The Labor Hub

Discussion Paper No. 19

ALGORITHMS AND ANTICIPATION: PREDICTIONS OF RE-EMPLOYMENT BY UNEMPLOYED WORKERS, THEIR CASEWORKERS AND MACHINE LEARNING

Gerard van den Berg, Arne Uhlendorff, max kunaschk, gesine stephan, and julia lang

Algorithms and anticipation: predictions of re-employment by unemployed workers, their caseworkers and machine learning

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

Abstract

Individual information about future idiosyncratic events may be used to improve AI-based predictions. Newly unemployed individuals as well as their caseworkers were asked for the perceived individual re-employment chances within 6 months. These are compared to algorithmic predictions from machine learning trained on large administrative data. Correcting the algorithm if the unemployed themselves predict long-term unemployment leads to a "super-predictor" with superior accuracy. The sources of this information advantage are obscured by anticipated future events not always being observable. We use individual risk aversion as indicator of incentives to collect information on future idiosyncratic events. We find that, indeed, the gain in accuracy of the super-predictor is concentrated among the risk-averse.

*University of Groningen, IFAU Uppsala, University Medical Center Groningen, IAB Nuremberg, ZEW, CEPR, J-PAL, Labor Hub.
†IAB Nuremberg.
‡IAB Nuremberg.
§IAB Nuremberg, Friedrich-Alexander-Universität Erlangen-Nürnberg.
¶CNRS and CREST, IAB Nuremberg, DIW, Labor Hub.

Keywords: expectations, prediction, unemployment, random forest, risk aversion, gender, information. **JEL codes:** J64, J65, C55, C53, C41, C21.
Acknowledgements: October 2026. We thank Eve Caroli, Xavier D'Haultfoeuille, Ivan Fernandez-Val, Anders Forslund, Pieter Gautier, Jeffrey Smith and participants at seminars in Amsterdam, Uppsala and Ghent and at conferences of the DFG SPP 1764, IZA/CREST, ESPE, IAAE, VfS, EALE, ZEW/CPHEB and a conference in Groningen on AI, for helpful comments and discussions. A preliminary version of this paper circulated under the title "Predicting re-employment: machine learning versus assessments by unemployed workers and by their caseworkers". We gratefully acknowledge support by the German Science Foundation as part of the SPP 1764 (grant numbers BE 4108/5-1 and STE 1424/5-1). We are grateful to the DIM unit of IAB for providing the data.

1 Introduction

The prediction of time spent in unemployment and risk of long-term unemployment convey important information for unemployment insurance (UI) agencies and their clients. Long-term unemployment is an adverse outcome of utmost societal relevance and can have lasting effects on an individual's career prospects and overall quality of life. Moreover, the agencies' budget and their mix of activities critically depend on the distribution and composition of unemployment.

UI agencies have traditionally relied on regression methods or on subjective profiling by caseworkers for such prediction purposes. The underlying idea is that each unemployed individual is assigned into one of a few categories, based on the expected unemployment duration as predicted by a regression or as perceived by the caseworker.[1,2] Recently, UI agencies have become interested in machine learning methods for predicting whether new clients leave unemployment within a certain time frame (see van Landeghem, Desiere, and Struyven, 2021). Denmark, New Zealand and Flanders in Belgium have implemented such an approach (Desiere, Langenbucher, and Struyven, 2019).

In this study, we compare three different types of predictions regarding the probability of finding a job within six months after unemployment entry: self-assessments by the unemployed individuals themselves, assessments by caseworkers, and predictions based on machine learning algorithms. We develop a methodological framework to understand differences between the predictors and we conduct an empirical comparison of their power. The empirical analysis is based on high-quality administrative data records from the German Federal Employment Agency, including caseworkers' decisions, merged with survey data for a sample of unemployed persons. For the machine learning approach we adopt random forest classifiers.

The predictive power of agents' subjective expectations has been documented for a wide range of economic outcomes (Manski, 2018; Mueller and Spinnewijn, 2023b). Dominitz (1998) and Stoltenberg and Uhlendorff (2022) demonstrate that expected earnings are correlated with future realizations after controlling for current earnings and a range of characteristics. Similar evidence has been provided for the risk of job loss (Campbell, Carruth, Dickerson, and Green, 2007; Hendren, 2017; Stephens, 2004) and longevity and death (Smith, Taylor, and Sloan, 2001). Mueller, Spinnewijn, and Topa (2021) provide evidence that self-assessed job find-

[1] We use "long-term unemployment" and "no re-employment within six months" interchangeably except where the difference is relevant.

[2] Examples of countries where employment agencies use logistic or probit regressions to categorize job seekers include Australia, Austria, the Netherlands and the US.

ing probabilities of unemployed workers are predictive for actual outcomes. They additionally document an on-average optimism bias among job seekers. While the literature analyzing the predictive power of subjective expectations typically makes comparisons to linear regressions controlling for a rather small set of observed characteristics, we are the first paper analyzing this question by employing more powerful machine learning techniques in combination with detailed administrative data that also contain caseworker information on the predicted reemployment probabilities.[3] Regarding caseworkers, it is useful to cite the influential study by Lechner and Smith (2007) which focuses on the role of caseworkers when assigning clients to ALMP. They find that assignment rules based on observed individual characteristics are superior when it comes to speeding up reemployment. This casts some doubt on the efficacy of caseworkers, and it is interesting to see if there is a relation with caseworker prediction of RE6.

We study to what extent combinations of the three data sources - administrative data on individuals, data on assessments by caseworkers, and self-assessments - improve the quality of predictions. It turns out that a combination of approaches along the lines of Granger and Ramanathan (1984) is more accurate in predicting those at risk of long-term unemployment than the predictors separately. Notably, taking the self-assessments of the unemployed into account improves on the machine learning algorithm, and a self-assessment of long-term unemployment should override a random-forest prediction of short-term unemployment. This leads to a number of interesting additional research questions and implications. Is it possible to describe the individuals whose self-assessment adds value to the super-predictor as compared to the machine learning predictor? In principle, the machine learning algorithm should account for the systematic determinants of the reemployment probability. There may be already-scheduled idiosyncratic events in the near future, which inform the self-assessment but which may not be observed in the administrative data (and indeed they may never materialize because of avoidance strategies). In the paper we study various approaches to deal with this. In particular, we design and apply an

[3]Mueller and Spinnewijn (2023a) study the prediction of long-term unemployment with machine learning techniques using administrative data from Sweden. They demonstrate that the predictive power can be substantially increased – compared to the use of a standard set of socio-demographic variables – by the inclusion of detailed information on employment histories. Bodilsen, Nielsen, and Rosholm (2025) use a large dataset containing administrative records as well as personality traits and attitudes of unemployed individuals, to predict long-term unemployment using machine learning methods. They conclude that machine learning algorithms trained on both sets of variables simultaneously provides the best predictions. Their population of interest differs somewhat from ours as it concerns hard-to-place individuals operating at the margin of society, with individual labor market histories that are plausibly less rich than those among the inflow into regular unemployment.

approach that relates a superior self-assessment to personality traits that affect the incentive to collect private information about the future. We find strong evidence for such an association. This result is strongly gender-dependent. This may be because males are more over-optimistic on average, but it may also be a result of females in the sample encountering more idiosyncratic events in the near future than males. In any case, the results lend credence to the notion that self-assessments reflect private knowledge about upcoming events that is not captured in the machine learning algorithm.

On a fundamental level, our analysis targets a conception that machine learning algorithms are superior predictors of human behavior because they are suppoed to embody every systematic determinant of the outcome of interest. What this conception potentially ignores is that humans are forward-looking and have every incentive to collect information about future events. To the extent that such future events are idiosyncratic and unrelated to those systematic determinants but are foreseen by the involved individuals, the self-assessments by those individuals have predictive value that goes beyond the machine learning algorithm.

In sensitivity analyses we examine the role of subsets of personal characteristics and individual histories as inputs into the random forest classifier. We also examine various other extensions, such as the inclusion of caseworker assessments in the data used to train the algorithm, and the usage of random forests to predict caseworker assessments. We describe how subjective self-assessments may be used as inputs for future machine learning algorithms to make individual predictions.

In monetary economics, a few recent studies on inflation forecasting use machine learning in conjunction with predictions by professional forecasters and experts (see Araujo, Bruno, Marcucci, Schmidt, and Tissot, 2023, for an informal overview). Framed in our context, predictions of such experts can be positioned in-between caseworker assessments and self-assessments by the unemployed. Throughout the paper we refer to findings in this literature that are relevant for our purposes.

The remainder of this paper is as follows: Section 2 describes the institutional background of the UI system in Germany and the caseworker assessment of job seekers. Section 3 gives an overview of the data we use. Section 4 presents the conceptual framework and Section 5 outlines the measures we use to assess the performance of the different predictors. Section 6 presents the baseline results. Section 7 constructs the super-predictor and Section 8 examines in which cases subjective assessments can be used to improve on the machine learning predictor. Section 9 contains a wide range of extensions and sensitivity analyses, including a study of the importance of various subsets of background characteristics and individual histories for training

the machine learning algorithm. Section 10 concludes.

2 The German unemployment insurance system

In Germany, upon becoming unemployed, individuals are entitled to benefits within the Unemployment Insurance (UI) system provided that certain conditions are met. Most importantly, workers must have been employed for at least 12 months during the 30 months (24 months at the time of our survey) prior to registering as unemployed. The amount of UI benefits is 60 percent of the previous net wage. This 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; for those aged above 50 the maximum duration increases to up to 24 months. To improve the labor market prospects of unemployed persons, employment agencies have a variety of tools and active labor market programs (ALMP) to choose from. These range from counseling to wage subsidies and additional vocational training.

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, where the level depends on the composition of the household. 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.

To receive UI benefits, the individual needs to register as a job seeker. Registration should take place within three months prior to the end of an employment relationship or three days after receiving notice of the end of the employment relationship the latest. Employment agencies offer an early meeting soon after registration. However, it is generally accepted if job seekers excuse themselves for such a meeting, e.g. because they do not want to miss attendance at their current employer. The first meeting with the caseworker often takes place around the date of actual unemployment entry. The mean duration of the first meeting was around 50 minutes.[4]

Caseworkers in employment agencies in Germany apply a so-called soft-profiling

[4]This comes from a survey of caseworkers that took place in 2012 and 2013 which overlaps with our observation window for unemployment entry; see van den Berg, Hofmann, Stephan, and Uhlendorff (2014). At the time, no virtual meetings took place yet. No data source on caseloads per caseworker is available for 2012 and 2013, but during the year 2016, a jobseeker-oriented caseworker was responsible for on average around 160 clients (Bundesregierung, 2019).

approach to categorize their incoming UI clients.[5] That is, during the first meeting, they subjectively assess whether they expect an unemployed person to find a job within six months or not and categorize jobseekers into different risk categories. During our observation period, caseworkers categorized job seekers into six distinct risk categories. The first and second category correspond to re-employment within 6 months (RE6) while categories three to six correspond to a later re-employment.[6]

3 Data

3.1 Survey data

For our analysis, we use a combination of survey data and administrative data. The survey data provide the starting point of our data description. A sample of newly registered unemployed individuals were asked to provide subjective assessments of their probability to re-enter employment within 6 months.[7] As explained in subsections below, the administrative data are used to apply machine learning methods and contain information on caseworker assessments, including those for the survey respondents.

The survey was conducted in five regions which were chosen to be jointly representative of the German labor market. Participants were interviewed around 4 to 6 weeks after unemployment entry between August 2012 and January 2013. We restrict the sample to individuals who, at the time of the interview, were over 25 years old and were registered as unemployed and as not having paid work and as receiving UI benefits, and who, in the survey interview, confirm that they are unemployed. We exclude individuals who were registered as unemployed during the three months prior to the onset of 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,

[5]Various approaches exist to categorize newly unemployed individuals; notably: rule-based profiling, caseworker-based profiling and statistical profiling (Desiere et al., 2019). Some countries rely on a combination of these approaches. Rule-based profiling uses administrative eligibility criteria, e.g., age or education. Caseworker-based profiling (soft profiling) relies on the caseworkers' assessment of the job seekers' reemployment chances. Statistical profiling uses statistical regression models to predict the expected unemployment duration of an individual or their probability of becoming long-term unemployed.

[6]Details are in the next sections. The first two categories do not differ in terms of the assessed time to re-employment, and neither do the third and fourth category. The fifth and sixth concern expected durations of more than a year but these categories constitute less than 0.4% of the sample.

[7]This survey was designed by the IAB Nuremberg for studying variations in the placement process; see Stöhr (2016). The latter found no relation between process details and perceived re-employment prospects.

and for whom we observe a recent assessment by a caseworker (see below for details). This results in 1158 observations. Subsection 3.3 presents a table with descriptive statistics.

For our purposes, the key survey question concerns the individual subjective probability to become re-employed within 6 months. The original survey question is as follows:

If you think about the future, how likely do you think it is that during the upcoming six months you will get a job again? Please give me a percentage. Here, a 0 means that during the upcoming six 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.[8] We see that 48% of the survey participants are 100% certain that they will find a job within six months while roughly 20% of the individuals state their chances to equal or to be below 50%.[9]

3.2 Administrative data

As second data source, we draw on administrative data encompassing the full population of unemployment entries in the five German labor market regions for selected time periods. We utilize data from the Integrated Employment Biographies (IEB v.12.01.00) that contain information on times in employment (due to social security contributions), registered job search, unemployment and welfare benefit receipt, as well as participation in labor market programs administered by the federal employment agency. The data also contain detailed information on the socio-demographic characteristics of individuals.

To the IEB data, we merge information on the assessment by the caseworker. This concerns the soft profiling that caseworkers are supposed to conduct when a person registers as unemployed. This entails an observation of whether the caseworker

[8]Note that due to the wording of the question, our outcome variable does not refer to the probability of re-employment within 6 months after entry but to the probability of re-employment after 7 months conditional on survival up to 1 month; see the next section.

[9]Note that the confrontation of the self-assessment with the realized outcome provides a measure of individual optimism bias. To appreciate this it is important to emphasize that the survey interviews were carried out by an independent company (using CATI) and not by caseworkers or other employees of public agencies. To avoid desirability biases, the interviewer informed the respondent at the outset that the survey is conducted by the company in cooperation with the Institute for Employment Research, and a guarantee was given that responses are treated with utmost confidentiality and remain anonymous. Also, it was emphasized that usage of the data for purposes other than specific academic research was prohibited. Notably, it was not provided to the caseworker. In the interview, permission was requested to link the responses to administrative data, under the above-mentioned provisos.

expects re-employment within 6 months (which may include a need for action to boost the motivation of the unemployed).[10] Accordingly, we compute a binary variable that takes the value 1 if the caseworker expects the individual to find a job within six months and 0 otherwise.

The administrative data we use contain entries into unemployment in the five labor market regions between August 2012 and January 2013, as well as entries in those regions one year earlier, so between August 2011 and January 2012.[11] These data provide explanatory variables and the individual outcome of interest. We use the entries from 2011/12 to train the machine learning algorithm to predict RE6, i.e., to predict whether an individual finds a job within six months.[12] This is motivated by the fact that 2011 and 2012 are comparable years in terms of labor market conditions and in terms of absolute levels of flows into and out of UI among individuals aged 25-64 in the five regions (see Statistics of the FEA, 2019). The year 2011 was slightly more favorable than 2012, but the relevant flow sizes differ only up to about 5% between the two years. Even when considering regions and 10-year age groups separately, differences in relevant flows between the two years do not exceed 10% in any subgroup. Below we also consider the alternative approach that is based on the training on data on the unemployment entries in the period during which the survey was actually conducted.

3.3 Descriptive statistics

Table 2 shows selected descriptive statistics: column (1) for the administrative data in 2011/12, column (2) for the administrative data in 2012/13, and column (3) for the survey sample. There are hardly any differences between columns (1) and (2). This is in line with the above-mentioned similarity of labor market statistics for 2011 and 2012. The largest difference concerns the average fraction finding a job within six months after the hypothetical interview date; this is about three percentage points lower in 2012/13 than in 2011/12. At first sight it might appear that machine learning classifiers based on the 2011/12 data are therefore too optimistic for outcomes in 2012/13. Such a concern ignores variation in background characteristics and histories that are part of the algorithm. In the paper we use

[10] In our main analysis, we allow for profiles that are up to one year old. Profiles that antedate entry were determined in a recent previous unemployment spell. When we omit these and only consider individuals with profiles that are 6 weeks old or less (and hence concern the current spell), the sample size decreases somewhat but the results are similar to the main results (see Subsection 9.2).

[11] In the remainder of the paper, we refer to these periods as 2012/13 and as 2011/12, respectively.

[12] For a detailed definition of the outcome variable, see Subsection 4.3.

various approaches to verify and/or correct for any residual over-optimism, e.g. by using different thresholds for the algorithm and by using different training data sets.

Comparing the administrative samples with the survey sample, we see that basic personal characteristics are fairly similar on average. The main differences are that the share of native Germans and the share of individuals with a vocational degree is higher in the survey sample than in the administrative samples. 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.[13]

4 Conceptual framework

To relate the three predictors to each other and to understand their differences we outline a unifying framework, where it should be emphasized that we do not impose any such framework when analyzing the actual data. Let time be continuous. 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 now, we therefore take the outcome of interest to be 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 at the individual level. We aim to predict this outcome.

We have three predictors, coming from different underlying information sources. Each of these can be related to $\mathrm{I}(T \leq 7 | T \geq 1)$, where the realization of T is not known in advance. Here, the key differences between the predictors concern the following:

(i) is the underlying information 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 (i), note that the caseworkers' information is binary. Regarding (ii), the

[13] See also Table 1. As it turns out, this positive selection does not influence the overall performance of our machine learning prediction model (see Subsection 4.3).

conditioning on $T \geq 1$ is relevant for the self-reported assessment. Obviously, the probabilities of $\mathrm{I}(T \leq 7|T \geq 1)$ and $\mathrm{I}(T \leq 6)$ are not equal except in simple settings.

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

The survey provides a self-reported version of $\Pr(T \leq 7|T \geq 1)$. In model settings, this individual conditional probability is more informative than just knowing whether the median or mean of $T|T \geq 1$ exceeds 7 or not. Indeed, in specific duration models, this conditional probability is informative on unobserved individual characteristics. However, the quality of the self-reported version also depends on whether the respondent understood the question. Appendix C provides a detailed exposition for specific models, allowing for measurement errors in the self-reported outcome.

We now translate the reported observation of $\Pr(T \leq 7|T \geq 1)$ 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.

In practice, individuals i may systematically over- or under-estimate these probabilities. Therefore, as a first step, we may 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 survey data, 9% (20%) of the individuals report that the probability is smaller than 0.5 (is smaller than or equal to 0.5). The fraction of individuals in these data with actual duration outcomes $T_i > 7|T_i \geq 1$ strictly exceeds 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 actual 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 the prediction $T_i > 7|T_i \geq 1$. We consider such an approach below. Note that this leads some individuals with a self-reported $\Pr(T_i \leq 7|T_i \geq 1)$ that exceeds 0.5 to be assigned the prediction that $T_i > 7$.

4.2 Predictor based on self-reported information from the caseworkers

From caseworker records on newly unemployed clients we observe whether the caseworker expects re-employment within 6 months. 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 re-employment probability within six months $\Pr(T \leq 6)$ exceeds 0.5. The former is closer to the meaning of the word "expectation" but has the disadvantage that the mean depends on the right-hand tail of the distribution. In model settings, the two approaches lead to similar predictors (see Appendix C).

Clearly, the value reported by the caseworker depends on the perceived re-employment rate within the first month, which makes it different from the predictor based on clients' perceptions. In the end, a comparison will reflect measurement errors as well, so that it is an empirical question which predictor performs best.

The caseworker's assessment may occasionally include a stipulation that the client's motivation will be stimulated in order to expect re-employment within 6 months. We do not further address this condition but simply assume that it is implicitly taken into account by everyone.[14]

4.3 Predictor based on machine learning

For individuals in the administrative data from 2011/12, we construct a hypothetical 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 the date of unemployment entry + 42 days.[15] We refer to $T_i = 7$ as "six months after the (hypothetical) interview".

Next, we construct the outcome of interest for the algorithm, by verifying whether

[14]Note that in such a situation a caseworker may choose one out of two possible predictions that are each valid depending on whether such a stumulating measure is provided or not. We assume that the caseworker adopts the policy regime with the stimulation and accordingly reports the highest re-employment probability, as this conforms to the agency's objectives. We then assume that this is common knowledge among the unemployed.

[15]Thus, we exclude individuals who exited unemployment within 42 days after registering as unemployed.

the individual has found a job within six months after this hypothetical interview.[16] Specifically, we measure this as a binary outcome, which is 1 if the person has found a job within six months and 0 otherwise. In our main specification, and in line with the subjective assessments, we do not condition on the individual still having the job 6 months after the interview. Thus, if someone finds a job within six months after the interview and then becomes unemployed again, the person is counted as having found a job. Consequently, right-censoring of unemployment spells is not a concern in our analyses.[17] We then use a random forest classifier to predict whether the individual is re-employed within six months.

