When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes

CHRISTOPH KERN, LMU Munich, Munich, Germany, Munich Center for Machine Learning (MCML), Munich, Germany, and University of Mannheim, Mannheim, Germany
RUBEN BACH, University of Mannheim, Mannheim, Germany
HANNAH MAUTNER, dmTECH, Karlsruhe, Germany
FRAUKE KREUTER, LMU Munich, Munich, Germany, Munich Center for Machine Learning (MCML), Munich, Germany, and University of Maryland, College Park, USA

Algorithmic profiling is increasingly used in the public sector with the hope of allocating limited public resources more effectively and objectively. One example is the prediction-based profiling of job seekers to guide the allocation of support measures by public employment services. However, empirical evaluations of potential side-effects such as unintended discrimination and fairness concerns are rare in this context. We systematically compare and evaluate statistical models for predicting job seekers' risk of becoming long-term unemployed concerning subgroup prediction performance, fairness metrics, and vulnerabilities to data analysis decisions. Focusing on Germany as a use case, we evaluate profiling models under realistic conditions using large-scale administrative data. We show that despite achieving high prediction performance on average, profiling models can be considerably less accurate for vulnerable social subgroups. In this setting, different classification policies can have very different fairness implications. We therefore call for rigorous auditing processes before such models are put to practice.

CCS Concepts: • **Applied computing** → **Sociology**; • **Security and privacy** → *Social aspects of security and privacy*; • **Human-centered computing** → *Empirical studies in HCI;*

Additional Key Words and Phrases: Algorithmic fairness, modeling decisions, statistical profiling, machine learning pipeline

ACM Reference Format:
Christoph Kern, Ruben Bach, Hannah Mautner, and Frauke Kreuter. 2024. When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes. *ACM J. Responsib. Comput.* 1, 4, Article 24 (November 2024), 30 pages. https://doi.org/10.1145/3689485

1 Introduction

Policymakers in public administration increasingly seek support from algorithmic decision-making systems to enhance the efficiency and effectiveness of government spending. Numerous

Authors' Contact Information: Christoph Kern, LMU Munich, Munich, Germany, Munich Center for Machine Learning (MCML), Munich, Germany, and University of Mannheim, Mannheim, Germany; e-mail: christoph.kern@stat.uni-muenchen.de; Ruben Bach, University of Mannheim, Mannheim, Baden-Württemberg, Germany; e-mail: r.bach@uni-mannheim.de; Hannah Mautner, dmTECH, Karlsruhe, Germany; e-mail: ha.mautner@dm.de; Frauke Kreuter, LMU Munich, Munich, Germany, Munich Center for Machine Learning (MCML), Munich, Germany, and University of Maryland, College Park, USA; e-mail: Frauke.Kreuter@stat.uni-muenchen.de.

ACM 2832-0565/2024/11-ART24
https://doi.org/10.1145/3689485

examples are documented in the literature. For example, in criminal justice systems algorithms support the allocation of intervention and supervision resources [4, 45]. Child protection services use algorithms to target risky cases and to allocate resources such as home inspections to identify and control health hazards [24, 77]. Immigration and border control use algorithms to filter and sort applicants seeking residence in the country [60]. Public employment services use algorithms to identify job seekers who may find it difficult to resume work and to allocate support programs [57].

A typical task in such automated or **algorithmic decision-making (ADM)** systems is to assess the risk that some event will take place and to recommend some preventive action for cases in a specific risk group (e.g., those with the highest risk scores). In the examples above, relevant events could be violent recidivism among convicted offenders, removal of a child from its home due to maltreatment, a fraudulent immigration application being granted, and a job seeker not resuming work for a long time. Those within the highest decile of risk scores, for instance, could then be recommended for support by a social worker (recidivism and child maltreatment), for an in-depth review by a human officer, or for participation in an active labor market policy program.

Research investigating human vs. statistical prediction has shown that statistical models and algorithms are often more accurate in estimating the risk of events such as academic failure, job performance, recidivism, psychiatric conditions, and long-term unemployment [see, e.g., 8, 25, 41, 62, 68]. Millions of data points resulting from increasingly digitized administrative processes paired with powerful machine learning models suggest that the performance gap between human and algorithmic risk assessment may increase in the future. Thus, ADM systems backed by statistical models and algorithms may enhance government efficiency and public service delivery by being more accurate in identifying those at risk [61].

However, concerns have been raised that ADM may result in unintended social, ethical, and legal consequences [11, 61]. An increasing number of scholars point out that ADM may foster existing biases or even introduce new ones by treating groups of people differently based on ascribed characteristics such as gender or ethnicity [11, 56, 69]. As ADM systems are typically fed with historical training data (e.g., past court or hiring decisions), biases in these data can be learned and replicated by the prediction models. As a result, algorithms may treat specific societal groups differently than others, or, in other words, algorithms may learn unfair association rules. Thus, while ADM systems may make more accurate risk assessments and, supposedly, be neutral and objective due to reducing human judgment in assessing risks, they nonetheless learn from data that may be full of biases and discrimination that was present when the data was generated. Moreover, an accurate risk assessment is only the first step in a typical ADM system, and additional biases may manifest when a decision is made based on the risk assessment [40, 58]. In other words, promises that technical solutions are consistent, neutral, and objective may not hold. Algorithmic risk assessments may be faster and more accurate than humans, but the resulting decision may likewise be biased and unfair.

The field of **fairness in machine learning (fairML)** has made considerable progress in proposing fairness notions and metrics to assess biases of prediction models [10, 65, 69, 70]. As the development of fairML methodology is often centered around a limited number of benchmark data sets [34], their systematic application in real-world scenarios, however, lags. This is particularly the case for ADM applications in labor market contexts as agencies may not disclose detailed documentation of their profiling models and data access is restricted. Nonetheless, ADM approaches such as the AMAS model to classify job seekers in Austria [44] have received considerable public attention due to concerns of algorithmic biases. Following preliminary work on fairness implications of algorithmic profiling of job seekers [2, 28], we set out to conduct a systematic fairness evaluation of profiling models using real-world administrative data with labor market histories of over 300,000 German job seekers.

Profiling of the unemployed is a particularly interesting use case for such an evaluation. **Long-term unemployment (LTU)**, that is, unemployment that lasts for more than 12 months, is a major societal challenge in many countries [32]. It has serious consequences for individuals not only in terms of economic deprivation but also for physical and mental health and overall well-being and it is one of the main causes of persistent poverty [1, 39, 55]. On the macro societal level, LTU is associated with high costs for health care systems and welfare services [64]. In Germany, for example, the share of LTU among all unemployed has decreased from 56% in 2007 to 28% in April 2020, but it remains a major social challenge [18, 19]. Facing limited resources, many **public employment services (PES)** apply profiling to improve the efficiency of social spending [57, 63]. Profiling is used to assess the chances of unemployed people to resume work. PES may then tailor their activities to specific individuals, for example, to those who are predicted to struggle with finding new employment. Profiling assesses a newly unemployed person's risk of LTU, that is, that (s)he will stay unemployed for more than 12 months. It is used at entry into unemployment such that a PES caseworker can intervene early on and, e.g., support individuals at risk of LTU in resuming work through targeted **active labor market policies (ALMP)**. ALMP are activation strategies such as vocational training, hiring subsidies for employers, and job creation schemes that are aimed at enabling unemployed individuals to quickly resume work [47]. In Germany, PES spending for ALMP measures summed up to a total of 4 billion EUR in 2021 [20].

Implementing an algorithmic profiling system to target job seekers in practice involves many critical design decisions, however [75, 81]. Questions that need to be answered include, for example, what type of prediction method should be applied? Which type of information should be used for model training? How should resources be allocated based on a prediction model's outputs? Eventually, such decisions can substantially affect the extent to which different societal groups are targeted or reached by support programs and public services. This especially includes the risk of perpetuating discrimination against historically disadvantaged groups. The AMAS profiling model that was built to classify job seekers in Austria, for example, exhibited a negative effect of being female on short-term re-employment propensities [44]. Based on such a model, different classification policies could be applied under which female job seekers could have higher (prioritize job seekers with low predicted re-employment propensities) or lower (prioritize job seekers with medium predicted re-employment propensities) chances of receiving extensive support by employment agencies, compared to their male counterparts.

Against this background, we compare and evaluate algorithmic profiling models for predicting job seekers' risk of becoming long-term unemployed (LTU) concerning (subgroup) prediction performance, fairness metrics, and vulnerabilities to data analysis decisions in this study. Focusing on Germany as a use case, we evaluate profiling models by utilizing administrative data on job seekers' employment histories that are routinely collected by German public employment services. Our contribution to the literature on algorithmic profiling and fairness in profiling is twofold: (1) We conduct a systematic *fairness auditing* of different prediction models and report on the implications of implementing algorithmic profiling of job seekers in Germany under realistic conditions. (2) We evaluate fairness implications of *design decisions* such as using different model types, classification thresholds and training data histories in the profiling context. This analysis shows how modeling decisions along the prediction pipeline can have group-specific downstream effects with a focus on the eventual allocation of support measures.

We use regression and machine learning techniques, specifically, logistic regression, penalized logistic regression, random forests, and gradient boosting to build profiling models. For each technique, we train multiple sets of prediction models that differ in the time frame and features that are used for model training. For each model, three classification policies for prioritizing job seekers are implemented that focus on very high, high, and medium predicted risks of LTU. Next to comparing the profiling models concerning group-specific prediction performance, we study the fairness implications of the models' classifications based on (conditional) statistical parity difference, false negative rate difference, and consistency in two evaluation data sets. We focus on four groups of job seekers: Female, non-German (i.e., foreign-born), female non-German, and male non-German individuals. A large body of literature shows evidence of discrimination in the labor market concerning gender [14] and ethnicity [87], which is likely to be reflected in historical labor market records and thus may be learned by a prediction model. Our fairness evaluation therefore aims to study whether discrimination against these groups could be perpetuated or mitigated under a given algorithmic profiling scheme.

This paper is structured as follows: Section 2 provides a brief introduction to statistical profiling implementations in various countries (Section 2.1) and discusses fairness concerns in the context of algorithmic profiling of job seekers (Section 2.2). Section 3 presents the data (Section 3.1) and the prediction setup including the range of modeling choices (Section 3.2) that are considered for predicting LTU in our empirical application. The results are summarized in Section 4, which includes the evaluation of subgroup prediction performance (Section 4.1) and fairness metrics (Section 4.2). We discuss our findings in Section 5.

2 Background

2.1 Statistical Profiling Implementations

Statistical profiling approaches used by **public employment services (PES)** across the globe typically aim to fight LTU by *preventing* it through identifying those at risk of becoming long-term unemployed at an early stage [63]. That is, statistical profiling approaches usually estimate the risk that a person who recently lost their job will remain unemployed for a predetermined period, such as 12 months. Based on the estimated risk, job seekers are segmented into risk groups, and support by PES is then determined based on the risk group an individual belongs to.

A variety of statistical profiling systems were developed in several countries. Comprehensive reviews of existing profiling implementations are presented in Loxha and Morgandi [63], Desiere et al. [27] and Körtner and Bonoli [57]. In summary, profiling approaches, are evaluated, tested, or used, for example, in Australia [21, 63, 66], Austria [44], Belgium [28], Denmark [21], Finland [84], Ireland [73, 74], the Netherlands [85], New Zealand [27], Poland [72], Portugal [26], Sweden [7], and the U.S. [15].

