Fairness in Focus: Quantitative Insights into Bias Within Machine Learning Risk Evaluations and Established Credit Models *

Jacob Ford *Solstice Power Technologies*

As the adoption of machine learning algorithms expands across industries, concerns arise about their potential to inadvertently perpetuate existing biases. This study critically examines the bias present in a machine learning risk assessment tool called EnergyScore, contrasting it with conventional credit score models. Our research pursues two objectives: (1) to scrutinize the extent of bias based on classical fairness benchmarks, and (2) to innovate in bias analysis by augmenting the number of protected categories observed. The study's outcomes offer a quantitative juxtaposition of potential discriminatory practices against these groups when using EnergyScore versus traditional credit scores. The insights presented emphasize the imperative nature of evaluating and mitigating bias in machine learning implementations across multiple protected classes. Our results show the machine learning model results in lower degrees of variance in various thresholds for most considered protected classes.

Keywords: bias, protected classes, machine learning

Introduction

As the use of machine learning algorithms continues to spread across industries, concerns about the potential for these algorithms to amplify existing biases have become increasingly relevant. Various techniques for measuring and mitigating bias in machine learning models have been developed. This study deploys these common bias metrics within a risk assessment algorithm and comparing it to the traditional credit score model. The study employs the algorithmic fairness definitions put forth by Hardt, Price and Srebro (2016) and aims to achieve two objectives: first, to assess the level of bias using traditional fairness criteria and second, to conduct a novel analysis by extending the number of protected classes considered. The results of this study will provide a quantitative comparison of the degree of discrimination faced by different protected classes in this specific use case.

Background

Fairness through unawareness, an approach to race-neutral threshold setting in multiple industries, has been widely criticized for being ineffective. Dwork et al. (2011). argue that researchers should embrace protected classes in the data to achieve a more accurate analysis of fairness. A similarity metric is used to describe the degree of similarity between individuals or groups, thereby revealing the ground truth of the distribution of the protected class. However, cases where protected class data is unavailable are not considered in this review.

Similarly, Zemel et al. (2013) acknowledge that bias cannot be completely eliminated from data or models, and that machine learning systems trained on historical data will inevitably inherit past biases. To address this, the authors propose a new algorithm that maps individuals to a probability distribution while preserving as much information as possible, while minimizing loss of identifiable information. The

*Replication files are available on the author's Github account (http://github.com/Jake-Ford). **Current version**: November 10, 2023; **Corresponding author**: jake@solstice.us.

results of this algorithm show improved accuracy and significant fairness, as measured by both individual and group definitions, compared to comparison models.

Furthermore, a recent study Suresh and Guttag (2019) highlights the importance of considering the entire life cycle of machine learning analysis to understand and mitigate sources of bias. The authors argue that blaming unfair results on biased data oversimplifies the complex processes involved in collecting, cleaning, processing, and modeling the data. These processes involve multiple human decisions that can collectively contribute to unintended results. The authors aim to increase awareness and focus on the cumulative sources of bias in machine learning, leading to the development of mitigation techniques.

EnergyScore

The machine learning algorithm, EnergyScore, has previously been described by Davuluri et al. (2019). I employ the original dataset used to create EnergyScore. To allow for an analysis of protected classes and to avoid data leakage, I use the training datset to construct and re-train the model, using the test dataset to quantify the effects on different protected classes.

EnergyScore was shown to be a more inclusive and accurate predictor of utility bill payment performance than a traditional credit score. This research will determine the extent to how different protected classes face discriminatory thresholds, using EnergyScore as the treatment compared to the tradiitionally used credit score.

Methods

Data

The machine learning model analyzed in this study is EnergyScore, an alternative risk assessment tool designed as an alternative to traditional credit scores. Previous work has shown the increase in overall accuracy and inclusion, particularly for low-to-moderate (LMI) populations Davuluri et al. (2019).

This study uses the data used to develop EnergyScore. Account-level credit account data collected between December 2009 and November 2016. Overall, over 800,000 observations of utility payment performance were collected across all fifty states and the District of Columbia. Credit history data, including FICO scores but also related utility payment performance history, was included as input variables.

Descriptive Statistics

Summary statistics are presented below in Table 1 for the four relevant variables used in this analysis: race, home ownership, education and race. Totals may not equal due to missing data. Notably, race suffers from a large degree of data marked as 'other', comprising 91.6% of the total. This category was dropped in the threshold analysis in the 'Results' section, so is not reported here.

To more accurately and inclusively quantify risk, several regression and machine learning algorithms were trained to predict risk of default payment. The best performing model was a random forest construction, which recorded the highest accuracy scores. This model also produced higher profits for lenders, as the EnergyScore provides a much more holistic metric on an individual's ability to pay utility bills. This better captures those individuals who would have been rejected due to lower FICO scores but able to pay (false negatives) while minimizing those who would have been accepted and failed to pay (false positives). Finally, the EnergyScore increases access for LMI customers, as an increase in both effectiveness for a larger applicant pool and an efficient manner of quantifying risk of late payment.

