Jnl. Soc. Pol. (2021), 50, 2, 367–385 © The Author(s), 2020. Published by Cambridge University Press
doi:10.1017/S0047279420000203

Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off

SAM DESIERE* **AND LUDO STRUYVEN****

*HIVA, KU Leuven, 3000 Leuven, Belgium
email: sam.desiere@kuleuven.be
**HIVA, KU Leuven, 3000 Leuven, Belgium
email: ludo.struyven@kuleuven.be

Abstract

Artificial intelligence (AI) is increasingly popular in the public sector to improve the cost-efficiency of service delivery. One example is AI-based profiling models in public employment services (PES), which predict a jobseeker's probability of finding work and are used to segment jobseekers in groups. Profiling models hold the potential to improve identification of jobseekers at-risk of becoming long-term unemployed, but also induce discrimination. Using a recently developed AI-based profiling model of the Flemish PES, we assess to what extent AI-based profiling 'discriminates' against jobseekers of foreign origin compared to traditional rule-based profiling approaches. At a maximum level of accuracy, jobseekers of foreign origin who ultimately find a job are 2.6 times more likely to be misclassified as 'high-risk' jobseekers. We argue that it is critical that policymakers and caseworkers understand the inherent trade-offs of profiling models, and consider the limitations when integrating these models in daily operations. We develop a graphical tool to visualize the accuracy-equity trade-off in order to facilitate policy discussions.

Keywords: profiling; statistical discrimination; public employment services; artificial intelligence; VDAB

1. Introduction

Big data applications are not yet part and parcel of service delivery in the public sector (Klievink *et al.*, 2017). Interest is, however, growing rapidly and experiments with artificial intelligence (AI) based applications are being set up in the public sector (Kim *et al.*, 2014; Wirtz *et al.*, 2018; AlgorithmWatch, 2019). AI innovations are already transforming benefit allocation and service delivery in the field of social and labour market policies (Shorey and Howard, 2016). It is considered a promising avenue to make welfare administration smarter and to offer tailored services to clients (Veale and Brass, 2019). The ultimate objective is improving the efficiency and effectiveness of the welfare state.

One area of the public sector where AI-based profiling models are currently being implemented is the classification and segmentation of jobseekers in different groups (Desiere *et al.*, 2019). Screening, profiling and targeting jobseekers

is considered useful for assessing individual needs with the aim of supporting work resumption (Barnes *et al.*, 2015). Public employment services (PES) or other administrative bodies have set up procedures for identifying and allocating jobseekers to different categories, which then often determines the activation measures they can access or are entitled to. Most PES aim to devote their limited resources to 'vulnerable' jobseekers. This objective is established for either efficiency considerations (that is, the conviction that vulnerable jobseekers benefit most from intensive support) or for normative reasons (that is, the principle that vulnerable jobseekers deserve more support). Either way a central operational challenge is identifying the 'vulnerable' jobseekers.

The literature distinguishes three types of profiling: rule-based (or administrative) profiling, caseworker-based profiling and statistical profiling (Loxha and Morgandi, 2014; Barnes *et al.*, 2015). Rule-based profiling uses administrative eligibility criteria, such as jobseekers' age, educational level, and/or unemployment duration to classify jobseekers into client groups. Caseworker-based profiling relies on caseworkers' judgement to profile jobseekers, frequently supported by quantitative and/or qualitative tools to assess jobseekers' skills and needs. Statistical profiling uses a statistical model to predict the likelihood of work resumption. In this case, a vulnerable jobseeker is defined as a jobseeker with a low probability of resuming work.

Statistical profiling is not a radically new idea (OECD, 1998; Eberts *et al.*, 2002; Hasluck, 2008; Loxha and Morgandi, 2014). In the 1990s, the US introduced the Worker Profiling and Reemployment Services (WPRS). This tool assigns to jobseekers a score which predicts the likelihood of exhausting unemployment benefits (Black *et al.*, 2003; Pope and Sydnor, 2011). Australia segments jobseekers into different service streams based on the Job Seeker Classification Instrument, an instrument that assigns higher scores to disadvantaged groups (Brady, 2018). The Netherlands relies on a profiling tool called the 'Work Profiler' which consists of 20 questions related to the jobseeker's skills and attitudes that are predictive of work resumption (Wijnhoven and Havinga, 2014; Dusseldorp *et al.*, 2018).

An attractive aspect of statistical profiling is its (supposed) neutrality (Martin, 2018). Jobseekers with a same predicted probability of becoming long-term unemployed can be treated similarly by the PES, whereas rule-based profiling rules require normative choices (e.g. prioritising young over old jobseekers) and are often path-dependent (Henman, 2004) and caseworker-based profiling leads to different outcomes for similar jobseekers (Fletcher, 2011; De Wilde and Marchal, 2019). For these reasons, statistical profiling is often considered an objective approach to prioritise jobseekers and to allocate resources.

AI-based profiling models can be considered the next step in the development of statistical profiling models. Just like the existing regression-based profiling models, AI-based profiling models predict a jobseeker's likelihood

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

of resuming work within a certain period. But, in contrast to previous models, they use machine learning techniques to predict this outcome and often include many more explanatory variables. In addition, the existing regression-based models in the US, Australia and the Netherlands use a standardized questionnaire that consists of a limited number of questions related to work resumption and that needs to be completed by the jobseeker or caseworker, whereas AI-based profiling models can be trained on existing, administrative datasets. While regression-based models using a rich set of explanatory variables can in principle be as accurate as AI-based models, AI profiling models are more flexible than the traditional approaches and will in general more accurately predict the likelihood of becoming long-term unemployed. Nevertheless, our findings are not specific to AI-based profiling, but hold for all statistical profiling models that accurately predict long-term unemployment regardless of the underlying statistical approach.

AI-based statistical profiling models raise new questions about fairness and discrimination (Eubanks, 2018). Ideally, one would like to develop profiling models that are simultaneously accurate and fair. One infamous example of a profiling model that fails on both accounts is COMPAS[1], a statistical model used in the US to predict the probability of recidivism, which is not more accurate than predictions made by (random) citizens (Dressel and Farid, 2018) and is biased against Afro-Americans (Angwin *et al.*, 2016).