The design and implementation of risk assessments vary considerably by country, however. Some approaches are aimed at predicting LTU (e.g., Belgium, Denmark, and the Netherlands), while others assess the likelihood of exit into employment (e.g., Ireland). Similar variation exists regarding the statistical models used. Examples are logistic regression models (e.g., Italy, Netherlands, Sweden) and popular machine learning algorithms, such as random forests and gradient boosting [e.g., Belgium and New Zealand, 27]. Typically, administrative labor market data are used as training data, but information collected from surveys is also used in some countries. Implementations also differ in their in- or exclusion of sensitive characteristics such as gender during model training (see Section 2.2 below). Due to differences in labor market policy and legislative frameworks, there is also considerable variation regarding the question of which risk groups are targeted by PES, based on their estimated risk scores. Many countries appear to target unemployed individuals with a high LTU risk [27]. In Austria, however, algorithmic profiling was supposed to be used to aim PES activities at unemployed individuals with a medium risk of LTU [2].

2.2 Fairness Concerns in Statistical Profiling

Although profiling approaches have been around for almost thirty years, concerns about the unequal treatment of job seekers based on ascribed characteristics such as gender and ethnicity have only recently caught attention. In the labor market setting, it is little surprising that fairness can quickly become a challenge in statistical profiling as numerous studies have shown that women and individuals with a migration background are disproportionately affected by unemployment and have lower job prospects [for Germany, see 9, 48, 54]. There is consistent experimental evidence that part of these differences can be attributed to statistical (stereotyping based on assumed group averages) and taste-based (prejudice against minority groups) discrimination in hiring decisions [71]. It is important to not study both attributes in isolation: In the German case, an additional ethnic disadvantage in labor market participation of women can be observed for specific groups of migrants (such as first-generation immigrants from less developed countries, [36]). Focusing specifically on discrimination in hiring, other studies suggest that ethnic minority men are particularly disadvantaged [6, 29].

Fairness notions. As discriminatory practices are manifested in (un)employment histories of women and migrants, prediction models trained with such data can pick up and incorporate historical bias [69, 70]. Moreover, even if sensitive characteristics of job seekers are not explicitly used for model training ("fairness through unawareness" [69]), predictions can nonetheless be affected. If labor market histories of, for example, women and men are distinct, then it is likely that an algorithm will learn different patterns and risks for women and men based on the correlation of gender and labor market histories. The degree and implications of such learned differences depend on the design and use of the broader profiling system [40, 59] as well as on the modeling decisions made in the implementation of the prediction model [75, 80] and thus need to be studied in context.

The fairness in machine learning literature has proposed numerous fairness notions and metrics to assess and quantify disparate social impacts of prediction algorithms. Fairness notions are commonly conceptualized on the group-, individual-, or multi/sub-group level [10]: Group-based fairness notions compare model outputs between groups commonly defined by protected attributes (such as gender and ethnicity) to, e.g., identify disparate model error [42, 65]. Individual fairness notions commonly require that individuals who are similar regarding the predictions task at hand should receive similar predictions [33]. Multi-group fairness imposes fairness requirements on large collections of subpopulations that may be defined by intersections of various protected and non-protected attributes [43, 51, 52]. Despite their apparent mechanistic differences, group and individual fairness can be motivated under the same normative principles [13].

Previous auditing studies. Fairness evaluations of profiling systems of the unemployed have not been discussed much until recently. Allhutter et al. [2, 3] conduct a document analysis with a focus on fairness concerns in the Austrian statistical profiling tool *AMAS*. This tool is based on a stratification procedure to assess short-term and long-term job prospects based on, among other variables, age, gender, citizenship, and health impairment [2]. Based on the predicted integration scores, job seekers are placed in one of three job prospects groups.[1] According to [2], those with mediocre job prospects are the focus of PES' measures to increase re-employment chances. Those in the highest group receive less intensive support from the PES as they are assumed to resume employment even without strong support and those in the lowest group are mostly referred to an external institution. Based on the stratification procedure, people of higher age, female gender, non-EU citizenship, or people with health impairment, are predicted lower prospects of finding a job in the short term. That is, in the AMAS algorithm ascribed (or protected) characteristics

[1]Specifically, the two underlying models aim to predict the likelihood of a jobseeker to find employment for at least 3 months within the next 7 months (short-term perspective) and the likelihood to find employment for at least 6 months within the next 2 years (long-term perspective) [2]. Rather than predicting LTU, the AMAS thus focused on the "inverted" task of predicting successful labour market integration along different time horizons.

of job seekers affect their integration score and thus potentially their chance of receiving support measures. As Allhutter et al. [2, p. 7] put it, "previously discriminated or marginalized groups are more likely to be classified as part of group C [the low job prospects group], which in turn reinforces existing inequalities as more discriminated populations of job seekers are more likely to receive less support." Alleged discrimination of this system was contested by the Austrian PES [17].

Desiere and Struyven [28] investigate fairness aspects of the statistical profiling system used by the Flemish PES *VDAB*. They document that job seekers belonging to historically disadvantaged groups such as migrants, disabled, and older age groups are more often incorrectly classified as high risk of LTU (here, unemployment that lasts for more than six months). Although the statistical profiling approach is more accurate in predicting LTU than a simple rule-based approach, it also shows more discrimination (defined as the ratio of false positive rates between groups) towards the aforementioned groups. This is the case even though sensitive characteristics are explicitly not included in the model. At the same time, discrimination depends on the threshold used to determine whether someone is high-risk or low-risk. For more restrictive thresholds, Desiere and Struyven find that discrimination against minority groups is highest. In this case, a large share of the job seekers with a high predicted risk is of foreign origin.

Building on these initial results, we set out to systematically investigate fairness concerns and their dependence on modeling decisions in algorithmic profiling of the unemployed. Before turning to our fairness evaluation, we first describe the data and prediction pipeline in detail.

3 Methods

3.1 Data

We use German administrative labor market records that we obtained from the Research Data Center of the German Federal Employment Agency at the **Institute for Employment Research (IAB)**. These data contain historical records of labor market activities (employment, unemployment, job search activities, and benefit receipt) for the majority of the German population [about 80% of the German labor force, 31]. Self-employed individuals and civil servants are not included as they are managed by a different institution [49]. The records go back as far as 1975 and cover all individuals who meet at least one of the following conditions in Germany: at least once in employment subject to social security (records start in 1975) or in marginal part-time employment (records start in 1999); received short-term unemployment benefits or participated in labor market measures under the German Social Code Book III (records start in 1975); received long-term benefits under the German Social Code Book II (records start in 2005); registered with the German PES as a job seeker (records start in 1997); participated in an employment or training measure (records start in 2000) [5]. Information is exact to the day and allows to create detailed labor market histories of individuals.

We use a 2% random sample of these records, the ***Sample of Integrated Employment Biographies*** [SIAB, 5]. It combines information from multiple sources, resulting in a dataset with detailed employment and unemployment information as well as unemployment benefits receipts (see previous paragraph). We use the factually anonymous version of the SIAB (SIAB-Regionalfile) – Version 7517 v1,[2] which was stripped of potentially sensitive information due to privacy regulations. Nonetheless, it is still well suited for predicting LTU due to the number of records and their granularity: detailed employment histories of 1,827,903 individuals are documented in a total of 62,340,521 rows of data.

[2]Data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

Our dataset comes in longitudinal form and often contains multiple entries per person. That is, each time a person's labor market status (e.g., registered as unemployed or started a job subject to social security) changes, a new entry is created. On average, we observe more than 34 data points for each of the nearly two million individuals. It is also possible that we observe only one entry for an individual, for example, if she was employed without any interruptions by the same employer. Depending on the type (e.g., employment episode, unemployment episode or benefit receipt episode) of an entry, socio-demographic characteristics such as age, gender, education, and occupation as well as information on the duration of the episode (e.g., duration of unemployment), information on income and industry (for employment episodes), information on participation in PES' sponsored training measures (for training measures episodes), or information on job search activities (for unemployment episodes) are available.

We restrict the SIAB data to include data points from the period between January 1, 2010, and December 31, 2016. We exclude data referring to periods *before 2010* as German legislators introduced fundamental labor market reforms between 2002 and 2005, which resulted in major socio-cultural, but also institutional changes in German labor market policies and fundamentally changed the way how unemployed people were supported by the German PES. In addition, new types of data were added to the SIAB during that time that challenged data comparability across longer periods.

Data collected *after 2016* are excluded because our objective is to predict unemployment that lasts for at least one year. Therefore, the last year of labor market histories available is needed to determine whether individuals who became unemployed by the end of 2016 became long-term unemployed or not. While one could include unemployment periods that started after 2016 but ended before December 31, 2017, it would introduce inconsistencies as we would obtain only non-LTU episodes in 2017 but no LTU episodes due to the right censoring of the data in December 2017.

In addition, we removed all individuals who never became unemployed during the period of observation. Since we predict LTU, individuals who were never either LTU or non-LTU would be irrelevant. These restrictions leave us with 303,724 unique individuals and 643,690 unemployment episodes.[3]

3.1.1 Definition of Long-term Unemployment. Our prediction outcome follows the definition of LTU employed by the German PES. According to the German Social Code Book III, article 18/1, individuals are long-term unemployed if they are continuously unemployed for more than one year. The same threshold is applied by Australia, Italy, and the Netherlands, among others [27]. Participation in labor market measures as well as periods of sickness or interruptions for other reasons of up to six weeks do not count as interruptions of an unemployment period.

LTU is therefore identified if a data point refers to an unemployment episode with a recorded length of more than one year.[4] If an unemployment episode's duration is less than one year, we define it as non-LTU. Unemployment periods are recorded in the administrative labor market data once an individual registers as unemployed with the PES. Therefore, they allow us to identify the exact date a person presents herself as unemployed to the PES. As the SIAB data also records the end date of an unemployment episode, we can recover the exact duration of an unemployment period and therefore identify LTU.

[3]Note that individuals can contribute more than one unemployment episode to our data as they may become unemployed more than once during the period of observation.

[4]Relevant episodes are those flagged as "job seeking while unemployed" and "job seeking while not unemployed" if "not unemployed" is caused by a parallel episode of participation in a PES labor market measure (German Social Code Book III, article 18 in combination with article 16).

Based on the definition from above, 97,599 (15.2%) out of a total of 643,690 unemployment episodes identified in the data are LTU episodes. We find that 79,361 (26.1%) out of the 303,724 individuals in our data who ever became unemployed between 2010 and 2016 experienced LTU at least once. Overall, the annual risk rates of entering LTU as calculated in our data roughly match the official rates of entry into LTU reported by the German PES [18].

3.1.2 Predictors. We transform our dataset into a one-observation-per-unemployment-episode form to be able to predict an individual's risk of LTU when becoming unemployed. That is, we consider the risk of LTU separately for each unemployment episode found in our data. This per-unemployment-episode solution closely follows PES practices as a new profiling would be conducted each time an individual registers as unemployed with the PES.

The SIAB data includes detailed information on employment and unemployment histories that we use to build predictor variables. To build features that comply with our per-unemployment-episode solution, we aggregate information over episodes before an individual enters unemployment. That is, we count, for example, the number of unemployment episodes an individual experienced in the past or the total duration of previous employment episodes. These predictors summarize individual *labor market histories*. In addition, we create a series of predictors that inform us about the *last job* held by a person, e.g., the industry branch of the job, the skill level required, and the (inflation-deflated) daily wage (if a person was ever employed). The choice of these predictors is inspired by other studies of statistical profiling and they commonly represent the main building block of profiling models that are already used in practice (see Section 2.1).

