Noname manuscript No.
(will be inserted by the editor)

Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models

Indrė Žliobaitė · Bart Custers

Received: 2015 November 26; Revised version 2016 March 19; corrected bias 2016 June 7

Abstract Increasing numbers of decisions about everyday life are made using algorithms. By algorithms we mean predictive models (decision rules) captured from historical data using data mining. Such models often decide prices we pay, select ads we see and news we read online, match job descriptions and candidate CVs, decide who gets a loan, who goes through an extra airport security check, or who gets released on parole. Yet growing evidence suggests that decision making by algorithms may discriminate people, even if the computing process is fair and well-intentioned. This happens due to biased or non-representative learning data in combination with inadvertent modeling procedures. From the regulatory perspective there are two tendencies in relation to this issue: (1) to ensure that data-driven decision making is not discriminatory, and (2) to restrict overall collecting and storing of private data to a necessary minimum. This paper shows that from the computing perspective these two goals are contradictory. We demonstrate empirically and theoretically with standard regression models that in order to make sure that decision models are non-discriminatory, for instance, with respect to race, the sensitive racial information needs to be used in the model building process. Of course, after the model is ready, race should not be required as an input variable for decision making. From the regulatory perspective this has an important implication: collecting sensitive personal data is necessary in order to

I. Žliobaitė (corresponding author)
Helsinki Institute for Information Technology HIIT, Espoo, Finland
Aalto University, Espoo, Finland
University of Helsinki, Helsinki, Finland
E-mail: indre.zliobaite@aalto.fi

B. Custers
eLaw, Center for Law and Digital Technologies, Faculty of Law,
Leiden University, Leiden, The Netherlands
WODC, Ministry of Security and Justice, The Hague, The Netherlands
E-mail: b.h.m.custers@law.leidenuniv.nl

guarantee fairness of algorithms, and law making needs to find sensible ways to allow using such data in the modeling process.

Keywords Non-discrimination · fairness · regression · data mining · personal data · sensitive data

1 Introduction

In the era of big data many online and offline decisions are made using predictive models learned on historical data. Examples include the prices we pay, the ads we see, the jobs we get, or the news we read online. Moreover, models can decide who gets a loan, who goes through an extra airport security check, or who gets released on parole. Normally such algorithms work as follows. Suppose the goal is to build a recommender system for deciding the level of salary, given a CV. Historical data is collected including a number of observations where the outcomes, e.g., education, qualifications and experience of a person, as well as the assigned salary, are known. A model is calibrated on this historical data such that deviations from the observed values in the modeling data are minimized. A model can be seen as a mathematical function taking inputs and outputting an estimate, like salary.

Growing evidence suggests (Barocas and Selbst, 2016; Citron and Pasquale, 2014; Edelman and Luca, 2014; Kay et al., 2015; Sweeney, 2013) that decision making by inappropriately trained algorithms can discriminate people. For example, automated matching of candidate CVs and job descriptions may propagate ethnicity related biases (Ajunwa et al., 2016). Discrimination can occur even if the computing process is fair and well-intentioned. This is because most data mining methods assume that modeling data is correct, and represents the population well, which is often not true in reality (Calders and Zliobaite, 2013). While some data mining techniques have already been proposed for discovering discrimination (Hajian and Domingo-Ferrer, 2013; Luong et al., 2011; Pedreschi et al., 2008) and removing discrimination (Calders et al., 2013; Feldman et al., 2015; Kamiran et al., 2010, 2013; Kamishima et al., 2012; Mancuhan and Clifton, 2014; Zemel et al., 2013), preventing such discrimination by computational means is subject of an active ongoing research.

At the same time an ongoing regulatory concern is how to prevent digital discrimination due to technologies. Discrimination on many grounds, and in many areas of life is forbidden by national and international legislation. For example, in Finland a new act (1325/2014) came into force in January 2015, substantially expanding the scope of protection against discrimination. The act applies to nearly all public and private activities protecting against discrimination based on ethnicity, age, nationality, language, religion, belief, opinion, health, disability, sexual orientation or other personal characteristics. Currently, the European Union (EU) is preparing a new non-discrimination directive (SEC(2008) 2181), further expanding the scope of protection. Insurances and banks have been explicitly forbidden to use age and gender in estimation of risks. For instance, from January 2013 using gender to set insurance premiums is not allowed within the EU (2004/113/EC directive exempted insurance before). However, it is not yet clear how to prevent potential digital discrimination, which is due to big data initiatives. For example, the US President Office calls for technical expertise in preventing disparate impact of big data analytics (House, 2014).

From the regulatory point of view there are two main tendencies in relation to potential digital discrimination: (1) monitor that no unnecessary data is collected and stored, and (2) monitor, that predictive models do not propagate discrimination when used for decision support. It has been commonly perceived that restricting access to sensitive information should prevent discrimination from happening. But the problem is more tricky. It has already been demonstrated that simply omitting the sensitive data does not guarantee that discrimination is removed (Edelman and Luca, 2014; Kamiran et al., 2010). Removing sensitive data removes a possibility for direct discrimination, because it is no longer possible to make different decisions for two persons, who are identical in all the observed attributes but differ in the sensitive characteristics. However, this may leave room for indirect discrimination, due to the redlining effect (Hillier, 2003; Squires, 2003), which may happen when some legitimate variables are correlated with sensitive characteristics, and thus may act as a proxy for the sensitive characteristics, e.g., a zip code may carry racial information, if race is removed, discrimination may continue based on zip code.

