Machine Learning Explainability & Fairness: Insights from Consumer Lending

Laura Blattner [1] **P-R Stark** [2] **Jann Spiess** [1] **Duncan McElfresh** [2] **Sormeh Yazdi** [2] **Georgy Kalashnov** [1]

Abstract

Stakeholders in consumer lending are debating whether lenders can responsibly use machine learning models in compliance with a range of pre-existing legal and regulatory requirements. Our work evaluates certain tools designed to help lenders and other model users understand and manage a range of machine learning models relevant to credit underwriting. Here, we focus on how certain explainability tools affect lenders' ability to manage fairness concerns related to obligations to identify less discriminatory alternatives for models used to extend consumer credit. We evaluate these tools on a "usability" criterion that assesses whether and how well these tools enable lenders to construct alternative models that are less discriminatory. Notably, we find that dropping features identified as drivers of disparities does not lead to less discriminatory alternative models, and often leads to substantial performance deterioration. In contrast, more automated tools that search for a range of less discriminatory alternative models can successfully improve fairness metrics. The findings presented here are extracted from a larger study that evaluates certain proprietary and open-source tools in the context of additional regulatory requirements (FinRegLab et al., 2022).

1. Introduction

In the context of consumer lending, model transparency largely functions as a means to an end in that it serves to further widely shared goals regarding anti-discrimination, consumer empowerment, and responsible risk-taking. For lenders and their regulators, model transparency is an essential instrument for evaluating whether a machine learning underwriting model can be used responsibly in that it helps firms enable internal and external oversight, manage risks, and document efforts to comply with law and regulation.

Consumer lending represents a "high stakes" use case for machine learning – one that can have a significant impact on people's financial lives, firms' safety and soundness, and communities' prosperity. Studying fairness in the context of consumer lending is compelling because machine learning models used to extend credit and additional techniques used to describe the behavior of those models must satisfy pre-existing anti-discrimination requirements. Accordingly, lenders, their regulators, and other stakeholders must address questions about whether certain technologies can be responsibly used.

Here, we present a subset of results from a larger study that considers the capabilities, limitations and performance of certain tools, proprietary and open-source, to help lenders manage machine learning underwriting models as required by law (FinRegLab et al., 2022). The main report evaluates certain model diagnostic tools with respect to properties of fidelity, consistency, and usability in the context of fair lending requirements and disclosures that must be given to recipients of certain kinds of adverse credit decisions. In this submission, we focus on the results pertaining to usability in the context of fair lending requirements: the ability of a tool to identify less discriminatory alternative models, a key component of fair lending requirements.

Participants of this study include seven financial technology companies and a set of open-source model explainability tools: SHAP (Lundberg & Lee, 2017a), LIME (Ribeiro et al., 2016), and permutation importance (Breiman, 2001). In keeping with industry practice, we calculate several metrics, including adverse impact ratio (AIR), standardized mean difference (SMD) for fairness, and AUC for model predictive performance.

Our study includes Logistic Regression and XGBoost underwriting models developed by the research team(the "Baseline Models") and models developed by some of the participating companies (the "Company Models"). Examples of Company Models include: Ensemble of Generalized Linear Models, Ensemble of Gradient Boosted

[1]School of Business, Stanford University, California, USA [2]FinRegLab, Washington D.C., USA. Correspondence to: Sormeh Yazdi <sormeh.yazdi@finreglab.org>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

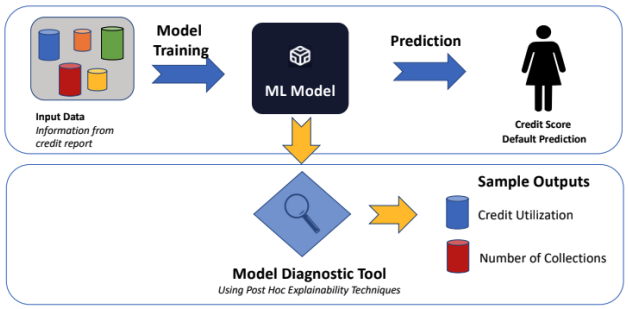

Figure 1. The methodology of this project is depicted in this graphic. Baseline and Company Models are trained with a given training dataset. Model explainability tools are applied, either providing some output requested by regulators, or resulting in models that have been altered to be more fair.

