Does Data-Driven Recruitment Lead to Less Discrimination? – A Technical Perspective

Kian Schmalenbach[1], Sven Laumer[1]

[1] Friedrich-Alexander-University, Schöller Endowed Chair for Information Systems,
Nuremberg, Germany
{kian.schmalenbach,sven.laumer}@fau.de

Abstract. Due to its large cost-saving potential, data-driven recruitment is becoming increasingly popular across various industries. However, several cases were reported where the use of corresponding technologies had caused systematic discrimination against certain candidate groups. While existing approaches to discover and prevent discrimination in data classification mostly perform well within a specific context, it remains unclear to what extent data-driven recruitment can be conducted discrimination-free in real-world business applications, where the respective context-specific assumptions do not necessarily hold. Hence, we first define a generic discrimination model that allows for arbitrary descriptions of job candidate characteristics, before applying two sophisticated discrimination-prevention algorithms on a sample data set generated from our model to evaluate the algorithms' performance. Our analysis shows that the amount of removed discrimination highly depends on the application context and its underlying definitions and assumptions, making it hard to provide a holistic answer to our research question.

Keywords: data-driven recruitment, discrimination-aware data mining, discrimination discovery, discrimination prevention

1 Introduction

Within the hiring process, decision makers in companies have already been using digital technologies to support their work for many years (e.g. [5], [10], [14]). However, the advancement of artificial intelligence and data mining offers the perspective to gradually automate the hiring process in its whole through the application of predictive classifiers, referred to as *data-driven recruitment* [2]. Since this technology allows for cost reduction through automation of large parts of previously manually conducted work of decision makers typically working in the HR department of a company, it is considered highly attractive by many large employers in different industry sectors, such as *Amazon*, *Ikea*, *Hilton*, or *Cisco* [2].

However, a recent case concerning the data-driven recruitment tool developed and applied by *Amazon* revealed that the company's decision-making is biased towards male applicants [4]. Apparently, the classifier trained to rank job candidates in descending order of suitability for specific vacant positions internalized a preference

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_q2-schmalenbach

of the responsible decision makers in favor of male job candidates. While ethically questionable, facing decisions depending on sensitive characteristics of candidates such as gender, race, or age is also considered illegal in many countries, including the United States [16]. Hence, *Amazon* finally stopped the development of its automated recruitment tool by 2017 [4].

The research area addressing the type of problem *Amazon* encountered is called *discrimination-aware data mining* and was first introduced by the authors of [11]. Since then, many authors have published methods and algorithms to discover discriminatory decision patterns in classifiers and to abolish and prevent further discrimination. However, these approaches often consider only one sensitive characteristic of applicants such as gender (e.g. [3], [7], [8]) or make restricting assumptions on the architecture of the considered classifier (e.g. [3]). Hence, the question whether it is technically possible to prevent discrimination in data-driven recruitment independent of the specific application context has not been clearly answered yet.

Within this paper, we will investigate to what extent existing approaches are able to perform discrimination-free data-driven recruitment in arbitrary scenarios, i.e. on input data that does not necessarily adhere to the assumptions formulated by the respective authors. For this purpose, we first propose a generic discrimination model, allowing for arbitrary descriptions of candidate characteristics. We then apply two of the most sophisticated discrimination prevention algorithms on a generic sample input data set generated from our model and evaluate their performance using the (adapted) metrics designed for either approach on both algorithms.

The following part of this paper is therefore structured as follows: Section 2 motivates our research question through a brief discussion of relevant literature, section 3 introduces our generic discrimination model and the algorithms considered in our analysis, section 4 describes the results of our analysis, and section 5 discusses our results in a broader research context.

2 Background and Literature Review

This section provides a brief overview about the background of discrimination-aware data mining as well as outcomes and limitations of previous work on this topic.

2.1 Discrimination and Data-Driven Recruitment

Data-driven recruitment generally refers to the process of making hiring decisions with the help of automated candidate data analysis conducted by some *classification* algorithm. To enable this classification process, the classifier first needs to learn characteristics of suitable candidates, which often happens by training the classifier on historical data sets containing information about previous hiring decisions faced by humans. Once completed, the classifier can then search a set of new job candidates and recommend (not) hiring each of them based on their characteristics. [2]

https://doi.org/10.30844/wi_2020_q2-schmalenbach

However, the human decisions comprised in the training data set may contain several forms of *bias*, which can lead to discrimination when internalized by an algorithmic classifier. *Direct discrimination* occurs when the hiring decision is influenced by *sensitive* candidate characteristics that must not be considered due to ethical or legal reasons, such as *gender*, *race*, or *age*. Facing a decision by legal attributes, which are correlated with some sensitive characteristics, is referred to as *indirect discrimination*. For instance, people of a certain race might commonly live in a certain neighborhood, identified by a specific ZIP code [11]. Even after removing the sensitive attribute *race* from the training data, discrimination is still possible by avoiding candidates with an address containing the specific ZIP code, who mostly belong to the deprived race. Moreover, the authors of [9] define *explainable discrimination* as a form of indirect discrimination, which is justifiable and hence legal, since the involved attribute contains objective information about the candidate's aptitude. They then define *illegal discrimination* as the difference between observed and explainable discrimination.

