Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Empirical study about the influence of social dimensions on the SCHUFA-Score

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science (M. Sc.)

eingereicht von: Philipp Waack
geboren am: 11.08.1991
geboren in: Hamburg

Gutachter/innen: Prof. Dr. Niels Pinkwart
Dr. Patrick Jähnichen

eingereicht am: ..... verteidigt am: .....

Contents

1. Introduction

The increasing importance of digitization leads to considerable change in many different social areas. Digital transformation includes the automation of processes and data-driven rationalization to assist are responsible person in making a decision. While this carries the potential of more effective and informed decision making, it can also cause social harm when such processes are not well accompanied by careful quality assurance. An automated process may incorporate biases in specific social groups and a lack of awareness can lead to decisions that disadvantage such group members. Automated processes assisting in decision making that affects persons in a society may include biases of this kind, lead to decisions that are viewed as unfair [33], [20]. Controversies concerning this problem are mainly discussed regarding machine learning techniques. This led to a growing field in the scientific discipline of computer science and machine learning which examines fairness and discrimination in data and statistical learning techniques [51], [58], [41].

Public debate about automated systems that aid decisions in a social context is difficult, because most of such systems are commercial and their logic is mainly opaque to people affected by them. An attempt to raise awareness on this issue was the openSCHUFA project which used crowd-sourcing to collect data about the solvency-scoring system of the SCHUFA [56]. SPON and BR published a series of articles based partly on the analysis of the data [16]. In this thesis we want to contribute to public discussion regarding the transparency of algorithms in socially relevant decision processes. We focus on the discussion about the influence of social group dimensions on a solvency-score in the context of discrimination. The basis of the discussion is a case study based on the data of the openSCHUFA-project. Concretely, we examine the solvency-scores of the SCHUFA credit bureau, which are mainly constituted by an algorithm and are used as a commercial product to help in different consumer-relevant contexts of decision making [2].

In section 2 we introduce core concepts that are relevant for the analysis and discussion. We then describe the concrete context of our case study and its relation to the defined concepts in section 3. section 4 gives a brief overview about related work of this thesis in the fields of discrimination-aware data mining and fairness in machine learning. In section 5 we turn to the methods used for the analysis. In section 6 we describe the actual analysis, starting with a description of the data set at hand. Then we derive concrete hypotheses that concretely consider the statistical influence of the social dimensions of sex, age and being located in either western or eastern

German states on the different versions of the bank and mail order sector scores of the SCHUFA credit bureau. We describe the proceeding in the analysis of the hypotheses and interpret our result. Finally, in section 7 we discuss the results and its limitations in the social context of discrimination.

2. Discrimination and Algorithm-assisted Decision Making

A main concern discussed regarding the use of algorithms to help in a decision making process is a morally wrongful decision made based on the output of an algorithm. To express this wrongfulness, there seem to be different terms that are partly used interchangeably. Often used terms are discrimination [18], bias [10], [15]. In this chapter we want to define the basic concepts we use in the context of this thesis and differentiate those concepts from similar or related concepts.

At first we will define discrimination and its relation to social inequality as a social context of the analysis of the influence of social dimensions in a decision process. Afterwards we briefly review relevant laws in the European union (EU) and Germany regarding discrimination. A second part of our conceptual basis regards algorithms in decision processes, where we focus on algorithms in the context of machine learning and specifically the application of classification. Finally, we combine the two core concepts and derive a definition of discrimination in algorithm-assisted decision making which we will apply in the context of our case study regarding the influence of social dimensions on the SCHUFA-score. It is important to note that the definitions and differentiation we describe in the following, are not exhaustive and do not claim to be complete. This section is meant to be a pragmatic clarification of the context of this work.

2.1. Conceptual Framework of Discrimination

There is no universally accepted definition of the concept of discrimination as the view on what discriminating practices are continuously shifts [3]. This can also be illustrated through differing legal definitions of discrimination [55]. But there seem to be similar essential dimensions in most definitions about discrimination [3]. In this section we will concentrate on a more modern definition focused on the explanation from Scherr [55]. We first define discrimination as a concept that does not implicate a morally

wrongfulness and afterwards describe a moral context of the terminology to motivate the discussion about discrimination.

While the verb *discriminate* is also used in fields such as statistics, it does not resemble the concept discussed in this section. This use of the word is more similar to the meaning of the distinguishing of individual instances based on observed characteristics. It can generally be described as a neutral differentiation [50]. In this thesis we will use the term *distinguish* instead.

Discrimination in our context is the consequence of views or acts of individuals or groups as well as social structures and institutions, policies and conditions. It consists of acts, practices or policies that impose a disadvantage on persons by restricting socially important resources based on their membership in a social group that is not individually modifiable compared to members of a reference group [55], [3].

For a more concrete understanding of this definition we now describe its components in more detail. A core aspect of discrimination is the comparison of a disadvantageous treatment between a social group and a contextually appropriate comparison group. We summarize the set of social groups compared as a *social dimension*. Generally, there are no universally defined social dimensions which are relevant to discrimination. Nevertheless, they are not arbitrary but are constituents of historical and contemporary social inequalities and power relations [55]. Often social dimensions are social or physical attributes that are perceived as not acquirable or not modifiable by the affected person itself such as color, age or sex [54]. The assignment of a social group can in some cases be viewed as part of the discrimination itself [55]. A disadvantageous treatment of people based on different ranges of income is different from relying on the race of a person. While the income can be viewed as modifiable by a person, the race can not be modified and furthermore may be part of the social identity of that person.

The *reference group* in discrimination-relevant social dimensions often depends on majority-minority-relations and is not constituted objectively, but in permanent formation which dependent on social conditions [25]. Discrimination therefore can be viewed as the disadvantageous distinction of a social group from a societal dominant group which constitutes assumptions of normality in a society. In this sense discrimination works as a rationalization of disadvantage and inequality as well as asymmetric power relations between a social group compared to a dominant group that constitutes social normality and regulates the exclusion of the social group from some resources [55].

In relation to the social dimension is also the idea of *intersectionality*. It states that the combination of different social groups a person can be attributed to may lead to a disadvantage and is therefore relevant to discrimination. For example, women might

not be discriminated against in general but women of a distinct ethnicity [50].

Another important component of the definition of discrimination is the *asymmetrical differentiation* between groups of a social dimension. This generally concerns a durable disadvantage to a social group regarding caused by the exclusion of that group from socially important resources such as education [3]. If an exclusion from a resource is discrimination-relevant depends like the definition of a social dimension on the context of the given subject.

Our definition of discrimination does not state if a asymmetrical differentiation along a social dimension is illegitimate as such. The discussion about legitimacy of a discrimination is again context-dependent and subject of social as well as legal debate. While we will refer to a notion of discrimination that does not value the legitimacy of a discrimination, we want to describe a moral view on discrimination to motivate the relevancy of a discussion about discrimination. A moral view is relevant to the discussion to argue for interventions against discrimination or to leave a subject of discrimination as socially accepted. We will explicitly refer to a moral notion as *wrongful or righteous discrimination*.

The perception of a wrongfulness of a discrimination is the result of social negotiation processes and social and historical exclusion. This wrongfulness does not need to be similar in any area of life and is dynamically negotiated in legislative adjustments [50]. A possible argument for wrongful discrimination can be inferred from a modern view on *social justice* that guarantees same rights and opportunities for every person as stated in the Universal Declaration of Human Rights (UDHR) (1948). This idea implies privilege and disadvantageous as problematic and in principle requires justification where disadvantage of a social group is part of social normality [29]. The discussion about wrongfulness of a discrimination can therefore be associated with social vindication of social positions and privilege [55].

2.1.1. Types of Discrimination

So far, we ignored the actors and acts that lead to a discrimination. This subsection focuses on the part of the definition concerning *individuals, groups and institutions* as actors and *views, acts, practices and policies* as *acts* and defines *types of discrimination*, which are based on different actors and acts [3].

Direct discrimination can be considered the most obvious type of discrimination. It involves the *disadvantageous treatment* of a social group by a direct use of the group membership or an intentional use of a proxy variable that is correlated with the group

membership. On an individual level this form of discrimination becomes more complex by differentiating between an *intentional* direct discrimination and an *unintentional* one, where the different treatment of the social group is intentional, but the caused disadvantage is not. Unintentional discrimination is led by some prejudice or bias against a certain social group, even when the bias does not involve an intention to treat the group disadvantageously [3]. Direct discrimination concerns the process of using a social group membership or a correlating variable to treat that group differently, which leads to a disadvantageous treatment [50].

When a decision maker has no intention to disadvantage the member of a group and no prejudice or bias motivating an act that leads to a disadvantage on the member of a social group, it can be described as *indirect discrimination*. Hence, the group membership is not intentionally nor unintentionally used directly in the disadvantaging treatment [3]. This considers the unequal distribution of a resource that leads to a disproportionate disadvantage for a social group relative to a reference group. This often is a indirect consequence of historical inequalities [55]. The association between social inequality and discrimination will be discussed later in the next subsection. Indirect forms of discrimination are harder to pin down, because it is not agreed upon which disproportionately worse effects on a certain group actually counts as indirect discrimination. For example, we could use the income of a person to decide for a specific treatment. This might impose a disproportionate disadvantage on the social group of women in comparison to men, because of a statistical correlation to the income distribution in a society according to which women in average earn less income than men. This will depend on the context of the potential discrimination. Furthermore, it is not clear if such a discrimination can be classify as illegitimate and depends on the legal negotiation in a society [50].

Until now we described the types of direct and indirect discrimination, which concentrate on the the process and the outcome of a discriminating act. These types do not define if a discrimination is is caused by individuals or groups. *Organizational discrimination* concerns a collective of persons participating in a decision process and a distribution of a resource [3]. Here a set of social processes through which organisational decision making, either implicitly or explicitly, results in a social group being disproportionately disadvantaged.

This can be also viewed on an institutional level called *institutional* or *structural discrimination*. This type of discrimination should be distinguished from the organizational view, because it does not only concern a collective of persons, but also the rules constituted by a collective. These rules form social structures and institutional practices

that are based on assumed normality. The resulting social norms are disadvantageous for a social group such as family norms [55]. There are accepted policies leading to institutional discrimination. Persons seeking for asylum for instance are excluded from the right to vote [25].

2.1.2. Social Inequalities

In this subsection we want to emphasize the connection between social inequality and discrimination as discussed in [54] to later use this connection to identify discrimination-relevant social dimensions and formulate hypotheses in the context of our case study regarding the influence of social dimensions on the SCHUFA score.

Social inequality describes social structures differentiated in terms of a hierarchy that privileges or disadvantages individuals or groups durably. A *social structure* can be defined as the interdependencies of a multi-dimensional segmentation of the society into different groups. This segmentation is based on relevant social features as well as relative and durable social relations between these groups [22]. A relevant *social feature* is an influential factor like occupation, qualification, gender or ethnicity that influences the social behavior of these constructed groups as well as their position in societal sections, institutions and social networks.

A hierarchical form of social structure is defined in the research of social inequality. It refers to regularly *unequal distributed resources*. Such resource needs to improve a persons living conditions. *Living conditions* are conditions of live and actions that constitute in external circumstances, which cannot be influenced in a short term. These circumstances in turn act as the possibility of a person to gain prevalent objectives in a society [30].

Manifestations of social inequality are multifaceted, which is why researchers categorize them into *dimensions of social inequality*. Basic dimensions identified are material wealth, power, prestige and education. These dimensions are in turn described and quantified by concrete features called *indicators*. Unequal distributions of such indicators implicate a manifestation of social inequality [30].

While the differentiation of dimensions of social inequality is a description of structures of social inequality, so called *determinants* go beyond the description. Determinants of social inequality create groups of persons with a common social feature that introduces or locks social opportunities. Determinants do not need to cause different opportunities but can be mediated by other circumstances the determinant is correlated with [30].

Theories are introduced to define hypothetical causes or serve as an explanation of

the emergence of hierarchical structures of social inequality. They enable the empirical investigation of hypotheses [22]. Traditional theories of social inequality are based on concepts such as class. Determinants such as employment are arranged vertically implying an inferiority and superiority regarding some resource such as wealth [30].

While vertical determinants follow a naturally vertical order of resource distribution often accepted by society, horizontal determinants indicate a unequally distributed resource without being necessarily acceptance socially such as sex, age, residential area or ethnicity. Often horizontal determinants do not follow a logic of effort legitimizing an inequality. It is common in social inequality research to use discrimination as an explanatory extension of types of disadvantages that are empirically not sufficiently explained by vertical determinants used to describe the typical conceptual classes of inequality. These are mainly the horizontal determinants [54]. Hence, socio-economic inequalities and discriminatory distinctions are entangled concepts because both are based on building social hierarchy that impact living conditions and opportunities. Furthermore, horizontal determinants of unequally distributed resources may lead to correlations with a resource subject to decision making process. This can lead to a disadvantageous treatment or outcome along a social dimension related to the horizontal determinant leading to a discrimination. This interdependence requires the contexts of the inequality and the discrimination to be related.

Note that there is a an important difference between the discourses of inequality and discrimination research. In inequality research it focuses on socio-economic inequalities between class concepts, while the center of debate in discrimination research concerns attributed group categories which are not modifiable by a person of that group. Since the concept of discrimination is essentially defined by a practice of disadvantageous distinction between social groups, it needs to be distinguished from disadvantages caused by inheritance of wealth and socio-economic positions [54]. But it can be argued that theory of inequality is important to determine dimensions that are relevant for discrimination as well as the analysis of the causes, types and consequences of discrimination. Regarding indirect, organizational and structural types of discrimination socio-economic inequalities can be viewed as distinct discrimination-relevant factors [54].

We will use the outlined connection as a ground for building hypotheses of social dimensions that can be considered relevant to discrimination. We furthermore argue that the observation of an influence of chosen social dimensions on a score of the SCHUFA in the given data set based on empirical results of social inequality research indicates at least an indirect form of discrimination.

2.1.3. Discrimination in European and German Law

In modern societies there is at least in legal texts consensus that any person should have equal opportunities and equal rights. Inequality between persons is only legitimate in terms of individual effort and commitment. Equal opportunities are also a principle of social justice. Hence, privilege and disadvantage is generally problematic and calls for justification [55]. Already in the United States Declaration of Independence (1776) and the french Declaration of the Rights of Man and of the Citizen (1789) contain the principle of liberty and equality of all human beings, and can be interpreted as an indirect prohibition of discrimination. As a reaction to the nazi regime and the Holocaust as well as cases of defiance of human rights like slavery and racism anti-discrimination law was made explicit by the Universal Declaration of Human Rights (UDHR) [40] (1948) [29]. It explicitly states that there should be made no distinction based on the race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. The European Convention on Human Rights (ECHR) [38] (1953) on the other hand prohibits discrimination based on sex, race, colour, language, religion, political or other opinions, national or social origin, association with a national minority, property, birth or other status. In the German Allgemeinen Gleichstellungsgesetz (AGG) [21] (2006) it is explicitly illegal to disadvantage a person based on race or ethnicity, sex, religion or ideology, disability, age or sexual identity .

Comparing these different examples of specific categories potentially affected by discrimination can be seen as an example of the societal learning process regarding discrimination. While the Universal Declaration of Human Rights did not mention the age, sexual orientation or disability the AGG did and the ECHR has the "other status" category which allows the extention of the article 14 to other ground such as sexual orientation. Hence, the law illustrates the ambiguity of the term discrimination as context-dependent and depending on historical behavior of societies as well as civil movements which fight for specific group-dependent rights. Furthermore, controversies and criticism regarding for instance the explicit exception of the citizenship from the prohibition of discrimination shows once more the shift of categories to legitimately discriminate upon [54]. But that does not mean that discrimination is an arbitrary phenomenon. It happens on grounds of a finite number of nameable structures, ideologies and prejudices [55].

Another legal device in ECHR and the AGG that is open for interpretation and is context-dependent concerns indirect discrimination which also includes structural forms

of discrimination. The AGG describes indirect disadvantages as seemingly neutral rules, criteria or methods that disadvantage in terms of one of the mentioned group memberships in comparison to other persons. The ECHR speaks of policies with disproportionate effects as potential discriminatory even if that is not the aim of the policies. Both legal documents state the exceptions that legitimate disproportionate effects on protected groups if it pursues a legitimate aim [3]. But which pursue and proportionality between means and aim legitimates a disproportional effect on a specific group is again context specific and open for interpretation and arguments.

We also want to shortly mention the legal terms of disparate treatment and disparate impact from US law, since it is used as a theoretical ground for much of the work in the scientific field of discrimination-aware data mining which will be later discussed in more detail because it is related to this thesis [58]. Under Title VII of the civil rights act of 1964 *disparate treatment* is the unequal treatment of persons on basis of membership in a protected group. This is similar to direct discrimination. *Disparate impact* on the other hand is related to indirect discrimination and involves policies or practices that are seemingly neutral but have a disproportionately adverse impact on protected groups. Disparate impact does not regard the intent of such practice or policy but requires a justification for that impact. The proportionality of disadvantageous effect is weighted against the justification to decide if it is discriminatory or not [5].

Lastly, we want to briefly discuss the German law specifically regarding scoring and its relation to anti-discrimination law, because it shows a connection between scoring as a special case of algorithm-assisted decision making and discrimination. While the ECHR and AGG regard scoring in its aim to help decision making, the data protection rights, namely General Data Protection Regulation (GDPR) and BDSG-neu [32] in part directly refer to solvency-scoring. The GDPR applies EU-wide and the BDSG-neu applies complementary as national German law [50]. Both involve special categories of personal data which are considered particularly relevant for protection. Such categories are partly similar to the protected categories mentioned by the ECHR and AGG.

The BDSG-neu also involves the §31 regarding scoring and solvency-reports called Schutz des Wirtschaftsverkehrs bei Scoring und Bonitätsauskünften. It is explicitly stated that scoring is only legally used if it satisfies data protection law. Furthermore the computation of a probability value must be done by using scientifically approved mathematical-statistical methods. The computation must not solely rely on address data or the affected person must be informed about that.

The computation of a probability value to aid decision making like in the case of scoring often involve an algorithm producing such value. Hence, we want to define

algorithms in the context of decision making in the next section.

2.2. Algorithm-assisted Decision Making

In this section we want to describe the role of algorithms in decision making to derive a conceptual basis of discrimination in decision process where algorithms are involved. Algorithm-assisted decision making generally involves the distribution of a resource based upon a decision process where an algorithm or a set of algorithms is involved to assist a person or collective of persons to make a decision. In this section we want to define the main components of this context of decision making that we want to focus on. We describe the application of machine learning algorithms with a focus on classification problems and its use in decision processes.

2.2.1. Machine Learning and Classification

An *algorithm* can be defined as a systematic procedure that consists of logical operations and which produces the solution to a problem in a finite number of ordered steps [5], [48]. Applications of algorithms in modern decision making processes are executed by a computer with continuously increasing resources. This enables an increasing complexity of the tasks that should be solved by algorithms. A *computer program* can be described as an unambiguous, ordered sequence of computational instructions necessary to achieve such a solution [48].

The algorithms used in many modern applications where decisions are influenced by algorithms can be referred to as *learning algorithms*. These kind of algorithms describe algorithms in the field of *machine learning*, a field which grew out of the scientific discipline *artificial intelligence* and draws from computer science [4] as well as statistics. Such algorithms can be described as techniques that allow a computer to acquire or improve its ability to perform a task by automatically extracting knowledge from data. A learning algorithm typically consists of a *mathematical model*, that can be viewed of as a mapping function from an input-space to an output-space. An *objective function* optimizes the parameters of the model given observed data and mathematical constraints. The model and the objective function need to be specified and come with assumptions about the underlying function that is aimed to be modeled. The learning algorithm defines how the objective funtion optimizes the parameters of the model. The process of optimization is also called training or learning.

While there is a variety of learning algorithms, we will concentrate on a specific class called *supervised learning*. The distinction between learning algorithms is made by the

type of data and the context of the algorithm. In supervised learning we have some input data X and output data Y, also called labels. The goal is to find a mapping function f that maps observed training data $X*$ to $Y*$. A test data set which was not used in the optimization step should now be able to be mapped with the input $X**$ to the label $Y**$. In supervised learning there is also distinction made between *regression* and *classification* which depends on the type of label data Y. In Regression the label is of a continuous type, while in classification labels are finite and discrete. Labels in classification are also called classes. In our discussion we will focus on classification, because the scoring procedure of our case study is based on a classification problem.

We now specifically provide a more formal definition of classification tasks with with one-dimensional label y. We will sometimes also refer to y as the target variable. We generally define the mapping function as $f : X \mapsto y$. The general model notation we are going to use is:

$$y = f(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

The mapping function f represents the model used to map the input X to the target y. X is a matrix consisting of n observations and d dimensions, which we also refer to as features.

To illustrate the basic ideas we will formally introduce logistic regression which is also the modeling approach used for the SCHUFA-score in our case study [2], which we will revisit in a later section.

Logistic regression can be viewed as a linear regression with target variable $y \in \{0, 1\}$ where a function is applied to the result, mapping the outcome into a continuous space $\hat{y} \in [0, 1]$ using the logistic function. This in turn can be interpreted as a probability. To obtain a mapping to model the target as a probability for the event $y = 1$ the logistic function is applied to a linear combination of the features X and model parameters θ that should be optimized, leading to:

$$p(y = 1|X) = \frac{exp(X\theta)}{1 + exp(X\theta)} \tag{2}$$

Formulating the log odds of the target y illustrates that the logistic regression model is a linear model in the log odds:

$$\log \frac{p(y = 1|X)}{1 - p(y = 1|X)} = X\theta \tag{3}$$

Hence, we can interpret the linear relationship of the features to the target in terms of the log of its odds [27] The objective function governs the optimization of the the weights θ to map the training input data X_* as good as possible to the training target variable y_* such that the outcome is close to 1 if we observe $y = 1$ and the outcome should be close to 0 if we observe $y = 0$. In terms of optimization we want to maximize the likelihood of the model given the data we observed.

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \tag{4}$$

In logistic regression this can be achieved by applying the likelihood principle, which works as the objective function:

$$L(\theta) = \prod_{i=1}^{n} p(y_i|\theta; x_i) \tag{5}$$

The core of the learning algorithm is the optimization step, where the optimal parameter setting $\hat{\theta}$ needs to be found. Generally we want to find the global minimum of the objective function. In the case of logistic regression we can use iterative optimization method [27]. Such numeric solutions update the parameters θ according to the objective function $L(\theta)$ and count as solved when the difference of the update θ_{t+1} at an iteration step $t + 1$ and the value θ_t of the past iteration t is smaller than a given threshold.

We can conclude from this illustrative example that a learning algorithm follows a clear mathematical formulation and is generally a deterministic algorithm that finds an optimal solution regarding a defined setting using statistical learning and optimization methods.

2.2.2. Learning Algorithms in a Decision Process

A *decision process* can be described as a collection of sequential or parallel tasks whose outcomes are combined to return a decision. In a algorithm-assisted decision process each task may involve an algorithm or natural person that decides about an outcome of a task [41]. An algorithm applied to a decision process can again be viewed as a collection of tasks. The process can consist of multiple tasks that are partly or fully solved by an algorithm. Between those distinct tasks, there might be natural persons that use the outcome of a task to act upon it.

A learning algorithm can also be described as a collection of tasks where the output is not based on explicit rules but implicitly depends on the setting chosen setting

of the learning algorithm. This setting consists of multiple adjustable components. The components mainly concerns the construction of the input data X and y, the choice of the modeling approach, the definition of the objective function and the actual learning algorithm optimizing the model. Those adjustments can be viewed as normative decision that imply assumptions and views a person has in the context of the application [14]. A person define and execute the resulting learning algorithm and uses the resulting trained model.

Hence, algorithm-assisted decision making involves a collection of tasks that lead to a decision about some action. A subset of tasks involves an automatic execution of rules that are explicitly or implicitly stated and defined by a natural person. The automatically produced outcome of a learning algorithm is interpreted and used to contribute to a decision upon some action by a natural person. Even in the case of a automatic decision the process that leads to that decision is constructed by a natural person and the evaluation of the decision is executed and interpreted by a natural person. The emphasis of natural persons leads to the term algorithm-*assisted* decision making.

2.3. Algorithm-assisted Decisions and Discrimination

As we have illustrated in the last subsection, an algorithm or specifically learning algorithm itself is a sequence of tasks which is based on mathematical formulation to find a mapping function. This does not involve discrimination on its own. Though, the application context is chosen by and the setting of the algorithm is adjusted by a person. This factors on the other hand may be relevant to the subject of discrimination depending on some conditions we want to explain in this subsection.

The *context of the application* needs to *affect natural persons*. This might be a direct task of modeling a concept about natural persons for example being a risk in some sense that implies a decision-relevant output to be interpreted [50] or a task that may not directly be linked to the inference of a decision but nevertheless leads to restrictions of resources if the process is flawed for specific groups such as in some examples of facial recognition concerning natural persons [35]. Furthermore, the application is situated in a decision process which aims to regulate the distribution of a socially important resource. Due to the restriction of resources a decision process becomes relevant to discrimination, because the systematic negative influence of a social group membership on a decision may lead to a durable disadvantage for that group.