In this study we take a further step and argue that sensitive data may actually be necessary in the modeling process, in order to be able to remove discrimination. For instance, in order to make sure that decision models are non-discriminatory with respect to race, race needs to be used in the model building process. We demonstrate this empirically and theoretically for linear regression, which is a mature and popular approach for building predictive models.

From the regulatory perspective this has an important implication: on the one hand laws protect privacy and restrict collection and use of sensitive personal data. But on the other hand, if the goal is to guarantee non-discrimination, it may be necessary to permit using sensitive variables (e.g., race) directly in the modeling. Our position is that if the law forbids to collect sensitive data, in many circumstances it would be impossible to sanitize regression models. The problem arises when sensitive data is correlated with legitimate data, e.g., when race is related to the number of years of education. If race is part of the modeling process, it becomes possible to separate the relations of race and education to salary, and isolate the contribution due to race from the model. If race is omitted, the education variable would pick up and carry forward race related biases. Unquestionably, after the model is ready, the sensitive variable should not be required for decision making.

This paper is intended as the first step towards resolving this important issue by providing proof-of-concept, facilitating a discussion and outlining important directions for research in this context. Our arguments are limited to regression models. The contributions of this paper are two-fold: (1) from the regulatory perspective we demonstrate that sensitive data is needed in order to guarantee non-discriminatory decision models, and (2) from the computing perspective we demonstrate a procedure for obtaining a baseline non-discriminatory regression model. We build upon a known theory on omitted variable bias in causality and statistics (see e.g., Pearl (2009)). The novel aspect from computing perspective is that in standard predictive modeling settings the theory focuses on incorporating the signal due to the omitted variable, while our goal in non-discriminatory data mining setting is to correctly exclude all such signals from the model. Similar approaches have been discussed in economic modeling (Pope and Sydnor, 2011), where the focus was on sanitizing regression models, our focus is on implications to data regulations.

2 Legal background

2.1 Personal data protection law

Decision-making models are often based on personal data. In the European Union (EU), the use of personal data is regulated by Directive 95/46/EC, the EU Data Protection Directive[1]. This directive was adopted in 1995 as a specification of the right to personal data protection in the EU Charter of Fundamental Rights and implemented in national legislation in all EU member states. The scope of the directive is personal data, which is defined in article 2 as any information relating to an identified or identifiable natural person. Such a person is referred to as the data subject. All processing of personal data, including the collection, recording, organization, storage, adaptation, alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction of personal data, should have a legal basis (article 7). The legal basis can be unambiguous consent of the data subject, performance of a contract, a legal obligation, a vital interest of the data subject, the public interest or other specific legitimate interests. For the processing of personal data, the directive introduces a set of principles for the fair processing of information:

- the collection *limitation principle*, stating that [t]here should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject[2];
- the *data quality principle*, stating that [p]ersonal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date;

[1] European directive 95/46/EG of the European Parliament and the Council of 24th October 1995, [1995] OJ L281/31. See also `http://europa.eu.int/eur-lex/en/lif/dat/1995/en_395L0046.html`

[2] This principle is sometimes referred to as the *principle of minimality*, see Bygrave (2002), p. 341.

- the *purpose specification principle*, stating that [t]he purposes for which personal data are collected should be specified [...] and that the data may only be used for these purposes;
- the use limitation principle, stating that [p]ersonal data should not be disclosed, made available or otherwise used for purposes other than those specified, [...] except a) with the consent of the data subject; or b) by the authority of law;
- the *security safeguards principle*, stating that reasonable precautions should be taken against risks of loss, unauthorised access, destruction, et cetera, of personal data;
- the *openness principle*, stating that the subject should be able to know about the existence and nature of personal data, its purpose, and the identity of the data controller;
- the *individual participation principle*, stating, among other things, that the data subject should have the right to have his personal data erased, rectified, completed, or amended[3];
- the *accountability principle*, stating that the data controller should be accountable for complying with measures supporting the above principles.

The former four principles focus on the data and the conditions under which processing of the data is allowed, and the latter four principles are duties of those responsible for the processing of personal data and rights of the data subjects.

Although there is separate legislation for addressing discrimination and equal treatment (see Section 2.2), the Data Protection Directive also addresses special categories of data. Article 8 of the directive states that the processing of personal data is prohibited when it concerns so-called sensitive personal data, i.e., personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life. Article 8 mentions several exemptions (such as explicit consent in some situations) to this strict ban on the use of sensitive personal data. Note that these categories of sensitive data are similar to the categories of sensitive data mentioned in many anti-discrimination law, but are not exactly the same. For instance, in many countries gender and age are also characteristics that are not allowed for particular decision-making, for instance, when hiring or firing employees.

It is important to note that the current legal framework for personal data protection is under revision. In 2012 the European Commission kicked-off the data protection reform by publishing its proposal of the EU regulation on personal data protection[4]. The objective of the new legal framework was to strengthen data protection, unify national legislation across member states

[3] Note that, in the European Data Protection Directive and the WBP, this principle applies only to incomplete or inaccurate data, or data that are irrelevant or processed illegitimately.

[4] Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 25.1.2012 COM(2012) 11 fi-

and adapt to the changed circumstances in a globalized and intra-connected world. The privacy principles described above remain unchanged in the proposed legislation, but new rights like the right to data portability and the right to be forgotten (or, more precisely, the right to erasure as it is phrased in the proposed legislation) are introduced, as well as mandatory privacy impact assessments for some situations. For more background on the proposed EU data protection regulation, we refer to further literature (Hornung, 2012; Kuner, 2012).