2.2 Discrimination-Aware Data Mining

As explained above, human-faced recruitment decisions may include several forms of discrimination, which a classifier might adopt during training [9]. Since the classifier will face all its decisions according to the same procedure it has learnt during the training period, the previously unsystematic bias of certain decision makers might turn into systematic discrimination by the classifier. Therefore, the authors of [15] describe the aim of *discrimination-aware data mining* as to discover discriminatory decision patterns of classifiers (*discrimination discovery*) and to prevent the application of such patterns within the classification process (*discrimination prevention*).

Discrimination Discovery. *Discrimination discovery* is the process of detecting discriminatory decision patterns in classifiers [11]. This process requires knowledge about the sensitive attributes, which should not be considered by the classification process due to legal restrictions, and an analysis if the classification decisions significantly depend on the value of some of these attributes [16].

Most existing approaches for discrimination discovery first require a domain expert to define the sensitive attributes in a data set (e.g. *gender*) and the accordingly deprived groups (e.g. *females*), and then propose procedures to discover discriminatory classifier decisions. The authors of [11] propose to check whether the confidence of a frequent classification rule with sensitive attributes (or attributes highly correlated with sensitive ones) exceeds a certain threshold α. The authors of [13] suggest clustering the candidate item sets according to the non-sensitive attributes and then comparing the probability of being hired for different values of the sensitive attributes within each cluster.

https://doi.org/10.30844/wi_2020_q2-schmalenbach

Discrimination Prevention. Whereas discrimination discovery proves useful to uncover discriminatory recruitment decisions, *discrimination prevention* is necessary to actually reduce or avoid discrimination in data-driven recruitment. According to [15], it can be realized by removing the discriminatory information from the training data set before learning a classifier (*data preprocessing*), by adapting the learning procedure (*model regularization*), or by modifying a trained classifier (*model post-processing*). Since data preprocessing works independent of the type of classifier to be trained, most previous work (e.g. [6], [9], [13], [17]) focuses on this approach.

A naive approach of performing data preprocessing could be to simply remove all sensitive attributes from the training data set. However, this is not sufficient to abolish discrimination, since legal attributes are often highly correlated with sensitive ones, an effect that is known as *redlining* [12]. In fact, a recent study [16] shows that it might actually be necessary to explicitly include the sensitive attributes in the training data set to be able to avoid redlining in the learning procedure.

To avoid both direct discrimination and redlining, the authors of [17] propose a local massaging technique, which ranks the candidates according to their probability of being hired and then modifies the labels of those close to the decision border. The authors of [9] extend this approach by duplicating or deleting data items close to the decision border (*preferential sampling*). The authors of [6] implement various algorithms that either change the value of sensitive attributes or the label, depending on the type of discrimination and confidence and support of the corresponding decision rules. A previous paper [3] presents three approaches to remove discrimination from trained naive Bayesian classifiers.

Discrimination Measurement. To evaluate the approaches mentioned above, computing the *accuracy* of the resulting classifiers is not sufficient, since it does not consider the amount of discrimination reduced by the corresponding procedure. Instead, each paper proposes different evaluation metrics, which are summarized and compared in [15]. Moreover, the authors of [7] prove that there is a linear trade-off between accuracy loss and discrimination reduction, independent of the specific discrimination prevention procedure. Hence, the general aim of discrimination prevention is to remove as much discrimination from the classifier as necessary, while keeping as much accuracy as possible [15].

2.3 Limitations of Existing Work

While the approaches presented above mostly perform well within their respective application context, they prove to have certain limitations, which are discussed below.

Assumptions on Sensitive Attributes. Many of the approaches describe above (e.g. [3], [7], [8]) are restricted to one binary sensitive attribute (e.g. *gender* = {*male, female*}). This assumption can be considered unrealistic since many anti-discrimination laws require more than one attribute to be excluded from the decision-making process [7]. In addition, there might be more than two possible values for the

https://doi.org/10.30844/wi_2020_q2-schmalenbach

sensitive attribute (e.g. *race* = {*white, black, hispanic*}). We aim at addressing this issue by defining a generic discrimination model, allowing for an arbitrary number of both sensitive and legal attributes as well as a probabilistic description of their interdependencies.

Explainable and Illegal Discrimination. As explained in section 2.1, legal attributes are often correlated with a set of sensitive attributes. For instance, male persons might tend to have a higher skills level or work experience for a certain position than females. If this would be the case, hiring equally many male and female candidates would lead to *reverse discrimination* [17], since it would mean that females with lower experience are hired just because of their gender. While the authors of [9] propose a model to explicitly calculate the amount of this so-called *explainable* and the truly *illegal* discrimination, they leave the task of combining their model with the algorithms proposed by different authors (e.g. [6], [13]) open for further research. Our paper addresses this topic by extending the algorithm proposed in [9] to be able to handle multiple sensitive attributes assuming that they are (pairwise) independent.

Lack of a Unified Discrimination Measure. A third problem throughout the literature on discrimination-aware data mining is the lack of a uniform way of measuring discrimination in a data set or within a classifier. The authors of [15] summarize and evaluate many common evaluation metrics for discrimination discovery and prevention, but also point out that the usability of each metric is dependent on the application context, making it difficult to create context-independent decision-support algorithms. Therefore, we evaluate both algorithms discussed within our analysis using a suitable extended version of the metrics proposed by each author, allowing for a more holistic discussion of the algorithms' performance.

