Machine learning for public policy: Do we need to sacrifice accuracy to make models fair?

Kit T. Rodolfa,[1] Hemank Lamba,[1] Rayid Ghani[1*]

[1]Machine Learning Department and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[*]To whom correspondence should be addressed; E-mail: rayid@cmu.edu.

One Sentence Summary: Trade-offs assumed to be inherent in machine learning may be small in practice, making reducing disparities more practical.

Growing applications of machine learning in policy settings have raised concern for fairness implications, especially for racial minorities, but little work has studied the practical trade-offs between fairness and accuracy in real-world settings. This empirical study fills this gap by investigating the accuracy cost of mitigating disparities across several policy settings, focusing on the common context of using machine learning to inform benefit allocation in resource-constrained programs across education, mental health, criminal justice, and housing safety. In each setting, explicitly focusing on achieving equity and using our proposed post-hoc disparity mitigation methods, fairness

was substantially improved without sacrificing accuracy, challenging the commonly held assumption that reducing disparities either requires accepting an appreciable drop in accuracy or the development of novel, complex methods.

There has been a rapid growth in the use of machine learning for applications with extensive impact on society, such as informing bail determination decisions (*1–3*), hiring (*4*), healthcare delivery (*5, 6*), and social service interventions (*7–9*). These wide-reaching applications have been met with heightened concerns about their potential for introducing or amplifying inequities, especially for racial minorities and economically disadvantaged individuals, motivating exploration of a range of potential sources and mitigation strategies for biases, including in the underlying data (*10*), labels (*5*), model training (*11–13*), and post-modeling adjustments to scores (*14, 15*). A common underpinning of much of this work is the assumption that trade-offs between equity and accuracy may necessitate complex methods or difficult policy choices (*16–19*), however little work to date has explicitly evaluated the magnitude (or existence) of these trade-offs in real-world problems.

Our study focuses on testing that assumption in resource allocation problems across several public policy domains. Organizations with limited resources are often only able to intervene and allocate benefits to a relatively small number of individuals with need, presenting a "top k" optimization problem where model accuracy is judged by $precision$ (also known as positive predictive value) among the k highest-scoring individuals. In such settings, we (*15*) and others (*14*) have argued that recall (also known as sensitivity) disparities are often an appropriate equity metric, reflecting a concept of "equality of opportunity." In a recent case study (*15*), we found that explicitly focusing on achieving equity and using subgroup-specific score thresholds as a post-hoc disparity mitigation method improved the equity of predictions with only a very modest decrease in accuracy. The empirical work here extends that surprising result to several policy contexts and modeling choices, suggesting that trade-offs between fairness and effectiveness may be negligible in practice, making improvement in equity easier and more practical across a wide range of applications. We come to this conclusion using a variety of projects we undertook over the past few years with government agencies across criminal justice, mental health, housing safety, and education:

Inmate Mental Health Seeking to break the cycle of incarceration for individuals with untreated mental health conditions, Johnson County, KS, partnered with us to prioritize limited resources for mental health outreach on individuals at risk of a future jail booking. We developed a model of risk for a booking in the next year, focusing on identifying 500 individuals for outreach in a 4-month window based on the resources available to the program. Disparities on race and ethnicity are particularly salient in the criminal justice context, and we focused on this attribute in our bias analyses.

Housing Safety The Code Enforcement Office in San Jose, CA is tasked with protecting occupants of properties with multiple units (such as apartment buildings) by conducting safety inspections, but doesn't have sufficient staffing to inspect all 4,500 properties every year. Using internal data supplied by the program, we developed a model for the risk that a serious violation would be found if a given property were prioritized for inspection. We focused on disparities between housing units in higher- and lower-income neighborhoods (median income above or below $55,000) where considerable disparities were observed in our initial models favoring higher-income areas.

Student Outcomes El Salvador's Ministry of Education seeks to support students to reduce the country's substantial dropout rates (recently as high as 29% in some years), but the budget for these programs is insufficient to reach every student. Student-level data was provided by the Ministry to develop a model of students at risk of dropping out, and our analysis here focuses on identifying the 10,000 highest-risk students in the state of San Salvador. The Ministry of Education was concerned with potential disparities in gender, age relative to grade level, and urban-rural divide. Initial analyses found large disparities with "over-age" students (those at least 2 standard deviations above the mean of their grade level), which we focus on in the present study.

Since the projects above used confidential and sensitive data and were done under data use agreements, we are not able to make that data publicly available. For our work to be easily reproducible, we include a fourth problem in this study where the data is available publicly:

Education Crowdfunding The non-profit DonorsChoose helps alleviate school funding shortages by providing a crowdfunding platform for teachers to post requests for their classroom needs. Here, we make use of a dataset DonorsChoose made publicly available in 2014 and posit an effort to assist projects at risk of going unfunded (for instance, providing a review and consultation) capable of helping 1,000 projects in a 2-month window. Reflecting the platform's goal of helping schools and teachers most at need, we focus in this context on disparities across school poverty levels (65% free/reduced lunch vs others).

Modeling methods are described in the Supplementary Materials, and the following results focus on using group-specific score thresholds to mitigate disparities observed in these models' predictions (see Figs. 1A and 1B for an illustration of these methods).

We started with the assumption, based on existing theoretical work, that there is a trade-off between "accuracy" and "fairness." As an initial experiment, we explored how model "accuracy" changes upon adjusting for disparities in the Inmate Mental Health setting using a single temporal validation split (with test set outcomes spanning 4/2018 to 4/2019). In Figs. 1C and 1D, each pair of points is a model specification: with results obtained without adjusting for equity in blue and those with the equity adjustment in orange. The x-axis shows the precision (positive predictive value) of each model on the 500 selected individuals and the y-axis shows the recall disparity between white and non-white individuals. Fig. 1C shows all models considered, while Fig. 1D provides a more detailed view of the better-performing models that might reasonably be selected. All of the unadjusted models had significant disparities, indicating that bias audits and model selection by themselves are insufficient for achieving fairness. However, applying our proposed bias mitigation method to adjust for disparities, we find little evidence of a fairness-accuracy trade-off: overall, the mean change in precision after adjustment is -0.0006 (std: 0.0087); Fig. 1E shows the distribution of these shifts.

We also investigated creating a composite model (based on the work of Dwork and colleagues (*13*)) by choosing the best-performing model for each subgroup and using the recall-equalizing set of individuals (maintaining a total list size of 500) from each subgroup-optimized model. The performance of this composite on the subsequent test set is shown as a red diamond in Figs. 1C and 1D. On both fairness and accuracy-related metrics, this composite appears to perform competitively with other models, but doesn't stand out as doing particularly well on either metric.

These initial results suggested disparity mitigation could be integrated into the process of model selection, and we next sought to compare strategies for doing so (summarized in Supplemental Table S1 and described in more detail in the Materials and Methods). In the strategy labeled "Mitigated - Single Model," model specifications are compared based on their $precision@k$ after applying group-specific thresholds to mitigate disparities and the chosen model is evaluated on new data. Selecting a model without regard to fairness and applying disparity mitigation only to the chosen model showed no practical difference in performance on either fairness or accuracy metrics in any of our analyses (see discussion in Supplementary Materials). As a baseline, the "Unmitigated" approach performs model selection without accounting for any disparities.

Additionally, the "Mitigated - Composite Model" strategy chooses the best-performing model on each subgroup at the model selection stage and picks recall-balancing thresholds across these. We also explored composite models created from fully-decoupled models trained only on subgroup-specific examples (also suggested in (*13*)), but saw no difference in performance from the composite method described here in these initial experiments and didn't pursue this line of investigation further. Likewise, in considering other fairness-enhancing methods that have been proposed, we found that some (such as the regularization method developed in (*12*)) were not well-suited to the "top k" problem setting and others (such as the methods in (*11, 20*)) could be shown to be equivalent to group-specific thresholds under certain conditions (see the Supplementary Materials for a discussion).

Applying our three strategies to a wide variety of models and hyperparameter combinations including random forests, logistic regression, boosting, and decisions trees built for each policy problem, appreciable recall disparities were present in the unadjusted models in all cases, ranging from 50% higher recall favoring white individuals in the Inmate Mental Health context to as much as a 250% disparity in the Student Outcomes setting. In the results shown in Fig. 2, strategies yielding larger disparities will have higher values along the y-axis while strategies yielding precision decreases would move left along the x-axis relative to the unmitigated models (solid squares marked with a 'U'). However, we see that mitigating disparities does not come at any appreciable cost to overall model performance on unseen data for any of the policy settings investigated: $precision@k$ is similar in magnitude for the unmitigated and mitigated models, with little difference between the fairness-enhancing approaches (large error bars in the unmitigated Student Outcomes result reflect the small number of temporal splits and a single cohort with low baseline disparity). This finding empirically demonstrates that any inherent cost or trade-off that might be present in adjusting for disparities appear to be relatively small.

To explore these results in more detail, Fig. 3 shows the effect of these bias mitigation strategies on the Inmate Mental Health setting over temporal validation cohorts (see Supplemental Figs. S2-S4 for results from the other policy settings). In this context, we also wanted to understand how these methods performed when adjusting for disparities across multiple subgroups, looking at race/ethnicity across white, black, and hispanic individuals.

Overall performance ($precision@500$) was similar for models selected from all three strategies over time (Fig. 3A) and recall disparities between white and black individuals were consistently and significantly improved by the adjustments (Fig. 3B). Without accounting for equity, recall for white individuals was 40-100% higher than black individuals in the chosen models, while both fairness enhancing strategies yielded a ratio near one (parity) for these two subgroups. The results for disparities between hispanic and white individuals (Fig. 3C) are broadly consistent, but show significantly more variation, likely due to the relatively small size of this group in the population (about 11% of each cohort), and appear particularly acute with the composite model approach.