Machines, and Monotonicity-Constrained XGBoost Model. The companies worked with these models in a set of defined explainability tasks. Four out of seven companies provided recommendations for producing less discriminatory alternative models for both the Baseline and Company Models. Furthermore, we included an additional open-source tool developed by the research team in the less discriminatory alternative model analysis.

2. Related Literature

Our approach and findings directly relate to current debates in the academic and policy literature on the fair, inclusive, and responsible deployment of machine learning models.

First, we relate to work on the importance and challenges of explainability, interpretability, and transparency of complex machine learning models in critical applications (e.g. Lundberg & Lee, 2017b; Slack et al., 2019). Our work aims to add an economic notion of model transparency in the context of specific use cases to existing mathematical notions of complexity and explainability. Specifically, we argue that when used in critical applications, model descriptions should be related to specific policy goals and be interpreted in their specific context. We document that reasonable model descriptions may disagree (related to recent work about disagreement by Krishna et al., 2022). Second, we speak to a debate in computer science, economics, and law on different ways of restricting models to ensure their fairness and avoid discrimination (Barocas & Selbst, 2016). Our work adds to a growing list of theoretical, simulation-based, legal, and empirical findings that discuss the limits of input restrictions and consider alternatives (Kleinberg et al., 2018a; Gillis & Spiess, 2019; Gillis, 2022). Specifically, we show the promise of an approach that directly optimizes for specific policy targets, such as lowering disparities across groups. Furthermore, we show that input restrictions can be costly for performance with limited gain in reducing disparities. Third, while not the main focus of our study, we also contribute empirical evidence to discussions around different notions of disparity and fairness metrics (Hellman, 2020). Specifically, we show that different natural measures of disparities lead to different rank-orderings between models, and thus confirm existing theoretical and empirical findings that suggest that different notions of fairness cannot all be fulfilled at the same time but represent inherent trade-offs (Kleinberg et al., 2016; Chouldechova, 2017). At the same time, in our study there are also groups of (typically more complex) models that perform well across measures and dominate other (typically simple) models, suggesting that the choice of measures matters, as does the choice of model class, and a combination of model and optimization can generally improve properties across the board (related e.g. to Coston et al., 2021). Finally, we also relate to a discussion about trade-offs between model complexity, performance, and fairness in the case of consumer finance (Fuster et al., 2022; Bartlett et al., 2022). In our study, notwithstanding the limitations in the development of our underwriting models noted in the introductory section of (FinRegLab et al., 2022), the more complex models in this study generally outperform simpler models across measures of both disparity and model fit criteria, confirming the general potential of modern machine learning methods. At the same time, we observe fairness–performance trade-offs within complex models, tracing out a Pareto frontier that joint optimization can achieve.

While our evaluation studies the ability of feature-based model diagnostic tools to respond to specific policy and regulatory needs, we note that there are two other aspects of the transparency of automated machine learning systems that hold promise for their safe and fair use, which we do not deal with in depth in our report. First, we consider descriptions of models themselves, and not of the algorithms that produced these models. But one advantage of increasingly automated model-building pipelines is that they can often be described more completely and more accurately than the hand-curation of features and models by human model-builders, thus offering an opportunity for procedural scrutiny and transparency (e.g. Kleinberg et al., 2018b). Second, even complex models can be evaluated by simulating model behavior across hypothetical distributions of applicants or validating on actual segments of particular interest, thus potentially making their performance and critical properties available to regulators even before deployment. This form of "discrimination stress testing" (Gillis & Spiess, 2019) offers an alternative way towards transparency that holds promise even as models become more complex.

3. Experiments and Results

We test the ability of certain model diagnostic tools to help lenders identify less discriminatory alternative models. Results presented in our working paper analyze data returned from each tool in performing a series of defined tasks on Logistic Regression and XGBoost Baseline Models and the Company Models across test and deployment test data sets (FinRegLab et al., 2022). Our evaluation process included the following stages:

First, companies provided recommendations on how to improve the fairness of the models and provide less discriminatory alternatives for lenders to use. We evaluated whether the less discriminatory alternative models proposed by each participating company reduce adverse impact when additional test data are run through the models – and at what cost to predictive performance.

Next, companies generated less discriminatory alternative improvements using a deployment test data set with a different applicant composition. This test allowed us to assess how well these tools generalize to a different environment.

