Working Paper 2024-06

Predicting re-employment: machine learning versus assessments by unemployed workers and by their caseworkers

Gerard J. van den Berg Max Kunaschk
Julia Lang Gesine Stephan Arne Uhlendorff

La Chaire de Sécurisation des Parcours Professionnels est gérée par la Fondation du Risque (Fondation de recherche reconnue d'utilité publique). Elle rassemble des chercheurs de Sciences Po et du Groupe des Écoles Nationales d'Économie et Statistique (GENES) qui comprend notamment l'Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE) et le Centre de Recherche en Economie et Statistique (CREST).
Les travaux réalisés dans le cadre de la Chaire « Sécurisation des Parcours Professionnels » sont sous l'égide de la Fondation du Risque.

Predicting re-employment: machine learning versus assessments by unemployed workers and by their caseworkers

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

June 2024

Abstract

We analyze unique data on three sources of information on the probability of re-employment within 6 months (RE6), for the same individuals sampled from the inflow into unemployment. First, they were asked for their perceived probability of RE6. Second, their caseworkers revealed whether they expected RE6. Third, random-forest machine learning methods are trained on administrative data on the full inflow, to predict individual RE6. We compare the predictive performance of these measures and consider how combinations improve this performance. We show that self-reported (and to a lesser extent caseworker) assessments sometimes contain information not captured by the machine learning algorithm.

Keywords: unemployment, expectations, prediction, random forest, unemployment insurance, information.
JEL codes: J64, J65, C55, C53, C41, C21.

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

Acknowledgements: We thank Eve Caroli, Xavier D'Haultfoeuille, Jeffrey Smith and participants at the final international conference of the DFG SPP 1764, the 17th IZA & 4th IZA/CREST Conference on Labor Market Policy Evaluation, ESPE 2023, IAAE 2023, VfS 2023 and EALE 2023, for helpful comments and discussions. 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). Furthermore, 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 outcomes (Manski, 2018; Mueller and Spinnewijn, 2023b). For example, 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 expected job finding probabili-

[1]For expositional reasons we often 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.

ties of unemployed workers are predictive for actual job finding probabilities. They additionally document an on average optimistic bias among job seekers.

While the literature analyzing the predictive power of subjective expectations usually relies on linear regressions controlling for a rather small set of observed characteristics, we are the first paper analyzing this question by employing potentially more powerful machine learning techniques in combination with detailed administrative data that also contain caseworker information on the predicted reemployment probabilities.[3]

We study to what extent combinations of the three data sources - administrative data on individuals, data on caseworker-profiling, and self-assessments - improve the quality of predictions. Our results show that the information provided by one of the three measures is not uniformly dominated by the information from the other measures. Thus, a combination of all three approaches may be most effective in accurately identifying those at risk of long-term unemployment. In sensitivity analyses we examine the role of subsets of inputs into the random forest classifier. As a practical recommendation, subjective self-assessments and/or caseworker assessments can be used as inputs for machine learning algorithms to obtain individual predictions. More modestly, machine learning predictions may be supplied to caseworkers on an individual-client basis.

The remainder of this paper proceeds as follows: Section 2 describes the institutional background of the UI system in Germany and of the statistical profiling of jobseekers. Section 3 gives an overview of the data we use for our analyses. 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 results. Section 7 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 unemployment benefits is 60 percent of the previous net wage. This increases to 67 percent if the unemployed person has children. The

[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.

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. Labor market 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 between caseworkers and unemployed workers was about 50 minutes.[4]

Caseworkers in labor market agencies in Germany apply a so-called soft-profiling 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

[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).

[5]Various approaches can be distinguished for categorizations of 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 different approaches. Rule-based profiling uses administrative eligibility criteria, e.g., age or education, to classify newly unemployed individuals into categories. Caseworker-based profiling (soft profiling) relies on the caseworkers' assessment of the job seekers' reemployment chances. Statistical profiling uses statistical models to predict the expected unemployment duration of an individual or their probability of becoming long-term unemployed.

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 are the starting point. Newly registered unemployed individuals were sampled and were subsequently 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 roughly 4 to 6 weeks after unemployment entry between August 2012 and January 2013. We restrict the sample to individuals who were over 25 years and who are registered as unemployed and receive UI benefits at the time of their interview. In addition, we exclude individuals who were unemployed during the three months before their current unemployment spell. Finally, we restrict the survey sample to individuals who gave permission to merge their survey answers with the administrative data, who have answered the survey question central to our analyses, and for whom we observe a 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.

[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.

Table 1 gives an overview of the distribution of answers for the self-assessment variable.[8]

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

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

[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. During the interview, permission was requested to link the responses to administrative data, under the above-mentioned provisos.

[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 6.2).

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 flows 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 training on data on the unemployment entries in the five regions 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.

[Table 2: Descriptive Statistics Across Samples]

Comparing the administrative data in 2011/12 (used to train the machine learning algorithm) with the administrative data in 2012/13 (containing the full sample of unemployment entries for the five regions where the survey was conducted), we hardly see any differences. This is in line with the above-mentioned similarity of labor market statistics for 2011 and 2012. The largest difference concerns the fraction who find a job within six months after the hypothetical interview date, which is about three percentage points lower in 2012/13 than in 2011/12.

Comparing the administrative samples with the survey sample, we see that most 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

[11] For 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, see Subsection 4.3.

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 present a unifying framework. 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 only binary. Regarding (ii), the conditioning on $T \geq 1$ is relevant for the self-reported assessment. However, clearly, 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, the value of the conditional survival function at $t = 7$ is more informative than just

[13] This positive selection does not influence the overall performance of our prediction model (see Figure A1 in Appendix A).

knowing whether the median or mean of $T|T \geq 1$ exceeds 7 or not. Indeed, in specific duration models, this predictor can be highly informative on unobserved individual characteristics. However, the quality of the prediction 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 caseworker predictor depends on the hazard in the first month. This makes it different from the predictor based on client's perceptions conditional on being unemployed for at least a month. 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 the condition that the client's motivation may be boosted in order to expect re-employment within 6 months. We do not further address this condition. We simply assume that it is implicitly taken into account by everyone.[14]

4.3 Predictor based on information from 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] For the remainder of the paper, we simply refer to $T_i = 7$ in words as "six months after the (hypothetical) interview".

Next, we construct the outcome of interest for the algorithm, by verifying whether 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 still counted as having found a job. Consequently, right-censoring of unemployment spells is not a

[14]Note that in such a situation a caseworker may make one out of two possible predictions that are each valid depending on whether such a boost is provided or not. We assume that the caseworker adopts the policy regime with the boost and the accordingly 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.

[16]We only observe dependent jobs that are liable to social security contributions. Thus, individuals who, for example, become self-employed or leave Germany, are counted as not having found a job. This may lead to a discrepancy between subjective assessments and outcomes observed in the administrative data. However, such cases should be rare in practice and it is not likely that they have a large impact on the results.

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 practice one could sidestep this layer and confine oneself to a comparison of the performance of three candidate predictors for the outcome variable of interest. In the remainder of the paper we pursue this comparison.

Random forest. Random forest is one of the best-performing off-the-shelf machine learning techniques (Biau, 2012). Several papers 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). The goal of a tree classifier is to grow a decision tree by recursive binary splitting. In each step, the classification algorithm chooses the variables and the split point that achieve the best fit. The most common criterion used for splitting nodes and pruning the tree is the Gini index, which indicates how mixed the classes are in the two groups created by a split. Then one or both of these groups are split into two more groups. This procedure continues until a stopping rule is met (Hastie, Tibshirani, and Friedman, 2011). Based on the majority vote, the classifier predicts a positive or a negative outcome. The individual trees are based on different random subsamples of the data and only a random subset of variables is used for each tree (Athey and Imbens, 2017). Thus, a random forest can be interpreted as an average of many separate tree classifiers that have all been estimated on a subsample of the data (Athey, 2017).[19]

We train our random forest models using data on our selection of unemployment entries from 2011/12.[20] As explanatory variables, we use sociodemographic charac-

[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 6.2. 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 Logistic Regression (see Subsection 6.2).

[19] 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.

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

teristics, educational background and variables relating individual employment and unemployment histories, and participation in active labor market programs, annually up to seven years back or for the elapsed lifetime.[21] For a list of explanatory variables, see Table A2 in Appendix A. Using these predictors, we predict the individual outcome for those in the survey sample, classifying all individuals with a predicted probability to find a job of $> 50\%$, as "positive".[22]

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. Furthermore, 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 substantially more training observations than in the survey sample are needed to achieve well-performing algorithms (see Subsection 6.2). 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. In our

[21] Additionally, we construct monthly labor market histories dating back up to 25 years before unemployment entry. However, this does not noticeably improve prediction performance (see Subsection 6.2).

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

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. The FPR measures the fraction of false positive classifications among all negative observations. 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 misclassifications. In that case, the above measures serve as inputs for further deliberations.

