

Causal Inference for Social Discrimination Reasoning

Bilal Qureshi¹, Faisal Kamiran², Asim Karim³, Salvatore Ruggieri⁴, and Dino Pedreschi⁴

¹ Addo.ai, Pakistan bilal@addo.ai

² Information Technology University of the Punjab, Pakistan
faisal.kamiran@itu.edu.pk

³ Lahore University of Management Sciences (LUMS), Pakistan
akarim@lums.edu.pk

⁴ Università di Pisa and ISTI-CNR, Italy {ruggieri, pedre}@di.unipi.it

Abstract. The discovery of discriminatory bias in human or automated decision making is a task of increasing importance and difficulty, exacerbated by the pervasive use of machine learning and data mining. Currently, discrimination discovery largely relies upon correlation analysis of decisions records, disregarding the impact of confounding biases. We present a method for *causal* discrimination discovery based on *propensity score analysis*, a statistical tool for filtering out the effect of confounding variables. We introduce causal measures of discrimination which quantify the effect of group membership on the decisions, and highlight causal discrimination/favoritism patterns by learning regression trees over the novel measures. We validate our approach on two real world datasets. Our proposed framework for causal discrimination has the potential to enhance the transparency of machine learning with tools for detecting discriminatory bias both in the training data and in the learning algorithms.

Keywords: Social Discrimination · Fairness, Accountability, and Transparency · Propensity Score · Causal Analysis.

1 Introduction

Discrimination can be instilled in automated decision-making systems at several places: it can be learned from training data; it can be a side-effect of biased algorithms; or, it can arise from “dumb-smart” programs that are used in contexts they are not designed or tested for. Understanding and mitigating forms of bias in data/web mining systems, and discrimination against protected social groups in particular, has been a growing research field in the last years, due to mounting awareness and concerns⁵.

The large body of research on discrimination and fairness [4,29], however, largely ignores that reasoning on observational data through correlation analysis may produce biased results, due to *confounding variables*. Consider, e.g., the well-known UC Berkeley gender bias case [7]: the admission outcomes for the fall of 1973 seemed to show that men were more likely than women to be admitted, a large difference indicating a potential discrimination. But dis-aggregating the numbers by the different departments,

⁵ See e.g., the AI Now report on *Discriminating Systems*, April 2019 (<https://ainowinstitute.org>).

it emerged a “small but statistically significant bias in favor of women”, due to fact that women tended to apply to competitive departments, whereas men tended to apply to less-competitive departments. Hence, the department confounds or biases the observed difference in admittance rates for males and females. The department here is a confounding variable because it is correlated both with the admittance decision and the gender of an applicant. It can also be thought of as an explanatory attribute that explains the overall preference for males, revealing that, in this case, gender discrimination did not occur. In a given analysis, there may be several confounding variables, and it is essential that their effect is minimized while estimating discrimination. Once accounted for confounding variables, we can reason on the *causal* effect of, say, gender, on the admittance decision.

Discovering and understanding causal influences among variables is a fundamental goal of any data analysis process. In general, randomized control trials (RCT) are the gold-standard for inferring such influences. This is the case, for example, of online controlled experiments [21], which are widely used to test for effectiveness of website personalization and of advertising strategies. Unfortunately, in discrimination discovery processes, RCTs are rarely possible or practical. In the vast majority of cases only observed data are available, along with domain expert’s knowledge. Nonetheless, causal influences are more likely to be acceptable in court of law than mere correlations or statistical test of hypothesis alone [16]. This is why recent works, discussed in Section 5, propose to reason on causal discrimination based on a network of causal relations among variables, specified with the intervention of a domain expert. However, this approach is unpractical in real high-dimensional problems, such as those arising in web analytics, as the difficulty of estimating causal inter-dependencies scales super-linearly with the number of features, quickly becoming intractable.

In this paper, we introduce an approach for causal discrimination discovery and understanding that does not require extensive expert’s intervention, based on *propensity score analysis* (PSA). The propensity score is the probability of membership to the protected group (e.g., females) given the individual’s or group’s attributes. PSA provides a principled way to ‘filter out’ the explainable effect of confounding variables and to quantify the causal effect of the protected attribute (e.g., gender). In the UC Berkeley example above, e.g., PSA accounts for the effect of the departments to which candidates apply and confirms that there is no evidence of women discrimination. We adopt propensity score weighting to balance the distributions of protected and unprotected group instances. Based on the modified distribution, we quantify the degree of discrimination or favoritism of each instance. More generally than a standard application of PSA, we learn a regression tree to understand the characteristics of instances falling under different discrimination/favoritism bands. We evaluate our approach on real-world datasets. The results confirm that the causal effect of group membership can be higher or lower than the effect computed by ignoring the influence of confounding variables. Thus, our approach promises defensible discrimination analyses in practice.

