

Fairness Beyond Binary Decisions: A Case Study on German Credit

Deborah D Kanubala¹, Isabel Valera^{1,2} and Kavya Gupta¹

¹Saarland School of Informatics, Germany

²MPI for Software Systems

Abstract

Data-driven approaches are increasingly used to (partially) automate decision-making in credit scoring by predicting whether an applicant is “creditworthy or not” based on a set of features about the applicant, such as age and income, along with what we refer here to as *treatment decisions*, e.g., loan amount and duration. Existing data-driven approaches for automating and evaluating the accuracy and fairness of such credit decisions ignore that treatment decisions (here, loan terms) are part of the decision and thus may be subject to discrimination. This discrimination can propagate to the final outcome (repaid or not) of positive decisions (granted loans). In this extended abstract, we rely on causal reasoning and a broadly studied fair machine-learning dataset, the German credit, to i) show that the current fair data-driven approach neglects discrimination in treatment decisions (i.e., loan terms) and its downstream consequences on the decision outcome (i.e., ability to repay); and ii) argue for the need to move beyond binary decisions in fair data-driven decision-making in consequential settings like credit scoring.

Keywords

algorithmic fairness, credit scoring, discrimination, path-specific counterfactual fairness

1. Motivation

In many areas such as hiring [1, 2, 3], law [4, 5, 6, 7], and finance [8, 9, 10, 11, 12], data-driven solutions are used in consequential decisions by predicting outcomes from historical data [13, 14]. The main assumption of these data-driven approaches for auditing or automating decision-making processes is the access to historical data $\mathcal{D} = \{s_i, \tilde{\mathbf{x}}_i, y_i\}_{i=1}^N$. In the context of loan approval [15, 16, 17], the available dataset is often assumed to contain a representative sample¹ of the random variables corresponding to i) sensitive attribute of applicants S ; ii) observed outcomes after a positive decision Y , which is used as a ground-truth label indicating the “creditworthiness” of applicants [19, 20, 21, 22]; and iii) features $\tilde{X} = \{X, Z\}$ that account for both the applicant characteristics X such as income, educational level, etc., along with the *treatment decisions* Z which in our case correspond to the loan terms such as duration, and loan amount, under which a historical positive decision (i.e., granted loan) was given.


EWAF’24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

*Corresponding author.

✉ dkanubala@aimsammi.org (D. D. Kanubala)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹For simplicity, we here assume that the data corresponds to independent and identically distributed samples and are sampled from an underlying data-generation process. However, we refer the reader to the literature on algorithmic decision-making under selective labels for relaxations of such an assumption [18].

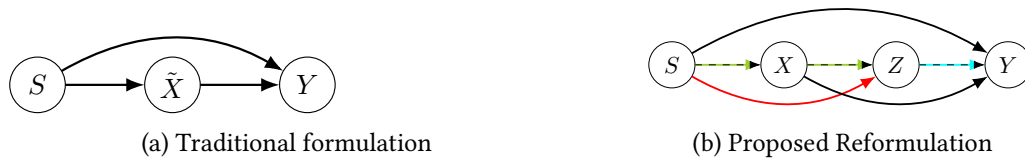


Figure 1: Causal graphs with different scenarios. a) without distinction between X and Z ; b) with distinction between X and Z . Green: paths considered unfair and are in the control of the decision maker (i.e bank). Red: Considered unfair as a result of direct discrimination and should not be accepted. Blue: downstream effect of discrimination from Z .

This setting, however, neglects that past decisions are not binary but also involve the treatment Z , which may, in turn, have a causal effect on the observed outcome Y . Studies have shown that *treatment decisions* Z can be discriminatory e.g., the authors of [23] found that while there were no significant differences in loan approval rates between binary gender identities, there was a substantial disparity in loan amounts between men and women. Similar studies [24, 25, 26, 27] also highlight the differences in loan terms between demographic groups. The results of these studies highlight that decisions are not binary and that discrimination against demographic groups may be neglected when considering binary decisions. That is while binary decisions may appear fair (e.g., in terms of acceptance rate [23] or true positive rate [25]) across groups, some demographic groups may still be subject to discrimination in the treatment they receive, i.e., loan terms. [23, 28, 29, 30].