An algorithm that influences or leads to a discriminatory decision is subject to a

setting which biases the resulting optimized model along a social dimension. This is generally a technical issue from the view of the adjustment of the algorithms setting. Hence, we collectively call adjustments which lead to such bias along a social dimension a *discrimination bias*. Note that even when the bias is caused by the structure of the data it still is a design decision to not prevent the bias to be incorporated into a model. The concrete causes for such biases will be further explained in the context of our case study in section 3.2.2. We use the term *bias* in its statistical sense of a systematic deviation to emphasize the technical background of discrimination bias similar to the use of bias in different sources[14]. In a statistical sense bias describes a systematic distortion of a statistical method or result due to a factor not allowed for its derivation [13]. There is a sociological and psychological notion of bias that we not refer to. It involves a persons distorted belief about some group feature causing him to favor a certain group. Social bias is related to the term *prejudice* [3].

We furthermore choose the term *discrimination* to emphasize the social context of the application the biased algorithm is situated in. We distinguish discrimination biases between direct and indirect discrimination bias depending on the intention of the person or collective of persons adjusting the algorithm. Intentionally adjusting the setting of an algorithm to treat a social group differently is *direct discrimination bias*. For instance, using a social dimension as an input feature leads to direct discrimination bias. Adjusting a algorithms setting so a bias along a social dimension occures unintentionally is called indirect discrimination bias. The use of an input variable which is correlated with a social dimension will lead to an indirect discrimination bias [50].

The output of such biased algorithm in a decision process may be systematically lower for a social group in a social dimension. Depending on the usage and interpretation of the output to derive a decision, the act upon the decision to regulate a resource may then lead to a durable disadvantage for a social group and therefore discriminating that group. The actual discrimination where a biased algorithm is involved in will is described by the term *discrimination in algorithm-assisted decisions*. Note that the discrimination in such context is not proven by identifying a discrimination biased model. The bias might be practically ignorable or mitigated by accounting for it when interpreting the output of the algorithm. Hence, a discrimination bias in a model only potentially follows a discriminating decision. It is also context-dependent and open for social debate and legal discussion if a discrimination bias is legitimate or not just like the discrimination itself.

The restricted resource may also be viewed from an inequality theoretic perspective on discrimination as described in 2.1.2. Thus, if a resource can be determined as

relevant for a dimension of social inequality leading to different opportunities or living conditions of groups of persons, a horizontal group dimension becomes relevant as a discriminatory dimensions as it lead to a restriction of such resource and is not modifiable by a person on its own.

In the context of discrimination biases in learning algorithms we want to emphasize the connection between inequality and discrimination. An unequally distributed dimension of inequality along a social dimension that is related to the resource which is regulated in a decision process may lead to a correlation that can be incorporated into a learned model. The biased model may in turn lead to a discrimination, hence retaining the inequality. Note that statistical correlations can also be used to legitimate a direct discrimination and can be viewed as assumptions that may constitute normality [54].

We also want to note the important role of organizational and institutional discrimination. The context of discrimination in algorithm-assisted decision processes can be often associated with such types, because companies and institutions are increasingly using outcomes of algorithms to automate some tasks in there decision processes.

The following chapters of the thesis will focus on the analysis of indirect discrimination biases in the context of a case study regarding the SCHUFA-score in a data set that is available to us. We will look at the influence of social groups in discrimination-relevant social dimensions that are durably disadvantaged in the distribution of wealth according to empirical results of social inequality research in Germany.

3. SCHUFA Solvency-Scoring and Discrimination Bias

In this section we introduce the background of our case study and discuss the relevancy of its context to the subject of discrimination bias. We start by defining solvency scoring and its use as a commercial service of the SCHUFA credit bureau. Afterwards we describe the use of learning algorithms in solvency scoring and the potential of discrimination bias in the given context. At the end we briefly describe the public discussion regarding transperancy in services of solvency-scoring and the connection to the openSCHUFA-project which made the data basis of this case study available.

3.1. Solvency-Scoring and SCHUFA Credit Bureau

The *SCHUFA credit bureau* (Schutzgemeinschaft für allgemeine Kreditsicherung - translated: General Credit Protection Agency) describes itself as “Germany’s leading credit bureau” that is a “source of information for corporate and private customers,

holding credit rating information about 67.7 million persons and 6 million companies" [2].

More than 9500 corporate customers use the services and rely on the information provided by the SCHUFA. The customers include banks, savings banks, trading firms and telecommunications companies. They are called *contractual partners*, because the SCHUFA builds with them a protection association. In this association the SCHUFA acts as an establishment that registers and stores the payment histories of companies and provides them to companies that are part of the protection association. Hence, the customers are supplier and recipients of data at the same time. This concept is expressed as the *principle of reciprocity* [2]. There are different notification proceedings that define different subsets of features out of the whole pool of features available to the SCHUFA company. The proceeding regulates the features provided to contractual partner and is chosen by payment services, risks and sector of the contractual partner. The principle of reciprocity makes the contractual partners the most important source of information to the SCHUFA. Further sources are for example official proclamations, defaulter books of district courts and a consumer himself. Most of the stored data is related to a natural person [24].

The data basis of the SCHUFA is unique for financial data of consumers in Germany because of the information exchange with its contractual partners [2]. But there are lacks of data because of many companies that are not part of the protection association. The data basis can be separated in three distinct categories: personal reference data, positive features and negative features. To *personal reference data* count non-financial information about a consumer like the name, address, date of birth or bank account number. This data is used to associate information received, with a natural person. *Positive features* regard contract data and can be separated in request features which concern requests of contractual partners about a consumer and report features which concern contract closings such as opening an account, handing out a credit card or information about a credit period. Report features are also relevant for the creditworthiness of a consumer. *Negative features* on the other hand are the result of violation of a contract. These features can be separated in soft negative features which concern judgments of contractual partners themselves and hard negative features which concern judicial decisions or legal requirements and do not depend on the judgment of contractual partners. The data is based on the aforementioned principle of reciprocal with the contractual partners and among others official proclamations. While partners use these information, some are also used for predicting the creditworthiness of consumers. The predicted value is also called SCHUFA-Score [24].

The company claims to report information 450.000 times per day and contributes in doing so to many business transactions. The SCHUFA does not directly perform any decisions, but delivers credit-relevant information and a solvency-score, sometimes generally referred to as the SCHUFA-Score [2]. While the SCHUFA bureau was founded in 1927, their first scoring-procedure was introduced in 1995, called the ASS (Auskunft-Scoring-Service - Report-Scoring-Service). With this service of the SCHUFA-Consumer Line [24] the SCHUFA uses a statistical-mathematical procedure to predict future behavior of credit borrowers based on past data about groups of consumers a single borrower is associated to [6].

Scoring in general can be defined as the assignment of a numeric value to a natural person for the purpose of predicting or controlling the persons behavior. This is called a score. The mapping of data to a numeric value is often obtained on a data basis by a computer program [50]. We focus on the solvency-score, because it is the subject of our case study. *Solvency-scores* are mostly assigned by commercial credit agencies to consumers with the aim of predicting their financial behavior. The prediction is often interpreted as a probability about the persons creditworthiness or the probability of default [50]. Its mapping to a consumer is governed by some defined criteria and features about the consumer as well as other modeling assumptions. The basic premise is that the behavior of groups of persons in the past conforms with future behavior and that the behavior of persons with similar features is also similar [24].

In the case of the SCHUFA credit bureau the solvency-score is not just one numeric number but includes score-values for a basis score and 12 different sectors determined by the types of contractual partners [2]. A score is mapped to a probability distribution and a risk class, illustrating a semantic interpretation of the numeric value. Each score is based on a different subset of features depending on statistical relevance for a sector and is the outcome of different modeled weightings also called score cards [24]. The features are a selected subset of the described data basis and include positive and negative features. While soft negative features depend on available positive features about a consumer, hard negative features already determine a bad rating [26]. The concrete features used by the SCHUFA to compute the scores are not publicly known, though there are some sources that provide a rough classification of used features [2]. Note that the SCHUFA wants to predict a statistical probability of default even when there is only few credit-relevant information [6]. This aim might be the cause for the SCHUFA to use geo-information such as the address data, which the SCHUFA claims to use only when there are no other information about a consumer [2]. Over time the scores where improved by the SCHUFA leading to 3 different versions of the sector

scores and some specific differentiation in the fields of sectors [1]. Old versions are partly still used by different contractual partners.

The solvency-score of the SCHUFA aims to avoid insufficiencies of traditional checks of creditworthiness because of lacking knowledge of a broader overview of financial relations. With a mutual data basis a standardized and efficient credit loaning is aimed to be enabled [24]. A motivational factor of the use of solvency-scores is that its application leads to a reduced amount of loan defaults, they lower transaction costs and are therefore important for the efficiency of finance markets. Furthermore, it is claimed that solvency-scores reduce the information asymmetry between borrower and lender and credit rationing is prevented because credit relevant information about borrowers is available [50].

The solvency score is thus used for decisions about the restriction of consumer-relevant resources, which are in some cases socially important. Since the SCHUFA seems to have information about more then three quarters of the German population, the score can be viewed as relevant for the German society. Features and scores provided by the SCHUFA are one of the most important components in the decision process of many companies working with consumers [26]. Hence, the SCHUFA-solvency-score is a feature that influences the access to resources for most consumers in Germany. Therefore, it can be claimed that the services of the SCHUFA can have an influence on the decision process of organizations and can potentially lead to disadvantageous treatment of social groups in German society. Disadvantageous treatment may lead to an unequal distribution of resources along a social dimension. This social context leads the application of the SCHUFA-solvency-score to be relevant for the subject of discrimination.

Furthermore, scoring procedures themselves are inherently at risk of leading to acts of discrimination, since they automatically produce a score mapped to a natural person that aims to aid in a decision process [50]. Therefore, the algorithm producing the score may lead to a discrimination bias.

3.2. Classification in Solvency-Scoring

In the following subsection we briefly discuss solvency-scoring generally as a classification task solved by a learning algorithm. We elaborate this methodological approach to illustrate that the context of solvency-scoring is relevant to our definition of discrimination bias. We then describe potential causes of discrimination bias to motivate the analysis of such biases in the context of solvency-scoring.

3.2.1. Computing a Solvency-Score

We now view a solvency-score as the outcome of a mapping function that aims to model the probability that a natural person does not default [50]. This mapping is based on a variety of modeling assumptions as well as an objective function that we use to optimize a mathematical model, given an observed data set.

The statistical model used most often in the field of scoring is the logistic regression [17]. In German law it is required to use statistical-mathematical methods to compute a score considering some measure of quality. Though, there is no legal regulation about the actual quality of the model [50]. Since the SCHUFA claims to use logistic regression as the statistical basis of their solvency-scoring method [2], we will briefly explain the modeling process in the context of logistic regression in general.

In scoring the input data X consists of n observations, each represents a natural persons with d observed attributes according to the used features. The label data is represented by the discrete target variable y. In the binary case y indicates if a person defaults or not. The aim of the logistic regression is to find a mapping function $f : X \mapsto y$ as described in 2.2.2 using a objective function and a defined optimization step. The resulting model consists of the optimized parameters θ. These can be represented as *score cards*, which assign a score value to a combination of feature attributes by weighting the attributes with the optimized parameters θ [17]. Predicting the probability of default for a new observed set of persons $X**$ the outcome $y**$ is obtained on the basis of the optimized mapping function. After mapping each case to a score-value y, the outcome may need to be transformed in multiple ways before it can be interpreted as a representative probability. Calibration is used transform the outcome to an expected probability distribution. Note that logistic regression is already well-calibrated. The probability distribution can furthermore be binned into ordinal rating levels. This process is also done for the sake interpretability and to retain the score-odds relationship between model versions. To transform the score value to a decision for example to give a person a specific resource the decision maker can simply define a threshold over which the person in question may not obtain that resource. This threshold is called a cut-off value [17].

The continuous value can be used as a feature in another algorithm-aided decision process. For example, a bank might use the solvency-score as one component in a separate scoring-system to decide to hand out a loan to a person [6]. It can be also directly used to assist a decision based on a defined cut-off value transforming the continuous outcome to a binary value.

3.2.2. Possible Causes of Discrimination Bias

While we have so far focused on the methodology of scoring, we now want to concentrate on the adjustment of the setting of the learning algorithm. We describe multiple normative design decisions that may cause a discrimination bias in the optimized model. For a more systematic take on this discussion we rely partly on the taxonomy of causes provided by O'Neil et al. [41] and Dobbe et al. [14] for a bias-related terminology. In each part of the taxonomy we turn back to the context of our case study to point out the relevancy of the discussion. We introduce biases concerning the specification of a model and learning algorithm, namely data bias, modeling bias, target bias and optimization bias. Then we explain biases in the implementation of the learning algorithm including emergent bias, data source bias and data quality bias.

Data bias: When a variable used to predict a target and is at the same time correlated with a social dimension a indirect discrimination bias may occur. Here, social inequalities and historical discrimination is potentially mirrored by the data. Such biases may be the most societal controversial ones, because they lead to a more accurate model but at the same time may strengthen inequality and disadvantage for a social group. The worst case would be a scenario where the data-generating distribution itself has discriminatory bias against some social group, hence even a perfect Bayes optimal classifier would incorporate this bias [41]. In solvency-scoring a unequal income distribution along the social dimension sex may lead to such a bias.

Sample bias: Indirect discrimination bias does not need to be the result of historical bias, but can also be caused by sample bias. In that case, the training data does not represent a social group well. In credit scoring it could be the case that disproportionately many male consumers in the training data set default, even though this may not reflect the data generating distribution. Hence, the learning algorithm may use the correlation in the data set to predict the probability of default.

Modeling bias: Modeling bias may occur in a flawed specification of the model. In terms of discrimination bias this for instance is the case when a social dimension is directly used to model a relationship. This also regards the intentional use of a proxy variable with no predictive power to implicitly model the relationship with a social dimension. One popular example in credit scoring is red lining [50], where the address data is used to disadvantage whole geographic areas or other social dimensions correlated with geographic areas.

Target bias: Label bias can be caused by the process of manually assigning class

labels as the outcome we want to model. This process is often not definite about the label assignment. A controversial example of label bias are predictive policing tools where historical data is used to predict locations in which crime is most likely to occur. Labeling violent and nuisance crimes alike will potentially drive a disproportionately amount of attention to poor neighborhoods [41]. In credit scoring the definition of default is also not clear but is a function of the way the credit industry has constructed the credit issuing and repayment system [5].

Optimization bias: Another normative decision that may introduce bias is the optimization function and optimization constraints. Depending on the metric the model is optimized for, it will capture different magnitudes of influence of predictors to model a target. In the discussion about fairness criteria, where specific constraints or data manipulations are used to match the rates of errors between social groups according to some fairness-metric is a concrete example. We will discuss this topic in a later section 4.1.

Emergent bias: A issue of the implemented model regards the feedback loop of a decision system. The person or organization in charge needs to monitor and verify the validity of the systems output. This is an especially important step if the system influences decisions regarding natural persons and might distort the data-generating distribution it is situated in. This can lead to input data of future learning iterations influenced by the current model which leads to self-fulfilling prophecies [41]. The environment in which the system is situated in and the kind of interaction between the system and its environment should be taken into account when designing a decision system. Continual societal changes can lead to reducing statistical correlations which should also be reflected by the model. Bias following from such situation is also called emergent bias and is important to account for in a context of use by real users as a result of a change in societal knowledge, user population, or cultural values [14].

Source bias: One aspect we would classify as a process issue concerns the quality of the data and the process of collecting the data. It involves the sources, methods of collecting and the association of data to a natural person. The SCHUFA for example has a data basis that includes features that are collected by exchange with other companies on a basis of trust. Hence, there might be a bias in the data caused by sources, providing the information. In the case of the SCHUFA that should be particularly noted for some of the soft negative features that depend on the judgment of the contractual partners.

Error bias: Regarding the association of data to a natural person there can emerge errors of mismatching. In the case of the SCHUFA mismatching and a limited access of

consumers to their data led to criticism [16]. The possible lack of personal information for instance may lead to ambiguous personal information, which in turn results in higher rates of mismatch. A mismatch can for example lead to a newly created data set for that person without any information about his solvency which might lead to bad approximatory features such as geo-data that can lead to a bad score [50]. This might cause a disproportionate amount of flawed scores correlated with specific groups.

As we have seen, there are different forms of biases that can lead to discrimination bias in the resulting model of a statistical scoring system. These biases partly intersect and imply different potential threads such as preserving discriminatory practices and inequalities for the sake of accuracy or inversely producing disproportionately many errors for specific social groups. Knowing these causes and reflect upon them offers a variety of possible solutions to at least reduce discrimination biases. But the decision regarding what kind of discrimination bias should be mitigated and how this should be done is not part of this thesis.

3.3. Transparency in Scoring and the openSCHUFA-Project

In the case of solvency-scoring *transparency* refers to the disclosure of information about a natural person that is affected by the scoring. Transparency is an important right, because a consumer needs to know what and how his personal information is used to be able to claim his rights. This concerns for example the right of correction of personal information used to obtain a score-value associated with the consumer [50]. This interest of transparency of information is contrasted with the interest of the company that sells the score as a product. A company might has an interest in confidentiality concerning the concrete process to compute the score. On one hand, this might be the case because of have an advantage over competing scoring-services. On the other hand some information might loose its predictive power, when a consumer knows about its use and changes her behavior to artificially improve the score. The *Score-gaming* is possible when a scoring procedure uses a variable, also called a proxy, to account for some construct that cannot be measured directly. For example, in solvency-scoring the score value might increase with the amount of credit cards of consumer. Hence, the consumer might purchase credit cards only to improve her score. But one might argue that such proxies should not be an argument for a lack of transparency and that proxies should be avoided in the first place [50]. Scoring reduces the complexity of qualitative judgement [50] and provides a quantitative procedure in a decision process. But the earlier discussed adjustable factors and normative design

decisions of a learning algorithm that potentially cause discrimination bias are often not appropriately discussed by the society affected by them. To initiate a constructive public discussion information about some aspects of the scoring procedure are necessary [50].

This is also the case regarding the score provided by the SCHUFA credit bureau, even though the solvency-score affects most of German society. The SCHUFA used to provide some information in the *SCHUFA-report*, a consumer can legally request regarding the assessments of five very roughly classified different data types and the sector-scores calculated from the data at the time of the creating the report. Though, since the EU-DSGVO came into force, the report only includes the basis score as a percentage value, which is only as a value of orientation [56]. Also the information stored by the SCHUFA regarding requested scores and their values by contractual partners and other information associated with a natural person are only freely available once a year. The rest of the time the information is only available as part of a product the SCHUFA sells to consumers [2].

As a reaction to the lack of transparency of credit bureaus the initiative *AlgorithmWatch* and *Open Knowledge Foundation* (OKFN) initiated the openSCHUFA project. The project used crowd-sourcing to obtain a set of SCHUFA-reports of persons in Germany scored by the SCHUFA with the aim to illuminate the opaque solvency-scoring of the SCHUFA credit bureau. Note that the sampling procedure was started to execute before the SCHUFA changed its policy about the information provided by a SCHUFA-report. The obtained data set was then provided to selected organizations and persons such as the editors of Spiegel Online and BR [16]. This data set is also the basis of the case study of this thesis where we aim to analyze disproportion in selected sector-score of the SCHUFA in terms of specific social groups to interpret those observed results in the context of solvency-scoring and the defined discrimination bias in the data set.

4. Related Work

Before we continue with our concrete methodology used in this thesis, we want to briefly discuss the scientific body that concerns similar issues of discrimination bias in learning algorithms and locate the work of the thesis in that scientific field. We start by describing different definitions of fairness in machine learning and the connection to discrimination. We then briefly explain methodological attempts to prevent

unfair behavior and the impossibility to fulfill multiple fairness-criteria. We then turn to discrimination discovery as a field of discrimination-aware data mining and its methodological approaches important to our analysis.

Most of the scientific literature regarding discrimination bias in learning algorithms is summarized as *fairness in machine learning*. The concept of *fairness* is not equal to the concept of discrimination as used in this thesis. In contrast to discrimination, fairness applies to the differential treatment of groups as well as individuals and is inherently valued morally [46], [28]. The legal terms often referred to as an argument for fairness in machine learning are *disparate treatment* and *disparate impact* as described in 2.1.3. Fairness regarding social dimensions may be also a morally valued discrimination depending on the social context and social dimension involved. In the field of fairness in machine learning group dimensions are often defined according to legally protected social group memberships. The group dimension is in many scientific work of the field exchangeable.

4.1. Attempts to Define and Maintain Fairness in Machine Learning

Most of the proposed fairness criteria in the literature regard classification tasks which fits our context of solvency-scoring. Criteria summarized as *group fairness* are especially interesting regarding discrimination bias, because they can also be viewed as discrimination-relevant, depending on the social context of the application. In the following description of fairness criteria we will denote a model as $f(x)$, non-protected variables as X and protected variables as A.

Individual Fairness: [15]

$$D(M_b, M_c) \leq d(b, c) \tag{6}$$

The notion of individual fairness mathematically formalizes the idea of treating similar individuals similarly. This is expressed by the Lipschitz-condition. Two individuals $b,\ c \in X$ of a set of individuals X are mapped to a set of outcomes y by a classifier described by the mapping from individuals to probability distributions over outcomes $M : X \mapsto \Delta(y)$. Furthermore a metric to measure the similarity between individuals $d : X \times X \mapsto \mathcal{R}$ is defined. The Lipschitz-condition defines the fairness-criterion of distributions assigned to similar people should be similar. The difficulty of this criterion

is the definition of a similarity measure between individuals.

Group metrics: [7]

Group criteria are explicit about the similar treatment across different group dimensions and can be associated with notions of discrimination. All these metrics have in common that they are notions of equal rates of the true and predicted outcomes in the context of classification. To describe selected group criteria we will formulate them in the simple case of binary classification, where the criteria reduce to ratios of the quantities of a confusion table. The confusion table consists of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) We will denote two different groups as of a group dimension A as b and c.

Positive Rate Parity

$$TPR_b = TPR_c, \quad TNR_b = TNR_c \tag{7}$$

where

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{FP + TN} \tag{8}$$

Generally, the criterion defines fairness to be maintained if the probability of an outcome is equal for the group in a group dimension, given the value of the true outcome. In the binary case this can be formulated as equal true positive as well as true negative rates across the groups.

Predictive Rate Parity

$$PPV_b = PPV_c, \quad NPV_b = NPV_c \tag{9}$$

where

$$PPV = \frac{TP}{TP + FP}, \quad NPV = \frac{TN}{FN + TN} \tag{10}$$

This criterion genereally states that a model is considered fair, when the probability of a true outcome is equal for the groups of a group dimension, given the value of the predicted outcome. In the case of binary classification this can be formulated as equal positive predictive values and negative predictive values across groups.

A field of fairness in machine learning concerns the prevention of unfair bias in a

model. The proposed techniques are mainly focused on maintaining group fairness, hence the field also referred to as *discrimination prevention*.

Discrimination prevention generally aims to define algorithmic designs that reduce or eliminate discrimination bias by implementing fairness-criteria into the design. Methods of discrimination prevention can be implemented in the preprocessing, optimization or postprocessing step of a learning algorithm [58].

4.1.1. Fairness Tradeoffs

An important case study regarding the aforementioned group fairness criteria is the controversy about the COMPAS recidivism score [33] . Essentially the propublica initiative found out that black persons where more likely to be falsely classified to again commit a crime compared to white persons. White persons in turn were more likely to be falsely classified to not commit a crime again. The result thus refers to the criterion of Positive Rate Parity. The company behind COMPAS however implemented the Positive Rate Parity criterion [7]. Further research revealed that it is impossible to satisfy the two notions at the same time under certain non-trivial conditions [34]. The specific impossibility theorem states that it is not possible to satisfy both positive rate parity and predictive rate parity when the base rates differ by protected group and when there is not perfect separation of outcome classes. The result has important implications on the definition of group fairness as well as techniques of discrimination prevention. Similar impossibility theorems where obtained for other group fairness criteria as well [7].

Another trade-off regards fairness criteria and the accuracy of a model. Since most criteria constrain a model to equalize ratios of quantities of the confusion table, they likely reduce the accuracy of the model. It can be viewed as controversial to incorporate fairness criteria by default into a model, because such technical definitions of fairness are motivated by different metrics, illuminating the inherent ambiguity and context-dependence of such issues [14]. The causes of unfair or discrimination biased models are not taken into account as well as the complexities of such phenomena. The trade-offs between criteria make the incorporation also difficult, because it likely leads to a unfair model according to another fairness-criterion. Eliminating technical biases for instance, will decrease discrimination bias while incorporating fairness criteria might only obscure a inherently faulty design process or lead to unreasonable behavior in the model. On the other hand, fairness criteria may increase the accuracy, when there is representational bias in the data, leading to a erroneous difference along a social

dimension. Hence, we also need to keep in mind the limited and relative meaning of accuracy regarding our observed data, especially in the case of predictions on natural persons. In the case of historical bias in the data fairness criteria may be viewed as an intervention against a social symptom. Though, such interventions will not solve the underlying problem of unequal social structure. The use of fairness-criteria thus needs to be discussed and decided socially and context-dependent.