In short, under the EU legal framework the processing of personal data is subject to several legal restrictions and the processing of sensitive personal data is basically not allowed unless a specific exemption exists. In most cases, the use of sensitive personal data for improving decision-making models and predictive models is not allowed. In fact, by restricting the collection and use of personal data to a minimum, data may not even be available to perform such modelling. After the models are ready, the use of sensitive personal data is usually not allowed as input for decision-making as this may involve discriminatory decision-making, as we will explain in the next subsection.

Note that the law only addresses the data, not the modeling process. The only legal requirement relating to the process is that people should not be subjected to (entirely) automated decision-making[5]. This mainly refers to using models rather than modeling itself.

2.2 Anti-Discrimination Law

The EU legal framework for anti-discrimination and equal treatment is constituted by several directives, including the Race Equality Directive (2000/43/EC), the Employment Equality Directive (2007/78/EC), the Gender Recast Directive (2006/54/EC) and the Gender Goods and Services Directive (2006/113/EC). These directives specify the anti-discrimination and equal treatment provisions in the EU Charter of Fundamental Rights. There is no general directive stating which attributes can and cannot be used for which types of decision-making, but EU member states can implement these directives in one single piece of legislation. All the directives mentioned above have in common that they prohibit the use of (a) particular characteristics, like gender, race, ethnicity, for (b) particular decision-making, like hiring or firing employees, providing particular goods and services and access to education. It is important to note that when assessing whether discrimination took place, it is always about the combination of characteristics and decisions: the use of sensitive characteristics for decision-making does not always or automatically result in discrimination, this depends on the type of decision that is made. An example of this may be so-called positive discrimination (also referred to as affirmative action) in

nal 2012/0011 (COD). Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:PDF

[5] Art. 15 of the EU directive on the protection of personal data.

which favoring members of disadvantaged groups is used as a policy. Furthermore, types of decisions that are not mentioned in anti-discrimination law (for instance, personal decisions with whom someone wants to associate) cannot constitute discriminatory acts, even when sensitive personal data are used for such decision-making.

Within EU anti-discrimination law, it is important to distinguish between legislation relating to discrimination on specific protected grounds (such as race, gender or age) and the general protection of equal treatment (Gellert et al., 2013). Equal treatment is based on the idea that similar situations should be treated similarly unless other treatment is objectively justified. Obviously, in practice no situations are exactly the same, which raises the question which characteristics are taken into account to assess whether situations are similar or different. In EU law, this is assessed only marginally: as long as a difference in treatment has some rationality and is not completely arbitrary, the quality of the underlying reasoning is not further assessed (Gellert et al., 2013; McCrudden and Prechal, 2009).

Most anti-discrimination law distinguishes between direct and indirect discrimination. Direct discrimination is discrimination on the basis of prohibited characteristics like religion, philosophy of life, political orientation, race, gender, nationality, sexual orientation, or civil status. Indirect discrimination is discrimination on the basis of other characteristics resulting in direct discrimination. An important phenomenon that may be mentioned in this context is so-called *masking*[6]. When particular characteristics are found to be correlated, it may be possible to use trivial characteristics as indicators of sensitive characteristics. For instance, when people living in a particular zip code area have a high health risk, insurance companies may use the zip code (trivial information) as an indication of a persons health (sensitive information), and may thus use the trivial information as a selection criterion. Using geographical characteristics for profiling is referred to as redlining (Hillier, 2003; Squires, 2003). Note that refusing insurance on the basis of a zip code may be acceptable, to some extent, as companies may choose (on the basis of market freedom) the geographic areas in which they operate. On the other hand, refusing insurance on the basis of sensitive data may be prohibited on the basis of anti-discrimination law. Masking may not be transparent for a data subject, as he may not know the consequences of disclosing trivial information, such as a zip code (Custers et al., 2013b).

In short, anti-discrimination law prohibits the use of several characteristics for several types of decision-making, whether direct or indirect. Data mining may reveal proxies for sensitive characteristics, but these cannot be used for decision-making when the results are similar to direct discrimination based on the sensitive characteristics.

[6] ECJ, C-127/07, 16 December 2008

3 Omitted variable bias

Next we will demonstrate empirically with a simple toy example, and theoretically for regression models, that omitting the sensitive characteristic (such as race) from the equation does not make a decision model free from discrimination, and, in order to remove biases, we actually need to use the sensitive characteristic in the modeling process.

3.1 Toy example

For illustration purposes imagine a simplified society, where monthly salary is assumed to depend on years of education, but in fact, due to some prejudice, it actually depends not only on education, but also on ethnicity of a person (native or immigrant). Suppose that the true underlying mechanism how salaries are decided is:

$$\textit{salary} = 1000 + 100 \times \textit{education} - 500 \times \textit{ethnicity}, \tag{1}$$

where *education* represents years of education, *ethnicity* is 0 for natives and 1 for immigrants. That is, people with no education get 1000 base salary, and for every year of education there is 100 extra. Immigrants get 500 less in all circumstances. This is a fictitious example.