Because this variability could reflect the small size of the hispanic sub-group in the list selected for intervention, we sought to explore the sensitivity of our results to changes in top k list size. In practice, each policy problem comes with a specific value of k determined by the resource constraints of the organization taking action. Fig. 4A-C shows the fairness and accuracy metrics for each strategy at different levels of program resource availability (k). Consistent with the findings above, Fig. 4C shows essentially identical $precision@k$ performance across all selection strategies, both bias-mitigated and unmitigated, and over all values of k explored. Across all k, the fairness improving strategy choosing a single model consistently showed improvement in disparity metrics for both black (Fig. 4A) and hispanic (Fig. 4B) individuals, regardless of whether disparities were adjusted for before or after model evaluation. By contrast, the composite strategy was less effective at equalizing recall across race/ethnicity subgroups, particularly at lower list sizes and with the smaller hispanic subgroup.

The under-performance of the composite model suggests a mechanism driving this result: evaluating a model's performance on such a small group may be especially prone to overfitting, choosing a model with an unreasonably high estimate of precision on the subgroup. Because this model is used to determine the number of hispanic individuals needed to equalize recall relative to other subgroups, the over-estimated precision biases towards selecting a smaller set of individuals than is actually needed (this can be observed in the Fig. S5B, which shows the fraction of hispanic individuals in the selected list). When the performance of the chosen model reverts to the mean on the unseen, future data, the under-estimated subgroup size results in a lower-than-expected recall and, thus, a higher disparity.

To further investigate the impact of subgroup size on the stability of the results (e.g., through sampling variation), we performed a resampling experiment to progressively increase the hispanic fraction in the population, ranging from 5% to 38% (focusing on $k = 500$ based on the actual program resources). As observed above, the composite model performs less well at reducing disparities between white and hispanic individuals for the baseline hispanic fraction of 11% (Fig. 4D). While this under-performance is also observed at other low fractions, this strategy becomes competitive with the other the single model strategy at hispanic fractions above 20%, consistent with our hypothesis that the small population of hispanic individuals in the underlying data may be creating a tendency towards overfitting in the formation of the composite models.

Taken together, these results lead to a promising and novel conclusion: mitigating disparities across a wide range of policy domains and resource sizes may neither require new and complex machine learning methods, nor be prohibitively costly in terms of sacrificing accuracy. Instead, it requires explicitly defining the fairness goal upfront in the machine learning process and taking active, practical steps, such as the post-hoc bias mitigation strategies investigated here, to achieve that goal. This finding contributes to a growing body of evidence that, in practice, straightforward approaches such as thoughtful label choice (*5*), model design (*8*), or post-modeling mitigation can effectively reduce biases in many machine learning systems. One factor that may contribute to the nominal trade-offs observed here may be the resource-constrained top k context itself, where there may be a sufficient number of additional positive examples to find in each subgroup with small threshold adjustments. Our focus on benefit allocation programs with limited resources reflects a setting that occurs very commonly in policy contexts and frequently encountered in our applied work (*6, 7, 9, 15*). Settings in which resources are less constrained may pose a greater challenge to disparity mitigation, typically operating deeper in the score distribution where the precision gradient may be flatter and larger adjustments may be required.

Much has also been written about the wide variety of fairness metrics that may be relevant depending on the context (*1, 14, 15, 21*), and further exploration of the fairness-accuracy trade-offs in those contexts is certainly warranted, particularly where balancing multiple fairness metrics may be desirable. Likewise, it may be possible that there is a tension between improving fairness across different attributes (e.g., sex and race) or at the intersection of attributes. Future work should also extend these results to explore the impact not only on equity in decision making, but also equity in longer-term outcomes and implications in a legal context such as those discussed in (*22*).

While much remains to be learned, the empirical findings presented here challenge the commonly held belief that there necessarily is a trade-off between "accuracy" and "fairness" in machine learning systems. We hope this work will inspire researchers, policymakers, and data science practitioners alike to explicitly consider fairness as a goal and take steps, such as those proposed here, in their work that can, collectively, contribute to bending the long arc of history towards a more just and equitable society.

References

1. A. Chouldechova, Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**, 153–163 (2017).

2. J. L. Skeem, C. T. Lowenkamp, Risk, race, and recidivism: predictive bias and disparate impact. *Criminology* **54**, 680–712 (2016).

3. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

4. M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, Mitigating bias in algorithmic hiring: evaluating claims and practices. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2020), pp. 469–481.

5. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **336**, 447–453 (2019).

6. A. Ramachandran, A. Kumar, H. Koenig, A. De Unanue, C. Sung, J. Walsh, J. Schneider, R. Ghani, J. Ridgway, Predictive analytics for retention in care in an urban HIV clinic. *Scientific Reports* 10.1038/s41598-020-62729-x (2020).

7. M. J. Bauman, R. Sullivan, C. Schneweis, R. Ghani, K. Boxer, T. Lin, E. Salomon, H. Naveed, L. Haynes, J. Walsh, J. Helsby, S. Yoder, Reducing incarceration through prioritized interventions. *Proceedings of the Conference on Computing and Sustainable Societies (COMPASS)* (ACM, 2018), pp. 1–8.

8. A. Chouldechova, E. Putnam-Hornstein, S. Dworak-Peck, D. Benavides-Prado, O. Fialko, R. Vaithianathan, S. Friedler, C. Wilson, A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* **81**, 134–148 (2018).

9. E. Potash, J. Brew, A. Loewi, S. Majumdar, A. Reece, J. Walsh, E. Rozier, E. Jorgensen, R. Mansour, R. Ghani, Predictive modeling for public health: preventing childhood lead

poisoning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 2039–2047.

10. I. Y. Chen, F. D. Johansson, D. Sontag, Why is my classifier discriminatory? *Advances in Neural Information Processing Systems 31* (NIPS, 2018), pp. 3539–3550.

11. L. Elisa Celis, L. Huang, V. Keswani, N. K. Vishnoi, Classification with fairness constraints: a meta-algorithm with provable guarantees. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), pp. 319–328.

12. M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment and disparate impact: learning classification without disparate mistreatment. *26th International World Wide Web Conference* (WWW, 2017), pp. 1171–1180.

13. C. Dwork, N. Immorlica, A. T. Kalai, M. Leiserson, Decoupled classifiers for group-fair and efficient machine learning. *Proceedings of Machine Learning Research* **81**, 119–133 (2018).

14. M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems 29* (NIPS, 2016), pp. 3315–3323.

15. K. T. Rodolfa, E. Salomon, L. Haynes, I. Mendieta, J. Larson, R. Ghani, Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2020), pp. 142–153.

16. H. Heidari, K. P. Gummadi, C. Ferrari, A. Krause, Fairness behind a veil of ignorance: a welfare analysis for automated decision making. *Advances in Neural Information Processing Systems* (NIPS, 2018), pp. 1265–1276.

17. S. A. Friedler, S. Choudhary, C. Scheidegger, E. Hamilton, S. Venkatasubramanian, D. Roth, A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), pp. 329–338.

18. M. Kearns, A. Roth, S. Neel, Z. S. Wu, An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), pp. 100–109.

19. M. B. Zafar, I. Valera, M. G. Rogriguez, K. P. Gummadi, Fairness constraints: mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (PMLR, 2017), pp. 962–970.

20. A. Krishna Menon, R. C. Williamson, The cost of fairness in binary classification. *Proceedings of Machine Learning Research* (PMLR, 2018), pp. 1–12.

21. S. Verma, J. Rubin, Fairness definitions explained. *International Workshop on Software Fairness* (IEEE/ACM, 2018), pp. 1–7.

22. A. Huq, Racial equity in algorithmic criminal justice. *Duke Law Journal* **68**, 1043–1134 (2019).

23. M. Hamilton, People with complex needs and the criminal justice system. *Current Issues in Criminal Justice* **22**, 307–324 (2010).

24. D. J. James, L. E. Glaze, "Mental health problems of prison and jail inmates" (Department of Justice, Bureau of Justice Statistics, 2006; https://www.bjs.gov/content/pub/pdf/mhppji.pdf)

25. E Fuller Torrey, A. D. Kennard, D. Eslinger, R. Lamb, J. Pavle, "More mentally ill persons are in jails and prisons than hospitals: a survey of the states" (Treatment Advocacy Center and National Sheriffs' Association, 2010; http://tulare.networkofcare.org/library/final_jails_v_hospitals_study1.pdf)

26. H. Holtzen, E. G. Klein, B. Keller, N. Hood, Perceptions of physical inspections as a tool to protect housing quality and promote health equity. *Journal of Health Care for the Poor and Underserved* **27**, 549–559 (2016).

27. E. Klein, B. Keller, N. Hood, H. Holtzen, Affordable housing and health: a health impact assessment on physical inspection frequency. *Journal of Public Health Management and Practice* **21**, 368–374 (2015).

28. S. Athey, Beyond prediction: using big data for policy problems. *Science* **355**, 483–485 (2017).

29. E. L. Glaeser, A. Hillis, S. D. Kominers, M. Luca, Crowdsourcing city government: using tournaments to improve inspection accuracy. *Am. Econ. Rev.* **106**, 114–118 (2016).

30. H. M. Levin, C. Belfield, "The price we pay: economic and social consequences of inadequate education" (Brookings Institution Press, 2007).

31. M. N. Atwell, R. Balfanz, J. Bridgeland, E. Ingram, "Building a grad nation" (America's Promise Alliance, 2019; https://www.americaspromise.org/2019-building-grad-nation-report)

32. H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, K. L. Addison, A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 1909–1918.

