Machine Learning and Subjective Assessments by Unemployed Workers and Their Caseworkers

Gerard J. van den Berg* **Max Kunaschk†**
Julia Lang‡ **Gesine Stephan§** **Arne Uhlendorff¶**

June 29, 2022. Preliminary and incomplete.
Please do not cite or circulate.

Abstract

The assessment of whether unemployed individuals will become long-term unemployed is important for the unemployment insurance budget as well as for fine-tuning active labor market policy packages. We analyze unique data on three sources of information on the risk of long-term unemployment (LTU). First, the individuals in the inflow into unemployment in five regions in Germany in 2012-2013 are asked for their perceived probability of LTU. Secondly, their caseworkers are asked whether they expect the individuals to become LTU. Thirdly, we use random-forest machine learning methods to predict individual LTU. The latter algorithms are trained on the full inflow one year earlier in these regions. We compare the predictive performance of each of these measures and we consider whether various combinations the measures improve this performance. The results suggest that caseworker assessments contain information not captured by the machine learning algorithm.

Keywords: unemployment, profiling, expectations, prediction, unemployment insurance.

JEL codes: J64, J65, C55, C53, C41, C21.

*University of Groningen, University Medical Center Groningen, IFAU Uppsala, IZA, ZEW, CEPR.

†University of Groningen, IAB Nuremberg.

‡IAB Nuremberg.

§IAB Nuremberg, Friedrich-Alexander-University Erlangen-Nuremberg.

¶CNRS and CREST, IAB Nuremberg, DIW, IZA.

Acknowledgements: We thank Eve Caroli and the participants of the final international conference of the DFG Priority Program 1764 for helpful comments and discussions. This project is part of the SPP 1764 program of the German Science Foundation (DFG). We gratefully acknowledge generous support of the DFG.

1 Introduction

Improving the reintegration chances of unemployed persons into the labor market remains an important challenge for policy makers as well as for the labor market administration. Of particular concern is the fact that a sizeable group of long-term unemployed individuals is deemed to be at a large distance from the labor market and to have a low probability of returning to work. A fraction of these individuals may have had better prospects of finding work earlier on in their unemployment spells. However, a combination of rather generous unemployment insurance benefits in combination with a lack of pressure to engage in active labor market policies and to keep search efforts high might result in an unemployment trap for these individuals.

The profiling of unemployed persons is an important tool that can contribute to an efficient use of placement resources and direct funds to where they are indeed effective. It can be used by caseworkers at the beginning and also during the placement process. The basic idea of profiling is that unemployed persons with different expected durations of unemployment require different intensities of counseling and placement activities. Often the outcome used for the profiling is the expected risk to become long-term unemployed.[1] Profiling thus assigns unemployed individuals into groups. The resulting categories shape the mix of active labor market policy (ALMP) treatments to which the individual is subsequently exposed. Individuals who can be expected to find a new job soon anyway do not need caseworker assistance.

By concentrating resources on those who are actually at high risks to stay unemployed for an extended period, services can be provided more effectively. Indeed, many labor market administrations have implemented and tried different profiling techniques to determine which jobseekers are in need of which further services. This approach as such is not new: Two prominent examples are the Worker Profiling and Reemployment System (WPRS) in the United States (e.g, O'Leary, 2015; Sullivan Jr et al., 2007), and the Job Seeker Classification Instrument (JSCI) in Australia (e.g, Lipp, 2005), which were both already implemented during the 1990s. In many US states, the WPRS ordered unemployed that could not expect a recall by the estimated probability to use up their unemployment benefit claims. Jobseekers were then assigned to local labor market offices, ranked by their probabilities to exhaust

[1] The survey question we base our main outcome on actually asks the jobseekers whether they think they will find a job within six months (see Section 3). Consequently, we operationalize our main outcome as reintegration into the labor market within six months. While this outcome only implicitly measures subjective predictions regarding the probability that an individual remains long-term unemployed (the alternative to reintegration within six months), we refer to our main outcome as long-term unemployment to improve readbility.

their benefits, until the capacities of those were exhausted. Black, Galdo, and Smith (2007); Black, Smith, Berger, and Noel (2003) analyze the WPRS in Kentucky, one of the most sophisticated systems then under WPRS, and conclude that the program accomplished its aim of shortening the duration of unemployment benefit receipt.

As administrative data sets encompassing unemployed persons have improved in size and quality, and computers have become able to handle larger datasets, statistical profiling has rapidly gained importance across labor administrations worldwide recently. The results of the profiling then provide some guidance about the services unemployed persons receive and their assignment to active labor market programs. A typical statistical approach would be that a model for the risk of becoming long-term unemployed is estimated based on past data using a regression approach. For newly registered jobseekers, predictions of the risk of long-term unemployment make use of the estimated coefficients.

Beyond traditional econometric techniques, economists have begun to take an interest in data mining and machine learning techniques, allowing for highly specific segmentations and supporting human decision making with automated algorithms (see also van Landeghem & Sam Struyven, 2021). Several countries such as Denmark and New Zealand, and regions such as Flanders have already adapted machine learning techniques for the prediction of long-term unemployment (Desiere, Langenbucher, & Struyven, 2019). Other countries (e.g. Germany) solely rely on caseworker-based profiling based on caseworkers' assessments of the labor market prospects of their clients.

A key aim of this paper is to compare the predictive power of different predictors. We are the first to contrast the predictive power of predictions based on machine learning methods, caseworker profiling results, and unemployed people's self-assessments of their employment prospects. We start with developing a methodological framework that allows us to contrast the predictors, expressing them in terms of underlying determinants of the correct duration distribution. We then conduct an empirical comparison of their predictive power. Our empirical analysis is based on high-quality administrative data records from the German Federal Employment Agency, including also caseworkers' profiling decisions, merged with survey data for a sample of unemployed persons. We also study to what extent combinations of the three data sources - administrative data on individuals, data on caseworker-profiling, and self-assessments - improve the quality of predictions.

Our empirical analysis finds that - in terms of overall predictive performance - random forest classifiers perform best, followed by self-assessments of the jobseekers themselves. One conclusion is that placement activities in Germany could be designed more efficiently and more accurately if statistical profiling results could be provided as a supportive tool for placement activities of caseworkers. We conclude our paper with a plea to test this in a field experiment in the German employment service.

2 Institutional background

2.1 Unemployment insurance system

In Germany, after becoming unemployed, individuals are usually entitled to unemployment benefits or welfare benefits, depending on their previous employment history. In order to receive unemployment benefits within the unemployment insurance system, certain conditions must generally be met. Most importantly, workers must have been employed for at least 12 months during the 30 months prior to registering as unemployed. The amount of unemployment benefits is 60 percent of the previous net wage. It increases to 67 percent if the unemployed person has children. The duration of unemployment benefit receipt depends on the duration of previous employment. For persons up to 50 years the maximum duration is 12 months.

Individuals who become unemployed and are not entitled to unemployment benefits (e.g. because they were not employed long enough or not employed at all before) may be entitled to welfare benefits, which are means tested and the level depends on the composition of the household the person lives in. People who have received unemployment benefits and continue to be unemployed after the end of the entitlement period for unemployment benefits may also be entitled to welfare benefits. It is an important goal of public employment services to prevent long-term unemployment and avoid welfare benefit receipt. To achieve this, employment agencies have a variety of tools to choose from. These range from more intensive counseling to wage subsidies or additional vocational training. Individuals who are at high risk of becoming LTU need special support to avoid this. Therefore, it is important to identify this high-risk group.