In general, however, theoretical work on labour market discrimination (Schwab, 1986) as well as the literature on machine learning (Friedler *et al.*, 2016; Kleinberg *et al.*, 2016) has shown that there is an inherent tension between model accuracy and discrimination. In this context, accuracy and discrimination are narrowly defined in technical terms. Accuracy refers to the predictive power of the model and is defined as the share of jobseekers correctly classified as either a low or high-risk jobseeker. Discrimination is defined as the proportion of jobseekers who belong to a particular group and find a job ex-post, but are misclassified as high-risk jobseekers, relative to this proportion among the dominant group (Žliobaitė, 2017). Importantly, this type of discrimination is independent from the existing inequalities in the historical data used to train the model, but purely stems from the mechanics of the statistical profiling model. Even in the absence of discrimination in the labour market, jobseekers belonging to a disadvantaged group such as migrants and older jobseekers will always be more likely to be misclassified as a high-risk jobseeker than the average jobseeker.

This paper confirms and illustrates the tension between model accuracy and model fairness. We obtained access to the output of an innovative, recently developed AI-based profiling model of the Flemish PES (VDAB), gradually replacing the existing rule-based approach. The output consists of the profiling score – the predicted probability of still being unemployed after six months from

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

a random forest model – of 288.756 jobseekers. We show that AI-based profiling improves the identification of jobseekers at-risk of becoming long-term unemployed compared to 'randomly' selecting jobseeker or compared to a standard rule-based approach. AI-based profiling is thus more accurate. At the same time, we confirm that AI-based profiling introduces 'statistical' discrimination: jobseekers who belong to a disadvantaged group (e.g. migrants, the disabled, older jobseekers) are more likely to be wrongly labelled as high-risk jobseekers. Improving accuracy comes at the cost of discrimination. Hence, there exits an accuracy-equity trade-off.

This tension between accuracy – or more generally – efficiency and equity in statistical profiling mirrors a decades-old debate in social policy (Okun, 1975; Le Grand, 1990). Social policy seeks to reconcile and overcome the tension between equality (in terms of a fair distribution of goods and services) and efficiency (in terms of an efficient allocation of resources). We raise Okun's big trade-off between equality and efficiency from a different angle than usual (the counterproductive effects of income redistribution on the economic process and the production of wealth). For PES, as one of the domains of social policy, the focus today is on prevention and early detection of at-risk groups (Ludwig-Mayerhofer *et al.*, 2014; Struyven and Van Parys, 2014; OECD, 2005). Policy makers want PES to be performant and (cost-)efficient, and continuously seek a balance between efficiency (restricting services to the most at-risk jobseekers) and equality (offering services to all jobseekers who need it). This is a major challenge because identifying at-risk groups is difficult and knowledge about the impact of policy interventions is limited.

This paper contributes to the growing literature on algorithmic fairness and discrimination (Calders and Verwer, 2010; Kleinberg *et al.*, 2016; Corbett-Davies *et al.*, 2017) and to the literature examining how ICT and AI-based applications are being integrated in the provision of public services (Busch *et al.*, 2018; Devlieghere *et al.*, 2019). The literature on (statistical) profiling emphasises that highly accurate profiling models can contribute to, but do not necessarily improve, the efficiency and effectiveness of service delivery. These models need to be integrated into daily management to be effective (Hasluck, 2008; Loxha and Morgandi, 2014; Desiere *et al.*, 2019). For instance, identifying 'vulnerable' jobseekers is only useful when policies are in place to support them. Similarly, profiling tools developed to support caseworkers that are not trusted, will not be used and will have no positive impact on service delivery (Lechner and Smith, 2007). Similar profiling models can be used for different purposes and can, hence, have a vastly different impact on service delivery (Marks, 2019). For instance, profiling models can be used to support caseworkers or to automate decisions. In the US, the profiling model automatically refers jobseekers to mandatory counselling or training programs (Black *et al.*, 2007). By contrast, jobseekers in the Netherlands are not obliged to

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

complete the Worker Profiler and the Flemish profiling model only determines the timing of the contact with the jobseeker. We will argue in the conclusion that a thorough understanding of the strengths and limitations of profiling models is essential when implementing these models in day-to-day management of PES.

The paper is structured as follows. We first introduce VDAB's profiling tool. We then present the data and the properties of the profiling model. After briefly explaining our methodology, we show that the model improves the identification of vulnerable jobseekers compared to rule-based profiling approaches, but is also unfair towards jobseekers of disadvantaged groups. This reflects the accuracy-equity trade-off. We develop a graphical tool to visualize this trade-off which can help to discuss this issue with stakeholders less familiar with statistical models. In the conclusion, we argue that it is critical to reflect on the integration of statistical profiling models into PES' decision-making processes to avoid (perceived) discrimination, while also harnessing its potential for improving the cost-efficiency of service delivery.

2. VDAB's profiling tool

The VDAB is the public employment service in Flanders (Belgium), responsible for mediation, referral and activation of jobseekers. The VDAB is one of the few PES which already developed an AI-based profiling model that estimates the probability of becoming long-term unemployed (Danneels and Viaene, 2015; Bouckaert *et al.*, 2017). The main objective is supporting caseworkers and line managers in deciding which jobseekers to prioritise (profiling). It is also a first step towards developing instruments that automatically recommend specific (online) training programs and counselling services to jobseekers that would increase their chances of finding a job (targeting) (Cockx *et al.*, 2019).

The VDAB experimented with different profiling models. The version from which we used the profiling scores in this study dates from January 2018. Like other profiling models, the instrument predicts the probability of resuming work within a given period. More specifically, VDAB's profiling tool assigns a jobseeker a profiling score 35 days after registration at the PES that gives the probability of gaining employment[2] within the next six months. A random forest model is trained on rich administrative datasets as well as on data collected at the time of self-registration at the VDAB. One advantage of this random forest model is its flexibility. It can easily be retrained as more recent data or new explanatory variables become available. Moreover, as the economy changes continuously, variables that had high predictive power in the past are not necessarily good predictors of work resumption today. By regularly retraining the model, the VDAB accommodates these evolutions and ensures that the profiling model remains accurate.