33. E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, K. L. Addison, Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time. *Proceedings of the Learning Analytics and Knowledge Conference* (ACM, 2015), pp. 93–102.

34. A. J. Bowers, R. Sprott, S. A. Taff, Do we know who will drop out? A review of the predictors of dropping out of high school: precision, sensitivity, and specificity. *The High School Journal* **96**, 77–100 (2012).

35. I. Morgan, A. Amerikaner, "Funding gaps 2018" (The Education Trust, 2018; https://edtrust.org/wp-content/uploads/2014/09/FundingGapReport_2018_FINAL.pdf).

36. M. Hurza, "What do teachers spend on supplies" (Adopt a Classroom, 2015; https://www.adoptaclassroom.org/2015/09/15/infographic-recent-aac-survey-results-on-teacher-spending/).

37. Statistics available from DonorsChoose at https://www.donorschoose.org/about (accessed: 23 June 2020).

38. Data available at https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data (accessed: 23 June 2020).

39. D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

40. T. Ye, R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, R. Ghani, Using machine learning to help vulnerable tenants in New York city. *Proceedings of the*

Conference on Computing and Sustainable Societies (COMPASS) (ACM, 2019), pp. 248–258.

Acknowledgments

We would like to thank the Data Science for Social Good Fellowship fellows, project partners, and funders as well as our colleagues at the Center for Data Science and Public Policy at University of Chicago for the initial work on projects that were extended and used in this study. We also thank Kasun Amarasinghe for helpful discussions on the study and drafts of this paper. **Author contributions:** KTR: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft; HL: Investigation, Software, Writing - Review & Editing; RG: Conceptualization, Supervision, Funding acquisition, Writing - Review & Editing **Competing interests:** The authors have no conflicts of interest to declare. **Data and materials availability:** Data from the Inmate Mental Health, Housing Safety, Student Outcomes contexts contains sensitive information, was shared under strict DUAs, and cannot be made publicly available. However, the Education Crowdfunding dataset is publicly available at: kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose. Additionally, the code used here for modeling, disparity mitigation, and analysis is available at github.com/dssg/peeps-chili.

Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1 to S7

Tables S1 to S5

References *(23–40)*

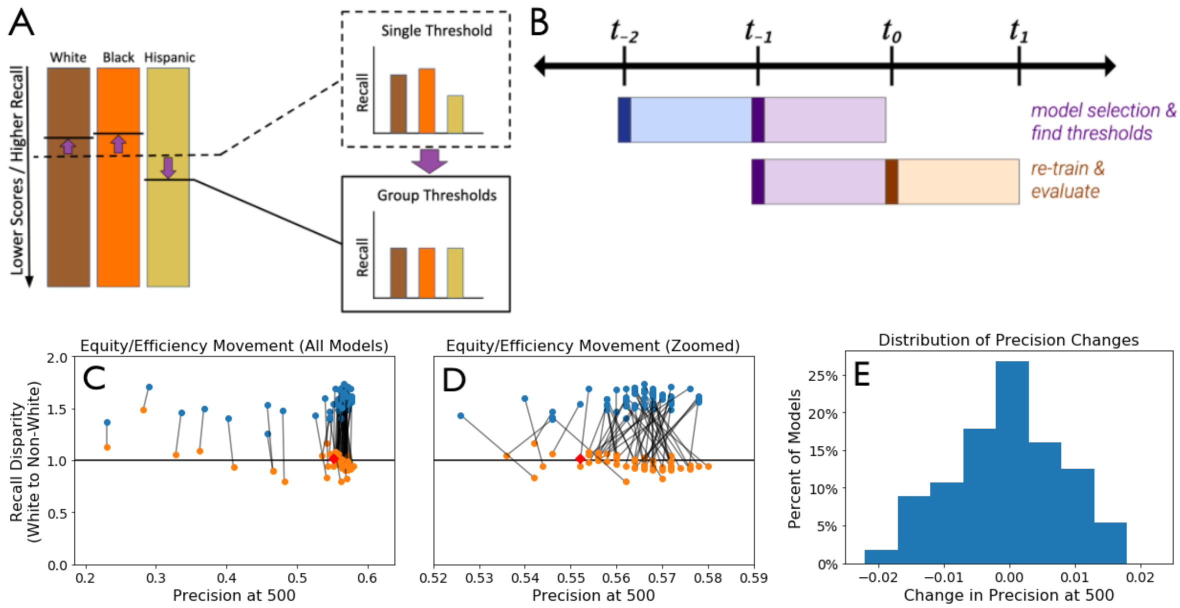

Figure 1: Illustration of the methods used and motivating results. (A) Subgroup-specific thresholds are applied to a modeled risk score to improve the recall equity among individuals chosen for intervention while maintaining a desired overall list size. (B) Temporal validation strategy: a grid of models is trained using examples as of t_{-2} (dark blue, with labels derived from the time shown in light blue) and predictions on a cohort as of t_{-1} (dark purple with labels derived from the time shown in light purple) are used to determine the equity-balancing thresholds described in (A). Models are then re-trained on this cohort for "current day" predictions as of t_0 (dark tan, with labels in light tan) used to evaluate model performance with equity adjustments. (C and D): Changes in race/ethnicity recall disparities before (blue) and after (orange) making post-hoc score adjustments for fairness in the Inmate Mental Health context. (C) shows all model specifications and (D) shows the cluster of well-performing models. The red diamond reflects the performance of a composite model combining the best-performing model for each subgroup. (E) Distribution of precision changes after adjusting for disparities for the models shown in (C).

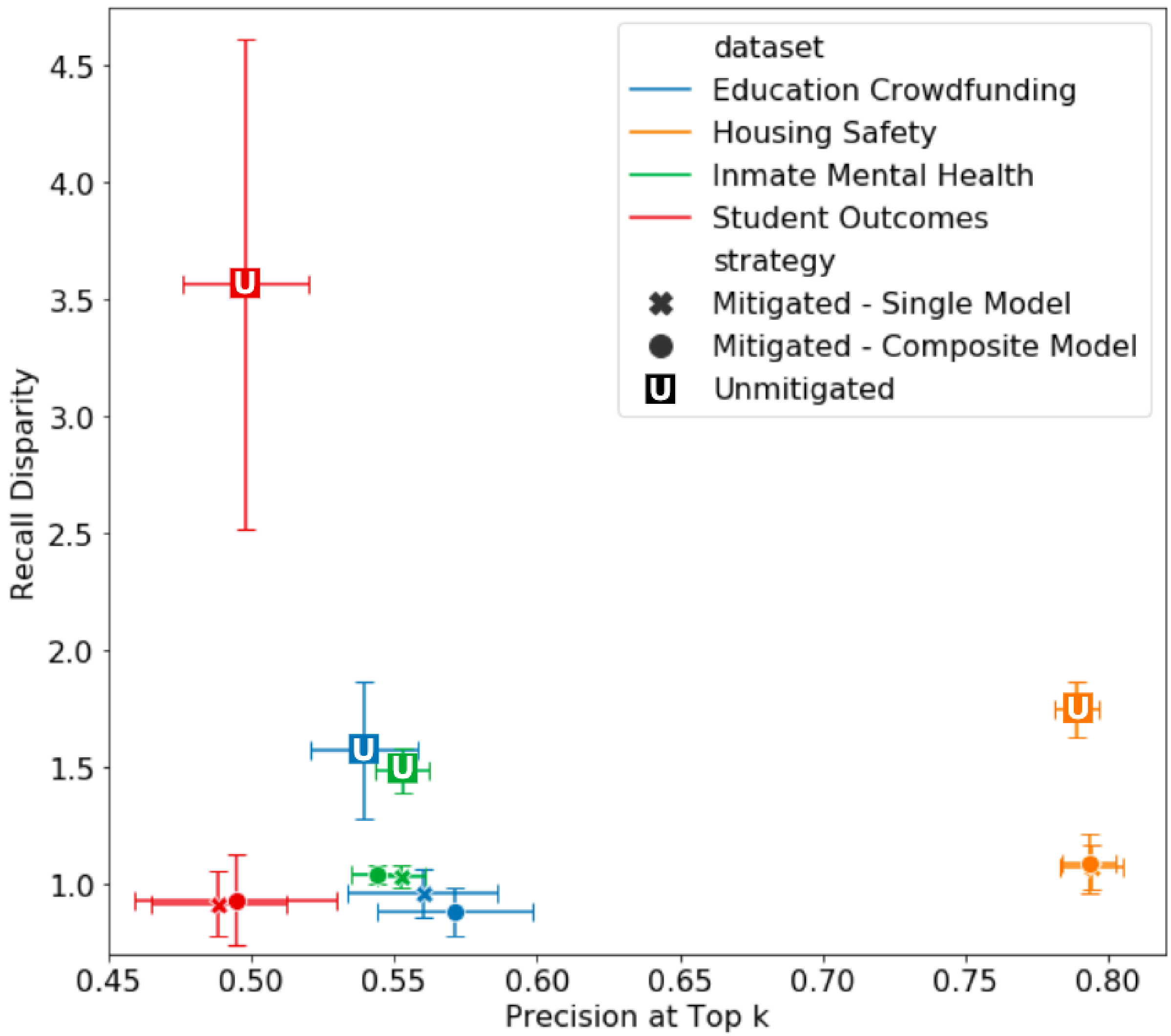

Figure 2: Comparing equity (recall disparity) and performance (precision@k) metrics for different model selection strategies between different policy contexts. In the Education Crowdfunding context, models are evaluated at $k = 1000$ over 10 cohorts; in the Inmate Mental Health context, $k = 500$ across 10 cohorts; in the Student Outcomes context, $k = 10,000$ across 5 cohorts; and in the Housing Safety context, $k = 500$ across 9 cohorts. Unmitigated (baseline) models are shown as solid squares marked with a 'U'. Decreases of strategies involving disparity mitigations relative to the y-axis demonstrate improvements in equity while showing little or no decrease in overall performance (that is, leftward movement on the x-axis). Error bars reflect 95% confidence intervals across temporal validation cohorts.