2.2 Profiling

The basic idea of profiling is that caseworkers of the employment services categorize jobseekers according to their expected duration of unemployment. Depending on the assigned category, caseworkers support jobseekers in different ways and with different intensities. Job seekers with a longer expected unemployment duration may require more intensive counseling and placement activities, whereas unemployed individuals with good reemployment prospects may need less caseworker assistance. Employment services need to use resources efficiently by concentrating on those people with a high risk of becoming long-term unemployed and therefore need special support.

There are different approaches for profiling the unemployed at entry into unemployment: rule-based profiling, caseworker-based profiling and statistical profiling (Desiere et al., 2019). Many countries rely on a combination of these different approaches. Rule-based profiling uses administrative eligibility criteria, e.g., age or education, to classify newly unemployed individuals into different categories. This method is easy to understand and implement and employment agencies often combine rule-based profiling with other types of profiling. Caseworker-based profiling (soft profiling) relies on the caseworkers' assessment of the job seekers' reemployment chances. Statistical profiling uses statistical models to predict the expected unemployment duration of an individual or their probability of becoming long-term unemployed based on administrative data and / or survey data.

Examples for countries where employment agencies use logistic or probit regressions to categorize job seekers include Australia, Austria, the Netherlands, and the US; Denmark, Belgium (Flanders), and New Zealand are among the few countries where employment agencies apply machine learning techniques to categorize job-seekers (Desiere et al., 2019).

In Germany, employment agencies use soft profiling. During the first meeting after unemployment entry, caseworkers assess whether they expect an unemployed person to find a job within six months or not and categorize jobseekers into different risk categories. We will return to this in Section 3.1.

3 Data and conceptual framework

We use a combination of administrative and survey data for our analyses (earlier used by van den Berg, Hofmann, Stephan, & Uhlendorff, 2016). We use the administrative data to predict long-term unemployment applying machine learning methods. Moreover, the administrative contain information on caseworker profiling. The survey data provide the unemployed individuals' assessment of their reemployment chances. In this section, we start with describing the data for the different predictors, followed by an explanation of how the predictors can be expressed in terms of the underlying individual unemployment duration distribution.[2]

[2]Note that we only consider unemployment entries into the unemployment insurance system. All individuals included in our final data are newly registered as unemployed and receive unemployment

3.1 Administrative register data

We use administrative data on the full population of unemployment entries in five German regions from Aug. 2011 - Jan. 2012 to train the machine learning algorithms we use to predict long-term unemployment.[3] The data stem from the Integrated Employment Biographies (IEB v.12.01.00) and contain the entire labor market histories of the unemployed job seekers, including detailed information on sociodemographic characteristics, employment and unemployment histories, and participation in active labor market programs.

We use the administrative data for four purposes: First, we use the data to construct our main explanatory variables and to train our machine learning models (details in Section 4). Second, we use the data to obtain information on our outcome of interest, the information whether the individual finds a job within six months.[4] Third, we use the data to obtain information from caseworker profiling activities. Fourth, we use the administrative data to compare model performance between the survey participants and the full population of unemployment entries in the employment agencies that participated in the experiment the survey data are based on.

3.2 Caseworker profiles

When a person registers as unemployed, caseworkers are supposed to assign job seekers different profiles. [5] These profiles, which are recorded in the administrative data, indicate, based on the caseworkers assessment, the chances of labor market integration and are the basis for the participation in different active labor market measures.

During our observation period, caseworkers categorized job seekers into six dis-

benefits.

[3]The survey that provides the individual predictions was conducted between Aug. 2012 and Jan. 2013 (for details, see Section 3.3). Therfore, we need the 2011-12 data for our machine learning approach to train the model (see Section 4). Additionally, we use the administrative data from Aug. 2012 - Jan. 2013 to compare model performance between the survey data and the administrative data.

[4]The participants of the survey were interviewed 4-6 weeks after unemployment entry and then asked about the coming six months (see Section 3.3). For individuals in the administrative data, we construct an artificial interview date to match the timeline of the survey participants. To do this, we select relevant new unemployment entries in the five German regions the survey was conducted in and define their hypothetical interview date as date of unemployment entry + 42 days. We then construct the outcome of interest by checking wether the individual is still unemployed six months after this hypothetical interview. For the remainder of the paper, we will simply refer to this as "six months after the interview".

[5]We only consider profiles which are up to one year old. Considering only very recent profiles that are 6 weeks old or less, the results do not change.

tinct risk categories. Two of those categories predict a reintegration into the labor market within six months, four predict an unemployment duration of more than six months.[6] Using those categories, we construct a binary variable that takes the value 1 if the caseworker expects the individual to find a job within six months and 0 otherwise.

3.3 Survey data

In addition to the administrative data, we have survey information from a sample of newly registered unemployed individuals in 5 regions in Germany. The jobseekers were interviewed about Integration Agreements, which are contracts between the employment agency and the unemployed, as part of a Randomized Controlled Trial conducted in 2012-2013. (earlier used by van den Berg et al., 2016) Participants were interviewed roughly 4 to 6 weeks after unemployment entry between August 2012 and January 2013. We restrict the sample to individuals over 25 years who are registered as unemployed and receive unemployment benefits at the time of their interview. In addition, we exclude individuals who were unemployed during the three months before their current unemployment spell. Finally, we restrict the survey sample to individuals who gave permission to merge their survey answers with the administrative data, who have answered the survey question central to our analyses, and for whom we observe a profile given by a caseworker no more than one year before the interview. We end up with 1198 observations.

The participants were asked a series of questions relating to their job search activities, their contacts with the federal employment agency, and some additional questions. For our purpose, the central question the participants were asked is related to their individual expectations regarding their chances to find a job.[7] The question reads as follows:

If you think about the future, how likely do you think it is that during the upcoming 6 months you will get a job again? Please give me a percentage. Here, a 0 means that during the upcoming 6 months you will with certainty NOT get a job, while 100 means that you will find a job with certainty.

Table 1 gives an overview of the distribution of answers for the self-assessment variable.

[6] For a detailed list of the categories and the distribution of the categories across samples, see Table A1.

[7] Due to this wording of the question, our outcome variable does not refer to the probability of becoming LTU, but to finding a job again in the relevant period.

[Table 1: Self-Assessed probability to find a job within six months]

We see that individuals tend to be optimistic. 48% of the survey participants are 100% sure that they will find a job within six months and only roughly 9% of the individuals think their chances are below 50%.

In Section 5, we will compare the predictive power of this self-assessment variable with the predictive power of a random forest classifier and with the predictive power of caseworker based assessments. To be able to compare the self-assessment with the fairly coarse caseworker profile categories (for more details, see Section 3.1), we construct a binary variable from the continuous self-assessment variable. For our main comparisons, the variable takes the value 1 if the self-predicted probability to find a job within six months is equal to or larger than 50% and the value 0 otherwise.[8]

3.4 Descriptive statistics

Table 2 shows selected descriptive statistics for the three different samples. Column (1) shows the descriptive statistics for the administrative data in 2011/12, Column (2) shows the descriptive statistics for the administrative data in 2012/13 and column (3) shows the descriptive statistics for the survey sample.

[Table 2: Descriptive Statistics Across Samples]

Comparing the administrative data in 2011/12 with the administrative data in 2012/13, we hardly see any differences in characteristics. The largest difference between those two samples is the fraction of individuals that find a job within six months after the interview, which is almost three percentage points lower in 2012/13 than in 2011/12. Comparing both administrative samples with the survey sample, we see that most characteristics are fairly similar across all three samples. The share of survey participants with a high-school degree is a bit higher in the administrative samples and the share of individuals with a vocational degree is higher in the survey sample than in the administrative sample. Finally, the fraction of individuals who find a job within six months is roughly 4-7 percentage points higher in the survey sample than in the administrative samples. Thus, the survey sample appears to be somewhat positively selected compared to the full population of unemployment entries in the employment agencies we consider.