Case studies are useful illustrations of levels of bias measurable in machine learning applications. This research follows the approach of Hardt, Price and Srebro (2016), building upon their work in two distinct areas. First, we compare bias measurements for the machine learning algorithm (EnergyScore) and the

Table 1: Descriptive Statistics

Variable	N	Frequency	Average FICO	Late Payment %
Late Payment				
Late	304,836	34.9%	452.587	NA
Not Late	567,420	65.1%	733.680	NA
Race				
Asian	464	1.8%	664.759	0.291
Black	7,625	29.3%	548.416	0.662
Hispanic	6,118	23.5%	666.046	0.254
White	11,850	45.5%	602.321	0.538
Home Ownershp				
Own	675,017	77.4%	667.555	0.265
Rent	197,137	22.6%	525.558	0.638
Income				
Low	270,417	31%	712.349	0.170
Medium	285,222	32.7%	550.488	0.546
Education				
High	316,515	36.3%	646.337	0.326
College	182,783	21%	681.666	0.226
Graduate School	97,746	11.2%	702.509	0.184
High School	587,566	67.4%	609.634	0.416
Vocational/Technical	4,059	0.5%	678.342	0.204

counterfactual where credit scores are applied. Secondly, we extend the analysis across multiple protected classes; including race, income, education and homeownership status.

Bias Measurements

Quantitative measurements of bias and fairness are provided for both EnergyScore and the comparison FICO score. Bias will be measured by disparate effects on sub-groups in threshold scenarios. Additionally, the threshold scenarios can be completed by optimizing either FICO scores or EnergyScore.

Using the methodology designed by Hardt, Price and Srebro (2016), the following scenarios will be applied to our protected classes: profit maximization, race blind, demographic parity, equal opportunity.

Profit Maximization

This measurement assumes that lenders will seek to minimize false positives. Hence, a cutoff for applicants is required. A FICO cutoff of 620 is used for good credit as classified by the Consumer Financial Protection Bureau. From Figure 1, setting the setting the FICO score to 620 results in the *total* non-default rate of 82%; or the default rate to 18%. Hence, for profit maximization we will set the FICO scores accordingly so that each group achieves maximum 18% default rate. Notice how from the graph below, individual FICO cutoffs will differ widely, Hispanics having the lowest and Whites having the highest.

To arrive at the EnergyScore threshold, the computation includes analyzing the share of the group approved, then setting the EnergyScore threshold to the same proportion. This concept is seen in the idea of demographic parity, and was taken from the EnergyScore whitepaper.

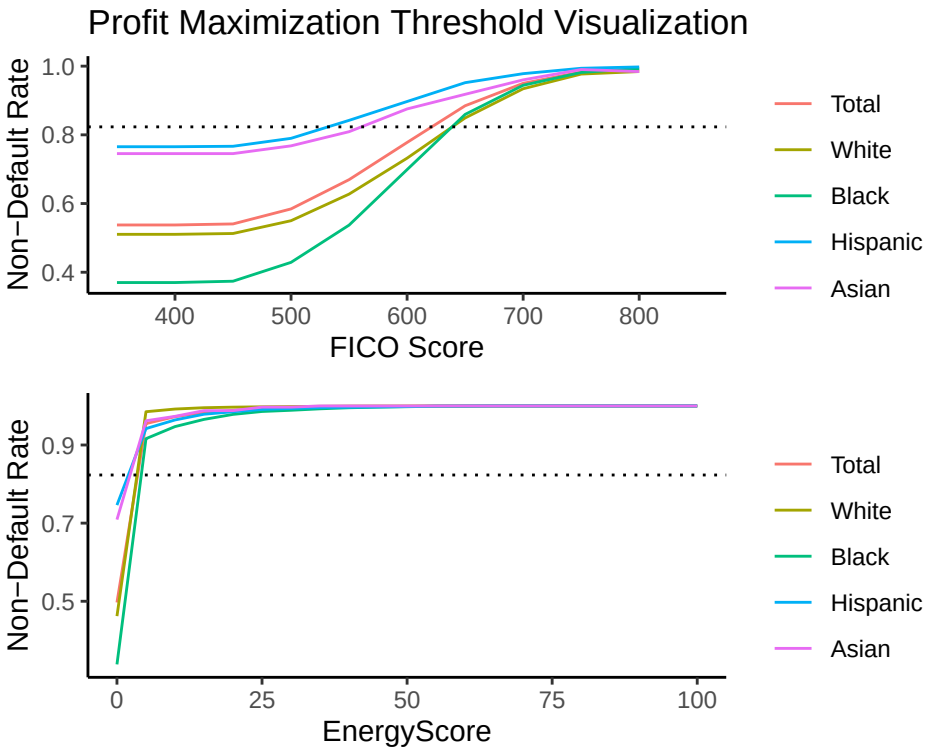

Figure 1: Profit Maximization Threshold

Race Blind

This criteria mimics the threshold construction steps completed in Profit Maximization however by only applying a single threshold to all groups. Hence to extend the previous example, all groups will be applied the same threshold of 620 for FICO. In threshold comparison section at end of each group, no variation will be observed.