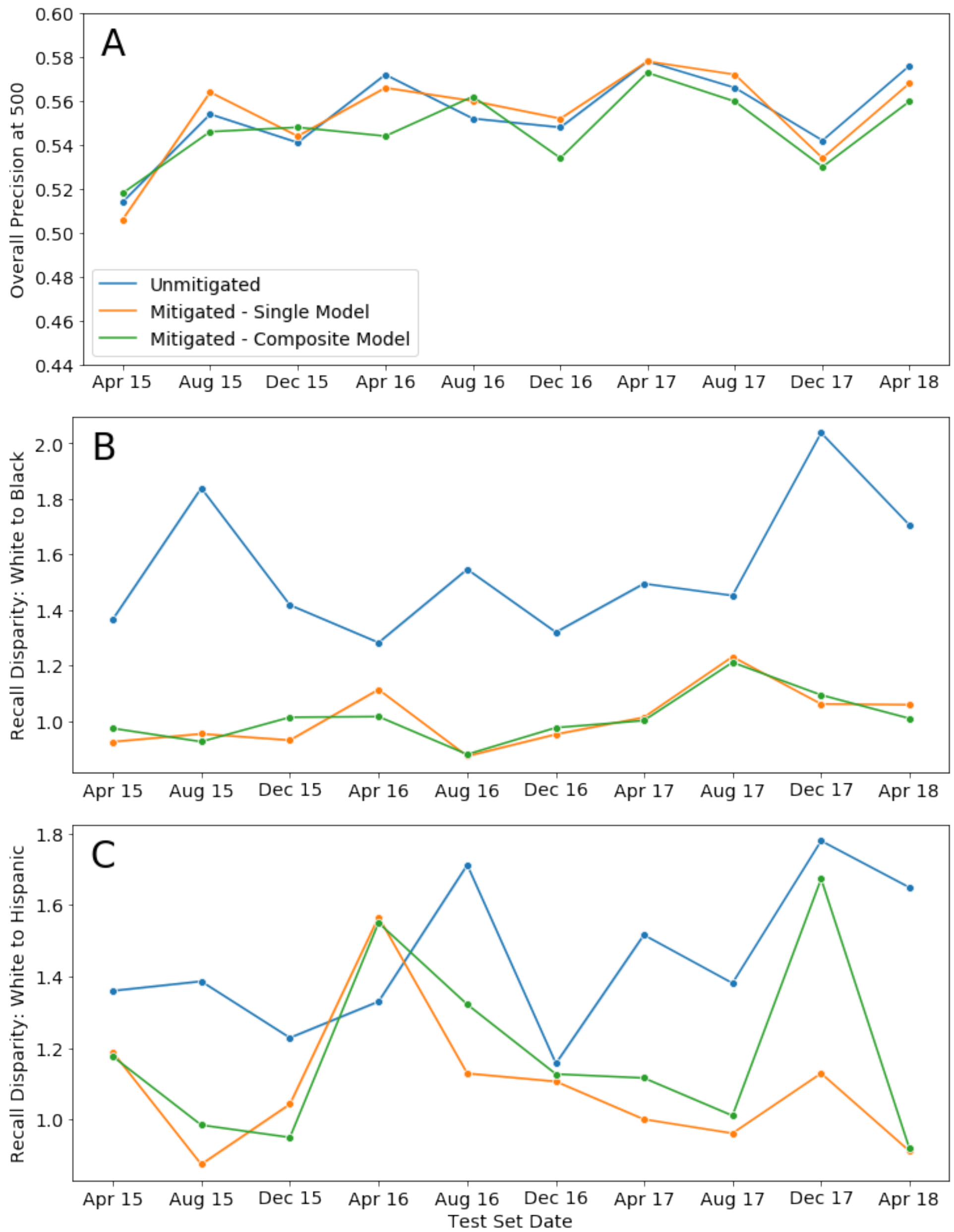

Figure 3: Comparing disparity and performance metrics over time for different model selection strategies. Results from the Inmate Mental Health policy setting using a total list size $k = 500$. (A) Model performance in terms of overall $precision@500$. (B) Recall disparities between white and black individuals. (C) Recall disparities between white and hispanic individuals.

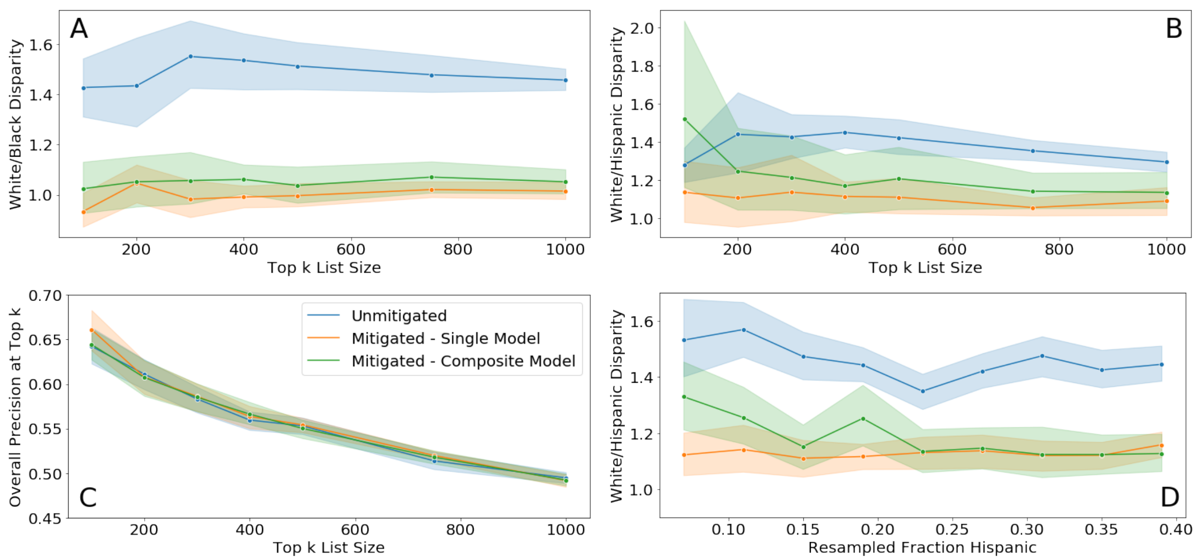

Figure 4: Comparing disparity and performance metrics across program scale and protected group size in the Inmate Mental Health policy setting. (A to C) Variation in results by list size: (A) white/black disparities, (B) white/hispanic disparities, and (C) overall $precision@k$. (D) Disparities between white and hispanic individuals by resampled size of the protected group in the overall population (using a list size of $k = 500$). Shaded intervals reflect variation across temporal validation splits (A-C) as well as bootstrap samples in (D).

Supplementary Materials for Machine learning for public policy: Do we need to sacrifice accuracy to make models fair?

Kit T. Rodolfa,[1] Hemank Lamba,[1] Rayid Ghani[1*]

[1]Machine Learning Department and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, 4902 Forbes Ave, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed; E-mail: rayid@cmu.edu.

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S5
References *(23–40)*

Materials and Methods

Policy Contexts and Data

A key aim of this work was to explore the fairness-accuracy trade-offs encountered in practice in the context of machine learning applications for public policy settings. As such, we drew on several projects we have worked on in partnership with government agencies across policy domains. We describe these contexts briefly in the main text and provide more details about each setting below:

Inmate Mental Health Untreated mental health conditions often result in a negative spiral, which can culminate in repeated periods of incarceration with long term consequences both for the affected individual and the community as a whole (*23*). Surveys of inmate populations have suggested a high prevalence of multiple and complex needs, with 64% of people in local jails suffering from mental health issues and 55% meeting criteria for substance abuse or dependence (*24*). The criminal justice system is poorly suited to address these needs, yet houses three times as many individuals with serious mental illness as hospitals (*25*). In 2016, Johnson County, KS, partnered with our group to help them break this cycle of incarceration by identifying individuals who might benefit from outreach with mental health resources and are at risk for future incarceration. While the Johnson County Mental Health Center (JCMHC) currently provides services to the jail population, needs are generally identified reactively, for instance through screening instruments individuals fill out when entering jail. The new program being developed will supplement these existing approaches by adding a new automatic referral system for people who are at risk of being booked into jail, with the hope that they can be outreached before they return to jail. Through our partnership, the county provided administrative data from their mental health center, jail system, police arrests, and ambulance runs. Modeling was focused on a cohort of Johnson County residents with any history of mental health need who had been released from jail within the past three years. Early results from this work were described in (*7*). A field evaluation of the predictive model is ongoing at the time of this writing, but validation on historical data demonstrated a 12% improvement over a baseline based on the number of bookings in the prior year and 4.8-fold increase over the population prevalence.