The VDAB collects standard information such as the jobseekers' age, educational level, nationality/origin and previous (un)employment spells, but also

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

records self-reported job preferences and participation in training programs. In addition, the PES tracks the jobseekers' activity and behaviour on their website. It monitors, for instance, how often jobseekers clicked on job vacancies or updated their online CV. For now, the model relies solely on administrative data and on data entered by jobseekers for other purposes. No new information is collected on soft skills, attitudes and job search strategies with the purpose of improving model accuracy.

In a more recent version of the profiling model, the number of explanatory variables has been reduced in order to simplify the model and to comply with privacy regulations and anti-discrimination law. Variables with a low explanatory power were removed from the model. Sensitive information such as origin/nationality and disability status was also discarded.[3] Omitting sensitive variables does not mean that discrimination disappears, because the model incorporates this information via other variables such as language skills. The pervasive nature of social identifiers means that such sensitive information is embedded in big datasets, even if it is not intentionally collected or is deleted (Williams *et al.*, 2018). Despite using fewer explanatory variables, a more recent version has the same level of accuracy as the January 2018 version from which we analysed the scores.[4]

The profiling model, rolled out as part of VDAB's new 'contact strategy', is currently used only to determine who should be contacted first. In the past, the ranking of whom to contact first was determined by rules and further interpreted by caseworkers. The contact strategy aims to reach all jobseekers within six weeks after registration. Based on their profiling score, jobseekers are divided in four groups from very unlikely to very likely to quickly resume work (with as thresholds a profiling score lower than 35%, between 35% and 50%, between 50% and 65%, and higher than 65%). Jobseekers most at-risk of becoming long-term unemployed are contacted first. Based on a phone interview, the caseworker then decides whether the jobseeker is self-reliant (and does not need close follow-up) or is to be referred to more intensive support. At the time of writing (December 2019), caseworkers do not have access to the profiling score. They are provided with automatically generated lists of jobseekers who need to be contacted by phone. This list gives the priority to jobseekers with a low profiling score. Hence, the profiling model only ensures that vulnerable jobseekers are contacted first and has no effect on the caseworkers' referral decisions. The current approach thus combines statistical and caseworker-based profiling.

3. Data and descriptive statistics

Our dataset consists of the population of jobseekers who registered at the VDAB in the course of 2016 (288.756 unique jobseekers). Each jobseeker is assigned a profiling score by VDAB's profiling model. As noted earlier, we use the January 2018 version of VDAB's profiling tool. Besides the profiling scores generated by

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

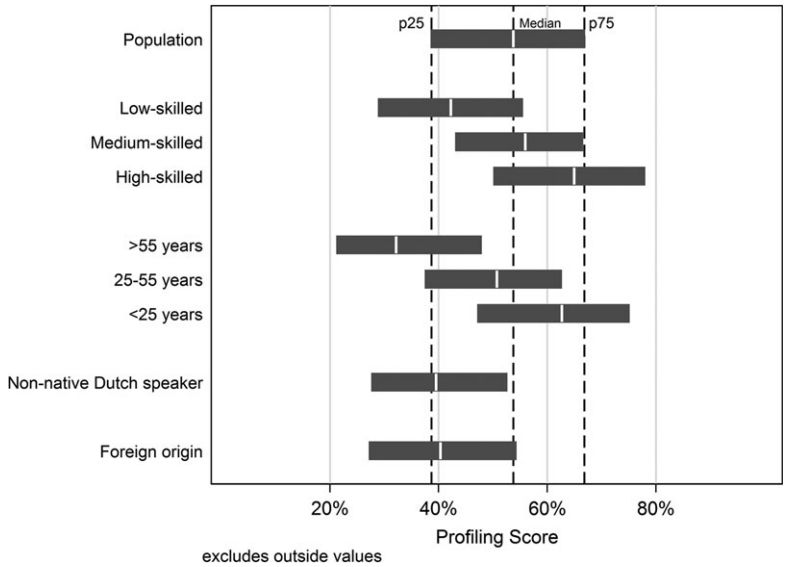

FIGURE 1. The distribution of the profiling score for different groups

VDAB's profiling model, we have access to data on jobseekers' characteristics such as previous work history, educational level, language skills, sex, age and nationality. The labour market position of each jobseekers is tracked from registration until January 2018. In each month, we know whether the jobseeker is employed or unemployed. We determine model accuracy by comparing the labour market outcome predicted by the model with the jobseekers' real labour market position seven months after becoming unemployed.

One particularly important variable for this study is 'migrant background'. VDAB defines jobseekers with a migrant background as jobseekers whose current or previous nationality is non-European. Twenty-three percent of the jobseekers have a migrant background, of which slightly more than half have the Belgian nationality. With a slight misuse of terminology, we will refer to jobseekers with a migrant background as 'migrants' or 'jobseekers of foreign origin[5]', while referring to the other jobseekers as 'jobseekers of Belgian origin[6]'.

The population is diverse, and includes jobseekers who will easily transition to a job as well as jobseekers who are rather unlikely to find a job in the near future. This is reflected in the profiling scores. Figure 1 shows the distribution of the profiling score with boxplots for the entire population as well as for specific groups. The median profiling score is 54%. Hence, slightly more than half of the 288.756 jobseekers are expected to resume work within six months. A quarter of the population has a profiling score lower than 39%, whereas a quarter has a profiling score higher than 67%.

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

Boxplots are also drawn by jobseekers' educational level, age, knowledge of Dutch and origin. There is quite some overlap in the distributions across groups. The average profiling score of low-skilled jobseekers is, for instance, lower than the average score of medium-skilled jobseekers. Nevertheless, roughly half of the low-skilled jobseekers have a higher score than a quarter of the medium-skilled jobseekers. In other words, many low-skilled jobseekers will more easily resume work than some of the more vulnerable medium-skilled jobseekers. This demonstrates that the profiling score gives a more nuanced picture of a jobseeker than a single criterion like educational level. The profiling score succeeds in capturing and summarizing in a single statistic a host of elements that determine the likelihood of resuming work.

4. Methodology

The objective of this paper is to illustrate how AI-based profiling models can improve the accuracy of identifying vulnerable jobseekers, but at the cost of 'discriminating' against individuals belonging to disadvantaged groups. We illustrate this in two steps.