Section 4 so far has discussed predictors in terms of an underlying duration distribution. In the remainder of the paper we sidestep this layer and confine ourselves to a comparison at face value of the three candidate predictors for the outcome variable of interest.

Random forest. Random forest is one of the best-performing off-the-shelf machine learning techniques (Biau, 2012). Several studies have shown that, in the context of classification of job seekers in Germany, random forest outperforms more traditional methods such as logistic regression or OLS (Kern, Bach, Mautner, and Kreuter, 2021; Kunaschk and Lang, 2022; Mühlbauer and Weber, 2022).[18] Random forest classifiers are based on a collection of tree classifiers that each cast a vote for the most popular class (Breiman, 2001).[19] The individual trees are based on different random bootstrapped subsamples of the data and only a random subset of background variables (or "features") is used for each tree (Athey and Imbens, 2017). Thus, a random forest can be interpreted as an average of many separate tree

[16] We only observe dependent employment in jobs that are liable to social security contributions. Thus, individuals who, for example, become self-employed or leave Germany are not counted as not having found a job at that point in time. This may lead to a discrepancy between subjective assessments and outcomes observed in the administrative data. Such cases should be rare in practice.

[17] To account for the possibility that individuals interpreted the question differently, we additionally report the results for a specification where we condition on the person still having a job exactly 6 months after the interview, in Subsection 9.2. The results based on that definition are similar to the main results.

[18] We additionally predicted our main outcome of interest using a Gradient Boosting Classifier and a classifier based on Logit Regression (see Subsection 9.2).

[19] A tree classifier grows a decision tree by recursive binary splitting. In each step, the classification algorithm chooses the variable and split-point that achieves the maximum variation between the groups created by the split. The most common criterion used for splitting nodes and pruning the tree is the Gini index. Next, one or both of these groups are split into two more groups. This procedure continues until a stopping rule is met (Hastie, Tibshirani, and Friedman, 2011). Based on the majority vote, the classifier predicts a positive or a negative outcome.

classifiers that have all been estimated on a subsample of the data (Athey, 2017).[20]

We train our random forest models using data on our selection of unemployment entries from 2011/12.[21] As explanatory variables (or "features", in machine learning terminology), we use sociodemographic characteristics, educational background, variables relating to individual employment and unemployment histories, including wages and benefits, and past participation in active labor market programs, annually up to seven years back or for the elapsed lifetime.[22] Table A2 in Appendix A provides a list of explanatory variables. Note that the random forest effectively allows for full interaction effects of these variables as well. Using the resulting algorithm, we predict the individual outcome for those in the survey, classifying all individuals with a predicted probability to find a job of $> 50\%$, as "positive".[23]

Next to the predictions that are only based on sociodemographic characteristics and individual labor market histories, we also train algorithms that add the caseworker profiling information, to see if this improves predictive power. Importantly, we cannot train models including the clients' self-assessment using past data, as we do not have survey information for unemployment entries in 2011/12.[24] Due to the modest size of the survey sample, it is not feasible to split the survey data into (even smaller) training and test data sets to perform a machine learning approach on that. Analyses using training sets of different sizes from 2011/12 indicate that

[20]We use the Python module scikit-learn, version 0.24.1 (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, and Dubourg, 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. A 90% fraction of the total sample is divided into five folds and for each combination of four out of these five folds a random forest classifier is built, where the bootstrapping takes place with replacement and root-p subsets of the number p of features are used, all for given hyperparameters. These classifiers are averaged. Subsequently, the hyperparameters are chosen by invoking grid-search cross-validation, using the remaining 10% of the total sample and the ROC-AUC criterium (see below). Most of the results in the paper apply a maximum tree depth of five.

[21]For our main analyses, we apply hyperparameter tuning, varying the maximum depth of each tree in the random forest, the minimum number of subsamples per leaf, and the minimum number of subsamples per split. This follows explorations in Kunaschk and Lang (2022). Results are found to be insensitive to tuning details.

[22]Additionally, we construct monthly labor market histories dating back up to 25 years before unemployment entry. This does not noticeably improve prediction performance (see Subsection 9.2).

[23]As a robustness check, we also predict our outcome of interest for the full population of unemployment entries in the five employment agencies 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 based on these samples is similar to the performance in the survey sample (see Figure A1 in Appendix A).

[24]This is a key difference with some recent studies on inflation forecasting. There, every observed inflation value used for training a machine learning algorithm can be related to a corresponding prediction from surveys among by professional forecasters (as well as to other determinants). See e.g. Schnorrenberger, Schmidt, and Valle Moura (2024), Kohlscheen (2022), Lenza, Moutachaker, and Paredes (2023) and Boaretto and Medeiro (2023).

substantially more training observations than in the survey sample are needed to achieve well-performing algorithms (see Subsection 9.2). We do use self-assessments in combination with machine learning algorithms in omnibus prediction models below. This should shed light on whether self-assessments enhance the predictive power of the algorithms.

5 Performance measures

To compare 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). In a sample, these are defined as follows:

$$\begin{aligned} \text{Accuracy} &= (\text{TP} + \text{TN}) \,/\, (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \\ \text{TPR} &= \text{TP} \,/\, (\text{TP} + \text{FN}) \\ \text{FPR} &= \text{FP} \,/\, (\text{FP} + \text{TN}) \end{aligned}$$

with TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives, and where TP+TN+FP+FN equals the sample size. TPR is also known as "sensitivity" while 1-FPR is known as "specificity". In our application, persons are classified as positive if they find a job within six months of unemployment and as negative if they don't.

"Accuracy" measures the fraction of individuals that is classified correctly. The TPR measures the fraction of true positives among all positives. TPR is particularly relevant if the interest is in preventing that an individual who takes up a job quickly was actually expected to become long-term unemployed. This interest may be driven by the costs of programs to prepare the individual for long-term unemployment. The FPR measures the fraction of false positive classifications among all negative observations. FPR is relevant if the interest is in preventing that an individual who does become long-term unemployed was actually expected to take up work quickly. This interest may be driven by the costs of dealing with the individual not being adequately prepared for long-term unemployment. Notice that caseworkers and employment agencies may have objectives that differ from the above measures. For example, they may be particularly concerned about budgetary implications of specific misclassifications. In that case, the above measures serve as inputs for further

deliberations.[25]

As seen in Table 2, the fraction of individuals in the sample who actually obtain a job within six months after the interview is 53%. Thus, a dumb predictor that predicts that every individual obtains a job within 6 months achieves an accuracy of 0.53. This provides a lower bound for the accuracy of other predictors. At the upper end, it is not realistic to expect accuracy values close to 1, for the following reason. Labor market frictions create inherent randomness in the matching of workers and firms. If individual characteristics and circumstances are highly favorable or highly unfavorable then this randomness is unlikely to dominate the actual outcome in an unpredictable way. But if individual characteristics and circumstances lead to an expected duration of approximately 6 months then the issue of whether the actual outcome exceeds 6 months becomes a lottery and will depend on the amount of time that the employer needs to prepare the paperwork for the job, etc.. If a substantial number of individuals leave around 6 months then this inevitably reduces the accuracy of machine learning predictors.

Measures to evaluate the performance of prediction methods can be related to measures of concordance. For example, in a binary setting, Kendall's tau can be shown to equal

$$\text{tau} = 2 \cdot \text{Accuracy} - 1 \tag{1}$$

Along this line, we may alternatively compare the predictive power of the three predictors by estimating separate linear probability models where we regress our binary outcome on the individual predictors and compare which of the predictors achieves the highest explanatory power. Furthermore, when investigating whether a combination of the three predictors improves predictions, we may regress outcomes on such combinations.

Note that the values of "accuracy", TPR, and FPR vary with the value of the classification threshold used to distinguish between fast and slow re-employment (recall the discussion about the threshold c in Subsection 4.1). To capture overall model performance irrespective of the classification threshold, the ROC-AUC Score

[25]See Black, Smith, Plesca, and Plourde (2002) for an early systematic application of such performance measures for comparisons of predictors of reemployment indicators, as well as an overview of the earlier literature. An early study of predictors of RE6 among unemployed in a European setting is in Bennmarker, Carling, and Forslund (2007). The predictors evaluated in these studies are obtained from parametric model estimates.

is a popular performance measure for classification tasks.[26] This measure can be calculated for the continuously distributed self-reported assessment probabilities that underlie the individual binary self-assessment indicator. It can also be calculated for the continuously distributed individual averages of classifiers used in the random forest. However, it cannot be calculated for the caseworker assessment, so we cannot use it to jointly rank the three predictors.

6 Baseline results

This section examines the performance of the three binary predictors separately, where for the random forest classifier we consider versions with and without the caseworker assessment among the algorithm training input. We start by applying a threshold value of 0.5 to the self-assessments as well as to the random forest predictions, when classifying individuals. Later sections examine combinations of binary predictors as well as the direct usage of the continuously distributed predictors from the self-assessments and the random forest.

Some preliminary aggregate descriptives. The random forest classifiers without and with caseworker assessments predict that roughly 67% / 66% of persons obtain a job within six months after the interview (Figure 1). As already seen in Table 1, the job seekers themselves are on average more optimistic: almost 80% of the survey participants predict that they obtain a job within six months. The caseworker assessments predict that only around 53% of the survey participants obtain a job within six months. Finally, as seen in Table 2, the fraction of individuals who actually obtained 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 who find a job within six months.

Next, we report the degree to which the different methods classify the same individuals as either positive or negative. We start by examining to what extent the predictors display independent variation. Table A3 presents their correlations. Clearly, these are rather small with the exception of the two predictors based on random forests. Table 3 presents the similarity in terms of prediction. The self-assessment and the machine learning approaches classify roughly 70% of the

[26] ROC-AUC stands for "Receiver Operating Characteristic - Area Under the Curve". The ROC curve plots the fraction of positive outcomes correctly identified (TPR) against the fraction of negative outcomes incorrectly identified as positive (FPR), displaying how these measures change if the classification threshold is varied. The ROC-AUC Score ranges from 0 - 1, with higher values indicating better prediction performance. A ROC-AUC Score of 0.5 is as good as a random guess, whereas a ROC-AUC Score of 1.0 indicates perfect prediction.

jobseekers identically. The machine learning algorithm without the caseworker information classifies individuals into the same category as the caseworker in roughly 60% of cases. Including the caseworker assessment in the machine learning algorithm increases the similarity of both approaches, as the number of identical classifications increases to about 70%. The two machine learning approaches classify a large share of people identically (almost 90%). Finally, caseworker- and self-assessment agree on the classification in close to 60% of the cases. Below, when discussing combinations of predictors, we consider characteristics of individuals for whom self-assessments and random forest predictors differ.

The performance of each of the three predictors. The above aggregate numbers do not reveal the share of correct predictions for each method. For this, we consider the "Accuracy" and we also consider the TPR and FPR as performance measures. Table 4 provides estimates of these three measures for each predictor. It also displays the corresponding standard errors obtained via bootstrapping (see also Figure 2).[27] Pairwise statistical tests of equality of the accuracy, TPR and FPR across predictors are also carried out by way of bootstrapping. Figures 3, 4 and 5 show the confidence intervals of the differences between the predictors.

In terms of accuracy, the random forest algorithm outperforms self-assessment and caseworker assessment. The random forest model without (with) caseworker assessment achieves an accuracy of 64.7% (64.6%), followed by self-assessment (63.0%), and caseworker assessment (59.7%). The estimated difference in accuracy between the random forest and the caseworker assessment is as large as 5pp. Figure 3 shows that this is significantly different from zero at the 5% level. Likewise, the self-assessment performs significantly better than the caseworker assessment. Other pairwise comparisons lead to p-values slightly above 0.05.

Regarding the range of possible accuracy values, it is useful to keep in mind that that a "dumb" predictor already achieves a value as high as 53%, whereas labor market frictions make values as high as 80% highly unlikely. It is also useful to point out that a predictor based on a simple parametric logit model with 85 covariates achieves an accuracy of 61% in our data (see below), which strictly exceeds the caseworker assessment accuracy.

[27] In a first step, we draw 500 bootstrap samples from the administrative data in 2011/12 and train a random forest classifier on each of the bootstrap samples separately. In a second step, we draw 500 bootstrap samples from the survey data. In a third step, we predict the outcome of interest for 500 unique bootstrapped random forest - survey data combinations and calculate the relevant performance measures for all bootstrap samples. We use this approach for the random forest accuracy, and we also use bootstrapping for the other accuracies and the differences between the various accuracies.

For certain objectives of policymakers and labor market administrators, the TPR and FPR may be interesting as well. The raw data showed that on average, clients are the most optimistic[28] and caseworkers the most pessimistic. This drives the rankings of TPR and FPR values across predictors.

As an alternative representation of the performance of the predictors, we estimate linear probability models that each include just one of our predictors (see Table 5). With 0.084 (0.082), the random forest without (with) caseworker assessment information has the highest R^2, followed by self-assessment, which achieves a R^2 of 0.076, outperforming caseworker assessment which achieves an R^2 of 0.036.

Subgroup analyses. The article by d'Haultfoeuille, Gaillac, and Maurel (2021) shows that the extent to which beliefs about future earnings are aligned with rational expectations differs across subgroups of individuals, notably across levels of education. With this in mind, we separate the predictions by education and also by gender and age group. Figures 6, 7 and 8 display mean prediction levels and predictor accuracy by subgroup.[29]

Looking at mean prediction levels, based again on thresholds of 0.5, we observe that job seekers are on average over-optimistic in each subgroup. For example, the mean self-reported value among younger unemployed is 88% whereas the actual mean is 58%. For the older unemployed these numbers are 71% and 48%, respectively. We also observe that females are on average less over-optimistic than males. Such a gender difference has been documented in other studies of comparisons of individual predictions and realizations of economically relevant events; see e.g. Bjuggren and Elert (2019).

Turning to "accuracy", the finding that caseworkers are worse at predicting RE6 than clients or random forest models holds for all subgroups. All three predictors perform better for older than for younger workers, and better for vocationally

[28]It is unlikely that over-optimism in clients' predictions is driven by a perception that labor market conditions were going to drastically improve in late 2012 and early 2013. The Eurozone debt crisis was slowly coming to an end, but the German economic and labor market performances had been insensitive to this. The ZEW leading business-cycle index for Germany turned positive in this period but the actual range of index values attained was not extraordinary from a historical point of view. Macro-economic statistics on the economy and the labor market (such as the labor force and the unemployment rate) did not change markedly. Note also that if changing conditions provided a rationale for optimism, then this should be reflected in caseworker assessments as well, which it did not. Finally, if macro expectations do play a role in self-assessments only, their effect may be captured in later analyses that allow an over-optimism correction parameter c.

[29]Findings by type of education (Figure 8) should be interpreted with some caution because of peculiarities of the German educational and vocational system and the ensuing educational classification. To maintain reasonable sample sizes we divide the total sample into two groups that effectively comprise workers who primarily followed a vocational track (high or low) and workers who primarily followed a general (or "academic") track.

educated workers than for more generally educated workers. The poorer performance of caseworker predictions is most evident in the latter group. The best performance of the random forest classifier is achieved for older workers, for males, and for vocationally educated workers. Interestingly, for females, self-assessments are actually more accurate than the random forest predictions. We return to this in Subsection 8.4. Table 6 provides the standard errors of the estimated key accuracies by gender.[30]

7 Combining the predictors to obtain a super-predictor

This section discusses the construction and performance of a super-predictor combining the three separate predictors. As a starting point we consider the estimation of multivariate regressions including different combinations of the three predictors (Table 7; for brevity we exclude the random forest with caseworker assessments among the training input).[31] The benchmark (first) column repeats the regression from Table 5 that only includes the random forest predictor. When adding the self-assessments, the R^2 increases from 0.084 to 0.130. This suggests that subjective expectations of the unemployed contain personal information about future events that is not reflected in the administrative data. When including the caseworker predictor as covariate in the baseline specification with only the random forest predictor, R^2 increases to 0.102.[32] This increase is much smaller than when adding the self-assessment. Finally, the most interesting column (column 4) includes all three predictors. This further increases the R^2 (to 0.140) indicating that combining all sources of information provides the largest explanatory power.[33]

Interestingly, the coefficients in column 4 are of highly unequal size. This has implications for a "super-predictor" of fast re-employment (i.e., of RE6=1, or, in

[30] Calculation of a bootstrapped standard error involves an immense computational burden of more than a week on a simple server. In the remainder of the paper we therefore use the asymptotic approximation of the standard error of Kendall's tau, invoking Equation (1).

[31] Along the lines of Granger and Ramanathan (1984)'s approach to optimal forecast combinations, we include an intercept in the regressions. The approach benefits from the absence of high collinearity between the separate predictors in the regression.

[32] This increase of R^2 may seem at odds with the fact that the inclusion of the caseworker assessment as input into the random forest does not improve the machine-learning prediction performance (recall Figure 2). However, the random forest that includes caseworker assessments as input is trained on a different, earlier and larger sample than the modestly sized sample used in the regression. In the regression, caseworker assessments of the respondents are included directly as a covariate, and it is possible that its regression effect captures individual variation not represented in the random forest algorithm.

[33] We repeat this exercise with the continuously distributed predictors (rather than the binary ones) from the random forest and the self-assessments (see below).

shorthand, of RE6) that can be based on the estimated regression. Here we adopt the natural rule that fast re-employment is predicted iff the fitted regression value exceeds 0.5. From the coefficients, it turns out that such values can only be achieved if *both* the random forest *and* the self-assessment *each* predict fast re-employment. For every other constellation of covariate values, the regression leads to the prediction of long-term unemployment. In particular, the caseworker assessment is not relevant for the super-predictor. In sum, the regression that combines all three predictors leads to the rule that RE6=1 is predicted if and only if both the random forest and the self-assessment predicted RE6=1.

It turns out that the accuracy of this super-predictor equals 66.2%. This exceeds the accuracy of each of the three separate predictors and, in particular, exceeds the accuracy of the random forest predictor. The standard error of the estimated accuracy is equal to 0.014. The self-assessments thus contain additional information on top of the information contained in the machine learning approach.[34] The added value compared to the random forest predictor necessarily comes from individuals who self-predict long-term unemployment while the machine learning algorithm predicts fast re-employment, as these are the only individuals for whom the super-predictor differs from the random forest predictor. Apparently, such individuals have information about their current personal situation that makes long-term unemployment likely although their fundamentals in terms of individual history and background characteristics are in line with fast re-employment. The size of this group is 63 (so 5.4% of the sample).

[34] A number of recent studies on inflation forecasting conclude that a combination of machine learning predictors and assessments by professional forecasters and experts is superior to each of those two predictors in isolation from each other. See, in particular, Araujo and Gaglianone (2023). The earlier-mentioned inflation studies by Schnorrenberger et al. (2024), Kohlscheen (2022), Lenza et al. (2023) and Boaretto and Medeiro (2023), which include survey-based expectations by experts into the machine learning model, find that the expert expectations improve the predictive performance of the machine learning model. Incidentally, the accuracy of the machine learning predictor with or without expert expectations is superior to other approaches. The study of firm survival in Miyakawa and Shintani (2022) echoes these findings: predictions by professional credit analysts add precision to those from random forest algorithms, but the latter in isolation do better than the former in isolation.

8 What is special about the individuals whose self-assessment can be used to improve the machine learning performance?

The super-predictor of the previous section motivates a search for specific features among the group of survey participants who correctly predict long-term unemployment if the random forest predicts the opposite. For this endeavor one might consider characteristics and outcomes in the register but in general these will not explain why the self-assessment is more informative or different, because that would already be captured by the random forest algorithm. To proceed, we consider two sources of potential additional information among the survey respondents:

(i) Administrative data on events occurring in the 6 months after the interview.

(ii) Personality traits that are self-reported in the survey.

8.1 Anticipated sickness

The occurrence of certain types of sickness may be orthogonal to personal characteristics and individual labor market histories but it may be anticipated by the individual. For example, the individual may have an upcoming appointment for cancer surgery involving hospitalization. Such future sickness absences will hamper a swift subsequent re-employment and the individual may include this information when providing the self-assessment of RE6. The latter will then tend to RE6=0 but if the sickness occurrence cannot be predicted by the variables underlying the random forest then the latter is more likely to predict RE6=1.

To proceed, we consider the subsample for whom the random forest incorrectly predicts RE6=1, and we assess whether those who themselves correctly predicted RE6=0 are systematically more likely to have experienced sickness absence in the relevant 6-month interval.

Information on sickness is obtained from the administrative data. Specifically, sickness that is deemed to hamper active job search in unemployment must be notified to the employment agency, and, thanks to this, our data provide sickness absence spells in the 6 months after the interview to the extent that the sickness took place while unemployed. The aforementioned subsample (with the random forest incorrectly predicting RE6=1) consists of 285 respondents. These all have RE6=0 so we observe their sickness for the full 6-month interval. Of the 285 respondents, 57 (i.e., 20%) experienced at least one recorded sickness absence spell, while 63 (i.e.,

22%) correctly predicted RE6=0. We apply regressions with as outcome variable the binary indicator of whether the self-assessment correctly predicted RE6=0 or not, and with as key covariate a binary indicator of sickness absence. Certain types of sickness absence might be captured by the random forest algorithm and may therefore not lead to a difference between the random forest predictor and the self-assessment.. For example, chronic diseases are likely to be reflected in individual labor market histories. We therefore control for the value of the continuously distributed random forest prediction (classification fraction) in the regression.