Housing Safety The Multiple Housing team in San Jose's Code Enforcement Office is tasked with protecting the occupants of properties with three or more units, such as apartment buildings, fraternities, sororities, and hotels. They do so by conducting routine inspections of these properties, looking for everything from blight and pest infestations to faulty construction and fire hazards (see (*26*) and (*27*) for a discussion of the importance of housing inspections to public health). Although the city of San Jose inspects all of the properties on its Multiple Housing roster over time, and expects to find minor violations at many of them, it is important that they can identify and mitigate dangerous situations early to prevent accidents. With more than 4,500 multiple housing properties in San Jose, CA – many of which comprise multiple buildings and hundreds of units – it is not possible for the city to inspect every unit every year. San Jose recently instituted a tiered approach to prioritizing inspections, inspecting riskier properties more frequently and thoroughly. Although the tier system helped focus inspections on riskier properties, the new system has its limitations. The city evaluates tier assignments for properties infrequently (every 3 to 6 years), and these adjustments require a great deal of expertise and manual work while leaving out a rich amount of information. In order to provide a more nuanced view of properties' violation risk over time and allow for more efficient scheduling of inspections, the Code Enforcement Office partnered with us to develop a model to predict the risk that a serious violation would be found if a given property was prioritized for inspection (similar tools have been developed for allocating fire inspections in New York (*28*) and health inspections in Boston (*29*)). Evaluation of the model on historical data indicated that it could provide a 30% increase in precision relative to the current tier system and the model's predictive accuracy was confirmed during a 4-month field trial in 2017.

Student Outcomes Each year from 2010 through 2016, 15-29% of students enrolled in school in El Salvador did not return to school in the following year. This high dropout rate is cause for serious concern, with significant consequences for economic productivity, workforce skill, inclusiveness of growth, social cohesion, and increasing youth risks (*30, 31*). El Salvador's Ministry of Education has programs available to support students with the goal of reducing these high dropout rates, but the budget for these programs is not large enough to reach every student and school in El Salvador. Predictive modeling has been deployed to help schools identify students at risk of dropping out in several contexts (*32–34*) and El Salvador partnered with us in 2018 to make use of these methods to focus their limited resources on the students at highest risk of not returning each year. Student-level data was provided by the Ministry of Education, including demographics, urbanicity, school-level resources (e.g., classrooms, computers, etc), gang and drug violence, family characteristics, attendance records, and grade repetition. For the present study, we focused on the state of San Salvador and identifying the 10,000 highest-risk students, considering annual cohorts of approximately 300,000 students and drawing on 5 years' of prior examples as training data.

Education Crowdfunding Many schools in the United States, particularly in poorer communities, face funding shortages (*35*). Often, teachers themselves are left to fill this gap, purchasing supplies for their classrooms when they have the individual resources to do so (*36*). The non-profit DonorsChoose was founded in 2000 to help alleviate these shortages by providing a platform where teachers post project requests focused on their classroom needs and community members can make individual contributions to support these projects. Since 2000, they have facilitated $970 million in donations to 40 million students in the United States (*37*). However, approximately one third of all projects posted on the platform fail to reach their funding goal. Here, we make use of a dataset DonorsChoose made publicly available for the 2014 KDD Cup (an annual data mining competition) including information about projects, the schools posting them, and donations they received. Because the other case studies explored here focused on proprietary and often sensitive data shared with us under data use agreements that cannot be made publicly available, we included a case study surrounding this publicly-available dataset. While we have not partnered with DonorsChoose to deploy the machine learning system described, we otherwise treated this case study as we would any of our applied projects. Here, we consider a resource-constrained effort to assist projects at risk of going unfunded (for instance, providing a review and consultation) capable of helping 1,000 projects in a 2-month window, focusing on the most recent 2 years' of data available in the extract (earlier data had far fewer projects and instability in the baseline funding rates as the platform ramped up). This dataset is publicly available at kaggle.com (*38*).

Machine Learning Details

All machine learning models, including feature engineering, model training, and performance evaluation were run using our open-source `python` ML pipeline package, `triage`. Machine learning methods used are from `sklearn` (a python package) or `catwalk` (a component of `triage` for baselines methods as well as `ScaledLogisticRegression`, which wraps the `sklearn` logistic regression to ensure input features as scaled between 0 and 1). The modeling grid for each project is described in Tables S2-S5, reflecting the modeling space explored by the teams working on each project. For each estimator in the tables, the grid search considered reflects the full cross-product of the hyperparameter values specified. As illustrated in Fig. 1B, we used a strategy of inter-temporal cross-validation (as described in (*39*) and (*40*)) to ensure that model evaluation and selection was done in a manner that reflected performance on novel data while guarding against "leakage" of information from the future affecting past results.

The method we used for mitigating disparities by post-modeling adjustment involving choosing sub-group specific thresholds (see Fig. 1A) was described in detail in (*15*) and draws on the idea of "equality of opportunity" discussed by Hardt in (*14*). In brief, because the notion of fairness relevant in these policy settings relies on equalizing recall across groups, and recall monotonically increases with depth traversed in a model score, unique score thresholds that balance recall across groups can be readily found for a given combined list size. For each model, we calculate within-group recall values up to each individual in an initial test set (purple / t_{-1} in Fig. 1B), order the combined set by within-group recall and take the top k individuals from this reordered set, calculating k_g for each group g such that $\sum k_g = k$ (the total top k list size desired) and recall is balanced across groups. To evaluate this process on novel data, models were tested on a future cohort (tan / t_0 in Fig. 1B) and the top k_g examples (ranked by score, then randomly to break ties) from each sub-group were selected to measure $precision@k$ and recall disparities. In the process of model selection, we explored applying these disparity-mitigating thresholds either before choosing a model specification ("Mitigated - Single Model") or after ("Mitigated - Unadj. Model Seln."), finding no substantive difference in performance (Fig. S1). For the "Mitigated - Composite Model" approach, a similar method was used, but within-group precision up to each individual is calculated for each model as well to determine the best model specification for each sub-group at each list depth (drawing on the ideas suggested by Dwork and colleagues in (*13*)), then k_g and group-specific model specifications are chosen for evaluation on novel data.

Code for all four projects including `triage` configuration files specifying the full feature sets used as well as code used to mitigate disparities and evaluate fairness is available at `github.com/dssg/peeps-chili`

Supplementary Text

The results discussed in the main text focus on post-hoc score adjustment methods, drawing on several previous lines of work (*13,14,15*) and finds these straightforward methods suffice to significantly reduce disparities with little accuracy trade-off in the settings considered. However, we also did some preliminary work exploring a number of other methods that have been proposed, particularly to enhance fairness during the modeling process itself. In public policy settings with limited resources, regularization methods (such as that described by Zafar and colleagues in (*12*)) provide a particular challenge. These techniques typically seek to find the best overall classifier subject to some fairness constraint. This may be well-suited to contexts where there are no hard constraints on the resources available to act on predictive positives (for instance, in deciding whether to order a relatively low-cost medical diagnostic test).

However, many applications, particularly in public policy contexts, are subject to the further constraint of limited resources — these applications are best formulated as "top-k" problems, but unfortunately this formulation introduces a non-convex optimization that isn't readily integrated into current fairness-constrained methods. Likewise, naively thresholding the resulting score from these methods to yield a set number of predicted positives provides no guarantee that the fairness constraints will hold. Fig. S7 provides an example with data from the Inmate Mental Health context: using the method described in (*12*) to generate a score with a false negative rate fairness constraint (note that $FNR = 1 - TPR$, so this constraint is equivalent to recall equity) and choosing the top 500 individuals performs no better at balancing equity than choosing the top 500 from an unconstrained score and far worse than choosing group-specific thresholds to balance equity as shown in the main text. Notably, this result isn't a criticism of Zafar's methods when applied in appropriate contexts, but rather an indication of the limitations of current fairness-constrained methods in the resource-constrained setting we frequently encounter in machine learning to support public policy decision making.

We also explored the methods proposed by Celis (*11*) and Menon and Williamson (*20*), but noticed that both methods could be shown to be equivalent to group-specific scaling or thresholding of an underlying estimate of the relevant probability distribution when applied to a single fairness metric and sub-group membership is known (notably, the method in (*11*) seems quite flexible to balancing multiple metrics or where group membership is itself being modeled as well). For a single, monotonic metric like recall/equality of opportunity, there will be a unique balanced solution of a given total size, so any method relying on post-hoc score adjustments should yield similar results.

To see this in the context of the method described by Celis and colleagues (*11*), we can start from their observation that any equity definition balancing a confusion-matrix statistic can be represented in the following form (following the notation used in their work):

$$q_{linf}^{(i)} = \frac{\alpha_0^{(i)} + \sum_{j \in [k]} \alpha_j^{(i)} P[f = 1 \mid G_i, A_j^{(i)}]}{\beta_0^{(i)} + \sum_{j \in [l]} \beta_j^{(i)} P[f = 1 \mid G_i, B_j^{(i)}]} \tag{1}$$

Here, i reflects subgroup membership, $A_j^{(i)}$ and $B_j^{(i)}$ are events such as $(Y = 1)$ and $\alpha_j^{(i)}$ and

$\beta_j^{(i)}$ are parameters defined to represent generalized form of equity function (likewise, $k, l \geq 0$ are integer values allowing the function to be written as linear combination of terms). Further, some metrics (including recall (that is, TPR) as we consider in this work) can be represented with a simplified linear form in which $\beta_0^{(i)} = 1$ and $\beta_1^{(i)} = 0$, giving:

$$q_{lin}^{(i)} = \alpha_0^{(i)} + \sum_{j \in [k]} \alpha_j^{(i)} P[f = 1 \mid G_i, A_j^{(i)}] \tag{2}$$

In this case, they show that fairness-improved score of an instance being classified is given by:

$$s_\lambda(x) = \eta(x) - 0.5 + \sum_{i \in [p]} \lambda_i \left(\sum_{j \in [k]} \frac{\alpha_j^{(i)}}{\pi_j^{(i)}} \eta_j^{(i)}(x) \right) \tag{3}$$

Where, $\eta(x) = P[Y = 1 \mid X = x]$, $\eta_j^{(i)}(x) = Pr[G_i, A_j^i \mid X = x]$ and $\pi_j^{(i)} = Pr[G_i, A_j^{(i)}]$, while λ_i represents langrangian parameters, used to solve the optimization problem.