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

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

Along this line, we may also compare the explanatory 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 as measured by R^2. Furthermore, to investigate whether a combination of the three predictors improves predictions, we 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 is a popular performance measure for classification tasks.[23] 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.

[23]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) and shows how these measures change if the classification threshold is varied. The ROC-AUC Score measures overall model performance, potentially ranging 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.

6 Results

6.1 Baseline results

We start this subsection by examining the performance of the three binary predictors, where for the random forest classifier we, in fact, have a version without and a version with the caseworker assessment among the algorithm training input. As mentioned above, 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 in this subsection we examine combinations of binary predictors as well as the direct usage of the continuously distributed predictors from the self-assessments and the random forest.

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

Some 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 already 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 investigate the degree to which the different methods classify the same individuals as either positive or negative. As a preliminary step, we examine 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 the random forest. Table 3 presents the similarity in terms of prediction. The self-assessment and the machine learning approaches classify roughly 70% of the 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 return to characteristics of individuals for whom self-assessments and random forest predictors differ.

[Table 3: Identical Predictions]

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 TPR, FPR, and "Accuracy" (Figure 2).[24]

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

In terms of accuracy, the random forest models outperform self-assessment and caseworker assessment. For a default threshold of 0.5, 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%). However, though absolute differences in accuracies are sometimes sizable (e.g., a 5pp difference between the best performing random forest model and the caseworker assessment), they are not statistically significantly different from zero.

For certain objectives of policymakers and labor market administrators, the TPR and FPR rate are more interesting. The raw data showed that on average, clients are the most optimistic[25] and caseworkers the most pessimistic. As clients largely believe in finding a job within six months, the TPR resulting from the self-assessment is the highest, while the random forest classifier performs second-best and caseworker assessment leads to the worst results. Finally, caseworker-based predictions are the least optimistic, resulting in the lowest FPR. The random forest classifier leads to the second-best results while individual self-assessments perform worst.

As an alternative way to compare the performance of the predictors, we estimate linear probability models that each include just one of our predictors. As before, for

[24] We obtain 95% confidence intervals via bootstrapping. 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. 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.

[25] 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. Admittedly, the Eurozone debt crisis was slowly coming to an end, but the German economic and labor market performances had always 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.

the random forest and for the self-assessment, the individual outcome variable takes the value 1 if re-employment is the most likely prediction and 0 otherwise. For the caseworker assessment, the variable takes the values 1 if the caseworker predicted RE6=1 and 0 otherwise. Table 4 shows the results from the separate regressions by predictor. 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.

[Table 4: Predictive power of the different prediction methods, separate models]

Combining the predictors to obtain a super-predictor. As a next step, we estimate multivariate regressions, including different combinations of the three predictors (Table 5; note that here we exclude the random forest with caseworker assessments among the training input). As a benchmark, the first column repeats from Table 4 for a regression model that only includes the random forest predictor. When we add the self-assessments, 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 we include the caseworker predictor as covariate in the baseline specification with only the random forest predictor, the R^2 increases to 0.102.[26] Clearly, this increase is much smaller than when adding the self-assessment. Finally, the most interesting column (column 4) concerns a model where all three predictors are included. This further increases the R^2 (to 0.140) compared to the baseline model and compared to the models that only include two of the predictors. This indicates that combining all sources of information provides the largest explanatory power.[27]

However, note that 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 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.

[26]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). Here it is important to keep in mind that 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 some individual variation not represented in the random forest algorithm.

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

For every other constellation of covariate values, the regression actually 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.

[Table 5: Joint predictive power of the prediction methods]

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 self-assessments thus contain additional information on top of the information contained in the machine learning approach. 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. After all, 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).

It is interesting to know if this special group of individuals is different from the other individuals in terms of other observable characteristics or outcomes. Of course, many such differences will be captured by the random forest algorithm. Among the survey respondents, we consider two sources of potential additional information:

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

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

In the first case, the only potentially useful variable that we identify concerns the frequency of sickness absence in the 6-month period. In theory, 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. Thus, the individual may know of future sickness absences that will hamper a swift re-employment and (s)he may include this information when providing the RE6 self-assessment. Unfortunately, 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.

The second source of potential additional information consists of self-reported measures of risk aversion, patience and locus of control.[28] We consider again the subsample for whom the random forest incorrectly predicts RE6=1. This subsample includes 285 respondents of whom 63 correctly predict RE6=0. We apply regressions with as outcome variable an indicator of whether the self-assessment correctly predicts RE6=0 or not, and with as covariates the personality traits. Much of the variation in these traits should be captured by the random forest algorithm, especially as these traits are often seen as time-invariant among adults. We therefore control for the value of the continuously distributed random forest prediction (classification fraction) in the regression. We find that of the three traits, only risk aversion matters. 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.[29]

It may be beyond the scope of the current study to explain why risk aversion has this effect while patience and locus of control do not have effects. Notice that the effect of risk aversion concerns events in the upcoming 6 months that cannot be explained by observed covariates and individual histories. The notion that the value of information is uniformly larger for risk averse individuals has been refuted in the literature but in many cases this notion is correct (see e.g. the overview study by Willinger, 1989). Risk averse individuals may therefore have a larger incentive to acquire information about future events, thus enabling them to provide a more accurate self-assessment.

Whether the super-predictor can be used in practice depends on the circumstances, because self-assessments are not routinely available. As mentioned above, the survey sample is too small to be divided into training and test data sets, so the super-predictor lacks external validation. Instead, our result can be seen as providing a motivation to collect self-assessment information routinely. With a sufficiently large database along these lines, self-assessments can be incorporated as inputs for the random forest classifier. We return to this in the concluding section of the paper. Alternatively, interviews may be used to identify observable conditions in which the self-assessed predictor of long-term unemployment dominates the predictor based on machine learning algorithms. We have seen that self-reported risk aversion is such a marker. It may be useful to find out what other features and events lead to

[28] 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.

[29] In this subsample, the occurrence of observed sickness absence in the relevant first 6 months of unemployment is not correlated to whether the self-assessment correctly predicts RE6=0 (whether one controls for the continuously distributed random forest fraction or not).

such predictions, and whether participation in some special ALMP can boost their re-employment chances.

Discrepancies between individual self-assessment and random forest prediction. Extending the analysis of the super-predictor in the above paragraphs, we shed light on discrepancies between the individual self-assessment and the random forest predictor more in general. Specifically, we consider the full group of individuals for whom the random forest prediction is incorrect and examine if the subgroup with a correct self-assessment is systematically different from its counterpart. Note that this extends the above analysis because it also includes individuals whose correct self-assessment does not lead to a correction on the random forest predictor in our super-predictor. The size of the subsample is now 401, of whom 171 correctly predict the value of RE6.

We again use the self-reported measures of risk aversion, patience and locus of control, and we again control for the value of the continuously distributed random forest prediction (classification fraction). The results are qualitatively identical to those for the personality traits in the previous paragraphs. 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.

Shifting the classification thresholds for binary predictors. The results presented so far are based on specific threshold values of 50% to classify individuals. However, this value is not always compelling. Consider, first, the random forest predictor. Viewed in isolation from the other predictors, the prediction performance of the random forest may actually be larger for a different threshold.

To proceed, Figure 3 presents the accuracy for all possible relevant threshold values. It turns out that the highest random-forest accuracy 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.[30] This is confirmed by the ROC-AUC Score (see Figure A2 in Appendix A).

[Figure 3: "Accuracy" for different models across various threshold values]

[30]Figure 3 also allows us to consider the performance of the models in terms of *maximum* accuracy, where the thresholds may differ for each predictor. Here we find that the ranking of the predictors is the same as for the threshold of 0.5: we find highest maximum accuracy for the random forest model without information on caseworker assessment (65.5% for a threshold of 0.54), followed by 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%).

Next, we consider shifting the threshold value for the self-assessments. In Subsection 4.1 we proposed taking over-optimism into account 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 3 shows that this leads to an accuracy that is only marginally below the highest attained accuracy of 64.8% across all thresholds for this predictor. At the same time, the accuracy is also close to the 63% value in the case of a 0.50 threshold. Thus, the correction does not appear to be very useful. 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.[31]

Continuously distributed random forest predictor and self-assessment. Above, when creating the super-predictor, we applied regressions on binary predictors which are themselves based on the random forest and the self-assessments. Specifically, these binary predictors are "rounded-off" versions of the underlying classifier fractions and the self-reported re-employment probabilities, respectively. The latter two are 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 4. Hence, the continuously distributed predictors dominate their binary versions. Notice also that the machine learning model including caseworker assessments outperforms the one not including those assessments, in terms of R^2 (0.148 with caseworker assessments vs. 0.134 without).

In Table A5 we regress RE6 jointly on the continuously distributed predictors and the binary caseworker assessment. The resulting R^2 are considerably higher than in Table 5. Subsequently, we derive a new super-predictor from the final column. The accuracy of this equals 67.4% which exceeds the value of 66.2% for the earlier super-predictor. This enhanced super-predictor is somewhat less transparent than the earlier version because the variables on the right-hand side span a wider range of values than before.