We find that, among those with RE6=0, sickness absence in the relevant first 6 months of unemployment is not related to whether the individual correctly predicted RE6=0. The regression coefficient for sickness absence equals 0.056 with standard error 0.064 and R^2=0.007. (If one does not control for the continuously distributed random forest fraction then these numbers are 0.053, 0.064 and 0.003, respectively.)

8.2 Anticipated recall to previous job

Regarding (i), the only other potentially useful variable that we were able to identify and observe in the administrative data is whether the newly unemployed in 2012 use a "job recall" option to return to the former employer, or more accurately, to the job at the establishment that the individual left upon becoming unemployed. If the unemployed know they can return at a future date then this may affect their assessment of the duration out of work. In particular, if the individual knows or learns in the first month of unemployment that the return option can be exerted but not within 6 months then this may motivate waiting until that moment of return, and hence to report RE6=0, which will then be confirmed by the realized RE6 value. If this is an idiosyncratic opportunity (e.g. if recalls usually take place within a few months or never at all) then RE6=0 will not be reflected in the random forest prediction.

The full sample contains 145 individuals who leave unemployment by way of a job recall. Following the approach regarding sickness spells, we may consider the subsample for whom the random forest incorrectly predicts RE6=1 and assess whether those who correctly predicted RE6=0 systematically more often return to their previous job. As it happens, not a single individual in this subsample used a job recall. Hence, the above hypothesis cannot be tested, and the underlying mechanism, if it exists at all, is not empirically relevant to explain the relevance of the self-assessment for predicting RE6=0.

In the subsample for whom the random forest incorrectly predicts RE6=1, job recalls may also affect self-assessments in a different way. If the individual is used to

receiving recalls within the first 6 months but unexpectedly does not receive a recall in the current spell, and if the respondent knows this at the survey date, then the individual may conclude that RE6=1 is much less likely than before and hence may predict RE6=0. In contrast, the random forest, incorporating the patterns in labor market histories, would lead to the prediction that RE6=1. Such a scenario would be in agreement to the observed absence of recalls among those with RE6=0. We cannot investigate this any further because our dataset does not contain information on the extent to which previous job spells were initiated by a recall.

8.3 Risk aversion

The second source (ii) of potential additional information consists of a small number of self-reported personality traits.[35] Much of the variation in these traits and their role in predicting long-term unemployment should be captured by the random forest.[36] Personality traits are often viewed as time-invariant characteristics among adults. So, for example, risk-averse individuals may have lower reservation wages in every past unemployment spell as well as in their current spell (Mortensen, 1986). However, the extent of risk aversion carries an additional relevance in connection to uncertain events in the future. Consider possible idiosyncratic events in the near future that cannot be explained or predicted by observed covariates and individual histories. The likelihood with which they occur, and the extent to which they affect life and, in particular, unemployment durations, may be gauged to some extent by the individuals themselves, by collecting information about the events and their determinants. Risk-averse individuals tend to attach a higher value to such information, and hence they have a higher incentive to acquire it and pay for this in utility terms. As a result, they will be better informed about such idiosyncratic events affecting their RE6 than risk-loving individuals. The relevance of the self-assessment in the super-predictor may then be larger among risk-averse individuals.

In other words, if we find an effect of risk aversion on the relevance of the self-assessment vis-à-vis the machine-learning prediction then this lends credence to the notion that self-assessments reflect private knowledge about upcoming events that is not captured in the algorithm. Notice that this does not presume that risk-averse

[35] These are quantified in the usual way. Risk aversion and patience are self-reported on scales from 0 to 10. The locus of control variable counts self-assessments of 8 statements, each ranging from 0 to 7. Because of the ordinal nature of these measures, our primary focus is on the sign and statistical significance of effects on predictions rather than their magnitude.

[36] Indeed, Bodilsen et al. (2025) recommend to collect personality traits and attitudes among unemployed individuals and to add these into machine learning algorithms for unemployment outcomes.

individuals face more idiosyncratic shocks or events in the near future. Rather, they may be better aware of the occurrence and prevalence of such shocks and events and therefore they may be better able to predict their current RE6 than risk-loving individuals or than a machine-learning algorithm.

In our context we are particularly interested in the role of those idiosyncratic events that prolong unemployment, so that lead to RE6=0, since this is where the self-assessments lead to a gain in prediction performance. Of course, risk averse individuals may also know more about events within or beyond the 6-month interval that would lead to a swifter re-employment and hence to a correct self-assessment of RE6=1. However, such cases do not contribute to the superior performance of the super predictor. Events beyond the 6-month interval are further into the future and thus the acquisition of information on their occurrence is more costly, making it is less likely that risk-averse individuals in response end up with RE6=1 if the random forest predicts RE6=0.

An important caveat is that it is not universally true that the value of information about future events is *uniformly* larger for risk-averse individuals. The theoretical literature provides some counter-examples; however, in many cases the notion that risk aversion increases the value of information about uncertain events in the future is correct (see e.g. the overview study by Willinger, 1989). Risk averse individuals then have a larger incentive to acquire information about future events.

In the survey data, the risk-aversion measure is based on the response to the following question:

"How do you assess yourself personally: on a scale from 0 to 10, are you in general a risk-taking person or do you try to avoid risks? Here, 0 means that you are absolutely NOT risk-taking and 10 that you are very much willing to take risks.

This is recorded for 1151 respondents. For expositional reasons we use 10 minus the reported value, so that a high value indicates more risk aversion. In the sample, its mean is 4.1 and the three quartiles are 2, 4 and 5, respectively. For robustness reasons we focus on a binary indicator (demarcating ≤ 4 from >4).

To proceed, we consider again the subsample for whom the random forest incorrectly predicts RE6=1. We apply regressions with as outcome variable an indicator of whether the self-assessment correctly predicts RE6=0 or not, and with, as key covariate, the measure of risk aversion. We also include the value of the continuously distributed random forest prediction (classification fraction) as covariate although we do not see a compelling reason for this to affect the outcome among those for whom the random forest prediction is incorrect.

Column (4) of Table 8 provides the regression results. Clearly, being risk averse has a strongly significant and substantially positive effect on the likelihood of correctly self-assessing that there will be no transition into employment in the upcoming 6 months. Thus, risk averse individuals provide particularly often valuable information related to future events if machine learning predicts fast reemployment. This finding also applies if the risk-aversion measure ranging from 0 to 10 is used and/or if the continuously distributed random forest prediction is omitted as covariate (see Columns (1)-(3)). The coefficient of the latter is always insignificant in these regressions.[37] This suggests that the information value added by self-assessments (on top of machine learning algorithms) is particularly high in the case of risk-averse individuals, if the upcoming 6 months contain idiosyncratic events that can not be predicted by extrapolation of past characteristics and outcomes.[38]

8.4 The role of gender in the effect of risk aversion

There is widespread evidence that women are on average more risk averse than men (see e.g. Eckel and Grossman (2008) and Filippin (2024) for overviews). Hence, it may be interesting to see if the results of the previous subsection are primarily driven by female individuals. An additional reason to consider results by gender is that women are less over-optimistic than men (see e.g. Dawson (2023) for an overview and Bjuggren and Elert (2019) for more specific examples in labor economics; recall also the finding in Section 6 for our own data) so that their distribution of self-assessments is less skewed towards the upper bound 1 of the support and displays more variation, providing more opportunities to detect the dependence on the individual risk aversion. By contrast, as many as 84% of the males in the sample assign probabilities exceeding 50% to the outcome that RE6=1 though in reality only 57% of them end up experiencing RE6=1. To put it bluntly, male self-assessments may be unresponsive to risk aversion because of their heavy over-optimism.

[37] The other two personality traits proxied with survey responses (patience and locus of control) do not contain explanatory power for the self-assessment over and above the random forest, in the subsample for whom RE6=0 while the random forest predicts RE6=1. The same applies to the time distance between the survey interview date and the date of the caseworker assessment. One could argue that a large difference in time means that at the interview date the individual has had more opportunities to get accommodated to unemployment and to learn about upcoming events. The mean and standard deviation of the time interval length are 24 and 6 calendar days, respectively. We do not find significant effects of this length on any of the differences between the predictors.

[38] The focus of our study is on forecasting and hence we do not claim to empirically identify a causal effect of risk aversion as a personality trait on the acquisition of information. A causal approach would need to address possible systematic variation in e.g. the cost of acquisition across the degree of risk aversion.

The subsample of 285 in the previous Subsection 8.3 is now divided into 122 females and 163 males. For females, the key coefficient, corresponding to the high-risk-aversion coefficient in Column (4) of Table 8 is 0.229 (standard error 0.083, p-value 0.007). For males, it is 0.078 (standard error 0.058, p-value 0.134). Hence, the results in Subsection 8.3 regarding risk aversion are driven by the women in the sample.[39] The estimated effect size for females is substantial. Among those for whom the random forest incorrectly predicts RE6=1, going from low to high risk aversion increases the probability of a correct self-assessment with around 23 percentage points, from around 0.2 to around 0.43. The finding is in line with the two explanations above (females being more risk averse and less over-optimistic, on average). Another possible explanation is that the females in the sample encounter more idiosyncratic events in the near future than males, so that their degree of risk aversion is more relevant than among males. Examples are childbirth or the onset of pregnancy, of the adult woman or of a close family member. Unfortunately, the data do not contain information on this.

8.5 General discrepancies between individual self-assessment and machine learning prediction

The super-predictor uses self-assessments of RE6=0 to enhance the random forest predictor if the latter states that RE6=1. There is a second possibility for discordance between the values of the two predictors, namely when the self-assessment is RE6=1 and the random forest predicts RE6=0. The latter constellation does not affect the super-predictor. Intuitively, a self-assessment of RE6=1 is exceedingly common due to the widespread over-optimism of workers, so that it is not advantageous to use such a self-assessment to overrule the random forest prediction that RE6=0. Nevertheless, it may be interesting to incorporate such cases in a study of discrepancies between the individual self-assessment and the random forest predictor in general, if the random forest prediction is incorrect, regardless of whether the discrepancy plays a role in the super-predictor. For this we consider the full group of individuals for whom the random forest prediction is incorrect and examine whether the subgroup with a correct self-assessment is systematically different from its counterpart.

This extends the analysis of Subsections 8.1-8.4 because it also includes some individuals whose correct self-assessment did not lead to a correction of the random forest predictor in our super-predictor. The size of the subsample is now 401, so

[39]The other coefficients by gender have similar features as in Table 8 and fully confirm this finding.

compared to previous subsections the subsample increases by 116 individuals. These 116 individuals all have realized RE6=1 while their random forest prediction is that RE6=0. Importantly, among these, a total of 93% (so 108 individuals) make the correct self-assessment of RE6=1. (To avoid confusion, note that this does not necessarily suggest that a super-predictor should always assign RE6=1 if the random forest predictor is RE6=0 and the self-assessment is RE6=1, because we also observe many individuals whose self-assessment incorrectly states RE6=1 if the random forest predicts RE6=0.) Because of the ubiquity of the self-assessment that RE6=1 among these 116 individuals, one may perhaps not expect these 116 individuals to provide much additional insight into specificities of the group that improves the random forest by their self-assessment.

We estimate regressions with as outcome variable an indicator of whether the self-assessment is correct or not, using the subsample of individuals for whom the random forest prediction is incorrect, and again including as a covariate the value of the continuously distributed random forest prediction (classification fraction). Regarding the role of sickness absence, the implementation of this idea is hampered by the fact that the administrative data at our disposal only record sickness absence during unemployment (to a sufficient degree of temporal granularity). Thus, early exits to work mechanically reduce the probability of observing sickness absence, creating an insurmountable selection problem among those with realized RE6=1.

Next, we consider the role of risk aversion. The results (see Table 9, with Column (4) containing the main results) are qualitatively identical to those in the previous subsections. Being risk averse has a strongly significant and substantially positive effect on the likelihood of correctly self-assessing the individual realization of RE6. Like above, this may be explained by differential incentives for gathering information. The results regarding risk aversion are driven by the cases with RE6=0 (so the cases that were analyzed in the previous subsections) and are again driven by female rather than male individuals.[40]

[40]Regarding the roles of recalls, patience, locus of control, and the time distance between the first caseworker meeting and the survey interview, the results reported in Subsection 8.2 and in footnote 37 continue to hold.

9 Generalizations, alternative performance measures, robustness checks and further analyses

9.1 More general classification schemes and modified predictors

Shifting the classification thresholds for binary predictors. Results presented so far are based on a specific threshold value of 50% for the dichotomous classification of self-assessments and random forest predictions. For prediction purposes, this 50% value is not compelling. Consider, first, the random forest predictor. Its prediction performance is likely to be be better in case of certain different thresholds. According to Figure 9, the highest accuracy (65.5%) is attained for threshold values slightly different from 50%. Incidentally, note that for a wide range of possible thresholds, the random forest model that includes caseworker assessments outperforms the random forest model without that information. This is confirmed by the ROC-AUC score (see Figure A2 in Appendix A).

Next, consider shifting the threshold value for the self-assessments. In Subsection 4.1 we proposed a method to correct for over-optimism by using a threshold such that the predicted fraction of "negatives" equals the observed fraction. In our data this leads to a threshold value $c = 0.87$ (see Tables 1 and 2). Figure 9 shows that this leads to an accuracy that is virtually equal to the highest attained accuracy of 64.8% across all thresholds for this predictor, exceeding the 63% value in the case of a 0.50 threshold. Whether the value of $c = 0.87$ has external validity is not known. In other settings it may provide a valuable correction for over-optimism.[41,42]

TPR and FPR values when shifting the thresholds for the binary predictors. Figure 10 displays the TPR and FPR for each predictor across all relevant thresholds. As the caseworker assessment is captured in a single binary indicator, its resulting TPR and FPR are by construction fixed across all thresholds. We use those caseworker TPR and FPR values as benchmark for comparisons.

[41] We may also use the corrected self-assessment predictor as a component for the super-predictor. The regression coefficients are close to those in column 4 of Table 7. The accuracy is virtually indistinguishable from the value for the original super-predictor (66.4% versus 66.2%).

[42] Figure 9 enables a comparison of the predictors in terms of highest achieved accuracy, where the thresholds may differ by predictor. It turns out that the ranking is the same as for the threshold of 0.5: the highest maximum accuracy for the random forest model without information on caseworker assessment (65.5% for a threshold of 0.54) is followed by those for random forest with caseworker assessment (64.9% for a threshold of 0.49) and self-assessment (64.8% for a threshold between 0.75 and 0.79) and caseworker assessment (59.7%).

Holding the FPR fixed at the level calculated based on the caseworker assessment (FPR=0.426), we find that the random-forest classifier and the self-assessed predictions exhibit a higher TPR, with the following ranking: caseworker assessment (TPR=0.603) < random forest with caseworker assessment (TPR =0.703) < random forest without caseworker assessment (TPR=0.719). The TPR for self-assessment is between 0.645 and 0.737 and is thus clearly superior to the TPR of caseworker assessment. (For the assessment by the survey participants, there is no FPR that exactly matches the FPR of the caseworkers, and the same applies to the TPR.) This exercise can also be performed the other way around: at the caseworker TPR level of 0.617, the FPR of the random forest classifier without/with caseworker assessment is considerably lower (FPR=0.338/0.339) than the caseworker FPR (=0.426). The same is true for the self-assessment (with a corresponding FPR between 0.347 and 0.358).

From this perspective, it appears that the machine learning predictor provides a reasonable balance between a high TPR and a low FPR. Furthermore, caseworker predictions are outperformed by the random forest classifiers and the self-assessment. This, however, does not necessarily hold over the entire range of potential thresholds. If the main objective is to attain a very high TPR or a very low FPR for an a priori chosen threshold, rather than a high accuracy at that threshold, then in some cases one may prefer caseworker assessments (or self-assessments) over machine learning methods.

Continuously distributed random forest predictor and self-assessment. The super-predictor is motivated by regressions on binary machine-learning and self-assessment predictors that are dichotomous "rounded-off" versions of the underlying classifier fractions and self-reported re-employment probabilities, respectively. The latter two are essentially continuously distributed and hence are potentially more informative. In Table A4 we regress RE6 on each of these continuously distributed predictors separately. This does indeed lead to higher R^2 than in Table 5. Hence, the continuously distributed predictors dominate their binary versions. Also, the machine learning model including caseworker assessments outperforms the one not including those assessments, in terms of R^2 (0.148 vs. 0.134).

In Table A5 we regress RE6 on the two continuously distributed predictors and the binary caseworker assessment and on subsets thereof. The resulting R^2 exceed those in Table 7. Consequently, we may construct a new super-predictor based on the final column. Its accuracy equals 67.4% which exceeds the value of 66.2% for the earlier super-predictor. The standard error of the estimated accuracy is 0.014.

This enhanced super-predictor is somewhat less transparent than the earlier version because the variables on the right-hand side of the regression span a wider range of values than before. Interestingly, this leads to a role for the caseworker predictor in specific cases. In particular, this occurs if the random forest predictor has a high classifier fraction value and the self-assessment is rather low. For example, suppose that the random forest classifier fraction for RE6 equals 0.7 and the self-assessment response is 40%. Then a caseworker assessment of RE6=0 leads to the super-predictor stating that RE6=0 whereas a caseworker assessment of RE6=1 leads to the opposite super-prediction of RE6=1. Such scenarios are relatively uncommon in the data, as the self-assessment typically exceeds the random forest prediction.

Predictor interactions in the super-predictor. We now consider a different type of extension of the super-predictor approach. The super-predictor in Section 7 is motivated by a regression that is additive in the various separate predictors. It is interesting to incorporate pairwise interactions between those predictors in this regression, since the relevance or precision of one predictor may depend on the value of another predictor. For example, the caseworker assessment might be more relevant for clients who are themselves pessimistic about their job prospects.

Table 10 reports the estimated coefficients for the regression with the three binary predictors and their interactions. The coefficients of the interaction terms are small and two of the three are insignificantly different from zero at the 10% level. The additive coefficient values are similar to those in the absence of interaction terms. The super-predictor arising from the regression is also similar to before, with one exception, namely that RE6=1 is now also predicted if both the random forest and the caseworker predict RE6=1, even if the self-assessment is that RE6=0. However, the latter finding is based on a set of positive coefficients of which one is actually insignificantly different from zero, namely the additive effect of the caseworker assessment. More importantly, the accuracy of the super-predictor does not improve with the extension, as it has the same value as in Section 7 (66.2%). Thus, the importance of interaction terms is minor.

Next, we replace the binary machine-learning and self-assessment predictors by their continuous versions. Again, the main insights do not substantially change. This lends further credence to the super-predictor version without interaction terms (where we refrain from formal tests because of intricacies due to the fact that the random forst classifier fractions are estimated rather than observed).

9.2 Robustness checks and further analyses of machine learning predictors

This subsection considers nine robustness checks and additional analyses. The first three deal with the outcome definition and the sample selection. The remaining six examine details of the implementation of the random forest.

Outcome variable. In the first check we modify the outcome definition. The main definition states that the individual has found a job within six months after the survey interview, which is not conditional on whether the person is still employed six months after the interview. As an alternative, we define the outcome as having a job exactly six months after the interview (see Figure A3). This hardly affects the results, except that with this definition, the random forest model with caseworker information has the highest accuracy.

We also explore to what extent machine learning predicts caseworker assessments as outcome. For this we build a new random forest. Figure A15 shows the results. In terms of accuracy and ROC-AUC score, this random forest does better than the predictors of RE6. However, the prediction is not perfect. Thus, caseworker assessments display variation not captured by the machine learning algorithm. Whether this variation adds to the prediction of the actual RE6 does not follow from this exercise, but the results from Subsection 7 suggests that it does not.

Sample selection. Regarding the caseworker assessments, we only included individuals with profiles that are no older than a year (with, as a result, a few profiles stemming from a previous unemployment spell in that year). We now restrict the sample by only including clients with profiles no older than six weeks (see Figure A4). The results using this restriction closely resemble the main results. However, interestingly, the accuracy of the caseworker assessments is slightly lower when only using more up-to-date profiles.

We also examine the roles of 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). Specifically, we repeat the analyses without these restrictions (see Figure A5). Except for the fact that the machine learning model not excluding caseworker information achieves the highest maximum accuracy, the results from this exercise are very close to the main results.