For recall specifically (defined as $TPR = P[f = 1 \mid G_i, Y = 1]$), the sum in q_{lin} has a single term ($j = 1$), with $\alpha_0^{(i)} = 0$, $\alpha_1^{(i)} = 1$, $A_1^{(i)} = (Y = 1)$. Further (and without loss of generality), for the case in which there are two sub-groups of the protected attribute (e.g., men and women), we can expand the sums and rewrite the adjusted score as:

$$\begin{aligned} s_{recall}(x) = P(Y = 1 \mid X = x) - 0.5 + \lambda_0 \left[\frac{1}{P[G_0, Y = 1]} P[G_0, Y = 1 \mid X] \right] \\ + \lambda_1 \left[\frac{1}{P[G_1, Y = 1]} P[G_1, Y = 1 \mid X] \right] \end{aligned} \tag{4}$$

This formulation can be quite useful for balancing recall equity in the case where group membership is itself being inferred. However, where we know $G = g$ for a given X, we note that this will pick out a single term as either $P[G_1, Y = 1 \mid X] = 0$ or $P[G_0, Y = 1 \mid X] = 0$ (note that this will be true if there were more sub-groups as well, simply adding additional terms). So, for a given example X, we obtain:

$$s_{recall}(x) = P(Y = 1 \mid X = x) - 0.5 + \lambda_g \left[\frac{1}{P[G = g, Y = 1]} P[Y = 1 \mid X] \right] \tag{5}$$

Rearranging leads to:

$$s_{recall}(x) = \left[1 + \frac{\lambda_g}{P[G = g, Y = 1]} \right] P(Y = 1 \mid X) - 0.5 \tag{6}$$

As it can be seen from the above equation, $1 + \lambda_g / P[G = g, Y = 1]$ is simply a group-specific constant with no dependence on the value of X beyond group membership. As such, this will result in scaling factor corresponding to squishing or scaling of an underlying predicted score $P(Y = 1 \mid X)$ around a threshold, but without reordering the score within a given subgroup itself. Further, because the monotonically increasing nature of recall as list depth is traversed, there will be a unique (up to exact ties in the score) recall-balancing solution of a given total list size, regardless of whether that solution is obtained by setting group-specific thresholds (as in the method we make use of in the current study) or group-specific stretching scores around a fixed threshold (as we see here), so long as the within-group ranking is not reordered.

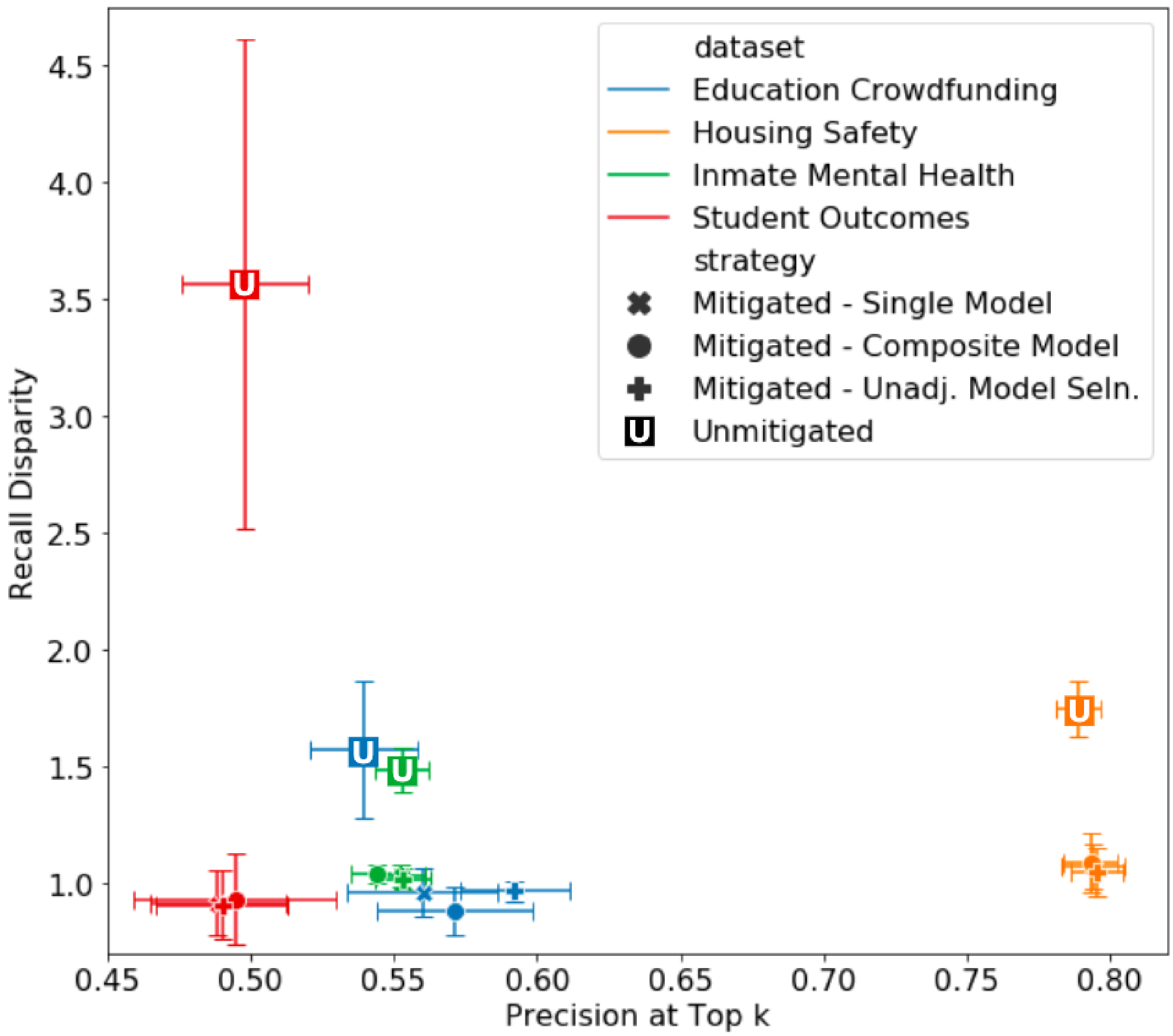

Fig. S1: Comparing equity (recall disparity) and performance (precision@k) metrics for different model selection strategies between different policy contexts, as in Fig. 2, including an additional model selection strategy (Mitigated - Unadj. Model Seln.) in which model specification is chosen without regard for disparities then group-specific thresholds are chosen for the selected model. Error bars reflect 95% confidence intervals across temporal validation cohorts. This strategy performed similarly to strategy accounting for disparities in the model selection itself in all policy settings considered.

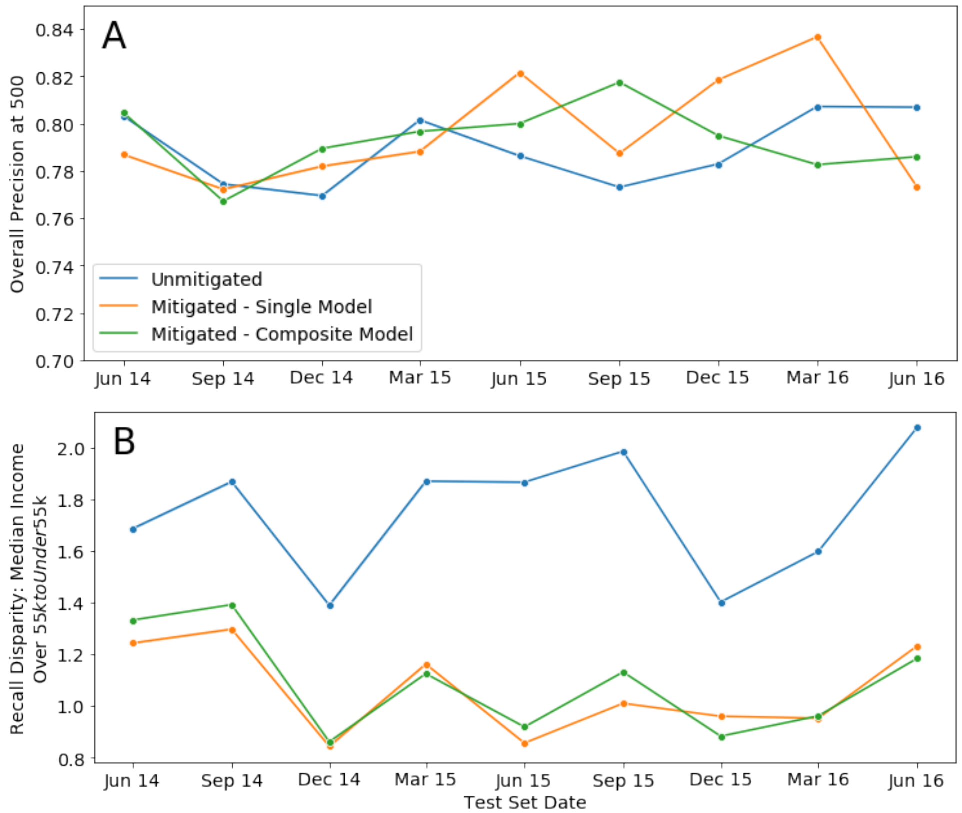

Fig. S2: Comparing model precision (A) and disparity (B) metrics over time for different model selection strategies in the Housing Safety policy context.

