Explaining Quantitative Measures of Fairness

Scott M. Lundberg
Microsoft Research
Redmond, WA 98052, USA
scott.lundberg@microsoft.com

Abstract

Quantifying the fairness of a machine learning model has recently received considerable attention in the research community, and many quantitative fairness metrics have been proposed. In parallel to this work on fairness, explaining the outputs of a machine learning model has also received considerable research attention. Here we connect explainability methods with fairness measures and show how recent explainability methods can enhance the usefulness of quantitative fairness metrics by decomposing them among the model's input features. Explaining quantitative fairness metrics can reduce our tendency to rely on them as opaque standards of fairness, and instead promote their informed use as tools for understanding model behavior between groups.

Author Keywords

Explainable AI; fairness; transparency; interpretability.

Introduction

Quantitative fairness metrics seek to bring mathematical precision to our definitions of fairness in machine learning [2]. Definitions of fairness however are deeply rooted in human ethical principles, and so on value judgements that often depend critically on the context in which a machine learning model is being used. This dependence on value judgements manifests itself in the mathematics of quantita-

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

tive fairness measures as a set of trade-offs between sometimes mutually incompatible definitions of fairness [3]. Since fairness relies on context-dependent value judgements it is dangerous to treat quantitative fairness metrics as opaque measures of fairness [1], since doing so may obscure important value judgment choices.

The danger of treating quantitative fairness metrics as opaque, black-box measures of fairness is strikingly similar to a related problem of treating machine learning models themselves as opaque, black-box predictors. While using a black-box is reasonable in many cases, important problems and assumptions can often be hidden (and hence ignored) when users don't understand the reasons behind a model's behavior [5]. In response to this problem many explainable AI methods have been developed to help users understand the behavior of modern complex models [6, 5, 4]. In this article we propose applying these explainable AI methods to quantitative fairness metrics.

Applying Explainable AI Methods to Quantitative Fairness Metrics

While there are many ways to explain the predictions of a machine learning model, the most popular methods are based on additive feature attribution [4, 5]. These methods explain the output of a model as a sum of effects attributable to each input feature. If this decomposition is exact then for a model f applied to a set of inputs x you can represent the model output as a sum of feature impacts

$$f(x) = \phi_0(f) + \sum_{i}^{M} \phi_i(f, x)$$

where $\phi_i(f, x)$ is the impact feature x_i has on the model's output. Note that $\phi_0(f)$ is a constant bias term that does not depend on the current input x.

An important property of additive feature attribution methods is that they transfer the units of the model's output (such as log-odds or probabilities) onto the model inputs (since the $\phi_i(f, x)$ terms have the same units as $f(x)$). This means that any operation we previously would have done once on the model output, we can now repeat M times for each of the model's inputs. Since quantitative fairness metrics are computed by measuring expected differences in model outputs between specific groups of samples, we can also apply these metrics to the "partial model outputs" represented by the attribution values $\phi_i(f, x)$. Doing this effectively decomposes the quantitative fairness metric into M different components that when added together reproduce the quantitative fairness metric as applied to the original model output.

Decomposing a fairness metric among each of a model's inputs reveals which input features my be driving any observed fairness disparities. Since users often have a much better semantic understanding of model inputs than they do of model outputs, a feature-level decomposition of fairness that highlights a small number of features is much more actionable and useful than a single overall measure of model fairness.

A Simulated Case Study on Explaining Statistically Parity

To demonstrate the usefulness of explaining quantitative fairness metrics we consider a simple simulated scenario based on credit underwriting. In the simulation there are four underlying factors that drive the risk of default for a loan: income stability, income amount, spending restraint, and consistency. These underlying factors are not observed, but they influence four different observable features in various ways: job history, reported income, credit inquiries, and late payments. Using this simulation we generate 10,000 random samples and then train a non-linear gradient boosting tree classifier to predict the probability of default (the full simulated setup is available online at tinyurl.com/wpffhhl).

By introducing sex-specific reporting errors into the simulation we can observe how the biases caused by these errors are captured by fairness metrics. For this analysis we use the classic statistical parity metric, though the same analysis works with other metrics.

As a baseline experiment we refrain from introducing any sex-specific reporting errors. This results in no significant statistical parity difference between the credit score of men and women (Figure 1A top). When we decompose the statistical parity difference using the SHAP feature attribution method [4] we also see no significant feature-level differences (Figure 1A bottom).

When we introduce an under-reporting bias for women's income into the simulation we see a moderately significant statistical parity difference appear in the model's output (Figure 1B top). If this were a real application, this statistical parity difference might trigger an in-depth analysis of the model to determine what might be causing the disparity. While this investigation is challenging given just a single statistical parity difference value, it is much easier given the per-feature statistical parity decomposition based on SHAP (Figure 1B bottom), where there is a clearly significant bias coming from the reported income feature.

