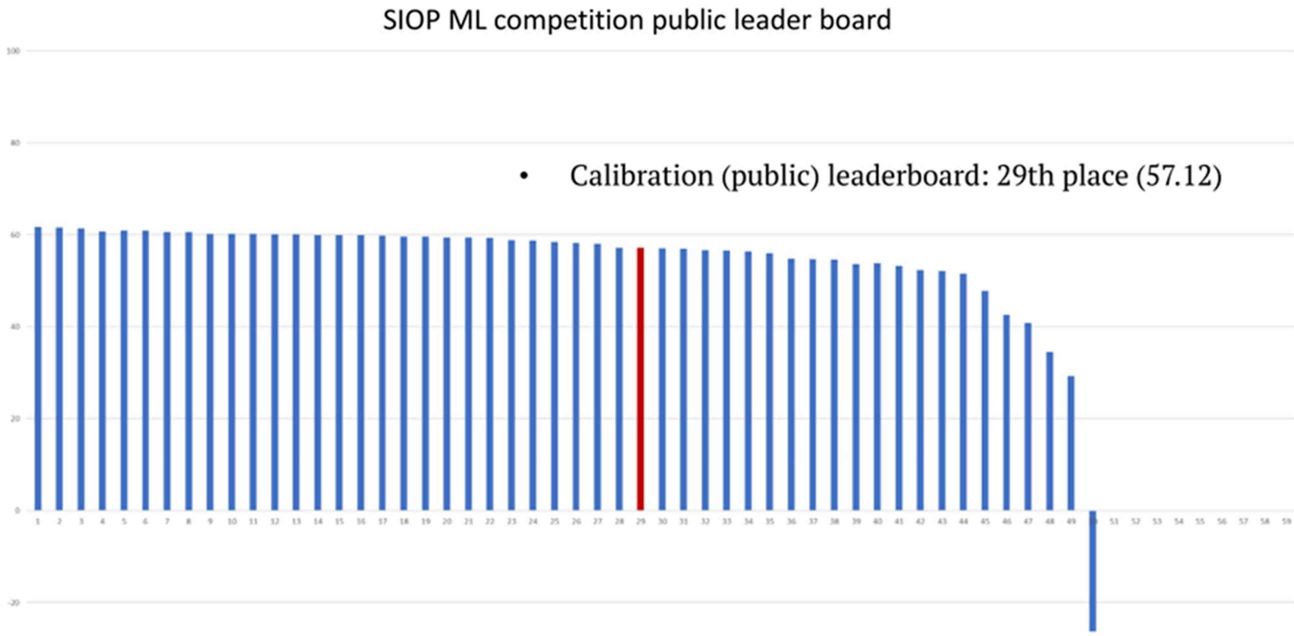
Fig. 1 Hypothetical competition rankings for two modeling competitions. This figure shows the architecture of the multi-criteria selection procedure followed by hypothetical rankings of 5 models (or candidates) across 5 criteria (C1–C5) with Borda values (in parentheses) and total Borda scores shown in the lower image

2 Results and discussion

Over 200 teams registered for the competition (53% from universities and 47% from industry) and were allowed to build models with the calibration data and submit up to 5 iterations to the public leader board, seeing live results from these submissions (we submitted only 1 iteration) [11]. Taken together there were over 1500 submissions to the public leaderboard and 60 teams submitted a final model for the competition. To our knowledge, all but our entry were some form of ML algorithm such as random forest classifiers.

When submitted to the public leaderboard of the competition "Models for All" obtained a combined criterion score of 57.12 ranking 29th, middle of the pack, compared to the 60 models entered. When generalized to the private leaderboard data set, "Models for All" score improved to 59.7 ranking 9th (see Fig. 2). Compared to the other 60 models entered, this simple and explainable model outperformed a majority of the models entered and scored relatively closely with the winning model (score = 62.53). That "Models for All" was one of the few models whose scores increased from prediction to competition highlights one of the advantages of simple rules which is that they avoid issues of overfitting and tend to be more generalizable (*see 10*). We note that adverse impact was not a direct consideration of "Models for All", but resulted as a natural byproduct of the prediction process. Because competition score breakdowns were not provided by the organizers, we cannot directly quantify the adverse impact of predictions made by "Models for All' but can safely assume that the number is at least comparable with other competitors.

While ML and AI technologies have added greatly to society and productivity, moving forward it is important to evaluate their drawbacks with clear eyes and to avoid the assumption that predictions made via ML are necessarily the best predictions that can be garnered from data. In addition to the results presented here which illustrate that simple models can predict just as well or better than modern ML models, there are more possible advantages of simple models