Now consider that the decision rules are unknown to the society, and data a data scientist wants to develop a salary recommendation system using observed data, given in Table 1. Either due to belief that salary should depend only on education, but not ethnicity, or due to legislation explicitly forbidding to use ethnicity, the data scientist decides to omit ethnicity, and use only education when building the model. The assumed model form is

Table 1 Toy example: synthetic data about salaries, generated according to Eq. (1).

education	*ethnicity*	*salary*	*education*	*ethnicity*	*salary*
1	1	600	1	0	1100
2	1	700	6	0	1600
3	1	800	7	0	1700
4	1	900	9	0	1900
10	1	1500	10	0	2000

$$\textit{salary} = b_0 + b_1 \times \textit{education}, \tag{2}$$

where b_0 is a coefficient denoting the base salary, and b_1 is a coefficient denoting how much extra a person should get for each extra year of education. The data scientist will find the coefficients from the data.

After running the standard regression fitting procedure (ordinary least squares) on the data in Table 1, the following model is obtained:

$$\textit{salary} = 602 + 128 \times \textit{education}. \tag{3}$$

We can see that people with no education would get 602 base salary, instead of 1000 which they are supposed to get according to the ground truth in Eq. (1). In addition, for every extra year of education a person would get 128, which is more than it is supposed to be according to the true underlying process (100). The fitted model punishes people with low education more than necessary, and rewards people with high education more than deserved, which is by itself already incorrect reflection of the underlying process, and would make incorrect recommendations, if used as a recommender system for salaries.

Moreover, in this fictitious society immigrants tend to have lower education (see Table 1). In other words, education variable is correlated with ethnicity. Hence, not only the learned model is incorrect, but immigrants suffer from that incorrectness more, because they tend to have lower education. This is an instantiation of *redlining*.

The example demonstrates that removing the ethnicity from the modeling process does not ensure that the model is free from discrimination. If instead the data scientist includes the sensitive attribute when fitting the model, and afterwards removes *the model component* related to ethnicity, the resulting model will be correct in terms of education and free from discrimination with respect to ethnicity.

Fitting a regression model on the complete data in Table 1 recovers $salary = 1000 + 100 \times education - 500 \times ethnicity$. Now we can remove the ethnicity component, replacing it by a constant which does not depend on ethnicity, in order to get a correct and discrimination-free model for recommendations:

$$salary = 1000 + 100 \times education - c. \tag{4}$$

The constant c can be zero, or computed from the data as will be discussed in the next section. We can argue that as long as c is the same for both ethnicities, this model will not discriminate, because the salary will correctly reflect the base level, and the variable component for extra education (as per underlying Eq.(1)).

3.2 Theoretical explanation

Omitted variable bias occurs when a regression model is fitted leaving out an important causal variable. The problem is well known in statistics, particularly in analyzing data from experimental trials aiming at discovering causal relationships (see e.g., Pearl (2009)), but these issues have not yet been vigorously discussed in the context of discrimination-aware data mining.

Let the true underlying model behind data be:

$$y = b_0 + b_1 x + \beta s + e, \tag{5}$$

where x is a legitimate variable (such as education), s is a sensitive variable (such as ethnicity), y is the target variable (such as salary), e is random noise with the expected value of zero, and β, b_1, and b_0 are non-zero coefficients.

Assume a data scientist decides to leave out the sensitive variable s, and fit the following model using the standard (OLS) procedure:

$$y = \hat{b}_0 + \hat{b}_1 x. \tag{6}$$

Then the estimates of the regression coefficients will be biased in the following way,

$$\begin{aligned} \hat{b}_1 &= b_1 + \Delta, \\ \hat{b}_0 &= b_0 + \beta\bar{s} - \Delta\bar{x}, \\ \textbf{where } \Delta &= \beta\frac{\hat{Cov}(x,s)}{\hat{Var}(x)}, \end{aligned} \tag{7}$$

where Δ is the bias that depends on the underlying data, $\bar{s}$ is the mean of the sensitive variable, and β is the true underlying bias towards ethnicity in the data. Proofs are given in Appendix A.

When s is omitted, $\hat{b}_1$ contains a bias, which does not go away even if we collect infinitely many observations for training the model. There would be no bias only in case the true $\beta = 0$, that is, the underlying data is free from inequalities in relation to the sensitive variable, or $Cov(x, s) = 0$, that is, the sensitive variable is not related to the legitimate variable (e.g., education variable is not related to ethnicity). For instance, if immigrants tend to have lower education, then the regression model would 'punish' low education even further by offering even lower wages to people with low education (who are mostly immigrants). Thus, in most realistic cases, not only removing the sensitive variable does not make regression models fair, but on the contrary, such a strategy is likely to amplify discrimination.

We advocate that a better strategy for sanitizing regression models would be to learn a model on full data including the sensitive variable, then remove the component with the sensitive variable, and replace it by a constant that does not depend on the sensitive variable.

3.3 A baseline for regression

We advocate the following procedure and model as a baseline for non-discriminatory regression. Firstly, build a regression model on a full dataset including the sensitive variable.

$$y = b_0 + b_1x_1 + \ldots + b_kx_k + \beta s, \tag{8}$$

where $x_1, \ldots, x_k$ are legitimate input variables (k is the number of variables, could be one or more), and s is the sensitive variable against which discrimination is forbidden, $b_0, \ldots, b_k, \beta$ are regression coefficients. The final model is obtained by replacing a component containing s with a constant c

$$y = b_0 + b_1x + \ldots + b_kx_k + c, \tag{9}$$

where c is a constant that depends on the assumptions about the source of the underlying inequalities in the data.