[8] As there is some bunching at 50%, we also used an alternative threshold that only defines self-predicted probabilities larger than 50% as positive prediction. The results are very similar to the main results. In addition, we also present different performance measures across all relevant thresholds in Section 5.

3.5 Three predictors

Let time be measured in continuous time. The unit of time is 1 month and its origin is taken as the moment of entry into unemployment. Clients' self-assessments are collected one month after the moment of entry. For ease of exposition we assume that the outcome of interest is the following: moving to employment before $t = 7$ conditional on being unemployed at $t = 1$. This can be expressed as

$$\mathrm{I}(T \leq 7 | T \geq 1)$$

where I(.) is the binary indicator function being 1 iff its argument is true and T is the unemployment duration (or more precisely the duration until work) which is a random variable even at the individual level. Of course we can also focus on $\mathrm{I}(T > 7|T \geq 1)$ which is $1 - \mathrm{I}(T \leq 7|T \geq 1)$. We aim to predict this outcome.

We have three predictors, coming from different underlying information sources (in a nutshell, from the unemployed individual, from the caseworker and from a machine learning (ML) algorithm). Each of these predictors can be related to $\mathrm{I}(T \leq 7|T \geq 1)$ but the underlying information does not provide three predictors of exactly the same variable or outcome. It is useful to examine this in detail. Here we emphasize that T is a random variable whose realization is not known in advance.

From a methodological point of view, the main differences between the three predictors we consider are as follows:

(i) is the underlying information is about *whether* $T \gtrless \tau$ for some number τ or about some other feature of the distribution of T?

(ii) is the underlying information conditional on $T \geq 1$ or not?

Regarding (ii), the main reason for considering the conditioning on $T \geq 1$ in the first place is that the self-reported assessment is only collected among those with $T \geq 1$. With other underlying information being about whether $T > 6$ among newly unemployed workers, the key issue becomes whether we can relate $\mathrm{I}(T \leq 7|T \geq 1)$ to $\mathrm{I}(T \leq 6)$. As we shall see, and as is intuitively clear, in any setting except for very simple cases, the probabilities of these two events are not equal.

We now discuss the three predictors in detail. This serves to improve our understanding of the various predictors. We emphasize that this section does not propose models or estimation approaches to be actually used in our empirical analysis.

3.5.1 Predictor based on self-reported information from the unemployed individuals

This predictor is based on survey interviews among newly unemployed workers, held around 6 weeks after entry into unemployment (which for convenience we refer to as "1 month"). The original survey question is as follows:
This provides a self-reported version of $\Pr(T \leq 7|T \geq 1)$. In model settings, such a value of the conditional survival function at $t = 7$ is more informative than just knowing whether the median or mean of $T|T \geq 1$ exceeds 7 or not. However, the quality of the prediction also depends on whether the respondent understood the question.

To proceed, we introduce the index i to denote an individual. Let $\theta_i(t)$ and $\Theta_i(t)$ denote the hazard rate and integrated hazard rate of individual i, respectively, so

$$\Theta_i(t) = \int_0^t \theta_i(u)du$$

We can write

$$\Pr(T_i > 7|T_i \geq 1) = \exp(-\int_1^7 \theta_i(u)du) = \exp(-\Theta_i(7) + \Theta_i(1))$$

or

$$\log\left(-\log \Pr(T_i > 7|T_i \geq 1)\right) = \log \int_1^7 \theta_i(u)du$$

The "log – log" transformation has one major advantage: it provides an expression that can attain every value between $-\infty$ and ∞. We use y_i to denote the left-hand side after the transformation.

We now connect this to the observed data. We observe a self-reported version of $\Pr(T_i > 7|T_i \geq 1)$, or, in other words, we observe a self-reported version of $\log\left(-\log \Pr(T_i > 7|T_i \geq 1)\right)$ which is y_i. We take the observable version $\widetilde{y}_i$ of y_i to equal the true y_i plus an error term ε_i which may have a normal distribution with mean zero and variance σ_ε^2. For example, a true conditional probability of 0.5 results in the observed prediction $\widetilde{y}_i = \log\log 2 + \varepsilon_i$.

In general, we may write

$$\widetilde{y}_i = \log \int_1^7 \theta_i(u)du \;+\; \varepsilon_i$$

Of course we may consider modifications to deal with heaping of $\widetilde{y}_i$ and to deal with $\widetilde{y}_i$ being ∞ or $-\infty$. In the latter cases, if the self-reported version of $\Pr(T_i \leq 7|T_i \geq 1)$

is 0% or 100% we may replace this by 0.5% and 99.5%, respectively, so that $\widetilde{y}_i$ is just a very large positive or negative number.

Next, we consider the predictor in the context of a Mixed Proportional Hazard (MPH) model. We emphasize that this is not a model we assume to be valid in practice. It is just a simple model to shape thoughts. In fact, since at the individual level the distinction between observed and unobserved covariates is irrelevant, we may effectively write the individual hazard rate as

$$\theta_i(t) = \lambda(t) \exp(v_i)$$

with v_i unobserved. This immediately leads to

$$\widetilde{y}_i = \log(\Lambda(7) - \Lambda(1)) + v_i + \varepsilon_i$$

where $\Lambda(t) = \int_0^t \lambda(u)du$. This is like a regression with a constant term $\log(\Lambda(7) - \Lambda(1))$. This illustrates that the self-reported predictor can be very informative on the individual characteristics v_i. However, the informativeness critically depends on $\mathrm{var}(v)$ versus σ^2_ε. Moreover, if we drop the MPH assumption and allow for time-varying v_i then anything is possible.

As the next stage, we need to translate the observation into a predictor of the ultimate outcome of interest $\mathrm{I}(T \leq 7|T \geq 1)$. A simple approach is as follows. The outcome of interest is binary. As a result, the expectation of the outcome of interest is

$$\mathbb{E}(\mathrm{I}(T \leq 7|T \geq 1)) = \Pr(T \leq 7|T \geq 1)$$

Thus, if $\Pr(T \leq 7|T \geq 1) > 0.5$ then it is more likely that the outcome is $\mathrm{I}(T \leq 7|T \geq 1)$ than that the outcome is $\mathrm{I}(T > 7|T \geq 1)$. If $\Pr(T \leq 7|T \geq 1) < 0.5$ then the converse applies. Along these lines, we may use as a predictor whether the observed self-reported probability is larger or smaller than 0.5.

Since individuals may systematically over- or under-estimate these probabilities, we may, as a first step, look at a regression (or tabulation) of the realized values of T_i against the self-reported $\Pr(T_i \leq 7|T_i \geq 1)$. In the data, 9% (20%) of the individuals reports that the probability is smaller than 0.5 (is smaller than or equal to 0.5). The fraction of individuals in the data with actual duration outcomes $T_i > 7|T_i \geq 1$ is actually higher than 20%. We may therefore correct the self-reported data by finding a threshold c for the self-reported probability such that the proportion of individuals with a self-reported probability below c equals the fraction of individuals with a duration $T_i > 7|T_i \geq 1$. All individuals with a self-reported probability below c can then be assigned a predictor $T_i > 7|T_i \geq 1$. Note that this leads some individuals with a self-reported $\Pr(T_i \leq 7|T_i \geq 1)$ that exceeds 0.5 to get the prediction that $T_i > 7$.

We finish this subsection with three remarks. First, in the MPH context, the informational value of the self-reported probability is reduced when we move to the predictor of the binary outcome of interest.

