Data & Policy (2023), 5: e32
doi:10.1017/dap.2023.29

RESEARCH ARTICLE

The impact of modeling decisions in statistical profiling

Ruben L. Bach[1], Christoph Kern[2], Hannah Mautner[3] and Frauke Kreuter[2,4]

[1]Mannheim Centre for European Social Research (MZES), University of Mannheim, Mannheim, Germany
[2]Department of Statistics, LMU Munich, Munich, Germany
[3]dmTECH, Karlsruhe, Germany
[4]Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA
Corresponding author: Ruben L. Bach; Email: r.bach@uni-mannheim.de

R.B. and C.K. contributed equally to this paper.

Received: 07 December 2022; **Revised:** 01 August 2023; **Accepted:** 11 September 2023

Keywords: algorithms; artificial intelligence; data science; statistical profiling; unemployment

Abstract

Statistical profiling of job seekers is an attractive option to guide the activities of public employment services. Many hope that algorithms will improve both efficiency and effectiveness of employment services' activities that are so far often based on human judgment. Against this backdrop, we evaluate regression and machine-learning models for predicting job-seekers' risk of becoming long-term unemployed using German administrative labor market data. While our models achieve competitive predictive performance, we show that training an accurate prediction model is just one element in a series of design and modeling decisions, each having notable effects that span beyond predictive accuracy. We observe considerable variation in the cases flagged as high risk across models, highlighting the need for systematic evaluation and transparency of the full prediction pipeline if statistical profiling techniques are to be implemented by employment agencies.

Policy Significance Statement

Data-driven profiling approaches are increasingly used in public administration to guide decisions such as the allocation of welfare state resources. Hopes are high that letting only the data speak will reduce biases humans may have and errors they may make and eventually result in decisions guided by objectivity only. Using statistical profiling of job seekers as an empirical use case, we show that modeling decisions in a typical data science pipeline have profound consequences for who is eventually suggested for support, however. That is, statistical models which promise to take human discretion and biases out of the equation may in fact be far less objective than one might think.

1. Introduction

Statistical profiling offers an increasingly important avenue for informing high-stake policy decisions, such as the allocation of scarce public resources. Examples include the allocation of intervention and supervision resources in criminal justice (Howard and Dixon, 2012), the allocation of in-person

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

investigations in the context of child protection services (Chouldechova et al., 2018), and the allocation of home inspections to identify and control health hazards (Potash et al., 2020). In these scenarios, statistical models are used to provide an initial risk assessment to guide a decision, for example, regarding the question which cases require special attention or which cases should be served first.

Profiling of job seekers is an application where statistical approaches have already been used for almost three decades. There, the goal of profiling is to identify job seekers who have a high risk of becoming long-term unemployed, such that public employment services (PES) can support them in finding a new employment and thereby lower their risk of becoming long-term unemployed. Fighting long-term unemployment (LTU, unemployment that lasts for more than 1 year), is a major societal challenge in many countries (Duell et al., 2016). It has serious consequences for affected individuals, not only in terms of economic deprivation but also regarding their physical and mental health as well as their overall well-being. From a societal perspective, LTU is associated with high costs for health care systems and welfare services. Therefore, being able to identify those with a high risk early on is a desirable goal, for both affected individuals and society.

PES use statistical profiling hoping that data-driven approaches will be more effective and objective as human deciders and to address calls for using the potentials of artificial intelligence in the government sector (Lepri et al., 2018; Engelmann and Puntschuh, 2020). These hopes are backed by psychological studies of decision-making that have demonstrated for more than 60 years that statistical models and simple rule-based predictions outperform humans in tasks such as predicting academic success, job performance, and psychiatric prognosis (Meehl, 1954; Yu and Kuncel, 2020; Kahneman et al., 2021). That is, statistical models are often simply more accurate in predicting, for example, who is likely to recidivate, and such models will do so consistently by returning similar outputs for similar instances.

Detailed documentation of profiling approaches is often not available to the public, however. In some cases, information about the final approach implemented is available, documenting, for example, the performance of the model used to identify those at risk (see, e.g., Allhutter et al., 2020, for a critique). For many approaches, it is difficult to understand why the models were implemented in their current form, how they were evaluated, and which approaches were tested but never made it into deployment. On those grounds, profiling practices received renewed attention in recent years, especially due to discussions about a lack of social responsibility and transparency in data-driven decision-making (Allhutter et al., 2020).

To address this shortcoming, we develop and systematically compare a series of statistical prediction models of LTU in this paper. We draw on German administrative labor market data, which are routinely collected by German PES as a by-product of their activities in supporting job seekers. These data could serve as a basis for statistical profiling in practice. Inspired by approaches already used in other countries, we employ regression and machine-learning methods with different levels of complexity to predict LTU. We then evaluate our prediction models and find that many design decisions made in the prediction pipeline such as who should be targeted and how groups should be defined based on estimated risk scores have major consequences that should be studied and documented. In fact, the models we use and the modeling decisions we make have major consequences regarding the question who is identified as being at risk of LTU and thus may receive special support by employment agencies: Although our models achieve comparable predictive performance, there is large variation in the predictors used by the models and in the cases flagged as high risk. Thus, while statistical models and algorithms promise to be neutral and objective, modeling decisions along the typical data science pipeline still leave more room for leeway than one may expect. Our work highlights the need for public administration agencies to document and evaluate the methodological choices underlying data-driven decision-making systems, which usually remain hidden to the public.

The remainder of this paper is structured as follows: We begin with a brief review of statistical profiling. Next, we present our prediction setup and our results. We conclude with a discussion of our findings where we emphasize that modeling decisions should not only be guided by the desire to train an accurate prediction algorithm. Instead, training an algorithm requires making a myriad of decisions that come with major consequences for those affected by the decisions. Moreover, we highlight that statistical profiling raises new questions related to interpretability, transparency, and fairness.

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

2. Background

PES in many countries use a variety of profiling techniques to fight LTU by *preventing* it through identifying those at risk of becoming long-term unemployed at an early stage (Loxha and Morgandi, 2014). Profiling techniques are usually designed to segment individuals at entry into unemployment into groups with similar risks of resuming work. PES can then *target* specific groups of individuals with treatments such as active labor market policies (measures aimed at increasing an individual's chance of finding a new job such as further vocational training and short classroom training, hiring subsidies for employers and job creation schemes). For example, those identified as having a low probability of resuming work could be given more extensive support than those with a high probability. Other PES may employ strategies that minimize overall LTU by targeting, for example, those who will profit the most.

Several OECD countries developed a variety of statistical profiling techniques with respect to estimating chances of reintegration of the unemployed into the labor market. However, detailed documentation of the statistical models is often not available (for an exception, see Holl et al., 2018). Comprehensive reviews of statistical profiling are presented in Loxha and Morgandi (2014), Desiere et al. (2019), and Körtner and Bonoli (2021). Here, we will briefly review some examples from the literature.