TPR and FPR measures. Figure 4 displays the TPR and FPR for each predictor across all relevant thresholds. As the caseworker assessment is captured in a single binary indicator, the resulting TPR and FPR are fixed across all thresholds. Therefore, we use the caseworker TPR and FPR as a benchmark for the comparisons.

[Figure 4: TPR and FPR across thresholds]

[31] 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 5. The accuracy is virtually indistinguishable from the value for the original super-predictor (66.4% instead of 66.2%).

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: 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.[32] We can also do the exercise 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). Thus, in terms of these comparisons, caseworker predictions also tend to perform worse than the random forest classifiers and the self-assessment. This conclusion does, however, not hold over the entire range of potential thresholds.

For our sample, this section so far has thus shown that for a broad range of potential thresholds, random forest classifiers and self-assessment out-perform predictions based on caseworker profiling in terms of accuracy. If one is, however, rather interested in a very high TPR or in a very low FPR, for a given threshold one might well conclude that self-assessments or caseworker assessments should be preferred compared to machine learning methods.

Subgroup analyses. d'Haultfoeuille, Gaillac, and Maurel (2021) provide evidence that whether beliefs about future earnings correspond to rational expectations differs across subgroups of individuals, for example by education level. With this in mind, we perform predictions by age, gender and education, to analyze potential heterogeneities in the performance of the predictors. Figures 5 and 6 display results by age and gender. Based on a threshold of 0.5, we observe that job seekers as an aggregate are over-optimistic in all sub-groups, regarding RE6. For example, the expected success rate among younger unemployed is 88% compared to an actual rate of 58%, while for the older unemployed the expected success rate corresponds to 71%, compared to an actual rate of 48%. In line with the results for the full sample, the average predicted share based on the random forest lies for all sub-samples between the actual share of unemployed that found a job within 6 months and the predicted shares based on the self-assessment of the unemployed. While the predicted shares based on caseworkers' assessments are the closest to the actual shares for all sub-groups, we observe some heterogeneity in the sign of the difference. Figure

[32]For the assessment by the survey participants, there is no FPR that exactly matches the FPR of the caseworkers. The same is true for the TPR.

7 displays results by type of education. This should be interpreted with caution due to 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") frack.

Turning to "accuracy", the finding that caseworkers are worse at predicting RE6 than clients or random forest models holds for all sub-groups. All three predictors perform better for older than for younger workers, and better for vocationally educated workers than for more generally educated workers. The poorer performance of caseworker predictions is most evident for the latter group. The greatest advantage of the random forest classifier is observed among older workers, males and vocationally educated workers, whereas females, younger workers and workers with a more general educational background achieve a similarly high or higher accuracy for self-assessment compared to random forest models. The finding that females are on average less over-optimistic than males, in comparisons of individual predictions and realizations of economically relevant events, has been documented in the literature; see e.g. Bjuggren and Elert (2019).

6.2 Robustness checks and further analyses

First, we check whether the results are robust to changes of the outcome definition. The main definition states that the individual has found a job within six months after the interview, not conditional on whether the person is still employed six months after the interview. As an alternative, we analyzed whether the results change if we define the outcome as having a job exactly six months after interview (see Figure A3). This hardly affects the results, except for the fact that for this definition, the random forest model with caseworker information has the highest accuracy. Besides this, our main conclusions remain unchanged.

Second, regarding the caseworker assessments, our main definition states that we only include individuals with profiles that are no older than a year (with 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 resemble the main results quite closely. However, interestingly, the accuracy of the caseworker assessments is slightly lower when only using more up-to-date profiles.

Third, we examine the role 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). In order to check whether these restrictions have an impact on prediction performance, we repeated our analyses without these restrictions (see Figure A5). Except for the fact that the machine learning model including caseworker information achieves the highest maximum accuracy, the results from this exercise are also very close to the main results.

Fourth, to investigate whether more detailed information on individual labor market histories can improve the predictive power of the machine learning models, we additionally constructed monthly employment, unemployment, and active labor market 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 models not including and including the caseworker assessment as input respectively.[33] 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 seem improve the predictive performance of the random forest classifier to a large degree compared to our main set of explanatory variables.[34]

Fifth, while the random forest classifier is a "black box" in terms of the predictive power of individual predictors, it is nevertheless interesting to see which sets of variables contribute to its ability to predict reintegration into the labor market. To investigate this, we present the ROC-AUC scores for models using different sets of predictors in Figures A8 and A9.[35] We see that, compared to the model that only includes 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

[33]Note that, 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.

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

[35]None of the models in Figure A8 contain caseworker information as a training input. All models in 9 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.

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 (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).

Sixth, we investigate how much training data 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 models with increasing training set sizes do not show systematic improvements 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.

Seventh, to investigate whether we can improve performance using different classification methods, we repeated our prediction exercises using a gradient boosting classifier and a logistic regression classifier using our main set of variables. Figure A11 shows the overall model performance, measured as the ROC-AUC score, for these classifiers compared to the random forest classifier. While the overall performance of the logistic regression classifier is considerably worse, the ROC-AUC score of the gradient boosting classifier is almost identical to that of our main random forest classifier. However, despite a similar overall performance of the random forest and the gradient boosting classifier, we achieve the highest maximum accuracy using the random forest classifier (see Figure A12).[36]

Eighth, we explore the question whether models using contemporary data achieve better results than using past data by training our models using administrative data from 2012/13 instead of data from 2011/12. If this does indeed improve the performance of our models, then this could imply that one should include leading economic indicators to improve the predictions. Figure A13 shows that the results in terms of accuracy across classification thresholds are fairly close to our main results. To further investigate this issue, we also present the ROC-AUC scores for a holdout sample of the administrative data and for the survey data, again using the administrative data from 2012/13 as training input (see Figure A14. Compared

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

to the results presented in Figure A2, we see that the ROC-AUC scores improved, 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.

Finally, we explore the question whether we can use machine learning to predict the caseworker assessments. If we are able to predict this perfectly, one could argue that caseworkers do not incorporate additional information not captured by the machine learning algorithms and could therefore be replaced entirely. Figure A15 shows the results for this exercise. We see that in terms of accuracy and ROC-AUC score, the machine learning algorithm provides better predictions than the one trained to predict RE6. However, the prediction is far from perfect, suggesting that the predictions of the caseworkers incorporate (unobserved) information in their decision process that is not available in the administrative data.

7 Conclusion

Machine learning predictors based on random forests perform better in terms of accuracy than predictors based on client self-assessments or caseworker assessments. This is a robust finding in our analyses using extensive administrative data to train the algorithm. In addition, the machine learning predictor provides a reasonable balance between a high TPR and a low FPR. Nevertheless, we identified cases in which the other predictors were not dominated by the machine learning predictor. For example, among females, self-assessments score a slightly higher accuracy. In most other cases, the self-assessment predictor performs almost equally well as the random forest predictor.

"Accuracy" is a comprehensive prediction performance measure, while TPR and FPR focus on more specific types of misclassification. 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. For TPR, self-assessment actually outperforms machine learning (and both outperform the caseworker assessment).

Conversely, 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. For FPR, the TPR ranking of predictor performances is reversed, so the caseworker assessment outperforms machine learning, and both strongly outperform self-assessment. In sum, if there are overriding reasons to prefer TPR (FPR) then the self-assessment (caseworker assessment) is the best predictor. Without such overriding concerns, machine learning prediction seems preferable.

In the paper we developed a correction on the self-assessment predictor by taking over-optimism into account. However, this only slightly improves the prediction performance. In otehr settings the procedure may provides 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.

The paper provides a number of additional results and insights. First, combining predictors increases the predictive power. A regression-inspired super-predictor, predicting short-term unemployment if and only if both the random forest and the self-assessment predict short-term unemployment, attains a higher accuracy than its components. This reflects the fact that subjective expectations of the client contain information about future events that is not captured by the administrative data. It appears that this is particularly relevant for risk averse individuals. Moreover, along some dimensions, the random forest benefits from the inclusion of caseworker assessments. The performance can be further enhanced by using the underlying continuously distributed predictors from the machine learning and/or the self-assessments. The results are qualitatively insensitive to changes in specificities of the outcome variable.

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 less than 50 individual labor market history variables suffices for the random forest. Here it is useful to include annual individual history variables dating back 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 a maximum of 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 better performance than less recent ones.

From a practical point of view, the results in the paper lead to recommendations for obtaining good individual-level predictions. This starts with the application of random forest algorithms trained on administrative data. These should be routinely made available upon clients' entry into unemployment. The algorithm may include

the caseworker's own subjective assessment as input. It is advisable to obtain a self-assessment of the new client as well and use this to further improve the overall prediction. Here we think it is important that the question is not asked by the caseworker or another representative of the employment agency, to avoid desirability bias and strategic responses. After a while, the responses can then be used to include the client's self-assessment as input into the random forest algorithm. The algorithm should be regularly updated.