Our method has relevant impact towards data-driven machine learning processes aimed at developing decision support systems. *First*, causal discrimination discovery provides a solid reasoning framework to detect biases contained in the training data, which may be inherited by the learned decision model. In turn, this prevent constructing

black-box systems that incorporate discriminatory rules (see, e.g., [14]) and whose decisions, by definition of black-boxes, cannot be easily interpretable (see [18]). *Second*, our proposed regression tree model for estimating the discrimination risk score offers a method for understanding the causal structure of the biases, and therefore repairing or sanitizing the training data for avoiding unwanted discrimination in the learned model. *Third*, similar considerations apply to the validation of a learned model on test data, in that causal discrimination discovery may be applied to the test examples labeled by the learned model, to assess any potential algorithmic bias introduced by the adopted machine learning technique. Thus, our proposed causal discrimination framework may help improving the quality of the data/web mining process, enhancing trust and transparency in the outcomes.

This paper is organized as follows. First, the related work is discussed in Section 2. Section 3 introduces basic tools for discrimination analysis and presents the propensity score approach for causal discrimination against individuals, which is extended to causal discrimination against groups in Section 4. Experiments are reported in Section 5. Finally, Section 6 summarizes our contribution and future work.

2 Related Work

Early approaches for discrimination discovery have primarily relied upon simple statistical measures of correlation between the protected group attribute and the decision attribute [38]. Albeit causation has been recognized as significant in the legal circles [16] and in medical research [17] for a while, the perils of correlation analysis have been pointed out in the data mining area only recently. When decisions are made automatically, e.g. by a machine learning model, the problem of preventing discriminatory decisions is known as *fairness*. See [6,32] for surveys of quantitative measures of fairness/discrimination. A recent overview on causal approaches to fairness is [25]. Counterfactual fairness [24], in particular, requires that the prediction of a machine learning model for an individual to be the same as that in the counterfactual world, where the individual had belonged to a different social group. More closely to our approach, propensity score based *stratification* was adopted in [12] to filter out the effect of confounding variables before learning fair linear regression models. In our work, however, we employ propensity score *weighting* and address the problem of discrimination discovery and quantification from data, rather than the fairness problem in the design of machine learning models. Thus, direct comparison can be made with causal approaches for discrimination discovery.

In [9] the causal structure of attributes (a Bayesian Network) is learned from data which is then analyzed to identify patterns of discrimination or favoritism. The construction of the causal network is however impractical for a large number of attribute-value pairs, and it also requires human intervention in providing a temporal order between pairs. [36] extends, as we do, the situation testing approach of [26] to causal reasoning. They adopt a Causal Bayesian Network, which, again, must be specified by a domain expert, to adjust the distance measure between tuples to include only the directly causal attributes. Our approach, instead, relies on the re-weighting of tuples to uniform the distribution in the protected and unprotected groups, without requiring hu-

man intervention. On the other hand, graphical and structural causal models [34,35], are able to distinguish and to explain the effects of direct, indirect, and spurious discrimination. I.e., the causal influence of the protected attribute (direct discrimination), of correlated attributes (indirect discrimination), and of common causes (spurious discrimination) on the decision attribute. In this paper, instead, we deal only with direct discrimination. Finally, some approaches (see e.g., [37]) are extended to tackle training data sanitization. While this is out of the scope of this paper, the method of the original paper [26] can be extended to data sanitization w.r.t. our causal risk difference measure.

In addition to causal graphical models, several other methods are available from the statistical [1] and econometric literature [15] to control for confounders in observational data, including restriction, matching, stratification, decomposition, regression, ANCOVA, etc. Propensity score analysis (PSA) was originally proposed in [30] and it has been widely applied in economics, medicine, and social sciences. See [2,19] for an overview. Uses of propensity scores include: matching homogeneous samples (matches are pairs/sets of individuals from treatment and control groups with similar characteristics), stratification or sub-classification (strata are sets of individuals whose propensity score in a same bin interval), weighting cases to account for the difference in the distribution of the observed covariates between treatment and control groups. Let us relate these notions to our approach. *First*, matching is a form of nearest-neighbor search, but done on the uni-dimensional space of propensity scores, and for tuple(s) of different groups. We also build a sub-set of available data in the neighborhood of a tuple for assigning a discrimination label to the tuple (individual discrimination), according to the legal practice of situation testing [5,26]. Our neighborhood definition (called *kset*), however, works in the multi-dimensional space of tuple attributes and looks for tuples of any group. *Second*, regarding weighting, our key definition of causal risk difference (Definition 2), turns out to be an estimator of the average treatment effect for the treated (ATT) [27]. In our context, the ATT is the expected difference in negative decision (also known as potential outcome) among individuals of the protected group (also known as treatment group). Hence causal risk difference is a counter-factual measure of suffered discrimination by the specific tuple under consideration. In classical propensity score weighting, instead, the ATT is considered with regard to the whole available data. *Third*, standard PSA targets a single contingency table for the whole dataset at hand, so considering discrimination at aggregate, level. Our approach extends standard PSA to extract a description of the groups with the highest discrimination by learning regression trees over the dataset enriched with an attribute measuring causal individual discrimination. Paths in the regression trees unveil specific conditions of (discovered) discrimination.