Statistical profiling is evaluated, tested, or used in countries such as Australia (McDonald et al., 2003; Caswell et al., 2010; Loxha and Morgandi, 2014), Austria (Holl et al., 2018), Belgium (Desiere and Struyven, 2021), Denmark (Caswell et al., 2010), Finland (Viljanen and Pahikkala, 2020), Ireland (O'Connell et al., 2009, 2012), the Netherlands (Wijnhoven and Havinga, 2014), New Zealand (Desiere et al., 2019), Poland (Niklas et al., 2015), Portugal (de Troya et al., 2018), Sweden (Arbetsförmedlingen, 2014), and the U.S. (Black et al., 2003). While all rely on statistical prediction approaches to segment users, the outcomes to be predicted vary. Some models are trained to predict LTU (e.g., Belgium, Denmark, and the Netherlands), others are trained to predict exit into employment (e.g., Ireland). Regarding the statistical methods used, most of the profiling approaches are based on logistic regression (LR) models (e.g., Austria, Italy, the Netherlands, and Sweden) but systems based on machine-learning algorithms such as random forests (RFs) and gradient boosting are also used in some countries (e.g., Belgium and New Zealand; Desiere et al., 2019). Most of the approaches are based on administrative labor market data like those used in this paper, but information collected from interviews conducted when newly unemployed individuals register with their PES are also used in some countries. In Belgium, data from interactions with the website of the PES, such as clicking on job vacancies posted on the website, are included as well (Desiere and Struyven, 2021). Regarding prediction performance, accuracy values from 60% to 86% and receiver operating characteristic-area under the curve (ROC-AUC) scores between 0.63 and 0.83 are reported (Desiere et al., 2019). Overall, prediction performance does not seem to vary much between different statistical prediction models (Matty, 2013; Desiere et al., 2019). Compared to human predictions of LTU, statistical models achieve much higher prediction performance, however (Arbetsförmedlingen, 2014; Arni and Schiprowski, 2015). In Germany, statistical profiling is used to a very small extent only. Integrating a more extensive data-driven profiling has been discussed by the agency since the early 2000s (Rudolph and Müntnich, 2001).

Predicting the risk of LTU or similar outcomes (see above) is only the first step in developing a statistical profiling system. The second step is implementing a decision routine based on predicted probabilities. Due to differences in labor market policy and legislative frameworks, there is considerable variation regarding the question which risk groups are targeted by PES, based on their estimated risk scores. To date, many countries appear to target unemployed individuals with a high LTU risk (Desiere et al., 2019).

In this paper, we explore the consequences of allegedly minor modeling decisions when predicting LTU with statistical approaches and using German administrative labor market data. Specifically, we provide an evaluation of several machine-learning models for predicting job-seekers' risk of becoming long-term unemployed (LTU) with respect to the effects typical modeling choices may have. The

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

availability of detailed administrative labor market data, described below, allows us to mimic a realistic use case.

3. Methods

3.1. Data

Data are obtained from German administrative labor market records. These records are maintained by the Research Data Center of the German Federal Employment Agency at the Institute for Employment Research (IAB). Briefly speaking, they contain historic records of labor market activities (employment, unemployment, job search activities, and benefit receipt) for large parts of the German population (about 80% of the German labor force; Dorner et al., 2010). Not included are self-employed individuals and civil servants as their data are processed by a different institution (Jacobebbinghaus and Seth, 2007). Overall, the data are of high quality because they are used by the German Statutory Pension Insurance to calculate pension claims and for assisting unemployed in finding a new job matching their specific profile (Jacobebbinghaus and Seth, 2007). Moreover, even information that is not included in the data such as personality traits and motivation is implicitly captured by the information available (Caliendo et al., 2017).

In more detail, the data contain historic information from 1975 to 2017 on all individuals in Germany who meet one or more of the following conditions: 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 in accordance with the German Social Code Book III (records start in 1975); received long-term benefits in accordance with 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; Antoni et al., 2019). These information are recorded exact to the day and allow the creation of detailed individual labor market histories.

We use a 2% random sample from the administrative labor market records, called *Sample of Integrated Employment Biographies* (SIAB; Antoni et al., 2019). They are *integrated* as they combine information from various sources such as employment information, unemployment information, and unemployment benefits receipt (see previous paragraph). Specifically, we use the factually anonymous version of the SIAB (SIAB-Regionalfile) – Version 7517 v1. 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). Some potentially sensitive information from the original SIAB file were removed in this version of the SIAB data to meet privacy regulations. Still, the data are well suited for the purpose of predicting LTU due to their enormous volume and granularity: they contain detailed employment histories of 1,827,903 individuals documented in a total of 62,340,521 rows of data.

The SIAB dataset comes in a longitudinal form. That is, we often observe multiple entries per person. Each time a person enters a new relevant labor market status (e.g., registered as unemployed or started a job subject to social security), a new entry is created. On average, we observe more than 34 data points for each of the nearly 2 million individuals. Note that it 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 a data point, 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 data points referring to the period between January 1, 2010 and December 31, 2016. We exclude data referring to periods *prior to 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

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

unemployed people were supported by the German PES. In addition, new types of data were added to the SIAB during that time that capture individuals' labor market behavior in response to the reforms.

Data collected *after 2016* are excluded because our objective is to predict unemployment that lasts for at least 1 year. Therefore, the last year of labor market histories available is needed to determine whether an individuals who became unemployed by the end of 2016 became long-term unemployed or not. While one could include unemployment periods that start 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. Therefore, we consider only unemployment episodes that started before 2017.

In addition, we remove 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 completely irrelevant. These restrictions leave us with 303,724 unique individuals and 643,690 unemployment episodes. Since individuals may become unemployed more than once during the period of observation, some individuals may contribute more than one unemployment episode to our data.

3.1.1. Definition of long-term unemployment

We follow 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 1 year. Participation in labor market measures for the unemployed as well as periods of sickness or interruptions for other reasons of up to 6 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 1 year. Specifically, episodes are relevant if 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). If an unemployment episode's duration is less than 1 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 and they allow us to identify the exact date of the start of an unemployment episode. Moreover, as the records we use are historic, we also observe the end date of an unemployment episode, which allows us to obtain the exact duration of an unemployment period and therefore to identify LTU. Note that the end of an unemployment episode does not necessarily imply exit into employment. It is also possible that unemployed individuals simply no longer consult with the PES.

Using the definition from above, 97,599 (15.2%) out of a total of 643,690 unemployment episodes identified in the data are LTU episodes. Regarding individuals, we find that 79,361 (26.1%) out of the 303,724 individuals in our data who ever became unemployed during the period considered experienced LTU at least once. LTU episodes and the number of affected individuals by year are shown in Table 1.