4.2. Discrimination Discovery

While group fairness criteria are used in discrimination prevention techniques, they can also be converted to measures of *discrimination discovery*. This field is associated with *discrimination-aware data mining* which also takes part in *discrimination prevention*. Discrimination discovery concerns techniques to identify discrimination bias in data or models [58]. It specifically concerns how to quantify discrimination, the identification of causal relationships in the data to attribute the discriminatory effect to. This especially regards to identifying direct forms of discrimination, since in theory the whole modeling process needs to be known. In the field discrimination is often defined by the legal terms of disparate treatment and disparate impact and like the fairness research defines social dimensions relevant to discrimination as containing legally protected groups [58]. Therefore discrimination in the field of discrimination-aware data mining and group fairness seem to follow the same conceptual framework.

The measurement of such discrimination assumes some kind of direct or indirect dependency between group membership and the outcome. Like the fairness literature, discrimination discovery mainly proposes measures for classification [58]. Since our case study focuses on a model producing a continuous score value that is part of the data set of the study, we will now focus on methods to measure discrimination bias in regression tasks.

4.2.1. Measures

In general there are different notions to measure discrimination classified by [58] into absolute, conditional and statistical measures. We will use mainly statistical measures in our case study, but will mention chosen absolute and conditional measures, because they will be related to the statistical measures used.

Absolute Measures: Those measures are used to quantify the overall difference of the outcome variable between group memberships of a social dimension. The most popular measure in early work on discrimination-aware data mining is the *mean difference*, which

we generalize here to also include regression tasks [41] There are major shortcomings of the measure. It is difficult to interpret the measures, since magnitude of a difference is no universally interpretable meaning. It is a observational measure and does not imply any notion of uncertainty and we do not account for the difference that can be explained by non-protected variables.

Conditional Measures: Such measures account are also quantifying the difference of an outcome between group memberships but at the same time account for difference that can be explained by non-protected variables. Generally this is done by building cohorts of equivalent covariate attributes and measuring the mean difference per cohort [58]. One approach to this are propensity scores [53], where the assignment of a treatment is modeled for example by logistic regression. In discrimination discovery the treatment is the social category in question. Then data points are stratified by propensity scores for example sorted by quantiles [10].

Statistical Measures: Measures that emphasize more on the confidence in a result are often statistical measures. They can be used to test the significance of an observational measure or provide a notion of uncertainty. Absolute measure and conditional measures can be used as a basis for a statistical test or model. Statistical significance does not imply practical significance [58]. This can be accounted for by computing effect sizes of the statistical tests which provide a magnitude of a effect. Measuring the difference in group dimension can also be done by a statistical model such as linear regression using dummy variables. The standardized coefficient can be interpreted as a effect size and also using other input variables can be interpreted as conditioning on those variables when looking at the coefficient of the group dimension. There are multiple examples of statistical modeling and hypothesis testing in discrimination discovery like in economics [36], police force [19] or credit market [51].

Finally, we want to note an important limitation of discrimination discovery in the context of our case study. While it is relatively easy to measure a statistical correlation between two variable x and y, it does not state anything about the cause that leads to a change in x or y. A statement about a direct influence and direction can be called a causal effect [41].

A typical framework to test for a causal effect of a on y would be to control for all other factors that influence y and then measure the influence of a in y. The idea of a parallel world, where only one variable a differs regarding a variable y is used in the counter-factual framework [41], [53]. To discover a direct discrimination bias in a model, a causal link between the group dimension a and the target variable must be proven. Indirect discrimination bias on the other hand describes a correlation between

a and y. It is assumed that the correlation can be explained by non-protected variable.

Except from randomized experiment, it is difficult to control for all factors. In the case of a learning algorithm the covariates are known though and can be controlled for. In the case of our study it is not practically possible to proof a causal link between a group dimension and the a target y. This is considered an exploratory analysis because we have no complete information about the predictors used in modeling the target [41]. We can only identify indirect discrimination bias which may be explained by a covariate we accounted for. If we cannot explain a bias with other covariates it does not imply a direct discrimination bias, since we might have an incomplete set of covariates. We at most can strengthen our belief in a direct discrimination bias.

5. Methodological Approach

In this section we explain the methodology that we will use to estimate the influence of social dimensions on the sector-scores of the SCHUFA credit bureau in a given data set. This goal implies important challenges we will discuss to account for them in the our analytic approach. To generate appropriate subsamples to measure the influence of a social dimension on a score, while conditioning on covariates, we will briefly describe propensity score matching. We will continue with our conditional measurement approaches, where one is based on the parametric linear regression modeling and the second approach uses non-parametric Gaussian processes.

Our goal is to estimate the influence of chosen social dimensions on sector-scores based on a given data set to obtain evidence for discrimination biases. The properties of the data set and the aim to model the sector-scores leads to the following important error sources we need to keep in mind.

(1) Representation

We cannot assume that the data set is, mostly because of the sampling process which is crowd-founding and so is not controlled by the sampler and is potentially biased by a specific sub-population that is frequently reached by the samplers who are publicly acting initiatives. Furthermore we can to some extend validate the data set for representation because we assume the German population to the approximate population of the data generating process i.e. the SCHUFA score. As we will see, a comparison with the German population with respect to relevant social dimensions show some biases in the data set such as a highly over-representative male persons. This complicates the identification of both

indirect and direct discrimination.

(2) Omitted variable bias
The dimensions used by the SCHUFA to compute the score are not publicly known. Therefore we cannot be sure about a causal relation of a social dimension and the score, because even when we control for all known covariates, we cannot be completely sure that there isn't a covariate we did not incorporated which is responsible for the correlation between a social dimension and the score. This problem makes it difficult to proof a causal connection and hence direct discrimination [51].

(3) Sampling bias and low support
Since we have a very small sample of the population that because of the sampling process alone has a potentially high selection bias in some dimensions, the data set as it is might have a high bias in estimates of effect in a social dimension. This is especially the case for combinations of feature attributes in one group that is not available in the other group. To estimate causal effects we are particularly interested in comparing examples that only vary in the social dimension subject to the analysis [41].

(4) Unknown modeling assumptions
Another issue with the used dimensions is the actual features that can be arbitrary transformations of these dimensions. Especially in the case of modeling assumptions such as linearity this becomes a problem since we don't know the type of relationship between dimensions which are most likely not linear in most cases. This affects the analysis of both indirect and direct discrimination, because a social dimension such as age might not be correlated linearly with the score but for example logarithmically, even when it is not directly used as an input-dimension.

We need to keep the issue of representation and omitted variable bias in mind when interpreting our results. To reduce bias of the estimation caused by low support and unknown modeling assumptions we introduce the methods in of the following subsections.

5.1. Subsampling

To account for sampling bias and low support between groups in a social dimension we apply a subsampling strategy to establish a sample consisting of similar covariate distributions between social groups of a social dimension [41].

We focus on *binary variables* representing a social dimension of two groups and use *propensity score matching* [53] to match persons with similar covariate attributes of the two groups. Since matching was introduced in the field of causal inference we will use the terms of treatment notated as A_1 and control group notated as A_0 in the further explanation. Treatment in our case refers to the group suffering from low support. We want to estimate the effect of A on a target y and assume that there is a set of shared covariates C that correlates with A and y. So goal is to isolate the effect of A on y by obtaining a set of similar group members of A_0 compared to A_1 according to the covariates C. To match similar persons of A_1 and A_0 with respect to C we want to model the probability $p(A = 1|C)$ of being in the treatment group A_1 given the covariates C.

We do that by optimizing a logistic regression model as introduced in 2.2.2 to estimate the function $e : C \mapsto A$. The output of the model in the context of matching is called a propensity score and is denoted as e. To finally obtain a subsample with similar covaraite distributions we choose the nearest neighbor of the propensity score e for each observation A_0 in A_1 in an iterative manner and stop after a given amount of iterations. A threshold of difference in the propensity score governs the acceptance of a matched partner [47]. Note that this matching strategy will lead to similar covariate distributions but not equal distributions. Hence, we view conditioning on the covariates when estimating the influence of A to y as still necessary.

5.2. Parametric Approach

We want to measure the influence of a social dimensions A on a score y given a set of covariates C in a data set. We define X to consist of A and C, which leads to our general notation to model a function $f : X \mapsto y$. In this subsection we describe a parametric linear regression model to estimate the function. This method and the associated frequentist view of statistical hypothesis testing in regression analysis is popular in observational studies concerning discrimination discovery [52]. The parameteric approach has the advantage of results that are easy to interpret. It provides a *t-test* to check if a model parameter significantly differs from 0 [57], which may be interpreted as a notion of statistical significance of an observed influence. The

parametric nature enhances its interpretability, because of the explicit assumptions about correlations. On the other hand, it requires much prior knowledge to explicitly incorporate non-linear relationships using feature transformations. There is a high risk of misspecifying the model assumptions.

The *linear regression* (LR) model is defined as:

$$y = X^T \hat{\theta} \tag{11}$$

We model a continuous variable y with a linear combination of a collection of variables V that are transformed to input features X and a set of weights θ. We will use an extension of the objective function of linear regression model, called *weighted least squares* (WLS) function to account for the potentially different numbers of matched partners in the matched subsamples the modeling will be based on. The objective function is:

$$L(\theta) = \sum_{i=1}^{n} \omega_i (y_i - X_i \theta_i)^2 \tag{12}$$

where

$$\omega_i = \frac{1}{n_{g,a}}, \quad \omega_i \in \Omega \tag{13}$$

n is the amount of group members, g is the matched group and a indicates the social group membership.

After the optimization step, we are interested in the parameter value θ_A of the feature A which represents a social dimension of interest. We interpret this as the influence of A on the score y conditioned on the covariates. The θ_A of a standardized data set can be interpreted as a effect measure. We can also use the frequentist framework to obtain a measure of statistical significance of the parameter and the model using aforementioned t-test for model parameters and a *F-test* to test if a model is significantly better then an intercept-model, given a defined significance level. These measures are only reliable under certain different assumptions, we need to check before interpreting the results.

- No Colinearity in the input features: There must be no strong correlations between the features. A violation causes the model to overestimate weights and consequently make the resulting measures unreliable.

- Normal distributed residuals: The residuals or error of the predicted target

instances $\hat{y}$ compared to the observed y must randomly spread around zero and approximately follow a normal distribution.

- No Heteroscendenacity in the residuals: The residuals must have equal variances
- No autocorrelation of residuals: The residuals of instances must not be related to each other

A critical point about this approach is the assumption of linearity. While it aids interpreting the model, linearity is often not an appropriate assumption of how variables correlate with a target variable, hence leads to a flawed specification of the model. To take non-linear relationships into account we introduce a process of feature expansion and selection to quantitatively choose suitable feature transformations for the linear regression model to specify reasonable relationships. We obtain a big set of features $\phi(X)$ by applying common transformation to the input variables X. This leads to many redundant features, because we transformed the same variable in multiple ways. To select appropriate features to model the score, we can not use ordinary least squares regression and select the highest weighed features, because of the limitations of this modeling approach concerning no colinearity. The LASSO regression is an regularization method and extension of linear regression. It introduces a penalty term in the objective function:

$$L(\theta) = (y_i - X_i\theta_i)^2 + \delta||\theta||_1 \tag{14}$$

The added term $\delta||\theta||_1$ penalizes high weight estimations in the training process. This eventually leads weights of redundant and irrelevant features to be pushed towards zero [27].

After the optimization step, we analyze the coefficients θ of the features and select a set of features according to the policy of the largest coefficient per variable-transformation in the range of a defined cut-off value. The selected features will be the input of the WLS-regression.

5.3. Non-Parametric Approach

The parametric approach has the main limitations of multiple assumptions that need to be fulfilled to reliably interpret its output and the explicit specification of relationships leading to the potential of wrong specifications even with a feature selection technique. Since we have no prior knowledge about the explicit relationships between the used

variables, we add a more flexible non-parametric approach to our analysis. We introduce the Gaussian process regression model (GP). With this modeling approach we do not explicitly assume relationships. In our analysis we can later check if the observed results of the linear regression and Gaussian process regression model are consistent.

Gaussian processes models are a Bayesian approach to model a function. In the case of GPs there are some important differences in the view and approach of a learning algorithm. One important difference concerns the treatment of parameters. While Frequentist approaches like the linear regression try to obtain an point estimate of a true parameter value θ_{true} , Bayesian approaches treat the parameter θ as a random variable that follows an unknown distribution. To estimate the distribution of θ, a some prior distribution is assumed and conditioned on observed data a updated distribution is obtained. The resulting distribution is called the posterior distribution. To compute the posterior distribution the *Baye's rule* is applied [49]

$$p(\theta|y, X) = \frac{p(y|X, \theta)p(\theta)}{p(y|X)} \tag{15}$$

where

- $p(y|X, \theta)$ is the likelihood,
- $p(\theta)$ is the prior,
- $p(y|X)$ is the evidence, with $p(y|X) = \int p(y|X, \theta)p(\theta)d\theta$,
- $p(\theta|y, X)$ is the posterior

Another important difference regarding GP models is, that the estimation does not consider parameters but the values of the function f, that we aim to model. This is done by assuming the data-generating function to be a multivariate Gaussian distribution, hence the function-values $f(X)$ are viewed of as draws from the distribution of the normal random variable f. Since the Gaussian process allows this for arbitrary inputs it is a generalization of the multivariate Gaussian distributions given the theoretic infinite dimensionality of the distribution. We can view the basic assumption as placing a Gaussian process as a prior distribution upon the function space $f(x)$ [49]

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

The properties of the function space that can be approximated are governed by the

mean-function $m(x)$ and the *covariance-function* $k(x, x')$.

$$m(x) = 0,$$

where we assume this to be the *zero-mean-function*, because we will standardize our data, hence having a mean of zero and a standard deviation of 1. The *covariance function* is also called a *kernel* and generally maps an input x and x' into a space of real numbers $k(x, x') \mapsto \mathbb{R}^{n \times n}$. It is crucial to the Gaussian process, because it defines the notion of similarity that is used to extrapolate new observations of the input space to the output space. This is based on the assumption that inputs X that are close, should have similar function values $f(X)$. Furthermore, by choosing a covariance function we encode assumptions about the function we want to learn. We introduce the 3/2-matern kernel, which will be relevant for this work:

$$k(x_p, x_q) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} \frac{|x_p - x_q|^2}{l})^\nu K_\nu (\sqrt{2\nu} \frac{|x_p - x_q|^2}{l}),$$

where $\Gamma(\nu)$ is the Gamma-function and k_ν is the Bessel-function. Our configurations of $\nu = \frac{3}{2}$ leads to a one time differentiable function of this general kernel, resulting in a rough shaped function. Furthermore we have the length-scale l, signal variance σ_f^2 and noise variance σ_n^2 for the covariance-function k_y used for the noisy target y as the hyperparameters of the model. Each hyperparameter can be interpreted and describes properties of the fitted data. The *length-scale* l describes the interval of variations in y for the input X. Thus, a low length-scale leads to more flexible and complex functions describing the data, while a large length-scale the functions become more simple. The scaling factor σ_f^2 can be interpreted as the *signal variance* which governs the uncertainty of the predictions leading to the overall space of possible functions that can be drawn, which in turn depends on the magnitude of covariances between nearby data points. σ_n^2 can be viewed as a *noise variance* which represents the deviance of the possible function-values from the observed data points. This deviance results from the limited flexibility of the function space governed by the kernel and its remaining parameters.

Given new instance f_{**} for a new data point x_{**} we can formulate the joint prior:

$$\begin{bmatrix} y_* \\ f_{**} \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} k(X_*, X_*) + \sigma_n^2 I & k(X_*, X_{**}) \\ k(X_{**}, X_*) & k(X_{**}, X_{**}) \end{bmatrix})$$

We can now condition the joint prior distribution on the data X, y, X_* to restrict the distribution to functions that agree with the observed data points:

$$f_{**}|X_*, y_*, X_{**} \sim \mathcal{N}(\bar{f_{**}}, cov(f_{**})),$$

where

$$\bar{f_{**}} = k(X_{**}, X_*)[k(X_*, X_*) + \sigma_n^2 I]^{-1} y$$

$$cov(f_{**}) = k(X_{**}, X_{**}) - k(X_{**}, X_*)[k(X_*, X_*) + \sigma_n^2 I]^{-1} k(X_*, X_{**})$$

By conditioning on the data we make use of the marginalization property of the normal distribution and take only the $n_* + n_{**}$-dimensional distribution defined by the n_* training points and n_{**} test points, ignoring the potentially infinite number of other dimensions. The predictive distribution has a variance and mean for each predicted target y_{**}. The resulting Gaussian process can be also described as the posterior process.

In the optimization step of the GP we want to optimize the hyperparameters of the kernel, conditioned on our observed data. The hyperparameters in training will be denoted as the vector $\theta = (M, \sigma_f^2, \sigma_n^2)^T$ Until now we treated l as a scalar or vector with equivalent entries. Since our main concern is an analysis of the relative relevancy of the features we want to assign a length-scale for each dimension in our data. This can also be referred to as *automatic relevance detection* (ARD) that determines l for each dimension leading the matrix M with diagonal entries of l. In extreme cases this will lead to a length scale with practically no effect of that dimension on the computation of the covariance, thus removing it from the inference.

To find the optimal hyperparameters we use the marginal likelihood, mentioned in the Bayes rule. The term marginal likelihood refers to the marginalization over the function values f. Under the Gaussian process model the prior is Gaussian:

$$f|X \sim \mathcal{N}(0, k(x, x')),$$

and the likelihood is a factorized Gaussian:

$$y| \sim \mathcal{N}(f, \sigma_n^2 I).$$

The logarithm of the product of those components yields the log marginal likelihood:

$$\log p(y|X) = -\frac{1}{2} y^T (k(x, x') + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log|k(x, x') + \sigma_n^2 I| - \frac{n}{2} \log 2\pi$$

The analytic tractability of the marginal likelihood is a major advantage of the Gaussian processes. To optimize the hyperparameters θ according to the marginal likelihood we incorporate θ into our notation leading to $p(y|X, \theta)$

The components of the marginal likelihood lead to an incorporation of the Occam's razor principle thus not only favors the best data fit but is a trade-off between simplicity of the model and fit to the data.

Since the optimization of the marginal likelihood might suffer from multiple local optima, we perform cross-validation on different hyperparameter initilizations to choose the best generalizing model. Cross-validation requires separating the observed data into a training and a validation set, optimizing the marginal likelihood with the training set and computing the error of predictions with the unseen data of the validation set. In our analysis we use the root mean squared error (RMSE):

$$\mathcal{L}(y, \hat{y})_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

This can be interpreted as an approximation of the ability to generalize well and we will refer to the RMSE of a test set as the generalization error (GE).

The estimation of the influence of a social dimension will be measured as a transformation of the length scale of the feature representing the social dimension. The disadvantage of this approach is the less intuitive interpretation of that measure compared to a linear regression weight. Also the optimization of the marginal likelihood does not yield a notion of uncertainty about the optimized length scale hyperparameters.

6. Empirical Study and Results

We introduced the relevant concepts and the context of the SCHUFA-solvency-score of our case study. We furthermore described related work regarding this thesis and the methodological approach we apply. Now, we turn to the concrete hypotheses of influences of social dimensions on the SCHUFA-score in the data set we want to analyze. We begin by describing the data set at hand and show representation biases in the data. Next, we will describe our process of deriving concrete hypotheses with the help of empirical results in social inequality research as motivated in 2.1.2. We further compare those results with the distribution of income in our data and with the score value as well. Then we describe the preprocessing steps which are important to reduce biases in the analysis of the derived hypotheses. Among others we briefly describe

our application of matching to lower the bias of low support. We continue with the inferential analysis, including the application of the parametric and non-parametric approach, followed by an interpretation of the results.

6.1. The Data Set

We will refer to the data set as the *openSCHUFA data set*, because it was sampled by the openSCHUFA project which consisted of the *AlgorithmWatch* Initiative and the *Open Knowledge Foundation* in 2018. We will briefly describe the sampling process to discuss potential sources of data biases and data errors. The sampling method was based on crowd-sourcing where the project members publicly called for free SCHUFA-reports (Datenkopie) required by section 15 of Datenschutzgrundverordnung (DSGVO). The report should contain the information stored by the SCHUFA about the natural person requested the report. Consumers interessted in more transperancy, regarding the SCHUFA solvency-scoring, were asked to take a picture of the aforementioned SCHUFA-report with a software designed for the openSCHUFA project and send it along with a filled out questionnaire with optional data fields to the servers of the openSCHUFA-project using the application [56].

The photographed documents were converted into a machine-readable format by an optical character recognition (OCR) software. The unstructured form of the SCHUFA-reports and the structured questionnaire data was provided by the openSCHUFA project for this thesis. While the output of the OCR-software, applied to the SCHUFA-reports, were unstructured texts, a very time-consuming process to again convert the data into structured data had been done by the editors of the data teams of SPIEGEL Online and Bayerischer Rundfunk (BR). This extracted form of structured data, combined with the data of the questionnaire, was then provided by the data journalists of SPIEGEL Online for this thesis.

The process of sampling and converting data into a structured and machine-readable form is important, because it implicates some assumptions and challenges regarding the data derived. First of all, the crowd-sourcing strategy likely leads to a sampling bias where some demographic groups are more likely to be in the sample then others. The group might be dominated by persons with a political interest in data and algorithm-assisted decision making due to the political context of the two initiatives forming the openSCHUFA project. The call of SPIEGEL Online and BR [37] might have lowered this sampling bias. However, the overall bias needs to be reflected and kept in mind through the analysis of the data.

Another critical aspect of the data set regards the conversion into a machine-readable format, where artifacts such as different brightness or different perspectives of the photographs of the reports partly led to symbols not interpretable in the machine-readable unstructured form. The extraction into structured data had to be carefully implemented. This circumstances lead to a potentially negative influence on the data quality and also needs to be kept in mind. Hence, we need to validate the plausibility of the data instances before analyzing the data.

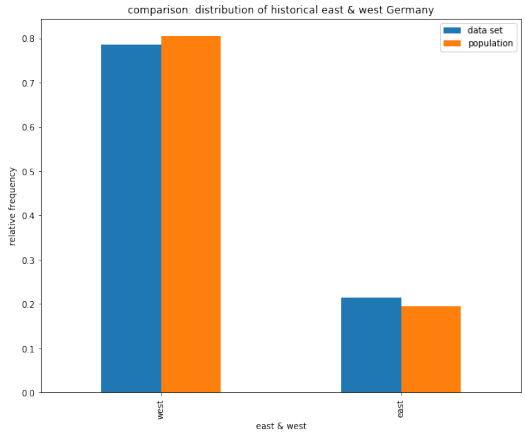
The questionnaire on the other hand has much less critical aspects about the quality of the data. The data was directly transformed into the structured JSON-format. Here, a major concern regards the information provided by the participants, since we can not proof the correctness of the information. We partly can validate the plausibility of the information as well. Nonetheless the mentioned concern motivates us to use as few information as possible from this part of the data set if there is a counterpart in the SCHUFA-report part of the data set because we assume that the SCHUFA score will be computed by some function of the information of the SCHUFA-report.

The questionnaire data consists of socio-demographic and economic data about the participant. Since this of course were optional information there is missing data in some instances. Socio-demographic data are important for our analysis of the influence of social dimensions in the SCHUFA-scores and are not included in the provided data of the SCHUFA-reports due to data privacy protection. The socio-demographic data is also important for the analysis of sampling and representation biases in the data which we will discuss in the following subsection.

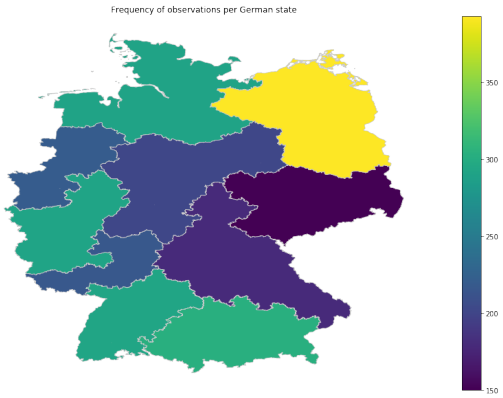
6.1.1. socio-demographic data

To get a general overview of the representation biases in the data, we compare the distributions of the social dimensions in the openSCHUFA data set, with the distribution of the German society according to the Zensus 2011 data [8]. We use the German society as a reference, because of our assumption that the solvency scores of the SCHUFA apply to a big proportion of German society [2]

Regarding the occurrences of persons living in eastern and western states in figure 1a shows a similar proportion compared to the German society. Examining the occurrences per German state as shown in 1b there is an obvious overrepresentation of persons living in upper eastern states and examining the full postal codes in more detail shows that many persons live actually in or around Berlin. Unfortunately there are too few observations for a more detailed analysis of location in Germany overall or in a city



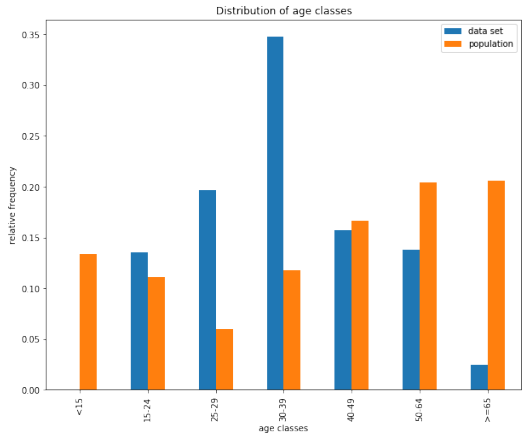
(a) Distribution of persons living in eastern and western states in Germany



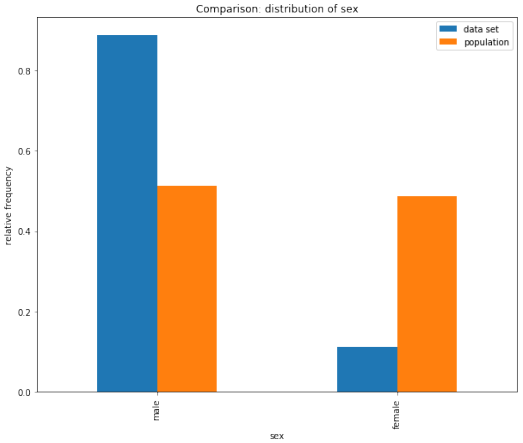
(b) Observed cases per German state

Figure 1: Distribution of observations in eastern and western states compared to the population

such as Berlin. Hence, we will focus on the differences between eastern and western states, but need to keep in mind that Berlin is here taken as part of eastern states, hence dominating the sample.