Demographic Parity

This theory sets the thresholds different for each group such that the proportion of accepted is equal across all groups. This leads to divergent thresholds per group. Figure 2 below visualizes the result, as different subgroups, race in this instance, would receive different thresholds. The graph below shows the same cumulative distribution curves. In demographic parity constraints, the threshold is calculated by setting the proportion of population above that value.

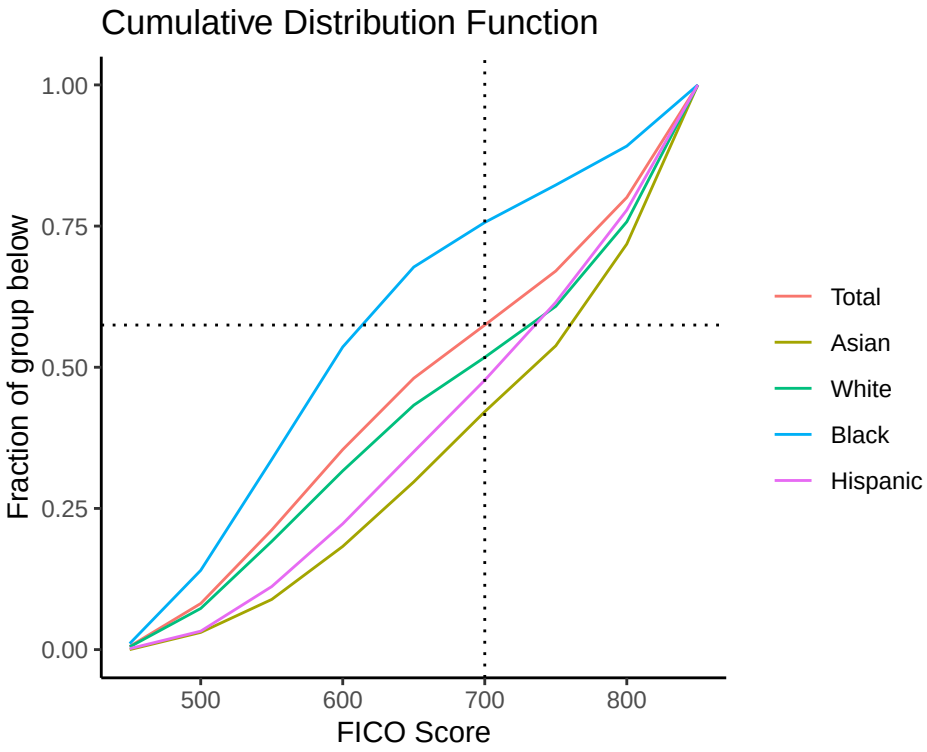

Figure 2: Demographic Parity Threshold

Equal Opportunity

The true positive rate, also known as sensitivity or recall, is the proportion of actual positive cases that are correctly identified by a model. The Equal Opportunity criteria sets the true positive rate equal across all groups. In the graph below, the true positive rate is shown with varying levels of FICO score cutoffs. In the example, the developer would choose a true positive level, in the below 70%, and apply the varying thresholds accordingly.