In a first step, we compare AI-based profiling to two selection-rules that are frequently used to prioritise jobseekers, namely (1) randomly selecting jobseekers and (2) prioritising low-skilled jobseekers. The first selection-rule is inspired by PES that do not use selection-rules, but have insufficient resources to support all jobseekers. Variants of the second selection-rule are frequently used by PES. As 33.8% of the jobseekers in the population are low-skilled, the second selection rule will by construction label 33.8% of the jobseekers as having a high and 66.2% of the jobseekers as having a low risk of becoming long-term unemployed. In order to make this selection rule comparable to the two other approaches, we will set the parameters of these approaches so that exactly the same proportion of jobseekers is labelled as high-risk jobseekers. This implies that the first selection rule randomly labels 33.8% of the jobseekers as having a high-risk, whereas the AI-based approach labels all jobseekers with a profiling score lower than 45% as high-risk jobseekers.

The accuracy and fairness of each approach will be calculated. Accuracy is defined as the share of jobseekers who are correctly identified as high or low-risk jobseekers. Several measures of fairness have been proposed (Romei and Ruggieri, 2014; Žliobaitė, 2017). We call a model 'unfair' if jobseekers of disadvantaged groups who find a job ex-post are more likely to be misclassified as high-risk jobseekers. In more technical terms, a model is fair if the false positive rate is equal across groups.[7] Discrimination is then measured as a ratio. For instance, if 40% of the migrants who find a job ex-post are misclassified as high-risk compared to 20% of the Belgian jobseekers, then the ratio-based measure of discrimination equals 2 (40%/20%). Žliobaitė (2017) refers to this

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

TABLE 1. The accuracy and fairness of AI-based profiling versus rule-based profiling

	Selection-rule 1 (randomly labelling jobseekers as high-risk)	**Selection-rule 2** (labelling all low-skilled jobseekers as high-risk)	**AI-based profiling** (labelling all jobseekers with a profiling score lower that 45% as high-risk)
Share of jobseekers labelled as high-risk jobseeker	33.8%	33.8%	33.8%
Accuracy (share of jobseekers correctly identified as low or high-risk)			
All jobseekers	50.2%	58.0%	66.0%
Belgian origin	51.5%	59.4%	65.4%
Foreign origin	45.8%	51.5%	66.0%
Discrimination (found a job ex-post, misclassified as high-risk ex-ante)			
Belgian origin	34.0%	23.2%	14.8%
Foreign origin	34.1%	42.9%	38.9%
Discrimination (ratio foreign origin/Belgian origin)	1.00	1.85	2.63

definition of discrimination as the 'impact ratio'. An alternative definition examines the difference between both shares (e.g. 40%-20%) rather than the ratios. Throughout the paper, we will use ratios. The implications of using difference-based measures are explored in Appendix 1.

In a second step, we play with the threshold used in the AI-based profiling model to distinguish low from high-risk jobseekers. We show how the share of jobseekers classified as high-risk jobseekers as well as the accuracy and fairness of the model is related to the choice of the threshold. By combining accuracy and fairness in a single graph, we derive the accuracy-equity trade-off.

In both steps, we compare the outcome of VDAB's profiling model for Belgian jobseekers to the outcome for jobseekers of foreign origin. The results also hold, however, for other disadvantaged groups in the labour market.

5. Results

5.1. AI-based profiling vis-à-vis rule-based profiling

We compare three profiling approaches: (1) randomly classifying jobseekers as high-risk; (2) classifying all low-skilled jobseekers as high-risk; and (3) classifying jobseekers with a profiling score lower than 45% as high-risk.[8]

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

These three approaches classify 33.8% of the jobseekers as high-risk jobseekers. Table 1 compares the share of jobseekers correctly identified as high or low-risk jobseekers (model accuracy) as well as the share of jobseekers who find a job ex-post but are misclassified as high-risk jobseekers ex-ante (discrimination). These statistics are presented by the jobseekers' origin.

The first selection-rule randomly labels 33.8% of the jobseekers as high-risk. This implies that the proportion of jobseekers of foreign origin classified as high-risk equals their proportion in the population. The accuracy of the random selection-rule is lower for jobseekers of foreign origin than for Belgian jobseekers: 46% of the jobseekers of foreign origin are correctly classified compared to 52% of the jobseekers of Belgian origin. In other words, too few jobseekers of foreign origin are classified as high-risk jobseekers. By definition, randomly classifying jobseekers as high-risk does not induce discrimination. One out of three jobseekers of foreign as well as Belgian origin who ultimately find a job are misclassified ex-ante.

The second selection-rule targets low-skilled jobseekers. All low-skilled jobseekers (33.8% of the population) are labelled as high-risk jobseekers, whereas the medium and high-skilled jobseekers are labelled as low-risk jobseekers.[9] Compared to randomly inviting jobseekers, this selection-rule improves the accuracy for both jobseekers of Belgian and foreign origin. The accuracy is, however, still lower for jobseekers of foreign origin than for jobseekers of Belgian origin (59% versus 52%). The higher accuracy induces statistical discrimination: jobseekers of foreign origin are 1.9 times more likely to be wrongly labelled as high-risk jobseekers than jobseekers of Belgian origin (43% versus 23%).

The third approach relies on AI-based profiling and classifies all jobseekers with a profiling score below 45% as high-risk jobseekers. Using AI-based profiling improves the accuracy of identifying at-risk jobseekers compared to randomly selecting jobseeker or compared to prioritising the low-skilled. Two out of three jobseekers are correctly classified. Moreover, model accuracy is the same for jobseekers of foreign and Belgian origin. However, the higher accuracy results in more discrimination: jobseekers of foreign origin are 2.6 times more likely to be misclassified as high-risk jobseekers compared to jobseekers of Belgian origin (39% versus 15%).

The three different profiling approaches neatly illustrate the accuracy-equity trade-off. While statistical profiling as well as selection-rules help to identify jobseekers at-risk of becoming long-term unemployed, they inevitably entail statistical discrimination. Individuals who quickly resume work, but belong to a group that has *on average* a low probability of finding work, are more likely to be misclassified as high-risk jobseekers. Statistical profiling is the most accurate approach, but also misclassifies a higher share of jobseekers of foreign origin than the two selection-rules.