Table 1. *LTU episodes and affected individuals, by year*

	Unemployment episodes	LTU episodes	Individuals	Individuals experiencing at least one LTU episode
2010	105,137	15,872 (15.1%)	91,405	15,872 (17.4%)
2011	95,597	14,813 (15.5%)	83,177	14,813 (17.8%)
2012	90,408	14,865 (16.4%)	79,154	14,865 (18.8%)
2013	88,988	14,324 (16.1%)	78,527	14,324 (18.3%)
2014	87,158	13,529 (15.5%)	76,747	13,529 (17.6%)
2015	86,692	12,688 (14.6%)	76,187	12,688 (16.7%)
2016	89,710	11,508 (12,8%)	78,373	11,508 (14.7%)
Total	643,690	97,599 (15.2%)	303,724	79,361 (26.1%)

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

Overall, the annual risk rates of entering LTU as shown in Table 1 roughly match official rates of entry into LTU reported by the German PES (Bundesagentur für Arbeit, 2019).

3.1.2. Predictors

To predict an individual's risk of LTU when becoming unemployed, we require a dataset in a one-observation-per-unemployment-episode form. That is, we consider the risk of LTU separately for each unemployment episode found in our data. We do not consider LTU on a per-individual basis because individuals can become unemployed more than once. We prefer our per-unemployment-episode solution to a per-individual solution as a new profiling would be conducted each time an individual registers as unemployed with the PES.

Since the SIAB data come in a longitudinal form, we often observe multiple data points per individual prior to becoming unemployed. To reduce these multiple observations to one observation per unemployment episode and individual, we count, for example, the number of unemployment episodes an individual experienced in the past or the total duration of 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, for example, 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 cited in Section 2.

Socio-demographic information is derived in two ways. For information such as age, we consider the most recent data point containing the information prior to or at entry into an unemployment episode. For information such as education, we consider the highest value observed prior to or at entry into an unemployment episode as these characteristics are sometimes measured with some inconsistencies (Fitzenberger et al., 2005). Note that we do not consider gender and nationality as predictors in our main set of models (see Section 5 for a discussion and Mehrabi et al., 2021 for a review of fairness issues in machine learning and automated decision-making). We will refer to an additional set of models that include those features when presenting the results.

In summary, our feature generation procedures ensure that only information observed at or before entry into unemployment is considered for predicting LTU. Table 2 presents groups of predictors (socio-demographics, labor market history, and last job) with examples for each group. Due to the large number of predictors (157), we list all predictors in the Supplementary Materials only (Table A.1 in the Supplementary Material).

Table 2. *Groups of predictors, with examples*

Group	Example predictors
Socio-demographics	Age, state of residence, education
Labor market history	Total duration of unemployment episodes
	Mean duration of employment episodes
	Total duration of job seeking episodes, scaled by age
	Industry worked in the most
	Time since last employment episode
Last job	Industry
	Duration of employment
	Skill-level required for last job
	More than one job
	Part-time/full-time/marginal employment
	Fixed-term employment
	Inflation-deflated average daily wage

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

3.2. Prediction setup

We use the outlined variables to predict the risk of LTU for an individual unemployment episode. Specifically, the prediction task includes the following components:

- *Set of features* X. This set includes all predictors that are presented in Section 3.1.2.
- *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 $Pr(Y=1|X)$. The predicted risk of becoming long-term unemployed based on a given prediction model.
- *Prediction* $\widehat{Y} \in \{0,1\}$. Binary prediction of becoming long-term unemployed ($\widehat{Y}=1$) and not becoming long-term unemployed ($\widehat{Y}=0$). 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 $\widehat{Y}=1$ 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 $\mathbf{r}$:

$$\widehat{Y}^{(h_a)} = 1 \, \text{if} \, R \geq c_{10}, \text{else} \, 0.$$

Policy 1b (P1b). Assign $\widehat{Y}=1$ 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 $\mathbf{r}$:

$$\widehat{Y}^{(h_b)} = 1 \, \text{if} \, R \geq c_{25}, \text{else} \, 0.$$

Use cases from other countries vary regarding the choice of the classification thresholds/policies (Loxha and Morgandi, 2014; Desiere and Struyven, 2021). It is likely that budget constraints in a country's PES, such as the number of support measures to be distributed and the costs per supported job seeker, will be decisive factors in determining the actual threshold to be used in a statistical profiling system. To address this issue, the two policies we designed here represent different capacity implications. That is, P1b would flag 2.5 as many unemployment episodes as high risk than P1a and recommend many more cases for support by employment agencies, likely resulting in much higher costs.

3.2.1. Prediction models

We consider four prediction methods that are used by PES across the globe for predicting LTU. In addition to regression approaches, we focus on prominent *machine learning* or ensemble methods that are typically well-suited for prediction tasks with many features and potentially complex relationships (see also the prediction methods used in de Troya et al., 2018). In the following, we will refer to the regression approaches as *regression methods* and to ensemble techniques as *machine-learning methods.* Besides differences in flexibility, the methods also differ in the interpretability of the resulting models, which can be a decisive factor when algorithmic profiling of job seekers is put to practice by PES. In summary, we compute predictions based on:

- *Logistic regression.* Common (unpenalized) LR, only main effects for all predictors are included. Results in an interpretable set of coefficients, and is included as a benchmark.
- *Penalized logistic regression (PLR).* LR with a penalty on the (ℓ_1, ℓ_2) norm of the regression coefficients (Tibshirani, 1996). In the former case (ℓ_1 penalty), a more parsimonious model compared to unpenalized LR can be returned, which may increase both interpretability and prediction performance.

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

- *Random forest.* Machine-learning technique relying on ensemble of deep (uncorrelated) decision trees grown on bootstrap samples (Breiman, 2001). Results in a model that cannot be readily interpreted without further helper methods.
- *Gradient boosting machines (GBM).* Machine-learning technique relying on ensemble of small decision trees that are grown in sequence by using the (updated) pseudo-residuals in each iteration as the outcome (Friedman et al., 2000; Friedman, 2001). Similar to RF, additional techniques are typically needed to support the interpretation of results.

As outlined in Section 3.1, our SIAB data includes information from the beginning of 2010 up to the end of 2016. To mimic a realistic profiling task, we build prediction models with *training data* covering the years 2010–2015, and we compare and evaluate the resulting models with *evaluation data* from 2016. To ease computational burden related to model tuning (see below), a random sample of 20,000 unemployment episodes from each training year (2010–2015) is drawn to construct the training set. Final model evaluation is done on the full data from 2016 (89,710 episodes).

With the exception of unpenalized LR, we need to tune hyperparameters to build a well performing model for a given task. The hyperparameter settings considered for each method are listed in Table A.2 in the Supplementary Material. The respective best setting for each method is selected based on temporal cross-validation (Hyndman and Athanasopoulos, 2018). Training and test sets are constructed from the 2010 to 2015 data by successively moving the time point which separates the fit and test period forward in time. While this leads the training data 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 includes data from 2010 to 2014, and the last test period data from 2015. These data sets are used to repeatedly train and test models for each method-hyperparameter combination.

The hyperparameter setting that results in the best average test performance over time for each method is selected to re-train the respective final model with the *full training data* (2010–2015). Furthermore, we re-train a second set of models with *restricted training data* using only the year 2015. This is done to explore the performance implications of training LTU models with different training data histories. One might naturally expect limited performance of the second set of models as they have access to fewer training examples. However, another perspective is that the structural associations between the predictors and the outcome could change over time (“concept drift”), such that training only with newer data could be beneficial.