Details of the implementation of the random forest. To investigate whether more detailed information on individual labor market histories can improve the predictive power of the machine learning models, we additionally construct and use monthly employment, unemployment, and active labor market program participation histories going up to 25 years back, resulting in more than 1800 explanatory variables. Figures A6 and A7 show the development of the ROC-AUC score when we increase the number of variables available for prediction, for the random forests not including and including the caseworker assessment as input respectively.[43] While overall model performance increases sharply in the beginning, for the first roughly 100 variables, it does not improve much when increasing the number of variables further. Thus, including the more detailed labor market histories does not improve the predictive performance of the random forest classifier to a large degree compared to our main set of explanatory variables.[44]

Next, we consider more broadly defined sets of "features" used in the random forest. While the random forest classifier is a "black box" in terms of the predictive power of single individual background characteristics and history, it is interesting to see which sets of variables contribute to its performance. Figures A8 and A9 present the ROC-AUC scores for random forest "models" using different sets of background features as training inputs.[45] We see that, compared to the caseworker assessment (Model 0), all other models exhibit better prediction performance. The inclusion of basic sociodemographic characteristics (Model 1), short-term labor market histories (Model 2), or information on the last job (Model 3) all improves the predictive performance, to a similar degree. Adding all three sets of variables together further improves overall performance of the model (Model 4). Additionally adding long-term labor market histories (Model 5) improves model performance even further, in particular for the models that do not include caseworker information as a training input. Finally, adding extremely detailed labor market histories ranging up to 25 years back achieves a similar predictive power as the models not using these variables

[43]For this exercise we do not run any additional model tuning algorithm aside from varying the variables available for prediction, as this would lead to an extremely large increase in computation times. Consequently, the overall performance of the models is slightly worse for this exercise than for the main analyses.

[44]The results are similar for the administrative data test sample (see Figures B1 and B2).

[45]None of the models in Figure A8 contain caseworker information as a training input. All models in Figure A9 include the caseworker assessment as a training input. "Model 0" includes only the caseworker assessment; "Model 1" includes age, sex, and education variables; "Model 2" includes the labor market history of the last year before entering unemployment; "Model 3" includes information on the last job before unemployment; "Model 4" includes all variables from Models 1-3; "Model 5" includes all variables from Models 1-3 and yearly labor market histories up to seven year before entering unemployment; "Model 6" includes all variables from Model 5 and monthly labor market histories up to 25 years before entering unemployment.

(Model 6). Interestingly, these findings on the relevance of certain sets of covariates are in line with those on the conditioning set in propensity-score matching evaluation of active labor market policies; see e.g. Heckman and Smith (1999) and Lechner and Wunsch (2013).

We also investigate which training-data size is necessary in order to obtain well-performing prediction models. To that end, we trained random forest classifiers using an increasing number of observations as training inputs. The results of this exercise are presented in Figure A10. We see that the performance of the models increases for training set sizes up to 15,000. Beyond this point, the predicton models do not show systematic improvements with increasing training set sizes, they do not depend anymore on the seed value of the random forest, and the ROC-AUC-Score remains relatively stable. This indicates that we need a substantial amount of training observations in order to achieve a good prediction performance.

To investigate how the performance depends on classification methods, we repeated our prediction exercises using a gradient boosting classifier and using a logit regression classifier with our main set of 85 background variables. Figure A11 shows their overall model performances, measured by the ROC-AUC score. The overall performance of the logit regression classifier is inferior but the gradient boosting classifier performs almost as well as our main random forest classifier. We achieve the highest maximum accuracy using the random forest classifier (see Figure A12).[46]

As a next robustness check, we explore whether random forest trained with contemporaneous data do better than those using past data. To this purpose we train random forests using administrative data from 2012/13 instead of 2011/12. If this does indeed improve the performance of our models, then this may motivate the inclusion of leading economic indicators. Figure A13 shows that results in terms of accuracy across classification thresholds are fairly close to our main results.

To further investigate this issue, we also present ROC-AUC scores for a holdout sample of the administrative data and for the survey data, again using the data from 2012/13 as training input (see Figure A14). Compared to the results presented in Figure A2, the ROC-AUC scores improve, from 0.708 (0.719) to 0.724 (0.735) for the administrative data excluding (including) caseworker information and from 0.702 (0.711) to 0.712 (0.720) for the survey sample excluding (including) caseworker information. Thus there is some evidence that using contemporaneous data improves the overall predictive performance of the machine learning algorithm.

[46]These results also hold for the administrative data test sample (see Figures B3 and B4).

10 Conclusion

When ranking the three different predictors, we find that the one based on the random-forest machine learning algorithm is more accurate than the client self-assessment, while the latter is followed at some distance by the predictor based on the caseworker assessment. This is a robust finding. In this sense, the random forest wins the contest between the predictors. The estimated difference in accuracy between the random forest and the caseworker assessment is as large as 5pp which is substantial in the light of the fact that the effective range of meaningful accuracy values is a small as 20pp. These baseline findings also mostly hold for subgroups defined by gender, education or age.

The three predictors display substantial independent variation. We consider combinations in a regression model, and this leads to a regression-inspired super-predictor which dominates the three underlying predictors in terms of accuracy. The super-predictor deviates only in one specific way from the random-forest predictor: it predicts short-term unemployment if and only if both the random forest and the self-assessment predict short-term unemployment. Thus, a self-assessment of long-term unemployment overrides a random-forest prediction of short-term unemployment.

The added value of the super-predictor compared to the random-forest predictor necessarily comes from individuals who self-predict long-term unemployment while the machine learning algorithm predicts fast re-employment. These individuals have information about their current and future personal situation that makes long-term unemployment likely although their fundamentals in terms of individual history and background characteristics are in line with fast re-employment. Using the survey data we show that these individuals, who tend to think that there will be no fast transition into employment while machine learning incorrectly predicts the opposite, more often have a risk-averse personality. This association is strongly significant and its magnitude is large, notably among females. We see this as an important finding, because risk-averse agents typically attach a higher value to information on uncertain events in the future, and hence they have a higher incentive to acquire it and pay for this in utility terms. Apparently, this makes risk-averse individuals particularly well-informed about future idiosyncratic events hampering swift reemployment.

To summarize this more concisely, we find that the individuals who self-predict long-term unemployment while the machine learning algorithm predicts fast re-employment are the ones who drive the superior performance of the super-predictor. It turns out that these individuals relatively often have a personality trait that makes them particularly interested in idiosyncratic future events. The latter lends

further credence to the notion that self-assessments reflect private knowledge about upcoming events that is not captured in the algorithm.

It should be kept in mind that this does not presume that risk-averse individuals face more idiosyncratic shocks or events in the near future. Also, notice that risk averse individuals may know more about events that lead to a swifter rather than a slower reemployment, but such cases do not contribute to the superior performance of the super predictor. Events beyond the 6-month interval are further into the future and thus the acquisition of information on their occurrence is more costly, making it is less likely that risk-averse individuals in response end up with fast reemployment if the random forest predicts the opposite.

We find that the results on risk aversion are stronger for females than for males. This may be because males are more over-optimistic on average, but it may also be a result of females in the sample encountering more idiosyncratic events in the near future than males, so that their degree of risk aversion is more relevant than among males.

On a fundamental level, our analysis addresses a conception that machine learning algorithms are superior predictors of human behavior because they should embody every systematic determinant of the outcome of interest. What this conception potentially ignores is that humans are forward-looking and have every incentive to collect information about future events. To the extent that such future events are idiosyncratic and unrelated to those systematic determinants but are foreseen by the involved individuals, the self-assessments by those individuals have predictive value that goes beyond the machine learning algorithm. As seen from this angle, individual responses to the likelihood of future events and outcomes can not be uniformly dominated by algorithms, when it comes to predicting those outcomes. In our analysis we did not have rich data on idiosyncratic future events that affect future employment outcomes. Rather, we develop and adopt an ingenious approach that exploits that some individuals are more likely to collect information about future events than others. Specifically, using self-reported risk aversion as an indicator of the incentive to collect such information, we showed that the extent to which self-assessments strictly add to the predictive value of the algorithms is larger if the individual is more risk averse.

These insights apply to contemporary settings in which administrative data are used to train machine learning algorithms for the prediction of individual outcomes, in the sense that the administrative data at our disposal are more or less representative of those that are commonly used for these purposes. However, advanced data collection procedures may reveal at least some of the information that individuals

may collect about their future that is not embodied in available administrative data. It is conceivable that more comprehensive registers, including those relying on intrusive surveillance of individual behavior, are sufficiently rich to accurately predict the relevant future events, to the same extent as the individuals are able themselves. Such registers may also be able to predict the extent to which the individuals anticipate such events. As seen from this angle, it is possible that the predictive value-added of self-assessments may decline in the more distant future of the usage of AI.

Our findings show that the caseworker assessments do not play a role in the super-predictor and have a lower accuracy than the other predictors. This does not mean that caseworkers' assessments are never useful. For example, in some cases, the accuracy of the random forest predictor slightly improves when individual caseworker assessments are included as a covariate in the training of the algorithm with administrative data. Further, in an extension of the super-predictor, with continuous self-assessments and random forest scores, the caseworker assessment affects the value of the super-predictor in rare cases (although statistical significance of this contribution is not demonstrated). It should also be noted that if the main objective is to attain a very high TPR or a very low FPR instead of a high accuracy, then in some cases one may prefer caseworker assessments over the random forest prediction. More in general, caseworkers provide a range of important services that may or may not be dependent on their assessment of a client's employability, but the evaluation of this is beyond the scope of this paper.

We now turn to policy implications for employment agencies that go beyond the straightforward conclusion that machine learning algorithms provide more accurate predictions of long-term unemployment than other predictors. (Note that this motivates the provision of random forest scores to caseworkers at the intake meeting.[47]) To proceed, it is useful to start discussing the practical feasibility of the super-predictor, because self-assessments are not routinely available. Our results motivate the collection of self-assessment information, and with a sufficiently large database along these lines, this can be incorporated as input when training random forest classifier algorithms. It is not clear if questions about self-assessments should be posed by the caseworker or another representative of the employment agency, as this carries risks of desirability bias and strategic responses.

The paper provides a number of additional results and insights. First, we develop a correction on the self-assessment predictor by taking over-optimism into account.

[47]In a sample of caseworkers in Germany in 2023, Dietz, Osiander, Sirman-Winkler, and Tepe (2026) find that caseworkers oppose the mandatory use of algorithmic recommendations.

However, this only slightly improves the prediction performance. In other settings the procedure may provide a valuable correction for over-optimism in classifiers based on self-assessment. It might be an interesting topic for further research to develop this further.

Secondly, we obtain some interesting insights into the machine learning approach. In general, random forest outperforms other formal classification methods. In our setting, a modest number of around 50 individual labor market history variables suffices for the random forest. Here it is useful to include annual individual history variables dating back up to 7 years, in particular if caseworker assessments are not used in the random forest. However, there are few or no gains of increasing the temporal granularity to a monthly level and/or further increasing the historical time window up to 25 years. In terms of training data size, the predictive performance improves up to 15,000 individuals but does not improve noticeably beyond that. More recent training data provide a slightly better performance than less recent ones.

Some topics for further research arise. First, as shown by the recent literature on inflation forecasting, if the prediction horizon is very short, self-reported forecasts from experts frequently outperform machine-learning forecasts. In our context, this may suggest that self-assessments gain additional usefulness if the prediction horizon is less than 6 months, whereas for unemployment durations exceeding one year the opposite may hold. A related topic for further research concerns the extent to which the caseworker assessment aids in predicting outcomes in the tails of the unemployment duration distribution. Some support for this is given by the finding that the contribution of this assessment to the predictive power of the machine learning algorithm is larger if continuously distributed predictions from the algorithm are used. More in general, it may be interesting to identify the types of caseworkers who attain the most accurate predictions and to stimulate a transfer of their expertise towards other caseworkers. Unfortunately, the data at our disposal do not allow for the linkage of clients sharing the same caseworker, so that we cannot pursue this in the current study.

Other topics concern obvious generalizations such as using prediction algorithms that are trained on continuously distributed duration outcomes instead of binary events like RE6. Alternatively, it may be interesting to use LLM's to obtain predictions based on speech recordings of intake meetings, athough this may adversely affect the quality of communication during the meetings (and may violate confidentiality regulations). Finally, it may be interesting to collect the inputs needed to quantify how incorrect predictions translate into the size of UI agencies' budgets.

References

Araujo, D., G. Bruno, J. Marcucci, R. Schmidt, and B. Tissot (2023): "Machine learning applications in central banking," *Journal of AI, Robotics & Workplace Automation*, 2, 271–293.

Araujo, D. and W. Gaglianone (2023): "Machine learning methods for inflation forecasting in Brazil: New contenders versus classical models," *Latin American Journal of Central Banking*, 4, 100087.

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

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

Bennmarker, H., K. Carling, and A. Forslund (2007): "Vem blir långtidsarbetslös?" Discussion paper, IFAU, Uppsala.

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

Bjuggren, C. and N. Elert (2019): "Gender differences in optimism," *Applied Economics*, 51, 5160–5173.

Black, D., J. Smith, M. Plesca, and S. Plourde (2002): "Estimating the duration of unemployment insurance benefit recipiency," Research bulletin, U.S. Department of Labor, Washington, DC.

Boaretto, G. and M. Medeiro (2023): "Forecasting inflation using disaggregates and machine learning," Working paper, PUC-Rio, Rio de Janeiro.

Bodilsen, S., S. Nielsen, and M. Rosholm (2025): "Measuring employment readiness for hard-to-place individuals," *Humanities and Social Sciences Communications*, 12, 835.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Bundesregierung (2019): "Antwort auf Kleine Anfrage der Abgeordneten Sabine Zimmermann u. a. und der Fraktion DIE LINKE, betreffend "Personelle Ausstattung und Arbeitsbedingungen in den Jobcentern und Agenturen für Arbeit"," *BT-Drs. 19/10601*.

Campbell, D., A. Carruth, A. Dickerson, and F. Green (2007): "Job insecurity and wages," *The Economic Journal*, 117, 544–566.

Dawson, C. (2023): "Gender differences in optimism, loss aversion and attitudes towards risk," *British Journal of Psychology*, 114, 928–944.

Desiere, S., K. Langenbucher, and L. Struyven (2019): "Statistical profiling

in public employment services: An international comparison," *OECD Social, Employment and Migration Working Papers, No. 224*.

d'Haultfoeuille, X., C. Gaillac, and A. Maurel (2021): "Rationalizing rational expectations: Characterizations and tests," *Quantitative Economics*, 12, 817–842.

Dietz, M., C. Osiander, M. Sirman-Winkler, and M. Tepe (2026): "Biased by design? Case managers' multidimensional preferences toward the design of algorithmic decision support systems," *Public Administration Review*, 0, 1–14.

Dominitz, J. (1998): "Earnings expectations, revisions, and realizations," *The Review of Economics and Statistics*, 80, 374–388.

Eckel, C. and P. Grossman (2008): "Men, Women and Risk Aversion: Experimental Evidence," in *Handbook of Experimental Economics Results*, ed. by C. R. Plott and V. L. Smith, Amsterdam: North-Holland.

Filippin, A. (2024): "Gender differences in risk attitudes," *IZA World of Labor*.

Granger, C. and R. Ramanathan (1984): "Improved methods of combining forecasting," *Journal of Forecasting*, 3, 197–204.

Hastie, T., R. Tibshirani, and J. Friedman (2011): *The Elements of Statistical Learning 2nd edition*, Springer Science Business Media, LLC.

Heckman, J. and J. Smith (1999): "The pre-program earnings dip and the determinants of participation in a social program: implications for simple program evaluation strategies," *Economic Journal*, 108, 313–348.

Hendren, N. (2017): "Knowledge of future job loss and implications for unemployment insurance," *American Economic Review*, 107, 1778–1823.

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

Kohlscheen, E. (2022): "What does machine learning say about the drivers of inflation?" Discussion paper, BIS, Basel.

Kunaschk, M. and J. Lang (2022): "Can algorithms reliably predict long-term unemployment in times of crisis? Evidence from the COVID-19 pandemic," *IAB-Discussion Paper 08/2022*.

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

Lechner, M. and C. Wunsch (2013): "Sensitivity of matching-based program evaluations to the availability of control variables," *Labour Economics*, 21, 111–121.

Lenza, M., I. Moutachaker, and J. Paredes (2023): "Forecasting euro area inflation with machine learning models," Research Bulletin 112, ECB,

Frankfurt.

Manski, C. F. (2018): "Survey measurement of probabilistic macroeconomic expectations: progress and promise," *NBER Macroeconomics Annual*, 32, 411–471.

Miyakawa, D. and K. Shintani (2022): "Disagreement between human and machine predictions," Discussion paper, Bank for International Settlements, Basel.

Mortensen, D. (1986): "Job Search and Labor Market Analysis," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and R. Layard, Amsterdam: North-Holland.

Mueller, A. I. and J. Spinnewijn (2023a): "Expectations data, labor market, and job search," in *Handbook of Economic Expectations*, ed. by R. Bachmann, G. Topa, and W. van der Klaauw, Academic Press, 677–713.

——— (2023b): "The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection," NBER Working Paper No. 30979.

Mueller, A. I., J. Spinnewijn, and G. Topa (2021): "Job seekers' perceptions and employment prospects: heterogeneity, duration dependence, and bias," *American Economic Review*, 111, 324–63.

Mühlbauer, S. and E. Weber (2022): "Machine Learning for Labour Market Matching," *IAB-Discussion Paper 03/2022*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg (2011): "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Schnorrenberger, R., A. Schmidt, and G. Valle Moura (2024): "Harnessing Machine Learning for Real-Time Inflation Nowcasting," Discussion paper, Netherlands National Bank, Amsterdam.

Smith, V. K., D. H. Taylor, and F. A. Sloan (2001): "Longevity expectations and death: Can people predict their own demise?" *American Economic Review*, 91, 1126–1134.

Statistics of the FEA (2019): "Entries into and exits out of unemployment benefit receipt," Data Warehouse of the Statistics of the Federal Employment Agency, accessed on October 9, 2019.

Stephens, M. (2004): "Job loss expectations, realizations, and household consumption behavior," *The Review of Economics and Statistics*, 86, 253–269.

Stöhr, J. (2016): "Eingliederungsvereinbarungen in der Arbeitslosenversicherung und das Suchverhalten der Arbeitslosen: Eine empirische Analyse," Friedrich-

Alexander-Universität Erlangen-Nürnberg, Mimeo.

Stoltenberg, C. and A. Uhlendorff (2022): "Consumption Choices and Earnings Expectations: Empirical Evidence and Structural Estimation," IZA Discussion Paper 15443.

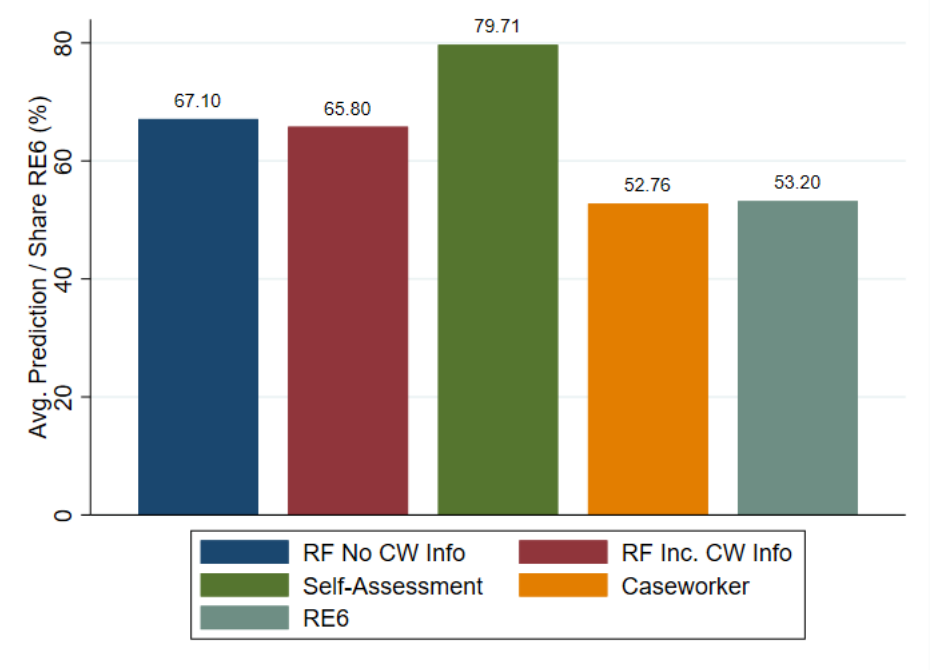
van den Berg, G. J., B. Hofmann, G. Stephan, and A. Uhlendorff (2014): "Was Vermittlungsfachkräfte von Eingliederungsvereinbarungen halten: Befragungsergebnisse aus einem Modellprojekt," *IAB-Forschungsbericht 11/2014*.

van Landeghem, B., S. Desiere, and L. Struyven (2021): "Statistical profiling of unemployed jobseekers," *IZA World of Labor*.

Willinger, M. (1989): "Risk aversion and the value of information," *Journal of Risk and Insurance*, 56, 320–328.