The companies used a variety of proprietary methods to identify less discriminatory alternative models. The techniques ranged from more automated tools, such as joint optimization techniques and adversarial debiasing, to feature dropping.[1]

Two responses suggested dropping the features most related to disparities based on information produced by the model diagnostic tools. These responses differ both in the number and identity of the features that were dropped. One response suggested a feature reweighting approach based on an open-source fairness tool. This response suggested new sample weights to be used in model re-training.

Three methods relied on some degree of automation in their search for less discriminatory alternative models. These approaches differ in whether and how they use protected class information in the search for and construction of less discriminatory models, which reflects different judgments about the permissibility of protected class information in the model building process.

One method, built by the research team, explicitly incorporates a version of the SMD statistic into the model training process. Here the machine learning algorithm now optimizes for a weighted sum of high predictive performance and low adverse impact. By varying the weight on the adverse impact in the algorithm's objective function, this joint optimization approach can trace out a menu of models that have different disparate impact and performance properties.

[1]In a prior part of the evaluation, the participating companies provided a list of top-10 features driving disparities in these models.

The second method combines this joint optimization approach with an adversarial debiasing technique.

A third method incorporates automation but does not explicitly consider protected class information the search for alternative models. Rather, this method searches over possible feature and hyperparameter combinations to identify a set of alternative models which can then – by a separate compliance team for example – be evaluated on the basis of their predictive performance and fairness properties.

For the company models, one company used a dual objective optimization approach. This algorithm is similar to adversarial debiasing, albeit with different implementation details. This debiasing routine considers two objective functions. The first function computes the AUC (or similar) metric which needs to be maximized, conditional on reaching the goal on the second objective which computes the bias metric (in this case, the AIR metric) is minimized below a target threshold. Both objectives are functions of model parameters. This resulting optimization problem is solved by an iterative mixed gradient approach.

Among the key findings, our results show that the ability to describe features that drive disparities with respect to a protected class does not automatically lead to models that are less discriminatory alternatives when this information is used mechanically. In other words, automated tools perform significantly better than strategies based on dropping features that were identified as drivers of disparities in the model. Furthermore, the more automated tools considered in this study generalize well to new data sets, and none outperforms at identifying a fairer alternative model at the lowest cost to predictive performance. We find that the automated tool that performs best depends on both the type of underwriting model and the specific metric of adverse impact considered.

4. Conclusion

While we did not test the full spectrum of potential bias mitigation approaches, implementation of specific company recommendations to drop a few individual features identified as most important in creating disparities does not significantly improve fairness and indeed imposes a cost in the form of significantly reduced predictive performance.

While additional research is warranted, our findings to date suggest that the traditional nexus between being able to identify key drivers of disparities and using that information for mitigation may be less applicable to managing disparate impact risks in machine learning underwriting models. Methods that rely on more automated approaches in their search for less discriminatory alternative models offer a notable contrast. Complex models in combination with tools that rely on some degree of automation can produce a menu of

model specifications that efficiently trade off fairness and predictive performance because they assess a broader range of features and incorporate fairness considerations into the model's development from the start.

Further Research Our work to date suggests a number of paths for additional inquiry. We plan to supplement this report later in 2022. The next publication will include two main additions:

- An extension of our evaluation of the capabilities, limitations, and performance of various model diagnostic tools in the context of the consumer protection requirements regarding adverse credit decisions and anti-discrimination requirements based on stakeholder input and further testing.
- An application of the evaluation framework used herein to assess the model diagnostic tools in the context of prudential model risk management expectations.

Once those further analyses are done, we will also consider holistically the implications of our findings across all three risk areas for the fair, responsible, and inclusive use of machine learning underwriting models and for the evaluation of approaches to explaining and managing machine learning models more generally.

References

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143(1):30–56, 2022.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Coston, A., Rambachan, A., and Chouldechova, A. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pp. 2144–2155. PMLR, 2021.

FinRegLab, Blattner, L., and Spiess, J. Machine learning explainability & fairness: Insights from consumer lending. 2022.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77 (1):5–47, 2022.

Gillis, T. B. The input fallacy. *Minnesota Law Review, forthcoming*, 2022.

Gillis, T. B. and Spiess, J. L. Big data and discrimination. *The University of Chicago Law Review*, 86(2):459–488, 2019.

Hellman, D. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pp. 22–27, 2018a.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018b.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017a. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-pre pdf.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017b.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. How can we fool lime and shap? adversarial attacks on post hoc explanation methods. 2019.