We used Stata (15, StataCorp, 2017) and R (3.6.3, R Core Team, 2020) for data preparations. Model training and evaluation were done with Python (3.6.4), using the scikit-learn (0.19.1, Pedregosa et al., 2011) package. Code for replication purposes is available at the following OSF repository: https://osf.io/ke538/?view_only=a1f08b62b18a4ef49b8fd97d5158c53c.

3.2.2. *Performance metrics*

A key aspect when considering statistical profiling of job seekers is whether the underlying prediction models can accurately identify individuals that face high LTU risks. This perspective considers accurate predictions as a prerequisite for optimal allocation of support programs to unemployed individuals. Prediction performance may be evaluated based on the predicted classes $\hat{Y}$ (accuracy, precision, and recall) or based on risk scores R (ROC-AUC and PR-AUC). For class predictions, it is important, not only to understand the profiling models’ overall accuracy, but also specifically their ability to identify true LTU episodes (recall) and the efficiency of their LTU predictions (precision). In the present setting, low recall would imply that many job seekers with high LTU risks may not be targeted with support, while low precision implies an inefficient allocation of support measures to job seekers that might not need it. We use risk-score based metrics to evaluate the models’ prediction performance across all applicable classification thresholds with respect to overall discrimination (ROC-AUC) and with a focus on the LTU class predictions (PR-AUC).

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

- *Accuracy.* Classification accuracy of predictions compared to observed outcomes. In range $[0,1]$:

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{y}_i = y_i),$$

where we sum over $i = 1, \ldots, n$ instances and $\mathbf{1}$ represents the indicator function.

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

$$\mathrm{Prec} = \frac{1}{k}\sum_{i=1}^{n} y_i \mathbf{1}\left(r_i \geq r_{[k]}\right),$$

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

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

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

- *ROC-AUC.* Area under the ROC curve. In range $[0,1]$, with 0.5 representing a random model.
- *PR-AUC.* Area under the precision-recall curve. In range $[0,1]$.

3.2.3. Interpretation techniques

In addition to prediction performance, different prediction models—ranging in complexity—can be compared with respect to the eventual lists of job seekers that are predicted as high risk and regarding the way the training data are utilized. We evaluate models based on the similarity of their LTU predictions using Jaccard similarities. We focus on Jaccard similarity as we consider our outcome of interest to be asymmetric (we are interested in the overlap among positive predictions) and different types of non-matches do not need to be weighted differently (as we are comparing classifiers and not against ground truth). This investigation allows us to evaluate whether different models may result in (dis)similar lists of job seekers that would eventually be targeted with support programs. Against this background, extracting which features were most important in the model building process is very informative for our purposes. From a methodological perspective, it allows us to judge the models' dependence on certain types of inputs. Practically, it provides supporting information on the models' (potentially dissimilar) decisions, which we use to demonstrate the variation that models similar in performance can generate.

- *Jaccard similarities.* Intersection divided by the union of list of LTU predictions ($\hat{y}_i$) of two models a and b (Tan et al., 2018):

$$JS = \frac{|\hat{y}_i{}^{(a)} \cap \hat{y}_i{}^{(b)}|}{|\hat{y}_i{}^{(a)} \cup \hat{y}_i{}^{(b)}|}.$$

- *Permutation feature importance.* Decrease in model performance when randomly shuffling a feature (Fisher et al., 2019):

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

$$FI_j = s - \frac{1}{R}\sum_{r=1}^{R} s_{r,j},$$

where j is the feature of interest, s is the performance score, and we sum over $r = 1, \ldots, R$ repetitions. We use 10 repetitions and ROC-AUC as the performance measure.

4. Results

4.1. Performance comparison

We present the prediction performance of the trained models in two steps. First, the model selection criterion and results from the temporal cross-validation procedure (with data from 2010 to 2015) are discussed to provide some context on the final prediction models that were chosen in this process. Second, we present the performance of the selected models in the evaluation set (data from 2016).

4.1.1. Temporal cross-validation

Model selection for PLR, RF, and GBM was done based on the average ROC-AUC over all test periods in the temporal cross-validation loop. Specifically, the respective hyperparameter setting with the highest average ROC-AUC was chosen for each method. The selected settings are highlighted in Table A.2 in the Supplementary Material. Note that a modest penalty on model complexity was optimal for PLR, which resulted in a final model with (still) 131 non-zero regression coefficients.

Table A.4a in the Supplementary Material shows the overall hold-out ranking performance (ROC-AUC) of the selected best models for each method over time. Overall, we observe ROC-AUC's in the range [0.694,0.774], which largely aligns with performance results that have been reported for LTU prediction in other countries (e.g., Belgium and New Zealand, Desiere et al., 2019). In comparison, we see that the LR models are slightly outperformed by PLR, RF, and GBM. Given the difference between LR and PLR, we suspect that this is likely due to model specification issues (e.g., many correlated and potentially uninformative predictors) in the unpenalized, "naive" logistic model. In addition, we observe a mildly positive trend for most models (except LR) with increasing test set ROC-AUC's over time, indicating that the models benefit from the increasing amount of historical training data that is used as we progress to more recent years.

Classification performance of the selected best models based on policies 1a and 1b is shown in Tables A.4b and A.4c in the Supplementary Material. That is, precision and recall are computed based on class predictions in which unemployment episodes with risk scores that are within the top 10% (policy 1a) or top 25% (policy 1b) of all scores are classified as LTU episodes. Generally, the best results are achieved with PLR, RF, and GBM (except for the first test set, 2011), with GBM consistently showing the highest precision and recall scores. As with ROC-AUC, the recall scores of PLR, RF, and GBM tend to increase over time, while there is no clear trend of the corresponding precision scores.

4.1.2. Evaluation set

Ranking and classification performance metrics of the selected best prediction models (re-trained with data from 2010 to 2015) for the evaluation set (data from 2016) are listed in Table 3. Starting with overall ranking performance (Table 3a), the findings confirm the temporal cross-validation results, with PLR, RF, and GBM outperforming the unpenalized LR model. Among those three best approaches, we see little differences in ROC-AUC scores. However, the tree-based machine-learning models improve over PLR in terms of PR-AUC, indicating higher precision over the range of applicable classification thresholds.