Finally, we acknowledge that PSA, and a fortiori our approach, suffers from some limitations. *First*, it cannot handle unknown or unmeasured covariates in the data, unless they are correlated with measured ones. Such a limitation, however, is also true for causal approaches [20,33]. *Second*, it works on assumption of “strong ignorability”, namely that protected attribute (treatment) is independent of the decision (potential outcome) conditional on the observed covariates. This is not easy to establish [28, Section 11.3.5], and, if it does not hold, bias reduction is not guaranteed. *Third*, the precise assumption that the protected attribute of race can be considered as “treatment” in the context of counterfactual causal analysis is being debated in the legal literature [22].

group	decision		
	\ominus	\oplus	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad RD = p_1 - p_2$$

Fig. 1. 4-fold, contingency table for $kset_{\mathcal{R}}(\mathbf{r}, k)$.

3 Causal Individual Discrimination

Individual Discrimination. Consider an individual \mathbf{r} from a protected-by-law group (from now on, protected group) suffering from a negative decision, i.e., $dec(\mathbf{r}) = \ominus$, and the central question: was \mathbf{r} discriminated? A *prima-facie* answer is to contrast *ex-post* the decisions made for individuals sharing similar features in a way inspired by the legal approach of *situation testing* – a form of randomized experiments, where individuals with the same characteristics except for the protected ground are sent to apply for a job or for a loan. Consider the contingency table in Figure 1 for the set of k -nearest neighborhood tuples $kset_{\mathcal{R}}(\mathbf{r}, k)$ of \mathbf{r} . Here, individuals are modelled by the vector of their features, and a similarity function between tuples is assumed to be defined. The risk difference RD defined in Figure 1 is a measure of the degree of bias towards negative decisions for members of \mathbf{r} to the protected group. If such a degree is above a threshold $\alpha \geq 0$, i.e., $RD > \alpha$, then the individual \mathbf{r} was discriminated. The threshold α is a parameter of the analysis. A dual statement for an individual \mathbf{r} of the unprotected group with a positive decision, i.e. $dec(\mathbf{r}) = \oplus$, is the question: was \mathbf{r} favored?

While we concentrate on RD , our approach readily applies to other measures defined in terms of contingency tables such as risk ratio, odds ratio, etc., and to their tests of statistical significance (see surveys [29,38]). A major drawback of directly using a discrimination measure over the contingency table as in Figure 1 is that it does not account for differences between the individuals in the neighborhood. Although we consider in the neighborhood tuples that are close to each other, differences in the attribute values of tuples in the protected and those in the unprotected group may still support different decisions for such groups. For example, in an employee salary dataset several factors may contribute to the high/low salary of an employee (education, work hours, age, etc.). While it is desired that males and females with similar credentials should have similar salaries, observed difference in salaries of males and females in the neighborhood of an employee could signify potential discrimination or favoritism. The calculated risk difference of the employee, however, would be biased if it is found, for instance, that females in the dataset work fewer hours than males and employees working for fewer hours obtain lower salaries. Thus, the factor work hours confounds the estimated risk difference. We aim, instead, at estimating the *causal* effect of belonging to the protected group on the outcome by handling the effect of confounding factors.

P propensity Score Analysis. Our problem commonly arises in the statistical analysis of observational data where assignment of individuals to treatment and control groups cannot be assumed to be at random [31]. In our context, we consider the protected

group as the treatment group, and the unprotected group as the control group. A method for causal analysis over observational data is provided by propensity scores. Propensity score analysis is a principled way to handle multiple confounding variables and to ‘filter out’ the explainable effect of these variables.