Other topics for further research include the extent to which the caseworker assessment aids in predicting more extreme unemployment duration values. 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. Yet other topics concern a randomized controlled trial to investigate the usefulness of the availability of machine learning predictions, and the extent to which the size of the UI agencies' budget depends on incorrect predictions.

References

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.

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.

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.

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.

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

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*.

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

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

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," *the Journal of Machine Learning Research*, 12, 2825–2830.

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*.

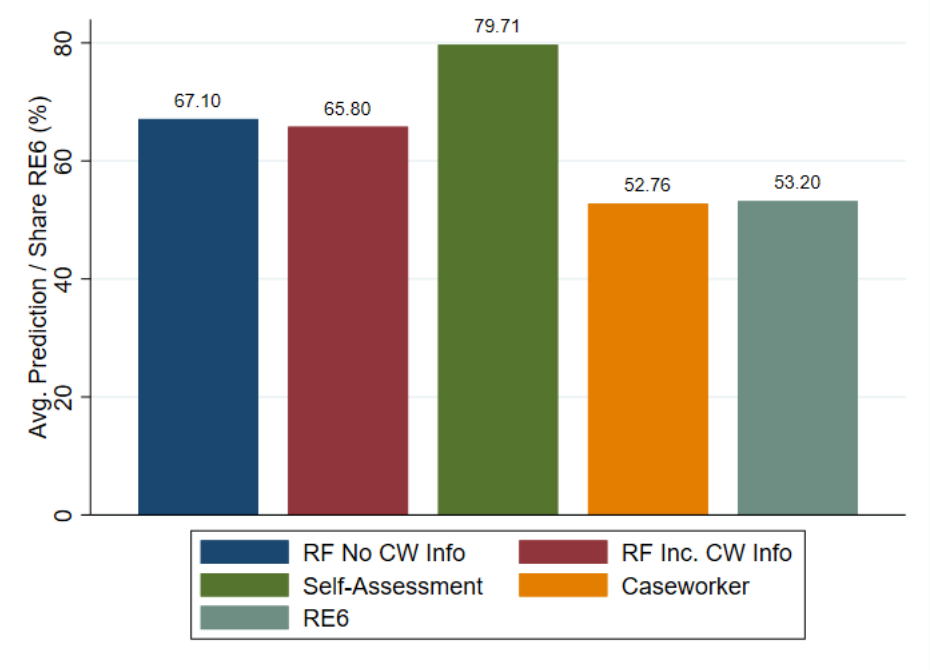
van Landeghem, B., S. Desiere, and L. Struyven (2021): "Statistical profil-

ing 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.

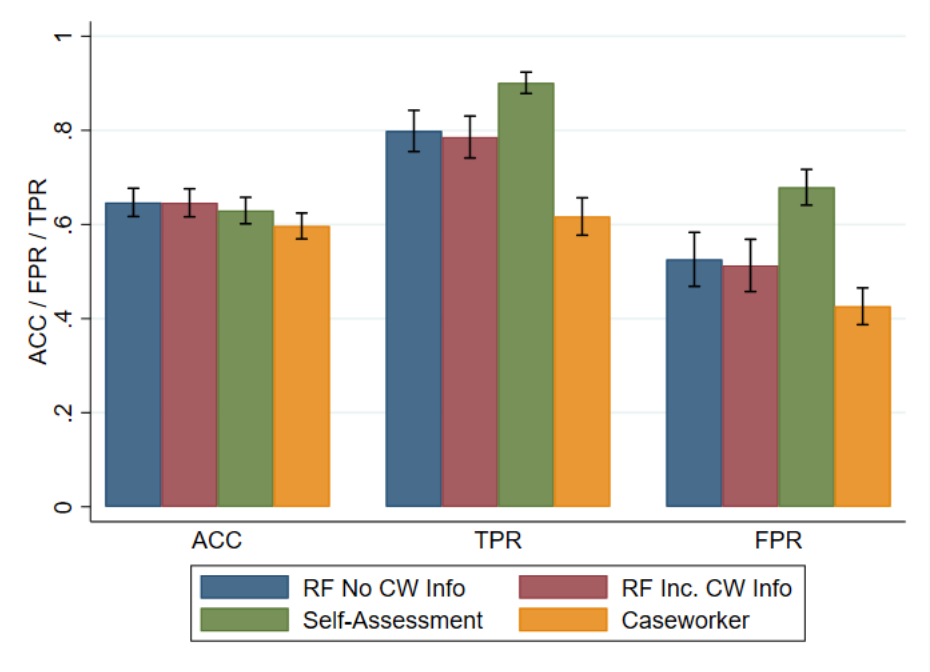
Figures and Tables

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



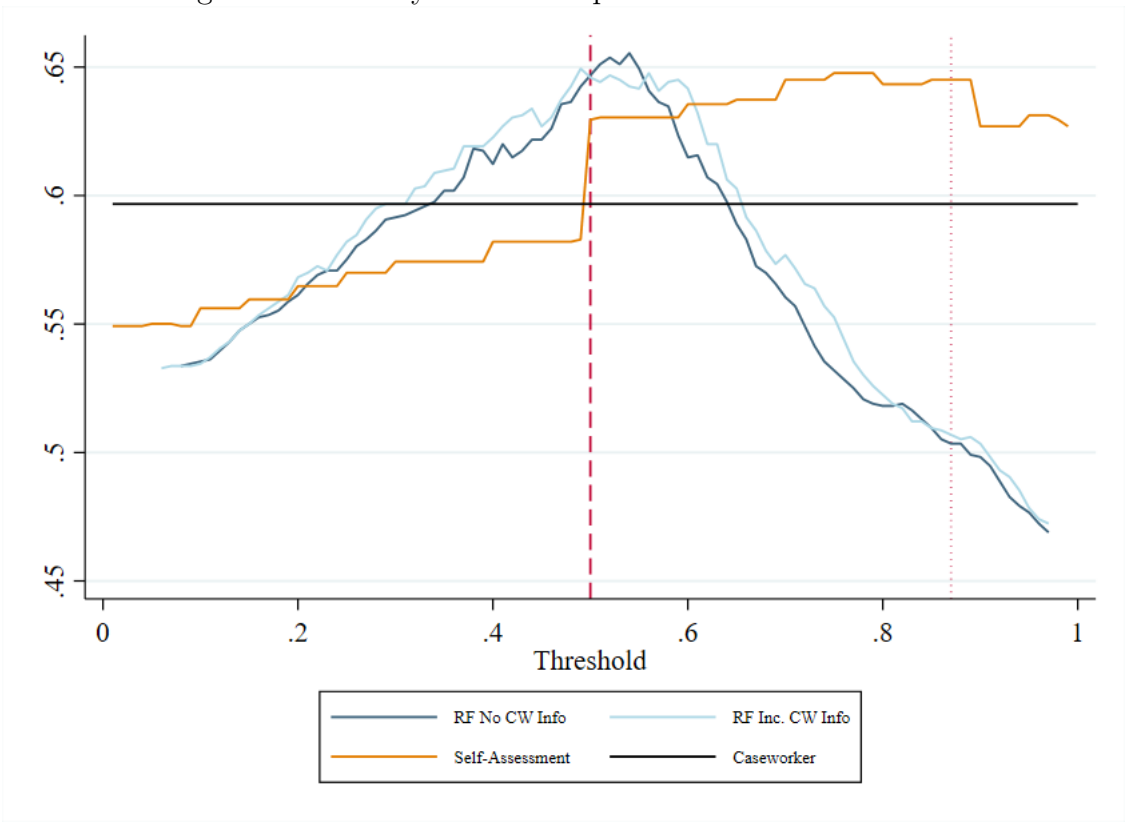
Note: This figure shows, for the survey sample, the average predicted share finding a job for each of the four predictors and the actual share of individuals finding a job within 6 months. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest 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



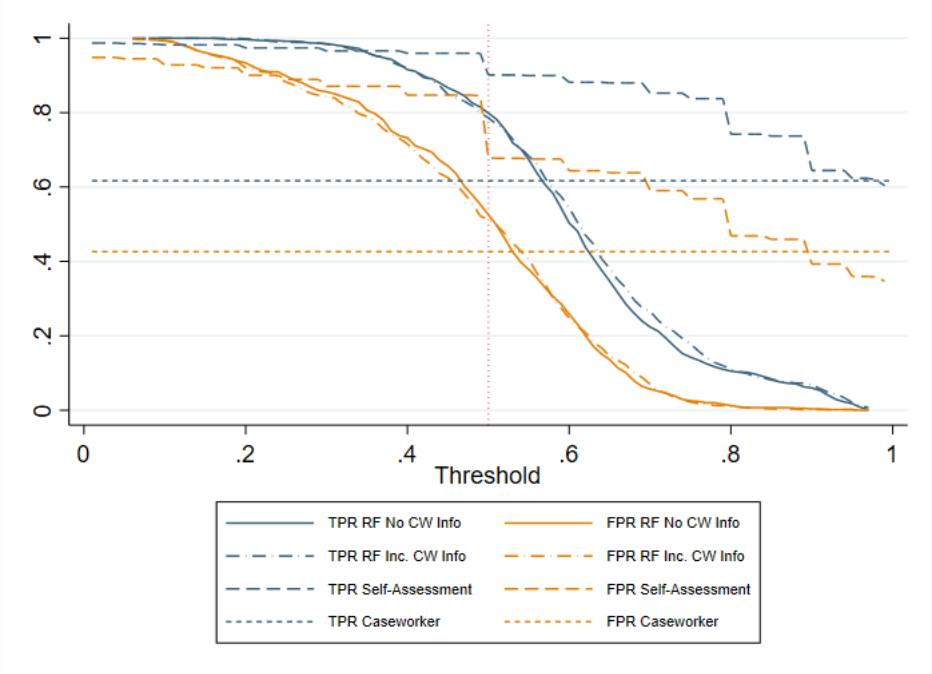
Note: This figure shows the Accuracy (ACC), False Positive Rate (FPR), and True Positive Rate (TPR) for the survey sample. 95% confidence intervals are obtained via bootstrapping (500 bootstrap samples of the administrative data from 2011-12 to train the random forest classifiers and 500 bootstrap samples 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: Accuracy of the four predictors across thresholds