Classification performance of prediction models is shown in Tables 3b and c. GBM consistently performs best under both policies. Under policy 1a, 29% of all observed LTU episodes are correctly detected and classified as such by the GBM model (recall). Conversely, 37.2% of all episodes that are

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

Table 3. *Prediction performance of selected prediction models (in 2016), trained with 2010–2015 data*

(a) Ranking performance

	ROC-AUC	PR-AUC
LR	0.700	0.256
PLR	0.760	0.298
RF	0.763	0.312
GBM	0.770	0.325

(b) Classification performance based on policy 1a

	Accuracy	Precision	Recall
LR	0.837	0.328	0.256
PLR	0.842	0.351	0.274
RF	0.845	0.366	0.286
GBM	0.846	0.372	0.290

(c) Classification performance based on policy 1b

	Accuracy	Precision	Recall
LR	0.745	0.246	0.479
PLR	0.767	0.290	0.565
RF	0.766	0.289	0.564
GBM	0.770	0.296	0.577

predicted by the GBM model to be LTU episodes are indeed LTU episodes (precision). Under policy 1b, GBM's recall increases considerably to 57.7%, however, at the cost of a decrease in precision to 29.6%. Nonetheless, differences in classification performance across the prediction models remain small. We also note that in- or excluding sensitive features (gender and nationality) from the prediction models hardly makes a difference in our setting, as we observe only very modest improvements in performance when models had access to additional (sensitive) attributes at training time (see Table A.6 in the Supplementary Material).

Complementing results for models that were trained using only data from 2015 are shown in Table A.5 in the Supplementary Material. Overall, it can be seen that there is a modest decrease in prediction performance in terms of ROC-AUC and PR-AUC when restricting the training data to only include the most recent year before the evaluation date cutoff. Although some improvements in precision and recall (at top 10%) can be observed for LR and PLR, the GBM model still shows the best results both with respect to overall ranking and classification performance. Precision and recall curves over the full range of applicable classification thresholds for all eight prediction models (LR, PLR, RF, GBM trained with 2010–2015 and 2015 data, respectively) are presented in Figure A.1 in the Supplementary Material.

4.2. Model interpretation

Next, we compare the prediction approaches with respect to the way they make use of the training data. We compare similarities in the LTU predictions and feature importance rankings to study whether choosing between regression and machine-learning methods leads to structurally different LTU prediction models. This tells us how much variation models that are similar in prediction performance produce when it comes to using the available information and flagging individuals as high or low risk.

Table 4. *Jaccard similarities between class predictions (in 2016)*

(a) Classification based on policy 1a

	LR	PLR	RF	GBM
LR	1.000			
PLR	0.544	1.000		
RF	0.540	0.562	1.000	
GBM	0.448	0.560	0.651	1.000

(b) Classification based on policy 1b

	LR	PLR	RF	GBM
LR	1.000			
PLR	0.539	1.000		
RF	0.571	0.718	1.000	
GBM	0.529	0.730	0.814	1.000

Table 4 presents Jaccard similarities between LTU predictions. It can be seen that the different models often result in considerably different lists of unemployment episodes that are classified as LTU. Among the prediction models, the agreement between the two tree-based machine-learning approaches is highest, with a similarity of the two lists of 65.1% (policy 1a) and 81.4% (policy 1b). Considerable differences can be observed among the other prediction approaches, especially when comparing the predictions of the unpenalized LR to PLR, RF and GBM. Even when comparing LR to PLR we only observe Jaccard similarities of 54.4% (policy 1a) and 53.9% (policy 1b) and thus considerable disagreement in the lists of job seekers that might eventually be supported by employment agencies given the models' predictions. While the general pattern of (dis)similarities matches the previously outlined performance results (i.e., the performance scores of RF and GBM are closest, and LR performs worst), the degree of disagreement in the LTU predictions is notable. This suggests that the similar accuracy and precision scores, in combination with the rather low recall, of PLR, RF and GBM under policy 1a is driven by the models identifying different subsets of the full list of true LTU episodes in the test set. For less restrictive classification thresholds (i.e., policy 1b), recall as well as the overlap between the predicted high-risk sets increases. This is supported by the comparably small Jaccard similarities between the LTU predictions and the *true outcome* (observed LTU classes), which range from 0.168 (LR) to 0.196 (GBM) under policy 1a and from 0.194 (LR) to 0.243 (GBM) under policy 1b.

Figure 1 shows the top 10 feature importance of the four prediction models (LR, PLR, RF and GBM). Each prediction model utilizes the set of predictors in different ways, especially when comparing the regression approaches (LR and PLR) to the machine-learning models (RF and GBM). It is evident that age is the most important predictor in both tree-based machine-learning techniques, RF and GBM, followed by various features that are predominately characterizing the (un)employment history of job seekers (e.g., the total duration of previous job seeking episodes and previous receipt of unemployment benefits). For LR and PLR, features about the unemployment history are dominant in the four most important predictor, including the total duration of previous unemployment benefits receipt episodes, the total duration of previous employment episodes (both unadjusted and scaled by age), and the total duration of previous job seeking episodes (scaled by age). For these models, job-seekers' age on its own plays a less prominent role. The differences in the top features may explain part of the dissimilarities in the LTU predictions across model types as observed above. Also note that the importance scores of the top features vary considerably across the different methods, indicating that, for example, RF and GBM are less dependent on their most important features in terms of performance, compared to LR and PLR. In summary, the results indicate that information about previous unemployment episodes in combination with age is decisive for accurate LTU predictions.