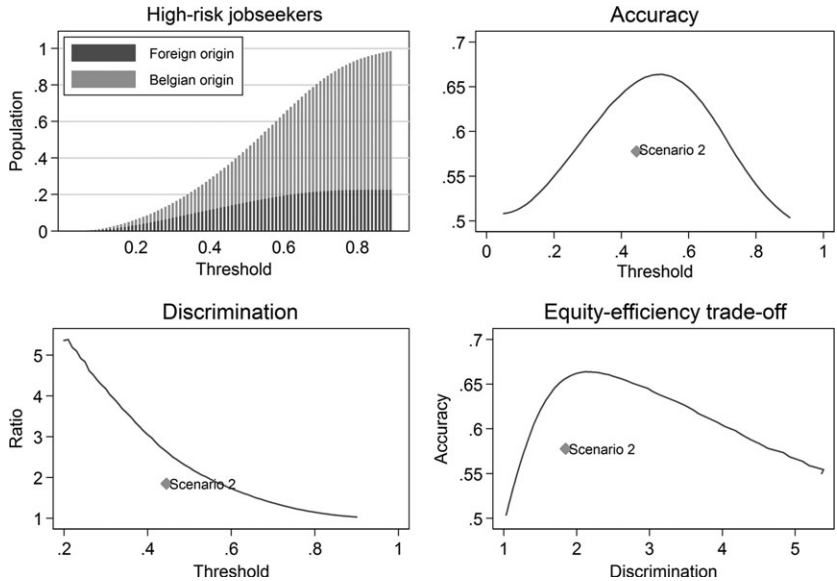

FIGURE 2. The trade-offs of a AI profiling model
Note: Scenario 2 classifies all low-skilled jobseekers as 'high-risk'

5.2. The accuracy-equity trade-off

The previous section illustrated the accuracy-equity trade-off of an AI profiling model when jobseekers with a profiling score lower than 45% are labelled as high-risk jobseekers. Depending on its resources and objectives, the PES could also set another threshold to distinguish between low and high-risk jobseekers. This threshold determines (1) the share of jobseekers labelled as high-risk, (2) the accuracy of the profiling model, (3) its fairness and (4) the accuracy-equity trade-off. This is illustrated with four figures (Figure 2).

The first figure (top, left) shows the relation between the threshold and the share of jobseekers identified as high-risk jobseekers. The higher the threshold, the more jobseekers labelled as high-risk and, thus, the more jobseekers prioritised by the PES. For instance, the proportion of jobseekers classified as high-risk is 20% if the threshold is set at 35% and increases to 80% if the threshold is set at 70%. The figure also shows the proportion of jobseekers of Belgian/foreign origin among the high-risk jobseekers. This proportion decreases with the threshold. If the threshold is low, most of the high-risk jobseekers are of foreign origin. For instance, if the threshold is 30%, 41,140 jobseekers (14% of the population) are classified as high-risk, of which 19,644 or nearly half are of foreign origin.

The second figure (top, right) shows the inverse U-shaped relation between the threshold and model accuracy. The inverse U-shape illustrates the well-

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

known trade-off between prioritising too few jobseekers (implying that many jobseekers who need support will not be supported) and too many jobseekers (implying that many jobseekers are being supported who would find a job anyway, i.e. the so-called deadweight effects). Accuracy reaches a maximum of 66% when the threshold equals 52%.

The third figure (bottom, left) shows the negative relation between the threshold and discrimination. Discrimination is defined as the share of jobseekers of foreign origin who find a job ex-post, but are wrongly labelled as high-risk jobseekers ex-ante, relative to this share for Belgian jobseekers. For instance, if the threshold is 30%, 14% of the jobseekers of foreign origin that find work ex-post are misclassified as high-risk ex-ante, compared to 3.4% of the jobseekers of Belgian origin. In this case, the level of discrimination is 4.2 (=14%/3.4%). Discrimination decreases as the threshold increases because the share of jobseekers of foreign origin among the high-risk jobseekers also decreases with the threshold.

Finally, the fourth figure (bottom, right) combines the second and third figure. It illustrates the accuracy-equity trade-off. This is again an inverse U-shaped relation. Higher accuracy increases discrimination up to a point when accuracy starts decreasing, while discrimination continues to increase. The intuition behind the results is that for high values of the threshold many jobseekers are considered high-risk and both the accuracy and the level of discrimination is low. For low values of the threshold, few jobseekers are considered high-risk, but those that are considered high-risk are predominantly of foreign origin. This explains the low accuracy and high level of discrimination for low values of the threshold.

The figures also highlight the accuracy and fairness of the selection-rule that labels all low-skilled jobseekers as high-risk, one of the rules discussed in the previous section (scenario 2 in Figure 2). This rule identifies 33.8% of the jobseekers as high-risk. This corresponds to a threshold of 45% if one wants to classify the same proportion of jobseekers as high-risk with the AI-based profiling model. The figure shows that this selection-rule is less accurate, but more fair than the AI profiling model. Interestingly, the accuracy-equity trade-off shows that the selection-rule is not optimal. Switching to AI-based profiling could increase accuracy without increasing discrimination (by reducing the threshold), or could reduce discrimination while keeping the same level of accuracy (by increasing the threshold).

As noted earlier, discrimination is defined here as the ratio of the share of jobseekers of foreign origin who find a job ex-post but are misclassified as high-risk ex-ante relative to this share for Belgian jobseekers. Rather than examining ratios, one could also define discrimination as the difference between both shares. Using difference-based measures of discrimination changes the relationship between the threshold and the level of discrimination as well as the shape of

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

the accuracy-equity curve. It imposes symmetry on both curves. In contrast to the ratio-based measure, discrimination measured in differences is not only low for high, but also for low values of the threshold. For low values, few jobseekers are misclassified as high-risk. Hence, the difference in the share of misclassified jobseekers of foreign origin versus Belgian jobseekers is small and, so is discrimination. The shape of the accuracy-equity curve changes to an ellipse: accuracy and discrimination are low for low as well as high values of the threshold; and a maximum level of accuracy and discrimination is reached at an intermediate level of the threshold. Appendix 2 presents these results in more detail. Importantly, regardless of the measure of discrimination, we do observe an accuracy-equity trade-off.