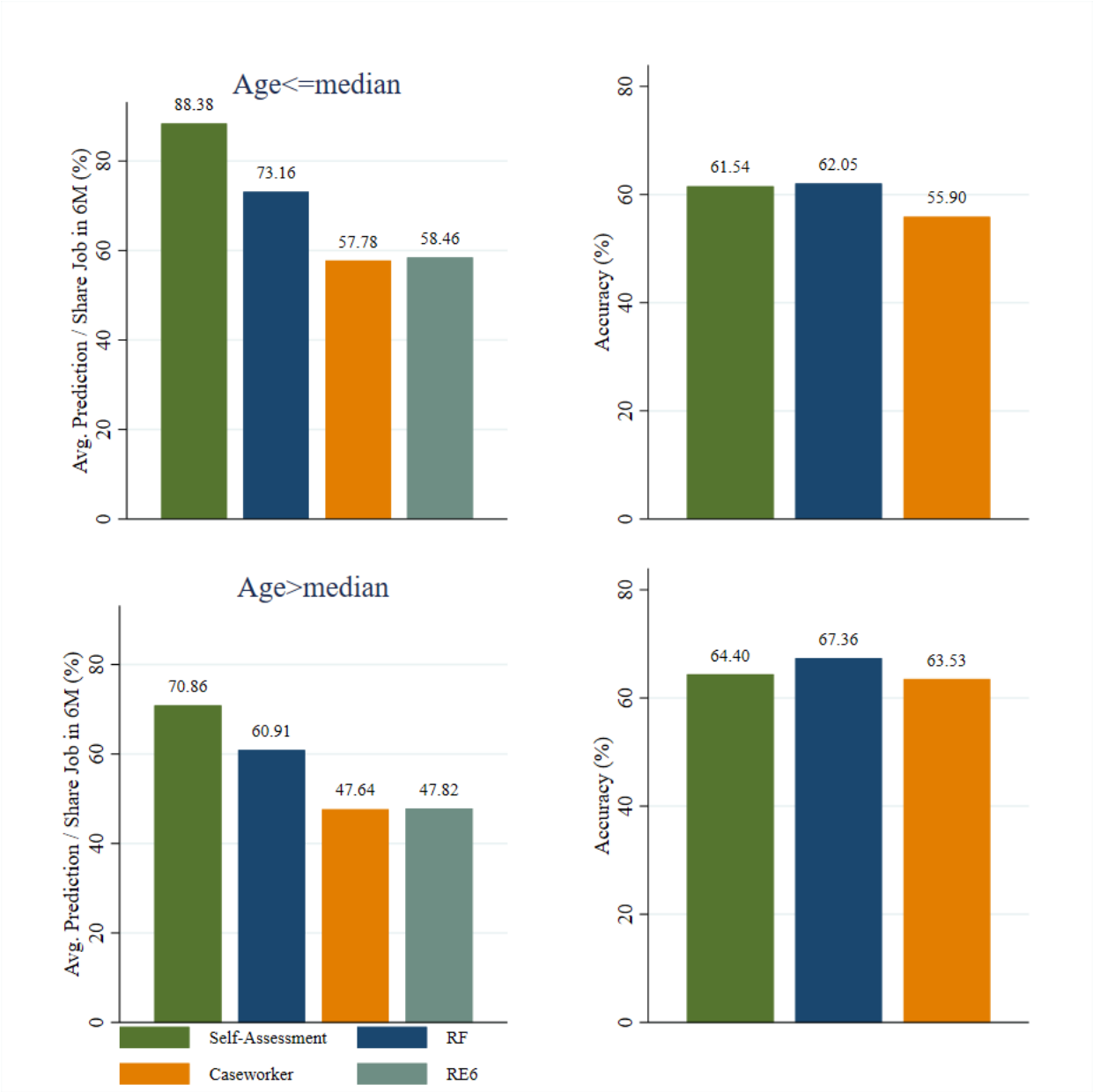
Note: This figure shows the accuracies of the four different predictors across all relevant classification thresholds for the survey sample. The dashed red line marks the (default) 50% threshold. The dotted red line marks the threshold at which the fraction of individuals that classify themselves as 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 4: TPR and FPR of the four predictors across thresholds



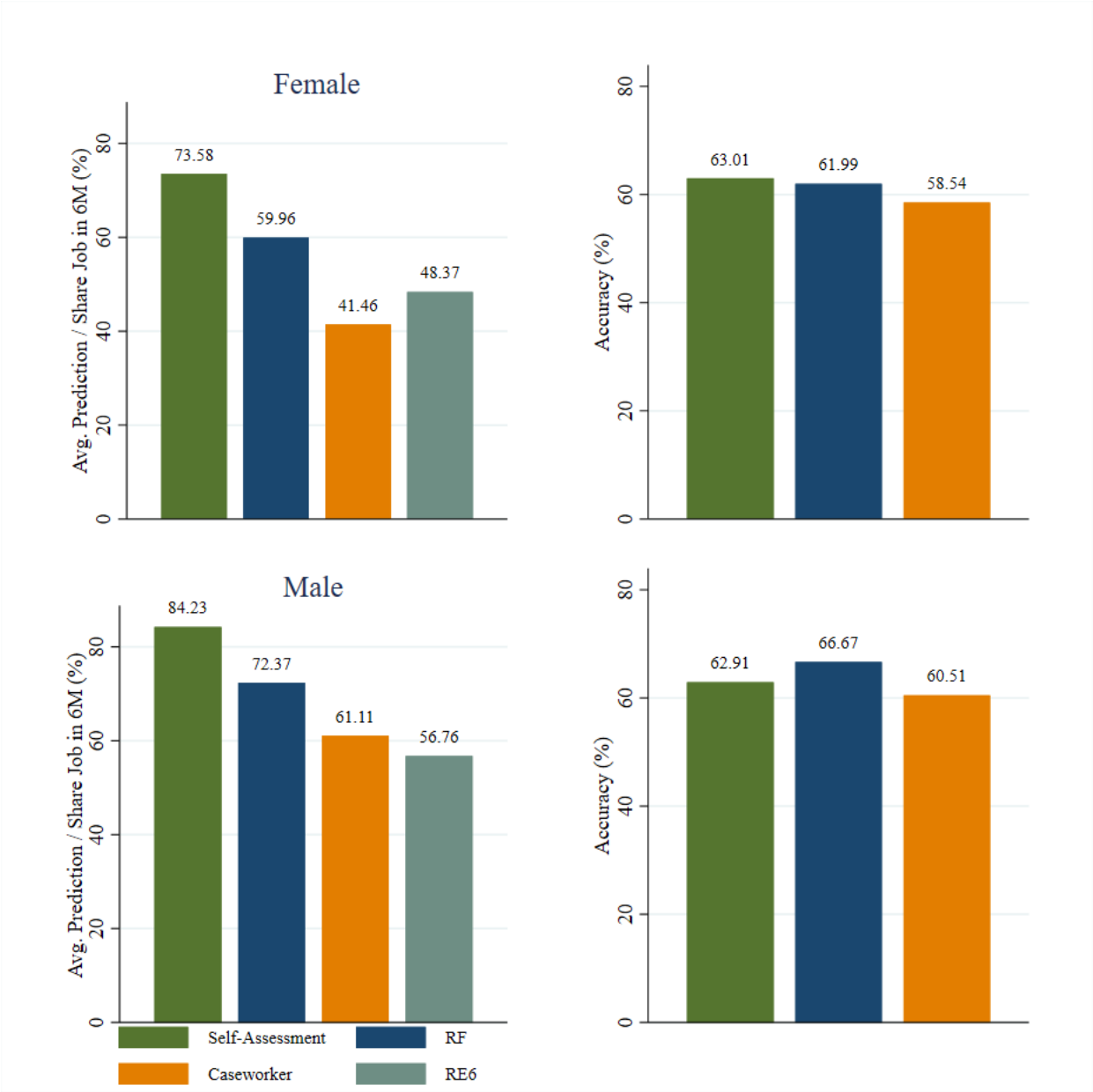
Note: This figure shows the True Positive Rate (TPR) and the False Positive Rate (FPR) of the four different predictors across all relevant classification thresholds for the survey sample. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest 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.