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

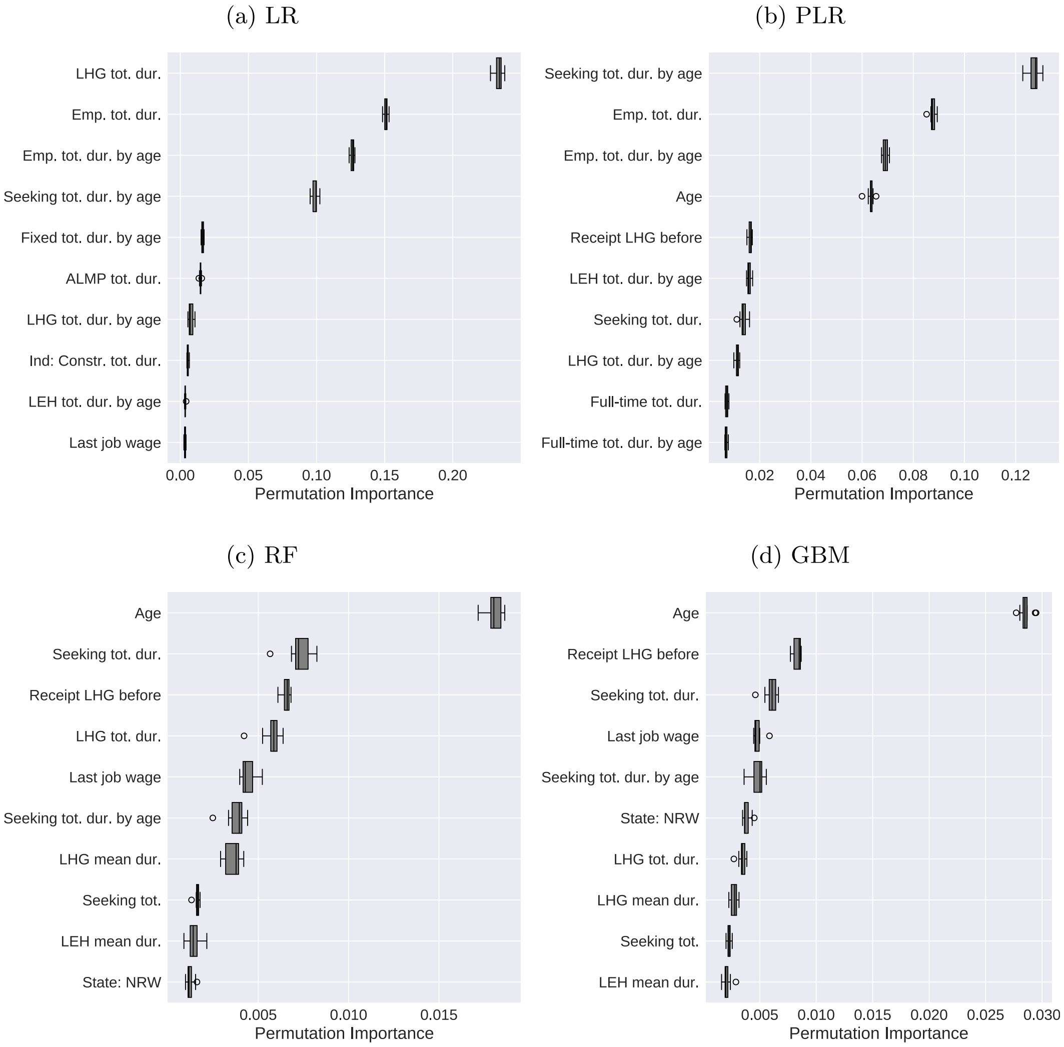

Figure 1. Top 10 feature importance for selected prediction models (in 2016). Panel (a) shows feature importance for the logistic regression approach, panel (b) shows feature importance for the penalized logistic regression approach, panel (c) shows feature importance for the random forest approach, and panel (d) shows feature importance for the gradient boosting machines approach. See Table A.1 in the Supplementary Material for a detailed description of the variable labels.

Table 5 provides further insights into the composition of job seekers that are predicted to become long-term unemployed. The predicted LTU episodes differ in various characteristics. The predictions of LR are most distinct with respect to age and focus on older job seekers under policy 1a, and younger individuals under policy 1b, compared to PLR, RF, and GBM. The prediction profiles of PLR, RF and GBM are similar under policy 1b, but show notable differences under (the more strict) policy 1a. Episodes that are classified as LTU by PLR belong to job seekers who are younger and earned less in their last job, compared to job seekers that would be targeted based on RF and GBM. GBM, in turn, focuses on job seekers who spent more time in employment, compared to RF. The different weighting of features across the prediction models thus has notable consequences on which group of job seekers is eventually classified to be at high risk of LTU and thereby potentially eligible to support programs. In summary, older job seekers would benefit more from a profiling approach based on LR and policy 1a, whereas job

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

Table 5. *Arithmetic means of selected features for predicted LTU episodes (in 2016)*

(a) Classification based on policy 1a

	Seeking	Employment	Last job					
	Total duration	Total duration	Wage	Age	Edu.: High	Non-German	Female	No. of obs.
LR	2,719.72	3,094.92	24.95	47.75	0.07	0.09	0.49	8,972
PLR	2,405.34	2,391.68	24.90	45.05	0.07	0.10	0.50	8,972
RF	2,450.67	2,650.73	28.31	45.53	0.08	0.11	0.51	8,972
GBM	2,270.42	2,967.83	29.17	45.88	0.09	0.10	0.51	8,972

(b) Classification based on policy 1b

	Seeking	Employment	Last job					
	Total duration	Total duration	Wage	Age	Edu.: High	Non-German	Female	No. of obs.
LR	1,677.85	2,578.46	28.15	40.45	0.08	0.21	0.45	22,428
PLR	1,722.03	2,841.26	30.51	44.75	0.09	0.15	0.49	22,428
RF	1,834.78	2,910.65	30.90	44.07	0.09	0.12	0.48	22,428
GBM	1,762.01	2,874.23	30.44	43.89	0.10	0.12	0.49	22,428

seekers with comparably short job seeking histories would benefit from GBM-based profiling (under the same classification policy). Note that for the same model, varying the classification threshold (comparing policy 1a vs. 1b) also induces considerable differences in the composition of job seekers who would be prioritized.

We further compare the predicted LTU episodes with respect to two attributes that were not used in the main set of prediction models: Nationality (coded as German and non-German) and gender (male and female in our data). While there are little differences along those dimensions across models under policy 1a, we observe that the LR and PLR predictions, compared to RF and GBM, are more often targeting non-German job seekers under policy 1b. This highlights that the selection of the prediction model has implications that span beyond the attributes that were used for model training. We also note that the outlined pattern is similarly observed when the models were allowed to explicitly use sensitive attributes as additional features (see Table A.7 in the Supplementary Material).

5. Discussion

In this paper, we set out to explore the use of statistical profiling for identifying job seekers at risk of becoming long-term unemployed. Specifically, we were interested in the question how different prediction models and supposedly minor methodological choices along a typical data science pipeline influence predictions from the models and the classifications based on them. We developed a series of regression and machine-learning models that predict a job seeker's risk of LTU using German administrative labor market data that are routinely collected by German PES, thereby, mimicking a realistic use case.

We could show that our algorithms achieve competitive predictive performance compared to approaches already used in other countries. We also found that predictive performance did not vary much depending on the model type used and time frame of the training data. However, while results showed that information about the (un)employment history of job seekers and characteristics of their previous unemployment spells is decisive for correctly predicting the risk of LTU, the importance of structural predictors varied considerably between the implemented models. There, age was the most important feature in both machine-learning approaches (random forest and gradient boosting). Despite

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

largely similar prediction performance, we observed that applying different profiling techniques led to considerable differences in the lists of job seekers that are eventually predicted to be at high risk of LTU. These disagreements, even among the regression and machine-learning models and across classification thresholds, underline that (even small) methodological decisions can have profound impacts in practice if such models were used to, for example, allocate labor market support programs based on predictions.

Hopes are high that data-driven algorithmic approaches will make decision processes more objective and efficient. Our results, however, indicate that predictive multiplicity (Marx et al., 2020) in statistical profiling contexts can introduce (new) forms of ambiguity that can lead to conflicting predictions and thus decisions for the same individual just by switching between equally accurate model types. In this case, model selection becomes a highly impactful task, and potentially measures that span beyond pure prediction accuracy should be considered to carefully investigate the downstream implications of different modeling regimes. Thus, we highlight that the discretion developers have in deciding, for example, which prediction model to use can have major consequences for the classifications based on them. Statistical models which promise to take human discretion and biases out of the equation may in fact be far less objective than one might think.