Socio-demographic information is derived in two ways. Information such as age, gender, and German nationality is derived from the most recent data point containing such information observed before or at entry into an unemployment episode. For information such as education, we consider the highest value observed before or at entry into an unemployment episode as these characteristics are sometimes measured with some inconsistencies [35].

In summary, our feature generation procedures ensure that only information observed at or before entry into unemployment is considered for predicting LTU. A list of the full set of predictors (157 in total) is provided in the appendix (Table A1). We provide further detail on the design decisions we made during data processing, model building and evaluation in Table A2 in the appendix.

3.2 Prediction Setup

Our prediction pipeline takes the outlined variables as input to predict the risk of LTU for an individual unemployment episode. Specifically, the prediction task includes the following components:

- Set of **nonsensitive attributes** X. This set includes all predictors that are presented in Section 3.1.2.
- **Protected attribute** S. Members of the unprivileged group, $S = s^*$, and members of the privileged group, $S = s$. Following Germany's main anti-discrimination regulation, Article 3 of the German constitution (Grundgesetz), we consider gender and German nationality as protected attributes, with female and non-German individuals representing the unprivileged groups. We furthermore consider two (unprivileged) subgroups based on the intersection of both attributes: non-German females (compared to German females and non-German males) and non-German males (compared to German males and non-German females). We build different *features sets* in which the protected attributes are either used or not used as additional predictors.
- **Observed outcome** $Y \in \{0, 1\}$. True binary label of long-term unemployed ($Y = 1$) and not long-term unemployed ($Y = 0$), as outlined in Section 3.1.1.

- **Risk score** $R \in [0, 1]$. Estimate of $Pr(Y = 1 \mid X)$. The predicted risk of becoming long-term unemployed is based on a given prediction model.
- **Prediction** $\hat{Y} \in \{0, 1\}$. Binary prediction of becoming long-term unemployed ($\hat{Y} = 1$) and not becoming long-term unemployed ($\hat{Y} = 0$). Generally, we assume that individuals whose unemployment episodes are classified as LTU would be eligible for labor market support programs. The classification is based on the risk score R and can be assigned along different *classification policies*:

  Policy 1a (P1a). Assign $\hat{Y} = 1$ to the top 10% episodes with the highest predicted risk scores. The classification threshold c_{10} is the $(0.1 \times n)$-th largest element of the risk score vector $\mathbf{r}$.

  $$\hat{Y}^{(P1_a)} = 1 \text{ if } R \geq c_{10}, \text{ else } 0$$

  Policy 1b (P1b). Assign $\hat{Y} = 1$ to the top 25% episodes with the highest predicted risk scores. The classification threshold c_{25} is the $(0.25 \times n)$-th largest element of the risk score vector $\mathbf{r}$.

  $$\hat{Y}^{(P1_b)} = 1 \text{ if } R \geq c_{25}, \text{ else } 0$$

  Policy 2 (P2). Assign $\hat{Y} = 1$ to the 50% of episodes with medium predicted risk scores. The classification threshold c_{75} is the $(0.25 \times n)$-th smallest element of the risk score vector $\mathbf{r}$.

  $$\hat{Y}^{(P2)} = 1 \text{ if } c_{25} \geq R \geq c_{75}, \text{ else } 0$$

Among the three classification policies, P1a and P1b align with the common rationale of classifying high-risk episodes to the LTU class. As we assume that being predicted as LTU would eventually result in interventions, e.g., special support by PES in practice, P2 focuses on a scenario in which such interventions are targeted to medium-risk cases. This scenario is inspired by the Austrian AMAS example which, allegedly, focused support measures on job seekers with a medium risk of LTU (see Section 2.2).

3.2.1 Prediction Models. We consider four methods for building prediction models of LTU. In addition to regression approaches, e.g., used in Italy, the Netherlands, and Sweden [27], we focus on prominent ensemble methods that are typically well-suited for prediction tasks with many features and are already used for profiling purposes in some countries. Specifically, the VDAB system in Belgium employs random forests [28], while both random forests and boosting approaches are considered in New Zealand [27]. In summary, we compute predictions based on the following *model types*:

- **Logistic Regression (LR)**. In common (unpenalized) logistic regression, only the main effects for all predictors are included. Results are in an interpretable set of coefficients and are included as a benchmark.
- **Penalized Logistic Regression (PLR)**. Logistic regression with a penalty on the (ℓ_1, ℓ_2) norm of the regression coefficients [83]. In the former case (ℓ_1 penalty), a more parsimonious model compared to unpenalized logistic regression can be returned, which may increase both interpretability and prediction performance.
- **Random Forest (RF)**. An ensemble of deep (uncorrelated) decision trees grown on bootstrap samples [16]. Results in a model that cannot be readily interpreted without further helper methods.
- **Gradient Boosting Machines (GBM)**. An ensemble of small decision trees that are grown in sequence by using the (updated) pseudo-residuals in each iteration as the outcome [37, 38]. Similar to RF, additional techniques are typically needed to support the interpretation of results.

Model training and evaluation. As outlined in Section 3.1, our SIAB data includes information from the beginning of 2010 up to the end of 2016. To robustly assess the fairness implications of LTU profiling, we evaluate prediction models in two *evaluation data* sets which include data from 2015 and 2016, respectively. The corresponding *training data* cover the preceding years, i.e., 2010–2014 (models evaluated with data from 2015) and 2010–2015 (models evaluated with data from 2016). To ease the computational burden related to model tuning (see below), a random sample of 20,000 unemployment episodes from each training year is drawn to construct the respective training set. The final model evaluation is done on the full data (86,692 unemployment episodes in 2015, and 89,710 episodes in 2016).

Model tuning and selection. Hyperparameter tuning for PLR, RF, and GBM is based on temporal cross-validation [46]: Training and test sets are constructed from the training data by successively moving the time point which separates the fit and test period forward in time. While this leads the training set to grow over time, we fix the respective test period to a single year. That is, the first fit and test periods include data from 2010 (fit) and 2011 (test). The last fit period covers data up to the last training year (2010–2013, 2010–2014), and the last test period includes data of the last training year (2014 and 2015, respectively). The hyperparameter setting with the highest average ROC-AUC over all test periods is chosen for each model type.

Training histories. The selected hyperparameter settings are used to re-train prediction models with the *full training data* (2010–2014, 2010–2015). Furthermore, we re-train additional sets of models with *restricted training data* using only the most recent training year (2014 and 2015, respectively). This is done to explore the fairness implications of training LTU models with different training data histories: One may argue that with the restricted data prediction models have fewer chances to learn discriminatory practices concerning the effects of gender and nationality on LTU propensities if those practices are more commonly observed in older (training) data.

Trained models. Model re-training with the full and restricted training data is done with and without protected attributes, respectively. Thus, we train a total of 16 final prediction models (model type × full/restricted training data × with/without protected attributes) for each training horizon (2010–2014 and 2010–2015) for predicting LTU in the respective evaluation set (2015 and 2016). 48 sets of class predictions are obtained per evaluation set by applying the three classification policies to each model.

Software. We used Stata (15, [82]) and R (3.6.3, [78]) for data preparations. Model training and evaluation was done with Python (3.6.4), using the scikit-learn (0.19.1, [76]) and aif360 (0.4.0, [12]) packages.[5]

3.2.2 Performance and Fairness Metrics. **Performance metrics.** The implementation of statistical profiling systems critically depends on the ability of the underlying prediction models to correctly identify individuals at risk who should receive preventive interventions. In the present context, accurate predictions are a prerequisite for an effective allocation of support programs to unemployed individuals. From a fairness perspective, high accuracy should not only hold overall but also for subgroups defined by protected attributes and their intersection. We posit that while technically less training data might be available for small subgroups, there is no adequate justification for an unequal distribution of prediction error as eventually predictions across all groups are used to guide decisions in practice [58]. We evaluate subgroup prediction performance concerning the predicted classes $\hat{Y}$ (balanced accuracy, F1 score). We further evaluate our prediction models

[5]Code for replication purposes is available at the following OSF repository: https://osf.io/9b4mp/?view_only=d625065eca2d428e9b3c3507a6c3579a

using additional classification measures (precision, recall) and based on risk scores R (ROC-AUC and PR-AUC) for comparison purposes.

- **Balanced Accuracy**. The arithmetic mean of sensitivity (recall) and specificity. In range $[0, 1]$, with 0.5 representing performance at random.

$$\text{Bal. Acc.} = \frac{1}{2} \times (\text{Sens.} + \text{Spec.})$$

- **F1 Score**. Weighted average of precision and recall. In range $[0, 1]$.

$$\text{F1} = 2 \times \frac{\text{Prec.} \times \text{Rec.}}{\text{Prec.} + \text{Rec.}}$$

- **Precision (at k)**. The proportion of correctly identified LTU episodes among all predicted LTU episodes. In range $[0, 1]$.

$$\text{Prec.} = \frac{1}{k} \sum_{i=1}^{n} y_i \mathbf{1}(r_i \geq r_{[k]})$$

Where k is a constant (i.e., the number of instances with a predicted positive outcome) and $r_{[k]}$ denotes the k-th largest element of the risk score vector $\mathbf{r}$.

- **Recall (at k)**. The proportion of correctly identified LTU episodes among all LTU episodes. In range $[0, 1]$.

$$\text{Rec.} = \frac{1}{\sum_{i=1}^{n} y_i} \sum_{i=1}^{n} y_i \mathbf{1}(r_i \geq r_{[k]})$$

- **ROC-AUC**. **Area under the receiver operating characteristic (ROC)** curve. In range $[0, 1]$, with 0.5 representing performance at random.
- **PR-AUC**. **Area under the precision-recall curve**. In range $[0, 1]$.

Fairness metrics. Our fairness evaluation follows the disparate impact framework and aims to investigate potential disadvantageous outcomes of statistical profiling processes for individuals according to their sensitive attributes [65]. On this basis, we focus on (multi-)group fairness notions but also consider individual fairness. Regarding group fairness, unemployed individuals who are members of unprivileged groups should not be disproportionately (falsely) excluded from labor market programs. This perspective considers targeted support from PES as an assistive intervention to which access should not be blocked or delayed just by being a member of a group that is defined by a protected attribute. In this context, we consider parity-based metrics that are defined solely based on predictions of long-term unemployment, and differences in false negative rates to measure the extent to which true LTU episodes are not correctly detected across groups. Regarding individual fairness, unemployed individuals with similar (nonsensitive) attributes should be assigned similar predictions. This perspective requires predictions that eventually make similar unemployed individuals equally eligible to be assigned to support programs.