Definition 1. *Assume that the dataset \mathcal{R} is a sample over a distribution of covariates \mathbf{x} . The propensity score $e(\cdot)$ is the conditional probability of belonging to the protected group given the covariates:*

$$e(\mathbf{x}) = Pr(\text{protected} \mid \mathbf{x})$$

The estimation of propensity scores requires two key decisions: the model or functional form of $e(\cdot)$ and the (confounding or explanatory) variables to include in \mathbf{x} . For binary valued groups, as in our case, the logistic regression model is commonly adopted [13]. The decision regarding which covariates to include in the logistic regression model is less standardized. It is recommended that the selected variables should influence simultaneously the group membership and the decision of individuals. Furthermore, it is advised against selecting too many covariates such that estimates of propensity scores for most individuals becomes exactly one or zero which violates the condition of overlap between protected and unprotected group individuals w.r.t. propensity score necessary for effective propensity score analysis [11]. Following these guidelines, we employ the following estimation procedure:

1) The propensity score is given by a logistic regression model learned over the dataset \mathcal{R} , i.e.,

$$e(\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \mathbf{x})}$$

Here, \mathbf{x} is the vector of selected covariates and β is the corresponding weight vector.

2) The variables that are correlated with both group membership and decision are candidates for inclusion in the propensity score model. The correlation measure and its threshold is decided by a domain expert. For example, variables producing information gain ≥ 0.05 w.r.t. both group and decision variables could be candidates for selection.

3) Remove from the selected candidates, variables that are proxies for the group membership, i.e., variables that are almost duplicates of the group variable, as they can cause the propensity scores to be either zero or one. For example, variables producing an entropy ≥ 0.95 for predicting the group variable are removed.

Propensity scores can answer the following counter-factual question: “what kind of outcomes would we have observed had the decision involving unprotected individuals involved protected ones instead?” Let us introduce now weights $w(\mathbf{x})$ such that the distribution of the protected and unprotected group becomes identical, i.e., such that:

$$Pr(\mathbf{x} \mid \text{protected}) = w(\mathbf{x}) \cdot Pr(\mathbf{x} \mid \text{unprotected})$$

By simple algebra and the Bayesian rule, we obtain:

$$w(\mathbf{x}) = K \cdot \frac{e(\mathbf{x})}{1 - e(\mathbf{x})} \quad (1)$$

where $K = Pr(\text{unprotected})/Pr(\text{protected})$ is a constant term that will not be relevant in subsequent analysis. $w(\mathbf{x})$ is the weight that a tuple \mathbf{x} of the unprotected group should count for in case it would belong to the protected group. This approach is known as *propensity score weighting* [30]. Rosenbaum and Rubin showed that, conditional on the propensity score, all observed covariates are independent of group assignment and, in large samples, they will not confound estimated treatment effects, i.e., in our context the decisions on the protected group can be compared with the decisions on the unprotected group once the latter is re-weighted using propensity scores.

Propensity Score Weighting. Let us now apply propensity score weighting in the context of discrimination measures over a contingency table for an individual \mathbf{r} of the protected group. See Figure 1. Instead of comparing the average negative decision p_1 of the protected group with the average negative decision p_2 of the unprotected group, we compare p_1 with the average negative decision p_2^c that would be obtained for individuals of the unprotected group if they were instead in the protected group. We define the weighted average negative decision of the unprotected group p_2^c as:

$$p_2^c = \frac{\sum_{\mathbf{s} \in \mathcal{S}, \text{dec}(\mathbf{s}) = \ominus} w(\mathbf{s})}{\sum_{\mathbf{s} \in \mathcal{S}} w(\mathbf{s})}$$

where $\mathcal{S} = \text{kset}_{\mathcal{R}}(\mathbf{r}, k, m) \cap \mathcal{R}(\text{unprotected})$ is the set of tuples in the neighborhood that belong to the unprotected group. Notice that the constant term K in (1) is ruled out in the calculation of the ratio p_2^c .

Definition 2. *The causal risk difference is $RD^c = p_1 - p_2^c$.*

The causal risk difference compares the proportion of negative decisions for the neighborhood tuples in protected group with the proportion of negative decisions for the neighborhood tuples of the unprotected group re-weighted to the distribution of the protected group. RD^c close to zero means that differences in the outcomes of the decision between the protected and unprotected group are motivated by confounding factors in the selection of the two groups which are covered by the covariates of the propensity scores. If RD^c is not close to zero, then there is a (causal) bias in decision value due to group membership, or there are covariates that have been not accounted for in the analysis (*omitted variables*). In the former case, $RD^c > 0$ denotes causal discrimination, and $RD^c < 0$ denotes causal favoritism (e.g., due to positive/affirmative actions).