Secondly, knowing the probability distribution of a duration variable does not suffice to predict an individual drawing from it. Therefore it is not realistic to aim for a 100% correct prediction score.

Thirdly, the framework in this subsection may open up opportunities to study self-reported survival probabilities in different contexts, e.g. by introducing observed covariates (such as over-optimism as a personality trait) or even by exploiting multiple-spell data. The former may provide an opportunity to make good use of survey data.

3.5.2 Predictor based on self-reported information from the caseworkers

This predictor is based on profiling among caseworkers, ideally carried out very close to the date of entry into unemployment. We basically observe whether the following statement is assessed to be correct:

The caseworker expects re-employment within 6 months. This may include a need for action to boost the motivation of the unemployed.

There are various ways to translate this into properties of the distribution of T. For example, it may relate to whether $\mathbb{E}(T) \leq 6$ or to whether the probability of finding work within 6 months $\Pr(T \leq 6)$ exceeds 0.5 (or some other specific number in-between 0.5 and 1 for $\Pr(T < 6)$). The former is closer to the phrasing of the statement but has the disadvantage that the mean duration also depends on the right-hand tail of the distribution. The latter only depends on the hazard from 0 to 6.

We first consider the former. We use that for any duration variable,

$$\mathbb{E}(T_i) = \int_0^\infty \exp(-\Theta_i(u))du$$

To relate $\mathrm{I}(\mathbb{E}(T_i) \leq 6)$ to the actually observed assessment, we may introduce a latent variable model. Let the expectation as perceived by the caseworker be denoted by $\widetilde{\mathbb{E}}(T_i)$, with

$$\log \widetilde{\mathbb{E}}(T_i) = \log \mathbb{E}(T_i) + \epsilon_i$$

where ϵ_i is a measurement error term which may have a normal distribution with mean zero and variance σ_ϵ^2. The caseworker agrees with the statement iff $\widetilde{\mathbb{E}}(T_i) \leq 6$, so iff

$$\log \mathbb{E}(T_i) + \epsilon_i \leq \log 6$$

This has a probit-like probability equal to

$$\Psi \left[\frac{\log 6 - \log \left(\int_0^\infty \exp(-\Theta_i(u)) du \right)}{\sigma_\epsilon} \right]$$

with Ψ the c.d.f. of a standard normal distribution. In the above MPH setting, this equation does not simplify. However, if we impose a Weibull duration dependence with $\lambda(t) = \alpha \lambda t^{\alpha-1}$ we obtain

$$\Psi \left[\frac{\gamma_0(\alpha, \lambda) + \frac{1}{\alpha} v_i}{\sigma_\epsilon} \right]$$

where γ_0 is a complicated function of the two Weibull parameters. This looks like a probit. At face value, it is much less informative on v_i than the first predictor. However, note that any comparison also depends on the variation in the measurement errors σ_ϵ^2 and σ_ε^2 and on α. Also note that having a good predictor it is not the same as having a precise estimate of v_i. After all, any estimate of v_i would have to be fed back into $\Pr(T_i < 6)$.

Now consider the approach in which the caseworker statement concerns whether the median $M(T)$ satisfies $M(T) \leq 6$ or not. The median is defined by $\Pr(T \leq M(T)) = 0.5$ so

$$M(T_i) = \Theta_i^{-1}(\log 2)$$

We may adopt again a latent variable model to connect the true median to the perceived median. With a Weibull duration dependence, we obtain

$$\Psi \left[\frac{\gamma_1(\alpha, \lambda) + \frac{1}{\alpha} v_i}{\sigma_\epsilon} \right]$$

where the function γ_1 is almost the same as the earlier function γ_0. Thus, the approach based on the median is virtually identical to the approach based on the expectation, in case of Weibull duration dependence.

Abstracting from the Weibull case, it is clear that the relevant expressions in this subsection depend on the hazard in the first month even if v_i is absent. This makes the usage of this predictor fundamentally different than the usage of the first predictor based on the clients' perceptions conditional on being unemployed for at least a month. Nevertheless, if σ_ϵ is much smaller than σ_ε then the caseworker prodictor may still perform better. Also, what the Weibull case illustrates is that the smoother the data-generating process, the more informative the caseworker assessments will be for $\Pr(T \leq 7|T \geq 1)$.

Another issue we need to address is the condition that the client may need a boost for motivation. An easy way out is that this is implicitly taken into account by everyone.

3.5.3 Predictor based on information from machine learning

Here we use variation in x_i to obtain a prediction model that follows a random forest trained on a different dataset. Either we predict $T_i|T_i \geq 1$ to get a prediction of $\mathrm{I}(T_i \leq 7|T_i \geq 1)$ or we directly predict $\mathrm{I}(T_i \leq 7|T_i \geq 1)$.

In the above model settings, one could interpret this as an approach where x_i provides information on Θ_i or on v_i.

We finish this subsection by pointing out that it has expressed the predictors in terms of underlying determinants of the correct duration distribution. In practice, one may sidestep this layer and instead simply choose some outcome to be predicted, define three candidate predictors and allow each of them to predict a part of the variation of the outcome. Subsequently, the relation between the three may be estimated. This is of course what we carry out in the remainder of the current paper. We emphasize that the latter approach is not inconsistent with the modeling approach in this section. The conceptual model framework simply serves to support the understanding of the empirical approach.

4 Machine learning methods

For the machine learning predictions we predict the probability that an individual becomes unemployed using a random forest classifier. In this section, we briefly introduce the random forest classifier we use for our prediction exercises. Furthermore, this section illustrates the different performance measures we use to compare the three different predictors.

4.1 Random forest

As it is one of the best-performing off-the-shelf machine learning techniques (Biau, 2012), and as several papers have shown that, in the context of classification of job seekers, random forest out-performs more traditional machine learning methods, such as logistic regression or OLS (Kern, Bach, Mautner, & Kreuter, 2021; Mühlbauer & Weber, 2022), we use a random forest classifier for our prediction exercises.

Random forest classifiers are based on a collection of tree classifiers that each cast a vote for the most popular class (Breiman, 2001). The goal of a tree classifier is to grow a decision tree by recursive binary splitting. In each step, the classification algorithm chooses the variables and the split point that achieve the best fit. The most common criterion used for splitting nodes and pruning the tree is the Gini index, which indicates how mixed the classes are in the two groups created by a split. Then one or both of these groups are split into two more groups. This procedure continues until a stopping rule is applied (Hastie, Tibshirani, & Friedman, 2011). Based on the majority vote, the classifier predicts a positive or a negative outcome. The individual trees are based on different random subsamples of the data and only a random subset of variables is used for each tree (Athey & Imbens, 2017). Essentially, a random forest can be interpreted as an average of many separate tree classifiers that have all been estimated on a subsample of the data (Athey, 2017).[9]

We train our random forest models using data on unemployment entries from one year prior to the experiment, i.e., 2011/12. As explanatory variables, we construct a series of variables on sociodemographic characteristics and educational background. Furthermore, we construct variables relating individual employment and unemployment histories, and participation in active labor market programs, yearly up to seven years back and lifetime. For a detailed list of all main explanatory variables, see Table A2 in the Appendix. Using these predictors, we then proceed to predict our outcome of interest for the survey sample, classifying all individuals, with a predicted probability to find a job of $>= 50\%$, as positive.[10] In addition to the models only based on sociodemographic characteristics and individual labor market histories, we also

[9]We use the Python module scikit-learn, version 0.24.1 (Pedregosa et al., 2011) (Pedregosa et al., 2011) for all analyses. Rather than letting each classifier vote for a class individually, the scikit-learn implementation of the random forest classifier averages probabilistic predictions of the individual classifiers.