- **Statistical Parity Difference**. The difference in the probability of being predicted LTU – i.e., being eligible for support programs – between an unprivileged and a privileged group.

$$Pr(\hat{Y} = 1 \mid S = s^*) - Pr(\hat{Y} = 1 \mid S = s)$$

- **Conditional Statistical Parity Difference**. The difference in the probability of being predicted LTU between an unprivileged and a privileged group, conditional on nonsensitive attributes. We condition on education (i.e., having a high school diploma).

$$Pr(\hat{Y} = 1 \mid S = s^*, X = x) - Pr(\hat{Y} = 1 \mid S = s, X = x)$$

— **False Negative Rate Difference**. The difference in false negative rates (one minus recall) between an unprivileged and a privileged group.

$$Pr(\hat{Y} = 0 \mid Y = 1, S = s^*) - Pr(\hat{Y} = 0 \mid Y = 1, S = s)$$

— **Consistency**. The average similarity of individual predictions and the predictions of their k-nearest neighbors [86]. The neighbors are defined based on the full set of (nonsensitive) attributes. We use $n_{neighbors} = 5$. Higher scores indicate more consistent predictions.

$$1 - \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - \frac{1}{n_{neighbors}} \sum_{j \in \mathcal{N}_{n_{neighbors}}(x_i)} \hat{y}_j|$$

4 Fairness Auditing

4.1 Subgroup Prediction Performance

We start by briefly presenting the overall prediction performance of the trained models to provide some context for the following fairness evaluations. Overall ranking and classification performance of the final prediction models after model tuning is presented in Table B1 (models trained with data from 2010–2014 and evaluated in 2015) and B2 (models trained with data from 2010–2015 and evaluated in 2016). In summary, we observe ROC-AUC scores in the range $[0.694, 0.774]$ and PR-AUC in $[0.252, 0.355]$, which largely aligns with performance results that have been reported for LTU prediction in other countries [27]. Comparing model types, we see that logistic regression is outperformed by PLR, RF, and GBM. Restricting the training data to include only the most recent year leads to somewhat lower performance levels while in- or excluding protected attributes has little effect on overall performance. Comparing prediction performance between the two evaluation data sets, we see some indication of lower performance when predicting LTU in 2016 compared to 2015.

Subgroup-specific performance results are presented in Figure 1. In each subplot, the distribution of subgroup (and overall) performance scores is shown for the full set of LTU predictions, i.e., for all combinations of model type, training horizon (full/restricted training data), feature setting (with/without protected attributes) and classification policy. In these comparisons, we focus on the two high-risk classification policies (P1a and P1b) as the medium-risk policy (P2) is deliberately set to not optimize performance but to identify unemployment episodes of job seekers who might be most 'susceptible' to support measures. We provide supplemental Figures B1 and B2 in the appendix which plot the subgroup performance scores grouped by model type.

The performance results allow for the following three conclusions. First, strong differences in balanced accuracy and F1 scores across groups can be observed (Figure 1). Taking overall performance as the baseline, the LTU predictions are similarly accurate for female job seekers but less accurate for non-Germans. An additional drop in performance can be observed when restricting the evaluation to non-German males. Second, the degree of subgroup-specific performance loss depends on the model type and classification policy. We observe stronger differences in balanced accuracy under the less restrictive classification threshold (P1b), particularly for logistic regression-based predictions (Figure B1). This result indicates that the overall improvement in performance under policy 1b comes at the cost of higher variation in performance scores across groups, introducing a delicate trade-off for employment agencies. At least for balanced accuracy, there might be an incentive to consider (more) restrictive thresholds for profiling practices although this does not protect against considerable variation in F1 scores across groups for all model types in our case (Figure B2). Which performance measure should be preferred is another conundrum as it depends on whether the focus of an employment agency is on correctly predicting both outcome categories

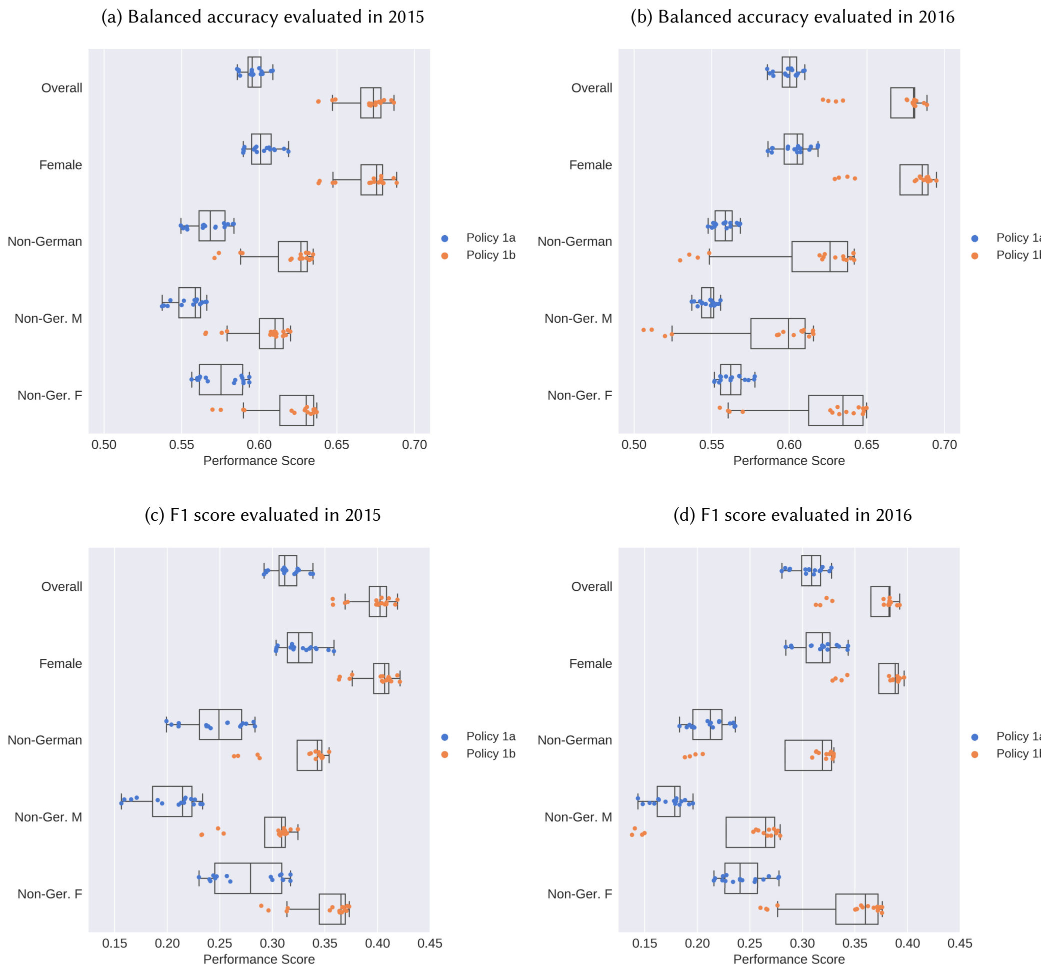


Fig. 1. Distribution of performance scores for different sets of LTU predictions, overall and by groups.

(balanced accuracy) or more specifically on efficiently identifying cases with a high risk of LTU (F1 score). Third and finally, similar performance patterns can be observed for both evaluation years, indicating that low subgroup performance is a systematic issue in profiling and not tied to the temporal specifics of a single evaluation year.

4.2 Fairness Metrics

We next evaluate the LTU predictions concerning fairness metrics, with a focus on group differences in the potential to receive support programs under different policies and on group-specific prediction error.

Figure 2 shows the distribution of statistical parity, conditional statistical parity, and false negative rate differences of the various LTU prediction sets, evaluated in 2015 and 2016. Similar to Figure 1, each subplot shows group-specific fairness scores for all combinations of model type, training horizon, feature setting, and classification policy. As this evaluation aims to study the composition of job seekers that would eventually be assigned to interventions, high-risk (P1a, P1b) as well as medium-risk (P2) classification policies are considered.

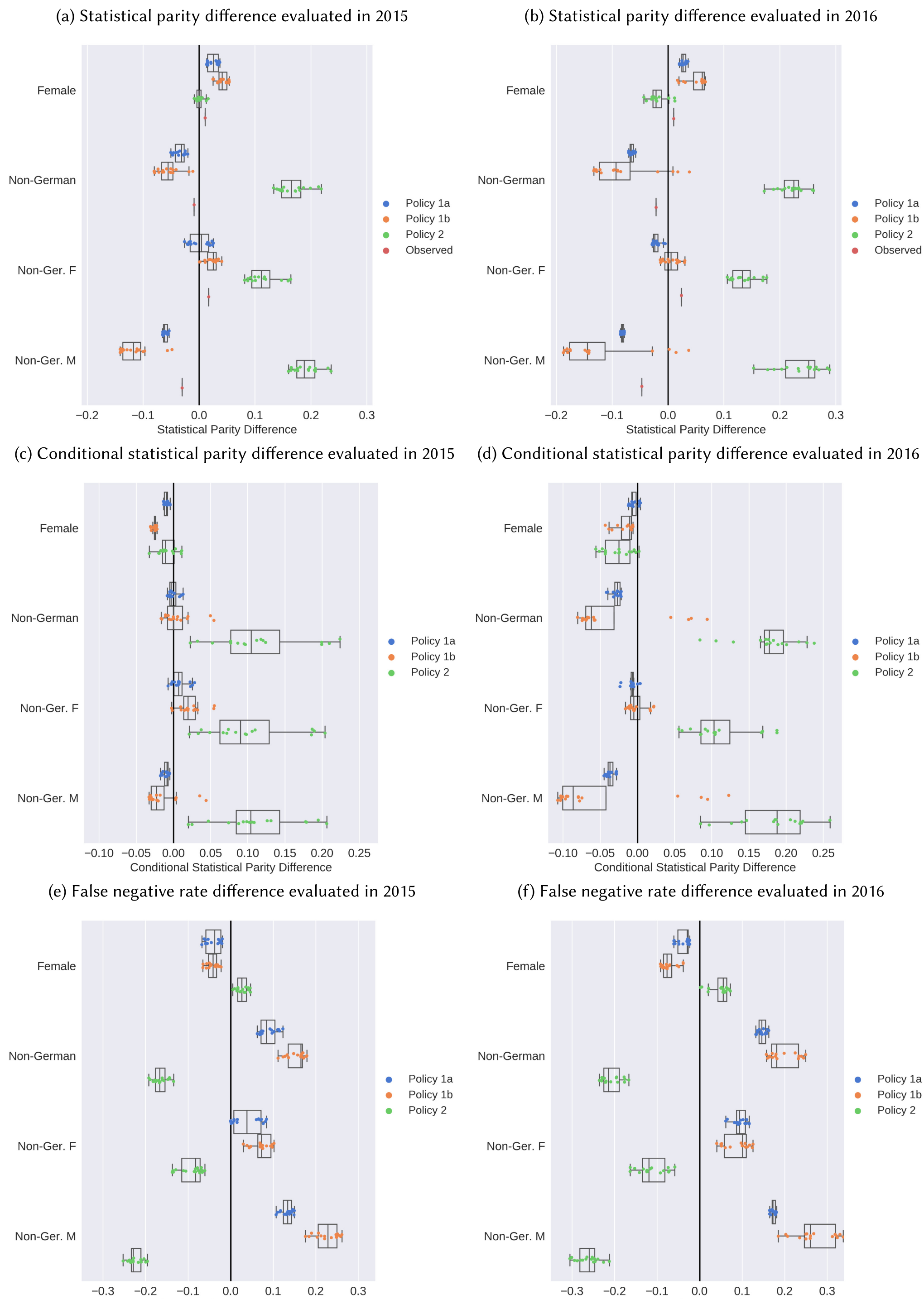


Fig. 2. Distribution of fairness metric scores for different sets of LTU predictions by groups.