While these works explored discrimination in *treatment decisions*, to the best of our knowledge none of these has considered the holistic analysis of the downstream effects of *treatment decisions* on the outcome. In this work, we go one step further and propose a setting to systematically answer the following questions about historical data of the form $\mathcal{D} = \{s_i, \mathbf{x}_i, z_i, y_i\}_{i=1}^N$, where we have now made explicit the distinction between applicant features X and treatment Z :

- **Research Question (RQ) 1 :** Does discrimination exist in the assignment of treatment Z (i.e., loan terms) across demographic groups?
- **Research Question (RQ) 2 :** In such a case, what are the downstream effects of such treatment discrimination on decision outcome Y (i.e., repayment probability)?

Implications: The common assumption that Y is sufficient for making lending decisions is inadequate. In other words, relying on discriminatory treatment decisions will propagate to the outcomes e.g., demographics perceived as higher risk may be offered loan terms that negatively affect their repayment probability, thus reinforcing the negative perception of their risk. In summary, our study aims to provide a compelling case for rethinking the existing fairness in the machine learning pipeline. Next, we detail how we can answer the above questions using a causality-based approach.

2. Discrimination in Treatment and its Effect on Outcome

We employ causal reasoning framework [31, 32, 33] for our work. More specifically, we contrast the causal graph in Figure 1a, which shows the current fair decision-making framework, with

the one in Figure 1b, where we consider the treatment decisions as part of the decisions made by the decision-maker. In our revisited data generation process in Figure 1b we make the following assumptions:

1. The *treatment decisions* Z is a causal child of both X and (potentially) S , and we assume the causal mechanism that *generated* Z to be potentially discriminatory due to a direct (solid red line Figure 1b) or indirect (dashed green) causal path. We refer to the direct causal effect of S on Z as *Direct Treatment Discrimination* (DTD), the indirect effect (mediated by X) of S on Z as *Indirect Treatment Discrimination* (ITD), and the combination of the two as *Treatment Discrimination* (TD).
2. The *treatment decisions* Z is a causal parent of the outcome variable Y , and thus TD may propagate to Y through the causal path(s) from S to Y mediated by Z (dashed blue line Figure 1b).

Based on these assumptions, we answer **RQ 1** and **RQ 2** as follows.

RQ 1 Discrimination in treatment: We assume that for each sample *factual* applicant in the dataset, which we denote by (s^F, x^F, z^F, y^F) , we can rely on the *action-prediction-abduction* steps by Pearl [34] to estimate its *sensitive counterfactual* $(s^{CF}, x^{CF}, z^{CF}, y^{CF})$. Then, we can define the (total) discrimination in treatment as the treatment difference between the factual individual z^F and its sensitive counterfactual z^{CF} , i.e.,

$$TD = z^{CF} - z^F, \quad \text{where } z^{CF} = Z(s^{CF}, X(s^{CF})). \quad (1)$$

Here, $Z(s, x)$ and $X(s)$ denote the causal function² that define the value of respectively the random variables Z and X , respectively, given their causal parents according to the causal graph in Figure 1b. Equation 1 quantifies the total discrimination. To disentangle the sensitive attribute's effect through different pathways, we follow path specific in [32] to rewrite to as:

$$TD = DTD + ITD, \text{ where} \\ DTD = z^{CF} - Z(s^F, X(s^{CF})) \quad \text{and} \quad ITD = Z(s^F, X(s^{CF})) - z^F. \quad (2)$$

RQ 2 Effect of discrimination in treatment (Z) on outcome (Y): This is simply the repayment odds³ of a *factual* applicant (s^F, x^F, z^F) *being treated* according to its true sensitive attribute to as it would have been treated according to its *counterfactual*, i.e., giving a *treatment decision* Z of a factual (female) to its counterfactual (male). We refer to this as *Treatment Discrimination Effect* (TDE) which we computed as:

$$TDE = \frac{\text{odds}(p^{CF})}{\text{odds}(p^F)} = \exp [Y(s^F, x^F, z^{CF}) - Y(s^F, x^F, z^F)] \quad (3)$$

²We make implicit the dependence of the causal functions to the exogenous variables. In addition, assume the absence of a hidden confounder, or equivalently, we assume causal sufficiency. For more details on structural causal models, refer to [35].