For example, one could assume that the salary paid to natives is correct, and the salary paid to immigrants is lower than it is supposed to be. In this case $c = 0$ (assuming that $s = 0$ denotes natives). Alternatively, one could assume that the salary paid to immigrants is correct, and natives get extra bonus. In this case $c = \beta$. Likely, the salary that is paid for the majority is correct, therefore, we suggest to use a weighted average of both, where the weight reflects the balance between both population groups in the data (this works for numeric sensitive variables as well, e.g., age). We suggest using $c = \bar{s}\beta$, where $\bar{s}$ is the mean value of the sensitive attribute over historical data, and β is the regression coefficient from the full regression model in Eq. (8).

3.4 Implications

The procedure we just described is referred to as a baseline, not a complete solution, since the effectiveness of this procedure heavily depends on the underlying data. This baseline assumes that inequalities have happened and are reflected in data in a certain way. It assumes that bias is additive, and only depends on the sensitive attribute, but not on the other input variables. This is modeled by linear regression. This is a simplifying assumption, and we anticipate that coming from this reasoning more sophisticated approaches can be developed.

The theoretical conclusion is only valid for linear models, because we have a theoretical proof for linear models. Formal conclusion for non-linear models remains a subject for future investigation. However, our intuition from working in this field and observing the behaviour of various data mining and machine learning models is that similar principles apply, but to what extent, and what models are more or less sensitive, remains to be researched.

An existing pilot study on discrimination prevention for regression (Calders et al., 2013) relates topic-wise, but has a different focus. Our focus is to analyze the role of the sensitive variable in removing discrimination, and demonstrate that it is necessary to use it for discrimination prevention. The study by Calders et al. (2013) is, of course, using the sensitive variable for formulating non-discrimination constraints, which are enforced during model fitting. But a discussion about the role of the sensitive variable is not the subject of their study. Similar approaches have been discussed in economic modeling (Pope and Sydnor, 2011), where the focus was on sanitizing regression models, our focus in this paper is on implications to data regulations.

We deliberately did not introduce or reuse any discrimination measures, since measuring potential discrimination by algorithms is a subject of scientific discussion (see e.g., a broad review by Romei and Ruggieri (2014), or a recent survey by Zliobaite (2015) for a discussion of various existing measures). Instead, we built our reasoning upon assuming an underlying decision process, and then checking, how learning a model from observed data would

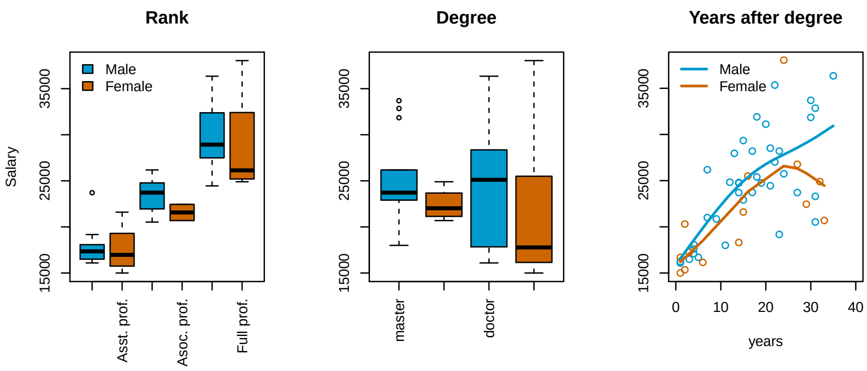

Fig. 1 Summary of the dataset characteristics.

behave with respect to that ground truth. This approach would not guarantee non-discrimination if the assumption about the underlying decision process is substantially different, but we suggest this approach as a starting baseline in search for advanced solutions. Overall, our main purpose with this discussion is to provide evidence for a need to collect and use sensitive data for discrimination prevention, and initiate a scientific discussion about regulatory actions thereof.

4 Case study

Now that we have discussed computational issues, we present an experimental case study with real-world data. The purpose is to illustrate our arguments with experimental analysis of a realistic case.

We use a dataset recording salaries of professors at a US college (Weisberg, 1985)[7]. Our predictive modeling task for the case study is to predict the salary for a person, controlling for discrimination with respect to gender, taking as input a rank (assistant, associate, full), a degree (masters, doctorate), and years since degree (a proxy for job experience). As a disclaimer, our goal is not to discover any particular discrimination in this data, but only to analyze the situation from the computational point of view.

The dataset records 52 persons, out of which 14 (27%) are female, and 38 male. Figure 1 provides a summary of data characteristics. We can see that there is some imbalance in the data between salaries for different genders at similar levels of qualification, therefore, there is a potential for a learned model to pick up and carry forward this imbalance.

Table 2 gives correlations between the input variables. We can see that salary is mostly correlated with rank and years of experience. Gender is mostly correlated with salary and rank. Hence, there is a relation of the sensitive

[7] Obtained from: `http://data.princeton.edu/wws509/datasets/\#salary`

Table 2 Correlations between input variables.

	gender	rank	degree	years	salary
gender	1.00	−0.23	0.08	−0.09	−0.25
rank		1.00	−0.01	0.70	0.87
degree			1.00	−0.48	−0.07
years				1.00	0.67

variable with the target variable, and a legitimate variable, therefore, we can expect omitted variable bias to occur, if we leave gender out of the model.

On this data we learn the following predictive models:

M1 - a standard model using all the variables including gender,
M0 - a blind model leaving out the sensitive attribute,
MM - a model trained only on male data,
MF - a model trained only on female data.