6. Conclusion

Facing a continuous flow of jobseekers, PES have to decide how to allocate resources to different groups of jobseekers. Selection-rules are used to distinguish between jobseekers who will receive (intensive) support and jobseekers who are referred to the digital services. These selection-rules aim to identify the most vulnerable jobseekers. As such, PES improve the cost-efficiency of service delivery and avoid wasting resources on jobseekers who will easily transition to a job. Traditionally, the selection-rules were relatively simple and identified vulnerable jobseekers based on observable characteristics such as educational level. More recently, PES have developed sophisticated AI-based profiling models to predict a jobseeker's likelihood of work resumption. The rationale is that 'big data models' are more accurate in identifying jobseekers at-risk of becoming long-term unemployed than simple selection-rules which only consider a few characteristics of jobseekers. The profiling score is then used to segment jobseekers in groups in order to prioritise vulnerable jobseekers and/or (automatically) refer them to different service streams.

The central question of this paper is whether AI-based profiling improves the early identification of jobseekers at-risk of becoming long-term unemployed while not increasing discrimination compared to more traditional ways of classifying jobseekers. An inherent feature of simple selection-rules as well as AI-based profiling models is (statistical) discrimination. Jobseekers who belong to a disadvantaged group (such as jobseekers of foreign origin) and find a job ex-post are more likely to be misclassified as a high-risk jobseeker. This can be considered discrimination as group characteristics determine the outcome of an individual belonging to this group. More performant profiling models misclassify fewer jobseekers and suffer therefore less from statistical discrimination. However, even the best profiling models are not perfect. An international comparison suggests that today's best performing profiling models correctly classify roughly 70% of the jobseekers (Desiere *et al.*, 2019).

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

Policymakers face an accuracy-equity trade-off. This trade-off is inherent to any form of profiling. Both rule-based and statistical profiling help to identify jobseekers at-risk of becoming long-term unemployed, but suffer from discrimination. In Flanders, for instance, the accuracy of labelling all low-skilled jobseekers as high-risk is 58%, implying that 58% of the jobseekers are correctly identified as low or high-risk jobseekers. However, this selection-rule also implies that 43% of the jobseekers of foreign origin that ultimately find a job are misclassified as high-risk jobseekers compared to 23% of the jobseekers of Belgian origin. Using an AI-based profiling model further improves accuracy, but also decreases the 'fairness' of the model. The maximum level of accuracy is 66%. At this level, jobseekers of foreign origin are 2.6 times more likely to be misclassified than jobseekers of Belgian origin (39% versus 15%).

Excluding sensitive variables (such as origin) from AI-based profiling models does not necessarily reduce discrimination because other variables are correlated with the sensitive variables (Pope and Sydnor, 2011; Žliobaitė and Custers, 2016). In our example, being of foreign origin is correlated with being a native Dutch speaker. Nearly eight out of ten jobseekers of foreign origin report that Dutch is not their native language compared to 15% of the Belgian jobseekers. We therefore do not expect that VDAB's current profiling model – which no longer includes nationality/origin, but still includes knowledge of Dutch – discriminates substantially less than the January 2018 version, which was used in this study and does include nationality/origin as an explanatory variable. Even in the unlikely case that sensitive variables are not correlated with other variables, removing them will reduce discrimination, but will also reduce model accuracy. The accuracy-equity trade-off is again unavoidable.

Does the accuracy-equity trade-off raise real concerns about fairness, or did we identify a purely theoretical problem without practical implications? In our view, the answer depends on (1) how the profiling model is operationalized and (2) whether the services that are being offered are considered helpful. In countries like the US, Australia and, to a lesser extent, the Netherlands, the outcome of the profiling model/selection-rule automatically determines the type of (often compulsory) services. In Flanders, the profiling model is not used to automate decision-making, but is used to prioritise jobseekers and support caseworkers. A new contact strategy, rolled out in November 2018, aims to reach all jobseekers within six weeks. Jobseekers most at-risk of becoming long-term unemployed, as predicted by the profiling model, are contacted first. Based on a phone interview, the caseworker decides whether the jobseeker is referred to more intensive support and follow-up or is 'self-reliant'. For now, the aim is to reach all jobseekers, starting with jobseekers with the lowest profiling score. But one can imagine that, when the caseload increases due to an economic downturn, only jobseekers with low profiling scores will be reached within six weeks. Arguably, the issue of 'fairness' is less of a concern when it only determines

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

the timing of contact with the PES (as is the case in Flanders) than when it is used to automate decisions (see Marks (2019) who distinguishes between 'soft-touch' and 'firm-hand' interventions).

The perceived usefulness of the PES' support is a second element that determines whether 'discrimination' matters. Misclassifications matter less if the services are considered helpful. By contrast, if the (compulsory) services are considered burdensome or unhelpful, the 'misclassified' jobseekers might resent them. Most PES, including the VDAB, have a dual mandate: supporting jobseekers and monitoring their job search. Discrimination is more of an issue if the PES focuses predominantly on monitoring job search with possible loss of benefits as a result. In this case, jobseekers of foreign origin – who are more likely to be misclassified – might feel targeted and discriminated.

Depending on the value of the services offered, discrimination is 'negative' or 'positive'. If services are only offered to high-risk jobseekers, then jobseekers of foreign origin are more likely to receive support. We can speak of 'positive' discrimination if the services are valuable and of 'negative' discrimination if the support is perceived burdensome. So far, we emphasized the implications for jobseekers of foreign origin. The other side of the coin is that jobseekers of Belgian origin who remain unemployed ex-post are more likely to be misclassified as low-risk ex-ante. Some jobseekers of Belgian origin will therefore not be offered support, although they need it. If support is very generous, Belgian jobseekers might feel excluded and discriminated.