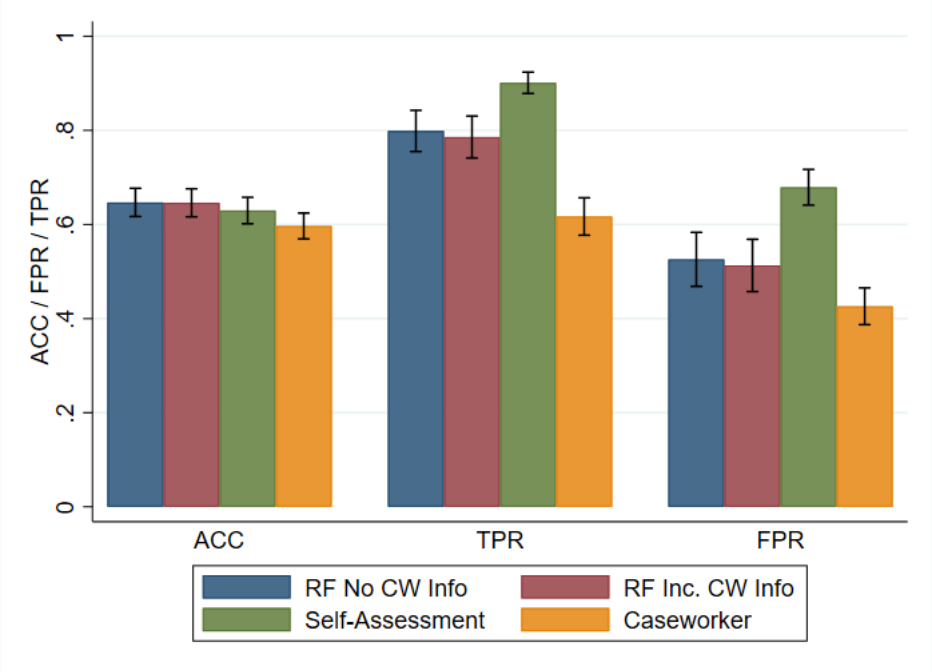
Figures and Tables

Figure 1: Average and predicted shares of RE6 at a threshold of 0.5



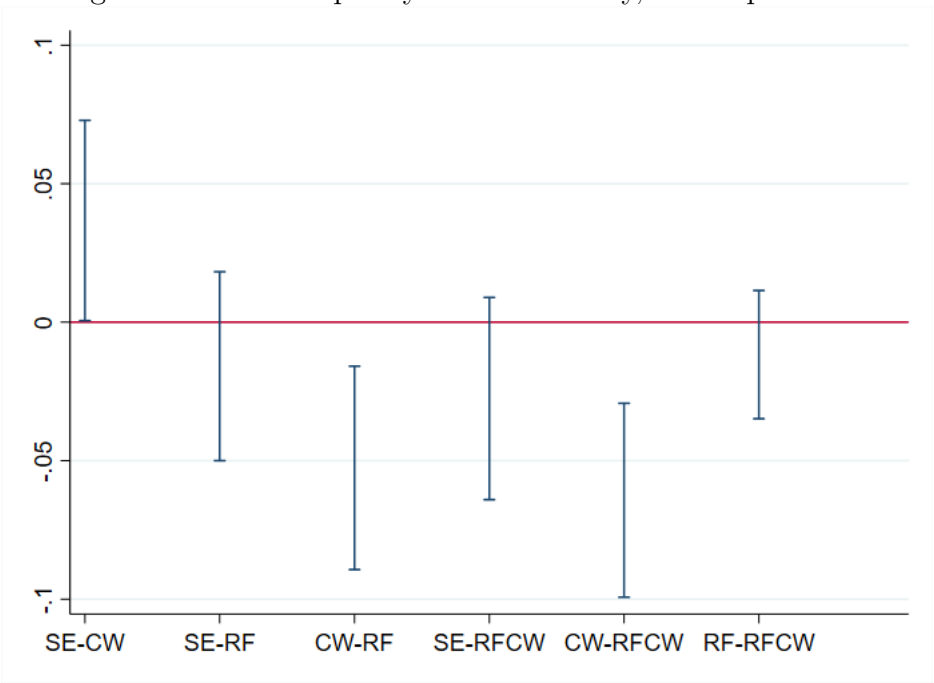
Note: This figure shows, for the survey sample, the average predicted share finding a job for each of the four predictors and the actual share of individuals finding a job within 6 months. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment, RE6 = actual share that found a job within six months. Source: EVA; IEB v.12.01.00.

Figure 2: Accuracy, TPR, and FPR of the four predictors at a threshold of 0.5



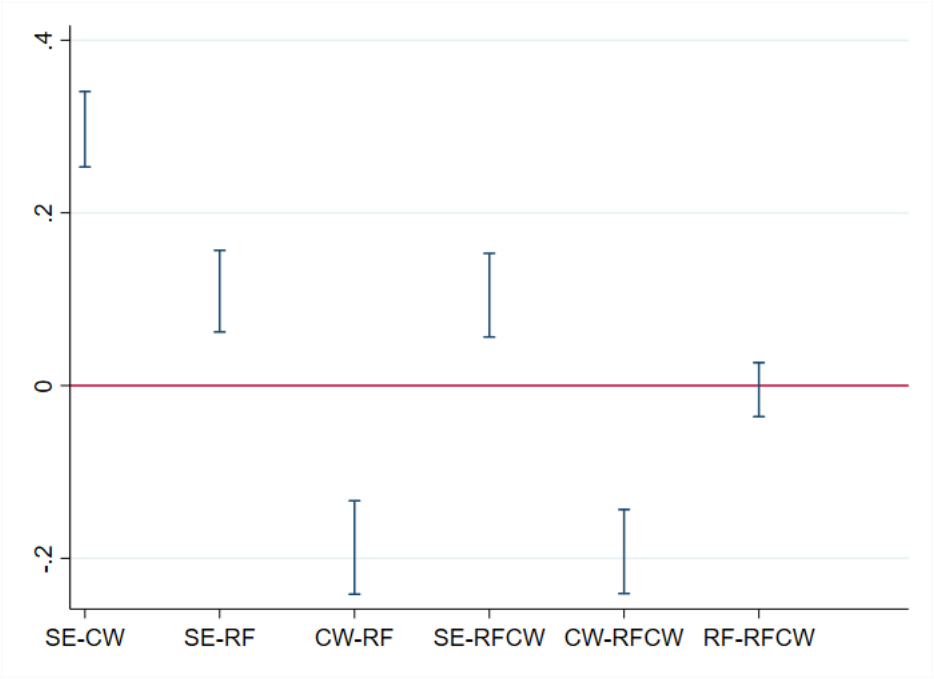
Note: This figure shows the accuracy (ACC), False Positive Rate (FPR), and True Positive Rate (TPR) for the survey sample. 95% confidence intervals are obtained via bootstrapping (500 bootstrap samples of the administrative data from 2011-12 to train the random forest classifiers and 500 bootstrap samples of the survey data for the predictions). RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Figure 3: Tests of equality of the accuracy, across predictors



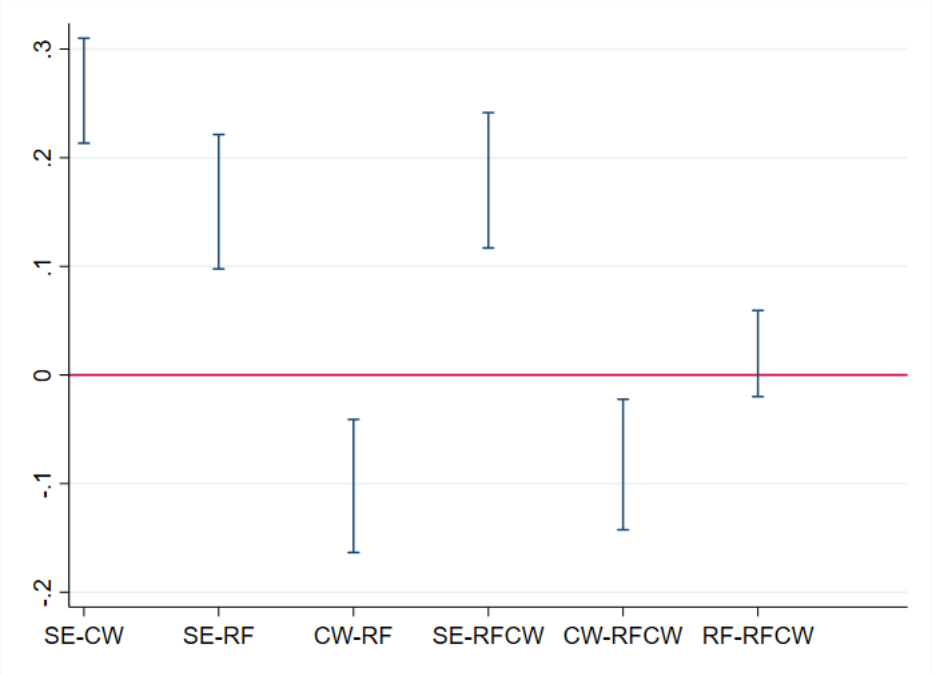
Note: 95% confidence intervals, obtained via bootstrapping. SE = self-assessment. CW = caseworker assessment. RF = random forest trained without past caseworker assessments, RFCW = random forest trained incl. past caseworker assessments.

Figure 4: Tests of equality of the TPR, across predictors



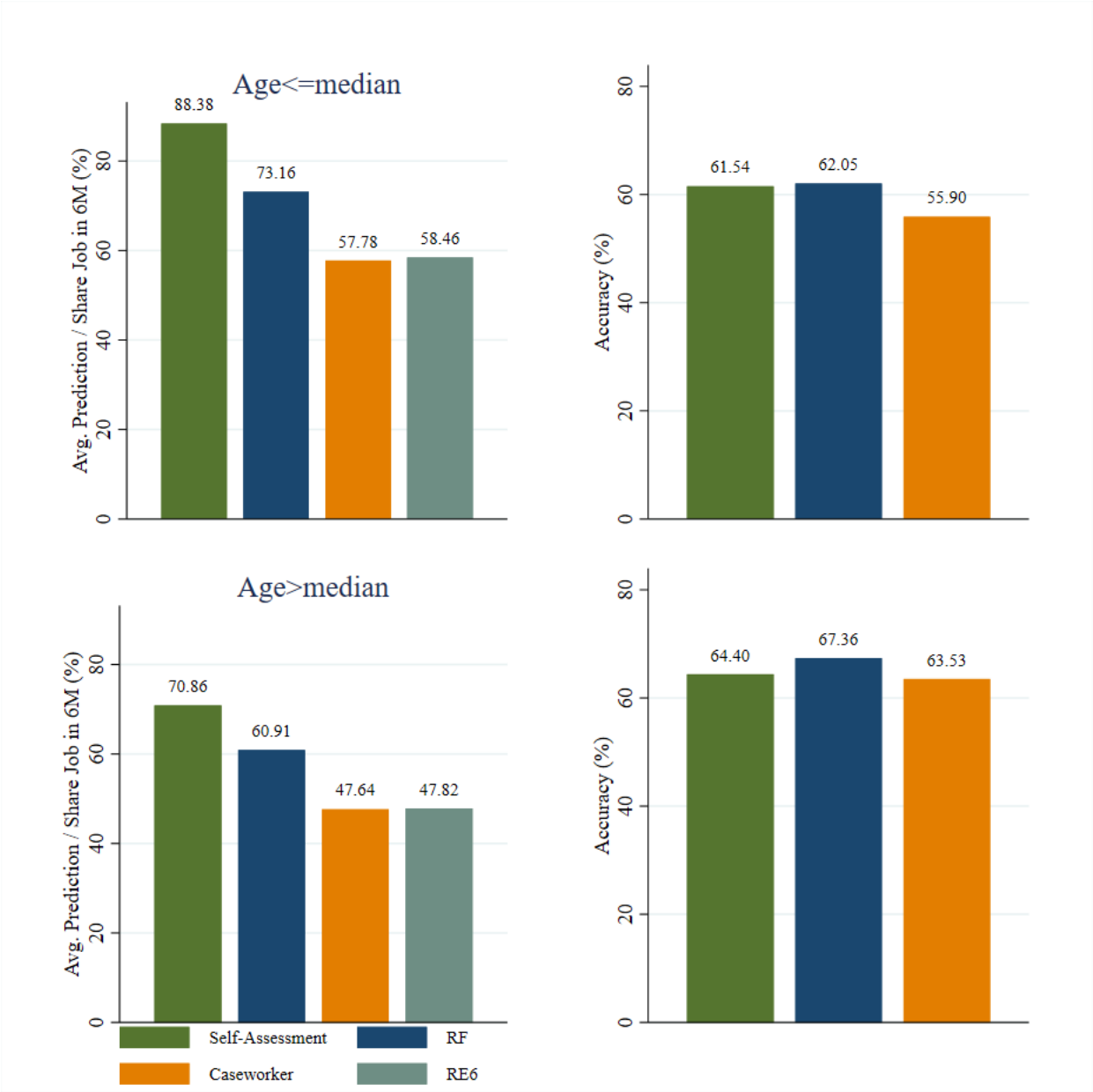
Note: 95% confidence intervals, obtained via bootstrapping. SE = self-assessment. CW = caseworker assessment. RF = random forest trained without past caseworker assessments, RFCW = random forest trained incl. past caseworker assessments.

Figure 5: Tests of equality of the FPR, across predictors



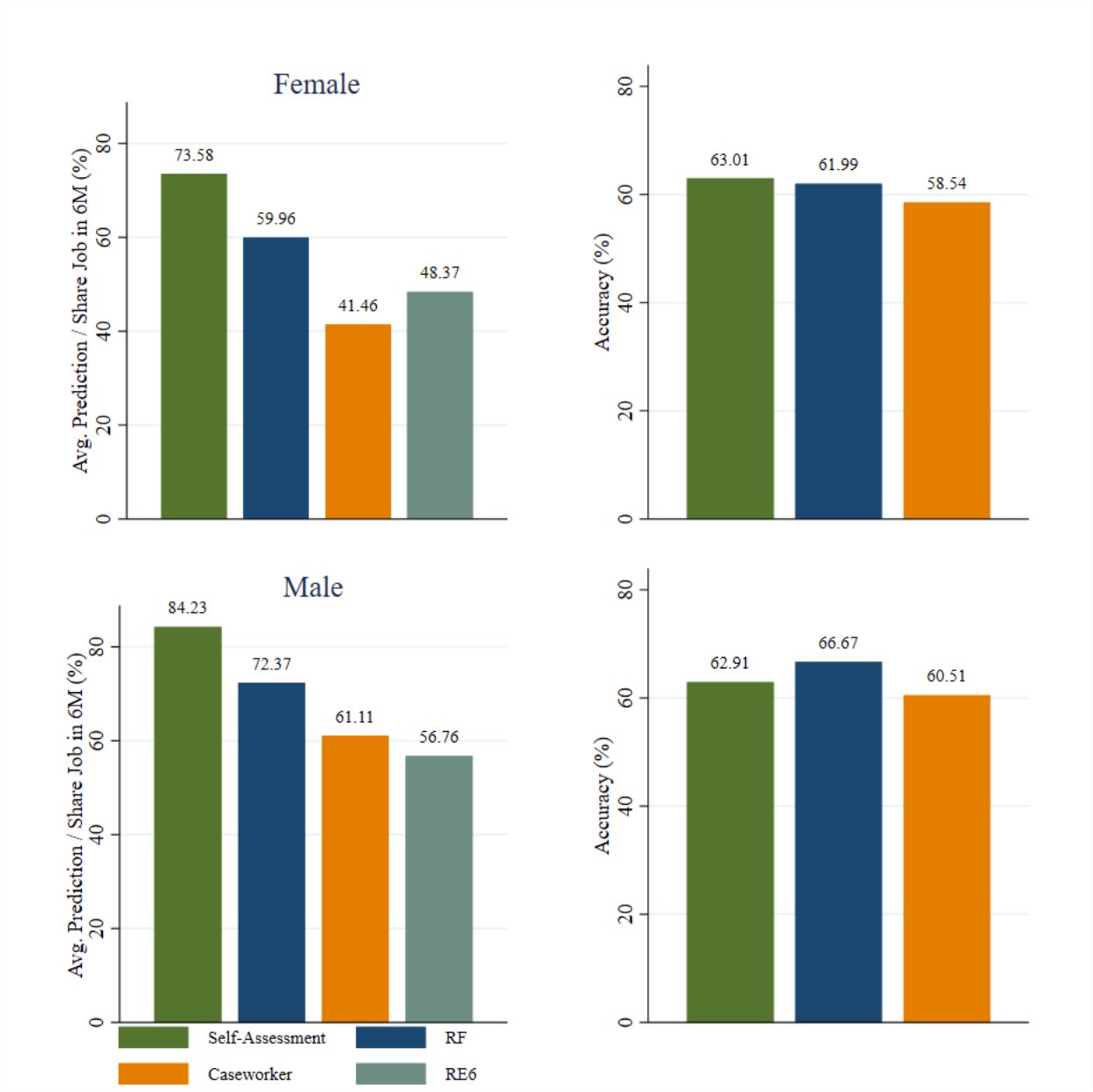
Note: 95% confidence intervals, obtained via bootstrapping. SE = self-assessment. CW = caseworker assessment. RF = random forest trained without past caseworker assessments, RFCW = random forest trained incl. past caseworker assessments.

Figure 6: Predictions and accuracies by age group



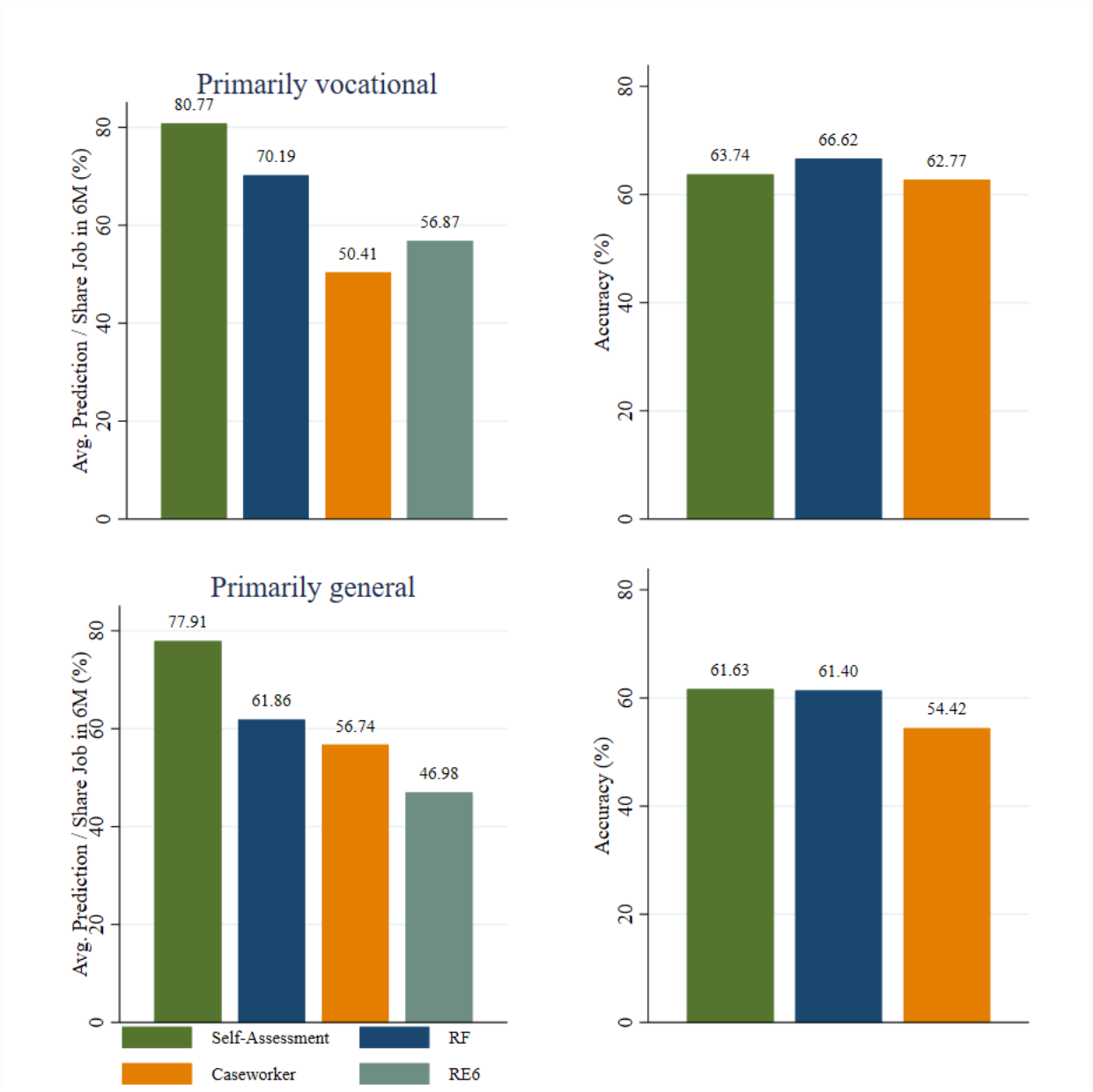
Note: This figure shows the average of positive predictions by age group (left panels) and the corresponding accuracies (right panels). The median age is 43. 585 (50.52%) individuals in our sample are up to 43 years old, 573 (49.48%) individuals are older than 43. The random forest results presented here exclude caseworker information as predictor. Source: EVA, IEB v.12.01.00.