Example 1. Let us contrast on a fictitious example risk difference and causal risk difference with the variation of propensity score. Fix a tuple of the protected group *with a negative decision*, and fix $k = 15$. Suppose the set of the k nearest neighbors of this individual contain 7 females, with 4 having a negative decision, and 8 males, with 3 having a negative decision. The (non-causal) risk difference of this individual is $RD = p_1 - p_2 = 4/7 - 3/8 = 0.196$. Suppose that males have a propensity score of 0.5. Their weights will be $0.5/(1 - 0.5) = 1$ and then $p_2^c = p_2$, hence $RD^c = RD$. Intuitively, if there is no difference in the distribution of males and females, causal risk difference boils down to risk difference. Assume now that the 5 males with positive decision have a greater propensity score, i.e., their characteristics are closer to the female

group distribution than to the one of the male group. Then their weights will become greater than 1 and $p_2^c < p_2$, hence $RD^c > RD$. For instance, when the propensity score of the positive males is 0.8, the weight for the positive males is equal to 4, and the causal risk difference of the individual becomes $RD^c = 4/7 - 3/23 = 0.441 > RD$. That is, the individual faces higher discrimination than the one quantified by risk difference, because similar males are given positive decisions even though they have ‘femalish’ characteristics. The opposite case is also possible, namely $RD^c < RD$, i.e., non-causal risk difference overestimates discrimination. For example, if the 5 males with positive decision have propensity score of 0, i.e., they are definitely not following the distribution of females, their weight is also 0, and then $RD^c = 4/7 - 3/3 = -0.429 < RD$.

4 Causal Discrimination Discovery

In the previous section, we assigned to every tuple \mathbf{r} of the protected group with negative decision a causal measure RD^c of discrimination (for $RD^c > 0$) observable when comparing its decision to that of its neighbors. For tuples of the unprotected group, a similar measure can be defined as $RD^c = p_2^c - p_1$, so that again $RD^c > 0$ denotes discrimination. Dually, $RD^c < 0$ denotes favoritism for individuals of the protected and unprotected groups experiencing a positive decision. This approach is useful for quantifying bias in discriminating/favoring an individual (*individual discrimination*). However, in many cases, a global description of who was discriminated/favored is also required. In order to infer such descriptions, we proceed by extracting a regression model from a modified version of the dataset \mathcal{R} at hand. Consider the problem of characterizing discrimination in the protected group. We restrict to tuples of the protected group with negative decision. We augment the dataset with a new attribute, the class attribute, setting its value to the RD^c for each tuple in it. Over such a labeled dataset, a regression tree [10] is extracted, which can be used for descriptive purposes. A path in the regression tree ending in a prediction of (large) positive values describes a context of discrimination for the protected group. Similarly, patterns of favoritism for the unprotected group with positive decision can be studied by extracting regression trees and looking at paths predicting large negative values of the class. In summary, we adopt regression to lift from causal individual discrimination analysis to causal global discrimination analysis.

5 Experiments

We perform experiments⁶ on two commonly-used datasets in discrimination analysis research: Adult, and Census Income. Both datasets are publicly available in the UCI Machine Learning Repository⁷. In both datasets, we use income as the decision attribute with income $> 50K$ as the positive decision and sex as the sensitive attribute considering females as the protected group.

⁶ Code at: <https://github.com/anam-zahid/Causal-inference-for-social-discrimination-reasoning>

⁷ <http://archive.ics.uci.edu/ml>

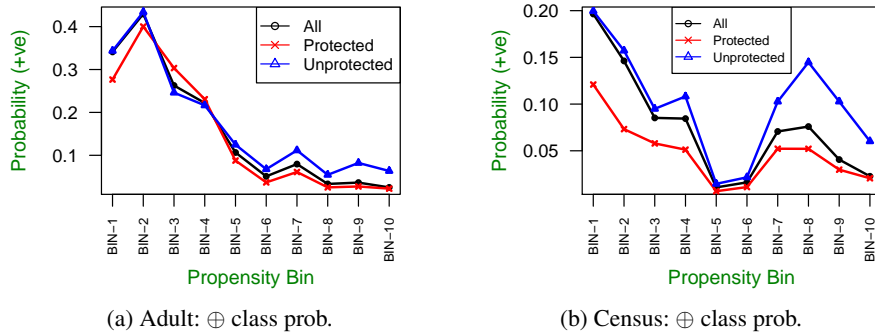


Fig. 2. Variation of positive decision probability with propensity score.

Propensity scores are estimated by a logistic regression model learned over the dataset for predicting membership to the protected group. The attributes used for propensity score estimation are those correlated with both decision and group attributes. Information gain is used as the correlation measure and the selection threshold is set to 0.05.

The neighborhood of a tuple (namely, $kset_{\mathcal{R}}(\mathbf{r}, k)$) is defined by setting $k = 15$. The Euclidean distance with attributes normalized (min-max normalization) in the interval $[0, 1]$ is used as the distance function.