[10]As a robustness check, we also predict our outcome of interest for the full population of unemployment entries in the five employment agencies that participated in the experiment using the full administrative data. We do this, both, for a holdout sample in 2011/12 and for the full sample of unemployment entries in 2012/13. The performance across these samples is similar to the performance in the survey sample (see Figure A1 in the Appendix).

train machine learning models using our main predictors in combination with the caseworker profiling information in order to investigate whether information based on caseworker profiles can improve predictive power.

4.2 Performance measures

For the comparison of the performance of the different prediction methods (self-assessment, caseworker assessment, and machine learning prediction), we focus on three different measures: the accuracy, the true positive rate (TPR) and the false positive rate (FPR). The measures are defined as follows:[11]

$$\text{Accuracy} = (\text{TP} + \text{TN}) \ / \ (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$
$$\text{TPR} = \text{TP} \ / \ (\text{TP} + \text{FP})$$
$$\text{FPR} = \text{FP} \ / \ (\text{FP} + \text{TN})$$

Accuracy measures the fraction of individuals we classify correctly, TPR measures the fraction of positives we identify among all positives and FPR measures the fraction of false positive classifications among all negative observations.

The drawback of Accuracy, TPR, and FPR is that they vary with the threshold at which a person is classified as likely to find a job within six months. Usually, for overall model performance irrespective of the classification threshold, the ROC-AUC Score is a popular performance measure for classification tasks. The ROC-AUC Score measures the connection between TPR and FPR for different classification thresholds. While it is possible to calculate this measure for the continuous self-assessment and machine learning classifications, it cannot be calculated for the caseworker assessment categories. Therefore, we cannot compare this measure accross all three predictors.

As we cannot compare the ROC-AUC Score across all three predictors, we additionally present an alternative measure for the explanatory power of the three predictors when we discuss our main results. To this end, we run separate regressions of our outcome on the individual predictors and compare which of the predictors achieves the highest explanatory power, measured as the R^2.

[11] TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

5 Results

5.1 Main results

We start our analyses by comparing the average prediction of the different predictors and the share of individuals that did indeed find a job within six months after the interview (Figure 1).

[Figure 1: Average Prediction and Average Actual Outcome (Threshold = 0.5)]

The random forest classifier without caseworker profiling information predicts that 67% of observations will find a job within six months after the interview and the random forest classifier including the caseworker profiling information predicts that roughly 66% of the individuals will find a job within six months. The job seekers themselves are much more optimistic: 80% of the survey participants predict that they will find a job within six months after the interview. The caseworker profiles indicate that only around 52% of the survey participants will find a job within six months. Finally, the fraction of individuals that had actually found a job within six months after the interview is just over 53%. Thus, on average, the caseworker classifications come closest to the actual share of individuals that find a job within six months.

However, these averages do not tell us how many job seekers are actually classified correctly. To investigate the actual prediction performance of the different predictors, we next turn to the TPR, FPR, and Accuracy (Figure 2).

[Figure 2: TPR / FPR / Accuracy (Threshold = 0.5)]

Similar to the averages, we see that the individuals are the most optimistic. As individuals largely believe in finding a job within six months, the TPR and FPR resulting from the self-assessment are the highest. The random forest classifiers classify a smaller share of individuals as positive and, consequently, TPR and FPR based on the machine learning models are lower than the self-assed ones. Finally, caseworker-based predictions are the least optimistic, resulting in the lowest TPR and FPR. In terms of accuracy, the random forest models outperform self-assessment and caseworker profiling: For a default threshold 0.5, the random forest model which includes case worker profiling has the highest accuracy (classifying 65.6% of the observations correctly), followed by random forest without caseworker profiling (64.4%), self-assessment (62.9%), and caseworker profiling (59.2%).

As an alternative measure for the prediction performance of our different predictors, we next compare their explanatory power. For this exercise, we run separate linear probability models, each including one of our predictors. As before, for the random forest and for the self-assessed probability, the variable takes the value 1 if the prediction is >= 50% and the value 0 otherwise. For the caseworker profiling, the variable takes the values 1 if the caseworker classified the job seeker as "close to the labor market" and 0 if the caseworker classified the job seeker as "not close to the labor market". Table 3 shows the results from the separate regressions including the respective predictors.

[Table 3: Predictive power of the different prediction methods, separate models, dummies]

As in the analyses before, the ranking of the predictors remains unchanged when we use the R^2 as performance measure: With 0.094, random forest including caseworker profiling information has the highest R^2, followed by random forest without caseworker profiling, which achieves a R^2 of 0.080. Self-assessment achieves a R^2 of 0.073, thus out-performing caseworker profiling, which achieves an R^2 of 0.033.

The results so far are all for a classification threshold of 0.5. However, the ranking of the performance measures may change at different threshold values. Therefore, Figure 3 compares the accuracy of the different classification methods over all relevant threshold values.

[Figure 3: Accuracy for different models across thresholds]

Looking at the performance of the models in terms of maximum accuracy (where the thresholds may differ for each model), the order of the models is the same as for the threshold of 0.5: We find highest maximum accuracy for the random model including information on caseworker profiling (65.6% for a threshold of 0.5), followed by random forest without caseworker profiling (65.1% for threshold of 0.51), self-assessment (64.4% for a threshold between 0.75 and 0.79 or 0.85 and 0.89), and caseworker profiling (59.2%). Moreover, Figure 3 shows that, for the majority of possible thresholds, the random forest model using caseworker information outperforms the random forest model without using this information with respect to accuracy.

We repeat the same exercise for the TPR and FPR. Figure 4 shows the TPR and FPR across all relevant thresholds. As the caseworker profiling is based on a single binary indicator, the TPR and FPR are fixed across all thresholds. Therefore, we use the TPR and the FPR of the caseworker profiling as a benchmark for our comparisons.

[Figure 4: TPR and FPR across thresholds]

Holding the FPR fixed at the level calculated based on the caseworker profiling (FPR=0.421), we see that the random forest classifier and the self-assessed predictions exhibit a higher TPR: Caseworker profiling (TPR=0.603) < random forest with caseworker profiling (TPR =0.692) < random forest without caseworker profiling (TPR=0.716) < self-assessment (TPR =0.733/0.738). We can also do the exercise the other way around: at the caseworker TPR level of 0.603, the FPR of the random forest classifier without/with caseworker profiling is considerably lower (FPR=0.330/0.324) than the caseworker FPR (=0.421). The same is true for the self-assessment (corresponding FPR=0.346). Thus, in terms of these performance measures, caseworker profiling based predictions also tend to perform worse than the random forest classifiers and the self-assessment.

This section has shown that random forest classifiers and self-assessment tend to out-perform predictions based on caseworker profiling in terms of overall predictive performance. At the TPR / FPR levels based on caseworker profiling, random forest classifiers and self-assessment also tend to perform better than caseworker profiling based predictions. However, combining random forest with information based on caseworker profiling, we achieve the highest overall accuracy / explanatory power.

5.2 Robustness checks

We checked whether the results are robust to changing the outcome definition. The main definition states that the individual has found a job within six months after the interview, not conditional on whether the person still has the job exactly six months after the interview. As an alternative, we checked whether the results change if we define the outcome as having a job exactly six months after interview (individuals could potentially interpret the question this way). However, this change in the outcome definition hardly affects the results and the main conclusions remain unchanged.

Regarding the caseworker profiles, our main definition states that we only include individuals with profiles that are no older than a year (with older profiles stemming from a previous unemployment spell). Thus, in the main specification, this variable includes some fairly old profiles that may not be reflecting the current labor market prospects of the job seeker. In order to address this potential issue, we additionally used a sample only including individuals with profiles no older than six weeks. The results using this restriction resemble the main results quite closely.