³Repayment odds refers to the likelihood that a borrower will successfully repay a loan [36] and is computed as $\text{odds}(p) = p/(1-p)$

where the repayment probability for an individual (s, x, z) is given by $p = P(Y = 1|s, x, z) = \sigma(Y(s, x, z))$, with $\sigma(\cdot)$ denoting the logistic function. Importantly, we can interpret the *TDE* values as follows:

- a) If $TDE \leq 1$, then the odds of s^F repaying the credit are equal or lower, meaning no negative downstream effect of treatment on the outcome.
- b) Otherwise, if $TDE > 1$, s^F is more likely to repay credit than its sensitive counterfactual s^{CF} . In this case, we consider s^{CF} has been subject to discrimination.

3. A Case Study Using German Credit

We analyze the German Credit dataset [37], using loan amount and duration as *treatment decisions*. Assuming additive (noise) linear causal functions, we learn the parameters of our causal model. For **RQ 1**, we measure the discrimination in treatment along the different pathways. Our results align with existing literature [23, 25, 27] and reveal that discrimination exists in treatment. Table 1, shows that males receive on average an increase of 10% and 20% in duration and credit amount respectively. While this may allow extending the loan terms over a long period, this often also means paying interest over the life of the loan and poses a higher risk of defaulting [38]. On the other hand, females receive on average a significantly lower credit amount and shorter duration. Our second **RQ 2** was to measure how discrimination in treatment propagates to the outcome. Following our analysis, hypothetically treating a female (factual) like you would have treated a male (counterfactual) on average decreases the repayment odds by 9%. On the other hand, a hypothetical treating males (factual) like females (counterfactual) results in an increase of repayment odds by 10%. Thus, we conclude that even though males receive preferential treatment with a higher amount of loan and longer duration, this treatment however has a negative downstream effect on their ability to pay back the loan. As such, the disparity in treatment across groups increases risks for male borrowers and puts male borrowers in a higher risk situation.

Table 1

German Credit: Treatment Discrimination and its downstream effects. We provide both the discrimination measures and also the transformed values of their odds ratio by setting s^F to its s^{CF} . (mean: μ , standard deviation: σ).

Measure	Path	Duration : $\mu(\sigma)$	Amount : $\mu(\sigma)$	Repayment Odds : $\mu(\sigma)$
DTD	$s \rightarrow z \rightarrow y$	0.068(0.015)	0.1737(0.028)	-0.041(0.006)
ITD	$s \rightarrow x \rightarrow z \rightarrow y$	0.0941(0.024)	0.059(0.035)	-0.053(0.014)
TD	both	0.162(0.022)	0.232(0.057)	-0.094(0.008)
TDE	Male \rightarrow Female	\downarrow 9%	\downarrow 16%	\uparrow 10%
	Female \rightarrow Male	\uparrow 10%	\uparrow 20%	\downarrow 9%

4. Open questions

We have shown from our analysis that there is discrimination in treatment and this propagates to the predictive outcome. Our study provides a compelling argument for rethinking the entire pipeline of ML fairness. Although we restricted our analysis to the German Credit dataset, the

implications extend to various other domains such as criminal justice, hiring, education, etc. Furthermore, ensuring fairness in observed outcomes Y may be inadequate to mitigate bias as this is still a function of biased treatment. This could also lead to a never-ending loop and in the worst case can worsen the financial situation of discriminated groups. These results prompt us to question the current framework and raise several open questions: *Is there a need to develop new notions of fairness, considering that Y remains a composition of an unfair Z ? What does designing a fair policy for Z entail? What types of datasets are necessary to ensure fairness in non-binary decision-making processes?*

5. Acknowledgements

This work has been funded by the European Union (ERC-2021-STG, SAML, 101040177). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

- [1] E. Faliagka, K. Ramantas, A. Tsakalidis, G. Tzimas, Application of machine learning algorithms to an online recruitment system, in: International Conference on Internet and Web Applications and Services, 2012.
- [2] J. Silas, P. Udhan, P. Dahiphale, V. Parkale, P. Lambhate, Automation of candidate hiring system using machine learning, International Journal of Innovative Science and Research Technology (2023).
- [3] A. A. Mahmoud, T. A. Shawabkeh, W. A. Salameh, I. Al Amro, Performance predicting in hiring process and performance appraisals using machine learning, in: 10th international conference on Information and communication systems, 2019.
- [4] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022.
- [5] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, Northpointe Inc (2016).
- [6] M. Hamilton, The sexist algorithm, Behavioral sciences & the law (2019).
- [7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research (2021).
- [8] A. S. Almheiri, Automated loan approval system for banks, 2023.
- [9] S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research (2015).
- [10] D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla, B. R. Reddy, Experimental analysis of machine learning methods for credit score classification, Progress in Artificial Intelligence (2021).
- [11] V. Moscato, A. Picariello, G. Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Systems with Applications (2021).