Figure 7: Predictions and accuracies by gender



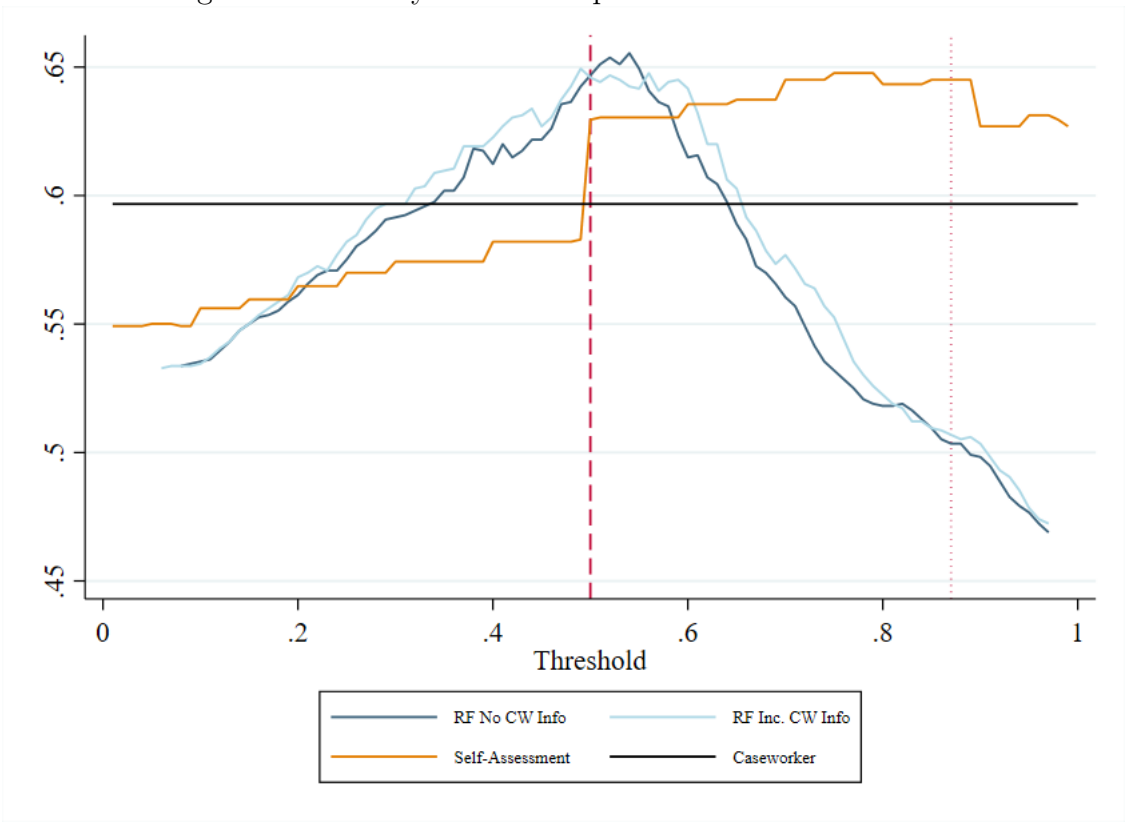
Note: This figure shows the average of positive predictions by gender (left panels) and the corresponding accuracies (right panels). 42% of the individuals are female. The random forest results presented here exclude caseworker information as predictor. Source: EVA, IEB v.12.01.00.

Figure 8: Predictions and accuracies by education



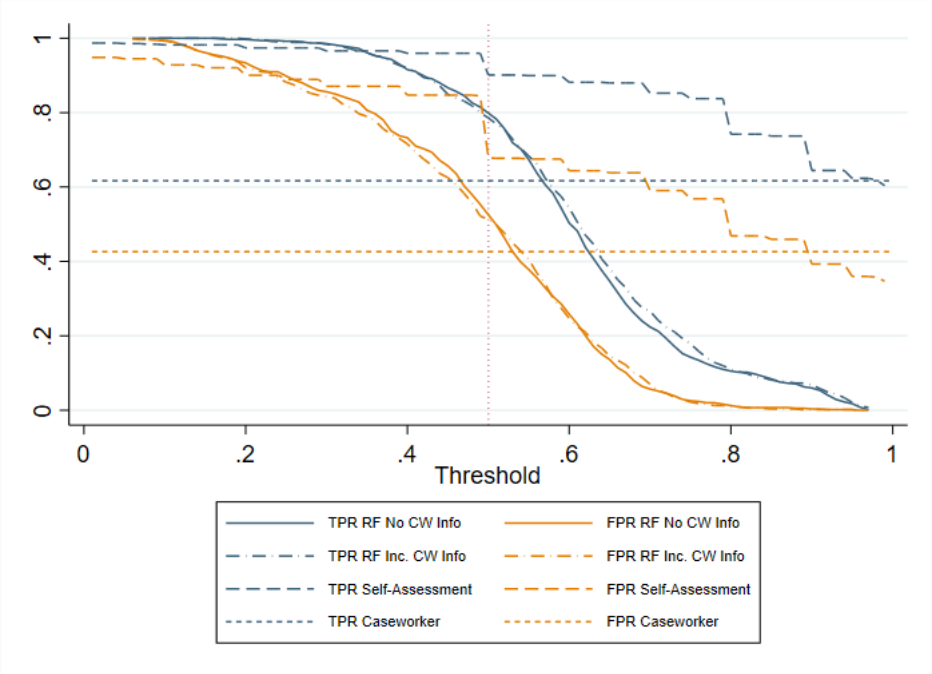
Note: This figure shows the average of positive predictions by education (left panels) and the corresponding accuracies (right panels). See Section 6 for a discussion of the educational classification. The group with primarily a general track (university degree or high school degree + vocational training) contains 37% of the sample. The other group has primarily followed a vocational educational track (high or low). The random forest results presented here exclude caseworker information as predictor. Source: EVA, IEB v.12.01.00.

Figure 9: Accuracy of the four predictors across thresholds



Note: This figure shows the accuracies of the four different predictors across all relevant classification thresholds for the survey sample. The dashed red line marks the (default) 50% threshold. The dotted red line marks the threshold at which the fraction of individuals that classify themselves as likely to find a job within six months roughly reflects the actual fraction of individuals that actually find a job. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Figure 10: TPR and FPR of the four predictors across thresholds



Note: This figure shows the True Positive Rate (TPR) and the False Positive Rate (FPR) of the four different predictors across all relevant classification thresholds for the survey sample. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Table 1: Self-assessed probability of RE6

Self-Assessed Prob. (%)	RE6 (%)	N	%	% Cumul.
0	22.9	35	3.02	3.02
1	0.0	1	0.09	3.11
5	33.3	3	0.26	3.37
8	100.0	1	0.09	3.45
10	10.0	10	0.86	4.32
15	0.0	4	0.35	4.66
20	31.3	16	1.38	6.04
25	0.0	6	0.52	6.56
30	33.3	15	1.30	7.86
40	23.5	17	1.47	9.33
49	0.0	1	0.09	9.41
50	28.6	126	10.88	20.29
51	0.0	1	0.09	20.38
55	50.0	2	0.17	20.55
60	39.3	28	2.42	22.97
65	25.0	4	0.35	23.32
70	39.5	43	3.71	27.03
75	42.9	21	1.81	28.84
80	52.2	113	9.76	38.60
85	37.5	8	0.69	39.29
90	61.3	93	8.03	47.32
95	41.9	31	2.68	50.00
98	75.0	4	0.35	50.35
99	60.0	15	1.30	51.64
100	66.4	560	48.36	100.00

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 and the actual fraction of individuals that found a job within six months after the interview. Source: EVA; IEB v.12.01.00.

Table 2: Descriptive statistics for the three samples

	Admin 2011-12	Admin 2012-13	Survey
RE6	49.26%	46.47%	53.20%
Age	42.10	41.99	43.44
Male	56.19%	56.68%	57.51%
German	73.72%	73.49%	82.82%
High school	37.51%	39.39%	37.56%
Vocational training	85.40%	85.01%	90.24%
University	21.78%	22.51%	21.93%
Daily wage in previous job	64.24	66.82	67.40
Total earnings in reg. empl. prev. year	15365.16	16287.46	16577.26
Total time in reg. empl. prev. year	231.68	236.55	247.09
Total time in unempl. previous year	57.01	56.38	52.43
N	29,130	30,255	1,158

Selected descriptive statistics for the three different samples. For both administrative samples, the RE6 refers to reintegration into the labor market within six months after the artificial interview date, defined as unemployment entry + 42 days (to match the timeline of the survey participants). Source: EVA; IEB v.12.01.00.

Table 3: Share of identical predictions

	RF (no CW info)	RF (incl CW info)	Self-assessment	CW assessment
RF (no caseworker info)	100%			
RF (incl caseworker info)	90%	100%		
Self-assessment	69%	69%	100%	
Caseworker assessment	60%	69%	60%	100%

Note: Share of individuals who are classified into the same category, by the different predictors. Source: EVA; IEB v.12.01.00.

Table 4: Estimates of predictor performance measures and their standard errors, for each predictor.

predictor	accuracy	TPR	FPR
random forest trained without past caseworker assessments	0.647	0.799	0.526
	(0.015)	(0.022)	(0.029)
random forest trained incl. past caseworker assessments	0.646	0.786	0.513
	(0.015)	(0.023)	(0.028)
self-assessment	0.630	0.900	0.679
	(0.014)	(0.012)	(0.019)
caseworker assessment	0.597	0.617	0.426
	(0.014)	(0.020)	(0.020)

Standard errors based on bootstrapping, in parentheses. Random forest and self-assessment dichotomized using threshold of 0.5.

Table 5: Explanatory power of different predictors - separate models.

	(1)	(2)	(3)	(4)
covariates				
RF Pred. (no CW info)	0.308***			
	(0.0296)			
RF Pred. (incl CW info)		0.302***		
		(0.0294)		
Self-Assessment			0.342***	
			(0.0329)	
CW Assessment				0.190***
				(0.0289)
Constant	0.325***	0.333***	0.260***	0.431***
	(0.0240)	(0.0237)	(0.0286)	(0.0212)
# observations	1,158	1,158	1,158	1,158
R^2	0.084	0.082	0.076	0.036

Note: Estimated coefficients from different linear probability models. The dependent variable in all four models is whether the individual actually found a job within six months after the interview. All predictors are included as binary variables. For the random forest classifiers and the self-assessment, the predictor takes the value 1 if the predicted probability exceeds 50% and 0 otherwise. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: EVA; IEB v.12.01.00.

Table 6: Estimates of accuracies and their standard errors, for each predictor, by gender.

predictor	males	females
random forest	0.667	0.620
	(0.019)	(0.020)
self-assessment	0.629	0.630
	(0.022)	(0.019)
caseworker assessment	0.605	0.585
	(0.022)	(0.019)

Asymptotic standard errors in parentheses. Random forest trained without past caseworker assessments. Random forest and self-assessment dichotomized using a threshold of 0.5.

Table 7: Explanatory power of different predictors - combined models.

	(1)	(2)	(3)	(4)
covariates				
constant	0.325***	0.143***	0.273***	0.119***
	(0.0240)	(0.0283)	(0.0261)	(0.0281)
random forest prediction	0.308***	0.254***	0.278***	0.237***
	(0.0296)	(0.0305)	(0.0305)	(0.0312)
self-assessment		0.274***		0.251***
		(0.0336)		(0.0345)
caseworker assessment			0.137***	0.101***
			(0.0289)	(0.0294)
# observations	1,158	1,158	1,158	1,158
R^2	0.084	0.130	0.102	0.140

Note: Estimation of linear probability models. The dependent variable is whether the individual experienced RE6=1. All predictions are included as a binary variable. The machine learning predictions and the self-assessed predictions take the value 1 if the predicted probability exceeds 50%. Random forest predictor excludes caseworker assessments. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 8: The effect of risk aversion on whether the self-assessment predicts slow reemployment or not, if machine learning incorrectly predicts fast reemployment.

covariates	(1)	(2)	(3)	(4)
constant	0.133***	0.329**	0.150***	0.348**
	(0.044)	(0.161)	(0.029)	(0.155)
risk aversion (0-10)	0.022**	0.022**		
	(0.010)	(0.010)		
high risk aversion (0/1)			0.153***	0.153***
			(0.049)	(0.049)
continuous random forest predictor included	no	yes	no	yes
# observations	285	285	285	285
R^2	0.016	0.020	0.034	0.038

Subsample with RE6=0 and with the random forest predicting that RE6=1. Outcome is whether the self-assessment is correct, which has value 1 (0) for 63 (222) individuals. Random forest predictor excludes caseworker assessments. Column (4) gives main results whereas columns (1)-(3) are robustness analyses. *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses.

Table 9: The effect of risk aversion on whether the self-assessment is correct or not, if the machine learning prediction is incorrect.

	(1)	(2)	(3)	(4)
covariates				
constant	0.351***	1.526***	0.365***	1.544***
	(0.047)	(0.090)	(0.033)	(0.084)
risk aversion (0-10)	0.018*	0.020**		
	(0.010)	(0.008)		
high risk aversion (0/1)			0.130***	0.131***
			(0.049)	(0.042)
continuous random forest predictor included	no	yes	no	yes
# observations	401	401	401	401
R^2	0.008	0.280	0.017	0.287

Subsample with an incorrect random forest prediction. Outcome is whether the self-assessment is correct, which has value 1 (0) for 171 (230) individuals. Random forest predictor excludes caseworker assessments. Column (4) gives main results whereas columns (1)-(3) are robustness analyses. *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses.

Table 10: Explanatory power of different predictors - combined models including interaction terms of the separate predictors.

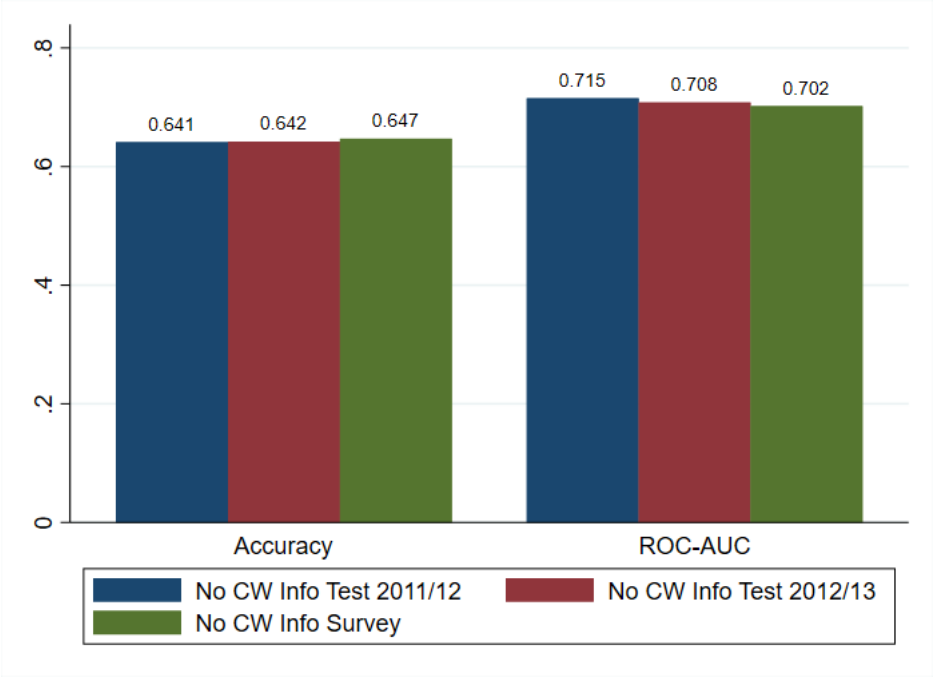
	(1)	(2)
covariates		
constant	0.119***	0.102***
	(0.0281)	(0.0283)
random forest prediction	0.237***	0.220***
	(0.0312)	(0.0595)
self-assessment	0.251***	0.332***
	(0.0345)	(0.0481)
caseworker assessment	0.101***	0.100
	(0.0294)	(0.0721)
random forest × self-assessment		-0.065
		(0.0700)
random forest × caseworker assessment		0.144**
		(0.0641)
self-assessment × caseworker assessment		-0.118
		(0.0757)
# observations	1,158	1,158
R^2	0.140	0.146

Note: Estimation of linear probability models. The dependent variable is whether the individual experienced RE6=1. Model (1) is without interaction terms and replicates Column (4) of Table 7. Model (2) includes interaction terms between the separate predictors. All predictions are included as a binary variable. The machine learning predictions and the self-assessed predictions take the value 1 if the predicted probability exceeds 50%. Random forest predictor excludes caseworker assessments. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix: For Online Publication

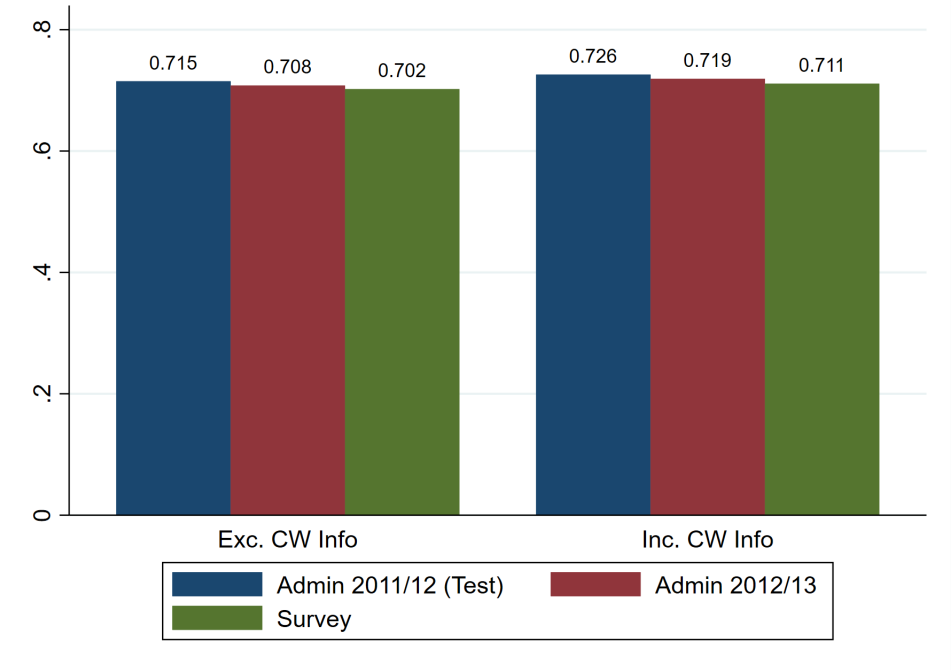
Appendix A. Figures and Tables

Figure A 1: Accuracy and ROC-AUC



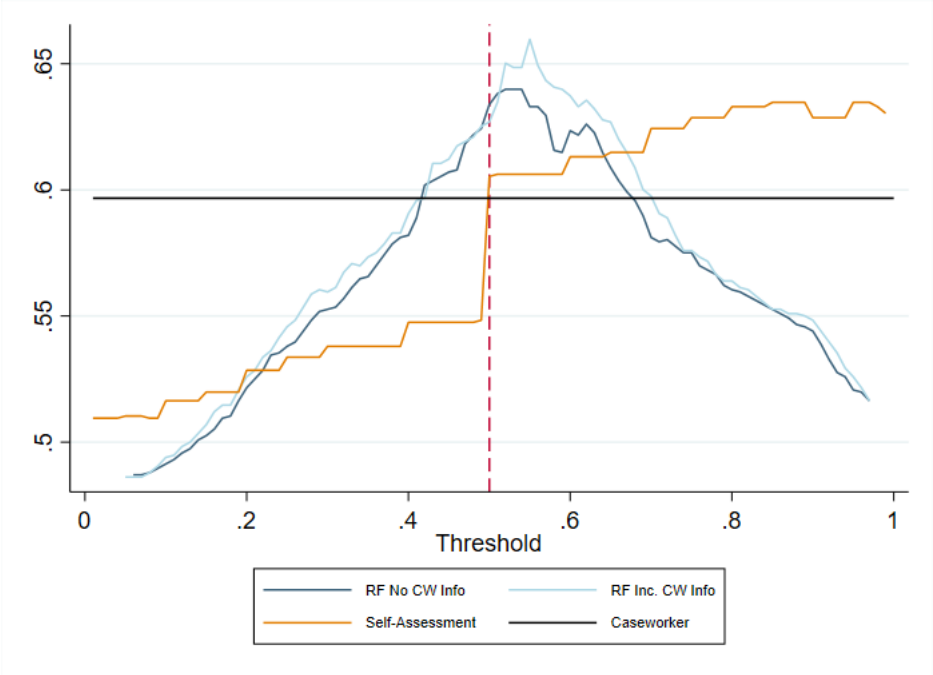
Note: This figure shows the Accuracy and ROC-AUC of the machine learning model for the admin data 2011-12 and 2012-13 and for the survey sample. The accuracies are measured at a classification threshold of 0.5. Source: EVA; IEB v.12.01.00.