As discussed in 3.3, the self-assessment variable takes the value 1 if the self-assessed probability is equal to or greater than 50%. However, around 10% of all survey participants chose exactly 50% (see Table 1). Therefore, the performance of the self-assessed prediction may be sensitive the the exact choice of this threshold. However, changing this definition to greater than 50% hardly affects the results and the main conclusions remain unchanged.

Furthermore, the sample restrictions regarding age (excluding individuals below 25 years), recent unemployment history (excluding individuals who were unemployed during the three months before the current unemployment spell), and benefit receipt (excluding individuals who do not receive benefits at the time of the interview) yield a somewhat selective sample. In order to check whether these restrictions have an impact on prediction performance, we repeated our analyses without these restrictions. The results from this exercise are also very close to the main results.

To check whether more detailed information on individual labor market histories can improve the predictive power of the machine learning models, we additionally constructed monthly employment, unemployment, and active labor market measure participation histories going up to 25 years back, resulting in more than 1800 explanatory variables. Including those more detailed labor market histories, however, does not improve the predictive performance of the random forest classifier compared to our main set of explanatory variables.

6 Policy considerations and implications

Our analysis based on past data shows that machine learning based algorithms can outperform caseworker-based profiling. However, implementing of such a tool in a real environment is a project that brings new challenges. For Germany, we advocate a randomized controlled trial, where it is tested whether a new profiling tool to be developed will have positive effects on the labor market prospects of profiled individuals.

We would like to highlight a few problems that may be taken into account in such a real environment. First, caseworker acceptance of a profiling tool is likely to be a key problem. For instance, Behncke, Frölich, and Lechner (2009) conducted a pilot study on statistically assisted program selection (SAPS) in Switzerland. In this study, randomly selected caseworkers received a prediction of the program with the largest expected impact on reemployment prospects. This information was supposed to support caseworkers‘ decisions, who had full discretion in choosing integration strategies. A main result of the experiment was that caseworkers mostly ignored the information. Underlying reasons might be that they did not trust the predictions, felt that statistical tools curb their discretion, and might even have feared their replacement through an automated system in the longer run.

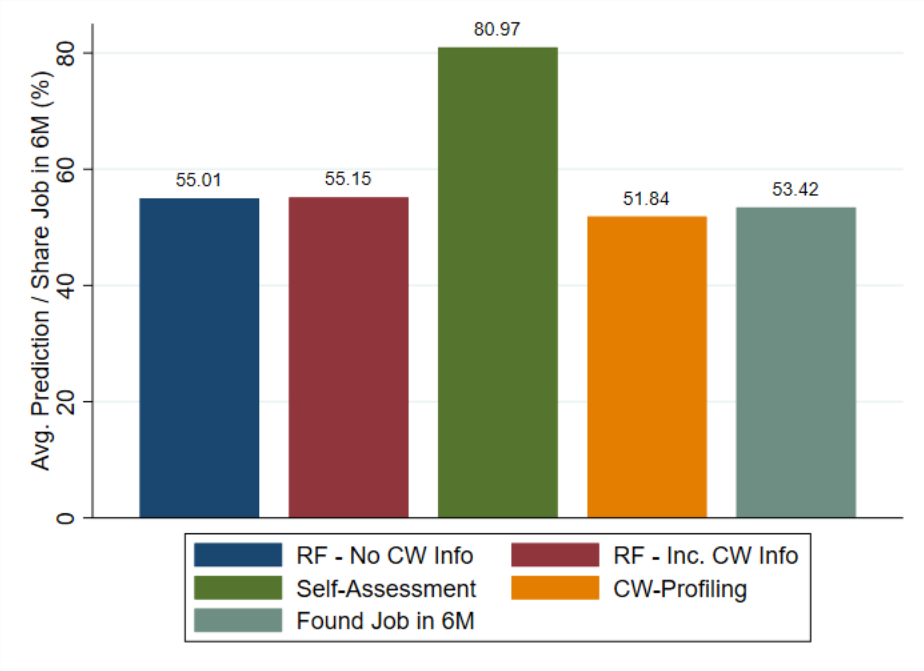
Second, data-based procedures can lead to discrimination (e.g, Desiere & Struyven, 2021; Kern et al., 2021). The most important predictive factor for the probability of a labor market integration of unemployed individuals is likely to be age. Older people are therefore likely to be almost always classified as having low chances of re-integration into the labor market by such procedures. If access to labor market programs is coupled with the classification, this is likely to severely affect their chances of being supported with labor policy instruments. The literature provides some suggestions to deal with this accuracy-equity trade-off.

Third, so-called scoring of individuals - which is what the profiling is ultimately about - can be legally questionable. As a matter of principle, such a prognosis should not be the sole deciding factor in determining which services individuals receive. In the EU, the general data protection regulation (GDPR) Article 22 states: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

The second and third problem can be demonstrated drawing on recent experiences of the Austrian labor market administration which tried to introduce a profiling tool (Holl, Kernbeiß, & Wagner-Pinter, 2018). It is based on a logistic regression model , where coefficients are estimated on the basis of data from the past and applied to people who are currently unemployed. Based on the criteria of gender, age group, state group, education, health impairment, care responsibilities, occupational group, labor market career, labor market type, a table with 81,000 prediction fields is created. On this basis, unemployed persons are segmented into "service customers" with good labor market prospects, "support customers" with low labor market prospects, and a residual category "counseling customers" with moderate labor market prospects. If a person is predicted to have at least a 66 percent probability of working at least 90 days in the 7 months following the start of unemployment, he or she is classified as a service customer. If a person is predicted to have less than a 25 percent probability of working at least 180 days in the 2 years following the onset of unemployment, he or she is classified as a support customer. The idea was that the client group should be decisive for the subsequent use of placement activities and active labor market policy instruments. Caseworkers would, however, be free to adjust the group assignment due to their own assessment. But the planned profiling led to strong protests in the Austrian public from the very beginning (Bachberger-Strolz, 2020). The use was then prohibited for data protection reasons, as the legal basis was not precise enough (application case of Art. 22 GDPR).