3 Research Model and Method

In this section, we define a generic discrimination model, and we present two discrimination prevention methods originally presented by [6] and [9], which we will then apply to a sample data set generated from our model.

3.1 Generic Discrimination Model

The discrimination model presented in this section is based on the model proposed by [9], but extended to allow for an arbitrary number of sensitive and legal attributes.

Basic Assumptions and Definitions. Let $\mathcal{A}$ be a set of sensitive attributes and let $\mathcal{B}$ be a set of legal attributes according to some legal definition provided by an external domain expert. Let $|\mathcal{A}| =: k$ and $|\mathcal{B}| =: l$. An item set $I = (A_1, \dots, A_k, B_1, \dots, B_l)$ consists of k sensitive and l legal attributes and their respective values. Let $\mathcal{DB}$ be a database containing n item sets, $\mathcal{DB} = \{I_1, \dots, I_n\}$.

The function $\mathcal{L} : I \mapsto C$ maps each item set $I \in \mathcal{DB}$ to a binary label $C \in \{\oplus,\ominus\}$. We write $\mathcal{L}(I) = \oplus$ to indicate that the job candidate represented by item set I is hired and $\mathcal{L}(I) = \ominus$ to indicate that the candidate is rejected. We assume that each candidate has a preference to be hired, i.e. $\mathcal{L}(I) = \oplus$ is a more preferential outcome for a candidate than $\mathcal{L}(I) = \ominus$.

For the sake of simplicity, we assume that, for every attribute $X \in \mathcal{A} \cup \mathcal{B}$, the decision maker favors a subset of the set of its possible values, which we denote with a lowercase letter. For instance, the decision maker favors $X = x$ over $X = \neg\, x$. We define $P(x) = n^{-1} \cdot |\{I \in \mathcal{DB} \mid X = x\}|$ as the share of item sets in $\mathcal{DB}$ where $X = x$, and we define $\neg x$ such that $P(\neg x) = 1 - P(x)$.

For all item sets $I \in \mathcal{DB}$ with $I = (A_1, \dots, A_k, B_1, \dots, B_l)$, we define a function $f(I): I \mapsto\ [0,1]$ as

$$f(I) = w_0 \cdot r + \sum_{i=1}^{k}\big(w_i \cdot \delta(a_i)\big) + \sum_{i=1}^{l}(w_{k+i} \cdot \delta(b_i)) \tag{1}$$

where $\delta(.)$ is a function that evaluates to 1 if its argument is true and to 0 otherwise, $w_i \in [0,1]$ are fixed weights with $\sum_{i=0}^{k+l} w_i = 1$, and $r \in [0,1]$ is a uniformly distributed random value.

Given a threshold $\theta \in [0,1]$, we then define:

$$\mathcal{L}(I) = \begin{cases} \oplus, & f(I) \geq \theta \\ \ominus, & otherwise \end{cases} \tag{2}$$

Direct Discrimination. Recall that all attributes $A \in \mathcal{A}$ are sensitive attributes, which may not be considered in the decision-making process due to legal restrictions. Hence, setting the values of the weights $w_i > 0$ for $i \in \{1, \dots, k\}$ in equation (1) is equivalent to conducting direct discrimination, because candidates with the preferred value of a sensitive attribute (e.g. *gender* = *male*) are favored over those with the respective complementary value (e.g. *gender* = *female*).

From a probabilistic perspective, this is equivalent to a positive correlation between the corresponding attribute $A_i \in \mathcal{A}$ and the class label $C =\oplus$ and can be expressed as

$$P(\oplus|\ a_i\,) > P(\oplus) > P(\oplus|\ \neg a_i\,) \tag{3}$$

where $P(\oplus|\ a_i)$ is the probability that a candidate represented by an item set $I \in \mathcal{DB}$ with $A_i = a_i \in I$ is hired, and $P(\oplus)$ is the probability that any candidate in $\mathcal{DB}$ is hired. Intuitively speaking, if A_i represents *gender* and a_i signifies *male*, equation (3) holds if and only if a male candidate has a higher chance of being hired than a female one. Note that this type of discrimination can be avoided by setting all weights $w_i = 0$ where $I \in \{1, \dots, k\}$, which is equivalent to removing all sensitive attributes $A \in \mathcal{A}$ from the input data set $\mathcal{DB}$.

Indirect Discrimination. Indirect discrimination or redlining occurs when the hiring decision is based on the values of a legal attribute $B_j \in \mathcal{B}$ that is correlated with some sensitive attribute $A_i \in \mathcal{A}$. For instance, male candidates may preferably graduate

https://doi.org/10.30844/wi_2020_q2-schmalenbach

from a certain university, causing a positive correlation between the sensitive attribute A_i (e.g. *gender*) and the legal attribute B_j (e.g. *graduation*).