Figure A 2: ROC-AUC scores for random forests with and without caseworker info



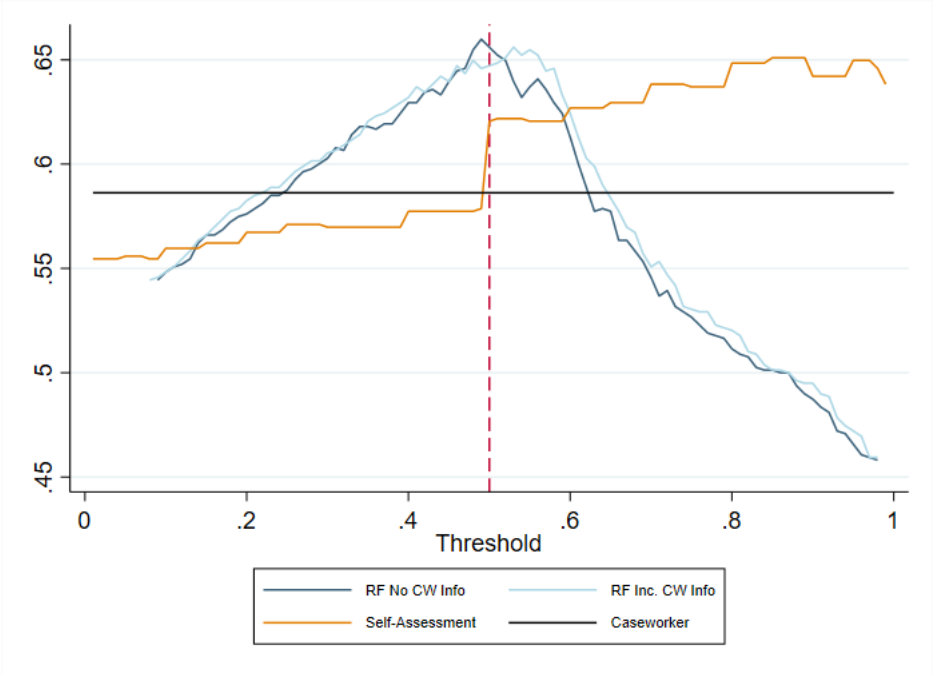
Note: This figure shows the ROC-AUC Scores of the machine learning model for the admin data 2011-12 and 2012-13 and for the survey sample. The first set of bars shows the results excluding the caseworker assessments as input for the machine learning model and the second set of bars shows the results including the caseworker assessments as input for the machine learning model. Source: EVA; IEB v.12.01.00.

Figure A 3: Accuracy of the four predictors across thresholds (outcome "having a job" instead of "finding a job")



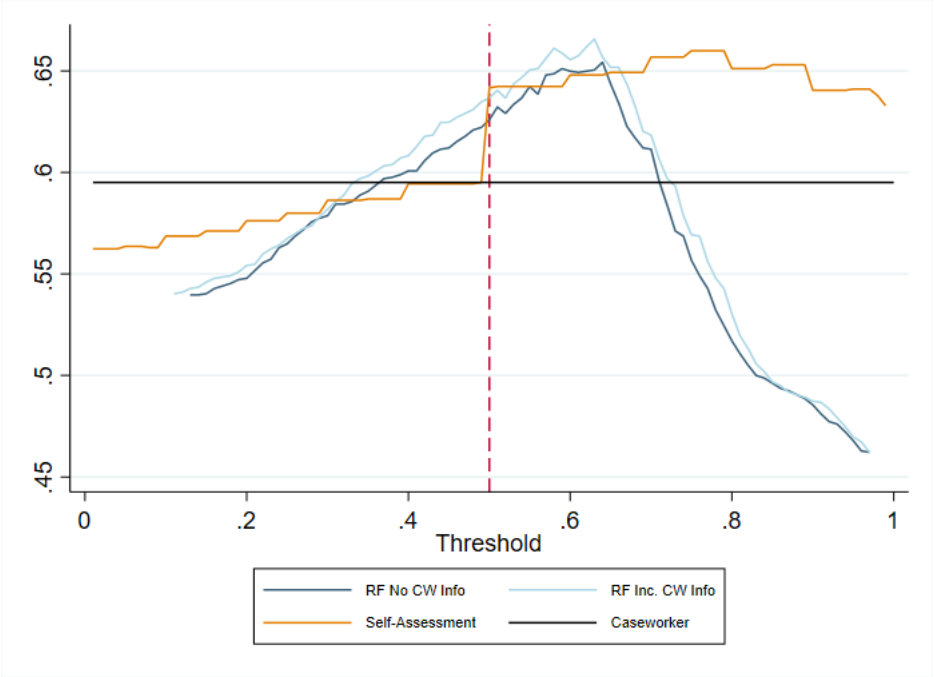
Note: This figure replicates Figure 3, but with the outcome defined as having a job six months after the interview instead of finding a job within six months. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Figure A 4: Accuracy of the four predictors across thresholds (profiles not older than 6 weeks)



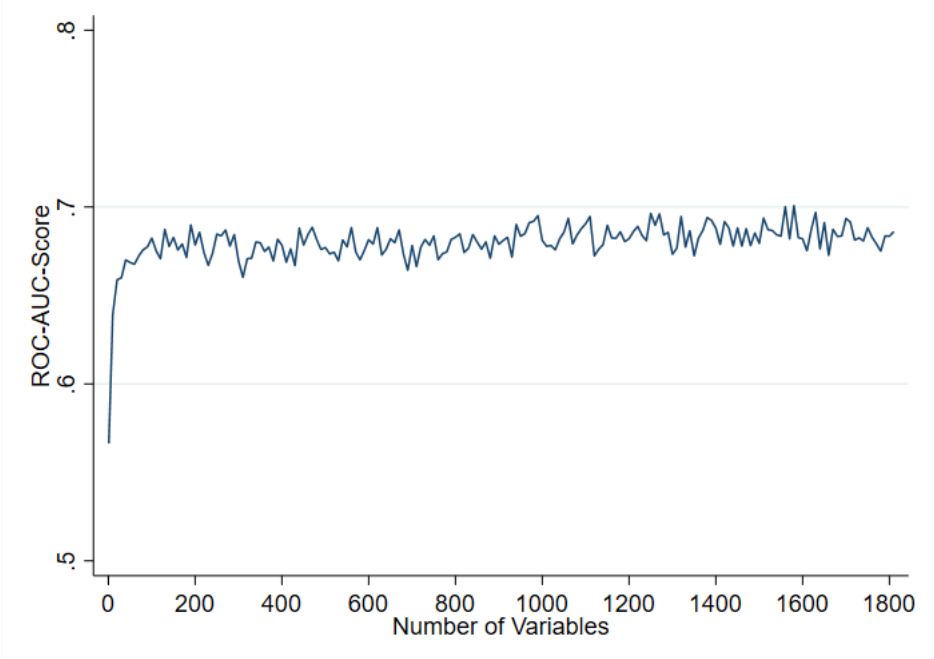
Note: This figure replicates Figure 3, but only with individuals with profiles not older than 6 weeks (instead of one year). RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Figure A 5: Accuracy of the four predictors across thresholds (less strict sample restrictions)



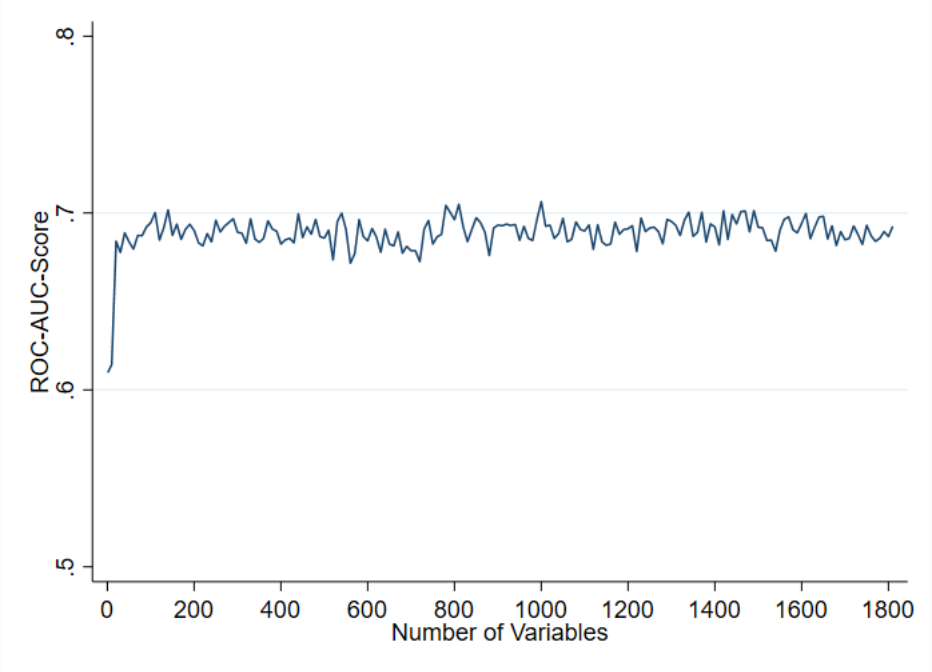
Note: This figure replicates Figure 3, but for a wider sample that is constructed using less strict sample definitions. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self assessed probability to find a job, Caseworker = caseworker assessment. Source: EVA; IEB v.12.01.00.

Figure A 6: ROC-AUC for an increasing number of variables - survey, no CW info



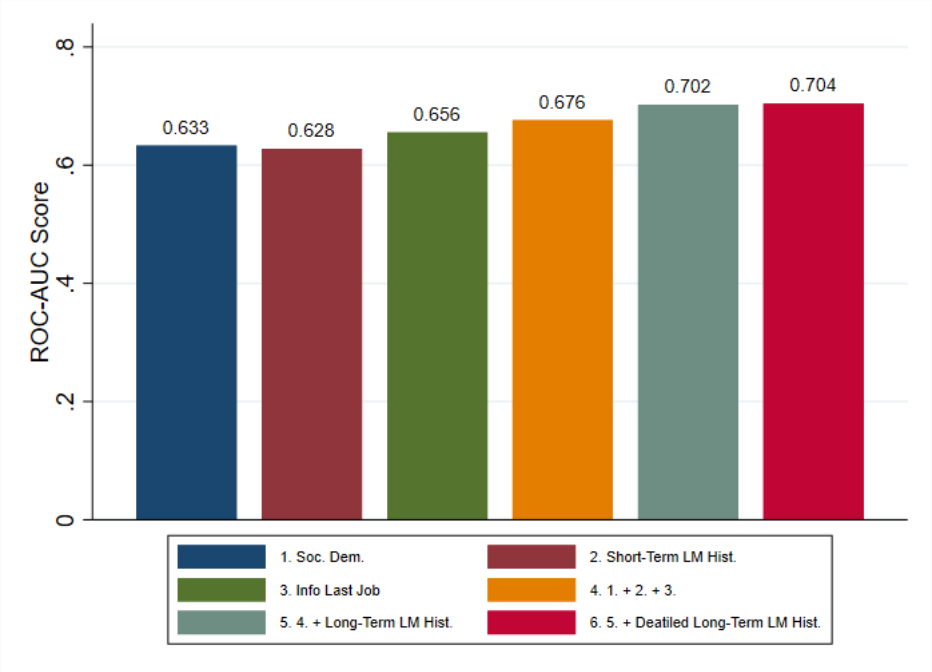
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of variables available for prediction (no additional model tuning). Source: EVA; IEB v.12.01.00.

Figure A 7: ROC-AUC for an increasing number of variables - survey, inc. CW info



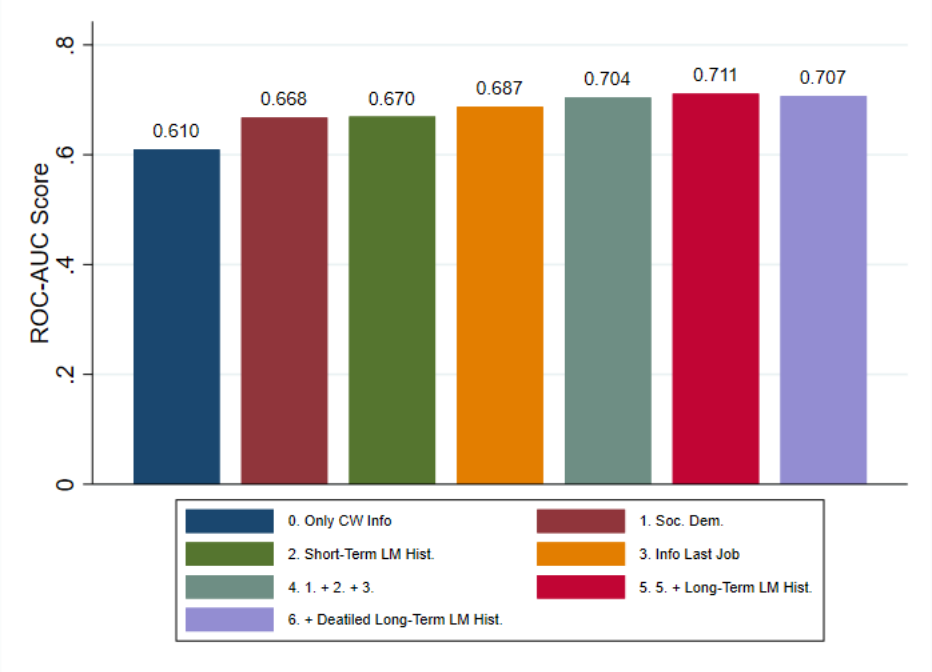
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of variables available for prediction (no additional model tuning). Source: EVA; IEB v.12.01.00.

Figure A 8: ROC-AUC for different variable inputs - no CW info



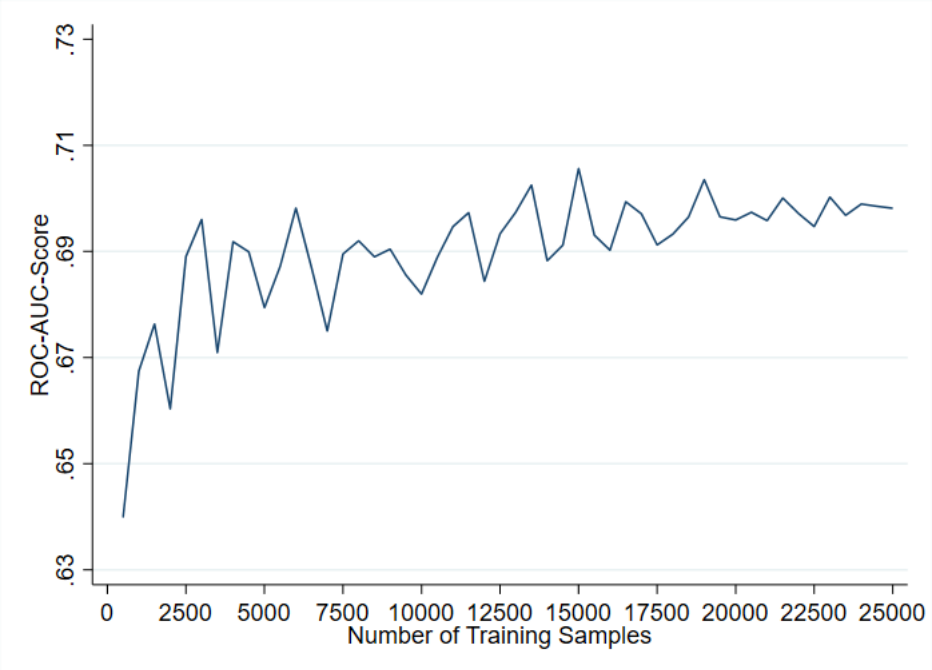
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample for different sets of variables included in the prediction algorithm. The models do not include caseworker assessments. Source: EVA; IEB v.12.01.00.

Figure A 9: ROC-AUC for different variable inputs - including CW info



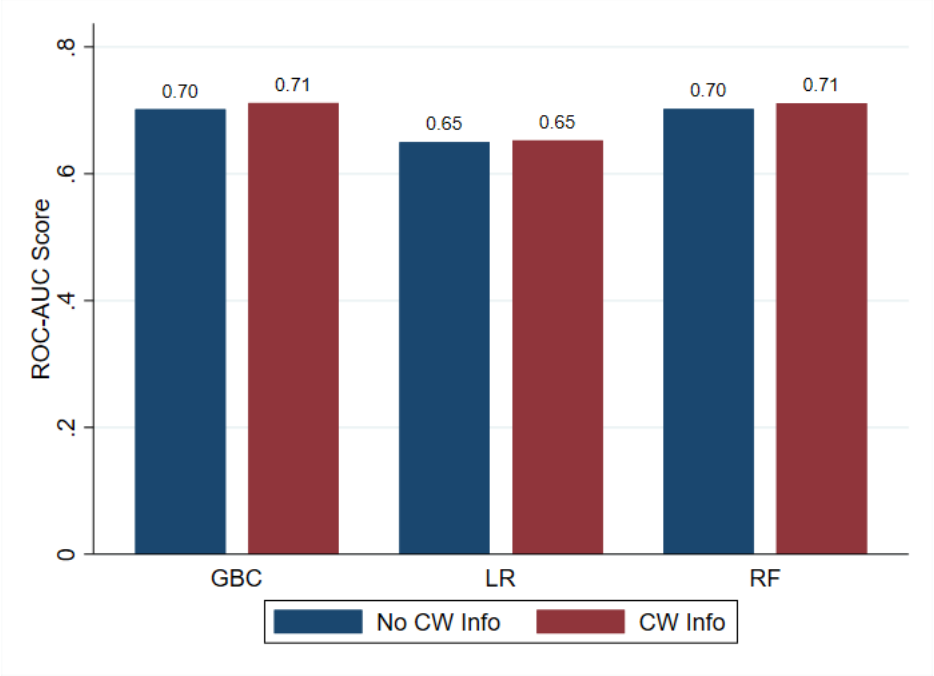
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample for different sets of variables included in the prediction algorithm. All of the models include caseworker assessments. Source: EVA; IEB v.12.01.00.

Figure A 10: ROC-AUC Scores for an increasing number of training samples (1000 - 25000 Training Samples)



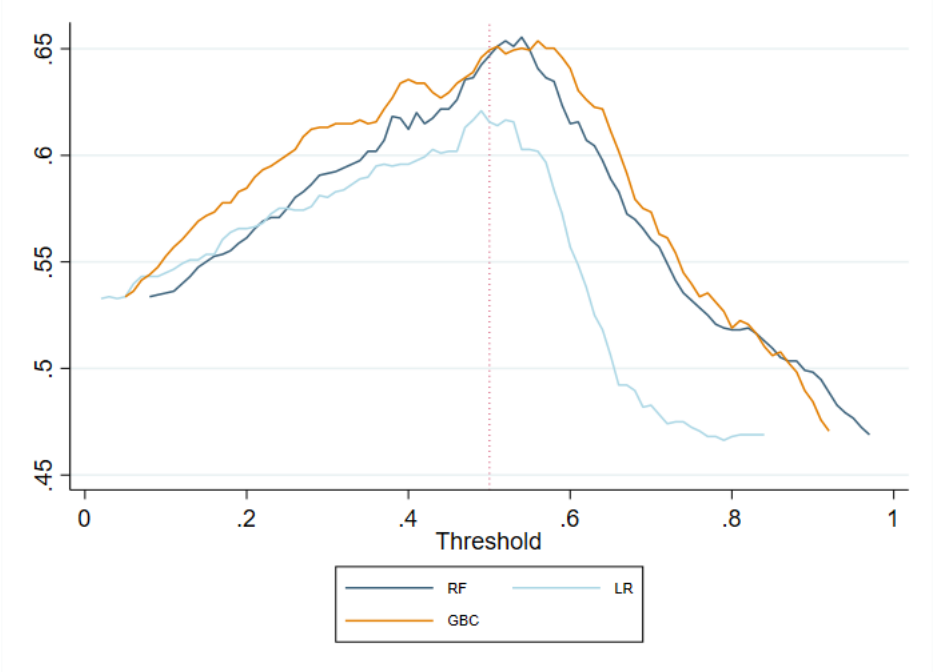
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of training samples available to train the algorithm (the algorithm was trained using a random subset of the admin data from 2011-12, as in the main specification). Source: IEB v.12.01.00.

Figure A 11: ROC-AUC by method - survey data



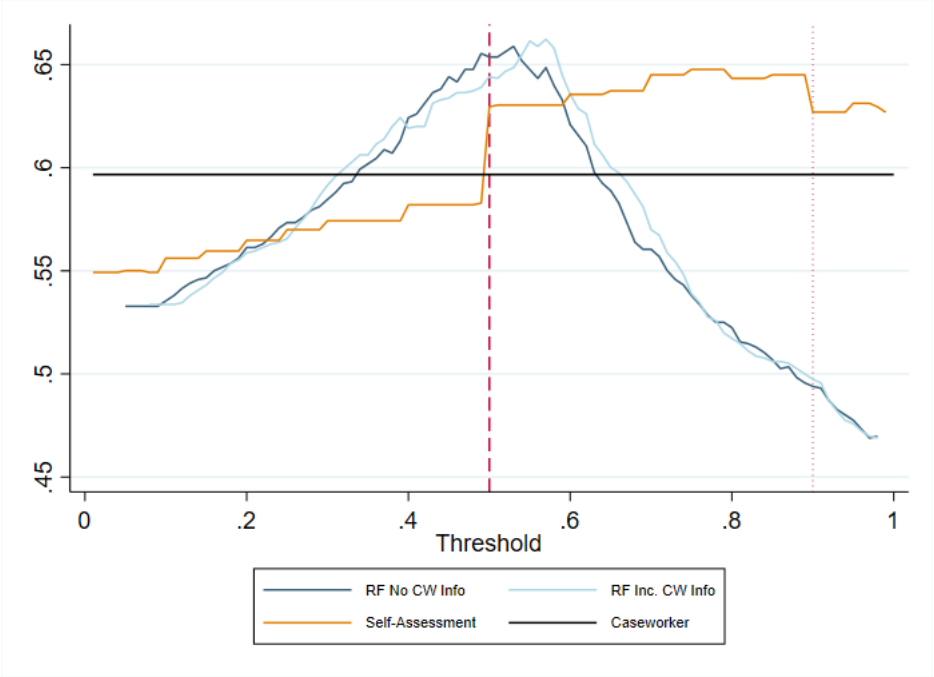
Note: This figure shows the ROC-AUC achieved by different classification methods for the survey sample. GBC = Gradient Boosting Classifier; LR = Logit Regression; RF = Random Forest. EVA; IEB v.12.01.00.