Starting with statistical parity in 2015 as shown in Figure 2(a), we see little differences in the average probability with which unemployment episodes of female job seekers are classified as LTU, compared to predictions for male job seekers. Stronger differences emerge concerning nationality. Unemployment episodes of non-German job seekers are less likely to be classified as LTU under the high-risk policies (P1a and P1b), but considerably more likely to be assigned to the medium-risk class (P2), compared to episodes of German job seekers. This suggests that foreign-born job seekers would have a higher chance of being eligible for support programs under a policy that classifies unemployment episodes with medium LTU risk as relevant, but a lower chance of being supported under high-risk policies. Statistical parity differences are particularly pronounced for non-German males. Similar patterns can be observed for the 2016 data (Figure 2(b)). Note that the composition of job seekers that in fact, experience LTU in the evaluation data is rather balanced concerning both gender and nationality (label "Observed" in Figure 2(a) and 2(b)) and thus observed group differences tend to be magnified by the profiling models. The degree of over-amplification is largely driven by the choice of the classification threshold, with the restrictive high-risk threshold (policy 1a) resulting in predictions that are closest to the observed group differences. Arguably, this might be expected as a strict threshold focuses on those LTU episodes for which the models are most confident. In combination with the previous results on model performance, this indicates that targeting a larger set of job seekers (policy 1b) in practice might increase the number of detected LTU episodes but also increases the risk of (inaccurate) group-based stereotyping.

Following the argument that unemployment episodes of demographic groups may be more (or less) likely classified as LTU due to structural differences in nonsensitive attributes between those groups, we re-calculated statistical parity differences conditional on education (i.e., having a high school diploma). In this case, parity differences concerning nationality are mitigated, particularly for the 2015 data (Figure 2(c)). Nonetheless, even among higher-skilled individuals, unemployment episodes of non-German job seekers are more often assigned to the medium-risk class than those of German job seekers. Conditioning on education has a less strong effect on parity differences in the 2016 data (Figure 2(d)).

The false negative rate differences in Figure 2(e) and 2(f) suggest that the outlined parity differences can be attributed to systematic prediction error. True LTU episodes of foreign-born job seekers are more often incorrectly classified as non-LTU episodes (i.e., higher false negative rates) under policies 1a and 1b, contributing to the parity differences that exceed differences in base rates as observed above. Conversely, lower false negative rates for foreign-born job seekers can be observed under policy 2. False negative rate differences are more pronounced in the 2016 evaluation data, and particularly strong for non-German males.

The outlined results highlight that choosing between classification thresholds has considerable fairness implications. To elaborate on this point, Figure B3 shows overall prediction performance (summarized by the F1 score) and statistical parity difference (based on nationality) over the full range of classification thresholds for selected prediction models. For thresholds that are less strict than policy 1a and 1b, more unemployment episodes of non-German (compared to German) job seekers are classified as LTU, whereas this difference is reversed as we increase the classification threshold. For thresholds that are more strict than policies 1a and 1b, the composition of job seekers who are predicted to experience LTU becomes more balanced.

In addition to group-based fairness metrics, we can also audit the LTU predictions concerning consistency (Table B3). In this case, we are interested in evaluating whether job seekers who are similar (in nonsensitive attributes) receive similar predictions on average. We observe rather high consistency scores for classifications that are based on policies 1a and 1b, indicating that job seekers with similar attributes would be largely treated similarly in these scenarios. Consistency is considerably lower under policy 2. Focusing on employment episodes with medium LTU risks thus decreases individual fairness and there is a higher chance that similar job seekers receive different predictions.

5 Discussion

5.1 Reflecting Design Decisions in Algorithmic Profiling

We evaluated the use of prediction models for profiling job seekers based on extensive German administrative data concerning subgroup prediction performance, fairness metrics, and implications of modeling decisions. We compared regression and machine learning approaches to predict long-term unemployment (LTU) using different classification thresholds, feature sets, and training horizons. Building on previous research in algorithmic fairness on the importance of design decisions [75, 80, 81], we particularly focused on the downstream effects of different profiling settings to evaluate how biases in historical labor market data are moderated or exacerbated by decisions made during data processing and model specification.

Our results show that applying a standard machine learning pipeline to administrative labor market data can have detrimental consequences for the individuals that would be affected by the models' predictions. While our profiling models achieve good overall performance scores that are comparable with results reported in other countries, strong differences in prediction performance across groups emerge. While the models perform similarly well for male and female job seekers, predictions are less accurate for foreign-born job seekers. These inaccuracies surface as over-amplifications of group differences in the models' predictions that exceed true differences in LTU rates between German and foreign-born individuals.

Among the design decisions we tested, two decision points stand out: Choosing the *model type* and the *classification policy*. Logistic regression showed the strongest drop in subgroup-specific prediction performance (especially under classification policy 1b). While this behavior might point to specification issues as already indicated by lower overall performance scores, its scale is only fully conceivable based on a careful model evaluation routine that explicitly takes vulnerable subpopulations into account. To some extent, the low subgroup performance of logistic regression stands in contrast to its apparent benefits in interpretability for public employment services, a trade-off that was less pronounced for penalized regression models in our study. Next to (and in interaction with) model types, choosing between different classification policies had considerable fairness implications: foreign-born (non-German) job seekers may have a higher (under policy P2) or lower (under policy P1a and P1b) chance of being eligible for support measures than German job seekers, depending on whether medium or high-risk individuals would be targeted by PES. Selecting a classification policy similarly determines which group experiences higher false negative rates. Compared to German job seekers, true LTU episodes of foreign-born job seekers are often not correctly detected by profiling models under high-risk classification policies while the opposite holds under a medium-risk policy. Thus, following the standard high-risk profiling theme would be detrimental for those groups that already experience various forms of disadvantage in the labor market. Among the three classification policies we tested, the strictest threshold (P1a) led to the most honest reflections of true group differences but still incurred considerable misclassifications for vulnerable social groups and thus should only be considered carefully in connection with additional mitigation procedures, safeguards, and sensible allocation strategies.

It is also important to note that eligibility for support measures does not necessarily imply positive labor market outcomes in practice. Different labor market programs are differently effective (for different groups) and can eventually also lead to negative outcomes such as vicious cycles of precarious employment or adverse mental health impacts [79]. Forced participation based on an incorrect risk assessment can similarly put an additional burden on individuals. The higher chance of a "positive" prediction for non-Germans jobseekers under the medium risk policy (P2) thus needs to be interpreted with great caution.

Selecting a classification threshold is only one of the many design decisions that need to be made when translating a policy problem into a tractable modeling task [75]. While we demonstrate the implications of selected options at specific decision points, other decisions were made without consideration of alternatives. As documented in Table A2, further critical decisions at the *data selection* step include the specific definition of the outcome variable of interest and the selection and definition of protected attributes, both of which directly tie to considerations of measurement bias and to the ability to adequately identify adverse impacts downstream. At the *preprocessing* step, we implemented a single processing pipeline although seemingly small changes at this stage can similarly have considerable fairness implications [23, 81]. While we considered a basic set of model types at the *modeling* step, our model tuning strategy only evaluated and optimized for prediction performance. We further note that at the *evaluation* step, considering evaluation data from two years might have increased robustness, but different time frames (e.g., monthly/ seasonal data) and subsets (e.g., evaluation by regions) could have been considered to probe model outputs for patterns of disparate impact more thoroughly.

5.2 Integrating Fairness Evaluations into Deployment Processes

While an employment agency may have some degrees of freedom when it comes to modeling decisions such as choosing the model type to be implemented, setting a classification policy, and deciding on the allocation scheme of support measures in practice strongly depends on the broader socio-institutional context, including the labor market policies, legislation, and budget constraints. However, we highlight that statistically, different thresholds do not only imply different precision-recall trade-offs but also different amplifications of group-specific biases. Thus, the critical discussions in public agencies implementing algorithmic profiling need to be centered around the broader socio-technical system and on the interplay between group-specific model error and the eventual use of the model's predictions. As structural differences in the labor market are (over)incorporated in profiling models, their predictions can be used to either mitigate or reinforce group differences, depending on the choice of the intervention regime. Choosing the "optimal" threshold or technical solutions such as group-specific thresholds [42, 50] cannot solve this conundrum alone as they do not differentiate among the various factors that can contribute to group-specific risks and require careful consideration of how differences can eventually be mitigated under which distributive justice principle [58]. Against this background, awareness of the learned group-specific patterns and errors is only an essential first step that can guide crucial discussions between developers, policymakers, and PES.

If biased predictions are discovered, one may typically want to correct them, for example, by pre-processing training data, by in-processing algorithms, or by post-processing predictions [see, e.g., 22, for an overview of debiasing techniques]. At this point, we cannot give recommendations regarding the question of how structural discrimination should be treated when found. For example, how should differences between men and women be treated when the German Social Code Book III, article 2/4 states that PES support should explicitly improve the labor market chances of women to remove existing disadvantages? From this perspective, distinguishing between the prediction and the decision step is essential [59]. For example, we may argue that any debiasing of profiling models should aim for high prediction accuracy across social groups [43, 53] rather than equalizing parity differences, such that the latter can be targeted by a sensible allocation of PES support. In the end, understanding biases and unequal treatment of social groups, especially of those that have been disadvantaged in the past, is a necessary precondition before any ADM system is implemented.

5.3 Limitations and Outlook

There are several limitations to our study. Germany currently implements case worker-based profiling, and the profiling outcomes cannot be reconstructed with the administrative data used in this study. We therefore cannot evaluate how our results compare to current profiling approaches used by the German PES, particularly in terms of fairness evaluations. However, previous literature comparing case worker-based and statistical profiling in other countries shows that statistical models outperform human predictions of LTU [7, 8]. Since the prediction performance of our profiling models is comparable to those of other countries, we assume that similar conclusions may be drawn for the German case. Nonetheless, our results show it is critical to acknowledge variation in performance across social groups and to carefully evaluate fairness implications rather than being solely guided by overall prediction performance. Moreover, to understand the larger societal impact of an algorithmic approach, both on the organizational side of PES and the influence on job seekers, one needs to extend the focus beyond fairness evaluations of the prediction step. Such assessments are beyond the focus of this paper.

Furthermore, given our focus on historical discrimination in the labor market and as the administrative data used in this study is somewhat limited concerning the measurement of detailed socio-demographic information, we only considered selected protected groups, and our results may therefore only provide a lower bound of potential biases in profiling of the unemployed. Computing fairness measures with respect to gender and nationality, operationalized with two simple binary measures and their combination, cannot cover the complexities of how intersectional discrimination manifests on the labor market. Further work could also consider the application of debiasing techniques in the present context to study their potential to correct group-specific prediction errors and advance toward fair algorithmic profiling of job seekers.