Table 3 presents the learned models. We can see that the standard model M1 subtracts 950 USD for females ($s = 1$). When we leave out the gender variable, the blind model M0 has different coefficients for rank, degree and years of service. There is an extra premium (228 USD) for each higher rank, a lower premium (219 USD) less for a higher degree, and a slightly lower premium (by 15 USD) for an extra year of service. As can be seen from correlations in Table 2, most of these adjustments in comparison to the base model will negatively affect salaries for females, as females have lower ranks, and higher degrees. Moreover, these automated adjustments will negatively affect males, which in their characteristics are more similar to an average female profile, that is, males in lower ranks. This is due to omitted variable bias, discussed in the previous section. A more convincing adjustment would be to remove the component that reduces salaries for females, while keeping all the other coefficients in tact, as discussed in the previous section.

Table 3 Learned models.

	salary				rank		degree		years		gender
Standard model (M1):	w	=	11956	+	$4993r$	+	$398d$	+	$103y$	−	$950s$
Blind model (M0):	w	=	11604	+	$5231r$	+	$179d$	+	$88y$		
Only males (MM):	w	=	11705	+	$5032r$	−	$31d$	+	$129y$		
Only females (MF):	w	=	10117	+	$4567r$	+	$2399d$	+	$116y$		

Separate models for males and females may be mistakenly seem neutral in a sense that they do not use gender variable as input. In fact, gender is used at the very beginning for splitting the data. Therefore, such a solution is not to be expected to be fair, but it may highlight interesting differences in decision mechanisms for the two groups. We can see that the female model MF has much lower base salary, and much higher coefficient for degree, while the male model MM has even a small negative addition for a higher degree,

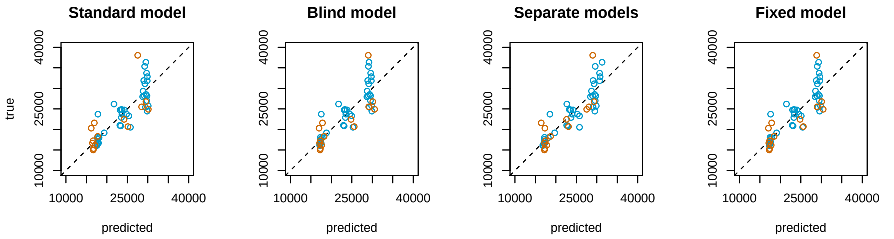

Fig. 2 Model outputs (predictions) vs. true values on the training data (orange - female, blue - male).

but the initial base salary is much higher. Hence, we see that there are some potential differences in decision making mechanisms.

Let us investigate how these models perform in terms of prediction accuracy. Figure 2 plots the model outputs against the true values in the data. On the horizontal axis we can see three clear salary groups emerging from the predictions, likely corresponding to the three levels of professorship (assistant, associate, full). We also observe in the standard model that females get assigned lower salaries within each group, particularly within the lowest salary group. This is not surprising given that rank is used as model input. As rank is strongly correlated with gender, this suggests that rank may potentially capture and propagate indirect discrimination, which will be verified using discrimination measures. Separate models seem to be disadvantaging females in the highest salary group, which may provide some insights into the decision making mechanisms. Blind model and fixed model visually perform similarly to each other. Generally we would expect fixed model to have a clear advantage, but in this particular case study the blind model is not performing too badly perhaps since coincidently there is a tradeoff (added and removed discrimination) due to opposing correlations of the legitimate variables with the gender (rank "-", degree "+", years "-").

We further investigate the performance via leave-one-out cross-validation, which builds a model on all but one observations, and tests the model on that remaining observation, and then moves to the next observation until all observations have served as testing points. We measure accuracy and discrimination by the models. Accuracy is measured on the original salaries corrected as follows: for each female 694 USD is added, for each male 255 USD is subtracted as per assumption of the correct salary discussed in Section 3.3 $c = \bar{s}\beta$, where $\bar{s} = 0.27$ is the proportion of females, and $\beta = -950$ is the coefficient from the learned regression for gender.

Discrimination is measured by propensity score matching as used in (Calders et al., 2013). This is a generalization over previously used conditional discrimination measure (Kamiran et al., 2013), the latter does not require explicitly designating explanatory variables. Conditional discrimination as well as propensity score take into consideration possible legitimate explanations for the differences in salaries between male and female, such as number of years in job experience. If we measured discrimination as a simple the difference between males and females, possible explanations of the differences would be ignored, and discrimination would be potentially overstated.

The idea behind propensity score matching is as follows. For each candidate (either male or female) we model the probability of being a female based on the profile. This allows us to group similar profiles together and measure differences in salaries within similar profiles. We use logistic regression. Rank, degree and years are used as inputs, and gender is the target variable. Given the propensity scores estimated by the logistic regression model, we divide the dataset into three groups with 17-18 candidates in each group. If there is no discrimination, distribution of salaries for males and females within the groups is expected to be similar, because qualifications of candidates within each group are similar. Within each group a normalized discrimination measure is computed $D = 2AUC - 1$ such that $D \in [1, -1]$, where 1 means maximum discrimination, 0 means no discrimination, and -1 means maximum reverse discrimination. AUC is the area under ROC curve of salaries with respect to gender.

Figure 3 presents the results. In terms of accuracy the desired result is as high as possible, and in terms of discrimination - as close to zero as possible, so being high on the dashed line in the plot is desired. We can see that on the training data (solid circles) the standard (M1) and separate models (MF and MM) are highly discriminatory, while the blind model (M0) and the fixed model (Mfixed) perform similarly well, with the fixed model having a small advantage of less reverse discrimination. This is an interesting result taking into consideration how different the model coefficients are (as analyzed in Table 3). When comparing training accuracies measured on modeling data with testing accuracies measured on unseen data via cross-validation, we see a common patter than accuracies on unseen data are lower. The highest drop is in the accuracy of separate models, which showed the highest accuracy on the training data. This suggests possible overfitting, when from a small sample training data is learned so well that not only underlying signal, but also a substantial amount of noise is captured. As a result, separate models does not generalize well. This is not surprising, since separate models (MF and MM) have half training data (only males or only females) than the rest of the models. Generally, the more data, the better generalization.