Figure 5: Predictions and accuracies by age group



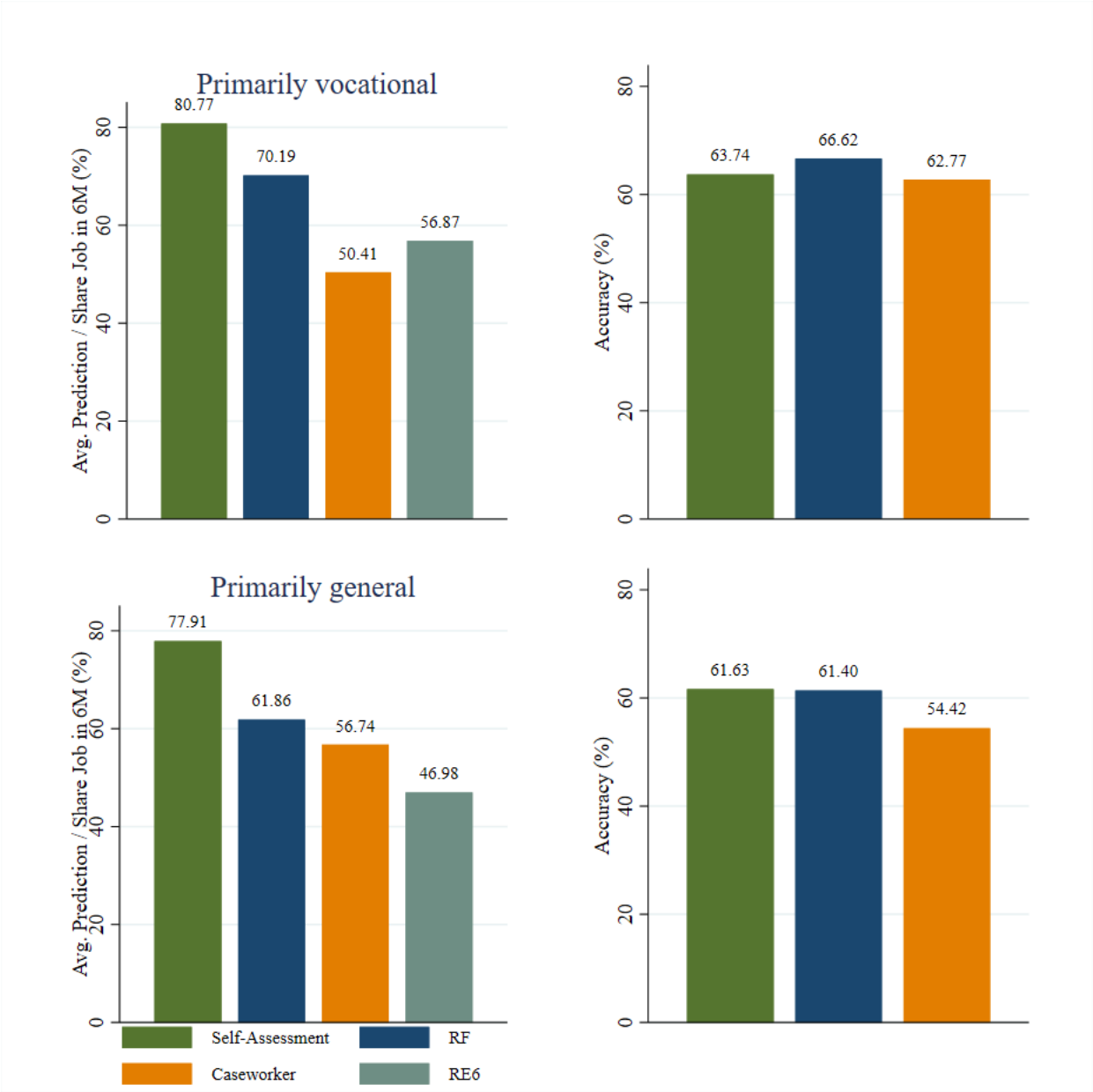
Note: This figure shows the average of positive predictions by age group (left panels) and the corresponding accuracies (right panels). The median age is 43. 585 (50.52%) individuals in our sample are up to 43 years old, 573 (49.48%) individuals are older than 43. The random forest results presented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

Figure 6: 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 7: 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 Subsection 6.1 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.

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%
Voc. Training	85.40%	85.01%	90.24%
University	21.78%	22.51%	21.93%
Daily Wage Last Job	64.24	66.82	67.40
Total Earn. Reg. Empl. Last Year	15365.16	16287.46	16577.26
Tot Dur. Reg. Empl. Last Year	231.68	236.55	247.09
Tot. Dur. Unemp. Last Year	57.01	56.38	52.43
N	29,130	30,255	1,158

Source: This table shows 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). EVA; IEB v.12.01.00.

Table 3: Share of identical predictions

	RF (no CW info)	RF (incl CW info)	Self-Assessed	CW Assessment
RF (no CW info)	100%			
RF (incl CW info)	90%	100%		
Self-Assessment	69%	69%	100%	
CW Assessment	60%	69%	60%	100%

Note: This table shows the share of individuals that get classified into the same category by the different predictors.
Source: EVA; IEB v.12.01.00.

Table 4: Explanatory power of different predictors - separate models

	(1)	(2)	(3)	(4)
VARIABLES	RE6	RE6	RE6	RE6
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-squared	0.084	0.082	0.076	0.036

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 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. For the caseworker assessment, the predictor takes the value 1 if the caseworker predicted reintegration within six months 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 5: Explanatory power of different predictors - combined models

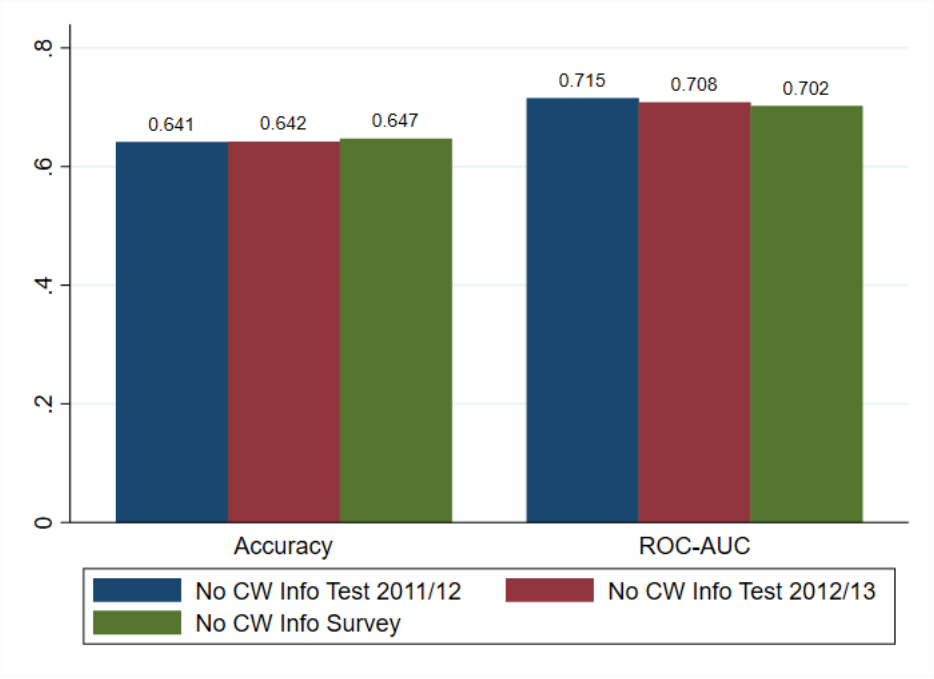
	(1)	(2)	(3)	(4)
VARIABLES	RE6	RE6	RE6	RE6
RF Pred. (no CW info)	0.308***	0.254***	0.278***	0.237***
	(0.0296)	(0.0305)	(0.0305)	(0.0312)
Self-Assessment		0.274***		0.251***
		(0.0336)		(0.0345)
CW Assessment			0.137***	0.101***
			(0.0289)	(0.0294)
Constant	0.325***	0.143***	0.273***	0.119***
	(0.0240)	(0.0283)	(0.0261)	(0.0281)
Observations	1,158	1,158	1,158	1,158
R-squared	0.084	0.130	0.102	0.140

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 (3) additionally includes the prediction from the caseworker assessment. Model (4) includes all three predictions. 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%. The caseworker predictions take the value 1 if the caseworker predicted reintegration within six months. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: EVA; IEB v.12.01.00.