Appendices

A Variables and Design Decisions

Table A1. List of Predictors Included in LTU Prediction Models

Group	Predictor
Socio-demo.	Age
	Vocational education, categorized (6 dummy variables)
	School education, categorized (7 dummy variables)
	State of residence
	Number of moves
Labor market history	In Employment six weeks before unemployment?
	Long-term unemployment benefits receipt six weeks before unemployment?
	Short-term unemployment benefits receipt six weeks before unemployment?
	Subsidized employment six weeks before unemployment?
	Registered as job-seeking while not unemployed six weeks before unemployment?
	Registered with PES for other reasons six weeks before unemployment?
	No information available six weeks before unemployment?
	Number of employers worked for
	Number of jobs without any vocational training held
	Mean duration of employment without any vocational training
	Total duration worked in industry x (14 types of industries)
	Total duration more than one job
	Total duration in marginal employment
	Total duration in full-time employment
	Total duration in fixed-term employment
	Total duration in temporary employment

(Continued)

Table A1. Continued

Group	Predictor
	Number of ALG II benefits receipt episodes
	Total duration of ALG II benefits receipt episodes
	Mean duration of ALG II benefits receipt episodes
	Number of ALG I benefits receipt episodes
	Total duration of ALG I benefits receipt episodes
	Mean duration of ALG I benefits receipt episodes
	Number of labor market program participation episodes
	Total duration of labor market program participation episodes
	Mean duration of labor market program participation episodes
	Total duration of subsidized employment episodes
	Number of job seeking episodes
	Total duration of job seeking episodes
	Mean duration of job seeking episodes
	Industry individual worked in for the longest time (14 dummy variables)
	Days since last employment, categorized (3 dummy variables)
	Days since last labor market contact, categorized (4 dummy variables)
	Days since last labor market contact (full-time), categorized (4 dummy variables)
	Time since last unemployment spell, categorized (6 dummy variables)
	Maximum skill-level required for all employment episodes, categorized (4 dummy variables)
	Total duration of employment episodes, scaled by age
	Total duration of employment episodes with more than one job, scaled by age
	Total duration of marginal employment, scaled by age
	Total duration of full-time employment episodes, scaled by age
	Total duration of fixed-term employment episodes, scaled by age
	Total duration of temporary work episodes, scaled by age
	Total duration of ALG II benefits receipt episodes, scaled by age
	Total duration of ALG I benefits receipt episodes, scaled by age
	Total duration of ALMP participation episodes, scaled by age
	Total duration of ALMP participation (activation) episodes, scaled by age
	Total duration of subsidized employment episodes, scaled by age
	Total duration of job seeking episodes, scaled by age
Last job	No info about previous jobs available
	Duration of last job
	More than one job at last job
	Inflation-deflated wage of last job
	Type of last job
	Type of last job missing
	Last job was part-time
	Last job part-time missing
	Skill-level required for last job, categorized (4 dummy variables)
	Last job was fixed-term
	Last job was fixed-term, missing
	Last job was temporary work, missing
	Last job was temporary work, missing
	Industry of last job (14 dummy variables)
	Commuted for last job?
	Commuted for last job, missing
	Last employment more than one job

Table A2. Summary of Design Decisions and Their Respective Implementation in Our Study (Adapted from [81])

Category	Decision	Our implementation	Alternatives considered?	Related fairness concept(s)
Data Selection	Definition of outcome	LTU following German PES standard definition	✗	Historical-, measurement- and representation bias [69]
	Features	Two feature sets (in- vs. excluding protected attributes)	✓	
	Training sample	Two training data sets (full vs. restricted by years)	✓	
	Selection and definition of protected attributes	Four sets of protected groups, ethnicity defined based on nationality	~	
Preprocessing	Coding of protected attributes	Attribute-specific comparisons against contrast group (e.g., non-German male vs. German male and non-German female)	✗	ML pipeline bias [23, 80]
	Scaling of continuous variables	No scaling	✗	
	Binning of continuous variables	No binning	✗	
	Encode of categorical variables	One-hot encoding	✗	
	Dealing with missing data	Median imputation and missing flags/categories	✗	
Modeling	Model types	Four model types (LR, PLR, RF, GBM)	✓	Algorithmic bias [30, 69]
	Model tuning	Fixed tuning strategy (temporal cross-validation w.r.t. ROC-AUC)	✗	
Post-Hoc	Classification threshold	Three classification policies (P1a, P1b, P2)	✓	
Evaluation	Evaluation sample	Two evaluation sets, fixed data splitting strategy (by year)	~	Evaluation bias [69], fairness hacking [67]

B Additional Results

Table B1. Prediction Performance of LTU Prediction Models, Evaluated in 2015

			Policy 1a				Policy 1b			
	ROC-AUC	PR-AUC	Bal. Acc.	F1 Score	Prec.	Rec.	Bal. Acc.	F1 Score	Prec.	Rec.
(a) Models trained with 2010-2014 data, without protected attributes.										
LR	0.715	0.294	0.588	0.298	0.367	0.251	0.648	0.372	0.295	0.503
PLR	0.761	0.327	0.596	0.313	0.385	0.263	0.676	0.406	0.322	0.550
RF	0.764	0.341	0.600	0.322	0.397	0.271	0.679	0.410	0.325	0.555
GBM	0.774	0.354	0.607	0.336	0.414	0.283	0.686	0.419	0.332	0.567
(b) Models trained with 2014 data, without protected attributes.										
LR	0.710	0.288	0.588	0.297	0.366	0.250	0.639	0.360	0.286	0.488
PLR	0.757	0.323	0.595	0.312	0.384	0.263	0.671	0.401	0.318	0.543
RF	0.758	0.327	0.596	0.313	0.386	0.264	0.672	0.402	0.319	0.544
GBM	0.767	0.338	0.601	0.323	0.398	0.272	0.681	0.413	0.327	0.559
(c) Models trained with 2010-2014 data, with protected attributes.										
LR	0.716	0.293	0.588	0.296	0.365	0.250	0.650	0.374	0.297	0.507
PLR	0.761	0.327	0.595	0.312	0.385	0.263	0.675	0.406	0.322	0.550
RF	0.764	0.341	0.601	0.325	0.400	0.273	0.678	0.409	0.324	0.554
GBM	0.774	0.355	0.608	0.338	0.417	0.285	0.687	0.421	0.333	0.570
(d) Models trained with 2014 data, with protected attributes.										
LR	0.712	0.288	0.589	0.299	0.368	0.251	0.639	0.360	0.286	0.488
PLR	0.757	0.323	0.595	0.311	0.383	0.262	0.671	0.401	0.318	0.543
RF	0.758	0.326	0.596	0.314	0.387	0.264	0.673	0.403	0.319	0.545
GBM	0.767	0.339	0.602	0.325	0.401	0.274	0.679	0.411	0.326	0.556

Table B2. Prediction Performance of LTU Prediction Models, Evaluated in 2016

(a) Models trained with 2010-2015 data, without protected attributes.

			Policy 1a				Policy 1b			
	ROC-AUC	**PR-AUC**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**
LR	0.700	0.256	0.589	0.287	0.328	0.256	0.632	0.325	0.246	0.479
PLR	0.760	0.298	0.600	0.308	0.351	0.274	0.681	0.383	0.290	0.565
RF	0.764	0.313	0.607	0.321	0.367	0.286	0.681	0.384	0.290	0.566
GBM	0.770	0.325	0.610	0.328	0.374	0.291	0.687	0.391	0.296	0.576

(b) Models trained with 2015 data, without protected attributes.

			Policy 1a				Policy 1b			
	ROC-AUC	**PR-AUC**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**
LR	0.695	0.253	0.591	0.291	0.332	0.259	0.627	0.319	0.241	0.471
PLR	0.756	0.298	0.602	0.312	0.356	0.278	0.680	0.382	0.289	0.563
RF	0.758	0.297	0.599	0.306	0.349	0.272	0.676	0.378	0.286	0.558
GBM	0.763	0.309	0.605	0.319	0.364	0.284	0.682	0.385	0.291	0.568

(c) Models trained with 2010-2015 data, with protected attributes.

			Policy 1a				Policy 1b			
	ROC-AUC	**PR-AUC**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**
LR	0.703	0.257	0.588	0.284	0.324	0.253	0.636	0.330	0.250	0.487
PLR	0.760	0.298	0.599	0.307	0.351	0.273	0.681	0.383	0.290	0.565
RF	0.764	0.312	0.606	0.320	0.365	0.284	0.681	0.383	0.290	0.565
GBM	0.771	0.326	0.611	0.329	0.376	0.293	0.689	0.393	0.297	0.580

(d) Models trained with 2015 data, with protected attributes.

			Policy 1a				Policy 1b			
	ROC-AUC	**PR-AUC**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**	**Bal. Acc.**	**F1 Score**	**Prec.**	**Rec.**
LR	0.694	0.252	0.591	0.292	0.333	0.259	0.623	0.315	0.238	0.464
PLR	0.756	0.298	0.603	0.313	0.358	0.279	0.680	0.382	0.289	0.563
RF	0.758	0.297	0.598	0.305	0.348	0.272	0.676	0.378	0.286	0.558
GBM	0.763	0.310	0.605	0.319	0.364	0.284	0.682	0.384	0.291	0.567

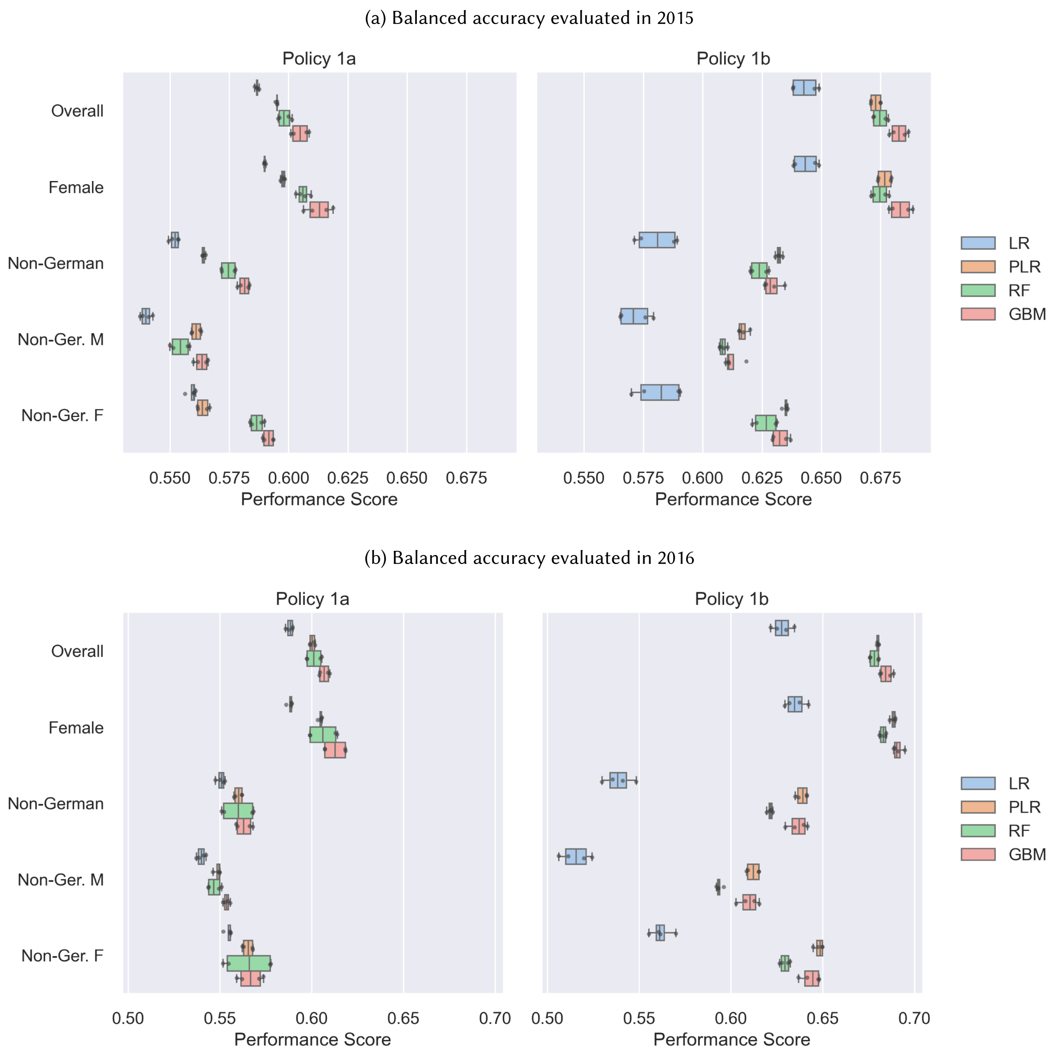
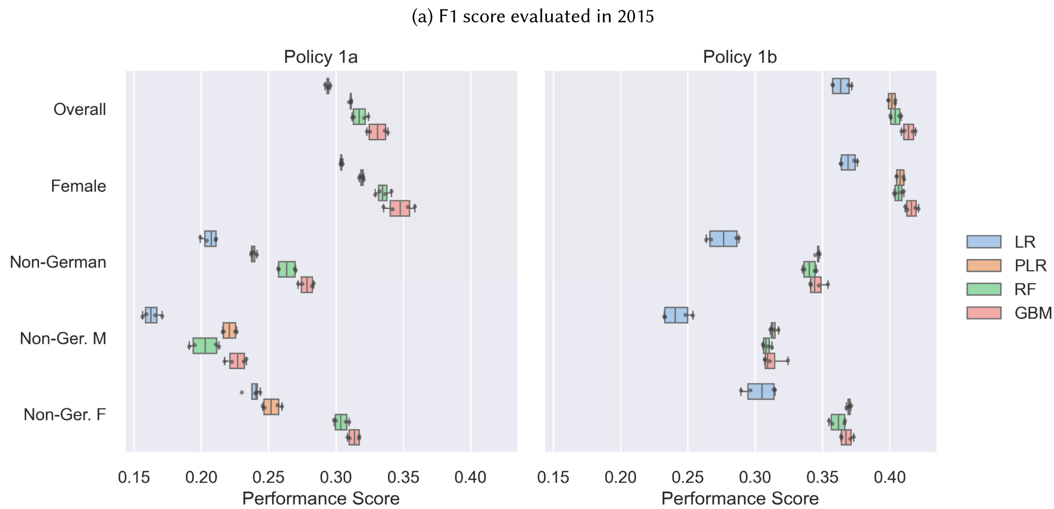