We discussed aggregated results, since our goal was to illustrate the necessity of sensitive data for making models free of discrimination. While at the aggregated level discrimination is removed as intended, looking closer into the subgroups of people with similar characteristics shows remaining biases at lower levels, that were not aimed at capturing in this case study. The principles that we have discussed for the whole population could be propagated to subgroups in a similar manner as suggested by Kamiran et al. (2013); a concrete solution for subgroups is a subject for future investigation. The main message from our case study is that Table 3 supports our theoretical argu-

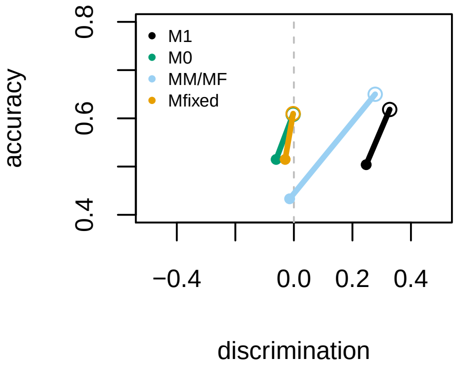

Fig. 3 Accuracies and discrimination of different models: standard model (M1), blind model (M0), separate models (MM and MF), and fixed model (Mfixed). Solid circles present performance on the training data, and empty circles present cross-validation results.

ments presented in Section 3 that availability of sensitive data is critical for being able to remove discrimination from predictive models.

5 Discussion from the legal perspective

As a general rule, more data may provide better models, as more attributes can be taken into account. More refined models may better describe actual situations. However, in some situations more accurate descriptions may not be desirable because they reveal sensitive information about people that they prefer not to have disclosed (Custers, 2012; Kosinski et al., 2013), or because they reveal patterns that society does not consider suitable for decision-making, for instance, if such patterns enable discrimination (Custers et al., 2013a). The general idea that excluding sensitive data from the process of creating decision-making models would yield non-discriminatory models does not hold (Kamiran and Calders, 2009). In the previous sections, we demonstrated that including sensitive data in the modelling process may not only result in more accurate models, but also in non-discriminatory models.

As a result we conclude that, to some extent, there exists a contradiction between data protection requirements limiting the use of personal data, specifically the use of sensitive personal data, and anti-discrimination law, calling for non-discriminatory models. Including sensitive personal data in decision-making models may yield less discrimination but is not always allowed under the current data protection law. There are some solutions, however.

For the first solution, it is important to note that anonymous data is not subjected to personal data protection law, as anonymous data does not match the description of personal data, i.e., data relating to identified or identifiable persons. Therefore, from a legal perspective it is allowed and possible to create models in the way described above on data, including sensitive data that is anonymized, that are non-discriminatory. Obviously, at the moment such models are used for decision-making regarding specific individuals, personal data has to be used as input for the model. At that moment, for instance, when ascribing credit scores or other risks to individuals, the data is no longer anonymous and the personal data protection law fully applies. Obviously, it is not allowed to use sensitive personal data as input for decision-making based upon such models.

In some cases, however, it may not be sufficient to use anonymous data for modelling processes (for instance, when identifying data is part of the model) or it may not be easy to anonymize the data (for instance, when so many attributes for per record are available that spontaneous recognition may occur) (Ohm, 2010). For these situations, other solutions are required to be able to use sensitive data. In some cases a solution in the existing legal framework can be found in the fact that the EU legal framework for personal data protection provides exemptions when data subjects provide consent. However, asking data subjects for consent may be unrealistic in the era of Big Data, with records of millions of data subjects or more (Schermer et al., 2014). Since the modelling process described above can provide non-discriminatory decision-making models, we argue that (as a second solution) future personal data protection law should provide an exemption to the use of sensitive personal data when used for creating models that explicitly intend to reduce discrimination. Given the sensitivity of such data, it is obvious that strict conditions should apply. The national data protection authorities could be mandated to handle such requests and impose conditions.

6 Conclusion

Using sensitive personal data in decision making by algorithms is subject to several legal regulations, including personal data protection law and anti-discrimination law. Regulations tend to restrict overall collecting and storing personal data to a necessary minimum, and at the same time to ensure that data-driven decision making is not discriminatory. We have argued that from the computing perspective these two goals are contradictory. We have demonstrated empirically and theoretically with standard regression models that, in order to make sure that decision models are non-discriminatory, e.g., with respect to race, the sensitive racial information needs to be used in the model building process. After the model is ready, race should not be required as an input attribute for decision making. From the regulatory perspective this has an important implication: collecting sensitive personal data is needed in order to guarantee fairness of algorithms, and law making needs to find sensible ways to allow using such data in the modeling process.

References

Ajunwa, I., Friedler, S., Scheidegger, C., and Venkatasubramanian, S. (2016). Hiring by algorithm: Predicting and preventing disparate impact. *SSRN.*

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104.

Bygrave, L. (2002). *Data protection law; approaching its rationale, logic and limits*, volume 10 of *Information law series*. The Hague, London, New York: Kluwer Law International.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling attribute effect in linear regression. In *Proc. of 13th IEEE ICDM*, pages 71–80.