(a) Distribution of age classes in Germany



(b) Distribution of male and female persons in Germany

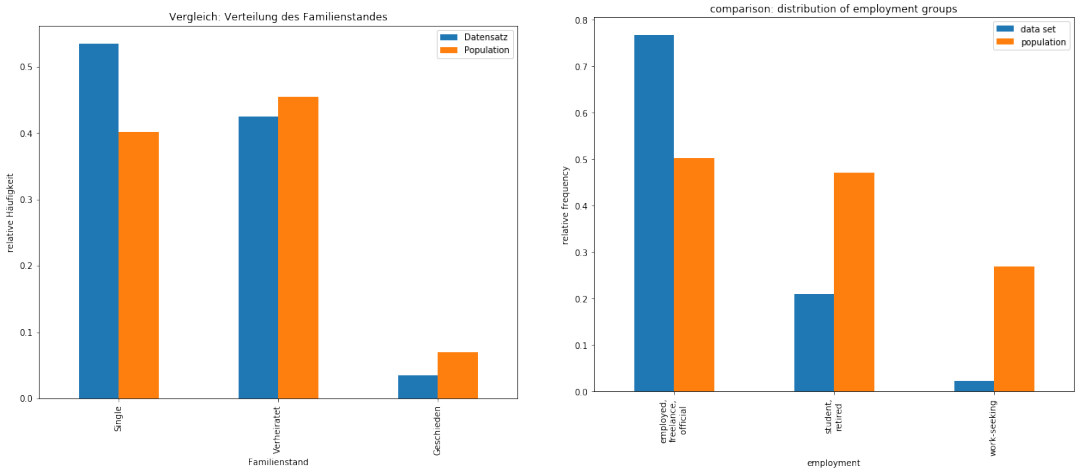
Figure 2: Distribution of observations of age classes and sex compared to the population

The figure 2a shows the proportions of age classes in the data set and the German population. While we can ignore the highly underrepresented group of persons under 15 since we assume that persons under 14 are not part of the SCHUFA data basis, the group of over 64 year old persons is also highly underrepresented and is assumed to be a relevant group in the SCHUFA data basis. The classes 40-49 and 50-64 years old

persons are slightly underrepresented and the class of 15-24 is slightly overrepresented. A bigger problem seems to be the classes 25-29 and 30-39 years old persons which are highly overrepresented. Hence, we can say that persons in the young and mostly in the middle age are overrepresented in the data set. For an analysis of the dimension age this representation bias needs to be taken into account.

The distribution of gender seen in figure 2b is obviously biased by a strong overrepresentation of men. The population distribution is approximately equal but in our sample women are only approximately 15% of the observations. To analyze this dimension we obviously need to take into account the representation bias. We also need to note that the attribute "non-binary" was unfortunately very infrequent and there were no data found for the German population regarding this attribute. Thus, we had to ignore this attribute in our analysis.

There is also a bias in the dimension of being a foreigner. We further note that the examples with the attribute of being a foreigner is very small with less than 2% of the data set. Hence, it is difficult to find an appropriate subsample for inference with less bias but still appropriately high sample size for the analysis. Therefore we unfortunately had to exclude it from our analysis.



(a) Distribution of the family status of persons in Germany

(b) Distribution of employment classes in Germany

Figure 3: Distribution of observations of family status and employment compared to the population

The distributions of family status in the data and the German population is shown in figure 3a. While there is an overrepresentation of approximately 10% of persons with family status single this bias is not particularly high and we assume a correlation

with the overrepresentation of young to middle aged persons that we assume to be more likely to have this family status.

In the dimension of occupation we see in figure 3b a strong overrepresentation of the category of "employment, freelancer, officials" as well as an underrepresentation of "student, retired" and "work-seeking". Since retired persons are mostly over 64 years old and we know our data is highly underrepresented in the age class over 64 years we assume that the underrepresentation mostly concerns retired persons. The overrepresentation of middle aged persons is most likely associated with the category "employment, freelancer, officials" and leads to its overrepresentation. We will briefly analyze this assumptions descriptively in 6.2.

As we have described different socio-demographic dimensions in the data set we formulate our assumption to have strong associations between the bias in the age class distribution and the dimensions of family status and employment. We assume that the SCHUFA has no information about family status and occupation [2]. Therefore we will concentrate on the age dimension. We furthermore concentrate on the dimensions of sex and geographic location. We have to ignore the dimension foreigner because of the low support in this dimension.

6.1.2. Economic data and the SCHUFA scores

The SCHUFA-scores and most of the economic data will be retrieved from the SCHUFA-report. We now want to shortly describe the SCHUFA-report and the SCHUFA-scores with the aim of selecting appropriate scores for our analysis.

The economic data in the questionnaire should be implicitly included in the SCHUFA-report, apart from the monthly income. We decided to use the information in the SCHUFA-report rather then the questionnaire data when available, because of our assumption that this is the information used in the function to compute the SCHUFA-scores. Furthermore the questions of the questionnaire might be misleading and participants might interpret some questions especially about economic data differently then others. To avoid biases because of semantically differing data we ignore the fields regarding loans from the questionnaire.

The SCHUFA-reports, used for this thesis, consist of three parts. A first table contains solvency-score values requested by contractual partners about a person in the past 12 months at the time when the report was created. We will refer to this part as *request-table*. The second table with the current solvency-score values we refer to as *score-table* and a third part, the *text-data*, with text regarding information about

further requests and information provided by contractual partners about a person. We assume the request-table as well as the text-data with requests and activities of a person to be important for the input of the SCHUFA-scores.

The solvency-score itself is not one score but there are 9 different scores regarding different sectors. We refer to these scores as *sector scores*. The sector potentially leads to a different weighting of important predictors per sector [1]. The differences and assumptions between the sector-scores are not transparent. Overall, there are the following sectors present:

- bank
- freelancer
- commerce
- mortgage banking
- small business
- mail order
- savings bank
- cooperative bank
- telecommunication company

There are three different versions of each sector-score. While version 1 is the oldest, version 3 is the newest and recommended by the SCHUFA. One obvious difference between version 1 and the other two versions is the score for savings bank and cooperative bank ("SCHUFA-Score für Sparkassen/Genossenschaftsbanken") which is in the other two versions split into two different scores savings bank ("SCHUFA-Score für Sparkassen") and cooperative bank ("SCHUFA-Score für Genossenschaftsbanken"). Since it would exceed the scope of this thesis to consider all sector-scores we need to decide for some sector-scores we want to analyze in detail. To describe the selection process we focus on the description and analysis of the score-table.

Each score in the score-table of the report is represented as a value which ranges from 1 to 1000 in version 1 and from 1 to 10000 in Version 2 and 3. Additional to the score-value, there are the fields *rating*, *repayment probability* and a text field called *meaning altogether*. All mentioned columns seem to be ordinal interpretations of the

score value. The bank score for example is assigned an A-rating and a repayment probability of 99.2% when the score value ranges from 9863 to 9999 and an M-rating and repayment probability of 39.55% when the score value ranges from 1 to 4927. This ordinal classification requires a person to have no open negative features else another filtering mechanism would be applied [1]. We note that it should not practically possible to get a score that would lead to 100% or 0% repayment probability due to the concept of probability which asymptotically approaches 0% or 100%, but there is no probability of 100% because of random events for instance which lead to an unexpected outcome.

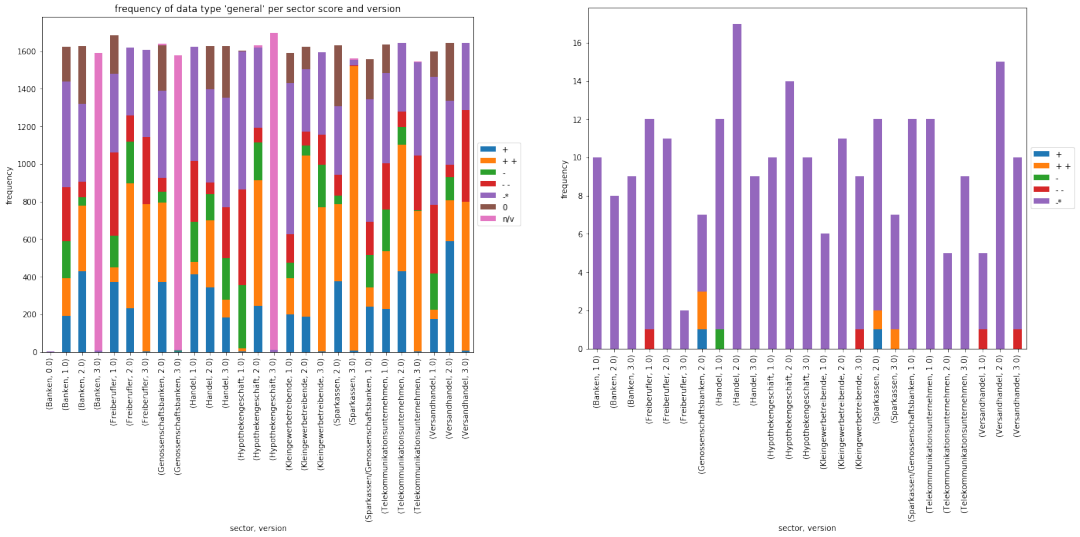
The other columns regarding the solvency-scores are summarized by *risk in data types*. The SCHUFA reports a rough information about the influence of attributes of dimensions in a category of data types. The influence of the attributes ranging from negative "–" to positive "++". There are four types of broad economic data types *previous payment problems*, *credit activity of last year*, *use of credit*, *length of credit history*. The last two categories seem to be most interesting for our analysis namely *general data* and *address data*.

We use the information of frequency of data types *general data* and *address data* as well as the number of requests for a score and the occurrence in the data set as criteria of relevance to decide the sector-scores. The two data types are assumed to be important indicators for a possible direct influence of social dimensions as predictors and the requests for a sector-score is viewed as an approximation for the importance of a score in consumer activities.

All scores and versions are approximately equally distributed in the data set. We ignore the special case of version 0 of the bank-score and the splitting of the score "Sparkassen/Genossenschaftsbanken", since it leads to difficulties in comparison of the scores.

For an overview of the usage of *general data* we can see in figure 4a that in Version 3 of bank ("Banken"), cooperative bank ("Genossenschaftsbanken") and mortgage banking ("Hypothekengeschäft") the data type is almost never used. In these cases it might be interesting to see if the scores nevertheless are correlated with social dimensions. We can further see the amount of positive ("+"/"++"), negative ("-"/"–" and "-*" for uncertain negativity introduced as a category by the SPIEGEL Online journalists) and neutral ("0") influence. Positive and negative cases seem partly negatively overweighting.

The data type *address data* in figure 4b is most of the time not used, hence we ignore this attribute to get an idea of the distribution of cases where the category is indeed used. Here the cases seemingly have a high frequency of uncertain negative attributes ("-*") which is why we cannot rely in detail on the given data. But we can see that



(a) Distribution of general data influence in the data set

(b) Distribution of address data in the data set

Figure 4: Distribution of influence of data types general data and address data per sector score version in the data set

each sector-score seems to use this data in some cases. Most often the sector-score of commerce ("Handel"), mortgage banking ("Hypothekengeschäft"), and mail order ("Versandhandel") uses this data type, even though these are few cases in the data set. Note that we can not say if the data categorized as *address data* is used for geo-scoring. Geo-scoring is applied in few cases where the SCHUFA has too few data about a person for a specific sector-score [2]. We also can not be sure of if the data type *address data* also includes the number of relocation of a person as it might also be an input of the score.

To approximate the usage of a sector-score we use the number of requests of the last two months. The bar plot in figure 5 shows a very high number of requests of all versions of the bank score. Another sector-score with higher number of requests in each version is the mail order score and the commerce score.

Since the mail order score seems to use the data types *general data* and *address data* in all three versions and is requested relatively often by contractual partners, we identify this score as relevant. Another score we choose to further analyze is the bank score, since it is most often requested by contractual partners of the SCHUFA and seems to almost never use the *general data* type in version 3 which makes it particularly interesting to compare the influence of social dimensions with the older versions as well as comparing it with the mail order score.

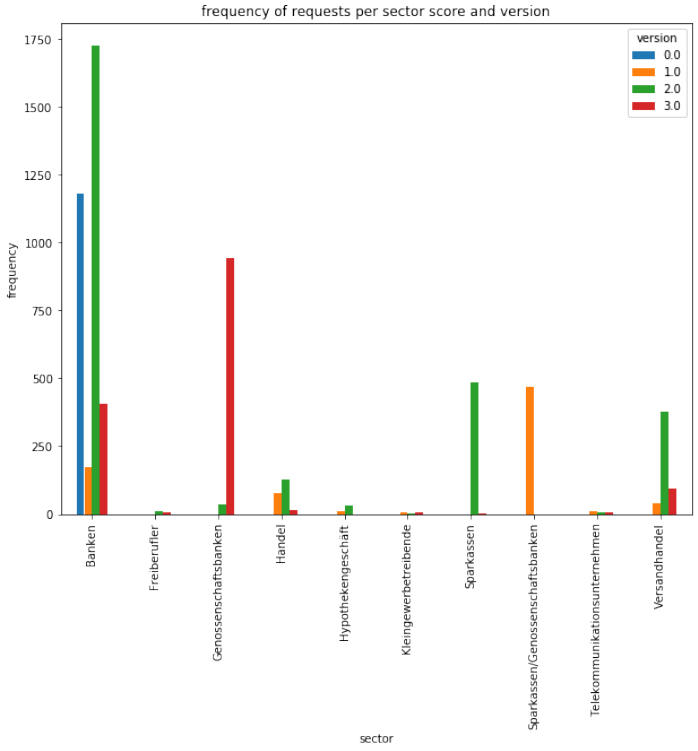


Figure 5: Distribution of requests frequency per sector score and version

6.2. Deriving Hypothesis

As a first step of the analysis we need to formulate possible hypotheses about social groups that might be distributed disproportionately in the SCHUFA-solvency-scores. To derive hypotheses about social dimensions in the data set, that are also social groups vulnerable to discrimination, we want to make use of empirical results of the field of social inequality research as motivated by 2.1.2. As an approximation of the possible inequalities, most likely incorporated in an indirect or direct way into a solvency-score, we use the inequality of wealth indicated by income and assets distributions in German society.

The scores may be viewed as regulating the exclusion and inclusion of wealth-related resources to some extend by influencing the distribution of resources affecting a high proportion of German society. On the other hand, we assume the SCHUFA-scores to be related to wealth, because high material wealth may lead to a fewer default risk. We therefore assume a relationship between the scores and wealth indicators such as income, because the indicators may directly or indirectly influence the score. Therefore, we assume a association in the distributions of income and solvency-scores.

Hence, to select hypotheses we first look at empirical social inequalities in Germany to identify possible social groups vulnerable to discrimination. Next, we want to find descriptive evidence for the identified income inequalities in our data set. As a last step, we compare the evidence of inequalities in income with the distribution of scores. If there is evidence for the assumption that the score incorporates such income inequalities it might even amplify patterns of inequality, because of the influence on wealth resources.

6.2.1. Building Hypotheses from Wealth- and Income-Inequalities

We will now briefly describe empirical results of inequalities in wealth to identify social dimensions that lead to inequalities and hence might also influence the SCHUFA-score. As mentioned before, we want to focus on horizontal determinants that represent social dimensions, as required by our definition of discrimination bias. Hence, we will not address observations in vertical dimensions such as occupation, because those inequalities are partly accepted by society in terms of effort.

The gain of a high economic status counts, in modern societies, to the most important goals to fulfill socially common objectives. In modern market economies, wealth is considered as the core of a high economic status. Money directly enhances economic status and wealth and therefore is an often used indicator to describe the distribution of wealth. Objectives like security, health, reputation, power and housing or working conditions are to some extend purchasable and therefore can be obtained with money. Moreover, post-materialistic values come often to the fore not until materialistic values are to a certain degree accomplished and in place. Hence, economic status is a very important dimension of social inequality that enables or prevents multiple opportunities of reaching socially common objectives [30].

There are two common measurable concepts that can describe wealth and its distribution in a society: income and assets. We will briefly discuss both indicators of wealth and will later refer to these indicators regarding empirical results observed in German society. First of all, income is a relative measure in that there are different levels of wealth in societies. Wealth and poverty have a different absolute level in Germany, compared to, for example, India. Hence, the distribution of wealth is referred only to the German society in the further discussion of this thesis.

The distributions of these indicators have direct societal and political consequences which is why materialistic inequalities can be seen as a core area of social inequality. If economic changes such as the degree of difference between rich and poor people are not

viewed as fair in a society, this directly leads to politically relevant popular discontent and social conflict [30]. This also illustrates the importance regarding solvency-scoring and discrimination-relevant social groups that are economically disadvantaged. We will now discuss empirical results of income inequality along social groups.

A structural change observed since the end of 20th century in Germany is the shift of the society from a industrial society to a post-industrial information and service society. This is expressed by an increasing importance of information, knowledge and qualification. But within the service sector enormous inequalities can be observed between high-qualified persons with high security and low-qualified persons having low incomes and few security because of a highly competitive work environment [31]. The inequalities that are largely related to the vertical dimension of occupation can be also associated with horizontal dimensions. We will briefly go into observed inequalities in sex, age, eastern and western states and family status to justify our potential hypotheses. In social inequality research the categories race, ethnicity, age and sex are seen as factors that produce social disadvantages and [54] argues that social inequality can not only be explained by vertical dimensions but also by the membership of discriminated or privileged social groups which are affected such as women, foreigners, people of color or people living in disadvantageous residential areas.

First of all, the income and asset inequality in Germany can roughly be structured by the vertical dimension of professions because a main part of income of a person is earned by occupation which roughly follows the order from high to low regarding: self-employed, pensioners, officials, employees, workers, retirees, farmer, unemployed and welfare recipients [30]. Furthermore the hierarchy of income can be explained by horizontal dimensions.

The income inequality in the horizontal dimension sex is an extensively discussed topic in the public discourse, which is also known as the gender pay gap. In [9] the inequality in income is estimated at 30% in favor for men. While most of the difference can be explained by factors like different performance groups and career choice, 6% of the difference could not be explained by such factors which might be caused by missing explainable variables or discriminating actions. Regardless of the cause this variable is obviously relevant for discrimination as a social dimension where women are the diadvantaged group and men are the reference group.

The age is a mostly accepted determinant of income and asset inequality because of normative stations a person is likely to undergo [30]. This mainly regards the dependence between age-classes, job career and income. While in the childhood labor is prohibited in Germany, school career normally ends with the age of 16 to 18. This

follows a job or academic training which determines children, juveniles and young adults to no or rather small income. On the other hand looking into the range between 18 and 64 is a diverse group which still has a income inequality overall. The group of old adults also exceeding the age of 64 and therefore including the age of retirement is the most controversially analyzed group in terms of social inequality [54]. Therefore the age can also be viewed as a social group dimension relevant to discrimination where according to empirical results of social inequality research young persons as well as old persons are disadvantaged and the reference group may be the group of middle aged persons. This may also be viewed as a continuous distribution with a peak at some age in the middle to old age.

An example of geographical difference in income is the regards the historical eastern and western states of Germany. While the structure of class between west and east seems to be more aligned the financial conditions of persons living in the east are in most social situations worse compared to persons living in the west of Germany [9]. The amount and structure of gross income differs between east and west where only 77% of western household gross income where available to households in the east [9]. Hence, the geographical dimension of eastern and western states may also be discrimination relevant. The financially disadvantaged persons living in eastern states can be compared to the reference group of western states citizens.

The income inequality can also be structured by the family type. Here, especially single mothers and female persons living alone are disadvantaged in comparison to other family types. An observed hierarchical order of income inequality structured by family type from high to low is as follows: Couples without children, men living alone, couples with children, women living alone, single parents [30]. Thus, having a partner is advantageous as well as having no children. This may be also viewed as a discrimination relevant variable. Though, we assume that the family status is associated with the dimension sex and age. Furthermore, the SCHUFA credit bureau should not have information about that social dimension. We will therefore focus on sex and age instead of family status.

Another important horizontal determinant of income inequality is the group membership of not having the German citizenship [9]. Since our data in this dimension is too scarce regarding persons of that group membership, we will not go into this inequality.

6.2.2. Selecting Hypotheses

We have described multiple intersecting inequalities in income along social dimensions. While employment is a vertical determinant, the horizontal dimensions age, sex, living in eastern or western German state, being a foreigner or the family status are also partly determinants of the income distribution.

Referring to the openSCHUFA data set as described in 6.1, we have to ignore the dimension of being a foreigner because of the high infrequency of foreign persons in the data set.

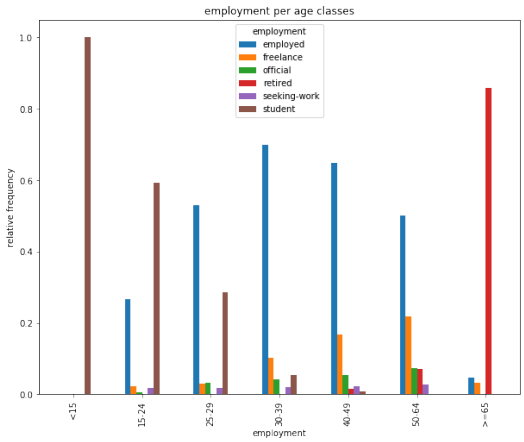
Since the horizontal dimensions gender, eastern/western states and age are unequally distributed with respect to income and wealth in the German population, they are also social groups which are possibly affected by discrimination. Especially age and gender are such categories, since social norms determine these characteristics and it is not possible to change these on a persons own behalf. These categories as well as the eastern/western states might also be part of a persons social identity.

Apart from the income distribution we will analyze the horizontal dimensions regarding the distribution of the dimension family status as well as the vertical dimension employment in the data set to find possible associations and biases that we need to keep in mind.