If we instead introduce an under-reporting bias for women's late payment rates we again see a moderately significant statistical parity difference for the model's output (Figure 1C top), and we now see a strong negative effect on women's default risk coming from the late payments feature (Figure 1C bottom).

Conclusion

Decomposing quantitative fairness metrics using explainable AI methods can reduce their opacity when the metrics are driven by measurement biases effecting only a few features. These "fairness explanations" enable users to better wrestle with the underlying value judgements inherent in fairness evaluation, and so may help reduce the risk of unintended consequences when using fairness metrics in real world contexts.

REFERENCES

[1] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[2] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

[3] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[4] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS 30*. 4768–4777.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*. 1135–1144.

[6] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.

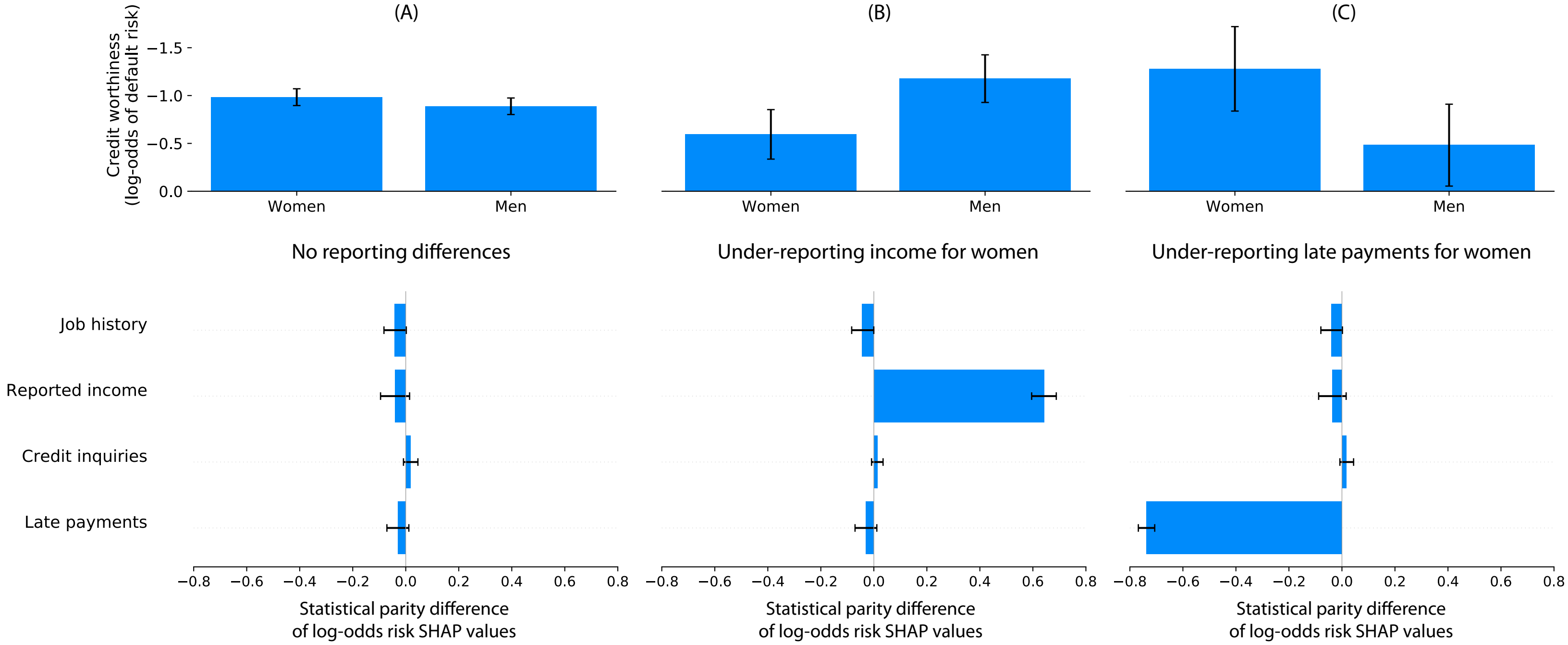

Figure 1: A simulation study on the effect of sex-specific reporting biases during credit scoring, and how SHAP values can be used to explain the resulting statistical parity differences. (A) On top is the mean credit worthiness of both men and women as predicted by a machine learning model. The difference between the two bar heights is the statistical parity difference (aka. demographic parity, one of the most common quantitative fairness metrics). On the bottom is the statistical parity difference among the SHAP values for each input feature to the model. The sum of the bars on the bottom equals the difference between the bars on the top. **(B)** The same analysis as done in (A), except that now the income of women is under-reported in the simulated data. This causes a moderately statistically significant disparity in the model's output, and a huge disparity in the SHAP values for the reported income feature. **(C)** The same analysis as (B) except now the number of late payments is under-reported for women (which causes the model to underestimate their risk).