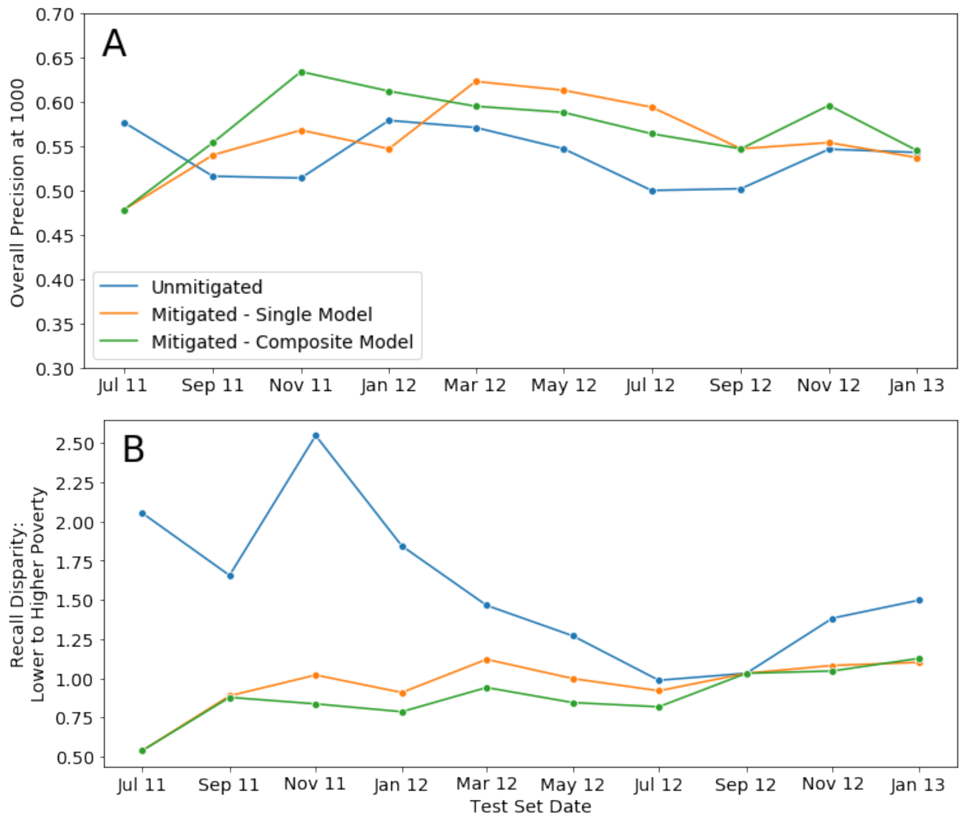

Fig. S3: Comparing model precision (A) and disparity (B) metrics over time for different model selection strategies in the Education Crowdfunding policy context.

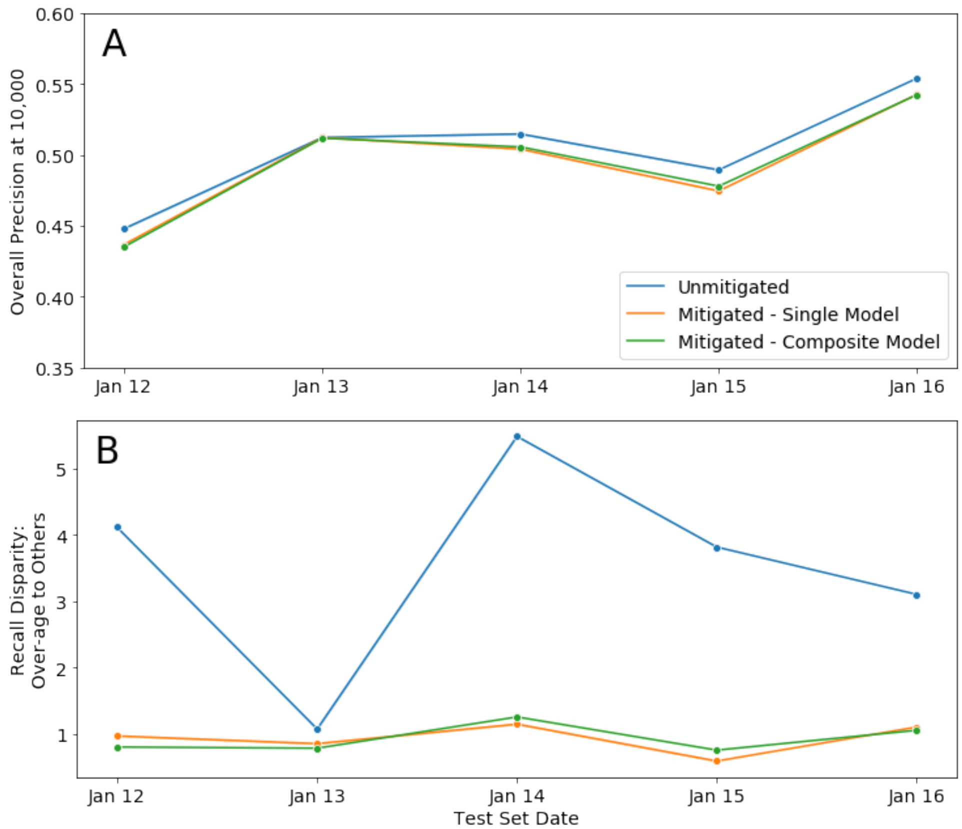

Fig. S4: Comparing model precision (A) and disparity (B) metrics over time for different model selection strategies in the Student Outcomes policy context.

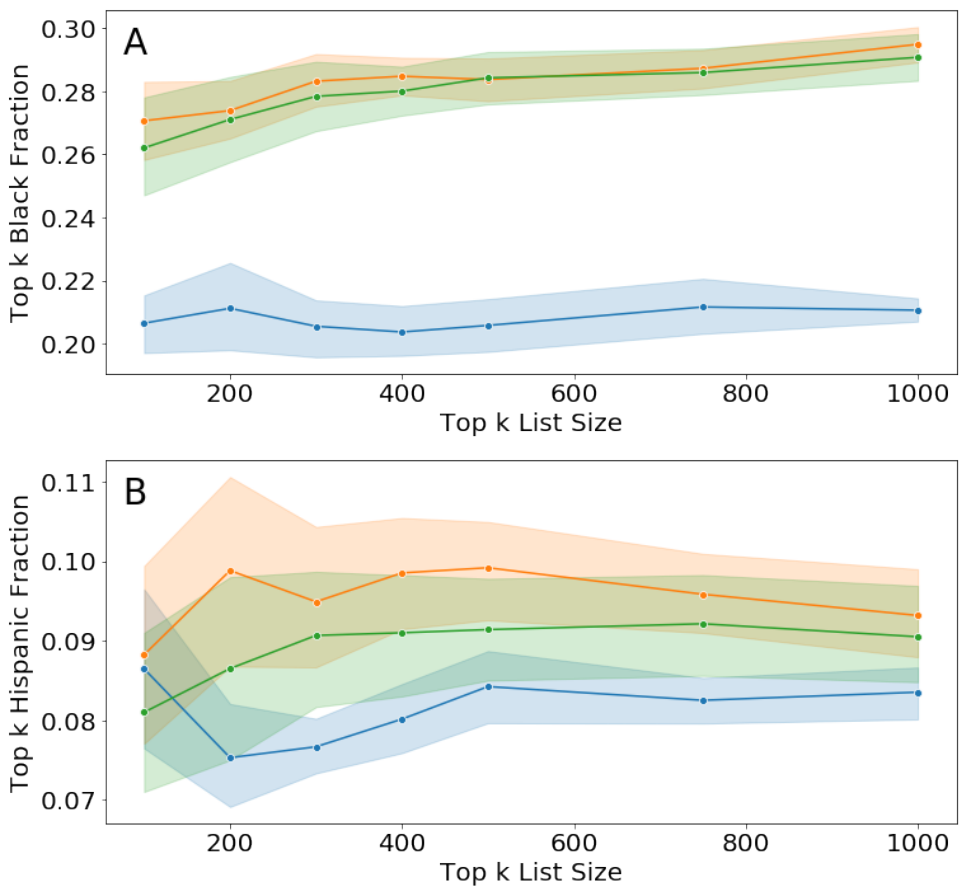

Fig. S5: Fraction of the selected "top-k" list that is African American (A) or Hispanic (B) by list size in the Inmate Mental Health setting (see Fig. 4A-C). Shaded intervals reflect variation across temporal validation splits.