Appendix: For Online Publication

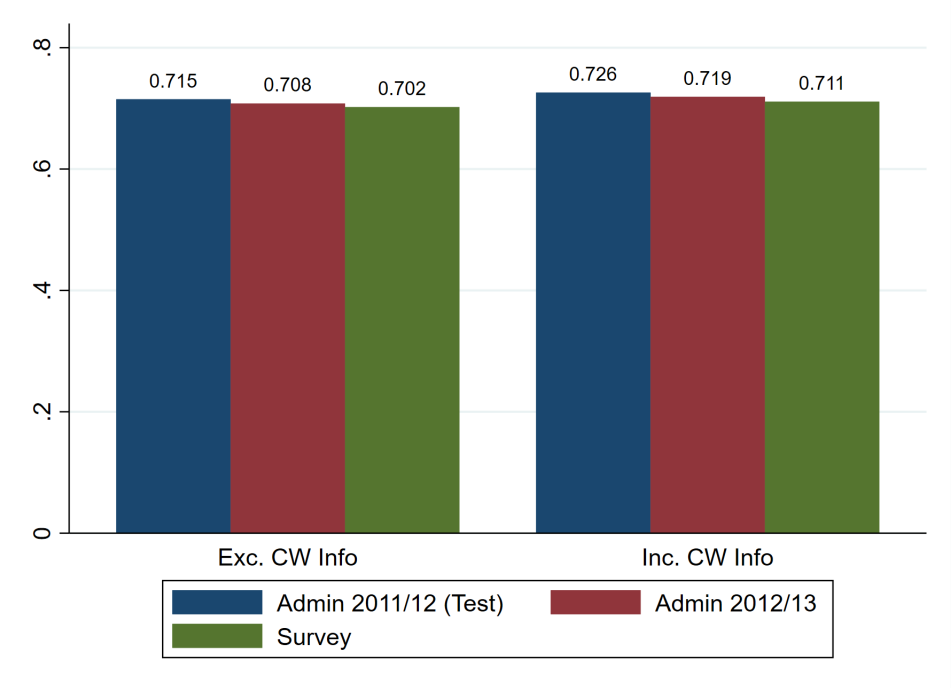
Appendix A. Figures and Tables

Figure A 1: Accuracy and ROC-AUC



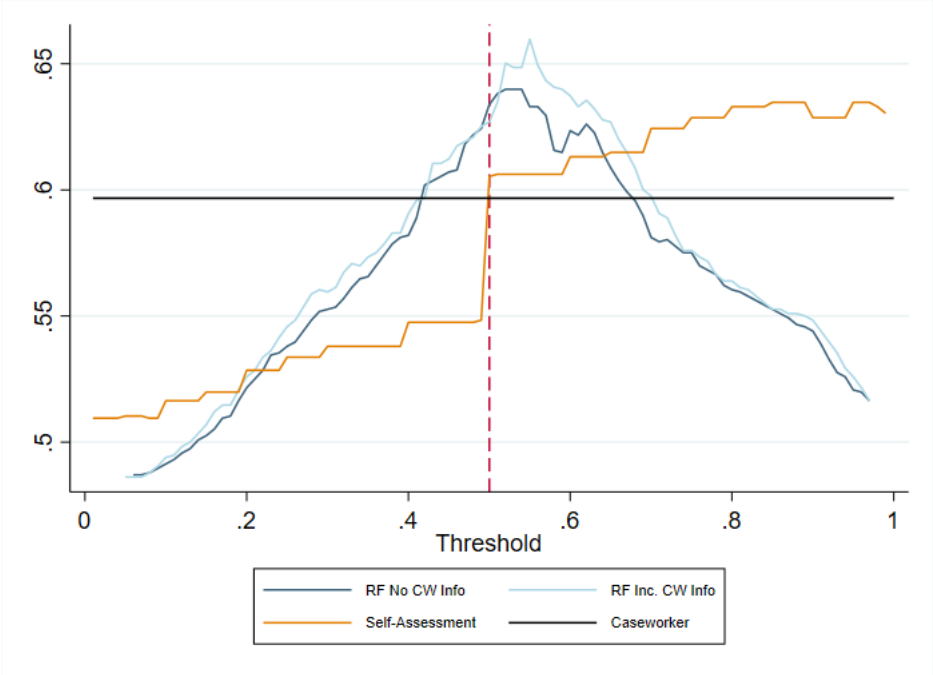
Note: This figure shows the Accuracy and ROC-AUC of the machine learning model for the admin data 2011-12 and 2012-13 and for the survey sample. The accuracies are measured at a classification threshold of 0.5. Source: EVA; IEB v.12.01.00.

Figure A 2: ROC-AUC Scores with and without caseworker info



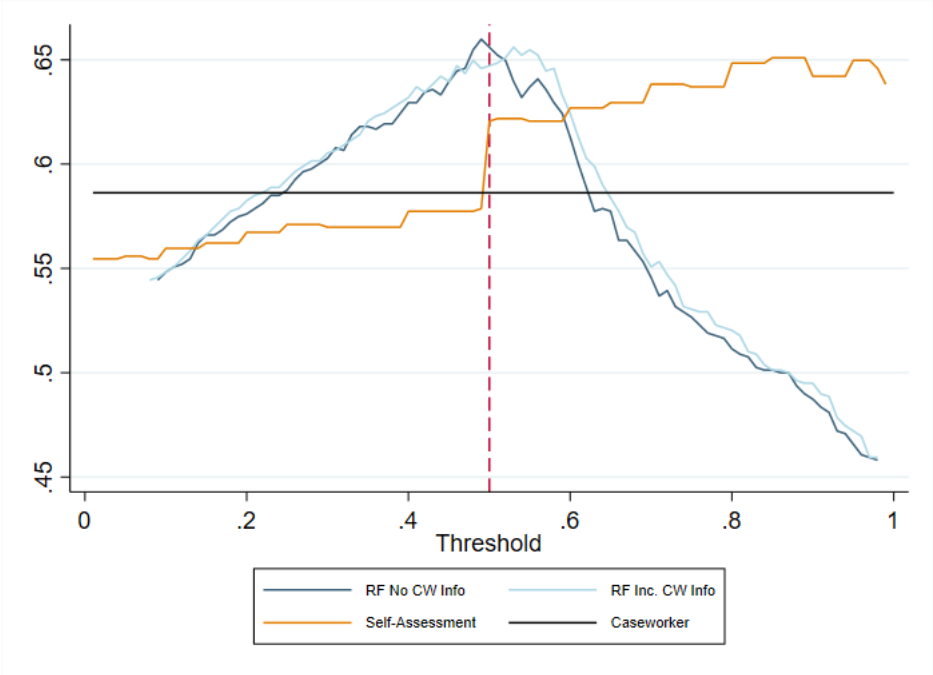
Note: This figure shows the ROC-AUC Scores of the machine learning model for the admin data 2011-12 and 2012-13 and for the survey sample. The first set of bars shows the results excluding the caseworker assessments as input for the machine learning model and the second set of bars shows the results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



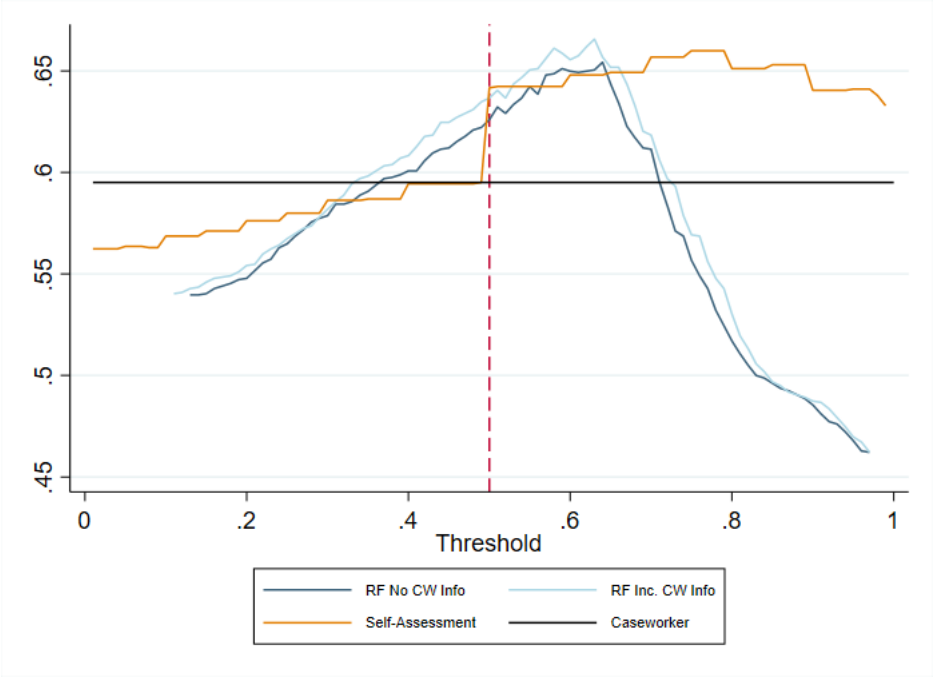
Note: This figure replicates Figure 3, but with the outcome defined as having a job six months after the interview instead of finding a job within six months. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as 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)