Variation of Positive Decisions with Propensity Scores. To study the variation of positive decisions with propensity scores, we discretized the datasets into 10 equal interval bins on the basis of propensity scores, e.g., *bin-1* instances have propensity (of protected group membership) less than 10%, *bin-2* with 10-20% and it ends with *bin-10* having propensity score between 90-100%. Individuals of bins 1-5 have characteristics that are more similar to those in the unprotected group (males) due to their low propensity of belonging to the protected group (females).

Figure 2 shows the difference in positive class probabilities of tuples based on their propensity score bins and group membership.

It turns out that: instances with male-like characteristics, i.e., bins 1-5, have considerably higher \oplus decision probabilities than female-like bins; and, that \oplus decision probability difference between protected and unprotected groups is also in favor of unprotected group instances in most of the bins. Differences are more marked for the Census dataset compared to the Adult dataset.

Group Wise Discrimination/Favoritism Trends. Plots in Figure 3 show group wise discrimination/favoritism trends measured through causal risk difference at the variation of propensity score. Propensity bins are on the x-axis while average RD^c 's are on the y-axis. Top row plots contain instances with positive RD^c (discriminated) while bottom row plots contain instances with negative RD^c (favored).

General assumption is that protected group instances are more discriminated than unprotected group instances. Figure 3 (a) support this assumption by showing that protected group instances face more discrimination in most of the propensity bins except with a few exceptions in male-like bins 2-5. Such a group-grounded bias becomes more

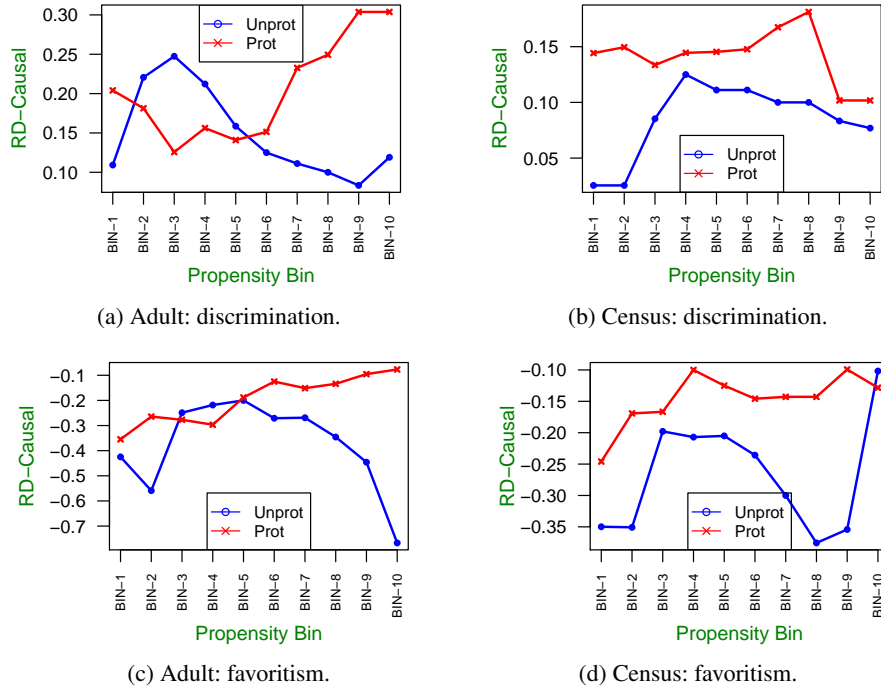


Fig. 3. Group wise discrimination/favoritism trends.

evident in the Census Income dataset, as shown in Figure 3 (b). Fig 3 (c & d) show a significant difference in the favoritism scores between both groups for both datasets. Unprotected group individuals (blue line) are highly favored in most of the propensity bins despite of having characteristics quite similar to protected instances. Fig 3 carries a lot of importance and shows very obvious discrimination/favoritism trends between both sensitive groups despite of similar propensity scores and characteristics.

RD and RD^c Comparison. Figure 4 shows a comparison between RD^c and its non causal variant RD . According to our discussion earlier in Section 3, RD overestimates both discrimination and favoritism scores (recall that higher absolute values denote higher discrimination/favoritism). Margin between RD and RD^c lines show causation magnitude. Top row shows discriminated instances with positive RD and RD^c and bottom row represents favored instances with negative scores. Figure 4 (a) shows higher discrimination scores by RD . Similar trends are seen in Figure 4 (b) and (d) while RD and RD^c are relatively closer in Figure 4 (c). We can conclude that RD is overestimating discrimination/favoritism scores in almost all the bins of both datasets but the margin is higher for the Census Income dataset.