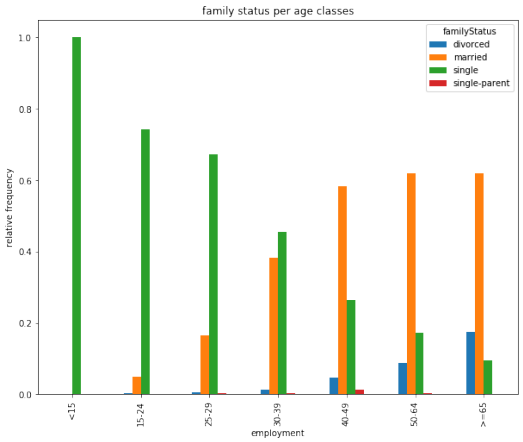
We now report results from the analysis of distribution of income for each of the identified horizontal dimensions and compare those with the population wide inequalities. We then look at the distribution of occupation and family status along the dimensions and finally compare the score distributions along the dimensions with the income inequalities in our data set as evidence for our hypotheses about the effect of horizontal dimensions on the solvency scores in the data.

monthly income	
Feature	Spearman correlation
age	0.42
score	0.25
score	
Feature	Pearson correlation
age	0.45

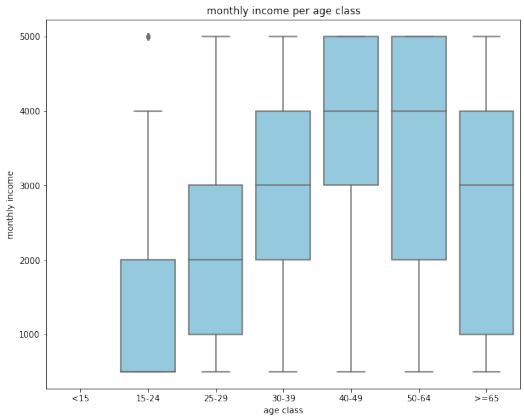
Table 1: Correlation between age, monthly income and bank score version 1



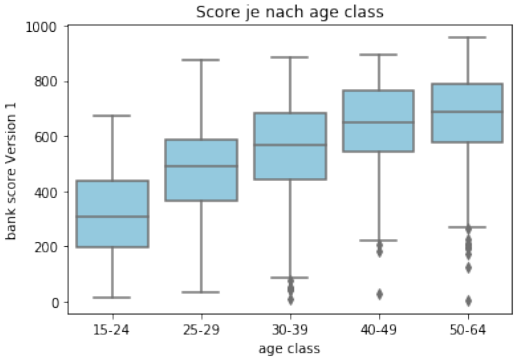
(a) Distribution of employment per age class



(b) Distribution of family status per age class



(c) Distribution of monthly income per age class



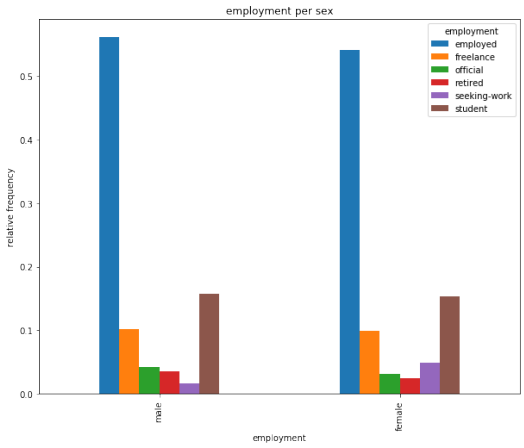
(d) Distribution of bank score version 1 per age class

Figure 6: Distribution of employment, family status, monthly income and bank score version 1 per age class in the data set

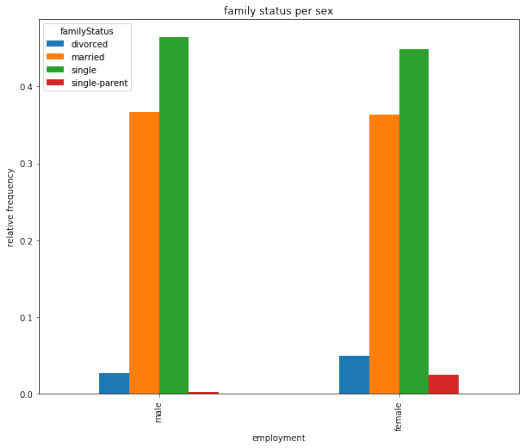
In figure 6a we can see that our assumptions about influences of the overrepresentation of middle aged persons in the data set regarding occupation and family status are descriptively evident. We can see in figures 6a and 6b a high proportion of employed singles in the group of middle-aged persons, especially 25-29 years old persons. The high proportions of singles or married persons indicate no big difference in income for an age class because these family statuses are in the upper part of income or assets distribution. We furthermore have a high proportion of employed persons in the overrepresented age classes. On the other hand the age group 15-24 has a high proportion of students and persons greater than 64 are mainly retired. These groups will likely have a smaller income due to their employment status.

As stated in social inequality research our data set also shows in 6c an income distribution that is unequally distributed along age and seems to increase with age. The class of persons over 64 years old indicate a decrease of income which may be caused by retirement. Since we have too few data instances for this category, we are going to drop it and no longer include it in our analysis of the score. Looking at the distribution of score values of the bank sector score, version 1, we see in figure 6d that the order of the distribution along age is very similar, the median of the score value is increasing with age. As seen in 1 the Spearman correlation coefficient of monthly income and age equals 0.42 which is a medium to high correlation. This interpretation of correlation coefficients or effect sizes are in studies of psychology and social science often guided by [11]. The pearson correlation coefficient of the score value and the age is 0.45 which is also a medium to high correlation. This may support our assumption that the income governs to some extend the solvency-score. On the other hand the spearman correlation of the score and the monthly income is only 0.25 which is only a small to medium correlation. Hence, the age might be used as a proxy variable for the income. Note that the monthly income variable is ordinal scaled with different intervals, hence it is not fine grained as the age variable. This can be another explanation for the different correlation strengths.

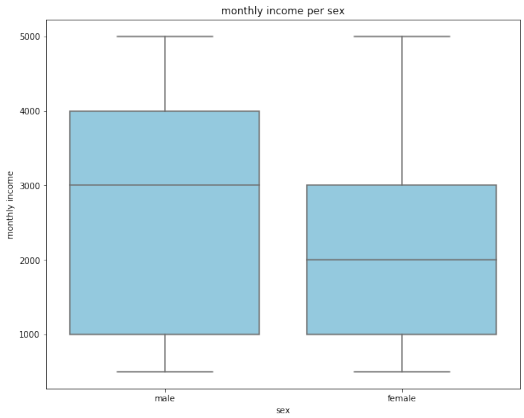
In our data as shown in figure 7a and 7b the proportions of employment and family status are descriptively almost similar. There is a higher proportion of women that are single-parents as well as a higher proportion of women that are work-seeking than men. The dimension of gender should be higher for men than for women, according to empirical results in income inequality. This can also be seen in figure 7c where the median monthly income interval of men is 3000 and the median monthly income



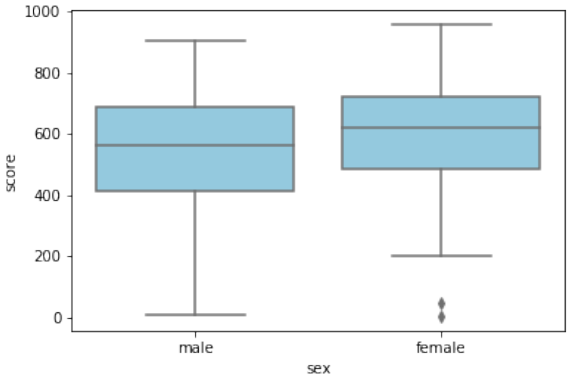
(a) Distribution of employment per sex



(b) Distribution of family status per sex



(c) Distribution of monthly income along sex



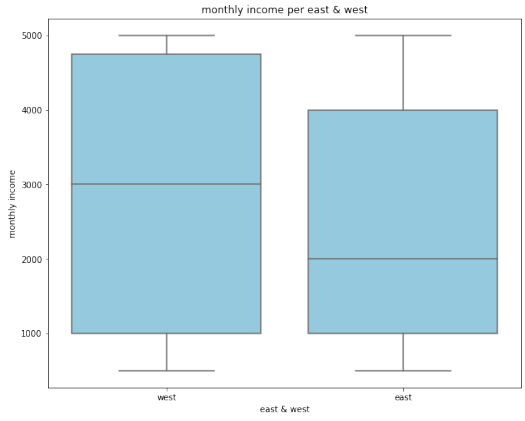
(d) Distribution of bank score version 1 along sex

Figure 7: Distribution of employment, family status, monthly income and bank score version 1 along the social dimension sex in the data set

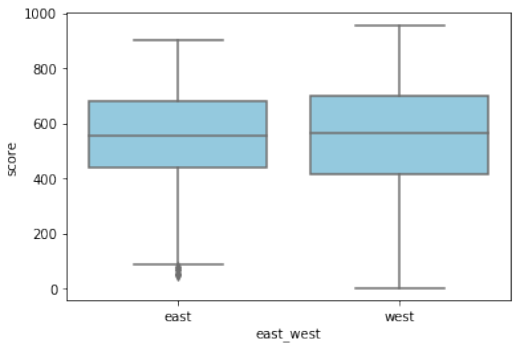
monthly income		
Feature	U-Statistic	p-Value
sex	46351	0.00
east west	112671	0.00
score		
sex	73635	0.005
east west	135101.5	0.664

Table 2: Mann-Whitney-U test result for the variables sex and east west states

interval of women is 2000. Interestingly, this descriptive difference in favor of men is reversed when we look at the distribution of score values of the bank score version 1 along sex in figure 7d. To examine the statistical significance of the descriptive results we computed the non-parametric Mann-Whitney-U test [12] to test if the samples of monthly income and score value between the groups male and female where drawn from the same distribution. A p-value near 0 implies different distribution and hence a significant difference in the samples. In 2 we can see the U-statistic and the resulting p-value which is approximately 0. This is evidence for our descriptive result of differing monthly income and score values in the bank score version 1. The mentioned small spearman correlation between monthly income and the score value and the observation of reverting inequality along the social dimension sex regarding the score contradicts our assumption of the income governing the solvency score. Though, we need to keep in mind that this is only a bivariate analysis ignoring possible covariates.



(a) Distribution of monthly income in eastern and western states



(b) Distribution of bank score version 1 in eastern and western states

Figure 8: Distribution of employment, family status, monthly income and bank score version 1 along historical eastern and western states of Germany in the data set

The distribution of income in 8 partly resembles the empirical results of the German population. The median income is lower in eastern states than in western states. The higher median income is located in the south of Germany and the lowest median income in both eastern German states. This similar to the empirical results of income distributions in social inequality research. The distribution of the bank score version 1 on the other hand is not seemingly different between eastern and western states which again differs from the monthly income distribution. To further analyze the difference we again compute the Mann-Whitney-U-Test. In table 2 we can see the resulting high p-value of the U-statistic of the distribution of income and bank score version 1. Both resulting p-values confirm our descriptive observations. While the monthly income distribution differs significantly between eastern and western states, the score distribution does not.

To conclude, the monthly income seems to be relevant to decide the default risk but it seems to be not a direct approximation to model such a risk because of the inconsistencies between associations of social dimensions and monthly income compared to the score value. Note that these descriptive results cannot lead to any conclusions regarding the direct or indirect influence of a social variable to the score due to the observed sampling and representation biases in the openSCHUFA data set as well as the difference in possible predictors. We conclude the results leading to our hypotheses we state about the influence of the social variables to the SCHUFA solvency-scores we want to further analyze:

- There are discrimination biases along social dimensions
  - Being older has a positive influence on the score value in the openSCHUFA data set
  - Being female has a negative influence on the score value in the openSCHUFA data set
  - Living in eastern states has a negative influence on the score value in the openSCHUFA data set
- The influence of a social dimension can be explained by constructed variables from the SCHUFA-report
- The strength of the influence of a social dimension reduces with the score version
- The strength of the influence of a social dimension differs between sector-scores

6.3. Preprocessing

Before we model the scores as a regression task to analyze the influence of the social dimensions we need to account for some concerns to adjust the data for in preprocessing steps. This mostly concerns the reduction of sampling bias and low support of a social group as well as a data basis that helps to explain the correlations between social dimensions and the score value.

We start with identifying possible covariates from the SCHUFA-report we want to account for to explain a possible influence of a social dimension on a score. We then describe our definition of outliers in the data set and argue for their removal. Then we describe transformations we apply to the data, to match the assumptions of our modeling approaches and the requirements of the analysis. Finally, we explain our approach to select meaningful subsamples for the analysis of discrimination biases regarding each social dimension.

6.3.1. Variable Discovery

The openSCHUFA data set in its unstructured form has multiple sources of possible variables used to compute the SCHUFA scores. We assume the score-table and the text-page of the SCHUFA report to be the main source, where we can identify covariates of the social dimensions. The structured form of the data provided by SPIEGEL Online already converted the text page of the SCHUFA report into structured information. We used the classified information as binary or count variables regarding a persons finance activity such as a contract for a giro account or a credit card.

We create count variables for the request in the past 12 month where we count the overall requests and the requests per sector score. To incorporate the time-related information for the length of financial history we compute the delta between the first activity concerning credits, credit cards or giro accounts and the date of when the SCHUFA-report was created.

A list that shows all variables inferred with a brief explanation can be found in the appendix A.

Note that the identified variables are only assumed to potentially have an influence on the SCHUFA-scores and are mainly inferred, based on the SCHUFA-report. Hence, we cannot be sure to control for all covariates of the social dimensions and cannot be sure about the true relationship between a social dimension and the scores. There may be an unobserved variable or feature that mediates the observed association.

6.3.2. Outlier Removal

The derived variables, that we assume to be potentially used, might be too infrequent to account for in the analysis. In our data set this concerns some of the count variables, which all have a modal value of 0. Since each variable adds more complexity to the model, we filter for the modal value in some count variables, which have only very infrequent values except from 0. In doing so we implicitly condition on those variables. Furthermore, we have to remove persons that are too underrepresented in the data set, as observed in 6.1. This concerns persons younger then 15 years or older then 64 year as well as persons with non-binary gender.

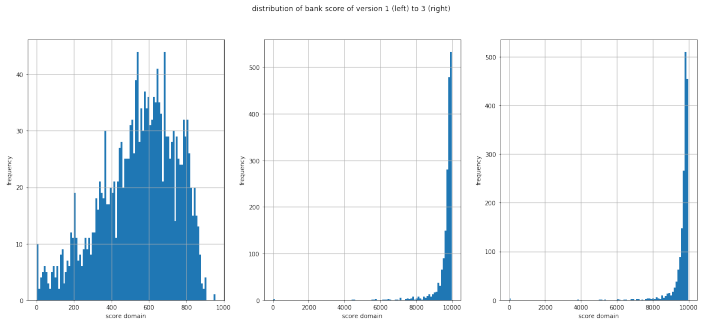
We also know that the SCHUFA scores are differently computed depending on the "negative features" by the SCHUFA [1]. If such feature exists for a person, he or she is assigned a different rating (N, O or P). According to the SCHUFA, this is for case, when there are debtor records publicly available of this person. Since we assume to not have such information in our data set and that this concerns a relatively small amount of people, we decide to remove persons with such ratings from the openSCHUFA data set. Hence, we will model the relationship of persons that have no "negative features" and we assume that there is no prior determining evidence for a high repayment risk.

Lastly, we remove instances with assumably not plausible information, because of wrong information given in the questionnaire or wrongly recognized information in the conversion process to machine-readable data. This concerned score-values that are not in the valid range of the score version, invalid ratings and not plausible creation dates of the SCHUFA-report.

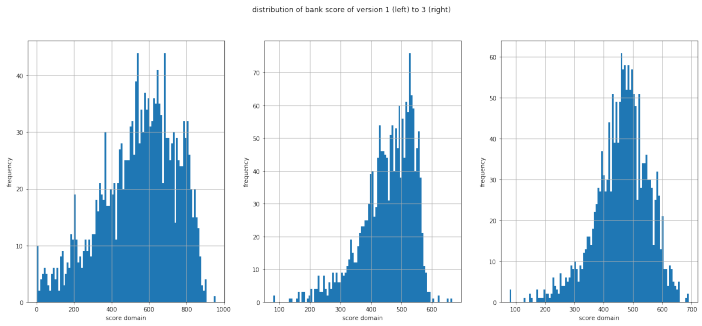
While the original data set had 2625 instances, the removal of outliers led to a removal of 28 observations. Hence, our cleaned data set contains data about 2597 persons.

6.3.3. Data Transformations

We now describe the transformations we need to apply to the data, to use the variables in the forthcoming models. In all modeling approaches we will assume the target variable to be approximately normal distributed. This is on one hand important for our parametric models where a normally distributed target might help to obtain randomly normal distributed residuals, which is important for a reliable interpretation of the linear regression used in the forthcoming analysis 5.2. Also the non-parametric approach in the analysis is modeling the target as a Gaussian random variable as explained in 5.3. The distribution of scores in figure 9a shows that the version 1 of the



(a) Original bank score distribution per version



(b) Transformed bank score distribution per version

Figure 9: Distribution of bank score versions in data set before and after transformation

score seems to be approximately normal-distributed with a left shift. The version 2 and 3 on the other hand seem exponentially distributed. To rescale the distribution to an approximate normal distribution similar to version 1, we apply the following function:

$$y_i^{scaled} = 1000 + \log(10000 - y_i) * 100$$

This function subtracts each y_i from the maximum of the target and applies the log-function. To get a similar range to version 1, the result is multiplied by 100 leading to a maximum of 1000 and adding the result to the maximum to revert the subtraction that we had to apply to correctly take the natural logarithm. The distributions after the transformation figure 9b show better approximation of the normal distribution. The left shift is slightly bigger then in version 1.

We also convert the ordinal dimension monthly income into a rank feature and the nominal dimensions sex and being from eastern states or western states as binary features that indicate being female or living in eastern states when the binary features equals 1. After the transformations we standardize the metric and ordinal dimensions such that all dimensions have a mean of 0 and a standard deviation of 1 by applying

the function:

$$x_i^{standardized} = \frac{x_i - \bar{x}}{s},$$

where $\bar{x}$ is the mean of x and s is the standard deviation of x.

6.3.4. Subsampling

Our final preprocessing step before we execute our analysis of the influence of social dimensions on the SCHUFA scores is to obtain an appropriate subsample for each chosen sector-score, its version and the social dimension considered. Our main goal is to find a subsample which has similar covariate distributions in the groups we want to compare as defined by the social dimension. By choosing comparable instances from all groups we aim to reduce the estimation bias in the influence of the social dimension on the score because of group members having low support in the combination covariate attributes in the openSCHUFA data set.

We want to obtain such subsamples by applying the propensity score matching. Since the subsampling process is equivalent for all combinations of scores and social dimensions, we exemplary describe the process in the case of the social dimension sex and the bank score, version 1 data set. Note that the subsampling of age is done by constructing a binary variable that encodes persons of the age between 18 and 39 as the control group, encoded by 1, and persons between 40 and 64 as the treatment group, encoded by 0, to account for the bias in the young and middle aged persons as discussed in 5.1.

before matching	
Feature	p-value
monthly income	0.00
finance year	0.033
after matching	
monthly income	0.61
finance year	0.601

Table 3: p-values of Mann-Whitney-U test smaller then 0.1 before and after matching of feature female in bank score version 1

To get an overview of the association between the social dimension sex and the covariates, we compute the Mann-Whitney-U-Test for the covariates distributions

between male and female groups. We use the test to validate if the matching process worked. We assume significant differences in the distributions of groups which should vanish in the obtained matched subsample after the matching process. The U-statistic and p-value per potential covariate was computed. We assume a significant difference with a p-value smaller than 0.05 and display only covariates with a p-value smaller than 0.1 to show only relevant cases. In the appendix A are also the results for the other social dimensions. As we can see in table there are 2 covariates with a lower p-value then 0.1. Since we want to have similar attributes in all covariates we try to match male and female persons on the basis of all covariates but match only the significantly different covariates if the matching process with all predictors fails.

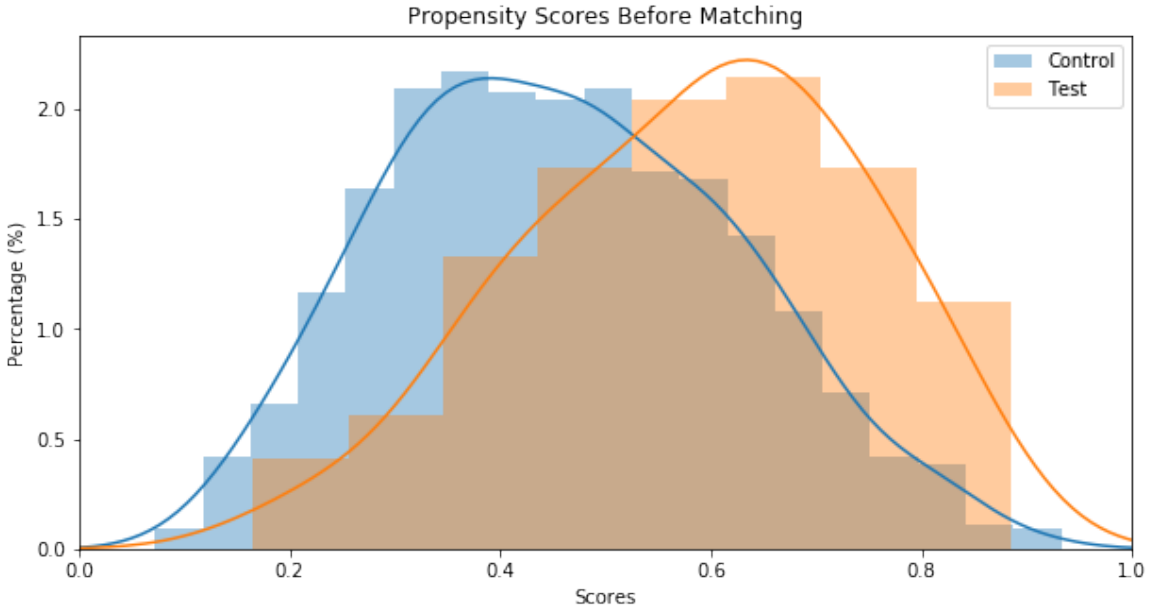


Figure 10: Distribution of propensity scores

We then use the logistic regression as described in 2.2.2 to compute the propensity score of the matching procedure as described in 5.1. In practice we use a manually modified version of the python library [39] to perform the matching procedure. We use the binary variable which indicates to be either female or not as the target variable which leads to the classification task of classifying the instances to be female considering the other covariates as predictors. We use female as the treatment group, because it is the smaller group. We want to find matching partners from the highly overrepresented group of males. In the figure 10 we can see the distribution of propensity scores for male and female instances respectively. The test distribution is the distribution of propensity scores for female persons and the control distribution represents the distribution of

propensity scores for male persons. Fortunately, there is a high intersection between both groups which should lead to a successful matching. To find one to at most five male matching partners for each observation with the attribute female, we find the nearest neighbors of the propensity scores of female and male. We iterate this process five times and accept a matching partner each time the minimal absolute difference of propensity score is smaller than a defined threshold of 0.003.

before matching		
Feature	p-value of bank score version 1	p-value of mail order score version 1
female	0.01	0.00
old	0.00	0.00
east	0.664	0.699
after matching		
female	0.192	0.00
old	0.006	0.00
east	0.857	0.027

Table 4: The p-value of the Mann-Whitey-U-Test for testing the difference of the distribution of both sector scores version 1 along the social features before and after matching

To validate the matched subsample we again use the p-value of the Mann-Whitney-U-Test to check if there are any significant differences in the covariate distributions. As we can see in table 3 the differences in the covariates after matching are not significant anymore which indicates a sufficient matched subsample.

We conclude that the matching worked successfully for all subsamples we wanted to obtain. The resulting data set sample sizes can be found in the appendix A. In the case of the social dimensions sex and age, the sample sizes partly reduced significantly, because of the matching process. In table 4 we can see exemplary the u-statistics p-value for the difference in the bank score and mail order score version 1 distributions along the social features we want to analyze. For the features female and old the p-value is lower than 0.05 while the east feature has a higher p-value. After the matching the distributions differ in some cases. The female feature in the case of the bank score has now a higher p-value than 0.05, thus indicating no difference between scores in this feature anymore, while the east feature has a lower p-value than 0.05, indicating a difference. These subsamples lead to more comparable instances of the social groups we want to analyze in the following section.

6.4. Inferential Analysis

We now formulate a regression problem to analyze the association of social dimensions to the score values depending on the covariates identified in 6.3.1. To reduce estimation bias in the effect of social dimensions on the scores we use the matched subsamples of the openSCHUFA dataset for each feature and each version of the bank score and mail order score. With the hypothesis testing we gathered evidence for indirect discrimination biases of the social dimensions sex regarding the version 1 of both sector scores and age regarding all considered scores. With the analysis of covariance we can strengthen our argument by depending on the covariates and find predictors that potentially mediate the discovered effects. Our goal is to find evidence to verify or falsify the defined hypotheses in 6.2 regarding the openSCHUFA data set.

We now generally view the procedure the SCHUFA applies to produce the score values as the data generating process which we aim to model to approximate the influence of a social dimension on the chosen sector-scores. Since the score value is a continuous variable we use regression methods to find a mapping function that best maps the observed variables V, consisting of the social dimension in question A and the identified covariate variables C, to the score value y. To obtain such approximation we use two modeling approaches. First, we use a parametric model specifically a linear regression 5.2. While this approach is easily interpretable one limitation is the assumed linearity of correlation. We try to select the most relevant feature transformations of the variables to model non-linear correlations with the target variable. We further use a non-parametric approach to obtain a more flexible model with fewer prior assumptions about the type of correlations. We use Gaussian processes as described in 5.3.

To check our concrete hypotheses, we use different quantities that we can obtain from the fitted models. To estimate a discrimination bias we examine the standardized coefficient of the social dimension in the case of the linear regression model or the length scale in the case of Gaussian process regression which we uniformly denote as θ. In both cases we use the standardized data to be able to compare the θ of the social variable with the θ values of the covariates in the model. The rank of the θ of the social variable can be compared between the two approaches. Furthermore, the θ of the standardized data enables us to compare the influence of a social dimension between versions of a sector-score or between sector-scores. We can use this comparability to answer questions about the discrimination bias regarding a specific sector-score as well as the change in the strength of discrimination bias between the versions or the sectors we analyze.

Another measure we want to use to check our hypotheses is the explained variance expressed by R^2. This measure and the standardized $RMSE$ can be also used as model quality measures that are comparable between models. We furthermore use the f-test of overall significance to check if a linear regression model or Gaussian process model is fitting the data significantly better then an intercept-only model.

To examine the influence of the covariates on a social dimension considered, we will examine and compare three models per approach. First, we model the score only with the social dimension and interpret the θ as well as the explained variance R^2 to obtain simple evidence for the question of indirect discrimination. We will refer to this model as a bivariate model. We then compute a model with the identified covariates which we call the covariate model and a model with the covariates and the social dimension in question denoted as the multivariate model. By comparing the R^2 of the two models we can examine the additional variance explained of the social dimension while depending on all covariates. Furthermore we can examine the difference in θ between the bivariate and the multivariate model to see how much of the observed effect in the bivariate model can be explained by the covariates. Because of the usage of the income distribution to formulate hypotheses we also control for the monthly income variable to examine its influence depending on the identified covariates and a social dimension.