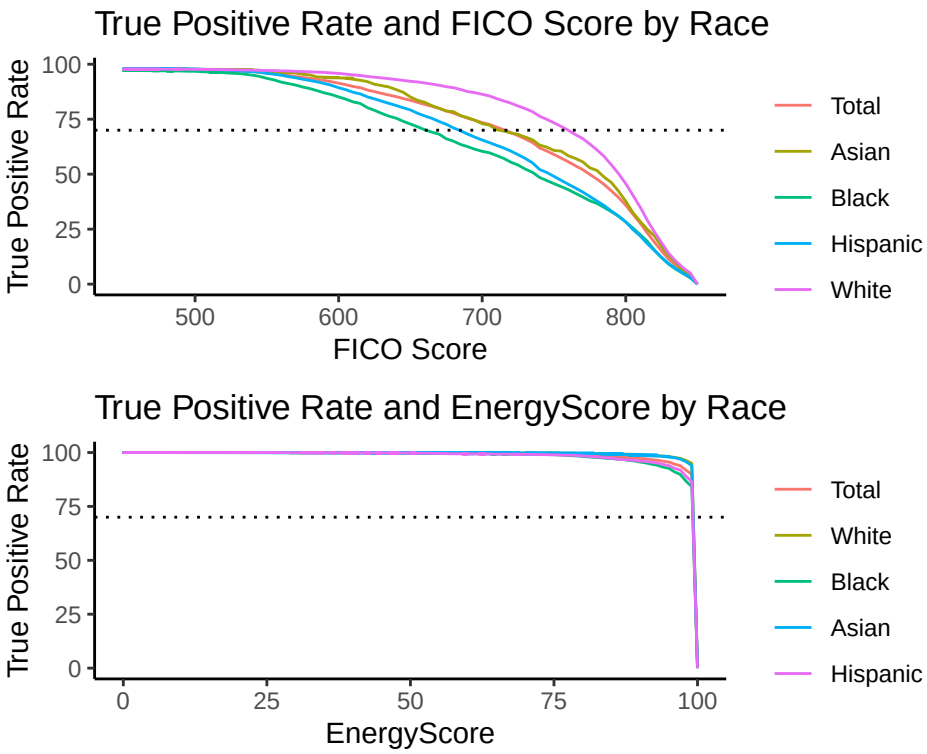

Figure 3: Equal Opportunity Threshold

Intra- and Inter-Group

Comparing distributions of thresholds from traditional bias measurements is a necessary but not sufficient step in our analysis. Expanding protected classes in this analysis provides more observations, but to determine how protected classes are treated by each algorithm, we propose using a common practice in mathematical statistics: intra- and inter-group variance.

Yang et al. (2022) previously found that expanding protected classes beyond race and gender is valuable in minimizing disparate impacts, particularly for marginalized groups. While the authors apply the intra- and inter-group theory to receiver operator curves (AUC), we apply a simplistic variance calculation for thresholds to determine how the respective risk classification metric treats different groups.

Results

To quantify how the four protected classes of interest are treated under the four threshold construction scenarios, we provide the following results. First, the thresholds for each protected class are shown on a continuum of the respective metrics range. This shows the relative difference in treatments within a particular metric.

Secondly, Intra-Group percentiles are shown to compare how individuals within a particular protected class are treated within the four threshold scenarios. For example, in Figure 6 the percentile value of the threshold is graphed. This second finding critically shows how individual protected classes are treated differently between the threshold scenarios.

Finally, the false positive curves are plotted. These visualize the predictive accuracy of the respective metric, particularly relevant here for organizations extending credit opportunities, as these examples represent errant approvals with associated losses.

Thresholds

In Figure 4 the thresholds are shown for Income and Race.

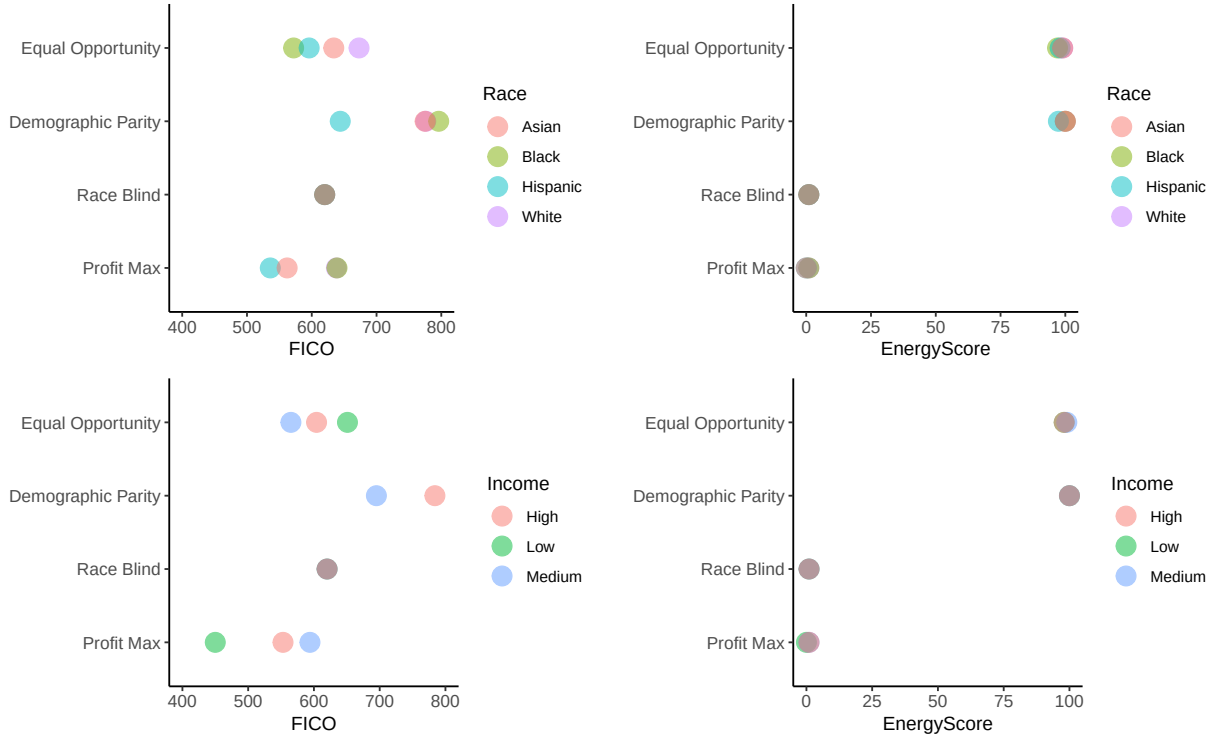

Figure 4: Cutoff Comparisons

In Figure 5 the thresholds are shown for Education and Homeownership.