From a probabilistic perspective, this can be expressed as follows:

$$P(\oplus \mid b_j) > P(\oplus) > P(\oplus \mid \neg b_j) \quad \text{and} \quad P(b_j \mid a_i) > P(b_j) > P(b_j \mid \neg a_i) \tag{4}$$

Due to the correlation between B_j and A_i, it is not sufficient to remove A_i from the input data set to avoid this type of discrimination. Instead, more advanced techniques are required to remove the correlation between legal and sensitive attributes with minimal accuracy loss (e.g. [7], [11], [16]).

Explainable Discrimination. In some cases, a legal attribute $B_j \in \mathcal{B}$, which is positively correlated with some sensitive attribute $A_i \in \mathcal{A}$, also contains objectively valid information about the aptitude of a candidate for a job position. For instance, graduates from a specific university (i.e. $B_j = b_j$) might have enjoyed an education of higher quality and therefore be more suitable for the job than others. If $P(b_j \mid a_i) > P(b_j \mid \neg a_i)$, removing all indirect discrimination according to equation (4) will hence lead to *reverse discrimination* in favor of candidates with $A_i = \neg a_i$ and $B_j = \neg b_j$.

According to [9], the total amount of discrimination, D_{all}, can therefore be split up in an explainable part, D_{expl}, and an illegal one, D_{ill}, such that $D_{all} = D_{expl} + D_{ill}$. Consequently, a classifier is considered discrimination-free if $D_{ill} = 0$ or $D_{all} = D_{expl}$.

3.2 Training Data Generation

Having introduced our discrimination model, we now describe a procedure to obtain a data set including all types of discrimination introduced in section 3.1 based on a simple Bayesian network, which we will use to evaluate the two discrimination prevention methods proposed in [6] and [9] discussed within this paper.

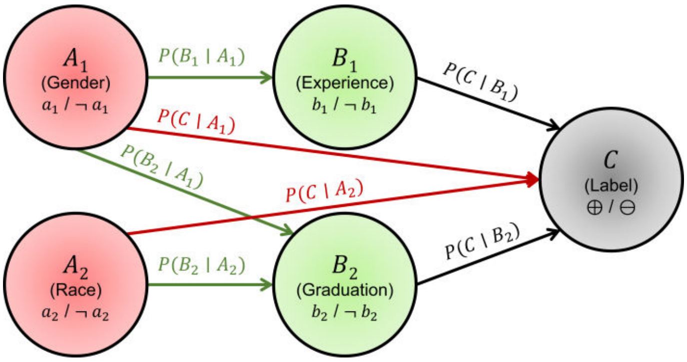

Figure 1. Bayesian network structure described in section 3.2. Sensitive attributes (*illegal discrimination*) are marked red, legal attributes (*explainable discrimination*) are marked green.

Network Structure. For the sake of simplicity, we assume our input data set to carry two independent sensitive attributes A_1 (representing *gender*) and A_2 (representing *race*) as well as two legal attributes B_1 (representing *experience*) and B_2 (representing *graduation*). We assume *experience* to only be correlated with *gender*, and *graduation* to be correlated with both *gender* and *race*. The corresponding Bayesian network structure is given in Figure 1. Note that this network just serves as an example instance of our model from section 3.1 and could be changed arbitrarily to match case-specific knowledge provided by an external domain expert.

Illegal and Explainable Discrimination. A data set sampled from our network will contain illegal discrimination if the weights w_1 and w_2 as defined in equation (1) are set to positive values, as this causes $P(\oplus | \, a_i) > P(\oplus | \, \neg a_i)$ for $i \in \{1,2\}$. In this scenario, the share $(w_1 + w_2)$ of the hiring decision modeled by equation (1) is based on the sensitive attributes *gender* and *race*, which is illegal according to our assumption. Furthermore, the data set will contain explainable discrimination if $P(B_i \mid a_j) > P(B_i \mid \neg a_j)$ for $(i,j) \in \{(1,1),(2,1),(2,2)\}$ and $w_3, w_4 > 0$, which means that the share $(w_3 + w_4)$ of the modeled hiring decision is based on the candidates' experience and graduation, which are here assumed to be better for male candidates. Under these assumptions, the authors of [9] argue that the alleged discrimination resulting from favoring males or persons of white race is explainable, since they were selected because of their higher experience and better graduation rather than their gender or race. Finally, setting $w_0 > 0$ will add some random noise to the data set, which intuitively means that the share w_0 of the hiring decision is based on some attributes or circumstances not covered by our model.

3.3 Discrimination Prevention Methods

In section 3.2, we described a procedure to obtain a data set of hiring decisions that contain direct and indirect discrimination. In a real-world scenario, this type of data is collected empirically, for instance by using records of previous hiring decisions, making it unclear to which extent it contains discrimination. Hence, a *discrimination prevention* algorithm needs to determine if a decision depicted in the data is discriminatory, and if so, remove it from the data set without sacrificing too much accuracy, i.e. with as little changes to the underlying decision criteria as possible.

In literature, two general approaches of removing discrimination can be found. The first one attempts to identify potentially discriminatory classification rules within the data set and modifies them by changing either the premise or the implication of the rules (e.g. [6], [11], [13]). The second one uses probabilistic methods to identify potentially discriminatory correlations between certain attributes and the hiring decision and tries to reduce or eliminate discrimination by changing the underlying probabilities (e.g. [3], [7], [9]). In this section, we present one advanced algorithm of each group, originally proposed by [6] and [9], whose performance we will then evaluate using our generic discrimination model from section 3.1.