Calders, T. and Zliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Inf. Society*, pages 43–57.

Citron, D. K. and Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89(1).

Custers, B. (2012). Predicting data that people refuse to disclose; how data mining predictions challenge informational self-determination. *Privacy Observatory Magazine*, 3.

Custers, B., Calders, T., Schermer, B., and Zarsky, T., editors (2013a). *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Heidelberg: Springer.

Custers, B., Van der Hof, S., Schermer, B., Appleby-Arnold, S., and Brockdorff, N. (2013b). Informed consent in social media use. the gap between user expectations and eu personal data protection law, SCRIPTed. *Journal of Law, Technology and Society*, 10:435–457.

Edelman, B. G. and Luca, M. (2014). Digital discrimination: The case of Airbnb.com. Working Paper 14-054, Harvard Business School.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proc. of 21st ACM KDD*, pages 259–268.

Gellert, R., Vries, K. d., Hert, P. d., and Gutwirth, S. (2013). A comparative analysis of anti-discrimination and data protection legislations. In *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Heidelberg: Springer.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.*, 25(7):1445–1459.

Hillier, A. (2003). Spatial analysis of historical redlining: A methodological explanation. *Journal of Housing Research*, 14(1):137–168.

Hornung, G. (2012). A general data protection regulation for europe? light and shade. *the Commissions Draft of 25 January 2012, 9 SCRIPTed*, pages 64–81.

House, T. W. (2014). *Big Data: Seizing Opportunities, Preserving Values.*

Kamiran, F. and Calders, T. (2009). Classification without discrimination. In *IEEE Int. Conf. on Computer, Control & Communication*, IEEE-IC4. IEEE press.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *Proc. of 10th IEEE ICDM*, pages 869–874.

Kamiran, F., Zliobaite, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Proc. of ECMLPKDD*, pages 35–50.

Kay, M., Matuszek, C., and Munson, S. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proc. of 33rd ACM CHI*, pages 3819–3828.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proc. of the National Academy of Sciences (PNAS)*, 110(15):5802–5805.

Kuner, C. (2012). The european commission's proposed data protection regulation: A copernican revolution in european data protection law. Privacy and Security Law Report.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of 17th KDD*, pages 502–510.

Mancuhan, K. and Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, 22(2):211–238.

McCrudden, C. and Prechal, S. (2009). The concepts of equality and non-discrimination in europe. european commission. DG Employment, Social Affairs and Equal Opportunities.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1765.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proc. of 14th ACM KDD*, pages 560–568.

Pope, D. G. and Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3):206–31.

Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowledge Eng. Review*, 29(5):582–638.

Schermer, B., Custers, B., and Van der Hof, S. (2014). The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics & Information Technology*, 16(2):171–182.

Squires, G. (2003). Racial profiling, insurance style: insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410.

Sweeney, L. (2013). Discrimination in online ad delivery. *Commun. of the ACM*, 56(5):44–54.

Weisberg, S. (1985). *Applied Linear Regression, Second Edition.*

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *Proc. of 30th ICML*, pages 325–333.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *CoRR*, abs/1511.00148.

A Omitted variable bias

We provide a theoretical expectation for the omitted variable bias in the ordinary least squares (OLS) estimation of linear regression coefficients. The theory is known in multiple statistical textbooks, we adapt the reasoning for discrimination prevention. For better interpretability we focus on a simple case with one legitimate variable, extension to more variables is straightforward.

Let the true underlying model behind data be

$$y = b_0 + b_1 x + \beta s + e, \tag{10}$$

where x is a legitimate variable (such as education), s is a sensitive variable (such as ethnicity), y is the target variable (such as salary), e is random noise with the expected value of zero, and β, b_1, and b_0 are non-zero coefficients.

Assume a data scientist decides to fit model $y = \hat{b}_0 + \hat{b}_1 x$.

Following the standard (OLS) procedure for estimating regression parameters the data scientist gets:

$$\hat{b}_1 = \tfrac{\hat{Cov}(x,y)}{\hat{Var}(x)}, \tag{11}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}, \tag{12}$$

where bar denotes the mean, and hat denotes that it is estimated from data.

Next we plug-in the true underlying model from Eq. (10)

$$\begin{aligned}
\hat{b}_1 &= \frac{\hat{Cov}(x, b_0 + b_1 x + \beta s + e)}{\hat{Var}(x)} && (13)\\
&= \frac{\hat{Cov}(x, b_0)}{\hat{Var}(x)} + \frac{b_1 \hat{Cov}(x, x)}{\hat{Var}(x)} + \frac{b_2 \hat{Cov}(x, s)}{\hat{Var}(x)} + \frac{\hat{Cov}(x, e)}{\hat{Var}(x)}\\
&= b_1 + \beta \frac{\hat{Cov}(x, s)}{\hat{Var}(x)},
\end{aligned}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = b_0 + b_1 \bar{x} + \beta \bar{s} - b_1 \bar{x} - \beta \frac{\hat{Cov}(x, s)}{\hat{Var}(x)} \bar{x} \tag{14}$$

$$= b_0 + \beta \bar{s} - \beta \frac{\hat{Cov}(x, s)}{\hat{Var}(x)} \bar{x}. \tag{15}$$

This demonstrates that when there is discrimination in the data, that is $\beta \neq 0$, unless $Cov(x, s)$ is zero, the estimates $\hat{b}_1$ and $\hat{b}_0$ will be biased by a component that depends on x and thus carries forward discrimination.