AI-based profiling models can be updated continuously. A flexible AI profiling model allows to quickly adjusting the profiling score to changing trends in the labour market. This strength of AI profiling is at the same time a weakness. Imagine that specific, vulnerable groups receive effective services that substantially increase their chances of work resumption. These jobseekers will not be identified as vulnerable because – due to effective support – they quickly resume work. As a result, the profiling model does not classify these jobseekers as high-risk, therefore excluding them from support in the future. Hence, there exists a risk that vulnerable jobseekers who are being supported effectively today will be misclassified as low-risk jobseekers tomorrow. In practice, this risk seems limited as most active labour market policies have at best a relatively small, positive effect on work resumption (Card *et al.*, 2017). VDAB could nevertheless consider to experiment with statistical profiling models that take support received by the PES into account and then predict the probability of resuming work in the absence of this support. This avoids that vulnerable jobseekers are misclassified as low-risk because they have been supported effectively in the past.

This paper did not explore whether and how profiling scores change caseworkers' behaviour. In the case of the VDAB, caseworkers establishing a

first contact do not see the exact profiling score on their dashboard. This may change in the future. Caseworkers will in any case remain entitled to overrule the profiling score. Whether the profiling score will influence their decision and whether caseworkers succeed in identifying individuals misclassified by the model is an interesting avenue for further research.

Profiling models offer tremendous opportunities to improve accuracy, but one should be aware of their limitations. Explaining the strengths and limitations of this new evolution to policymakers and caseworkers that are not familiar with statistical or AI models can be challenging. The four figures presented in section 5 offer a tool to visualize the trade-offs, to highlight strengths and weaknesses, to compare with the selection-rules that are currently being used and to guide the critical choice of the threshold. Given a threshold, the figures can be used to discuss key policy questions such (1) Is the level of 'discrimination' acceptable? (2) Can we redesign current selection-rules so that model accuracy increases while the level of discrimination remains unaltered? and (3) How can services be designed so that jobseekers of foreign origin do not feel 'targeted'? If the discrimination inherent in the model is deemed unacceptable, one could consider alternative strategies such as randomly selecting jobseekers, letting the jobseekers choose whether they want support, or invite all jobseekers but only once they have been unemployed for a few months.

Acknowledgement

The authors wish to thank the reviewers of the Journal of Social Policy for their thorough and pertinent feedback. The insightful comments from the participants of the Belgian Day of Labour Economists (2019), the participants of the Espanet Conference in Stockholm (2019) and the staff at the Flemish Department for Work were also greatly appreciated. We thank Nele Havermans for carefully reading and commenting on a previous draft. We are grateful to the VDAB, and in particular to Stijn Van De Velde, for making the data and the profiling score available.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S0047279420000203

Notes

1 Correctional Offender Management Profiling for Alternative Sanctions.

2 We follow VDAB's convention to define 'employment'. The conventions are complex because 33 different administrative codes are used to register labour market positions. The two most common codes defining employment are (1) jobseekers who accepted a regular job (with an open-ended or fixed contract) and (2) jobseekers who engaged at least 10 days in the last 28 days in temporary employment.

3 Excluding sensitive variables is often mandated by law. For instance, article 9 of the General Data Protection Regulation (GDRP), stipulates that processing of special categories of data (such as race, origin, political opinions) is prohibited, except under specific circumstances (see Goodman and Flaxman (2016) for an overview of GDPR's implications for algorithmic decision making).

4 According to the VDAB, the AUC, a common measure of model accuracy, of the February 2019 and the January 2018 version is, respectively, 70.2% and 74.3%. Using a threshold of 50%, the models correctly classify respectively 70.2% and 67.1% of the jobseekers. These parameters will further evolve as the model continuous to be improved.

5 Note that this definition departs from the standard definition of origin used in Belgium. The standard definition looks at someone's current and previous nationality as well as the nationality of the parents.

6 89% of the 'Belgian' jobseekers have the Belgian nationality; 11% of the 'Belgian' jobseekers have a European nationality.

7 This definition of fairness is known as '*predictive equality*'. Other popular definitions include '*statistical parity*', meaning that an equal share is identified as high-risk in both groups, and '*conditional statistical parity*', meaning that after controlling for a legitimate set of characteristics (e.g. education), an equal share is identified as high-risk in both groups (Corbett-Davies *et al.*, 2017).

8 The first selection-rule implies '*statistical parity*'; the second selection-rule implies '*statistical parity conditional on educational level*'.

9 This selection-rule satisfies statistical parity conditional on educational level. Given the educational level, the share of jobseekers labelled as high-risk is the same among Belgian jobseekers and migrants (i.e. 100% if low-skilled; 0% if medium or high-skilled).

References

AlgorithmWatch (2019). *Automating society: Taking stock of a automated decision making in the EU*. Berlin, AlgorithmWatch in cooperation with Bertelsmann Stiftung.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016), *Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica.

Barnes, S.-A., Wright, S., Irving, P. and Deganis, I. (2015), *Identification of latest trends and current developments in methods to profile jobseekers in European public employment services: final report*.

Black, D.A., Galdo, J. and Smith, J.A. (2007), 'Evaluating the worker profiling and reemployment services system using a regression discontinuity approach.' *American Economic Review*, 97, 2, 104–107.

Black, D.A., Smith, J.A., Plesca, M. and Shannon, S. (2003), *Profiling UI claimants to allocate reemployment services: Evidence and Recommendations for States*. Final Report to United States Department of Labor.

Bouckaert, D., Reussens, M., Larnout, D., Heene, L., Schoonbrood, S., Claes, R., Klewais, E. and Humbeeck, G.V. (2017), 'VDAB op koers voor een datagedreven aanpak met big data.' *Over Werk*, 2, 64–69.

Brady, M. (2018), 'Targeting single mothers? Dynamics of contracting Australian employment services and activation policies at the street level.' *Journal of Social Policy*, 47(4), 827–845.

Busch, P.A., Henriksen, H.Z. and Sæbø, Ø. (2018), 'Opportunities and challenges of digitized discretionary practices: a public service worker perspective.' *Government Information Quarterly*, 35, 4, 547–556.

Calders, T. and Verwer, S. (2010), 'Three naive Bayes approaches for discrimination-free classification.' *Data Mining and Knowledge Discovery*, 21, 2, 277–292.

https://doi.org/10.1017/S0047279420000203 Published online by Cambridge University Press

Card, D., Kluve, J. and Weber, A. (2017), 'What works? A meta analysis of recent active labor market program evaluations.' *Journal of the European Economic Association*, 16, 3, 894–931.