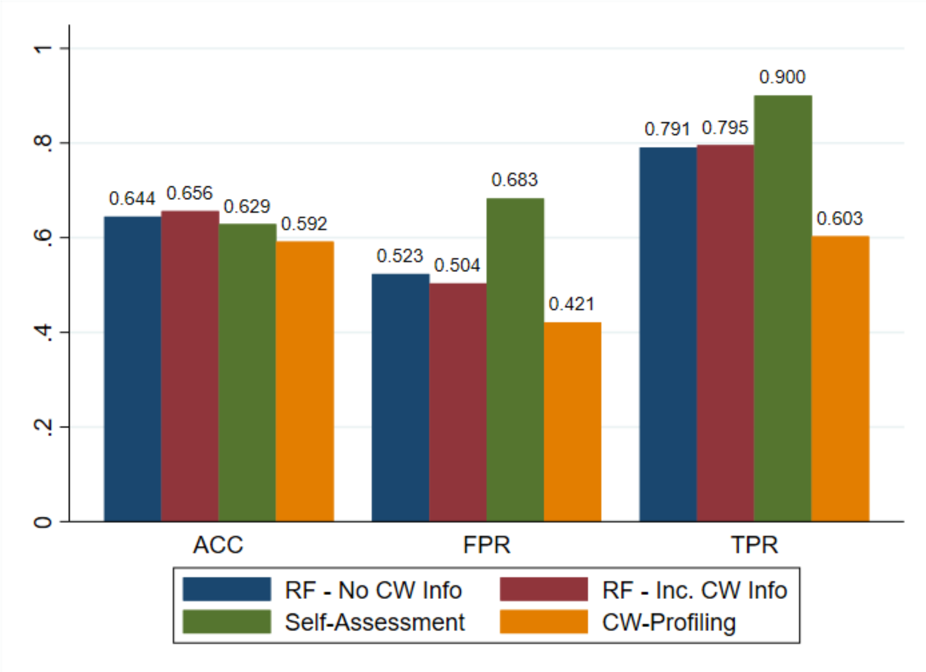
Figures and Tables

Figure 1: Average (predicted) Share Finding a Job within 6 Months



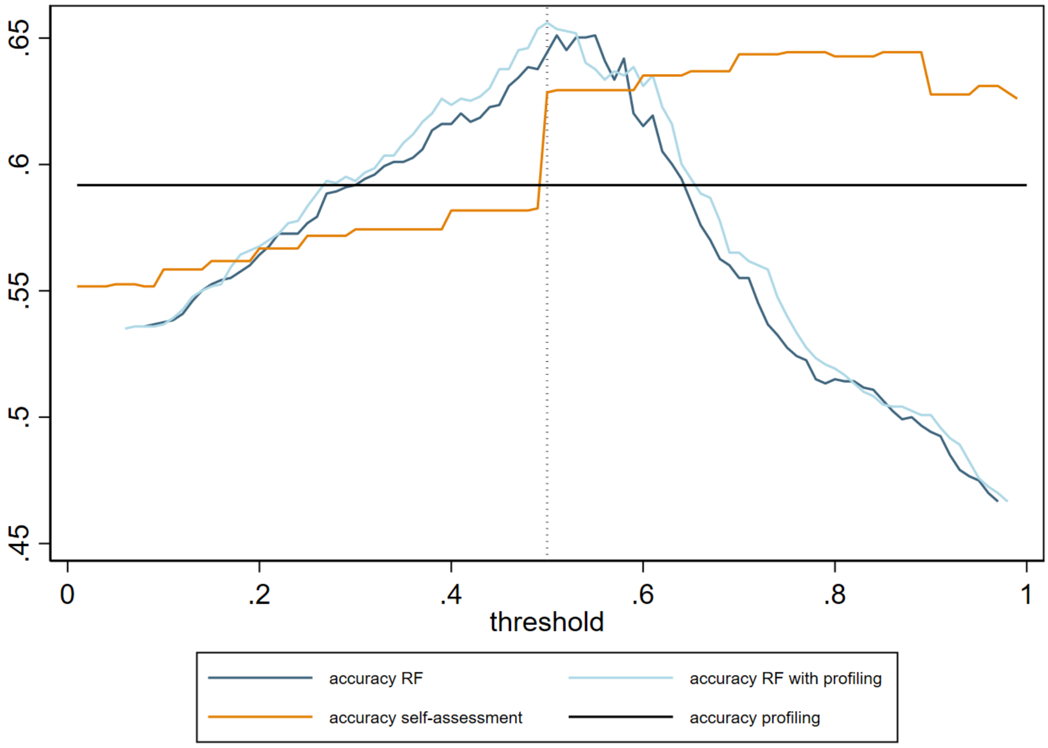
Note: This figure shows the average predicted share finding a job for each of the four predictors and the actual share of individuals finding a job within 6 months. Source: IEB v.12.01.00.

Figure 2: Accuracy, TPR, and FPR of the Four Predictors



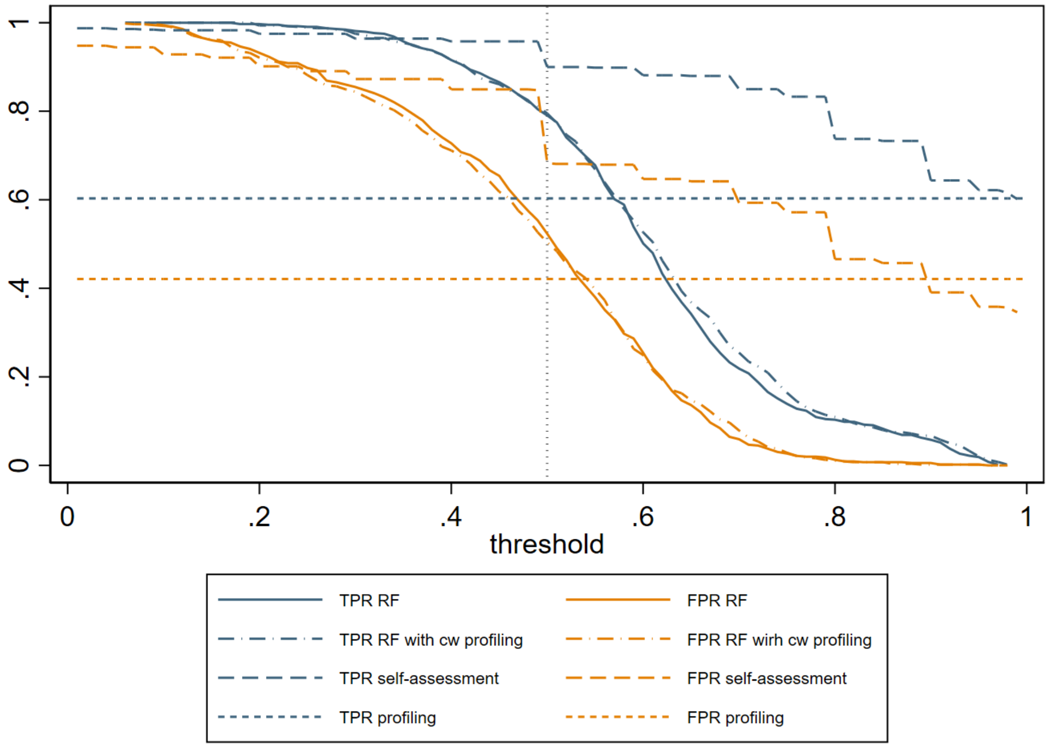
Note: This figure shows the Accuracy, TPR, and FPR for each of the four predictors. The threshold for a positive classification for the random forest models and for the self-assessment is 0.5. Source: IEB v.12.01.00.

Figure 3: Accuracy of the Four Predictors Across Thresholds



Note: This figure shows the TPR and FPR of the four different predictors across all relevant classification thresholds. Source: IEB v.12.01.00.

Figure 4: TPR and FPR of the Four Predictors Across Thresholds



Note: This figure shows the accuracies of the four different predictors across all relevant classification thresholds. Source: IEB v.12.01.00.

Table 1: Self-Assessed Probability to Find a Job within Six Months (in %)

Self-Assessed Prob.	N	%	% Cum.
0	36	3.01	3.01
1	1	0.08	3.09
5	3	0.25	3.34
8	1	0.08	3.42
10	10	0.83	4.26
15	4	0.33	4.59
20	16	1.34	5.93
25	6	0.50	6.43
30	17	1.42	7.85
40	17	1.42	9.27
49	1	0.08	9.35
50	129	10.77	20.12
51	1	0.08	20.2
55	2	0.17	20.37
60	29	2.42	22.79
65	4	0.33	23.12
70	46	3.84	26.96
75	23	1.92	28.88
80	120	10.02	38.9
85	8	0.67	39.57
90	94	7.85	47.41
95	32	2.67	50.08
98	5	0.42	50.5
99	15	1.25	51.75
100	578	48.25	100

Note: This table shows the subjective assessment whether on not an individual will find a job within six months, based on the answers of the jobseekers in the survey sample. Source: IEB v.12.01.00.