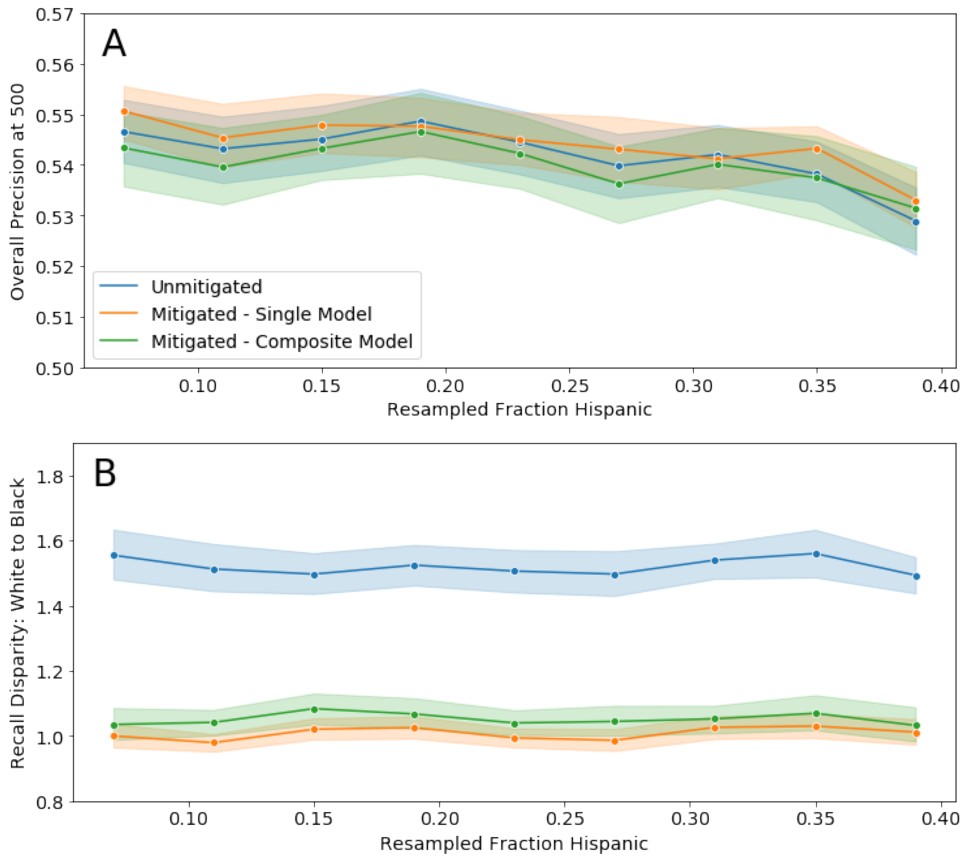

Fig. S6: Precision (A) and white-to-black disparities (B) by re-sampled hispanic fraction from the experiment shown in Fig. 4D. Shaded intervals reflect variation across bootstrap samples and temporal validation splits.

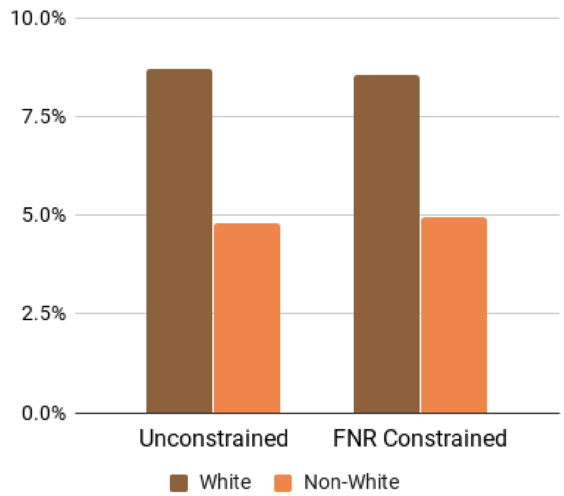

Fig. S7: Comparison of recall disparity after selecting the top 500 individuals from unconstrained and False Negative Rate (FNR)-constrained models following the methods described in Zafar et al. Results from the Inmate Mental Health policy context.

Table S1: Descriptions of Model Selection Strategies

Strategy	Description
Unmitigated	Baseline strategy with no equity adjustments
Mitigated - Composite Model	The highest-precision model is chosen for each subgroup and a composite model is formed combining these, with equity-balancing thresholds used when calculating test performance
Mitigated - Single Model	All models are adjusted for recall equity and then the best model is selected based on precision and equity-balancing thresholds are applied for calculating test performance
Mitigated - Unadj. Model Seln.	Model is selected based on precision then equity-balancing thresholds are applied for calculating test performance

Table S2: Model Hyperparameter Grid for Inmate Mental Health Policy Setting

Estimator	Hyperparamter	Grid Search Values
`RandomForestClassifier`	max_features	sqrt
	criterion	gini, entropy
	n_estimators	100, 1000, 5000
	min_samples_split	10, 25, 100
	max_depth	5, 10, 50
`ScaledLogisticRegression`	penalty	l1, l2
	C	0.001, 0.1, 1, 10
`PercentileRankOneFeature`	feature	Jail bookings last 1 or 5 years
`SimpleThresholder`	rules	Any mental health history, Jail releases in last: 1, 3, 6 months, or 1 year
	logical_operator	and

Table S3: Model Hyperparameter Grid for Education Crowdfunding Policy Setting

Estimator	Hyperparamter	Grid Search Values
`DecisionTreeClassifier`	criterion	gini
	min_samples_split	2, 5, 10, 100, 1000
	max_depth	1, 5, 10, 20, 50, 100
`RandomForestClassifier`	max_features	sqrt
	criterion	gini, entropy
	n_estimators	100, 1000, 2000, 3000
	min_samples_split	10, 50
	max_depth	10, 50, 100
`ExtraTreesClassifier`	max_features	log2
	criterion	entropy
	n_estimators	10, 50, 1000
	min_samples_split	5, 25, 50
	max_depth	20, 50, 100
`ScaledLogisticRegression`	penalty	l1, l2
	C	0.0001, 0.001, 0.01, 0.1, 1, 10
`AdaBoostClassifier`	n_estimators	500,1000

Table S4: Model Hyperparameter Grid for Housing Safety Policy Setting

Estimator	Hyperparamter	Grid Search Values
`DecisionTreeClassifier`	criterion	gini, entropy
	min_samples_split	10, 20, 50, 100
	max_depth	1, 2, 3, 5, 10, 20, 50
`RandomForestClassifier`	max_features	sqrt, log2
	criterion	gini, entropy
	n_estimators	100, 1000, 5000
	min_samples_split	10, 20, 50, 100
	max_depth	2, 5, 10, 20, 50, 100
`ExtraTreesClassifier`	max_features	sqrt, log2
	criterion	gini, entropy
	n_estimators	100, 1000, 5000
	min_samples_split	10, 20, 50, 100
	max_depth	2, 5, 10, 50, 100
`ScaledLogisticRegression`	penalty	l1, l2
	C	0.001, 0.01, 0.1, 1, 10
`PercentileRankOneFeature`	feature	Days since last routine inspection

Table S5: Model Hyperparameter Grid for Student Outcomes Policy Setting

Estimator	Hyperparamter	Grid Search Values
`DecisionTreeClassifier`	criterion	gini
	min_samples_split	2, 5, 10, 100, 1000
	max_depth	1, 5, 10, 20, 50, 100
`RandomForestClassifier`	max_features	sqrt
	criterion	gini
	n_estimators	100, 500
	min_samples_split	2, 10
	class_weight	˜, balanced subsample, balanced
	max_depth	5, 50, 100
`ExtraTreesClassifier`	max_features	log2
	criterion	gini
	n_estimators	100
	max_depth	5, 50
`ScaledLogisticRegression`	penalty	l1, l2
	C	0.0001, 0.001, 0.01, 0.1, 1, 10

References

23. M. Hamilton, People with complex needs and the criminal justice system. *Current Issues in Criminal Justice* **22**, 307–324 (2010).

24. D. J. James, L. E. Glaze, "Mental Health Problems of Prison and Jail Inmates" (Department of Justice, Bureau of Justice Statistics, 2006; https://www.bjs.gov/content/pub/pdf/mhppji.pdf)

25. E Fuller Torrey, A. D. Kennard, D. Eslinger, R. Lamb, J. Pavle, "More Mentally Ill Persons Are in Jails and Prisons Than Hospitals: A Survey of the States" (Treatment Advocacy Center and National Sheriffs' Association, 2010; http://tulare.networkofcare.org/library/final_jails_v_hospitals_study1.pdf)

26. H. Holtzen, E. G. Klein, B. Keller, N. Hood, Perceptions of physical inspections as a tool to protect housing quality and promote health equity. *Journal of Health Care for the Poor and Underserved* **27**, 549–559 (2016).

27. E. Klein, B. Keller, N. Hood, H. Holtzen, Affordable housing and health: a health impact assessment on physical inspection frequency. *Journal of Public Health Management and Practice* **21**, 368–374 (2015).

28. S. Athey, Beyond prediction: using big data for policy problems. *Science* **355**, 483–485 (2017).

29. E. L. Glaeser, A. Hillis, S. D. Kominers, M. Luca, Crowdsourcing city government: using tournaments to improve inspection accuracy. *Am. Econ. Rev.* **106**, 114–118 (2016).

30. H. M. Levin, C. Belfield, "The price we pay: Economic and social consequences of inadequate education" (Brookings Institution Press, 2007).

31. M. N. Atwell, R. Balfanz, J. Bridgeland, E. Ingram, "Building a Grad Nation" (America's Promise Alliance, 2019; https://www.americaspromise.org/2019-building-grad-nation-report)

32. H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, K. L. Addison, A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 1909–1918.

33. E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, K. L. Addison, Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time. *Proceedings of the Learning Analytics and Knowledge Conference* (ACM, 2015), pp. 93–102.

34. A. J. Bowers, R. Sprott, S. A. Taff, Do we know who will drop out? A review of the predictors of dropping out of high school: precision, sensitivity, and specificity. *The High School Journal* **96**, 77–100 (2012).

35. I. Morgan, A. Amerikaner, "Funding Gaps 2018" (The Education Trust, 2018; https://edtrust.org/wp-content/uploads/2014/09/FundingGapReport_2018_FINAL.pdf).

36. M. Hurza, "What Do Teachers Spend on Supplies" (Adopt a Classroom, 2015; https://www.adoptaclassroom.org/2015/09/15/infographic-recent-aac-survey-results-on-teacher-spending/).

37. Statistics available from DonorsChoose at https://www.donorschoose.org/about (accessed: 23 June 2020).

38. Data available at https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data (accessed: 23 June 2020).

39. D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

40. T. Ye, R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, R. Ghani, Using machine learning to help vulnerable tenants in New York city. *Proceedings of the Conference on Computing and Sustainable Societies (COMPASS)* (ACM, 2019), pp. 248–258.