Discrimination Prevention through Modification of Classification Rules. One possible approach to remove discrimination from our training data set was proposed in [6]. The algorithm belongs to the group of procedures that identify and eventually modify potentially discriminatory rules in the data set, here by using a parameter α that quantifies how much difference in the treatment of the favored and deprived groups is tolerated before a hiring decision is considered discriminatory. A selection of important measures used within the algorithm and their corresponding probabilistic notation adhering to our model from section 3.1 is listed in Table 1.

In short, the procedure works as follows:

1. Find frequent classification rules $\mathcal{FR}$ in $\mathcal{DB}$ and classify these into directly discriminatory rules $\mathcal{MR} \subseteq \mathcal{FR}$ and redlining rules $\mathcal{RR} \subseteq \mathcal{FR}$. The set of redlining rules also includes the correlated directly discriminatory rules.
2. For each redlining rule $r \in \mathcal{RR}$, find related directly discriminatory rules $r' \in \mathcal{RR}$.
   1. For each $r' \in \mathcal{MR}$, modify the class label of specific item sets $I \in \mathcal{DB}$ from $\neg C$ to C until $\text{conf}(B \rightarrow C) > \max(t_1, t_2)$ for thresholds t_1, t_2.
   2. For each $r' \notin \mathcal{MR}$, modify the class label of specific item sets $I \in \mathcal{DB}$ from $\neg C$ to C until $\text{conf}(B \rightarrow C) > t_1$ for threshold t_1.
3. For each directly discriminatory rule $r' \in \mathcal{MR} \setminus \mathcal{RR}$, modify the class label of specific item sets $I \in \mathcal{DB}$ from $\neg C$ to C until $\text{conf}(B \rightarrow C) > t_2$ for threshold t_2.

Discrimination Discovery. To find frequent rules, the *Apriori* algorithm is used. All discovered rules whose implication consists of the class label C are partitioned into *potentially discriminatory* (PD) and *potentially nondiscriminatory* (PND) rules, where $PD = \{r = (A, B \rightarrow C) \in \mathcal{FR} \mid A \in \mathcal{A}\}$, i.e. the premise includes a sensitive attribute. According to [11], a rule $r = (A, B \rightarrow C) \in PD$ is considered *directly discriminatory* or *α-discriminatory* if $\text{elift}(r) \geq \alpha$ for some threshold $\alpha > 0$. Furthermore, a rule $r = (B, D \rightarrow C) \in PND$ is considered a *redlining rule* if there is an α-discriminatory rule $r' = (A, B \rightarrow C)$ combined with some background knowledge rules of the form $r_{b1} = (A, B \rightarrow D)$ and $r_{b2} = (D, B \rightarrow A)$.

Table 1. Measures used in [6] and probabilistic equivalents

Measure	Formula	Probabilistic Equivalent
$\text{supp}(X)$	$\lvert\{I \in \mathcal{DB} \mid X \in I\}\rvert$	$P(X)$
$\text{conf}(X \rightarrow C)$	$\frac{\text{supp}(X, C)}{\text{supp}(X)}$	$P(C \mid X) = \frac{P(X, C)}{P(X)}$
$\text{lift}(X \rightarrow C)$	$\frac{\text{conf}(X \rightarrow C)}{\text{supp}(C)}$	$\frac{P(C \mid X)}{P(C)} = \frac{P(X, C)}{P(X)P(C)}$
$\text{elift}(A, B \rightarrow C)$	$\frac{\text{conf}(A, B \rightarrow C)}{\text{conf}(B \rightarrow C)}$	$\frac{P(C \mid A, B)}{P(C \mid B)} = \frac{P(A, B, C)}{P(A)P(B, C)}$

Discrimination Prevention. To remove discrimination from the training data set, the authors of [6] propose to change the class label C of certain item sets $I \in \mathcal{DB}$ from $\neg C$ to C until the confidence of the base rule $(B \rightarrow C)$ exceeds the following thresholds:

$$t_1 \coloneqq \frac{\text{supp}(r_{b2}) \cdot (\text{conf}(r_{b2}) + \text{conf}(r) - 1)}{\text{supp}(B \rightarrow A) \cdot \alpha} \quad \text{and} \quad t_2 \coloneqq \frac{\text{conf}(r')}{\alpha} \tag{5}$$

Threshold t_1 defines the condition that a potentially nondiscriminatory rule $r \in PND$ is a redlining rule, whereas threshold t_2 defines the condition that a potentially discriminatory rule $r' \in PD$ is a directly discriminatory rule. Note that the amount of discrimination removed by this algorithm depends on the parameter α: The higher its value, the lower are t_1 and t_2, leading to less discrimination prevention. According to [6], the amount of tolerable discrimination depends on national legislation (e.g. $\alpha_{max} = 1.25$ for the USA).

However, this definition does not allow for a distinction between illegal and explainable discrimination as proposed in [9]. We will therefore investigate how the amount of illegal and explainable discrimination removed depends on different values of α, and, hence, if the algorithm causes *reverse discrimination* according to [9].