Figure A 12: Accuracy across thresholds by method - survey data



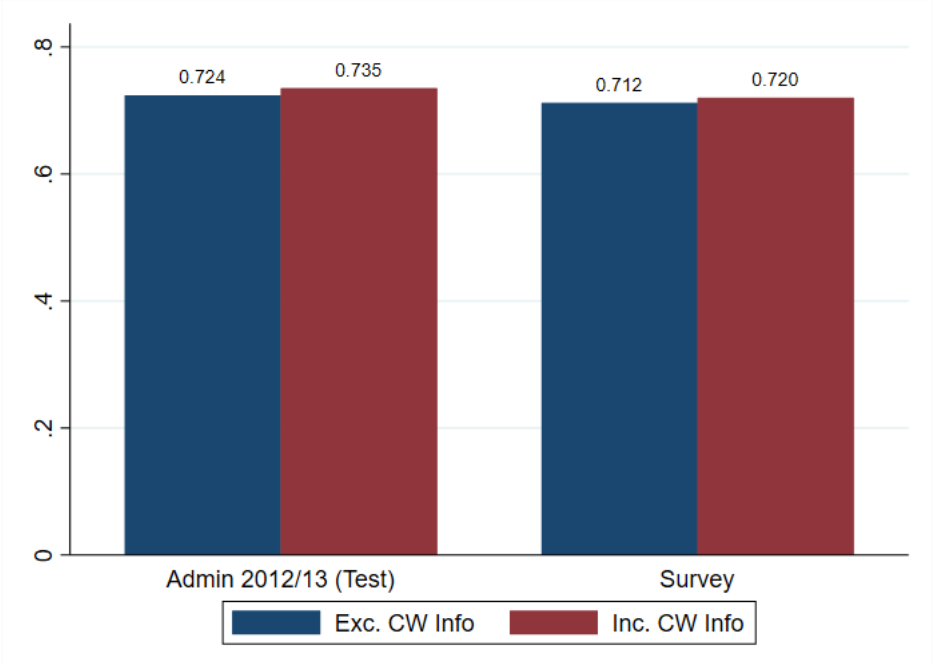
Note: This figure shows the accuracies of different classification methods across all relevant classification thresholds for the survey sample. GBC = Gradient Boosting Classifier; LR = Logit Regression; RF = Random Forest. Source: EVA; IEB v.12.01.00.

Figure A 13: Accuracy across thresholds using 2012/13 data as training input - survey data



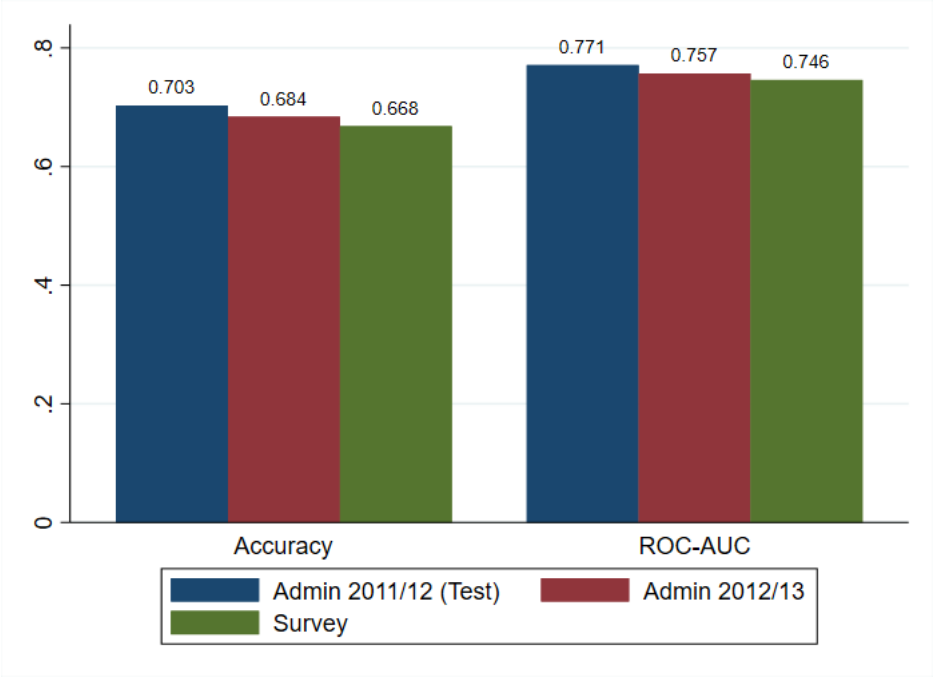
Note: This figure shows the accuracies of different classification methods across all relevant classification thresholds for the survey sample. Instead of using data from the year before (2011/12) to train the algorithm, we used data from 2012/13 as training input. Source: EVA; IEB v.12.01.00.

Figure A 14: ROC-AUC scores using 2012/13 data as training input



Note: This figure shows the ROC-AUC scores for the holdout sample in 2012/13 and for the survey sample. Instead of using data from the year before (2011/12) to train the algorithm, we used data from 2012/13 as training input. Source: EVA; IEB v.12.01.00.

Figure A 15: Accuracy and ROC-AUC scores for predictions of the caseworker profiles



Note: This figure shows the accuracies and ROC-AUC scores for the three different samples. The predicted variable in this case is the caseworker profile. Source: EVA; IEB v.12.01.00.

Table A 1: Caseworker profiles across samples

	Admin 2011/12	Admin 2012/13	Survey
Marktprofil	39.63%	36.54%	39.03%
Aktivierungsprofil	8.06%	8.80%	13.73%
Förderprofil	39.53%	41.37%	40.50%
Entwicklungsprofil	10.45%	11.02%	6.39%
Stabilisierungsprofil	1.46%	1.58%	0.26%
Unterstützungsprofil	0.87%	0.69%	0.09%

Note: This table shows detailed caseworker profiles for the different samples. The first two categories reflect the expectation of reintegration within six months, the last four categories reflect the expectation of the opposite. Source: EVA; 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	Total Days Employed Last 1-7 Years
	Total Days Employed Last 1-7 Years (Marg. Empl.)
	Total Days Employed Lifetime
	Total Days Employed Lifetime (Reg. Empl.)
	Total Days Employed Lifetime (Marg. Empl.)
	Total Earnings Last 1-7 Years
	Total Earnings Last 1-7 Years (Reg. Empl.)
	Total Earnings Lifetime (Lifetime)
	Total Earnings Last 1-7 Years (Marg. Empl.)
Unemployment History	Amount UE Benefits at Interview
	Total Days Receiving UE Ben. Last 1-7 Years
	Total Days Receiving UE Ben. Lifetime
	Total Amount UE Ben. Last 1-7 Years
	Total Amount UE Ben. Lifetime
	Total Days Registr. UE Last 1-7 Years
	Total Days Registr. UE Lifetime
	Total Days Labor Market Program Last 1-7 years
	Total Days Labor Market Program Lifetime
LHG History	LHG at interview
	Total Days LHG Lifetime
	Total 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: EVA; IEB v.12.01.00.

Table A 3: Correlations between the different predictors

	RF (no CW info)	RF (incl CW info)	Self-Assessment	CW Assessment
RF (no CW info)	1			
RF (incl CW info)	0.77	1		
Self-Assessment	0.23	0.25	1	
CW Assessment	0.20	0.38	0.22	1

Note: This table shows the correlations between the different predictors. Source: EVA; IEB v.12.01.00.

Table A 4: Explanatory power of different predictors - separate models (continuous predictors)

	(1)	(2)	(3)
VARIABLES	RE6	RE6	RE6
RF continuous (no CW)	0.011***		
	(0.0006)		
RF continuous (incl CW)		0.011***	
		(0.0006)	
Self-Assessment continuous			0.006***
			(0.0005)
Constant	-0.076**	-0.083**	0.074*
	(0.0384)	(0.0362)	(0.0415)
Observations	1,158	1,158	1,158
R-squared	0.134	0.148	0.093

Note: This table shows the estimated coefficients from separate linear probability models. The dependent variable in all four models is whether the individual actually found a job within six months after the interview. All predictors are included as continuous variables and are scaled to run from 0 to 100. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: EVA; IEB v.12.01.00.

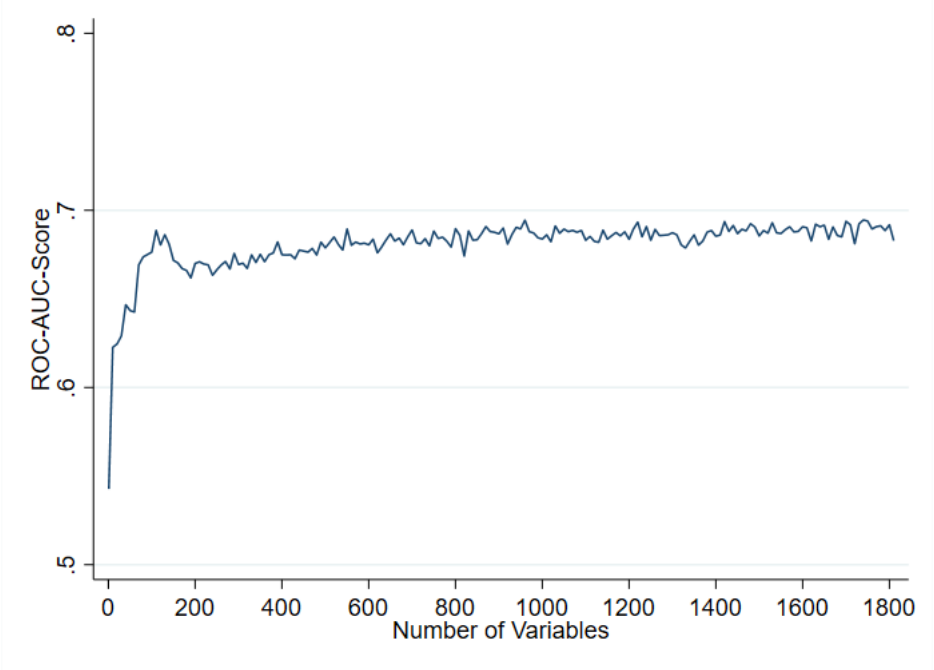
Table A 5: Explanatory power of different predictors - combined models (random forest predictor and self-assessment as continuous variables)

	(1)	(2)	(3)
VARIABLES	RE6	RE6	RE6
RF continuous (no CW)	0.011***	0.009***	0.0086***
	(0.0006)	(0.0007)	(0.0008)
Self-Assessment continuous		0.004***	0.0037***
		(0.0005)	(0.0005)
CW Assessment			0.066**
			(0.0296)
Constant	-0.076**	-0.283***	-0.277***
	(0.0384)	(0.0417)	(0.0417)
Observations	1,158	1,158	1,158
R-squared	0.134	0.174	0.178

Note: This table shows the estimated coefficients from separate linear probability models. The dependent variable all models is whether the individual actually found a job within six months after the interview. Model (1) only includes the prediction from the ML model without caseworker predictions. Model (2) additionally includes the prediction from the self-assessment of the individuals. Model (4) includes all three predictions. The machine learning and self-assessment predictions are included as continuous variables and are scaled to run from 0 to 100. The caseworker prediction as a binary variable. The caseworker predictions take the value 1 if the caseworker predicted reintegration within six months. *** p<0.01, ** p<0.05, * p<0.1. Source: EVA; IEB v.12.01.00.

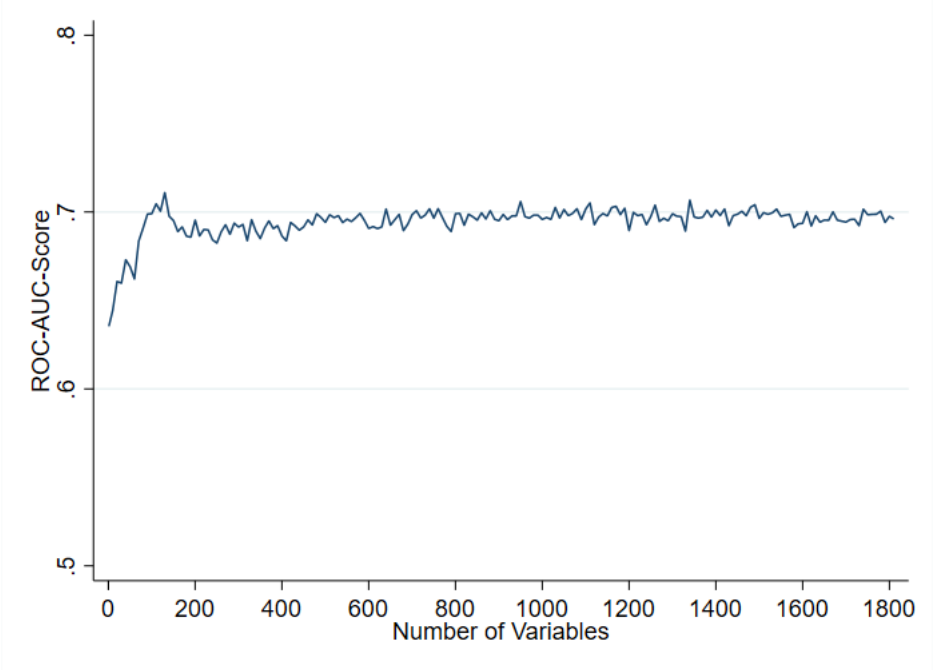
Appendix B. Additional Figures and Tables

Figure B 1: ROC-AUC for an increasing number of variables - admin, no CW info



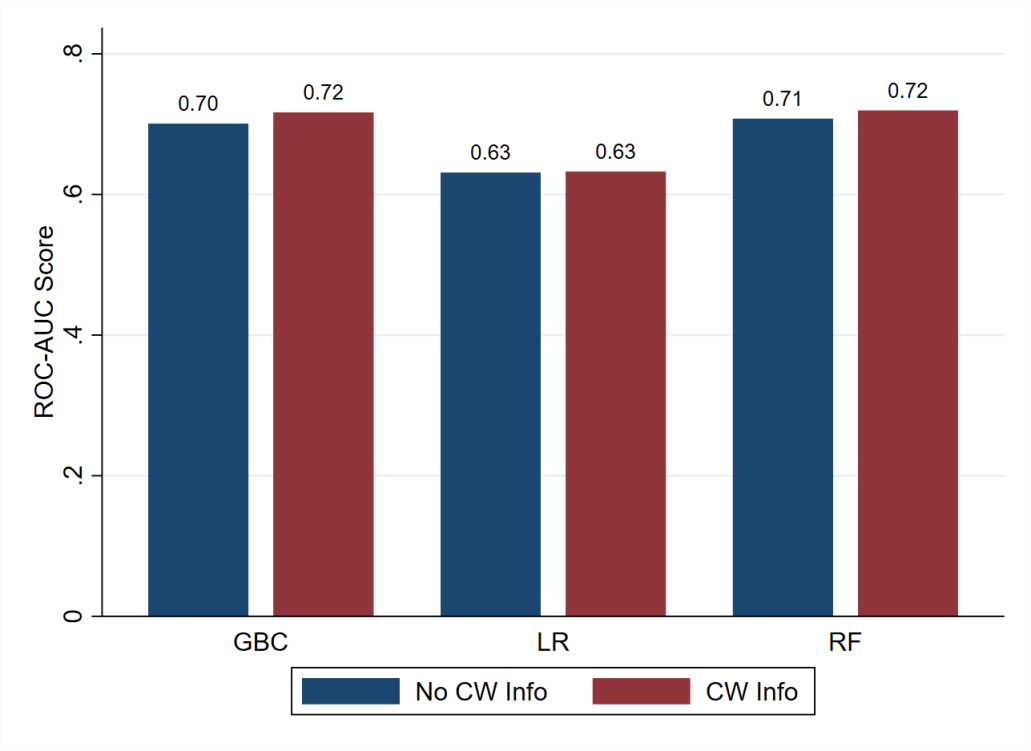
Note: This figure shows the ROC-AUC Scores of the random forest model for the admin sample when we increase the number of variables available for prediction (no additional model tuning). Source: IEB v.12.01.00.

Figure B 2: ROC-AUC for an increasing number of variables - admin, incl CW info



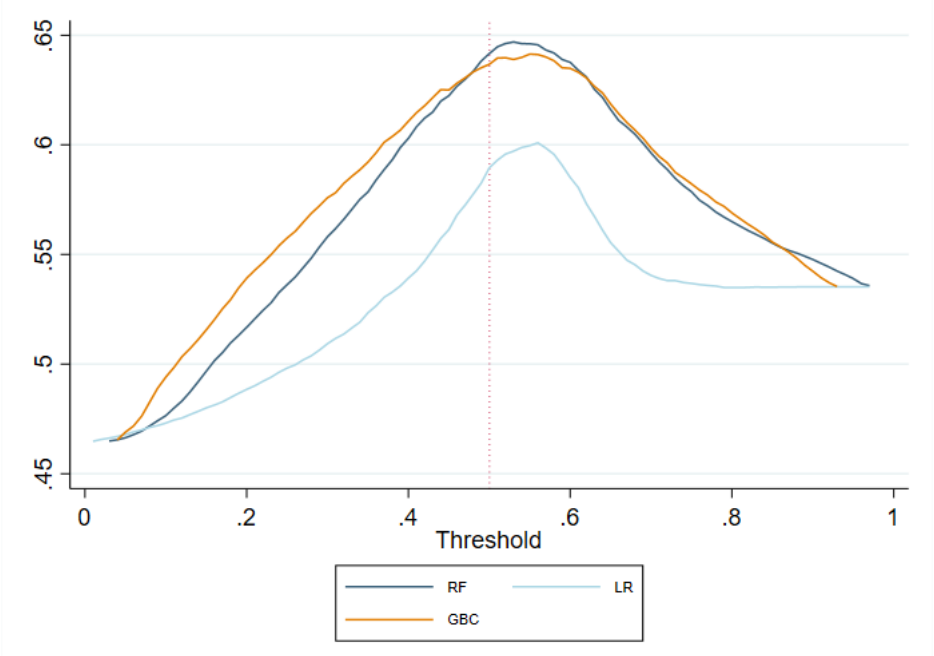
Note: This figure shows the ROC-AUC Scores of the random forest model for the admin sample when we increase the number of variables available for prediction (no additional model tuning). Source: IEB v.12.01.00.

Figure B 3: ROC-AUC by method - admin data 2012/13



Note: This figure shows the ROC-AUC achieved by different classification methods for the admin sample. GBC = Gradient Boosting Classifier; LR = Logit Regression; RF = Random Forest. Source: IEB v.12.01.00.

Figure B 4: Accuracy across thresholds by method - admin data 2012/13



Note: This figure shows the accuracies of different classification methods across all relevant classification thresholds for the admin sample. GBC = Gradient Boosting Classifier; LR = Logit Regression; RF = Random Forest. Source: IEB v.12.01.00.

Appendix C. Comparison in terms of a model framework

Let the index i denote an individual. Let $\theta_i(t)$ and $\Theta_i(t)$ denote the hazard rate and integrated hazard rate of the unemployment duration distribution of T of individual i, 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

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

This "log – log" transformation has the advantage that 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 this observed self-reported version $\widetilde{y}_i$ of y_i to equal the true y_i plus a measurement 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 now 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 $-\infty$. (For example, 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 as a simple model to shape thoughts. 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 illustrates that the self-reported predictor can be very informative on the individual characteristics v_i. However, the informativeness critically depends on $\text{var}(v)$ versus σ_ε^2. Moreover, if we drop the MPH assumption and/or allow for time-varying v_i then anything is possible.

We make three remarks. First, in the MPH context, the informational value of the self-reported probability is reduced when we move from the reported probability 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 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.

We now consider the translation of the caseworker assessment into whether $\mathbb{E}(T) \leq 6$ and into whether $\Pr(T \leq 6)$ exceeds 0.5, respectively. For any duration variable,

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

To relate $\text{I}(\mathbb{E}(T_i) \leq 6)$ 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 specification. At face value, it is less informative on v_i than the predictor based on the client's self-assessment. 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 for the caseworker assessment 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 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 is, the more informative the caseworker assessments will be for $\Pr(T \leq 7 | T \geq 1)$.

As a final note, in the above model settings, one could interpret machine learning as an approach where observed characteristics and past outcomes x_i provide information on Θ_i or on v_i.