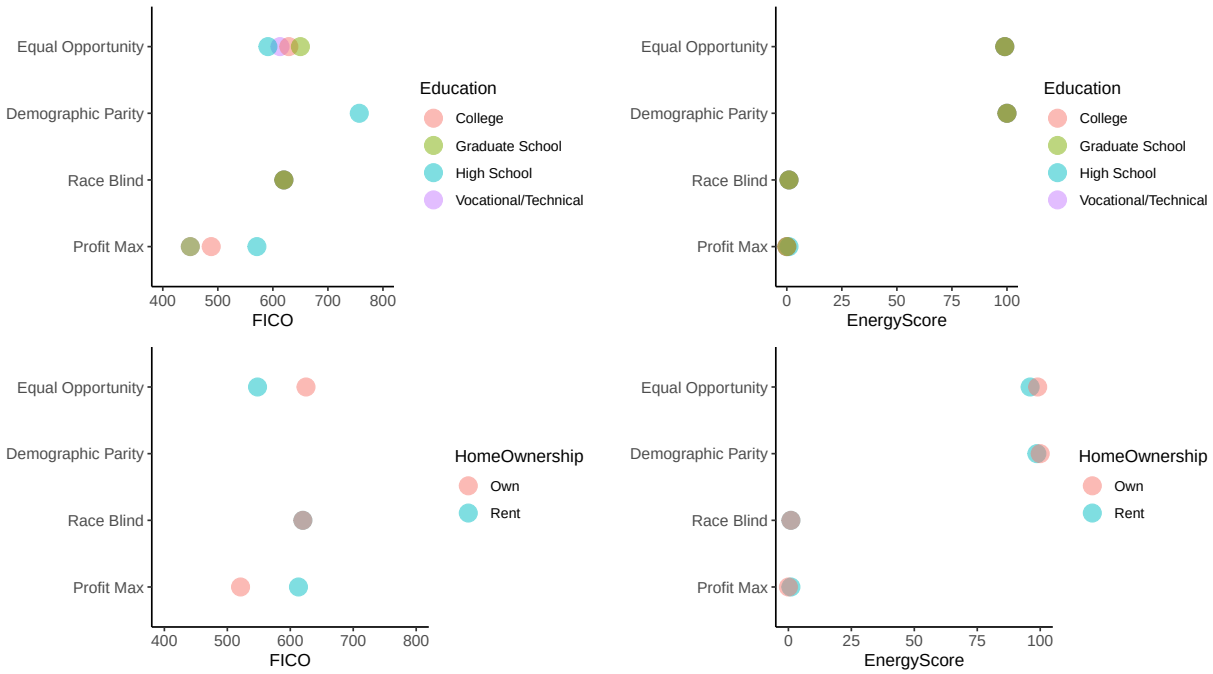

Figure 5: Cutoff Comparison Continued

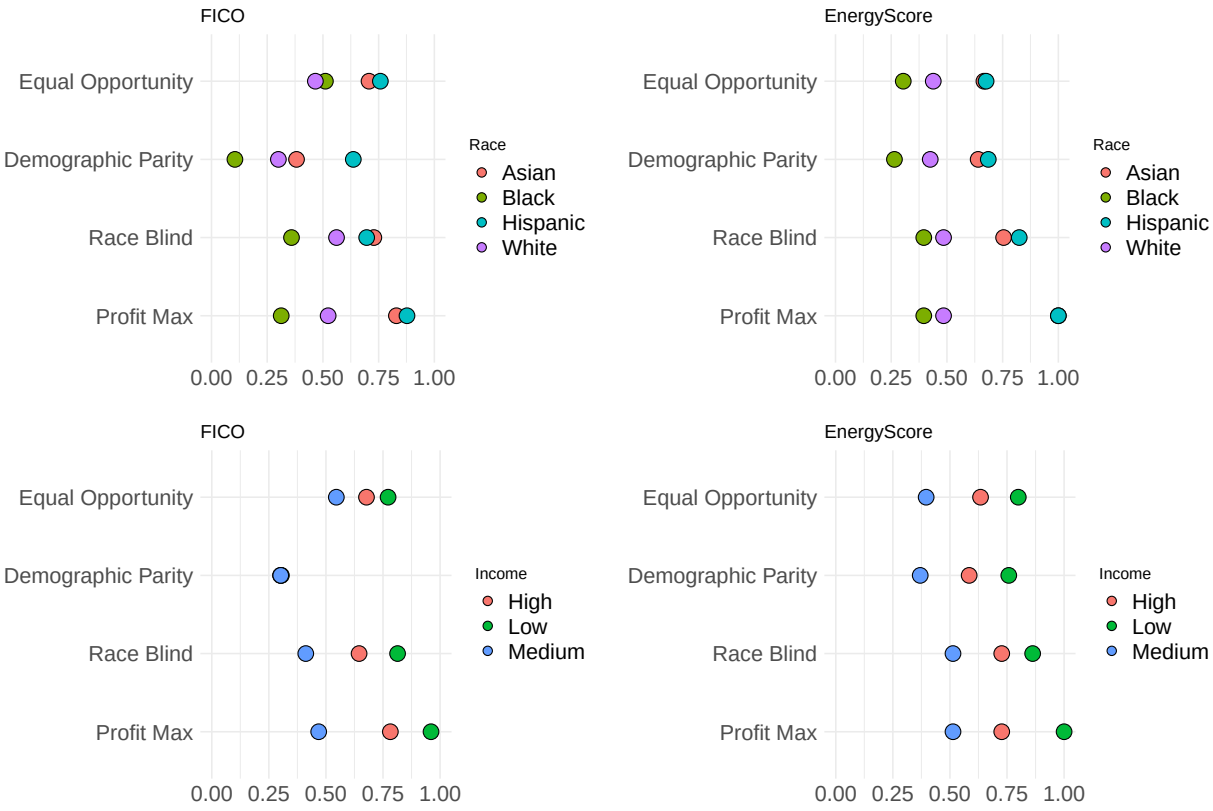

Figure 6: Intra-Group Percentiles Comparison