Implementing a statistical profiling approach in reality requires making many design and policy decisions beyond pure technical details of the statistical prediction model. Decisions such as who should be targeted and how groups should be defined based on estimated risk scores have major consequences for the quality of the predictions. For example, determining the classification threshold for discriminating between LTU and non-LTU cases cannot be done without consulting a country's socio-institutional context, for example, in terms of labor market policies, legislation and budget constraints. Here, we cannot make recommendations on which threshold should be used, but we highlight that different decisions imply different precision-recall trade-offs and structural differences in the composition of the classified job seekers. Similarly, as we discussed above, different techniques quickly result in considerable differences in the cases that are eventually predicted to be at high risk of LTU. Again, methodological decisions that do not seem to matter much from a pure predictive performance perspective can have profound impacts in practice. We do not believe that policy makers need to be experts in training statistical models, but those who are should closely collaborate with them to reach informed decisions regarding the distribution of errors and the consequences that different design decisions may have. Moreover, documentation of the choices made and the alternatives evaluated is more important than ever before.

Awareness of the consequences of design decisions in profiling approaches requires additional attention when considering fairness, an aspect that we did not study in detail in this paper. Statistical prediction models can quickly learn associations such as women and people with a migration background being disproportionately affected by unemployment and having lower job prospects. If, as a result, women and people with a migration background receive a different LTU risk score than men and non-migration background individuals, using these scores to prioritize individuals may result in discrimination based on gender and nationality. Arguably, simply excluding such attributes will not remove potential discrimination as predictions could nonetheless be biased if labor market histories of women and men are distinct. Thus, a key challenge for future research on the documentation and evaluation of statistical profiling models for labor market policy and other purposes will be to ensure that no unintended negative consequences for affected individuals arise, especially regarding discrimination. While research on such topics has received increased attention in recent years (see, e.g., Mehrabi et al., 2021), it has only recently caught attention in the literature on statistical profiling of the unemployed (Allhutter et al., 2020; Desiere and Struyven, 2021).

Likewise, addressing questions related to explainability and transparency (XAI) is an important avenue for future research on statistical profiling. While, once trained, both regression and machine-learning-type predictive models can calculate risk predictions in a matter of seconds, designing systems in ways that allow for human understanding of how the predictions were calculated remains challenging, at least, for machine-learning approaches. However, legislation like the General Data Protection Regulation, for example, requires statistical profiling tools to be transparent and explainable (GDPR art. 13 and 14), that is, that an individual affected by profiling has the right to "meaningful information about the logic involved." While a

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

profiling system based on a traditional econometric regression model may be easily understood, conditional on some technical fluency, machine-learning approaches are often *black-box-approaches* where the underlying model cannot be easily articulated and understood by a human. In our case, we found that gradient boosting outperformed PLR. However, as a machine-learning technique, boosting cannot be easily interpreted. Thus, there seems to a trade-off between performance and transparency and explainability in our statistical profiling approaches. On the one hand, as differences in predictive performance between the best regression and the best machine-learning technique are small, PES may be willing to accept slightly inferior predictive performance of traditional regression models in exchange for increased transparency. On the other hand, making black-box algorithms transparent and explainable is an active field of research in the machine-learning community (see, e.g., Arrieta et al., 2020). Thus, investigating and eventually achieving explainability and transparency in statistical prediction models of LTU is another avenue for future research. Such research may include both the perspectives of those relying on the recommendations of prediction models (case-workers in job agencies) and of those being affected by subsequent decisions (job seekers). Moreover, questions regarding the integration of statistical profiling approaches into PES' technical and organizational frameworks will require more research in the future. Here, approaches from the human–computer–interaction literature may prove to be useful.

Another topic worthy of further exploration is how changes in social and economic circumstances affect model performance, a concept sometimes referred to as *drifting* (see Section 3.2 and Salganik, 2019, Chapter 2.3.7). That is, changes in economic circumstances, such as employment shocks due to the financial crisis in 2007/2008 and the COVID-19 pandemic may affect how well a system performs and how consistent predictions are across different prediction models as structural associations between the predictors and the outcome may change in such times. While beyond the focus of our paper, our results suggest that, on the one hand, model performance did not change much when restricting training data to just 1 year. That is, using fewer but more recent data points does not decrease model performance and consistency in predictions across models. On the other, the data we considered did not cover major economic changes such as those caused by recent crises. Therefore, we can only speculate whether and how much drifting would be a challenge for statistical profiling. From a practical perspective, it seems that the efforts of re-training, evaluating, and deploying new models should be significantly lower once a prediction-based system has been deployed and integrated into PES' IT frameworks. With more and more research data being available for prolonged time periods and in timely manner, investigating drifting and its consequences will be an interesting and highly relevant avenue for future research.

There are several limitations to our study. We cannot evaluate how our results compare to current profiling approaches used by the German PES. No information on the prediction performance of current profiling approaches in Germany are available. However, previous literature studying human predictions versus predictions from simple rule-based and statistical models as well as the few studies that compared case worker-based and statistical profiling in other countries (see Sections 1 and 2) clearly show that statistical models outperform humans in various prediction tasks. The prediction performance of our statistical profiling approach is comparable to those of other countries (Desiere et al., 2019). For these reasons, similar conclusion regarding the superiority of statistical profiling to case worker-based profiling in terms of prediction performance may be reached for Germany if we were able to compare the approaches. In addition, we believe that our results provide only a lower bound in terms of prediction performance as we relied on an anonymized version of the administrative labor market data and because we did not include further external information such as regional unemployment rates and job vacancies in our models. Overall, our results demonstrate that statistical profiling for labor market policy is a topic worth of further exploration, particularly with respect to improving its transparency, explainability, and fairness.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/dap.2023.29.

Acknowledgments. A preprint covering some of the work reported in this paper is available as Kern et al. (2021).

Funding statement. This work was supported by the Baden-Württemberg Foundation under grant “Fairness in Automated Decision-Making—FairADM” and by the Volkswagen Foundation under grant “Consequences of Artificial Intelligence for Urban Societies (CAIUS).” The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare no competing interests exist.

Author contribution. Conceptualization: R.L.B., C.K., and F.K.; Data curation: R.L.B. and H.M.; Formal analysis: C.K. and H.M.; Funding acquisition: R.L.B., C.K., and F.K.; Investigation: R.L.B., C.K., and H.M.; Methodology: R.L.B., C.K., and H.M.; Project administration: R.L.B., C.K., and F.K.; Resources: R.L.B., C.K., and H.M.; Software: R.L.B., C.K., and H.M.; Supervision: R.L.B., C.K., and F.K.; Validation: R.L.B., C.K., and F.K.; Visualization: C.K.; Writing—original draft: R.L.B. and C.K.; Writing—review and editing: R.L.B., C.K., H.M., and F.K.

Data availability statement. 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). DOI: https://doi.org/10.5164/IAB.SIAB-R7517.de.en.v1.

References

Allhutter D, Cech F, Fischer F, Grill G and Mager A (2020) Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data 3*(5), 1–17.

Antoni M, Ganzer A and vom Berge P (2019) Factually Anonymous Version of the Sample of Integrated Labour Market Biographies (SIAB-Regionalfile)—Version 7517 v1. Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB). Available at https://doi.org/10.5164/IAB.SIAB-R7517.de.en.v1 (accessed 12 September 2023).