In the following section we will describe the practical modeling process and the validation of the resulting models generally and again exemplary with the the bank sector score version 1 to examine the discrimination bias of the social dimension sex in the openSCHUFA data set.

We then interpret the results of the models for the chosen three social dimensions age, sex, east/west states regarding our hypotheses and compare the computed quantities with the more recent versions 2 and 3 of the bank sector score to see if possible effects of social features reduce with higher score version. We also compare the bank-score with the mail order-score in our data set to see differences between sector-scores.

6.4.1. Parametric Analysis

We assume the data generating process to incorporate non-linear relationships. To illustrate evidence for this assumption we show in figure 11 a scatter plot with the score value on the y axis and the age on the x axis. Looking at the dense part of the distribution the relationship between these dimensions seem to be not linear but logarithmic. To take non-linear relationships into account we concentrate on a

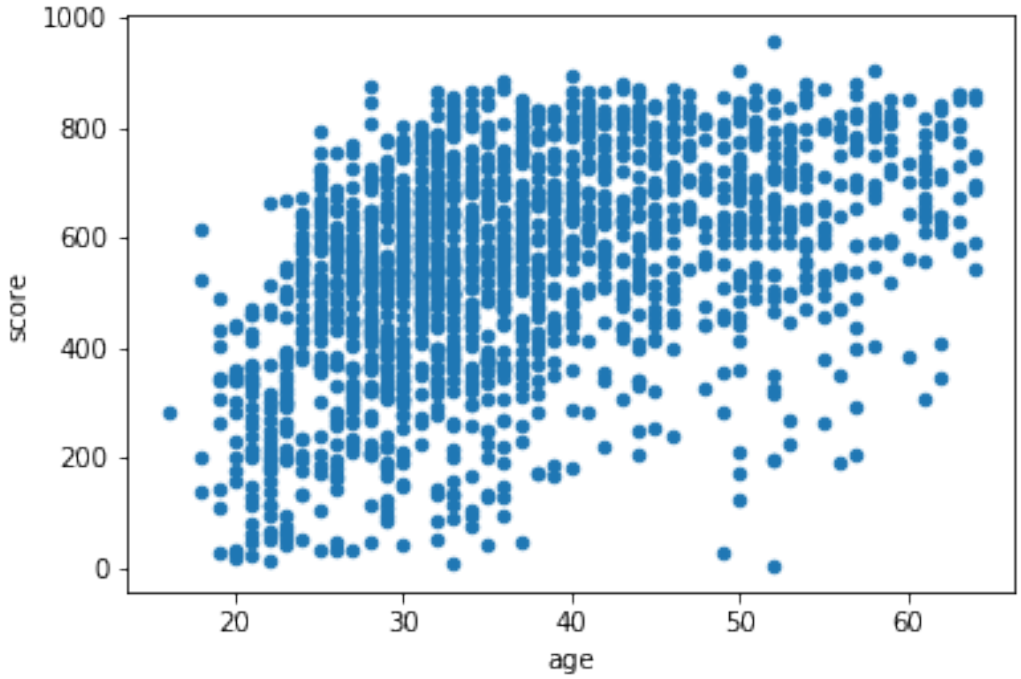


Figure 11: Scatterplot of the bank score version 1 on the y-axis and the age on the x-axis

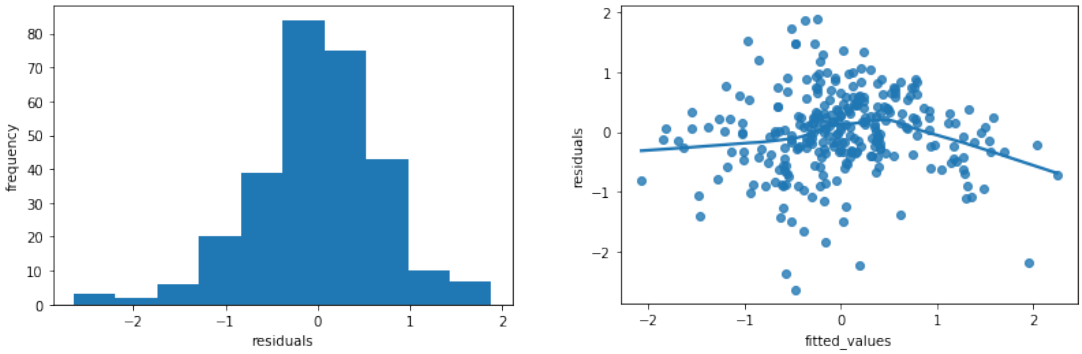
process of feature expansion and selection to quantitatively choose suitable feature transformations for our linear regression model. We apply the transformations log, exp, root and power on the variables to expand the space of possible features. Note that we also implicitly apply the identity-function to the variables to treat themselves as features. This assumeably leads to many irrelevant and redundant features, because we have many transformations of the same variable. To select appropriate features to model the score, we use the LASSO-regression as described in 5.2. Our goal is to find the feature transformation of each variable with highest weight over a defined cut off weight-value. Since the magnitude of the penalty term of the LASSO is governed by a hyperparameter we use cross validation to determine the most fitting value for the hyperparameter given the data observed. This implicates that we do not necessarily find the true transformation of each feature but the transformations that leads to the best estimation performance given the openSCHUFA data set.

We choose a set of possible hyperparameter values. For each hyperparameter value we randomly split the data set into a train and a test set and fit the regression with the train set. We then evaluate the quality of the model by computing the root mean squared error (RMSE) using the test set. In our case we repeat this process of splitting

the data set and evaluating the model 5 times and compute the mean of the errors. The hyperparameter value of the model with the lowest average RMSE is then chosen to fit the whole data set leading to the weights of the features we want to use for feature selection. We use the LASSOCV package of the sklearn library [42] to apply the described procedure. Since the transformations used by the data generating process might change with the sector score and its version we need to repeat this process and select features per sector score and its version. The identified features X are now used for the mapping function $f : X \mapsto y$.

For the bank score version 1 we determined the hyperparameter $\delta \approx 0.001$ and a cut off value of 0.04 due to the reduction of sorted weight values. For each variable we chose the transformation with highest weight. With the fitted linear regression model using [43] we compute the measures R^2, the p-value of the overall model significance and coefficients θ of the features as well as the rank of the social feature as described before. To take advantage of the interpretability of the parametric approach the unstandardized coefficient of the social feature Θ_a is computed to express the increase of the score-units with the increase of the social feature by one unit. Additionally, to examine the uncertainty of θ, we report the standard error (SE) and the p-value of a t-test to check if the coefficient of the social feature significantly differs from zero, which would implicate no discrimination bias.

After we fitted the model to the data we need to check the assumptions of linear regression to obtain reliable and interpretable results. First of all, the assumption of no colinearity in the predictors may be violated, because of highly correlated features. To test for multicolinearity we use the VIF value as an orientation of dependence between predictors. We determine the toleration value of the VIF score to be at most 5 [44]. In our example no VIF-score is above 5 and we assume the obtained standardized coefficients to be reliable. The assumption of random normally distributed residuals around zero is taken into account by plotting the distribution of residuals in figure 12a as well as the residual plot in figure 12b with the predicted score values $\hat{y}$ on the y axis and the residuals e on the x axis. The residuals are seemingly normally distributed and seem to not have any clear patterns which would speak against a random distribution. The mean of the residuals is near 0. Lastly, we use the Durbin-Watson-test to check for autocorrelation between residuals. This is a property the residuals are required to not have. A value of 2 indicates no autocorrelation, while a value of 0 or 4 indicates perfect positive or negative autocorrelation. The observed d-statistic of 1.835 is sufficiently near to 2 [45] , hence we assume no autocorrelation. In conclusion, the assumptions of linear regression seem to be fulfilled for our fitted model and the weights as well as the



(a) Distribution of regression residuals

(b) Residual plot with a line illustrating the lowess-regression fit on the residuals

Figure 12: Plots to examine the residuals of the multivariate regression model of the bank score version 1 to estimate the influence of the feature female

hypothesis tests should be reliable.

This procedure is applied to all matched data sets for each social dimension and sector-score version. In the appendix A the results for each procedure can be found with all covariate effects. Before we concentrate on interpreting the results for each social feature and comparing the results of the versions of the bank score in the next section, we will describe the procedure of deriving models and its quantities with the non-parametric approach.

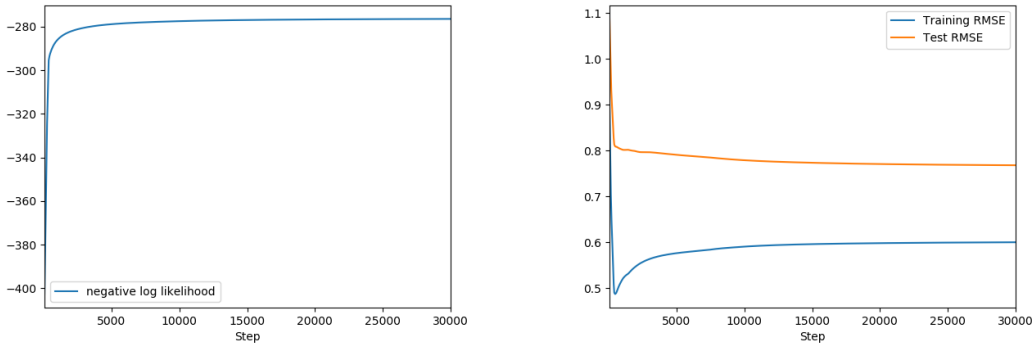
6.4.2. Non-Parametric Analysis

We want to compare the derived models with results from a non-parametric approach which places less assumptions upon relationships. We use the Gaussian process model as described in 5.3.

While the mean function is simply set to the zero-mean function due to the standardization of the data, the covariance function needs to be chosen depending on our data. The covariance is especially important because with the use of ARD length scales, we get a relevance measure for the importance of each feature regarding our score value. Similarly to the choice of the hyperparameter in LASSO-regression we use cross-validation to find the best fitting covariance function by optimizing a Gaussian process with different covariance functions and choose the function that leads to the smallest average test error according to the root mean squared error (RMSE).

We chose the $\frac{3}{2}$-Matern covariance function according to the results of the cross validation. As initial parameters 3 for the length scales was chosen, which leads initially

to less complex functions with fewer variation. The kernel variance was initially set to 0.01 which leads to a generally smaller deviation in the latent function values f and the initial likelihood variance of 3 causes a high noise-level added to f.



(a) Change in Negative log likelihood over 30000 optimization iterations

(b) Change in train and test RMSE over 30000 optimization iterations

Figure 13: Change of optimization and quality measures over 30000 training iterations

Each model was trained by maximizing the negative log-likelihood of the model w.r.t. the hyperparameters with at most 30000 training iterations. In the case of the social dimension sex and the bank score version 1 the train- and test-likelihood as well as the train- and test-RMSE through the training iterations are shown in figure 13. We can see that the train-error is initially low and increases while the test-error decreases which shows the improvement of generalization.

After training the optimized ARD length scales of the Gaussian Process model are transformed to weights by transforming each length scale l_i to $\theta_i = l_i^{-1}$ which we interpret as the weight of variable i for predicting the score y. These need to be interpreted relative to the other length scales in the model. Hence, they cannot be interpreted as easily as in the linear regression approach. To get a measure of relevance we add a reference feature which is randomly drawn from a standard Gaussian distribution $e \sim \mathcal{N}(0, 1)$. We obtain our relevance score r similar to a proposed measure by Ghoshal and Roberts [23].

$$r_i = \frac{\theta_i}{\theta_e}$$

When the weights θ are sorted descendingly, the most relevant variable is shown first. To get a sense of what weight would be uninformative we use the relevance score. A score $r <= 1$ would be less or equally relevant as gaussian noise and a score of $r > 1$ would be more informative than Gaussian noise, which we would interpret as relevant.

Additional to the measures described before 6.4, the r_a of the fitted GP will be reported. To further examine the uncertainty of a Gaussian process model we use the generalization error (GE). The GE is the RMSE when the model predicts a hold-out data set which is in our case 10% of the size of the training set. This measure should indicate if there is a pattern captured which exceeds the training data. We use the GE to report evidence that the learned hyperparameters represent some signal of the underlying data generating process.

Similar to the linear regression analysis, we want to validate the GP models to evaluate the reliability of the model results. In the Bayesian framework we can do so by generating samples of the latent function values with our input-data X. We added a small amount of Gaussian noise to the X matrix. If the properties of the resulting noisy sample-distributions $\hat{y}'$ are similar to the observed data y, we assume the GP model to represent the learned function well. This validation technique is called a posterior predictive check (PPC).

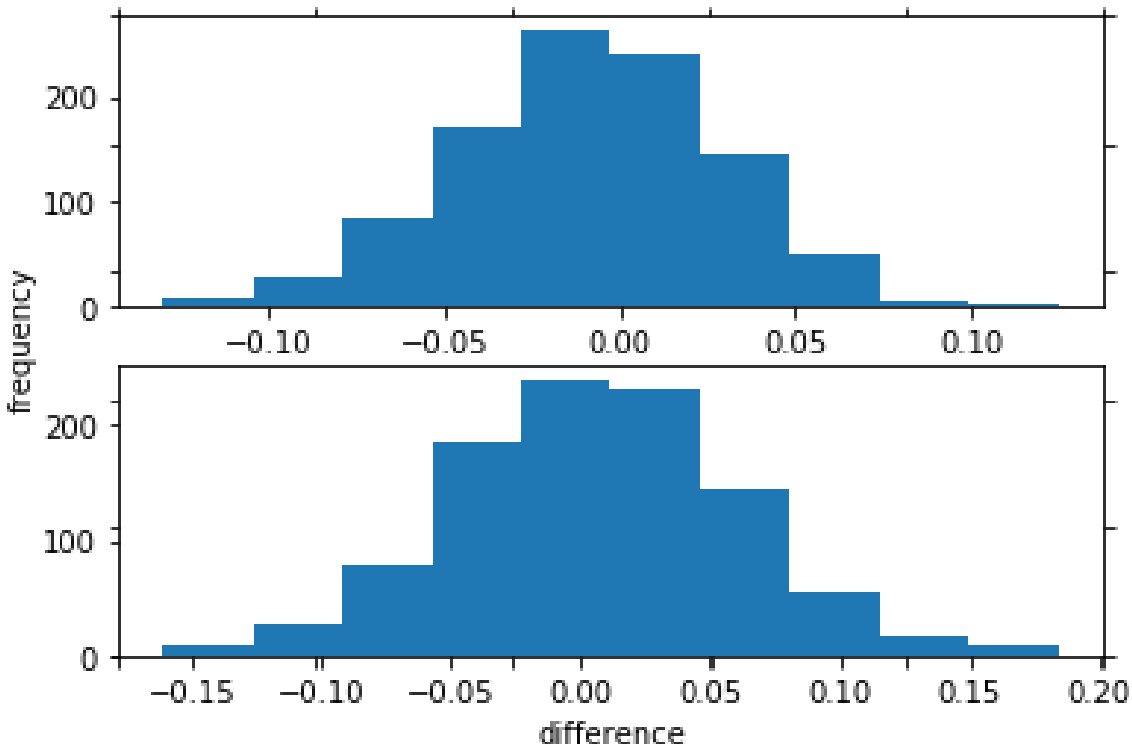


Figure 14: Mean and standard deviation distribution of 1000 noisy samples of the GP

Figure 15: Posterior predictive check plots for the GP to model bank score version 1 to estimate the influence of the feature female

We now illustrate the validation of the model fitted the data set to analyze the

bank-score version 1 in the case of the feature sex. Since we have standardized our data we assume the generated samples of the GP to have a mean of 0 and a standard deviation of 1. We can see in figure 15 that the means indeed center around 0 and the standard deviation centers around 1. For reasons of clarity, we will not examine all other models. The PPC plots of the other models can be found in the appendix A. Generally, all models yield similar results to the example as illustrated above.

6.5. Interpreting the Results

In the last section we explained the modeling process, the obtained measures we want to use to get evidence for possible discrimination bias in the openSCHUFA data set and the validation of the fitted models. We now focus on interpreting the obtained results regarding discrimination bias for each social variable to evaluate our stated hypotheses in 6.2.

In each subsection a chosen social dimension is in focus. We will concentrate on the bank-sector score in this part of the analysis and report for each score version the resulting measures described in 6.4 of the fitted linear regression model. We then interpret the results of the Gaussian process model relative to the linear regression results.

To analyze the change of importance of a social features in the score versions of the data set we use the magnitude of the social feature weight as well as the rank of the feature in the multivariate model. Finally, we compare the change in a social feature weight between the bank sector and mail order sector score according to its magnitude or rank.

We will not show the exact p-values of the hypothesis tests, but denote intervals for the sake of clarity in the following manner:

$0 > x < 0.001$: '***'

$0.001 > x < 0.01$: '**'

$0.01 > x < 0.05$: '*'

$0.05 > x < 0.1$: '_'

$0.1 > x < 1$: ' '

The p-values are the results of the two mentioned hypothesis tests. The interval of the p-value of the t-test is denoted next to the linear regression coefficient that was tested.

We report the p-value to have a notion of confidence in the observed influence. When we speak of a significant coefficient, we mean that the coefficient significantly differs from zero, which resembles the alternative hypothesis of the the t-test of the coefficient. We furthermore denote the p-value interval of the F-test of overall model significance next to the name of the model. When we speak of the model to be significant we mean the alternative hypothesis of the f-test that the model is explaining more variance than an intercept-only model, where all feature coefficients would be zero. This resembles the alternative hypothesis of this F-test. Note, that we only report p-values for the t-test of testing slope test of the social feature of the linear regression and the overall model significance of the bivariate and multivariate linear regression models as well as the multivariate Gaussian process models.

6.5.1. Feature Female

Here we examine the binary feature *female* which indicates if an observation is a female person or not. We use the feature to state about a possible discrimination bias in the bank score version 1 in the openSCHUFA data set along the social dimension sex.

version 1

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate_	0.995	0.01	/	0.189_	34.215	0.112	/
Covariate***	0.702	0.535	0.01	/	/	/	/
Multivariate***	0.686	0.557	0.003	0.284***	51.574	0.079	3/22

Table 5: Linear regression quantities for the effect of the feature female on the bank score version 1

The bivariate model in table 5 with the female feature as the predictor explains approximately 1% of the variance in the bank score version 1. The model has a p-value greater than 0.05 which we would interpret as not significant. Being female has a small effect on the score of 0.189 in our data while it is not significant. The standard error shows the uncertainty in the estimate which indicates that a value near zero is also quite likely. Hence, we are very unsure about a indirect discrimination bias in the simple regression case. Note that in the simple linear regression case, the t-test and

F-test are similar. We still interpret both separately for the sake of consistency with the rest of the models.

Comparing the covariate model with the multivariate model we can see an increase of 2.2% in the variance explained. Note that the increase is bigger than the R^2 of the bivariate model. Similarly, the coefficient of the female feature increases by nearly 0.095 standard deviations of the score which leads to a small to medium effect in the data set. Interestingly, the coefficient of the female feature is now statistically significant. This overall observed increase of relevance of the female feature indicates an association with some of the covariates to explain the score in the data. The unstandardized coefficient Θ_{female} shows an increase in the score value of approximately 52 score points when a person is female. Note that the insignificant effect of monthly income decreases when the female feature is added. This might indicate that the social feature explains some of the association between monthly income and the score in the data.

Hence, the hypothesis of an indirect discrimination bias for the version 1 bank score along the dimension sex, favouring female persons, in our data can be verified, even though we are very uncertain about it according to the bivariate model. The multivariate model shows a significant effect potentially associated with other covariates. Because we do not know which features are truly used to compute the score value, we cannot be sure that the influence is also included in the true score - maybe the SCHUFA uses other features which would not lead to a amplification of the influence of the feature female.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	0.997	0.005	/	/	/	/	/
Covariate	0.578	0.666	/	/	/	/	/
Multivariate***	0.581	0.662	0.768	0.428	0.001	3.763	7/22

Table 6: Gaussian process quantities for the effect of the feature female on the bank score version 1

The results of the Gaussian process approach are reported in table 6. The bivariate model explains only 0.5% according to its R^2. The bivariate model explains only 0.5% of the scores variance. This is only half of the explained variance of the linear regression model. A possible explanation of the difference is the choice of covariance function we had chosen based on the multivariate model. Another explanation is the property of

optimizing the likelihood, which includes a notion of generalization which might have led to a smaller R^2 of the training data.

Comparing the covariate model with the multivariate model we can see in the R^2 that the model quality reduces slightly when adding the feature of being female. This might indicate that the feature does not help explaining the variance in the bank score version 1 when incorporating the covariates. The small drop of variance explained when adding the female feature can also be explained by the added complexity. The generalization error with 0.786 indicates some captured structure and therefore generalizable results, even though it is higher than the training RMSE. Nevertheless, the rank of the the weight θ_{female} shows a rank of 7 out of 22 features incorporated. The relevance score of 3.763 indicates a small amount of information, since the feature is almost 4 times more informative than the Gaussian noise feature. Hence, the model indicates a small effect on the score value, even though it can be considered practically ignorable.

In comparison to the multivariate linear regression model the importance rank has also dropped to the seventh place. Since the overall explained variance according to R^2 is higher then the LR model we conclude to see a small observable discrimination bias for bank score version 1 in favor of female persons. But since the result is not consistent, we are uncertain about this bias.

version 2

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate	0.997	0.006	/	0.146	10.64	0.108	/
Covariate***	0.616	0.606	-0.052	/	/		/
Multivariate***	0.61	0.611	-0.055	0.149*	10.844	0.071	8/22

Table 7: Linear regression quantities for the effect of the feature female on the bank score version 2

While there was some uncertain evidence for a discrimination bias regarding the sex of a person in bank score version 1, the relevance in version 2 seems to drop in the data set. The bivariate regression in table 7 indicates a small but insignificant influence of the female feature. The R^2 is only 0.6% and hence explains less compared to the version 1 score. In the multivariate model we can see that the θ_{female} does not increase much compared to the bivariate model. Because the standard error is smaller in the multivariate model the result is more certain and is still considered significant.

This indicates a small but lower effect compared to version 1. Comparing the R^2 of the covariate and multivariate model we see a small increase. Hence, there is evidence for a small discrimination bias but we are less certain compared to version 1. Θ_{female} shows a score increase of roughly 11 points when a person is female in according to our data set.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	1.0	0.00	/	/	/	/	/
Covariate	0.512	0.74	/	/	/	/	/
Multivariate	0.51	0.74	0.693	0.364	0.00	4.561	14/22

Table 8: Gaussian process quantities for the effect of the feature female on the bank score version 2

The Gaussian process model results reported in table 8 indicate that the female feature does not explain any variance of the score in the bivariate model. The multivariate model seems to not improve when adding the female feature compared to the covariate model. It is more relevent than the Gaussian noise according to the r_{female} of 4.561 and a rank of 14 of 22 features. The generalization error indicates a sufficient generalization relative to the train RMSE.

While the linear regression shows a small but significant coefficient, the GP indicates a practically ignorable bias along the dimension sex. Hence, we observe a small discrimination bias concerning the sex in the data set but the effect is smaller than in version 1 and we are less confident about it given the data.

version 3

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate	1.0	0.002	/	0.093	7.175	0.112	/
Covariate***	0.6	0.67	0.002	/	/	/	/
Multivariate***	0.6	0.67	0.003	-0.033	-2.578	0.069	16/22

Table 9: Linear regression quantities for the effect of the feature female on the bank score version 3

Finally, in version 3 of the bank score in table 10 there seems to be no evidence for an effect of the feature female in the data set. In the bivariate model the explained variance is 0.2% which we consider ignorable and it is not significant. Comparing the covariate model with the multivariate model, there is no observable improvement of model quality according to R^2. The absolute standardized coefficient has no practical significance which is why the switch of sign of the coefficient is not reliable.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	1.0	0.00	/	/	/	/	/
Covariate	0.478	0.772	/	/	/	/	/
Multivariate***	0.478	0.772	0.591	0.333	0.00	1.0	24/24

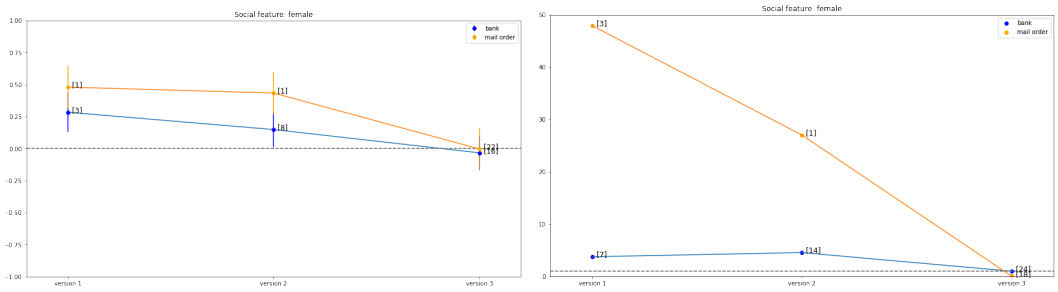
Table 10: Gaussian process quantities for the effect of the feature female on the bank score version 3

The Gaussian process model results in table 10 shows a similar result compared to the linear regression model. The bivariate model does not explain any variance in the score and there is no change in the multivariate model compared to the covariate model. We cannot see any relevance in the feature female as the relevance score r_{female} indicates that being female is not more informative than Gaussian noise. Thus, we are quite confident that there is no discrimination bias in the score along the dimension sex in the openSCHUFA data set.