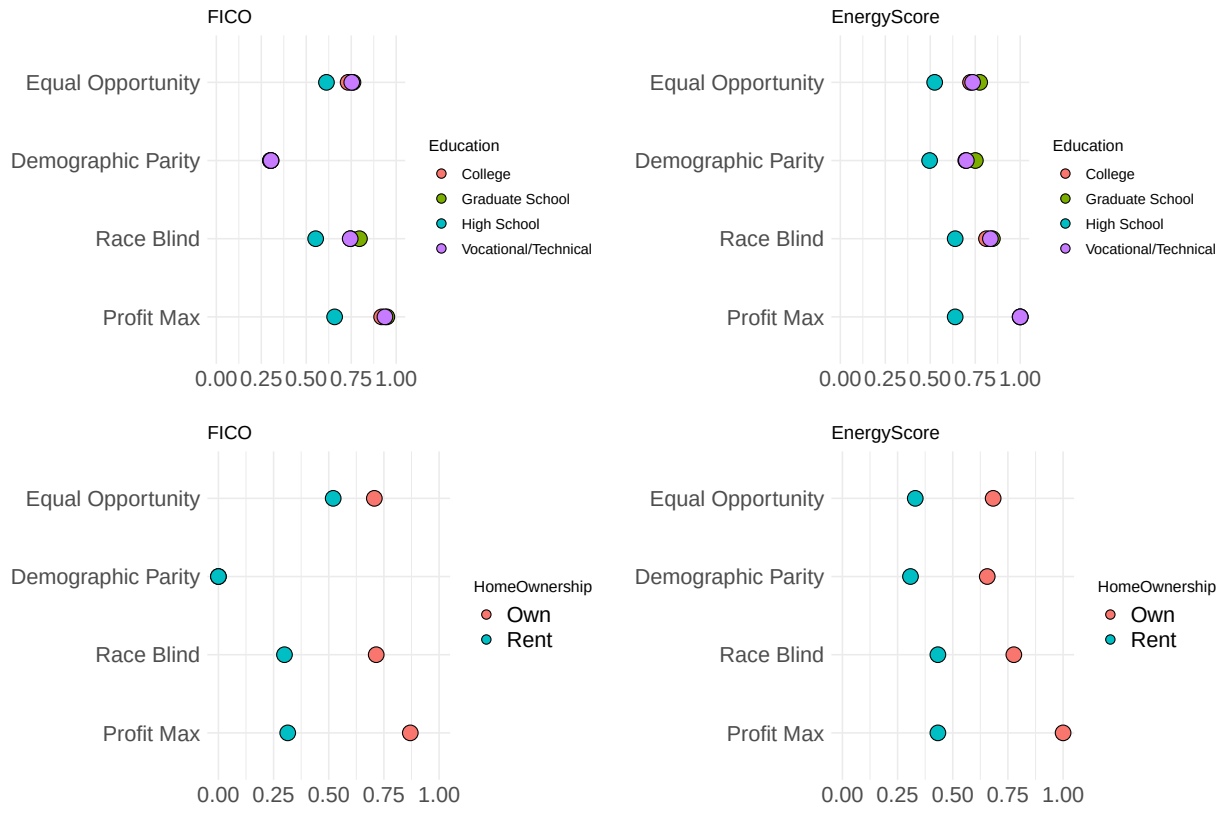

Figure 7: Intra-Group Percentiles Comparison Continued

Intra- and Inter-group

show the inter- and intra-group variance for our bias metrics, respectively.