Fig. B1. Distribution of performance scores for different sets of LTU predictions, overall and by groups.



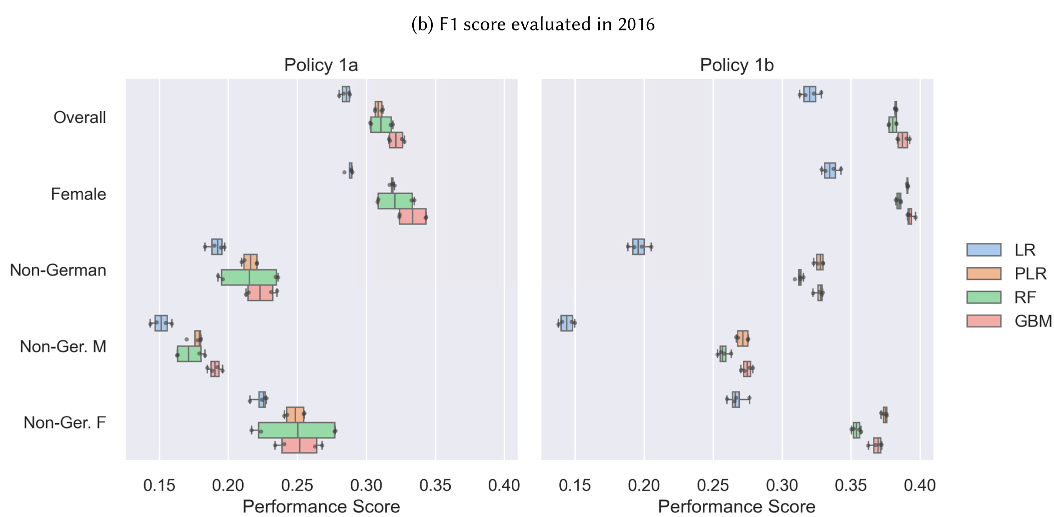


Fig. B2. Distribution of performance scores for different sets of LTU predictions, overall and by groups.

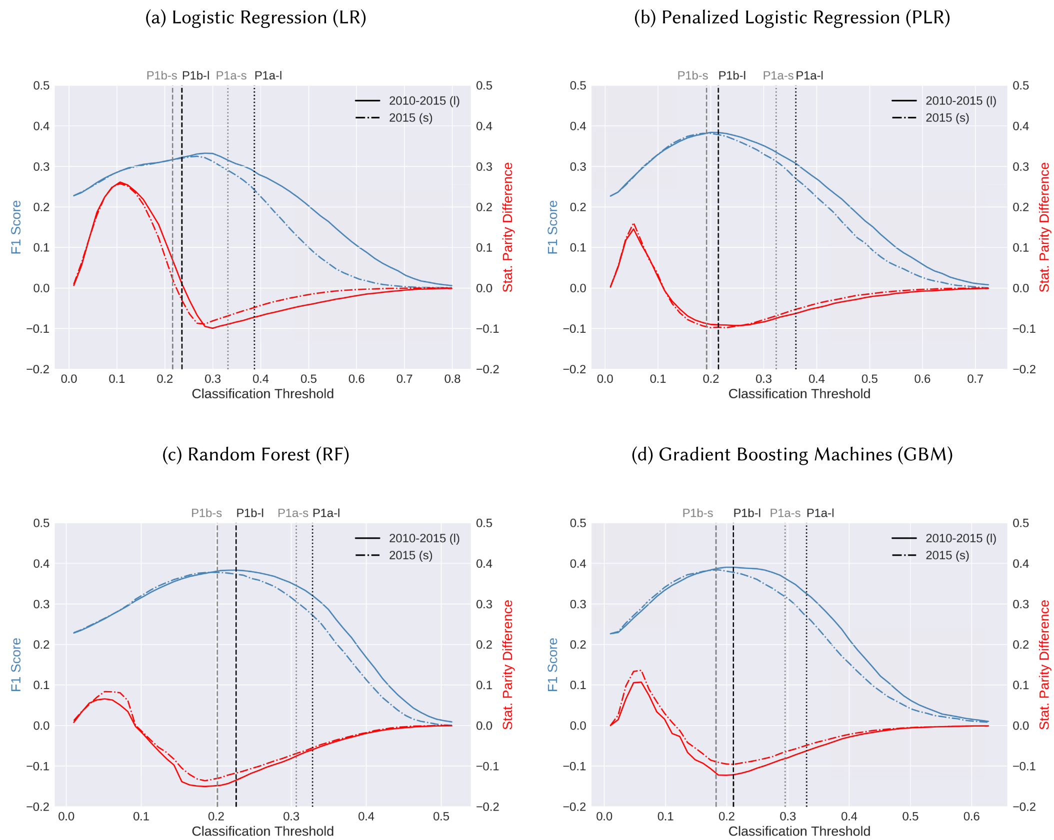


Fig. B3. F1 and statistical parity difference (non-German vs. German) versus threshold curves of LTU prediction models, trained without protected attributes in 2010–2015 and 2015, and evaluated in 2016. The classification threshold of policy 1a is indicated by a dotted line and the threshold of policy 1b by a dashed line.

Table B3. Consistency of LTU Prediction Models with Different Threshold Policies

(a) Models evaluated in 2015 | (b) Models evaluated in 2016

Model	Policy	Training data (a)	Consistency (a) without protected attributes	Consistency (a) with protected attributes	Training data (b)	Consistency (b) without protected attributes	Consistency (b) with protected attributes
Label			0.82	0.82		0.84	0.84
LR	P1a	2010-2014	0.96	0.96	2010-2015	0.96	0.96
	P1a	2014	0.95	0.95	2015	0.96	0.96
	P1b	2010-2014	0.92	0.93	2010-2015	0.92	0.92
	P1b	2014	0.92	0.92	2015	0.92	0.92
	P2	2010-2014	0.82	0.81	2010-2015	0.82	0.82
	P2	2014	0.81	0.80	2015	0.82	0.82
PLR	P1a	2010-2014	0.93	0.93	2010-2015	0.94	0.94
	P1a	2014	0.93	0.93	2015	0.94	0.94
	P1b	2010-2014	0.89	0.89	2010-2015	0.89	0.89
	P1b	2014	0.89	0.89	2015	0.89	0.89
	P2	2010-2014	0.76	0.76	2010-2015	0.76	0.76
	P2	2014	0.76	0.76	2015	0.76	0.76
RF	P1a	2010-2014	0.94	0.94	2010-2015	0.94	0.94
	P1a	2014	0.94	0.94	2015	0.95	0.95
	P1b	2010-2014	0.91	0.91	2010-2015	0.91	0.91
	P1b	2014	0.92	0.92	2015	0.91	0.91
	P2	2010-2014	0.79	0.79	2010-2015	0.80	0.79
	P2	2014	0.80	0.80	2015	0.80	0.80
GBM	P1a	2010-2014	0.93	0.93	2010-2015	0.93	0.93
	P1a	2014	0.93	0.93	2015	0.92	0.92
	P1b	2010-2014	0.89	0.89	2010-2015	0.89	0.89
	P1b	2014	0.89	0.89	2015	0.88	0.88
	P2	2010-2014	0.76	0.76	2010-2015	0.77	0.77
	P2	2014	0.76	0.76	2015	0.76	0.76

References

[1] Katharine G. Abraham, John Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2019. The consequences of long-term unemployment: Evidence from linked survey and administrative data. *ILR Review* 72, 2 (2019), 266–299.

[2] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic profiling of job seekers in Austria: How Austerity politics are made effective. *Frontiers in Big Data* 3 (2020). https://doi.org/10.3389/fdata.2020.00005

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. DER AMS-ALGORITHMUS. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Endbericht. https://doi.org/10.1553/ITA-pb-2020-02

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *Ethics of Data and Analytics* (2016), 254–264.

[5] M. Antoni, A. Ganzer, and P. vom Berge. 2019. Sample of Integrated Labour Market Biographies Regional File (SIAB-R) 1975 - 2017. FDZ-Datenreport, 04/2019 (en). http://doku.iab.de/fdz/reporte/2019/DR_04-19_EN.pdf. https://doi.org/10.5164/IAB.FDZD.1904.en.v1

[6] Mahmood Arai, Moa Bursell, and Lena Nekby. 2008. Between Meritocracy and Ethnic Discrimination: The Gender Difference. https://psycharchives.org/en/item/e228d29b-b537-4951-bb09-a0209fe8e9b1. https://doi.org/10.23668/psycharchives.9038. Accessed December 27, 2022.

[7] Arbetsförmedlingen. 2014. Arbetsförmedlingens Återrapportering 2014: Insatser för att förhindra långvarig arbetslöshet. https://arbetsformedlingen.se/download/18.3e623d4f16735f3976ea22/2.%20Insatser%20f%C3%B6r%20att%20f%C3%\B6rhindra%20l%C3%A5ngvarig%20arbetsl%C3%B6shet%201.0.pdf. Accessed December 27, 2022.

[8] P. Arni and A. Schiprowski. 2015. Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung - Teilprojekt 2: Pilotprojekt Jobchancen-Barometer. Erwartungshaltungen der Personalberatenden, Prognosen der Arbeitslosendauern und deren Auswirkungen auf die Beratungspraxis und den Erfolg der Stellensuche. IZA Research Report No. 70. http://ftp.iza.org/report_pdfs/iza_report_70.pdf

[9] Melanie Arntz and Ralf A. Wilke. 2009. Unemployment duration in Germany: Individual and regional determinants of local job finding, migration and subsidized employment. *Regional Studies* 43, 1 (2009), 43–61.