RD^c and Rule Generation. We use regression rules to explore interesting patterns in the datasets that lead to high values of discrimination/favoritism scores. Rules are created from protected and unprotected groups separately and only include instances with non-zero RD^c . In our analysis, every node (whether leaf or not) acts as a rule and its

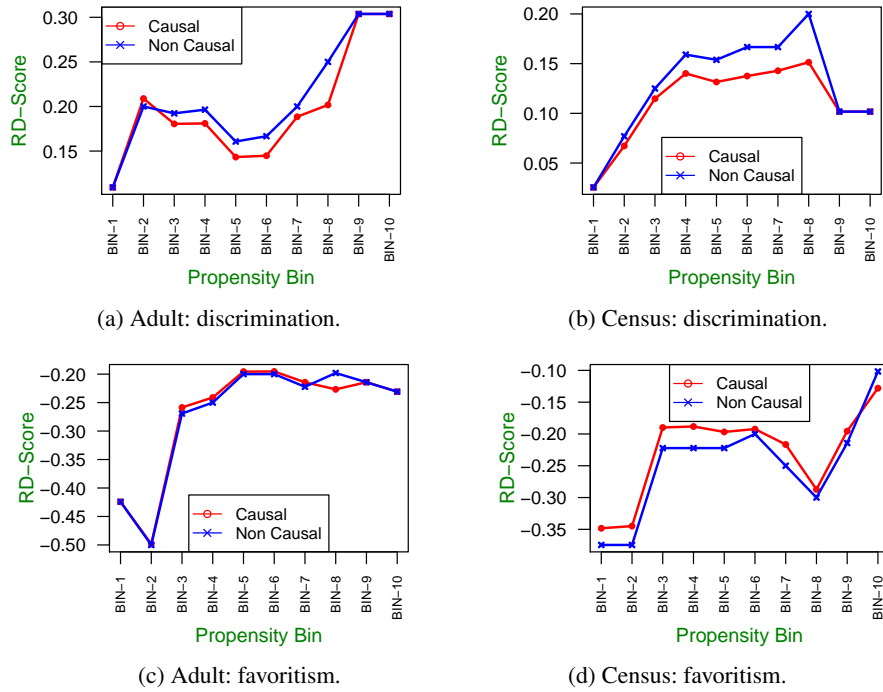


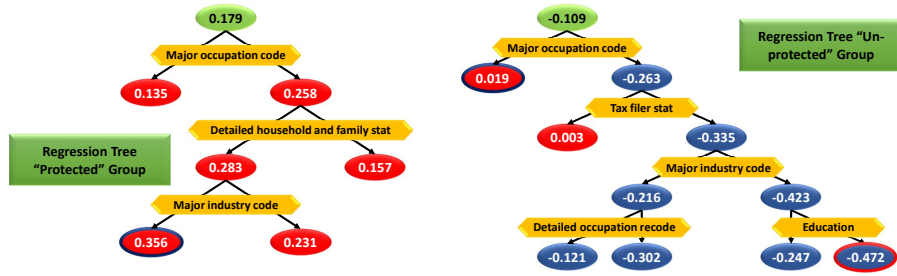
Fig. 4. RD and RD^c comparison.

label (the average risk difference of tuples reaching the node) acts as bias magnitude. Figure 5 (a & b) belong to protected and unprotected groups of the Census Income dataset. The green node is the root, red nodes show discriminatory rules while blue nodes show favoritism rules. Nodes with maximum discrimination/favoritism magnitude are highlighted with contrasting borders.

In Figure 5 (a), after the split on major occupation code, we get two subsets with 13.5% and 25.8% average causal risk difference respectively. Similarly we discover multiple other subsets that have different magnitudes of discrimination/favoritism. The most discriminated group suffers from a 35.6% causal risk difference. Similar trends are seen in Figure 5 (b) where most of nodes regard favoring sub-groups of unprotected individuals. E.g, in the rightmost node, males with comparable characteristics as females benefit from high income with a causal risk difference of -47.2%. This is a key observation of our causal discrimination discovery technique that highlights under which conditions (the path from root to the node) discrimination and favoritism do arise.

6 Conclusions

Based on propensity score analysis, our approach for discrimination discovery and understanding constructs a modified distribution of observed covariates that eliminates the effect of observed confounding factors. The degree of discrimination or favoritism



(a) Discrimination against the protected group. (b) Favoritism for the unprotected group.

Fig. 5. Census Income: regression trees.

is quantified by the risk difference computed over the modified distribution. This is the starting point for measuring and understanding discrimination bias in training data (preventing discrimination learning) or in test data (assessing discrimination in predictive models). While graphical/structural causation models are able to model finer-grained discrimination analyses, our approach does not require human intervention, and scales linearly with the number of features. This makes it feasible to apply the approach in practice. Moreover, our approach can be easily implemented on top of basic data analysis tools (logistic regression, nearest-neighbor set, regression trees). Future work is open towards several directions. A first one is to adapt training data sanitization [26] and fair machine learning models [6] to account for causal discrimination measures. A second possible direction is to extend our approach and experimentation to non-structured and high-dimensional data types which may also hide human biases, such as text [8], which originate in the web [3] and from social media [23].