False Positive Comparisons

In Figure 8 the false positives are shown for Race and Income

Table 2: Inter-Group Variance

temp	Variance_FICO	Variance_ES
Education		
College	0.069	0.019
Graduate School	0.077	0.013
High School	0.025	0.006
Vocational/Technical	0.072	0.018
Home Ownership		
Own	0.151	0.024
Rent	0.046	0.004
Income		
High	0.044	0.005
Low	0.080	0.011
Medium	0.010	0.006
Race		
Asian	0.038	0.027
Black	0.028	0.004
Hispanic	0.011	0.023
White	0.013	0.001

Table 3: Intra-Group Variance

variable	Variance_FICO	Variance_ES
Education		
Profit Max	0.019	0.033
Race Blind	0.012	0.009
Demographic Parity	0.000	0.013
Equal Opportunity	0.005	0.013
Home Ownership		
Profit Max	0.154	0.161
Race Blind	0.086	0.059
Demographic Parity	0.000	0.061
Equal Opportunity	0.017	0.062
Income		
Profit Max	0.062	0.059
Race Blind	0.041	0.031
Demographic Parity	0.000	0.038
Equal Opportunity	0.013	0.041
Race		
Profit Max	0.071	0.106
Race Blind	0.028	0.043
Demographic Parity	0.048	0.038
Equal Opportunity	0.021	0.033