Discrimination Prevention through Preferential Sampling of Item Sets. Another approach to remove discrimination from training data, which belongs to the group of procedures that change probabilistic values to reduce potentially discriminatory correlations between attributes in the data set, was proposed in [17] and further extended in [9]. The approach aims at removing illegal discrimination through *local preferential sampling* while maintaining explainable (legal) discrimination and accuracy as high as possible, but only accepts one sensitive attribute $A \in \mathcal{A}$ as input at a time, i.e. $|\mathcal{A}| = 1$. In short, the procedure works as follows:

1. Partition the training data set $\mathcal{DB}$ into one partition for each explainable attribute. An *explainable attribute* $E \in \mathcal{B}$ is set of legal attributes that are correlated with the sensitive attribute A. Let $\mathcal{P}$ be the set of resulting partitions, and for each $P \in \mathcal{P}$, let P^a the subset of P where $(A = a)$ for all $I \in P$ (and likewise $P^{\neg a}$).
2. For each partition $P \in \mathcal{P}$ and a given number $\Delta(A, E)$ as defined in equation (6):
   i. Delete $0.5 \cdot \Delta(a, E)$ item sets $I \in P^a$ with $(C = \oplus) \in I$ and duplicate $0.5 \cdot \Delta(a, E)$ item sets $I \in P^a$ with $(C = \ominus) \in I$.
   ii. Duplicate $0.5 \cdot \Delta(\neg a, E)$ item sets $I \in P^{\neg a}$ with $(C = \oplus) \in I$ and delete $0.5 \cdot \Delta(\neg a, E)$ item sets $I \in P^a$ with $(C = \ominus) \in I$.

For explainable attribute E and for $A \in \{a, \neg a\}$, $\Delta(A, E)$ is defined as follows:

$$\Delta(\text{A}, \text{E}) \coloneqq n \cdot P(A) \cdot \left| P(\oplus| \, A, E\,) - \frac{P(\oplus| \, a, E\,) + P(\oplus| \, \neg a, E\,)}{2} \right| \tag{6}$$

Note that for each partition $P^A \in \mathcal{P}$, the item sets $I \in P^A$ are ranked according to their acceptance probability $P(C \mid A, E)$. Then, only $\Delta(A, E)$ item sets $I \in P^A$ closest to the

https://doi.org/10.30844/wi_2020_q2-schmalenbach

border between $C = \oplus$ and $C = \ominus$ are modified. This means that the decision records are only changed for candidates very close to the "decision border," resulting in minimal accuracy loss [9].

4 Results and Evaluation

Given our discrimination model from section 3.1 and our training data model from section 3.2, we now analyze to what extent and at what cost the two algorithms presented in section 3.2 are able to remove discrimination from a given data set. Moreover, we also discuss the benefits and drawbacks of the discrimination model used to perform the evaluation of the algorithms.

4.1 Experimental Setup

To perform the evaluation, we use our training data model from section 3.2 to sample a generic training data set with 5000 items. We thereby apply the parameter settings listed in Table 2, which cause the training data to contain both illegal and explainable discrimination as described in section 3.2. Specifically, we assume that the hiring decision represented in the data set is partly based on the candidates' gender and race, and that candidates with favored gender and race also have a higher probability of having the desired graduation and work experience level. We then run both algorithms on copies of the training data set and analyze discrimination removal and accuracy loss. The training data set, source code, and evaluation raw data are available on *GitHub*[1].

Table 2. Parameter values applied during the evaluation

Parameter	Value
$P(a_1)$	0.5
$P(a_2)$	0.4
$P(b_1 \mid a_1)$	0.6
$P(b_1 \mid \neg a_1)$	0.4
$P(b_2 \mid a_1, a_2)$	0.9
$P(b_2 \mid a_1, \neg a_2)$	0.5
$P(b_2 \mid \neg a_1, a_2)$	0.7
$P(b_2 \mid \neg a_1, \neg a_2)$	0.4

Parameter	Value
n	5000
θ	0.6
$(\alpha_{min}, \alpha_{max})$	(0.6, 1.8)
w_0	0.25
(w_1, w_2)	(0.125, 0.075)
(w_3, w_4)	(0.225, 0.325)
conf_{min}	0.1
supp_{min}	0.05

[1] https://github.com/kianschmalenbach/discrimination-2019

https://doi.org/10.30844/wi_2020_q2-schmalenbach

4.2 Evaluation of the Discrimination Prevention Algorithms

Discrimination Removal. Figure 2 shows the amount of discrimination included in the input data set before and after executing each of the algorithms presented in section 3.2 for different values of α. Note that, to calculate the amount of explainable and illegal discrimination, we have to adopt the formulas suggested by [9] to allow for more than one sensitive attribute. We therefore define

$$D_{all} \coloneqq \mathrm{k}^{-1} \cdot \sum_{i=1}^{k} P(\oplus |\, a_i\,) - P(\oplus |\, \neg a_i\,) \quad \text{and} \quad D_{expl} \coloneqq k^{-1} \cdot \sum_{i=1}^{k} D_{expl}^{A_i} \tag{7}$$

where $D_{expl}^{A_i}$ is the value of D_{expl} for sensitive attribute A_i as defined in [9].