References

1. Agresti, A.: *Categorical Data Analysis*. Wiley Series in Probability and Statistics, Wiley-Interscience, 2 edn. (2002)
2. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46**(3), 399–424 (2011)
3. Baeza-Yates, R.A.: Bias on the web. *Commun. ACM* **61**(6), 54–61 (2018)
4. Barocas, S., Selbst, A.D.: Big data’s disparate impact. *California Law Review* **104** (2016)
5. Bendic, M.: Situation testing for employment discrimination in the United States of America. *Horizons Stratégiques* **3**(5), 17–39 (2007)
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., , Roth, A.: Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018)
7. Bickel, P.J., Hammel, E.A., O’Connell, J.W.: Sex bias in graduate admissions: Data from Berkeley. *Science* **187**(4175), 398–404 (1975)
8. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *NIPS*. pp. 4349–4357 (2016)

9. Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. *I. J. Data Science and Analytics* **3**(1), 1–21 (2017)
10. Breiman, L., Friedman, J., Olshen, R., Stone, C.: *Classification and Regression Trees*. Wadsworth Publishing Company (1984)
11. Bryson, A., Dorsett, R., Purdon, S.: *The use of propensity score matching in the evaluation of active labour market policies*. Crown (2002)
12. Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: *ICDM*. pp. 71–80. IEEE (2013)
13. Caliendo, M., Kopeinig, S.: Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* **22**(1), 31–72 (2008)
14. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. *Science Advances* **4**(1) (2018)
15. Fortin, N., Lemieux, T., Firpo, S.: Decomposition methods in economics. In: *Handbook of Labor Economics*, vol. 4, pp. 1 – 102. Elsevier (2011)
16. Foster, S.R.: Causation in antidiscrimination law: Beyond intent versus impact. *Houston Law Review* **41**(5), 1469–1548 (2004)
17. Grimes, D.A., Schulz, K.F.: Bias and causal associations in observational research. *Lancet* **359**, 248–252 (2002)
18. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5), 93:1–93:42 (2019)
19. Guo, X.S., Fraser, M.W.: *Propensity score analysis: Statistical methods and applications*. Sage Publications, Inc., 2nd edn. (2015)
20. Kilbertus, N., Ball, P.J., Kusner, M.J., Weller, A., Silva, R.: The sensitivity of counterfactual fairness to unmeasured confounding. In: *UAI*. p. 213. AUAI Press (2019)
21. Kohavi, R., Longbotham, R.: Online controlled experiments and A/B testing. In: *Encyclopedia of Machine Learning and Data Mining*, pp. 922–929. Springer (2017)
22. Kohler-Hausmann, I.: Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern Univ. Law Rev.* **113**, 1163–1227 (2019)
23. Kulshrestha, J., Eslami, M., Messias, J., Zafar, M.B., Ghosh, S., Gummadi, K.P., Karahalios, K.: Search bias quantification: investigating political bias in social media and web search. *Inf. Retr. Journal* **22**(1-2), 188–227 (2019)
24. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: *NIPS*. pp. 4069–4079 (2017)
25. Loftus, J.R., Russell, C., Kusner, M.J., Silva, R.: Causal reasoning for algorithmic fairness. *CoRR abs/1805.05859* (2018)
26. Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: *KDD*. pp. 502–510. ACM (2011)
27. Morgan, S.L., Todd, J.L.: A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* **38**(1), 231–281 (2008)
28. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2 edn. (2009)
29. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* **29**(5), 582–638 (2014)
30. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983)
31. Shadish, W.R., Cook, T.D., Campbell, D.T.: *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin (2002)
32. Verma, S., Rubin, J.: Fairness definitions explained. In: *FairWare@ICSE*. pp. 1–7. ACM (2018)
33. Wu, Y., Zhang, L., Wu, X.: Counterfactual fairness: Unidentification, bound and algorithm. In: *IJCAI*. pp. 1438–1444. ijcai.org (2019)

34. Zhang, J., Bareinboim, E.: Fairness in decision-making - the causal explanation formula. In: AAAI. AAAI Press (2018)
35. Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. *I. J. Data Science and Analytics* **4**(1), 1–16 (2017)
36. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: A causal inference approach. In: *IJCAI*. pp. 2718–2724 (2016)
37. Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: *KDD*. pp. 1335–1344. ACM (2017)
38. Zliobaite, I.: Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* **31**(4), 1060–1089 (2017)