Comparing the versions

In figure 16 we can see that the relevance of being female drops as the version becomes more recent in the openSCHUFA data set. This illustrates our conclusion that there is a small uncertain discrimination bias in the bank score in version 1 which gets even more uncertain with the version 2 and practically vanishes in version 3. Hence the version of the bank sector score is relevant for the strenght of the effect of the discrimination bias of the sex.

The figure also shows the relevance of the feature of being female in the mail order sector score. We can see that there is much stronger and therefore certain effect compared to the bank sector score. It is also a more consistent effect comparing the models Gaussian process and linear regression model. Interesstingly, the effect seems to get even higher in version 2 and then drops to an ignorable effect in version 3. Hence,



(a) Standardized coefficient of the female features per version of the bank and mail order sector score

(b) Relevance score of the female features per version of the bank and mail order sector score

Figure 16: Plots to illustrate the development of the influence of the female feature

the discrimination bias of the sex increases in the version 2 of the mail order sector score and significantly decreases in version 3.

We conclude that there is a small discrimination bias along the female feature in the openSCHUFA data set. Its strength differs between the version and sector of the score. In the bank score the bias decreases with the version examined and there seems no evidence for the bias in version 3. The mail order score differs with a higher magnitude of weight and because the version 2 seems to have even stronger evidence for discrimination bias than version 1. Though the bias again vanishes with version 3 of the mail order score.

6.5.2. Feature Age

Because the variable age is a metric feature we need to proceed differently in some aspects of the matching process and interpretation of the model quantities. In the section about descriptive analysis we have seen that young to middle aged persons are overrepresented. Because of this observation we binarized the age feature and define older underrepresented people as (41-64) as the treatment and younger persons (15-40) as the control group. Hence, we want to find observations with similar covariates between younger and older persons. In the modeling process we use the metric age variable of the matched observations since the variable contains more detailed information then the binarized feature.

version 1

Model	RMSE	R^2	θ_{income}	θ_{age}	Θ_{age}	SE	$rank$
Bivariate***	0.966	0.054	/	0.192***	4.194	0.047	/
Covariate***	0.81	0.334	0.064	/	/	/	/
Multivariate***	0.785	0.371	0.051	0.192***	3.552	0.041	3/22

Table 11: Linear regression quantities for the effect of the feature age on the bank score version 1

According to the linear regression results in table 11, the bivariate model seems to explain about 5.4% of the variance of the bank score version 1 in the openSCHUFA data set. The model as well as the coefficient are significant. The standardized weight of 0.192 can be interpreted as a small effect. Hence, there seems to be at least an indirect influence of the feature age in our data set regarding the score version 1.

The comparison of the covariate model and the multivariate model shows an increase of 3.7% of explained variance. The insignificant effect of the monthly income drops slightly, while the weight of age remains significant and equal to the bivariate model weight. The rank of 3 out of 22 predictors also strengthens our observation of a small influence of the age on the score version 1. The unstandardized coefficient indicates an increase of approximately 3.6 score points for one unit increase in the age feature. Controlling for the time based finance features regarding the credit history do not lower the effect of the age which could have explained the influence in the bivariate model.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.93	0.135	/	/	/	/	/
Covariate	0.724	0.476	/	/	/	/	/
Multivariate***	0.665	0.559	0.862	0.532	0.056	29.882	1/24

Table 12: Gaussian process quantities for the effect of the feature age on the bank score version 1

The bivariate Gaussian process model in table 12 explains with a R^2 of 13.5% more variance than in the linear regression case. Comparing the covariate with the multivariate model we can observe an increase of 8.3% when adding the age feature. The multivariate model ranks the age feature as the most relevant with a roughly 30 times more relevant than Gaussian noise. While this looks like a very clear influence of the age on the score, the GE is 0.862 which is quite high. Hence, we have a high influence of the age while the error of the model is quite high. This can be interpreted as a higher uncertainty in the model parameter configuration.

To conclude, the models consistently express at least an indirect discrimination bias along the dimension of age where older persons are favored in the data set. Though, we are uncertain about the strength of the discrimination bias.

version 2

Model	RMSE	R^2	θ_{income}	$\theta_{ln(age)}$	$\Theta_{ln(age)}$	SE	$rank$
Bivariate***	0.91	0.154	/	0.392***	110.061	0.045	/
Covariate***	0.757	0.444	0.06	/	/	/	/
Multivariate***	0.689	0.529	0.03	0.306***	85.958	0.037	2/22

Table 13: Linear regression quantities for the effect of the feature age on the bank score version 2

In version 2 of the bank score in table 13 the influence of the age seems even stronger compared to version 1. Note that this might be the case because of the different transformations applied to the age variable in the linear regression model. While the LASSO feature selection indicated the plain age variable as the most fitting feature in version 1, we transformed the feature with the natural logarithm for the score version 2 and 3. Specifically, the explained variance in the score of the bivariate model is 15.4% and the coefficient indicates a significant medium effect of the age feature to the score. This effect reduces when adding the covariates but still is significant and can be considered of medium strength, where the unstandardized $\Theta_{ln(age)}$ indicates an increase of approximately 110 score points when increasing $ln(age)$ by one unit. Comparing the covariate and multivariate model the R^2 increases by 8.5% which is smaller than the variance explained by the bivariate model, thus we assume that some of the influence of the feature age can be explained by some of the covariates in the model.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.858	0.264	/	/	/	/	/
Covariate	0.7	0.51	/	/	/	/	/
Multivariate***	0.575	0.67	0.698	0.415	0.016	38.039	1/24

Table 14: Gaussian Process quantities for the effect of the feature age on the bank score version 2

The GP results in table 14 shows that the bivariate model has a higher explained variance with an R^2 of 26.4% compared to the linear regression model. The increase of explained variance of 16% when adding the age feature again indicates the covariates explaining some of the effect of the age on the score. Since the relevance score r_{age} is about 11 units higher compared to the r_{age} of bank score version 1, we can infer a higher informativeness relative to Gaussian noise compared to the model in version 2. Hence, the log-transformation might not be the only explanation regarding the difference in the relevance of age. The GE which is small relative to the train RMSE also indicates a general pattern captured by the multivariate model.

Hence, there seems to be a discrimination bias in the score version 2 of the data set along the feature age. The uncertainty and significance measures as well as the consistency between the two model approaches indicate a more higher certainty in the result than in the result of version 1.

version 3

Model	RMSE	R^2	θ_{income}	$\theta_{ln(age)}$	$\Theta_{ln(age)}$	SE	$rank$
Bivariate***	0.935	0.108	/	0.332	106.240	0.05	/
Covariate***	0.755	0.422	0.109	/	/	/	/
Multivariate***	0.719	0.477	0.06	0.249	79.828	0.041	2/22

Table 15: Linear regression quantities for the effect of the feature age on the bank score version 3

The results in table 15 of the linear regression show a significant bivariate model explaining 10.8% of the version 3 score and a medium effect of the age feature. Comparing the effect with the multivariate model shows a drop of a still significant effect which can be considered to have small to medium strength on the score. A unit increase of $ln(age)$ yields an increase of about 80 score points. The comparison of the covariate model with the multivariate model shows an increase in R^2 of 5.5% and the age feature is ranked as second most relevant. Note that the coefficient of the monthly income decreases by about 0.05 standard deviations of the score.

The results of the Gaussian process model in table 16 show again a higher variance explained compared to the linear regression in the case of the bivariate model. The

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.914	0.165	/	/	/	/	/
Covariate	0.561	0.686	/	/	/	/	/
Multivariate***	0.413	0.829	0.692	0.373	0.048	65.169	4/24

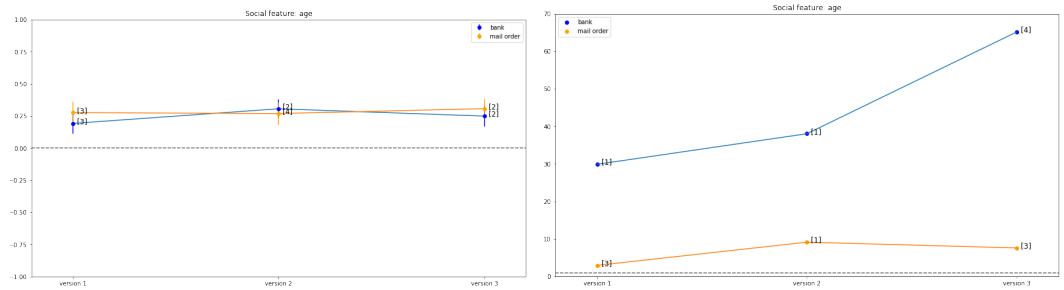
Table 16: Gaussian process quantities for the effect of the feature age on the bank score version 3

explained variance increases by 14.3% when adding the age feature. Because the GE indicates a quite high error relative to the train RMSE, we are more uncertain about the fitted parameters. While the r_{age} is at 65.169 higher than observed before, the relevance drops in rank compared to the model of the version 2 score.

Thus, like the linear regression model, the Gaussian process model also shows that the age feature is not as relevant as in score version 2, though it still certainly indicates a similar discrimination bias along the dimension of age in the openSCHUFA data set.

Comparing the versions

The figure 17 shows the different weight magnitudes between versions of the score.



(a) Standardized coefficient of the age features per version of the bank and mail order sector score

(b) Relevance score of the age features per version of the bank and mail order sector score

Figure 17: Plots to illustrate the development of the influence of the age feature

While the linear regression model weight shows a drop between version 2 and version 3, the relevance score of the Gaussian process model though, is higher for this version then for version 2. This incosistency shows the that we cannot be certain about the magnitude of weights between versions but we can state with high certainty that the bank score has a discrimination bias along age in the openSCHUFA data set.

The figure shows also the difference in score versions of the mail order score in comparison to the bank sector score. We can see that the relevancy of age is in both sector scores on a similar magnitude.

As we have discussed above, there is quite strong evidence for a discrimination bias along the dimension age in the data set. The variable explains much of the variance of the score value in the openSCHUFA data set on its own and because it still significantly improves the two model approaches when adding it to the identified covariates we are confident in our hypothesis that the score is biased along the age in the data set. The linear regression and the descriptive fitting indicate a more positive score the older a person gets. We have to note though that we cannot make any statements about persons older then 64 years, because we don't have enough data about this region of the age.

6.5.3. Feature East

The variable east is a binary feature which indicates whether a person is living in the previously eastern states of Germany or not. The descriptive analysis led us to the assumption that living in the east will at least indirectly lower the SCHUFA score in the openSCHUFA data set as the unequal income distribution of Germany and the distribution of score ratings suggested. Since the quantities derived from our model approaches to analysis this hypothesis do not show any interesting development, we will concentrate our discussion on the version 1 of the bank score.

version 1

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.0	0.00	/	0.039	7.692	0.075	/
Covariate***	0.666	0.569	0.08	/	/	/	/
Multivariate***	0.666	0.569	0.08	0.015	3.021	0.05	15/22

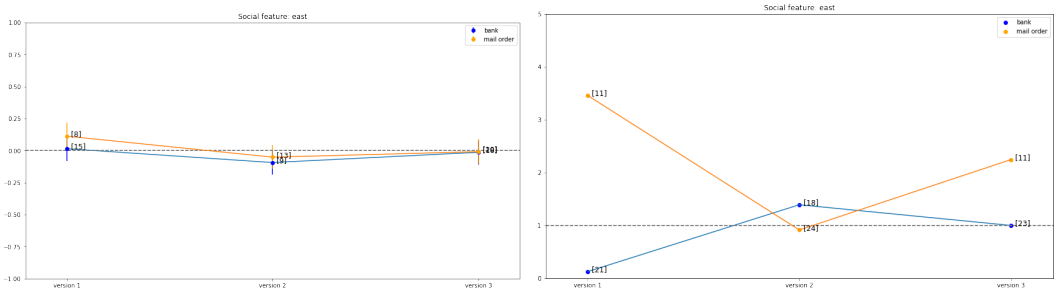
Table 17: Linear regression quantities for the effect of the feature east on the bank score version 1

The linear regression table 17 approach shows no evidence for an influence of the east feature to the score. The bivariate model shows that the feature does not explain variance of the score value and has no significant effect on the score. The comparison of the covariate and multivariate model confirms this observation since there is no increase in the coefficient of determination and the insignificant weight of the east feature drops.

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.0	0.00	/	/	/	/	/
Covariate	0.573	0.672	/	/	/	/	/
Multivariate***	0.573	0.672	0.761	0.36	0.00	0.127	21/24

Table 18: Gaussian process quantities for the effect of the feature east on the bank score version 1

The Gaussian process suggests similar conclusions compared to the linear regression approach. The results in table 17b show again no variance explained in the bivariate model and no increase in the R^2 when adding the east feature to the covariates. The relevance score of the feature is smaller then 1, indicating that the feature is not more informative than Gaussian noise.



(a) Standardized coefficient of the east features per version of the bank and mail order sector score

(b) Relevance score of the east features per version of the bank and mail order sector score

Figure 18: Plots to illustrate the development of the influence of the east feature

Since the other versions of the bank score lead to similar results regarding the two model approaches, we only attach the results to the appendix A and concentrate on the analysis of weights as shown in the figure 18.

Here, we also see no special trend of weight increase or decrease comparing the versions or the score sector of bank and mail order. Hence, we conclude that there is no influence of the east feature and therefore no evidence for discrimination bias whether indirect or direct.

The east feature shows no significant result, that would indicate a discrimination bias along being located in western or eastern states of Germany. Hence, there is no strong evidence for a worse or better score when living in eastern states. Since we have seen in the descriptive analysis that there is a disparity between eastern and

western states along income in our data set, there could have been the possibility of an indirect association of the score with the east feature. But our analysis indicates no informativeness of the east feature neither in the bivariate nor in the multivariate model. Hence, we could not find evidence for a discrimination bias along the east feature. Limitations especially of the analysis of this feature will be pointed out in the next section.

7. Discussion

In this section we will derive conclusions from our observations in the openSCHUFA data set as described in the last section. We point out critical limitations of our results and derivations of our answers. This mostly regards the quality of our data basis and the difficulty of reasoning on a population level, from the data set. Furthermore, the quantities used to get evidence for our questions are also critically reflected. We then will discuss conclusions of our results and possible consequences. Finally, we describe possible further work in the context of discrimination biases in an outlook.

7.1. Limitations

While we in so far have focused on the results of our modeling approaches, we now need to critically reflect the limitations of our analysis and interpretations. This most of all regards to the generalization of our results in terms of the data set and model estimations. Our results and interpretations always refer to the observations of the openSCHUFA data set, which can be interpreted as evidence for specific assumptions about scores of the SCHUFA. However, it cannot be interpreted as results that are true for the population and therefore the true SCHUFA-solvency-scores. Hence, the results are meant to start an evidence based discussion about the derivation of credit worthiness regarding social dimensions. The results are not sufficient for a general statement about the logic of the examined scores.

This is the case because of the data basis itself. Since the SCHUFA report data was obtained from scanned documents they needed to go through an error-prone process. This included optical-character-recoginition (OCR) based software which transformed the data into an unstructured document format and scripts to transform the unstructured data into structured data. Even though the process was accompanied by plausibility tests and obviously implausible data was removed, we cannot be sure about the correctness of the data set.

Another limitation regarding the data set is that it is not representative, especially regarding combinations of attributes. The descriptive analysis of the data set had shown that there is a over-representation of younger persons. Many are students as well as middle aged persons that are employed and married. Also a high proportion of persons live in Berlin.

To avoid the problem of sampling and representation bias we used the matching approach to reduce the bias and to concentrate on the influence of the chosen social dimensions in the context of solvency-scoring. The correct representation of attributes is not important for our question, when we chose an appropriate set of covariates for the matching process. Because the attributes are still differing between matches, we controlled them in the modelling process to take into account the differences between matched groups. This places high importance on the selected covariates. The subsampling of the data set furthermore comes with an important shortcoming, since it reduces the sample size significantly and controlling for the covariates inflates the complexity of our models. Hence, it is more difficult to obtain results with high confidence. In the linear regression we tried to account for that shortcoming by examining not only the weights and coefficient of determination, but also the the standard error and p-values. In the Gaussian process we used the RMSE of test data as an approximation of the generalization error and as a quality measure. An important property of the data we need to note regarding the estimation in our models is the type of many of our covariates. Features such as the number of credit cards or number of loans are count features. Since most features are zero-inflated, having their mode equal to zero, there is a risk of higher estimation errors, since the standardization leads to a higher range to fulfill the properties of having a mean of zero and standard deviation of one.

Especially the failure of correct representation and the matching process leads to the caution of not generalizing interpretations of the results as reflecting the true logic of modeling the credit worthiness by the SCHUFA scores. We can use the result only to observe the influence of the chosen social dimensions on the examined scores in the openSCHUFA data set and obtain at most evidence about the relevance of such social features regarding the scores.

As we have discussed in the previous section, we interpret the observed influence as evidence for a discrimination bias in the examined score. This on the other hand does not answer the question of an indirect or direct discrimination bias. While the bivariate models in our analysis are only used as evidence for indirect discrimination biases, the multivariate models neither can be interpreted as direct discrimination bias evidence.

Even though we control for the identified covariates. This is the case, because we do not know the correct features used to compute the score value. Hence, it is still possible that we missed covariates that would mediate the observed effect of the social feature, leading to an indirect discrimination bias. Our multivariate results can therefore only be interpreted as a stronger support of our observation of some kind of discrimination bias. Beside this general limitation of the interpretation of discrimination bias evidence in our analysis, there are limitations about the social features used to examine influence of the according social dimension on the scores. The east feature for instance was constructed from the first number of the zip-code of a person. The numnber 0 and 1 were classified as indicating eastern states and 2 to 9 indicated western states. The role of Berlin is complicated in this simplified classification and is also overrepresented in the data set. This could lead to wrong estimation of the effect of western and eastern location on the scores examined. Furthermore, the data set was too small for a more detailed analysis of locations. While the more detailed zip codes are too fine grained for an analysis, the used feature might be a too broad approximation of the location of a person and therefore does not represent the social dimension well. Our analysis of the age feature is limited to persons between 18 and 64 years, because we have too few persons to derive meaningful results of younger or older age. We also need to reflect the difference in the subsampling and inference proceeding. While we used a binarized feature to match overrepresented younger persons with underrepresented older persons, we used the metric feature of the matched groups for the analysis of the discrimination bias. This leads to a more fine grained and informative estimation of effect. Hence, while the age differs even between persons of one group younger or older persons, the covariates should be similar in matched groups of a old person and potentially multiple young persons. Therefore, controlling for the covariates is especially important for this analysis.

7.2. Discussion of Results

Keeping the previously discussed limitations in mind, we now want to conclude from our observations and think about possible consequences following from our observations. Generally, the observed results of the analysis of age and sex can be partly viewed as evidence for discrimination bias regarding the examined scores in the data set. The analysis of the age variable showed strong evidence in the data set for a discrimination bias according to the model comparison and weight examination. The older a person gets, the better his or her score will be. As we have seen in the descriptive analysis

of the data set, the age variable is associated with many other variables such as the employment or family status of a person, as well as the monthly income. An empirical correlation might be used as a justification even for a direct influence of the age to model credit worthiness. It may be reasonable to use the age as a ground to model risk to default depending on the sector. If a young person applies for a credit, the risk might be statistically higher to default. This may be justified with the assumption that younger persons probably have a lower income than older persons, because they might be studying or be in apprenticeship. Young persons are also more probable to relocate, which can also be valued as a risk factor. On the other hand, it can be viewed questionable, if a correlation of the age, which is a social attribution that cannot be changed by a person on its own, is a sufficient argument to use it directly. A further implication of the relatively high magnitude of the age variable is the question of how high an influence of a social dimension should be, if there is a reasonable statistical correlation with the target one aims to model. One might agree in a given context, that age is a reasonable feature to model credit worthiness, but this does not answer the question of how relevant this dimension should be at most. Hence, the analysis of the age leads us to the question of when the use of a correlation of a target with a social dimension is over-weighted by the social harm it implies on social groups. Another concern is the indirect bias of the age, because of correlations with finance-based features. The age is most likely correlated with features from the credit history data type. While we tried to incorporate such information, it is likely that we missed features that are actually used. But even the days passed since a person obtained the first credit card is correlated with the age. Hence, features of financial history may cause discrimination bias along age and act as proxies for this social dimension, which may be viewed as reasonable. In this case, the question of the magnitude of the effect of age still holds, but the trade-off between statistical use and social harm shifts to the potentially interacting finance history features in contrast to the indirect influence of the age when modeling the risk to default.

The analysis of the influence of the social dimension sex yielded much weaker evidence for a discrimination bias in the data set. This can be observed by the higher p-values of the linear regression. Another evidence for higher uncertainty is the inconsistency between the Gaussian process model and the linear regression model as well as the small magnitude of the relevance ratio. The result indicated a better score for a person to be female than being male. This observed discrimination bias was stronger in the first version and depending on the sector increased or reduced in the version 2 of the examined scores and eventually almost vanished in the version 3 of the scores.

The different magnitudes of relevance of a social feature can be best observed in this example. The mail order score in the data set evidently had a higher discrimination bias along the dimension sex than the bank score. From these observations we can derive multiple aspects to discuss. First of all, whether the bias is of direct or indirect nature, it changes with the version and drops significantly in the most recent version 3 of the examined scores. The observed development of the influence along the versions of the score may be explained by the change of role allocations between men and women. The social roles may have an influence on the financial behavior of women and men. Social changes are also called exogenous effects [50] and might explain the observed different magnitudes of the bias in the score versions and might be a relevant factor for differences that need to be accounted for. This observation implicates that it makes a difference which contractual partner of the SCHUFA wants to make a decision with the help of a SCHUFA score. If the contractual partner uses an old version, for example of the mail order score, a scored male person might be rejected for some resource but when the contractual partner uses the newest version 3, the scored male person might be accepted for the same resource. As a consequence it is debatable to only allow specific versions of the score to be used according to criteria that need to be developed. One criteria traditionally would be the increase of accuracy. According to our observation it might be also important to take the change in discrimination biases into account as a criterion to prohibit specific versions in a scoring system. Following our observations, the sector of the score is also important to take into account. The different change of magnitude in the versions observed in the data set when comparing the bank sector and the mail order sector scores. Hence, the importance of discarding old versions of a scoring mechanism may be of different priority. According to the observations in the openSCHUFA data set, we may view old versions of the mail order score to be more critical than old versions of the bank score in terms of discrimination biases. Another debatable aspect implicated by our observations is, if a discrimination bias along the social dimension sex is acceptable in general. Like age, the sex is socially attributed and unlike age, it does not change in most cases over time. Interestingly, the logic of favoring a sex is in our example inverted, even though the income and wealth distribution is oppositely directed. This also implies that the discrimination bias is contrary to the historical discrimination against women. Favoring women in such context can even be framed as a desired intervention. Another argument could be the exogenous effects mentioned earlier as a statistically supported rationale to favor women. These different arguments and perspectives about if a discrimination bias is desired or undesired shows the importance of a social debate about these biases in our

view.

The variable indicating a persons lives in eastern or western states of Germany did not lead to strong evidence for a discrimination bias along this location-dependent information in the openSCHUFA data set. Hence, we are very uncertain about its influence on the score. Even an indirect discrimination bias, which we had expected based on the unequal wealth distribution along this dimension, that we also observed in the data set, could not be shown. This might be the case because there actually is no bias regarding this dimension. This might be the case, because the score is not a function of the income. It is also possible that the feature itself is not precisely constructed as discussed in the section about limitations. Causes might be the assignment of Berlin to the eastern states higher complexity of location such as the difference of persons living in the city and persons living at the country side.

The aim of this observational case study was to yield socially relevant results and to illustrate different derivations from the results to start a discussion. The context of the SCHUFA is interesting, because it involves the influence of many decision making processes that affect the access to resources for many consumers in German society. Many consumer have an obvious interest in more transparency about composites of scoring procedures they are affected by. Transparency in this context can be viewed as essential for consumers to enforce their rights in terms of privacy and non-discrimination [50]. On the other hand, there is also interest on the scorers side to increase trust of the consumer to be correctly and fairly scored and to avoid misunderstandings or too shortened understandings about a scoring procedure. But on the other hand we need to note the trade-off between the socially motivated desire of more transparent decision processes and the commercial interest of the company who constructs and uses the score. This was briefly discussed in 3.3.

Transparency regarding discrimination biases may not mean to reveal all aspects of the construction of the score. First of all, it is important to differentiate between indirect discrimination biases and direct discrimination biases. Both are important to be transparent but accompanied by reasoning of a company about why those biases exist. This would most importantly raise awareness about the existence of discrimination biases which otherwise might have been not discovered intentionally or unintentionally. An exemplary transparency process could include a public component to raise trust in the score of the consumer by revealing an ordinal measure of magnitude of a indirect and direct discrimination bias accompanied by a rationale for the bias. Another important information might be the rank of the social dimension relative to the other features. No other feature and no concrete magnitude need to be revealed. The consumer then

can decide, if this bias seems reasonable to her or him.