Cockx, B., Lechner, M., Bollens, J. (2019), '*Priority to unemployed immigrants? A causal machine Learning evaluation of training in Belgium*.' IZA Discussion Paper No 12875.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017), *Algorithmic decision making and the cost of fairness*. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.

Danneels, L. and Viaene, S. (2015), 'Simple rules strategy to transform government: An ADR approach.' *Government Information Quarterly*, 32, 4, 516–525.

Desiere, S., Langenbucher, K. and Struyven, L. (2019), *Statistical profiling in public employment services*. OECD Working Paper.

Devlieghere, J., Bradt, L. and Roose, R. (2019), 'Electronic information systems as means for accountability: why there is no such thing as objectivity.' *European Journal of Social Work*, 1–12.

De Wilde, M. and Marchal, S. (2019), 'Weighing up work willingness in social assistance: a balancing act on multiple levels.' *European Sociological Review*, 35(5), 718–737.

Dressel, J. and Farid, H. (2018), 'The accuracy, fairness, and limits of predicting recidivism.' *Science Advances*, 4, 1.

Dusseldorp, E., Hofstetter, H. and Sonke, C. (2018), *'Landelijke doorontwikkeling van de UWV Werkverkenner: eindrapportage.'*

Eberts, R.W., O'Leary, C.J. and Wandner, S.A. (2002), *Targeting employment services*, WE Upjohn Institute.

Eubanks, V. (2018), *Automating inequality: How high-tech tools profile, police, and punish the poor*, St. Martin's Press.

Fletcher, D. R. (2011), 'Welfare reform, Jobcentre Plus and the street-level bureaucracy: towards inconsistent and discriminatory welfare for severely disadvantaged groups?' *Social policy and Society*, 10(4), 445–458.

Friedler, S.A., Scheidegger, C. and Venkatasubramanian, S. (2016), 'On the (im) possibility of fairness.' *arXiv preprint arXiv:1609.07236*.

Goodman, B. and Flaxman, S. (2016), 'European Union regulations on algorithmic decision-making and a "right to explanation".' *arXiv preprint arXiv:1606.08813*.

Hasluck, C. (2008), *The use of statistical profiling for targeting employment services: The international experience*. New European Approaches to Long-Term Unemployment: What role for public employment services and what market for private stakeholders.

Henman, P. (2004). 'Targeted! Population segmentation, electronic surveillance and governing the unemployed in Australia.' *International Sociology*, 19(2), 173–191.

Kim, G.-H., Trimi, S. and Chung, J.-H. (2014), 'Big-data applications in the government sector.' *Communications of the ACM*, 57, 3, 78–85.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016), 'Inherent trade-offs in the fair determination of risk scores.' *arXiv preprint arXiv:1609.05807*.

Klievink, B., Romijn, B.-J., Cunningham, S. and de Bruijn, H. (2017), 'Big data in the public sector: Uncertainties and readiness.' *Information Systems Frontiers*, 19, 2, 267–283.

Lechner, M. and Smith, J. (2007), 'What is the value added by caseworkers?' *Labour Economics*, 14(2), 135–151.

Le Grand, J. (1990), 'Equity versus efficiency: the elusive trade-off.' *Ethics*, 100(3), 554–568.

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. World Bank Group, Washington, DC.

Ludwig-Mayerhofer, W., Behrend, O. and Sondermann, A. (2014). 'Activation, public employment services and their clients: the role of social class in a continental welfare state'. *Social Policy & Administration*, 48, 5, 594–612.

OECD (1998), *Early identification of jobseekers at risk of long-term unemployment: the role of profiling*, OECD.

OECD (2005), *OECD Employment Outlook*, OECD.
Okun, A. (1975). '*Equality and Efficiency: The Big Tradeoff*.' Washington: Brookings Institution Press.
Marks, M., (2019), 'Artificial Intelligence based suicide prediction.' *Yale Journal of Health Policy, Law, and Ethics, Forthcoming*.
Martin, K. (2018), 'Ethical implications and accountability of algorithms.' *Journal of Business Ethics*, 1–16.
Pope, D.G. and Sydnor, J.R. (2011), 'Implementing anti-discrimination policies in statistical profiling models.' *American Economic Journal: Economic Policy*, 3, 3, 206–231.
Romei, A. and Ruggieri, S. (2014), 'A multidisciplinary survey on discrimination analysis.' *The Knowledge Engineering Review*, 29, 5, 582–638.
Schwab, S. (1986), 'Is statistical discrimination efficient?' *The American Economic Review*, 76, 1, 228–234.
Shorey, S. and Howard, P. (2016). 'Automation, big data and politics: A research review.' *International Journal of Communication 10*.
Struyven, L. and Van Parys, L. (2014). 'Revisiting the Pillars of the PES and Common Challenges.' In F. Leroy and L. Struyven (eds.), Building Bridges. Shaping the Future of Public Employment Services Towards 2020, 49–69. Brugge: die Keure.
Veale, M. and Brass, I. (2019), *Administration by algorithm? Public management meets public sector machine learning*. Public Management Meets Public Sector Machine Learning.
Wijnhoven, M. and Havinga, H. (2014), 'The Work Profiler: A digital instrument for selection and diagnosis of the unemployed.' *Local Economy*, 29, 6-7, 740–749.
Williams, B.A., Brooks, C.F. and Shmargad, Y. (2018), 'How algorithms discriminate based on data they lack: challenges, solutions, and policy implications.' *Journal of Information Policy*, 8, 78–115.
Wirtz, B.W., Weyerer, J.C. and Geyer, C. (2018), 'Artificial Intelligence and the public sector—applications and challenges.' *International Journal of Public Administration*, 1–20.
Žliobaitė, I. and Custers, B. (2016), 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models.' *Artificial Intelligence and Law*, 24, 2, 183–201.
Žliobaitė, I. (2017), 'Measuring discrimination in algorithmic decision making.' *Data Mining and Knowledge Discovery*, 31, 4, 1060–1089.