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

Jacob Ford, Senior Data Scientist

Solstice Power Technologies, LLC

160 Alewife Brok Parkway, #1048, Cambridge, MA 02138

jake@solstice.us

**Abstract.** 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.** Machine learning bias, protected classes, risk evaluation.

1. 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 increased. Various techniques for measuring bias in machine learning models have been developed. This study deploys commonly accepted machine learning bias metrics within a risk assessment algorithm, comparing the model’s bias comparisons to traditional credit score modeling. This research aims to achieve two objectives: first, to conduct a case study to assess the level of bias using traditional fairness criteria; and secondly, 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.

* 1. Background

Fairness through unawareness, an approach to protected class 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 cohesion 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 results of this algorithm show improved accuracy and significant fairness, as measured by both individual and group definitions, compared to alternative models.

Further, Suresh and Guttag (2019) highlight 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.

Other significant contributions to this field include Barocas, Hardt, and Narayanan (2019), who provide a comprehensive overview of fairness and machine learning, covering both technical and social considerations. They emphasize the need for transparency and accountability in the development and deployment of machine learning models. Kleinberg, Mullainathan, and Raghavan (2016) also discuss the theoretical underpinnings of fairness in machine learning, proposing fairness constraints that can be integrated into algorithm design to balance accuracy and equity.

Moreover, recent work by Mehrabi et al. (2021) offers a systematic survey of bias and fairness in machine learning, exploring various dimensions and sources of bias, as well as potential mitigation strategies. Their review highlights the complexity of achieving fairness and the ongoing need for interdisciplinary research to address these challenges effectively.

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. These measurements are described in greater detail in Section 2.3. First, we compare bias measurements for the machine learning algorithm (EnergyScore) and the counterfactual where credit scores are applied. Secondly, we extend the analysis across multiple protected classes; including race, income, education and homeownership status.

* 1. **EnergyScore**

For our case study, the machine learning algorithm analyzed, 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 dataset 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 traditionally used credit score.

1. Methods
   1. Data

This study uses the data used to develop EnergyScore, expanding the analysis into additional protected classes. Account-level credit profile data was collected between December 2009 and November 2016 to construct the model. 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. Demographic data on race, income, education and homeowner status was collected for each respondent, allowing for any credit risk model to be analyzed using commonly applied machine learning threshold techniques in determining how these protected classes are affected.

The output of the model is a probability of late payment on a utility tradeline account, used as a proxy for certain financed products such as community or residential solar, where often traditional credit scores are required.

* 1. Descriptive Statistics.

Summary statistics are presented below in Table 1 for the demographic protected classes. variables used in this analysis: 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.



**Table 1.** Descriptive Statistics

* 1. Bias Measurements

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.



**Fig. 1.** Profit Maximization Threshold for Race

### Race Blind

This criterion 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.



**Fig. 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.



**Fig. 3.** Equal Opportunity Threshold

### Intra- and Inter-Group Variance

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.

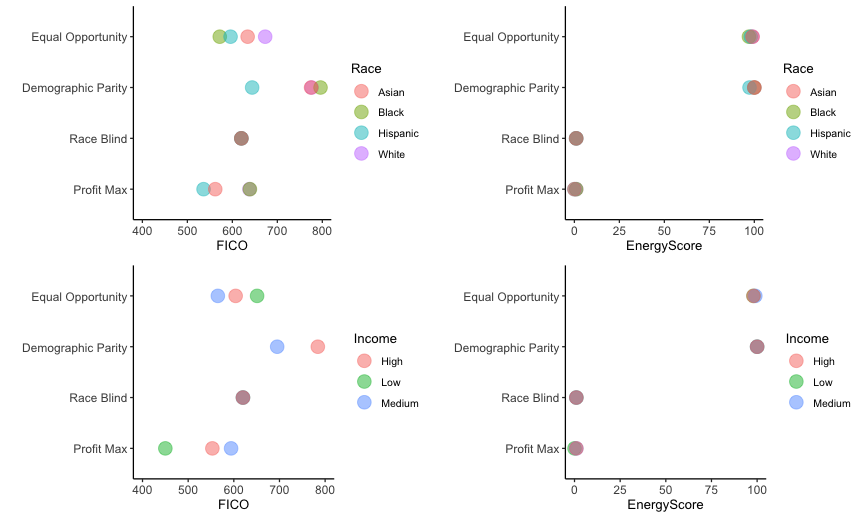
Yang et al. (2022) previously found that expanding protected classes beyond race and gender is valuable in minimizing disparate impacts, particularly for marginalized groups. While the authors apply the intra- and inter-group theory to receiver operator curves (AUC), we apply a simplistic variance calculation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