Arbetsförmedlingen (2014) Arbetsförmedlingens Återrapportering 2014: Insatser för att förhindra långvarig arbetslöshet. Available at 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 12 September 2023).

Arni P and Schiprowski A (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. Available at http://ftp.iza.org/report_pdfs/iza_report_70.pdf (accessed 12 September 2023).

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R and Herrera F (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion 58*, 82–115.

Black DA, Smith JA, Berger MC and Noel BJ (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), 1313–1327.

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

Bundesagentur für Arbeit (2019) *Berichte: Blickpunkt Arbeitsmarkt*. Arbeitsmarktsituation von langzeitarbeitslosen Menschen. Accessed at https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Langzeitarbeitslosigkeit/generische-Publikationen/Langzeitarbeitslosigkeit.pdf (accessed 12 September 2023).

Caliendo M, Mahlstedt R and Mitnik OA (2017) Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies. *Labour Economics 46*, 14–25.

Caswell D, Marston G and Larsen JE (2010) Unemployed citizen or ‘at risk’ client? Classification systems and employment services in Denmark and Australia. *Critical Social Policy 30*(3), 384–404.

Chouldechova A, Benavides-Prado D, Fialko O and Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. New York, NY: PMLR.

de Troya ÍMdR, Chen R, Moraes LO, Bajaj P, Kupersmith J, Ghani R, Brás NB and Zejnilovic L (2018) Predicting, explaining, and understanding risk of long-term unemployment. In *NeurIPSWorkshop on AI for Social Good*, Montréal, Canada. Available at https://aiforsocialgood.github.io/2018/pdfs/track1/97_aisg_neurips2018.pdf (accessed 12 September 2023).

Desiere S, Langenbucher K and Struyven L (2019) Statistical Profiling in Public Employment Services. OECD Social, Employment and Migration Working Papers, No. 224. Available at https://www.oecd-ilibrary.org/content/paper/b5e5f16e-en (accessed 12 September 2023).

Desiere S and Struyven L (2021) Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy 50*(2), 367–385.

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

Duell N, Thurau L and Vetter T (2016) *Long-Term Unemployment in the EU: Trends and Policies*. Gütersloh: Bertelsmann Stiftung Gütersloh.

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

Engelmann J and Puntschuh M (2020) KI im Behördeneinsatz: Erfahrungen und Empfehlungen. Berlin: Kompetenzzentrum Öffentliche IT, Fraunhofer-Institut für offene Kommunikationssysteme FOKUS. Available at https://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-6350714.pdf (accessed 12 September 2023).

Fisher AJ, **Rudin C and Dominici F** (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research 20*(117), 1–81.

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

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

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

Holl J, **Kernbeiß G and Wagner-Pinter M** (2018) Das AMS-Arbeitsmarktchancen-modell. Available at https://www.ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_dokumentation.pdf (accessed 12 September 2023).

Howard PD and Dixon L (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), 287–307.

Hyndman R and Athanasopoulos G (2018) *Forecasting: Principles and Practice*. Melbourne: OTexts.

Jacobebbinghaus P and Seth S (2007) The German integrated employment biographies sample IEBS. *Schmollers Jahrbuch 127* (2), 335–342.

Kahneman D, **Sibony O and Sunstein CR** (2021) *Noise: A Flaw in Human Judgment*. Boston, MA: Little Brown.

Kern C, **Bach R**, **Mautner H and Kreuter F** (2021) Fairness in algorithmic profiling: A German case study. *arXiv:2108.04134* preprint. Available at https://arxiv.org/pdf/2108.04134 (accessed 12 September 2023).

Körtner J and Bonoli G (2021) Predictive algorithms in the delivery of public employment services. *SocArXiv* preprint. Available at https://osf.io/j7r8y/download (accessed 12 September 2023).

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

Loxha A and Morgandi M (2014) Profiling the unemployed: A review of OECD experiences and implications for emerging economies. Social Protection and labor discussion paper, SP 1424. Available at https://openknowledge.worldbank.org/server/api/core/bitstreams/4236e278-57c1-5d49-a726-aeecb7a74779/content (accessed 12 September 2023).

Marx C, **Calmon F and Ustun B** (2020) Predictive multiplicity in classification. In Singh HD (ed.), *Proceedings of the 37th International Conference on Machine Learning, Volume 119 of Proceedings of Machine Learning Research*, pp. 6765–6774. Virtual Conference: PMLR.

Matty S (2013) Predicting likelihood of long-term unemployment: The development of a UK jobseekers' classification instrument. Department for Work and Pensions Working paper No 116. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/210303/WP116.pdf (accessed 12 September 2023).

McDonald C, **Marston G and Buckley A** (2003) Risk technology in Australia: The role of the job seeker classification instrument in employment services. *Critical Social Policy 23*(4), 498–525.

Meehl PE (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.

Mehrabi N, **Morstatter F**, **Saxena N**, **Lerman K and Galstyan A** (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*(6), 1–35.

Niklas J, **Sztandar-Sztanderskal K and Szymielewicz K** (2015) *Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making*. Technical report, Fundacja Panoptykon, Warsaw.

O'Connell PJ, **McGuinness S and Kelly E** (2012) The transition from short-to long-term unemployment: A statistical profiling model for Ireland. *Economic and Social Review 43*(1), 135–164.

O'Connell PJ, **McGuinness S**, **Kelly E and Walsh J** (2009) National profiling of the unemployed in Ireland. Research Series 10, Economic and Social Research Institute, Dublin.

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

Potash E, **Ghani R**, **Walsh J**, **Jorgensen E**, **Lohff C**, **Prachand N and Mansour R** (2020) Validation of a machine learning model to predict childhood lead poisoning. *JAMA Network Open 3*(9), e2012734–e2012734.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: *R Foundation for Statistical Computing*.

Rudolph H and Müntnich M (2001) Profiling" zur Vermeidung von Langzeitarbeitslosigkeit. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 34*(4), 530–553.

Salganik MJ (2019) *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

StataCorp (2017) *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.

Tan P-N, **Steinbach M**, **Karpatne A and Kumar V** (2018) *Introduction to Data Mining*. New York: Pearson.

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

Viljanen M and Pahikkala T (2020) Predicting unemployment with machine learning based on registry data. In Dalpiaz F, Zdravkovic J and Loucopoulos P (eds), *Research Challenges in Information Science*. Cham: Springer International Publishing, pp. 352–368.

https://doi.org/10.1017/dap.2023.29 Published online by Cambridge University Press

Wijnhoven MA and Havinga H (2014) The work profiler: A digital instrument for selection and diagnosis of the unemployed. *Local Economy 29*(6–7), 740–749.

Yu MC and Kuncel NR (2020) Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions 6*(2), 1–10.

Cite this article: Bach RL, Kern C, Mautner H and Kreuter F (2023). The impact of modeling decisions in statistical profiling. *Data & Policy*, 5: e32. doi:10.1017/dap.2023.29