- [12] J. Sirignano, A. Sadhwani, K. Giesecke, Deep learning for mortgage risk, arXiv preprint:1607.02470 (2016).
- [13] T. Scantamburlo, A. Charlesworth, N. Cristianini, Machine decisions and human consequences, arXiv preprint:1811.06747 (2018).
- [14] A. Coston, Principled Machine Learning for Societally Consequential Decision Making, Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA, 2023.
- [15] C. Hurlin, C. Pérignon, S. Saurin, The fairness of credit scoring models, arXiv preprint:2205.10200 (2022).
- [16] M. Rajesh, A. Lakshmanarao, C. Gupta, An efficient machine learning classification model for credit approval, in: Third International Conference on Artificial Intelligence and Smart Energy, 2023.
- [17] K. Bhatt, P. Sharma, M. Verma, K. Agarwal, Loan status prediction in the banking sector using machine learning, in: International Conference on Computational Intelligence, Communication Technology and Networking, 2023.
- [18] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, S. Mullainathan, The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.
- [19] R. Dobbe, S. Dean, T. Gilbert, N. Kohli, A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, arXiv preprint:1807.00553 (2018).
- [20] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions, *The quarterly journal of economics* (2018).
- [21] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, *Annual Review of Statistics and Its Application* (2021).
- [22] B. Green, L. Hu, The myth in the methodology: Towards a recontextualization of fairness in machine learning, in: Proceedings of the machine learning: the debates workshop, 2018.
- [23] I. Agier, A. Szafarz, Microfinance and gender: Is there a glass ceiling on loan size?, *World development* (2013).
- [24] C. L. Escalante, A. Osinubi, C. Dodson, C. E. Taylor, Looking beyond farm loan approval decisions: loan pricing and nonpricing terms for socially disadvantaged farm borrowers, *Journal of Agricultural and Applied Economics* (2018).
- [25] A. F. Alesina, F. Lotti, P. E. Mistrulli, Do women pay more for credit? evidence from italy, *Journal of the European Economic Association* (2013).
- [26] D. Aristei, M. Gallo, Are female-led firms disadvantaged in accessing bank credit? evidence from transition economies, *International Journal of Emerging Markets* (2022).
- [27] Y. Li, Gender differences in car loan access: An empirical analysis, in: Proceedings of the 12th International Conference on E-business, Management and Economics, 2021.
- [28] A. Cozarenco, A. Szafarz, Women's access to credit in france: how microfinance institutions import disparate treatment from banks, Available at SSRN 2387573 (2014).
- [29] I. Agier, A. Szafarz, Credit to women entrepreneurs: The curse of the trustworthier sex, Available at SSRN 1718574 (2010).
- [30] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, A. Walther, Predictably unequal? the effects of machine learning on credit markets, *The Journal of Finance* (2022).
- [31] D. Plecko, E. Bareinboim, Causal fairness analysis, arXiv preprint:2207.11385 (2022).

- [32] S. Chiappa, Path-specific counterfactual fairness, in: Proceedings of the AAAI conference on artificial intelligence, 2019.
- [33] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, *Advances in neural information processing systems* (2017).
- [34] J. Pearl, et al., *Models, reasoning and inference*, Cambridge, UK: Cambridge University Press (2000).
- [35] J. Pearl, M. Glymour, N. P. Jewell, *Causal inference in statistics: A primer*, John Wiley & Sons, 2016.
- [36] M. Szumilas, Explaining odds ratios, *Journal of the Canadian Academy of Child and Adolescent Psychiatry* (2010).
- [37] H. Hofmann, Statlog (german credit data) data set, UCI Repository of Machine Learning Databases (1994).
- [38] Z. Guo, Y. Zhang, X. S. Zhao, Risks of long-term auto loans, *Journal of Credit Risk* (2022).