1. Results

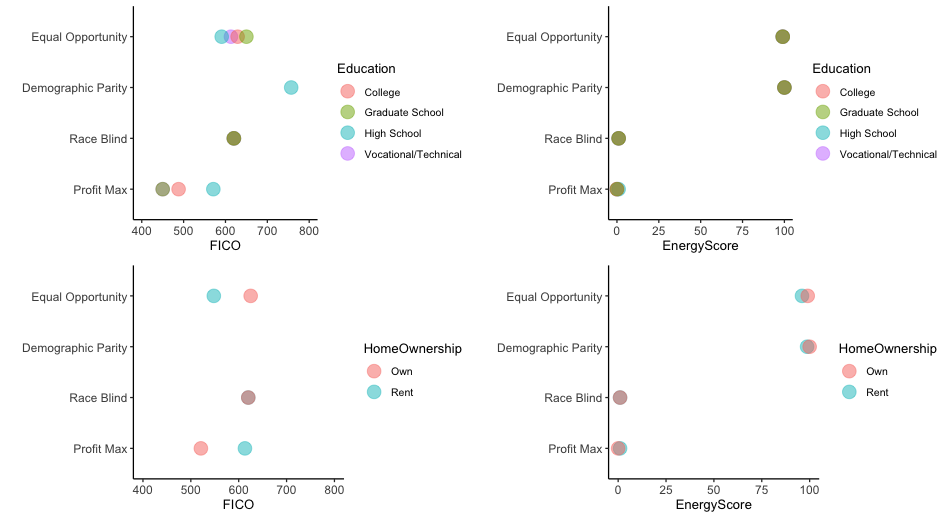
To quantify how the four protected classes of interest are treated under the four threshold construction scenarios, we provide the following results. First, the thresholds for each protected class are shown on a continuum of the respective metrics range. This shows the relative difference in treatments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

### Threshold Comparisons

In Figure 4 the thresholds are shown for Income and Race. Note how the thresholds for EnergyScore are nearly identical, whereas the FICO score distribution of thresholds at various optimization goals are widely dispersed. Similar distributions are seen in Figure 5 the thresholds are shown for Education and Homeownership.



**Fig. 4.** Income and Race Threshold Comparisons



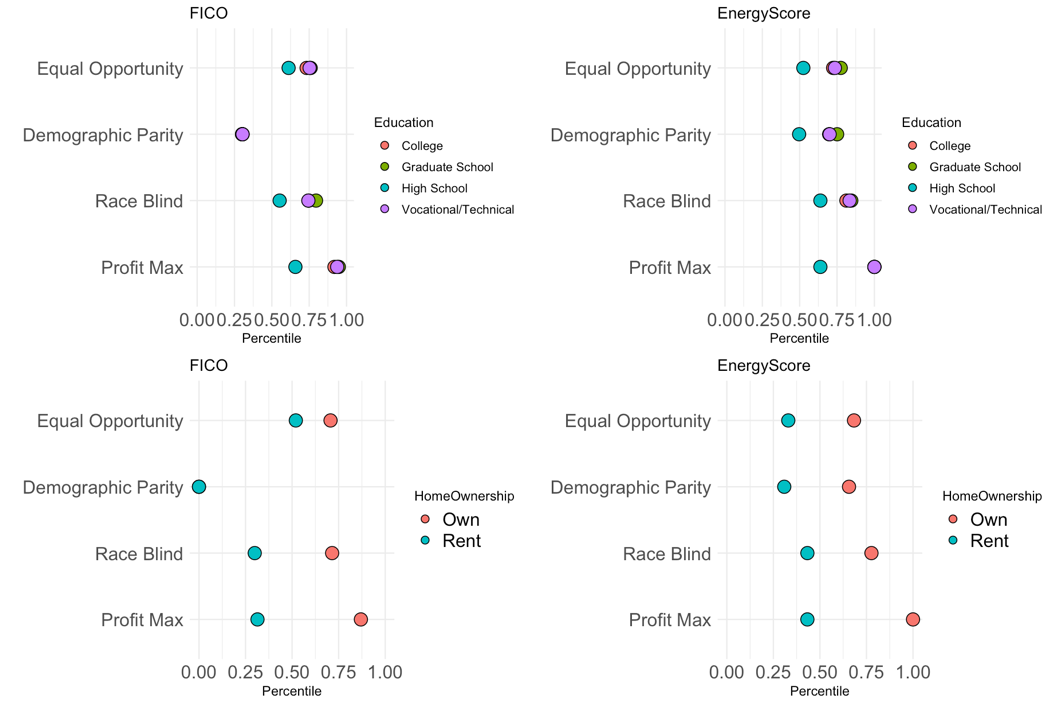
**Fig. 5.** Education and Homeowner Threshold Comparisons

### Intra-Group Percentiles

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.