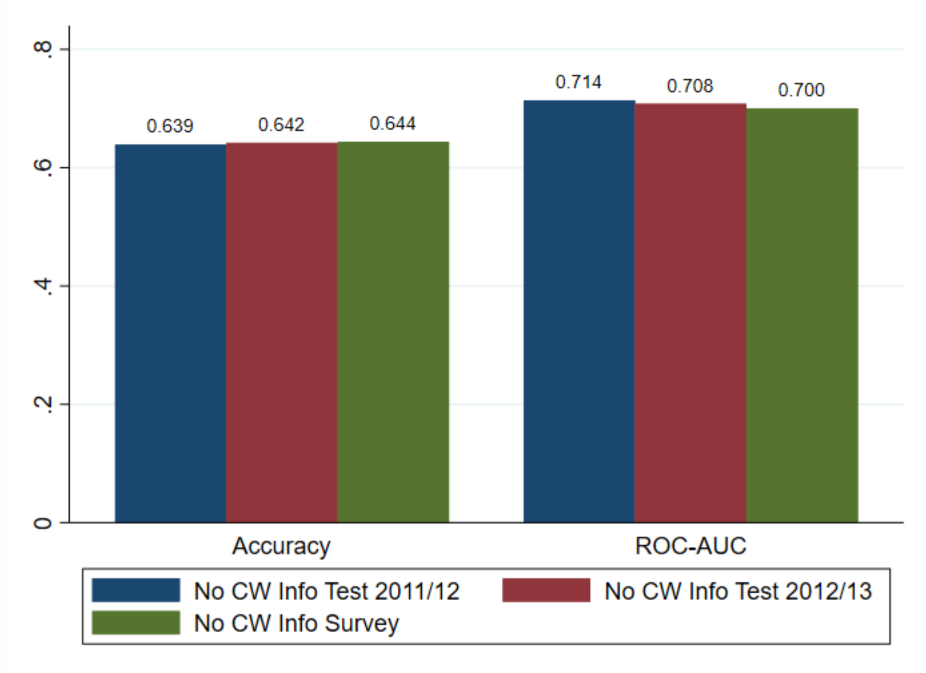
Table 2: Descriptive Statistics Across Samples

	Admin 2011-12	Admin 2012-13	Survey
Finds Job Within 6 Months	49.26%	46.47%	53.42%
Age	42.10	41.99	43.42
Male	56.19%	56.68%	57.18%
German	73.72%	73.49%	82.55%
High School	37.51%	39.39%	37.31%
Voc. Training	85.40%	85.01%	90.40%
University	21.78%	22.51%	21.45%
Daily Wage Last Job	64.24	66.82	67.06
Total Earn. Reg. Empl. Last Year	15365.16	16287.46	16450.62
Tot Dur. Reg. Empl. Last Year	231.68	236.55	246.91
Tot. Dur. Unemp. Last Year	57.01	56.38	52.86
N	29,130	30,255	1,198

Note: This table shows descriptive statistics for the different samples. Source: IEB v.12.01.00.

Appendix Figures and Tables

Figure A 1: Accuracy and ROC-AUC



Note: This figure shows the Accuracy and ROC-AUC of the machine learning model for the admin data 2011-12 and 2012-13 and for the survey sample. The accuracies are measured at a positive classification threshold of 0.5. Source: IEB v.12.01.00.

Table A 1: Caseworker Profiles Across Samples

	Admin 2011/12	Admin 2012/13	Survey
Marktprofil	39.63%	36.54%	38.23%
Aktivierungsprofil	8.06%	8.80%	13.61%
Forderprofil	39.53%	41.37%	41.57%
Entwicklungsprofil	10.45%	11.02%	6.26%
Stabilisierungsprofil	1.46%	1.58%	0.25%
Unterstutzungsprofil	0.87%	0.69%	0.08%

Note: This table shows detailed caseworker profiles for the different samples. The first two categories implicitly predict a reintegration into the labor market within 6 months, the last four categories implicitly predict a reintegration within more than six months. Source: IEB v.12.01.00.

Table A 2: Predictors

Predictor Group	Predictor
Sociodemographic Characteristics	Age
	Sex
	High School
	Vocational Degree
	University Degree
Info on Last Job / Current Minijob	Daily Wage Last Job
	Commute Last Job
	Part-Time / Full Time Last Job
	Minijob at Interview
	Earnings Minijob at Interview
Employment History	Tot. Days Employed Last 1-7 Years
	Tot. Days Employed Last 1-7 Years (Marg. Empl.)
	Tot. Days Employed Lifetime
	Tot. Days Employed Lifetime (Reg. Empl.)
	Tot. Days Employed Lifetime (Marg. Empl.)
	Tot. Earnings Last 1-7 Years
	Tot. Earnings Last 1-7 Years (Reg. Empl.)
	Tot. Earnings Lifetime (Lifetime)
	Tot. Earnings Last 1-7 Years (Marg. Empl.)
Unemployment History	Amount UE Benefits at Interview
	Tot. Days Receiving UE Ben. Last 1-7 Years
	Tot. Days Receiving UE Ben. Lifetime
	Tot. Amount UE Ben. Last 1-7 Years
	Tot. Amount UE Ben. Lifetime
	Tot. Days Registr. UE Last 1-7 Years
	Tot. Days Registr. UE Lifetime
	Tot. Days Labor Market Program Last 1-7 years
	Tot. Days Labor Market Program Lifetime
LHG History	LHG at interview
	Tot. Days LHG Lifetime
	Tot. Days LHG Last 1-7 Years
Caseworker Profile	Integration within 6 / 12 / >12 Months

Note: This table shows the predictors used to train the main random forest models used to predict reintegration into the labor market within six months. Source: IEB v.12.01.00.

References

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Bachberger-Strolz, K. (2020). Profiling, targeting, algorithmen, künstliche intelligenz–über die irrwege einer debatte in der arbeitsmarktpolitik. *Wirtschaft und Gesellschaft*, *46*(3), 329–363.

Behncke, S., Frölich, M., & Lechner, M. (2009). Targeting labour market programmes—results from a randomized experiment. *Swiss Journal of Economics and Statistics*, *145*(3), 221–268.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063–1095.

Black, D. A., Galdo, J., & 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., Berger, M. C., & Noel, B. J. (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.

Desiere, S., Langenbucher, K., & Struyven, L. (2019). Statistical profiling in public employment services: An international comparison.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning 2nd ed. t. hastie, ed.* Springer Science Business Media, LLC.

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). Das ams-arbeitsmarktchancen-modell.

Kern, C., Bach, R. L., Mautner, H., & Kreuter, F. (2021). Fairness in algorithmic profiling: A german case study. *arXiv preprint arXiv:2108.04134*.

Lipp, R. (2005). Job seeker profiling: The australian experience. In *Eu-profiling seminar.*

Mühlbauer, S., & Weber, E. (2022). *Machine learning for labour market matching* (Tech. Rep.). Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for ....

O'Leary, C. J. (2015). Evaluating public employment programs with field experiments: A survey of american evidence.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Sullivan Jr, W. F., Coffey, L., Kolovich, L., McGlew, C. W., Sanford, D., & Sullivan, R. (2007). Worker profiling and reemployment services evaluation of state worker profiling models: Final report. *ETA Occasional Paper*(2007-15).

van den Berg, G. J., Hofmann, B., Stephan, G., & Uhlendorff, A. (2016). *Eingliederungsvereinbarungen in der arbeitslosenversicherung: Nur ein teil der arbeitslosen profitiert von frühen abschlüssen* (Tech. Rep.). IAB-Kurzbericht.

van Landeghem, B. D., & Sam Struyven, L. (2021). Statistical profiling of unemployed jobseekers. *IZA World of Labor*.