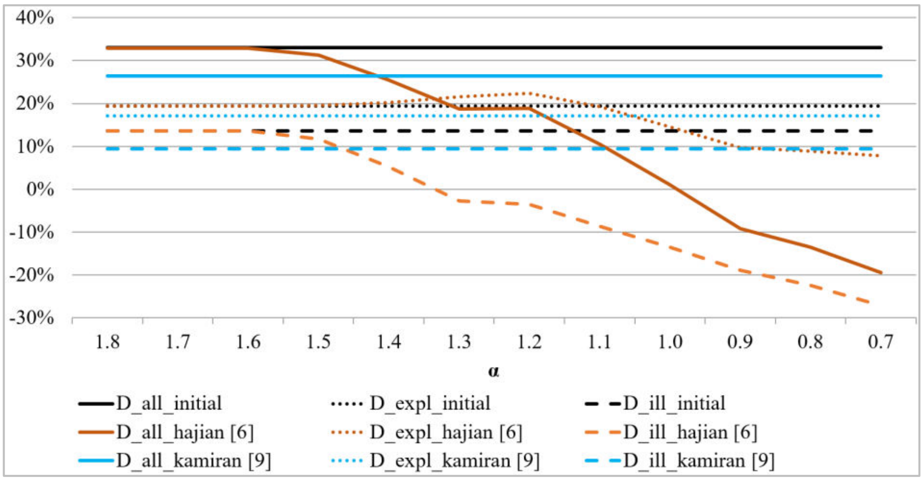

Figure 2. Amount of total, explainable, and illegal discrimination present in the training data set before and after the execution of each of the two algorithms for different values of α

Our evaluation shows that, as expected, the amount of discrimination removed by the algorithm of [6] is negatively correlated with the value of α, leading to *reverse discrimination* ($D_{ill} < 0$) for $\alpha < 1.3$. We also observe that the algorithm of [9] retains almost all explainable discrimination in the data set as desired, but does not remove all illegal discrimination, since it is not designed to handle more than one sensitive attribute simultaneously. However, both algorithms are able to significantly reduce the number of illegal discrimination initially contained in the training data set if $\alpha < 1.5$.

https://doi.org/10.30844/wi_2020_q2-schmalenbach

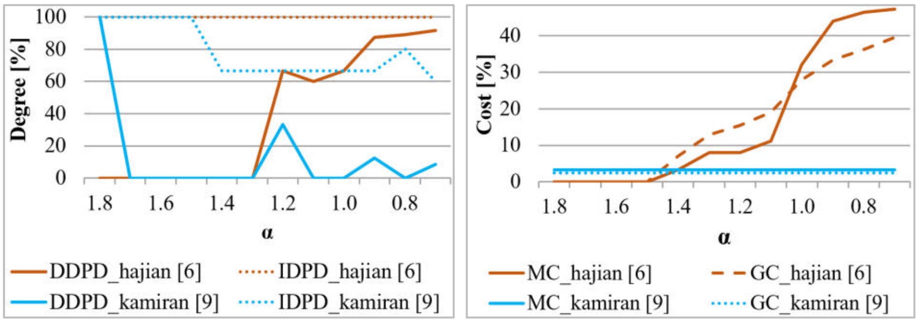

Figure 3. $DDPD$, $IDPD$, MC, and GC scores of the two algorithms, dependent on α

A second way to measure discrimination removal is to calculate the *direct* and *indirect discrimination prevention degree* as proposed in [6]. These measure the percentage of directly (resp., indirectly) discriminatory rules that vanished after the execution of the algorithms relative to the initial number of such rules. To compute the respective values, we use the *Apriori* algorithm to mine frequent rules both before and after the algorithms' execution. The results are given in the left diagram of Figure 3. Note that the results of both algorithms are dependent on the parameter α since it is used to determine whether a rule is considered discriminatory. Hence, we can observe that, as α decreases, the amount of discrimination removal increases for the algorithm proposed in [6], while it decreases for the (α-independent) one proposed in [9].

Accuracy Evaluation. Apart from assessing the discrimination removal performance, measuring the accuracy is essential for the evaluation of the algorithms. Since both algorithms modify the training data rather than a classifier, the authors of [6] propose to measure accuracy as the cost of losing information initially contained in the training data. The *misses cost* (MC) measures the percentage of frequent rules that are no longer extractable from the modified data set, and the *ghost cost* (GC) measures the percentage of new ones that were not extractable from the original data set. The right diagram of Figure 3 shows both scores for the two evaluated algorithms, dependent on α.

While the accuracy of the algorithm proposed by [9] is independent of α and comparably low due to the *local preferential sampling* technique described in section 3.2, we can observe that the cost of the algorithm proposed by [6] dramatically increases if α is set to values below 1.0. In that case, the large amount of discrimination removal significantly modifies the information originally contained in the training data set.

4.3 Evaluation of the Generic Discrimination Model

The evaluation described above is based on our generic discrimination model, which we introduced to address the limitations of previous work listed in section 2.3. Having understood its structure and its application to data sets that contain information about discriminatory hiring decisions, we can conclude that our model overcomes the main disadvantage of many existing approaches that are restricted to only one sensitive attribute, which is a highly unrealistic assumption as shown by our examples from section 3. Using a Bayesian network as the underlying probabilistic foundation, our model also allows a quantified description of explainable discrimination similar to the approach introduced by [7] and [9], but even for multiple sensitive attributes. In addition, the model can capture that part of a real hiring decision is made “at random” in the sense that the reason cannot be expressed by a purely deterministic function.