**Fig. 6.** Race and Income Intra-Group Threshold Comparisons



**Fig. 7.** Education and Homeowner Intra-Group Threshold Comparisons

| Inter-Group Variance | | |
| --- | --- | --- |
| **Variable** | **FICO** | **EnergyScore** |
| **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 |

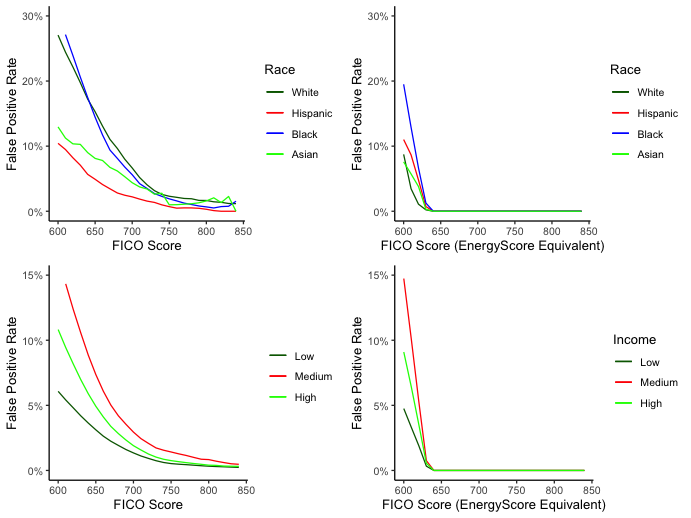
**Table 2.** Inter-Group Variance

| Intra-Group Variance | | |
| --- | --- | --- |
| **Variable** | **FICO** | **EnergyScore** |
| **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 |

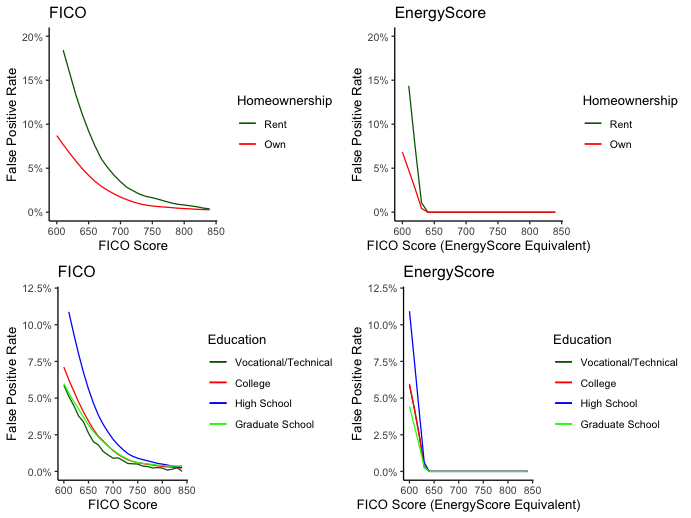
**Table 3.** Intra-Group Variance

### False Positive Curves

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.



**Fig. 8.** Race and Income False Positive Comparison



**Fig. 9.** Education and Homeowner False Positive Comparison

1. Conclusion

In summary, this study underscores the critical necessity of quantifying bias in machine learning applications using a variety of protected class definitions. Our findings highlight the potential for machine learning models to offer more equitable outcomes when appropriately designed and evaluated. The quantitative insights gained from our analysis reveal that the algorithms analyzed tends to exhibit lower variance in bias across different thresholds compared to conventional credit scores, suggesting a more consistent treatment of protected classes. Furthermore, our results emphasize the importance of extending bias analysis to include a broader range of protected categories, which can uncover subtle yet significant discriminatory practices that might otherwise go unnoticed.

This research contributes to the growing body of literature advocating for fairness in algorithmic decision-making and reinforces the imperative for ongoing vigilance and innovation in bias detection and mitigation techniques. Future research should continue to explore and refine methodologies for bias detection and mitigation to ensure fairness and inclusivity in predictive analytics. By doing so, we can work towards a more just and equitable application of machine learning technologies in various domains.

* 1. Funding

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

* 1. Data Availability

Data is not shared publicly due to ongoing patent application for PCT US2020 056147

References

1. Hardt, Moritz, Eric Price and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning.” CoRRabs/1610.02413.URL: <http://arxiv.org/abs/1610.02413>
2. Davuluri, Sruthi, Ren. Garc.a 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>
3. 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>
4. 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>
5. 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>
6. Zemel, Rich, YuWu, Kevin Swersky, Toni Pitassi and Cynthia Dwork. 2013. Learning Fair Representations .In Proceedings of the 30th International Conference on Machine Learning, ed. Vol. 28 of Proceedings of Machine Learning Research Atlanta, Georgia, USA: PMLR pp. 325–333. URL: <https://proceedings.mlr.press/v28/zemel13.html>
7. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org.
8. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores." arXiv preprint arXiv:1609.05807.
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). "A Survey on Bias and Fairness in Machine Learning." ACM Computing Surveys (CSUR), 54(6), 1-35.