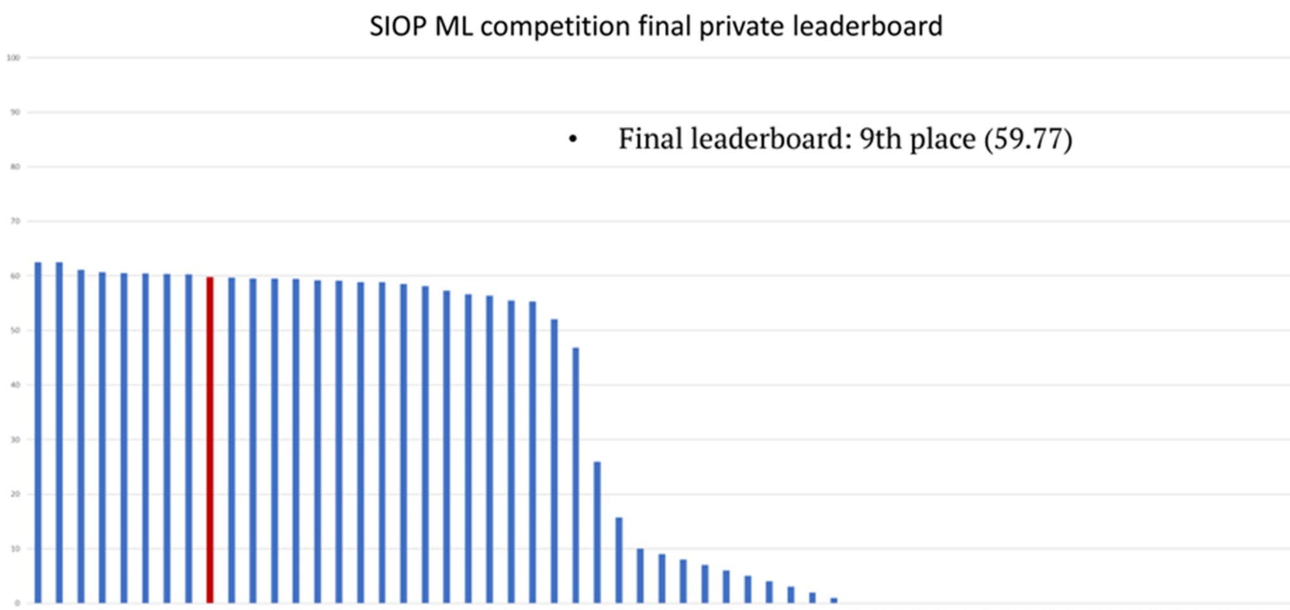


Fig. 2 Results from the 2020/21 SIOP Machine Learning competition. The final ranking of "Models for All" is shown in the public leaderboard (above) and the final private leaderboard (below)

over ML. Primarily, simple models such as "Models for All" are both explainable and transparent, a common criticism of modern ML technologies [1]. Because decisions are based on simple Borda counts, "Models for All" makes explaining why certain applicants are selected over others straightforward, with the components of the Borda scores readily available and comparable. In other words, the decision criterion (Borda count) is directly decomposable into ranks on all criteria making decisions immediately explainable. Additionally, if protected groups were disadvantaged in the selection decisions, the same process can be used to explore the specific measures in which these disadvantages appear. The multi-criterion evaluation procedure outlined by Harman et al. [10] which was the inspiration for our model entry, has the advantage of allowing criteria such as protected status to be directly evaluated in predictive models [12]. For example, facial recognition algorithms could be directly evaluated not only on the proportion of correct identifications, but on whether a model predicts differentially based on race or gender.

Author contributions JLH: conceptualization, visualization, funding acquisition, supervision, writing—original draft. JLH, JS: methodology, investigation, project administration, Writing—review and editing. Both authors read and approved the final manuscript.

Funding Louisiana Board of Regents (BOR) ATLAS Grant AWD-AM210614 (JLH).

Data availability All data, code from the top models, and materials from the 2021 SIOP Machine Learning Competition is available at https://github.com/izk8/2021_SIOP_Machine_Learning_Winners code for models for all can be found at https://github.com/jaelle/siop_ml_competition_2021

Declarations

Competing interests Authors declare that they have no competing interests.

References

1. Shadbolt N. A matter of trust. IEEE Intell Syst. 2002;17:2–3.
2. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. ProPublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
3. Benjamin R. Assessing risk, automating racism. Science. 2019;366(6464):421–2.
4. Kirchner L, Goldstein M. How automated background checks freeze out renters. The New York Times. 2020. https://www.nytimes.com/2020/05/28/business/renters-background-checks.html. Accessed 15 Aug 2022.
5. Martignon L, Katsikopoulos KV, Woike JK. Categorization with limited resources: a family of simple heuristics. J Math Psychol. 2008;52:352–61.
6. Frederik J, Martijn M. The new dot com bubble is here: it's called online advertising. The Correspondent. 2019. https://thecorrespondent.com/100/the-new-dot-com-bubble-is-here-its-called-online-advertising/13228924500-22d5fd24. Accessed 15 Aug 2022.
7. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol. 2011;62:451–82.
8. Katsikopoulos KV. How to model it? Review of "cognitive modeling" (J. R. Busemeyer and A. Diederich). J Math Psychol. 2011;55(2):198–201.
9. Katsikopoulos KV, Durbach IN, Stewart TJ. When should we use simple decision models? a synthesis of various research strands. Omega—Int J Manag Sci. 2018;81:17–25.
10. Harman JL, Yu M, Konstantinidis E, Gonzalez C. How to use a multi-criteria comparison procedure to improve modeling competitions. Psychol Rev. 2021;128:995–1005.
11. Koenig N, Thompson I. The 2020–2021 SIOP machine learning competition. Presented at the 36th annual Society for Industrial and Organizational Psychology conference in New Orleans, LA. https://github.com/izk8/2021_SIOP_Machine_Learning_Winners. Accessed 15 Aug 2022.
12. Harman JL, Scheuerman J. Multi-criteria comparison as a method of advancing knowledge-guided machine learning. Proceedings of the Association for the Advancement of Artificial Intelligence 2022 Fall Symposium on Knowledge Guided Machine Learning. 2022.