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[11] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[12] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943. Accessed December 27, 2022.

[13] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[14] Sebawit G. Bishu and Mohamad G. Alkadry. 2017. A systematic review of the gender pay gap and factors that predict it. *Administration & Society* 49, 1 (2017), 65–104. https://doi.org/10.1177/0095399716636928

[15] Dan A. Black, Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel. 2003. Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. *American Economic Review* 93, 4 (2003), 1313–1327.

[16] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[17] H. Buchinger. 2019. AMS Antwort an GBA Arbeitsmarktchancen. https://epicenter.works/document/2055. Accessed December 27, 2022.

[18] Bundesagentur für Arbeit. 2019. Berichte: Blickpunkt Arbeitsmarkt. Arbeitsmarktsituation von langzeitarbeitslosen Menschen. https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Langzeitarbeitslosigkeit/generische-Publikationen/Langzeitarbeitslosigkeit.pdf. Accessed December 27, 2022.

[19] Bundesagentur für Arbeit. 2021. Berichte: Arbeitsmarkt kompakt – Auswirkungen der Corona-Krise auf den Arbeits- und Ausbildungsmarkt. https://statistik.arbeitsagentur.de/Statistikdaten/Detail/202111/arbeitsmarktberichte/am-kompakt-corona/am-kompakt-corona-d-0-202111-pdf.pdf?__blob=publicationFile&v=2. Accessed December 27, 2022.

[20] Bundesagentur für Arbeit. 2022. Einnahmen und Ausgaben des BA-Haushalts. https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Fachstatistiken/Einnahmen-Ausgaben/Produkte/Alle-Produkte-Nav.html. Accessed December 27, 2022.

[21] Dorte Caswell, Greg Marston, and Jørgen Elm Larsen. 2010. Unemployed citizen or 'at risk' client? Classification systems and employment services in Denmark and Australia. *Critical Social Policy* 30, 3 (2010), 384–404.

[22] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. https://arxiv.org/abs/2010.04053. arXiv:2010.04053 [cs.LG].

[23] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of imputation strategies on fairness in machine learning. *J. Artif. Int. Res.* 74 (Sep. 2022), 25 pages. https://doi.org/10.1613/jair.1.13197

[24] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[25] Robyn M. Dawes, David Faust, and Paul E. Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[26] Í. Martínez de Rituerto de Troya, Ruqian Chen, Laura O. Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B. Brás, and Leid Zejnilovic. 2018. Predicting, explaining, and understanding risk of long-term unemployment. In *32nd Conference on Neural Information Processing Systems*.

[27] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. 2019. Statistical Profiling in Public Employment Services. OECD Social, Employment and Migration Working Papers, No. 224. https://doi.org/10.1787/b5e5f16e-en. https://www.oecd-ilibrary.org/content/paper/b5e5f16e-en

[28] Sam Desiere and Ludo Struyven. 2021. Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy* 50, 2 (2021), 367–385. https://doi.org/10.1017/S0047279420000203

[29] Claudia Diehl, Michael Friedrich, and Anja Hall. 2009. Young adults with immigrant background and their transition to the German system of vocational training. The role of preferences, resources, and opportunities. *Zeitschrift für Soziologie* 38, 1 (2009), 48–67. https://doi.org/doi:10.1515/zfsoz-2009-0103

[30] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. 2024. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems* 36 (2024).

[31] M. Dorner, J. Heining, P. Jacobebbinghaus, and S. Seth. 2010. The sample of integrated labour market biographies. *Journal for Applied Social Science Studies* 130, 4 (2010), 599–608.

[32] Nicola Duell, Lena Thurau, and Tim Vetter. 2016. *Long-term Unemployment in the EU: Trends and Policies*. Bertelsmann Stiftung Gütersloh.

[33] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS'12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[34] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[35] B. Fitzenberger, A. Osikominu, and R. Völter. 2005. Imputation rules to improve the education variable in the IAB employment subsample. *Journal for Applied Social Science Studies* 125, 3 (2005), 405–436.

[36] Fenella Fleischmann and Jutta Höhne. 2013. Gender and migration on the labour market: Additive or interacting disadvantages in Germany? *Social Science Research* 42, 5 (2013), 1325–1345. https://doi.org/10.1016/j.ssresearch.2013.05.006

[37] J. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232.

[38] J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* 28, 2 (2000), 337–407.

[39] Duncan Gallie, Serge Paugam, and Sheila Jacobs. 2003. Unemployment, poverty and social isolation: Is there a vicious circle of social exclusion? *European Societies* 5, 1 (2003), 1–32.

[40] Frederic Gerdon, Ruben L. Bach, Christoph Kern, and Frauke Kreuter. 2022. Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society* 9, 1 (2022), 1–13. https://doi.org/10.1177/20539517221089305

[41] William M. Grove, David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment* 12, 1 (2000), 19.

[42] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv:1610.02413 [cs]* (Oct. 2016). http://arxiv.org/abs/1610.02413

[43] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948.

[44] J. Holl, G. Kernbeiß, and M. Wagner-Pinter. 2018. Das AMS-Arbeitsmarktchancen-modell. https://ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

[45] Philip D. Howard and Louise Dixon. 2012. The construction and validation of the OASys Violence Predictor: Advancing violence risk assessment in the English and Welsh correctional services. *Criminal Justice and Behavior* 39, 3 (2012), 287–307.

[46] R. J. Hyndman and G. Athanasopoulos. 2018. *Forecasting: Principles and Practice*. Melbourne: OTexts.

[47] Herwig Immervoll and Stefano Scarpetta. 2012. Activation and employment support policies in OECD countries. An overview of current approaches. *IZA Journal of Labor Policy* 1, 1 (2012), 1–20.

[48] Marita Jacob and Corinna Kleinert. 2014. Marriage, gender, and class: The effects of partner resources on unemployment exit in Germany. *Social Forces* 92, 3 (2014), 839–871.

[49] Peter Jacobebbinghaus and Stefan Seth. 2007. The German integrated employment biographies sample IEBS. *Schmollers Jahrbuch* 127, 2 (2007), 335–342.

[50] Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. 2021. Group-Aware Threshold Adaptation for Fair Classification. arXiv:2111.04271 [cs.LG].

[51] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[52] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 100–109.

[53] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, 247–254. https://doi.org/10.1145/3306618.3314287

[54] Irena Kogan. 2011. New immigrants—old disadvantage patterns? Labour market integration of recent immigrants into Germany. *International Migration* 49, 1 (2011), 91–117.

[55] Katja Kokko, Lea Pulkkinen, and Minna Puustinen. 2000. Selection into long-term unemployment and its psychological consequences. *International Journal of Behavioral Development* 24, 3 (2000), 310–320.

[56] Anton Korinek. 2019. Integrating ethical values and economic value to steer progress in artificial intelligence. *National Bureau of Economic Research*.

[57] John Körtner and Giuliano Bonoli. 2021. Predictive Algorithms in the Delivery of Public Employment Services. https://osf.io/j7r8y/download. Accessed December 27, 2022.

[58] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap is There? https://arxiv.org/abs/2105.01441. arXiv:2105.01441 Accessed December 27, 2022.

[59] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology* 7 (2022). https://doi.org/10.3389/fsoc.2022.883999

[60] Maciej Kuziemski and Gianluca Misuraca. 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy* 44, 6 (2020).

[61] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. The premise, the proposed solutions, and the open challenges. *Philosophy & Technology* 31 (2018), 611–627.

[62] Zhiyuan Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science Advances* 6, 7 (2020).

[63] Artan Loxha and Matteo Morgandi. 2014. Profiling the unemployed: A review of OECD experiences and implications for emerging economies. *Social Protection and Labor Discussion Paper* SP 1424 (2014).

[64] Stephen Machin and Alan Manning. 1999. The causes and consequences of longterm unemployment in Europe. In *Handbook of Labor Economics*, O. C. Ashenfelter and D. Card (Eds.). Vol. 3. Elsevier, Amsterdam, 3085–3139.

[65] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the Applicability of ML Fairness Notions. http://arxiv.org/abs/2006.16745. Accessed December 27, 2022.

[66] Catherine McDonald, Greg Marston, and Amma Buckley. 2003. Risk technology in Australia: The role of the job seeker classification instrument in employment services. *Critical Social Policy* 23, 4 (2003), 498–525.

[67] Kristof Meding and Thilo Hagendorff. 2024. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology* 37, 1 (2024), 4.

[68] Paul E. Meehl. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

[69] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2021). https://doi.org/10.1145/3457607

[70] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[71] David Neumark. 2018. Experimental research on labor market discrimination. *Journal of Economic Literature* 56, 3 (September 2018), 799–866. https://doi.org/10.1257/jel.20161309

[72] Jedrzej Niklas, Karolina Sztandar-Sztanderskal, and Katarzyna Szymielewicz. 2015. Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making. https://panoptykon.org/biblio/profiling-unemployed-poland-social-and-political-implications-algorithmic-decison-making. Accessed December 27, 2022.

[73] Philip J. O'Connell, Seamus McGuinness, and Elish Kelly. 2012. The transition from short- to long-term unemployment: A statistical profiling model for Ireland. *The Economic and Social Review* 43, 1 (2012), 135–164.

[74] P. J. O'Connell, S. McGuinness, E. Kelly, and J. Walsh. 2009. National Profiling of the Unemployed in Ireland. Research Series 10, Economic and Social Research Institute, Dublin. https://www.esri.ie/system/files?file=media/file-uploads/2015-07/RS010.pdf

[75] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT*'19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[77] Eric Potash, Rayid Ghani, Joe Walsh, Emile Jorgensen, Cortland Lohff, Nik Prachand, and Raed Mansour. 2020. Validation of a machine learning model to predict childhood lead poisoning. *JAMA Network Open* 3, 9 (2020). https://doi.org/10.1001/jamanetworkopen.2020.12734

[78] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[79] Tania Raffass. 2017. Demanding activation. *Journal of Social Policy* 46, 2 (2017), 349–365. https://doi.org/10.1017/S004727941600057X

[80] Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani. 2020. Bias and fairness. In *Big Data and Social Science* (2nd ed.), Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane (Eds.). CRC Press, Boca Raton, FL, Chapter 7. https://textbook.coleridgeinitiative.org

[81] Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. arXiv:2308.16681 [stat.ML].

[82] StataCorp. 2017. *Stata Statistical Software: Release 15*. StataCorp LLC, College Station, TX.

[83] R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288.

[84] Markus Viljanen and Tapio Pahikkala. 2020. Predicting unemployment with machine learning based on registry data. In *Research Challenges in Information Science*, Fabiano Dalpiaz, Jelena Zdravkovic, and Pericles Loucopoulos (Eds.). Springer International Publishing, 352–368.

[85] Martijn A. Wijnhoven and Harriët Havinga. 2014. The Work Profiler: A digital instrument for selection and diagnosis of the unemployed. *Local Economy* 29, 6-7 (2014), 740–749.

[86] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, 325–333. http://proceedings.mlr.press/v28/zemel13.html

[87] Eva Zschirnt and Didier Ruedin. 2016. Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies* 42, 7 (2016), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279

Received 17 May 2023; revised 18 May 2024; accepted 5 August 2024