However, our model has the drawback of assuming that the sensitive attributes are pairwise conditionally independent, and, more importantly, supposes that the illegal attributes are known in advance. Moreover, our model assumes that a decision maker only distinguishes one favored and deprived group per attribute, which might be unrealistic for attributes with many potential values. Therefore, the following section addresses the evaluation results with regard to implications for future research.

5 Discussion and Implications

Discrimination in data-driven recruitment typically occurs when the unsystematic bias of (human) decision makers is learnt and hence internalized by an (algorithmic) classifier, leading to both ethical and legal problems. Even though a variety of approaches already exists to remove discrimination from training data sets before, while, or after learning classifiers, their performance is mostly depending on the application case. The fact that this is often defined according to restrictive assumptions, such as a limited number of sensitive attributes, highly impedes the process of measuring and evaluating the algorithm’s context-independent performance.

In this paper, we presented a generic discrimination model, allowing for an arbitrary number of both sensitive and legal attributes, and including a procedure to generate instances. We then compared the performance of two sophisticated algorithms, originally proposed in [6] and [9], using adapted evaluation metrics suitable for our generic model. We showed that the first algorithm generally performs better at removing discrimination from input data, but at the cost of lower accuracy and a high probability of significant *reverse discrimination*. While these two factors are mostly eliminated by the second one, it has the crucial drawback of not performing well on data sets with multiple sensitive attributes.

In conclusion, it can be said that the performance of any discrimination prevention algorithm highly depends on the underlying discrimination model and its restricting assumptions. Hence, answering the question whether data-driven recruitment leads to less discrimination first requires a concise uniform definition of discrimination, taking into account both legal and illegal forms of discrimination. Furthermore, due to a trade-off between accuracy and discrimination removal [7], it is not possible to remove discrimination from classifiers without distorting the original decision patterns.

https://doi.org/10.30844/wi_2020_q2-schmalenbach

Consequently, the question whether data-driven recruitment leads to less discrimination cannot be clearly answered from a technical point of view. As promising areas for further research, we therefore suggest investigating the following topics:

- How can the idea of illegal and explainable discrimination [9] be brought together with prevention algorithms allowing for several sensitive attributes in the training data (e.g. [6], [13])?
- How could a uniform measure of discrimination in training data look like, preferably applicable for different legal and technical contexts?
- How can discriminatory patterns in a data set be discovered without additional knowledge provision through an external domain expert?

As a final remark, we would like to point out that our discussion of discrimination-aware data-driven recruitment is limited to the algorithmic perspective of removing discriminatory patterns from classifiers. However, there are alternative approaches for avoiding discrimination in data-driven recruitment, such as exploratory discrimination-aware data mining [1], as well as different dimensions of the overall subject, such as its legal, ethical, or social dimension, that go beyond the scope of this paper.

References

1. Berendt, B., Preibusch, S.: Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artif. Intell. Law 22(2), 175–209 (2014)
2. Bogen, M., Rieke, A.: Help wanted: an examination of hiring algorithms, equity. Tech. rep., and bias. Technical report, Upturn (2018)
3. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21(2), 277–292 (2010)
4. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazonscraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (Accessed: 07.07.2019)
5. Eckhardt, A., Laumer, S., Maier, C., Weitzel, T.: The transformation of people, processes, and it in e-recruiting: Insights from an eight-year case study of a German media corporation. Employee Relations, 36(4), 415–431 (2014)
6. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. 25(7), 1445–1459 (2013)
7. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33(1), 1–33 (2011)
8. Kamiran, F., Calders, T., Pechenizkiy, M.: Techniques for discrimination-free predictive models. In: Custers, B., Calders, T., Schermer, B.W., Zarsky, T.Z. (eds.) Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases, Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 3, pp. 223–239. Springer (2013)

https://doi.org/10.30844/wi_2020_q2-schmalenbach

9. Kamiran, F., Zliobaite, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. 35(3), 613–644 (2013)
10. Laumer, S., von Stetten, A., Eckhardt, A.: E-assessment. Business & Information Systems Engineering, 1(3), 263–265 (2009)
11. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Li, Y., Liu, B., Sarawagi, S. (eds.) Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008. pp. 560–568. ACM (2008)
12. Squires, G.D.: Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. Journal of Urban Affairs 25(4), 391–410 (2003)
13. Thanh, B.L., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Apté, C., Ghosh, J., Smyth, P. (eds.) Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011. pp. 502–510. ACM (2011)
14. Wirtky, T., Laumer, S., Eckhardt, A., Weitzel, T.: On the untapped value of e-HRM: A literature review. Communications of the Association for Information Systems, 38(1) (2016)
15. Zliobaite, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Discov. 31(4), 1060–1089 (2017)
16. Zliobaite, I., Custers, B.: Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artif. Intell. Law 24(2), 183–201 (2016)
17. Zliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Cook, D.J., Pei, J., Wang, W., Zaïane, O.R., Wu, X. (eds.) 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011. pp. 992–1001. IEEE Computer Society (2011)

https://doi.org/10.30844/wi_2020_q2-schmalenbach