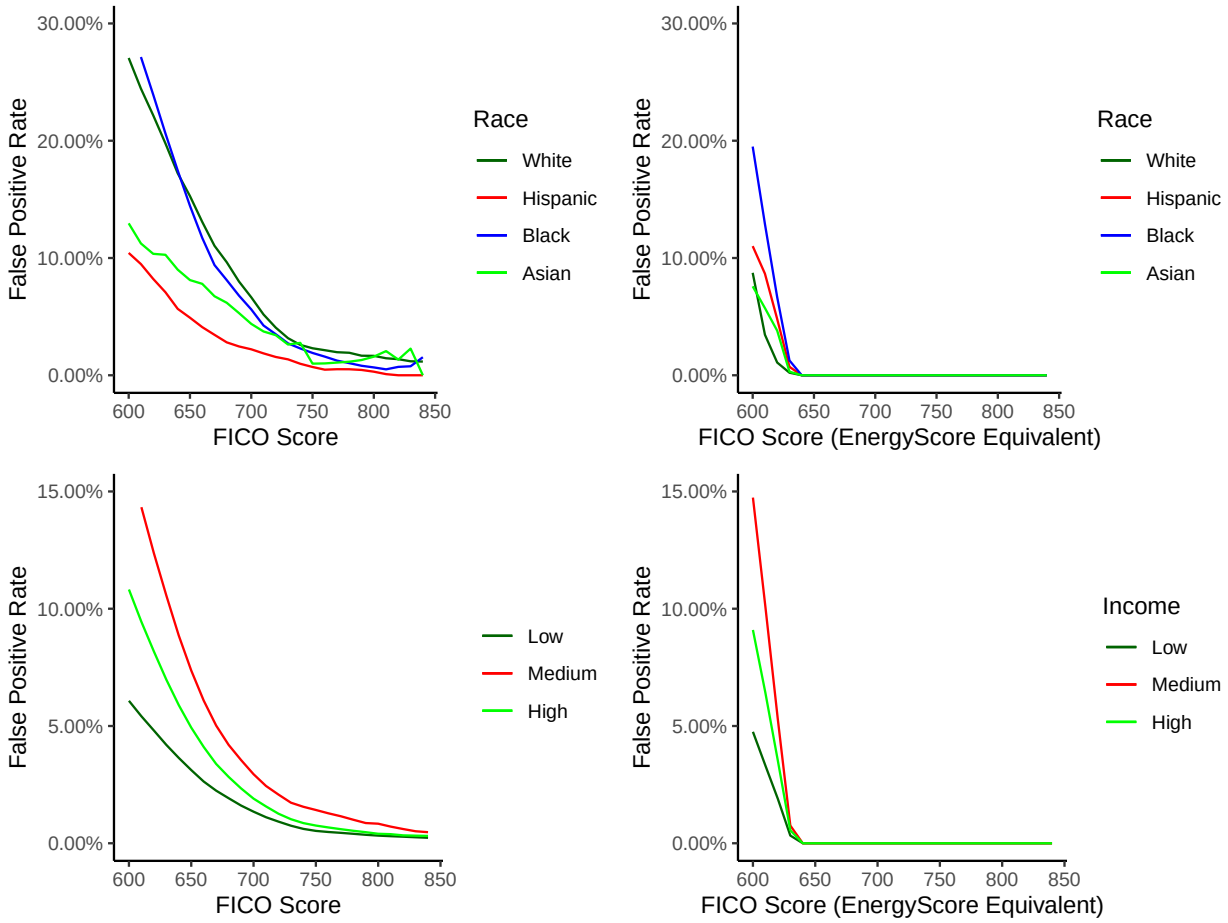

Figure 8: False Positive Comparison

In Figure 9 the false positives are shown for Race and Income

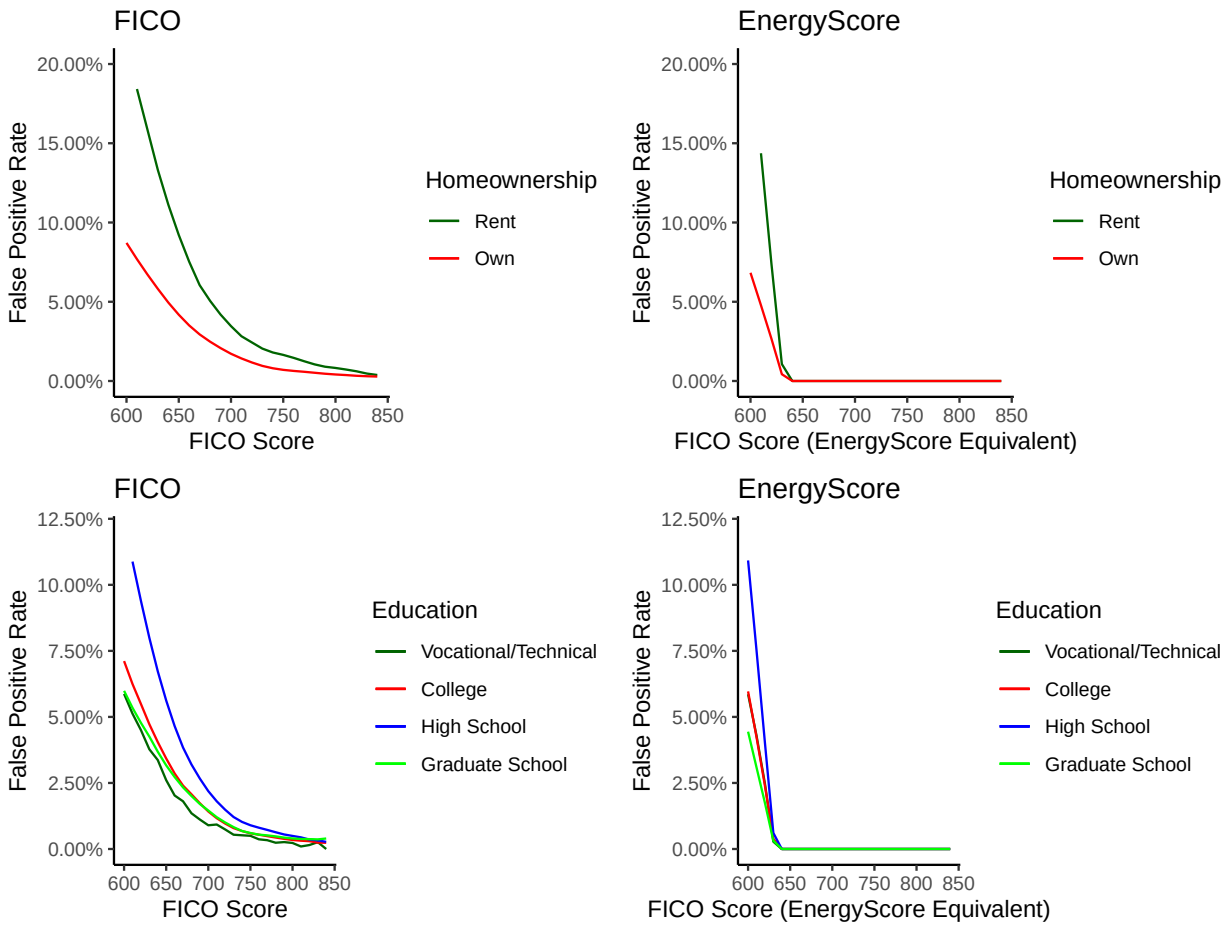

Figure 9: False Positive Comparison, Continued

Conclusion

Funding

This work was supported by the Tides Foundation grant TF2112-104450.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available as they include information relating to a pending patent application.

References

Davuluri, Sruthi, René García Franceschini, Christopher R Knittel, Chikara Onda and Kelly Roache. 2019. Machine Learning for Solar Accessibility: Implications for Low-Income Solar Expansion and Profitability. Working Paper 26178 National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w26178*

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard S. Zemel. 2011. "Fairness Through Awareness." *CoRR* abs/1104.3913.
URL: *http://arxiv.org/abs/1104.3913*

Hardt, Moritz, Eric Price and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *CoRR* abs/1610.02413.
URL: *http://arxiv.org/abs/1610.02413*

Suresh, Harini and John V. Guttag. 2019. "A Framework for Understanding Unintended Consequences of Machine Learning." *CoRR* abs/1901.10002.
URL: *http://arxiv.org/abs/1901.10002*

Yang, Zhenhuan, Yan Lok Ko, Kush R. Varshney and Yiming Ying. 2022. "Minimax AUC Fairness: Efficient Algorithm with Provable Convergence.".
URL: *https://arxiv.org/pdf/2208.10451.pdf*

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, ed. Sanjoy Dasgupta and David McAllester. Vol. 28 of *Proceedings of Machine Learning Research* Atlanta, Georgia, USA: PMLR pp. 325–333.
URL: *https://proceedings.mlr.press/v28/zemel13.html*