While our empirical analysis had only the aim to discover potential discrimination biases in some chosen scores of the SCHUFA credit bureau in the observed data set, it did not address the causes of such biases. As we have briefly discussed in 3.2.2, there are diverse causes for discrimination biases which might concern, for example the data basis, the construction of the target or the objective function of the optimization procedure. Hence, a second non-public component of a possible transparency process could be an auditing procedure, which might be performed by an independent third party company or a governmental institution. This procedure could evaluate the scoring procedure under a confidentiality clause. The auditing could be used to analyze possible causes for discrimination biases and hand out recommendations for actions to the company and a test certificate with a rating for the public as a further instance of transparency.

Such an auditing procedure might also lead to an obligation to the use of specific versions of a score. This especially could be the case, when specific social groups have a certain disadvantage compared to others such as a specific sex. The criteria that should be audited do not have to be limited to discrimination biases but can also regard the accuracy of the score as well as other aspects that should be accounted to audit a algorithm in the context of scoring and automating tasks in decision processes.

7.3. Outlook

While there is much research about fairness in machine learning regarding methods to discover or prevent learning algorithms to learn discrimination biases, especially in classification tasks, there is only few work related to applied learning algorithms. Work such as the research of Julia Angwin [33] has shown the importance of empirical studies to see the complexity of such problems. The analysis of the SCHUFA scoring procedure could be another possibility to relate theoretical work on fairness in machine learning with real cases and can also function as a link between perspectives of different scientific and non-scientific fields.

This thesis can be viewed as a starting point on a case-based discussion, because it only shows a glimpse of the aspects that could be analyzed and discussed regarding discrimination bias in algorithm-assisted decision processes. Even our case study just shows a fragment of the whole solvency-scoring procedure of the SCHUFA, since we tried to examine only two out of the 9 sector scores. Furthermore, the analyzed discrimination biases are quite general, because of the exploratory nature of the examination.

Therefore a further step could be to formulate more advanced hypotheses that might

for example aim to examine intersectional discrimination biases, where a combination of specific features is analyzed as interaction terms in a regression model. Such hypotheses could be also formulated as causal relations to examine either mediation of effects or moderation. For example one could test if a feature of credit history mediates the effect of age or it could be tested if the credit history features moderate the effect of age and maybe strengthens the effect further according to the openSCHUFA data set.

While the case study was used to discuss the discovery of potential discrimination biases, one could try to remove the discovered biases with prevention techniques as briefly discussed in 4.1. An important part of the discussion would be the contextual application of such techniques, depending on the bias we want to remove, as well as the fairness criterion defined. The discussion in the conclusion of this thesis has pointed out the difficulty of deciding if a bias should be removed completely or if its strength should be controlled but not removed, because it might be reasonable to do so. For such advanced analysis of the SCHUFA scores the data set might need to be more extensive in the sample size as well as its representation of social groups. Hence, an additional iteration of sampling of SCHUFA reports would be necessary, which is on the other hand difficult, since the current freely accessible SCHUFA report contains even fewer information then the one used in our analysis.

Another problem occurring when we want to further examine the scores in the openSCHUFA data set is the accessibility of the data. Due to reasons of privacy the data set is not publicly accessible. A possible solution could be to use techniques of privacy preserving data mining to make it securely accessible to the data science community.

Another step following this work might be to develop a concept of transparency with respect to discrimination biases. An important challenge on this aspect might be the incorporation of different perspectives about what needs to be transparent and what is the aim of the concept. Another challenge could be to account for the complexity of discrimination biases because of the nature of relationships as correlates and causalities as well as the ethical emphases regarding desired and undesired biases from different points of view.

While we discussed discrimination biases in the context of algorithm-assisted decision making to emphasize the importance of such biases, we finally want to also point out some opportunities following these form of decision procedures and even the biases discovered. First of all it is important to keep in mind that learning algorithms only assist humans in making decisions. These are often based on statistical inference and optimization techniques. They do not discriminate on basis of social biases, but are

dependent on the settings adjusted by a human being or collective of human beings. This process potentially leads to discrimination bias in a resulting model due to faulty specification or normative decisions of the designer of the algorithm. To develop socially desired algorithm-assisted decision process strategies of refinement based on specific criteria can be an important step towards a constructive use of algorithms.

References

[1] S. H. AG. SCHUFA-Branchenscores, 2019. URL https://www.schufa.de/de/unternehmenskunden/leistungen/bonitaet/geschaeft-privatkunden/schufa-branchenscores/.

[2] S. H. AG. SCHUFA: Über uns, 2019. URL https://www.schufa.de/de/ueber-uns/.

[3] A. Altman. Stanford Encyclopedia of Philosophy: Discrimination. Feb. 2011. URL https://plato.stanford.edu/archives/win2016/entries/discrimination/.

[4] S. Barocas and A. Rosenblat. Data & Civil Rights: Technology Primer. page 7, 2014.

[5] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. 2014.

[6] I. Becker. *Datenschutzrechtliche Fragen des SCHUFA Auskunftsverfahrens*. 2006.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [stat]*, Mar. 2017. URL http://arxiv.org/abs/1703.09207. arXiv: 1703.09207.

[8] B. f. p. Bildung. Zahlen und Fakten – Volkszählung/Zensus 2011, 2013. URL http://www.bpb.de/nachschlagen/zahlen-und-fakten/soziale-situation-in-deutschland/169557/themengrafik-demografische-merkmale.

[9] D. Bpb. Datenreport 2018, 2018. URL http://www.bpb.de/nachschlagen/datenreport-2018/.

[10] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling Attribute Effect in Linear Regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 71–80, Dallas, TX, USA, Dec. 2013. IEEE. ISBN 978-0-7695-5108-1. doi: 10.1109/ICDM.2013.114. URL http://ieeexplore.ieee.org/document/6729491/.

[11] J. Cohen. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, N.J, 2nd ed edition, 1988. ISBN 978-0-8058-0283-2.

[12] T. S. community. scipy.stats.mannwhitneyu — SciPy v1.3.0 Reference Guide, May 2019. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html.

[13] O. Dictionaries. English Dictionary, Thesaurus, & grammar help | Oxford Dictionaries, 2019. URL https://en.oxforddictionaries.com/.

[14] R. Dobbe, S. Dean, T. Gilbert, and N. Kohli. A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. *arXiv:1807.00553 [cs, math, stat]*, July 2018. URL http://arxiv.org/abs/1807.00553. arXiv: 1807.00553.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, Apr. 2011. URL http://arxiv.org/abs/1104.3913. arXiv: 1104.3913.

[16] C. Elmer, A. Kruse, M. Pauly, P. Seibt, P. Stotz, U. Köppen, O. Schnuck, R. Schöffel, J. Steule, and M. Zierer. Exklusive Datenauswertung: Blackbox Schufa - SPIEGEL ONLINE - Wirtschaft, 2018. URL https://www.spiegel.de/wirtschaft/service/schufa-so-funktioniert-deutschlands-einflussreichste-auskunftei-a-1239214.html.

[17] S. Finlay. *Credit Scoring, Response Modeling and Insurance Rating*. 2012.

[18] W. E. Forum. How to Prevent Discriminatory Outcomes in Machine Learning. 2018.

[19] R. G. Fryer. An Empirical Analysis of Racial Differences in Police Use of Force. July 2016.

[20] W. Fröhlich and I. Spiecker. Können Algorithmen diskriminieren?, Dec. 2018. URL https://verfassungsblog.de/koennen-algorithmen-diskriminieren/.

[21] B. für Justiz. AGG - nichtamtliches Inhaltsverzeichnis. URL https://www.gesetze-im-internet.de/agg/index.html#BJNR189710006BJNE000100000.

[22] R. Geissler. *Die Sozialstruktur Deutschlands*, volume 7. 2014.

[23] S. Ghoshal and S. Roberts. Extracting predictive information from heterogeneous data streams using Gaussian Processes. *Algorithmic Finance*,

5(1-2):21–30, June 2016. ISSN 21585571, 21576203. doi: 10.3233/AF-160055. URL http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/AF-160055.

[24] C. Giesswein. *Die Verfassungsmaessigkeit des Scoringverfahrens der Schufa*. 2012.

[25] M. Gomolla. Diskriminierung. In *Handbuch Diskriminierung*. 2014.

[26] Guellich. Die Ermittlung relevanter Kundenmerkmale zur Kreditwürdigkeitsprüfung. In *Fuzzy Expertensysteme zur Beurteilung von Kreditrisiken*. 1997.

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. 2 edition, 2008.

[28] D. Hellman. FAT* 2018: Keynote by Deborah Hellman - What is discrimination, when is it wrong and why?, Aug. 2018. URL https://www.youtube.com/watch?v=qomsX8ZvvIY.

[29] U. Hormel and A. Scherr, editors. *Diskriminierung: Grundlagen und Forschungsergebnisse*. VS Verlag für Sozialwissenschaften, Wiesbaden, 1. aufl edition, 2010. ISBN 978-3-531-16657-5.

[30] S. Hradil. *Soziale Ungleichheit in Deutschland*, volume 8. 2001.

[31] S. Hradil. Soziale Ungleichheit: Soziale Schichtung, 2012. URL http://www.bpb.de/politik/grundfragen/deutsche-verhaeltnisse-eine-sozialkunde/138439/soziale-schichtung.

[32] intersoft consulting. EU-Datenschutz-Grundverordnung als übersichtliche Website. URL https://dsgvo-gesetz.de/.

[33] J. L. Julia Angwin. Machine Bias, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[34] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, Sept. 2016. URL http://arxiv.org/abs/1609.05807. arXiv: 1609.05807.

[35] S. Lohr. Facial Recognition Is Accurate, if You're a White Guy. *The New York Times*, Feb. 2018. ISSN 0362-4331. URL https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[36] M. Luca and B. Edelman. Digital discrimination: The case of airbnb.com. 2014.

[37] M. Marquart and A. v. Hove (Grafiken). Umstrittene Bonitätsbewertung: Wie Sie die geheime Schufa-Formel knacken können. *Spiegel Online*, Feb. 2018. URL https://www.spiegel.de/wirtschaft/service/kreditwuerdigkeit-wie-die-schufa-formel-zu-knacken-ist-a-1193522.html.

[38] S. Meisen. Europäische Menschenrechtskonvention. URL https://www.menschenrechtskonvention.eu/.

[39] B. Miroglio. benmiroglio/pymatch, Aug. 2019. URL https://github.com/benmiroglio/pymatch. original-date: 2017-09-20T20:57:05Z.

[40] U. Nations. Universal Declaration of Human Rights, Oct. 2015. URL https://www.un.org/en/universal-declaration-human-rights/.

[41] C. O'Neil, B. d'Alessandro, and T. LaGatta. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. 5(2), 2017.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, page 6, Nov. 2011.

[43] J. Perktold, S. Seabold, and J. Taylor. statsmodels.regression.linear_model.WLS — statsmodels, 2018. URL https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.WLS.html.

[44] J. Perktold, S. Seabold, and J. Taylor. statsmodels.stats.outliers_influence.variance_inflation_factor — statsmodels, 2018. URL https://www.statsmodels.org/devel/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html?highlight=vif.

[45] J. Perktold, S. Seabold, and J. Taylor. statsmodels.stats.stattools.durbin_watson — statsmodels, 2018. URL https://www.statsmodels.org/devel/generated/statsmodels.stats.stattools.durbin_watson.html.

[46] C. U. Press. FAIRNESS | meaning in the Cambridge English Dictionary, 2019. URL https://dictionary.cambridge.org/dictionary/english/fairness.

[47] J. J. Randolph, K. Falbe, A. K. Manuel, and J. L. Balloun. A Step-by-Step Guide to Propensity Score Matching in R. 19(18), Nov. 2014.

[48] B. Rankin and R. Nelson. Automata theory, 2019. URL https://www.britannica.com/topic/automata-theory.

[49] C. E. Rasmussen and K. I. Williams. *Gaussian Processes for Machine Learning*. 2006.

[50] L. Reisch, G. Gigerenzer, and G. G. Wagner. Gutachten - Verbrauchergerechtes Scoring, 2018.

[51] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, Nov. 2014. ISSN 0269-8889, 1469-8005. doi: 10.1017/S0269888913000039. URL https://www.cambridge.org/core/product/identifier/S0269888913000039/type/journal_article.

[52] A. Romei and R. Salvatore. Discrimination Data Analysis: A Multi-disciplinary Bibliography. 2013.

[53] P. R. Rosenbaum and D. B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. 1983.

[54] A. Scherr. *Diskriminierung und soziale Ungleichheit*. 2014.

[55] A. Scherr. *Diskriminierung*. 2016.

[56] M. Spielkamp. OpenSCHUFA, May 2019. URL https://openschufa.de.

[57] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Oct. 2004.

[58] I. Zliobaite. A survey on measuring indirect discrimination in machine learning. 2017.

List of Figures

List of Tables

Appendices

A. Appendix

A.1. Mathematical Conventions

The following mathematical symbols are frequently used:

Symbol	Meaning
n	the sample size of a data set
d	the amount of features of an observation in a data set
V	a set of variables represented as a $n \times d$ Matrix
X	a set of input features represented as a $n \times d$ Matrix
x	a feature observation represented as a $1 \times d$-vector, $x \in X$
x'	another feature observation, $x'\ inX$
Y	a multidimensional set of output labels, also called a target, represented as a $n \times m$-matrix
y	a 1-dimensional set of output labels (target), represented as a $n \times 1$-vector
$X_*,\ y_*$	a training data set
$X_{**},\ y_{**}$	a test data set
A	depending on the context this is a variable or feature representation of a social dimension represented as a $n \times 1$-dimensional vector. $A \in V$ or $A \in X$
C	depending on the context this is a set of variables or features potentially covarying with a social dimension and a target. It is represented as a $n \times o$-dimensional vector. $C \in V$ or $C \in X$
f	is a function that maps some X to y. Depending on the context, this may be also viewed as a random variable
$f(X)$	function values at X.
θ	model parameters or hyperparameters aimed to optimize represented as a $1 \times d$-vector
$L(\theta)$	an optimization function used to optimized the model parameters θ of a model
$e(x)$	propensity score function
$\|\|\theta\|\|_1$	l_1-norm of the parameters θ
$m(x, x')$	a mean-function in the context of Gaussian processes. $m : n \times 1$
$k(x, x')$	a covariance-function (also called kernel) in the context of Gaussian processes. $k : n \times n$
$p(y\|X)$	reads as the probability of y given X
$\mathcal{N}(m, k)$	Normal distribution with a mean-vector m and a covariance-matrix k

Table 19: Description of mathematical symbols and their meaning

A.2. Description of Covariate Variables

This is a brief description of the variable discovery subsection. Most of the listed variables are provided by the structured data set of Spiegel Online.

SCHUFA report

- previous payment problems
  - debt collection: Amount of requests concerning debt collection management
  - debt relief: residual debt discharge announced in a insolvency proceeding
  - insufficient assets: Financial status not sufficient to satisfy creditor
  - solvency: Amount of requests for creditworthiness evaluation
  - undisclosed assets: Debtor failed to provide information on financial status
- credit activity of last year
  - last year credit: Days since first credit activity in the time span of the last 365 days
- use of credit
  - global credit: Count of global credit allowed
  - lease purchase: Amount of contracted lease agreements
  - line of credit: credit line granted
  - rent request: request about tenancy
  - rent signed: Request about
  - secured loan: secured loan concluded
  - unsecured loan: unsecured loan concluded
- length of credit history
  - request: Amount of overall score requests
  - sector request: Amount of score requests of a specific sector
  - banking connection: Request on confirmation of a bank connection
  - business credit card: credit card contract for freelancers concluded
  - business giro: account for freelancers opened
  - business relation: Request on a business relation

  - conditions: condition request
  - credit card: credit card contract concluded
  - giro: account opened
  - mail order: mail order accounts used
  - permanent account: permanent account opened
  - seizure protection account: seizure protection account opened
  - settlement account: settlement account concluded because of contractual violation
  - telecom: telecommunication contract concluded
  - credit card days: Days past since first credit card contract concluded
  - giro days: Days past since first account opened
  - credit days: Days past since first credit contract concluded
- address data
  - address update: request on address update
- general data
  - identity: Request on identity or age verification

Questionnaire

- address data
  - postal code
  - east west: either located in historical eastern or western states of Germany
- general data
  - age of the observation
  - sex of the observation
  - migration background of the observation
  - foreigner: Does not posses a German passport
  - family status: Single, married, divorced, single-parent
  - occupation: Employed, official, freelancer, retired, student, work-seeking
  - relocation: Amount of relocation

A.3. Subsampling Sample-sizes

score	social dimension	sample size
mail order score version 1	age	360 observations
mail order score version 2	age	376 observations
mail order score version 3	age	375 observations
bank score version 1	age	419 observations
bank score version 2	age	410 observations
bank score version 3	age	362 observations
mail order score version 1	sex	260 observations
mail order score version 2	sex	324 observations
mail order score version 3	sex	281 observations
bank score version 1	sex	289 observations
bank score version 2	sex	331 observations
bank score version 3	sex	301 observations
mail order score version 1	eastern, western states	693 observations
mail order score version 2	eastern, western states	734 observations
mail order score version 3	eastern, western states	702 observations
bank score version 1	eastern, western states	700 observations
bank score version 2	eastern, western states	743 observations
bank score version 3	eastern, western states	646 observations

Table 20: Sample size of each obtained matching subsample

A.4. Model Results

This section contains all model results for each the different social features and sector score versions.

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.00	0.00	/	0.039	7.692	0.075	/
Covariate***	0.666	0.569	0.08 ∗ ∗	/	/	/	/
Multivariate***	0.665	0.569	0.08 ∗ ∗	0.015	3.021	0.05	15/22

Table 21: Linear regression quantities for the effect of the feature east on the bank score version 1

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.00	0.001	/	−0.045	7.692	0.073	/
Covariate***	0.638	0.608	0.01	/	/	/	/
Multivariate***	0.637	0.610	0.01	−0.093*	3.021	0.047	9/22

Table 22: Linear regression quantities for the effect of the feature east on the bank score version 2

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.00	0.001	/	0.067	7.692	0.079	/
Covariate***	0.618	0.643	0.07*	/	/	/	/
Multivariate***	0.618	0.643	0.07*	−0.014	3.021	0.048	18/22

Table 23: Linear regression quantities for the effect of the feature east on the bank score version 3

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.00	0.00	/	/	/	/	/
Covariate	0.573	0.672	/	/	/	/	/
Multivariate***	0.573	0.672	0.708	0.389	0.00	0.126	21/22

Table 24: Gaussian process quantities for the influence of the feature east on the bank score version 1

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.00	0.00	/	/	/	/	/
Covariate	0.547	0.701	/	/	/	/	/
Multivariate***	0.546	0.702	0.761	0.36	0.00	1.391	18/22

Table 25: Gaussian process quantities for the influence of the feature east on the bank score version 2

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.00	0.00	/	/	/	/	/
Covariate	0.537	0.712	/	/	/	/	/
Multivariate***	0.537	0.712	0.908	0.363	0.00	1.00	23/24

Table 26: Gaussian process quantities for the influence of the feature east on the bank score version 3

Model	RMSE	R^2	θ_{income}	θ_{age}	Θ_{age}	SE	$rank$
Bivariate***	0.933	0.11	/	0.332***	282.445	0.05	/
Covariate***	0.812	0.338	0.108*	/	/	/	/
Multivariate***	0.762	0.410	0.081_	0.277***	235.417289	0.043	3/22

Table 27: Linear regression quantities for the effect of the age female on the mail order score version 1

Model	RMSE	R^2	θ_{income}	θ_{age}	Θ_{age}	SE	$rank$
Bivariate***	0.951	0.083	/	0.283***	2.788995	0.049	/
Covariate***	0.773	0.400	−0.003	/	/	/	/
Multivariate***	0.726	0.467	−0.025	0.269***	2.648532	0.04	4/22

Table 28: Linear regression quantities for the effect of the feature age on the mail order score version 2

Model	RMSE	R^2	θ_{income}	θ_{age}	Θ_{age}	SE	$rank$
Bivariate***	0.933	0.111	/	0.339***	142.99	0.05	/
Covariate***	0.831	0.314	0.066	/	/	/	/
Multivariate***	0.772	0.399	0.023	0.307***	129.391	0.044	2/22

Table 29: Linear regression quantities for the effect of the feature age on the mail order score version 3

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate***	0.972	0.054	/	0.443***	99.396	0.115	/
Covariate***	0.698	0.49	−0.08	/	/	/	/
Multivariate***	0.660	0.550	−0.078	0.48***	107.493	0.084	1/22

Table 30: Linear regression quantities for the influence of the feature female on the mail order score version 1

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate***	0.980	0.050	/	0.443***	40.912	0.108	/
Covariate***	0.712	0.449	0.093_	/	/	/	/
Multivariate***	0.683	0.494	0.092_	0.434***	40.025	0.084	1/22

Table 31: Linear regression quantities for the influence of the feature female on the mail order score version 2

Model	RMSE	R^2	θ_{income}	θ_{female}	Θ_{female}	SE	$rank$
Bivariate	1.00	0.00	/	0.005	0.571	0.119	/
Covariate***	0.684	0.536	0.134∗	/	/	/	/
Multivariate***	0.684	0.536	0.134∗	−0.003	−0.305	0.085	22/22

Table 32: Linear regression quantities for the influence of the feature female on the mail order score version 3

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate**	0.998	0.013	/	0.232**	52.284	0.076	/
Covariate***	0.687	0.530	0.078∗	/	/	/	/
Multivariate***	0.684	0.533	0.076∗	0.113*	25.47	0.054	8/22

Table 33: Linear regression quantities for the influence of the feature east on the mail order score version 1

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.0	0.00	/	−0.04	−3.933	0.074	/
Covariate***	0.636	0.619	0.029	/	/	/	/
Multivariate***	0.636	0.620	0.03	−0.052	−5.108	0.047	13/22

Table 34: Linear regression quantities for the influence of the feature east on the mail order score version 2

Model	RMSE	R^2	θ_{income}	θ_{east}	Θ_{east}	SE	$rank$
Bivariate	1.0	0.00	/	−0.002	−0.245	0.076	/
Covariate***	0.655	0.576	0.053_	/	/	/	/
Multivariate***	0.655	0.576	0.053_	−0.009	0.977	0.05	20/22

Table 35: Linear regression quantities for the influence of the feature east on the mail order score version 3

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	0.972	0.056	/	/	/	/	/
Covariate	0.590	0.652	/	/	/	/	/
Multivariate***	0.563	0.683	0.69	0.431	0.026	47.933	3/24

Table 36: Gaussian process quantities for the influence of the feature female on the mail order score version 1

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	0.980	0.040	/	/	/	/	/
Covariate	0.584	0.659	/	/	/	/	/
Multivariate***	0.487	0.763	0.596	0.369	0.044	27.015	1/24

Table 37: Gaussian process quantities for the influence of the feature female on the mail order score version 2

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{female}	r_{female}	$rank$
Bivariate	1.0	0.00	/	/	/	/	/
Covariate	0.43	0.815	/	/	/	/	/
Multivariate***	0.43	0.815	0.845	0.362	0.00	0.035	18/24

Table 38: Gaussian process quantities for the influence of the feature female on the mail order score version 3

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.913	0.166	/	/	/	/	/
Covariate	0.677	0.542	/	/	/	/	/
Multivariate***	0.656	0.570	0.683	0.552	0.011	2.941	6/24

Table 39: Gaussian process quantities for the influence of the feature age on the mail order score version 1

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.919	0.155	/	/	/	/	/
Covariate	0.678	0.54	/	/	/	/	/
Multivariate***	0.580	0.663	0.703	0.494	0.033	9.12	1/24

Table 40: Gaussian process quantities for the influence of the feature age on the mail order score version 2

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{age}	r_{age}	$rank$
Bivariate	0.583	0.172	/	/	/	/	/
Covariate	0.677	0.542	/	/	/	/	/
Multivariate***	0.583	0.66	0.873	0.511	0.017	1.185	3/24

Table 41: Gaussian process quantities for the influence of the feature age on the mail order score version 3

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	0.997	0.005	/	/	/	/	/
Covariate	0.611	0.627	/	/	/	/	/
Multivariate***	0.609	0.629	0.75	0.437	0.0	3.456	11/24

Table 42: Gaussian process quantities for the influence of the feature east on the mail order score version 1

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.00	0.00	/	/	/	/	/
Covariate	0.539	0.71	/	/	/	/	/
Multivariate***	0.538	0.71	0.655	0.355	0.0	0.913	24/24

Table 43: Gaussian process quantities for the influence of the feature east on the mail order score version 2

Model	RMSE	R^2	GE	$\bar{\sigma}$	θ_{east}	r_{east}	$rank$
Bivariate	1.00	0.00	/	/	/	/	/
Covariate	0.570	0.676	/	/	/	/	/
Multivariate***	0.556	0.690	0.782	0.401	0.002	2.244	11/24

Table 44: Gaussian process quantities for the influence of the feature east on the mail order score version 3

A.5. Digital Appendix

Additional results not reported in the main part of the thesis can be found in the digital version of the appendix.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den September 13, 2019 .....