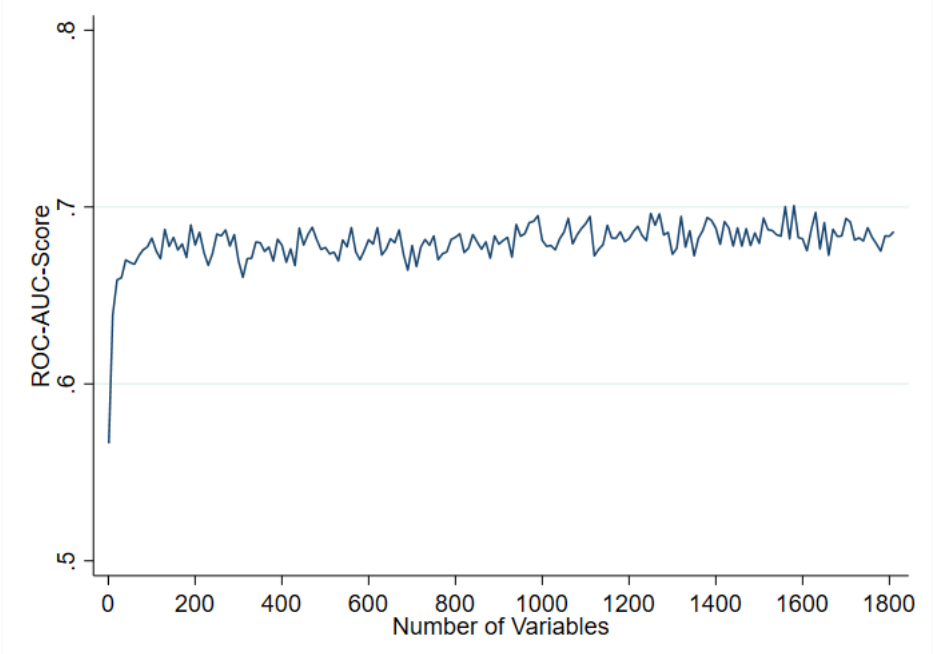
Note: This figure replicates Figure 3, but only with individuals with profiles not older than 6 weeks (instead of one year). RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = 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)



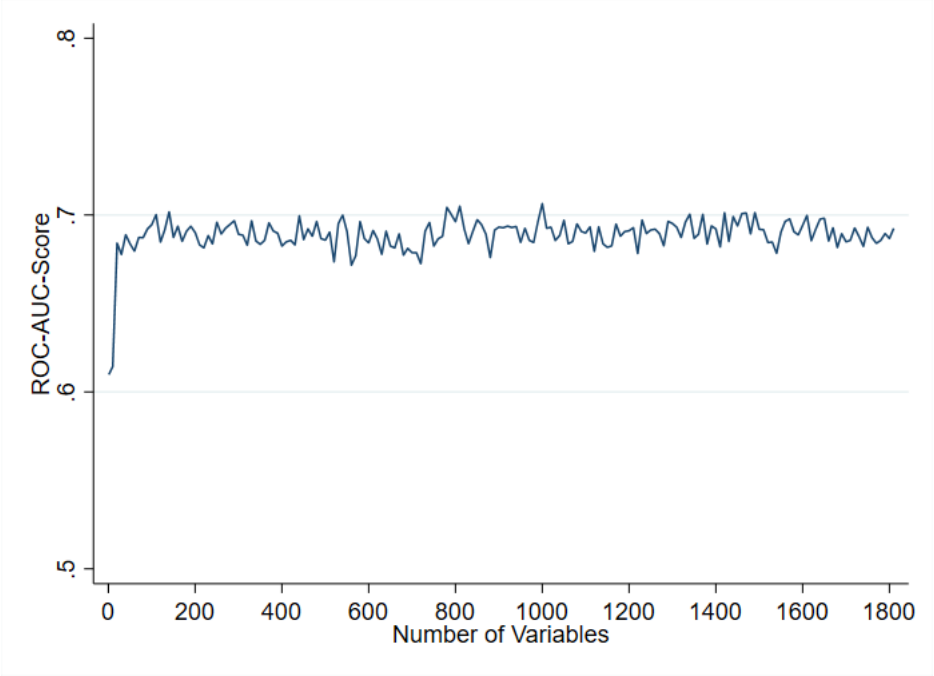
Note: This figure replicates Figure 3, but for a wider sample that is constructed using less strict sample definitions. RF No CW Info = random forest without caseworker assessments as training input, RF Inc. CW Info = random forest with caseworker assessments as training input, Self-Assessment = self 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



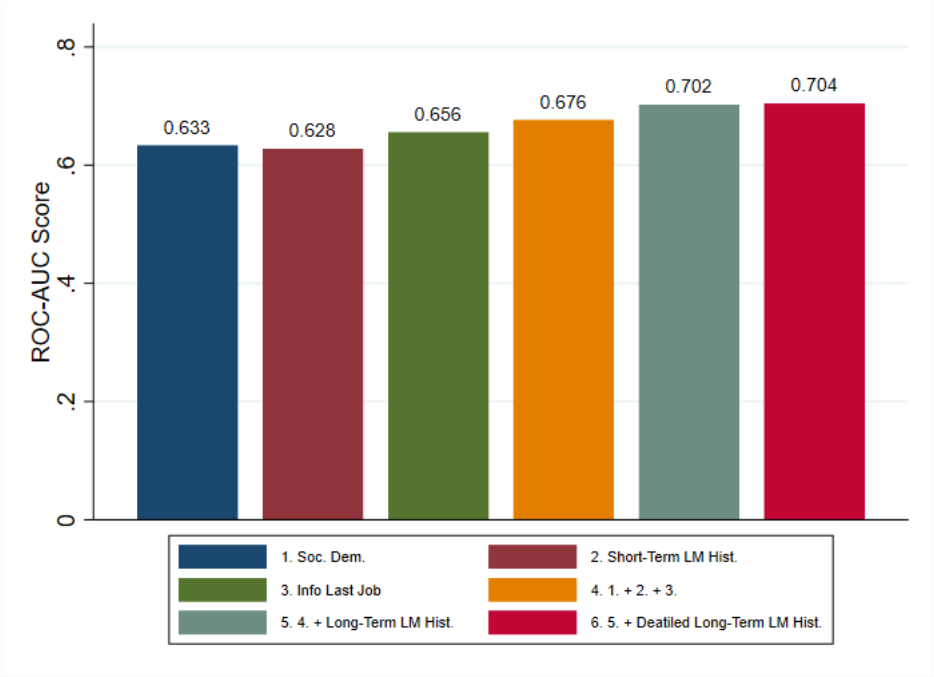
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of variables available for prediction (no additional model tuning). Source: EVA; IEB v.12.01.00.

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



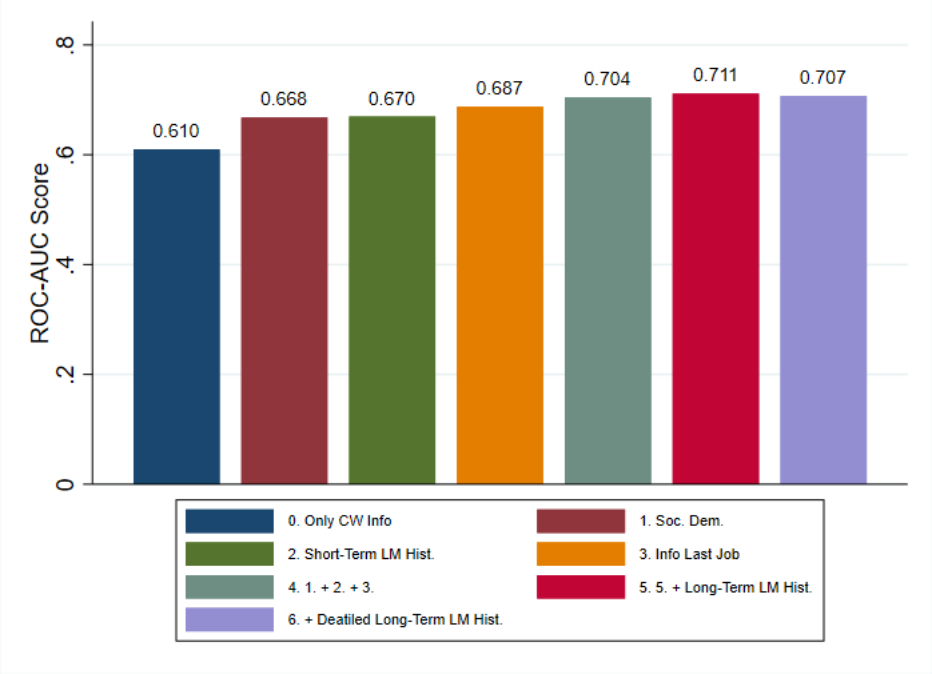
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of variables available for prediction (no additional model tuning). Source: EVA; IEB v.12.01.00.

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



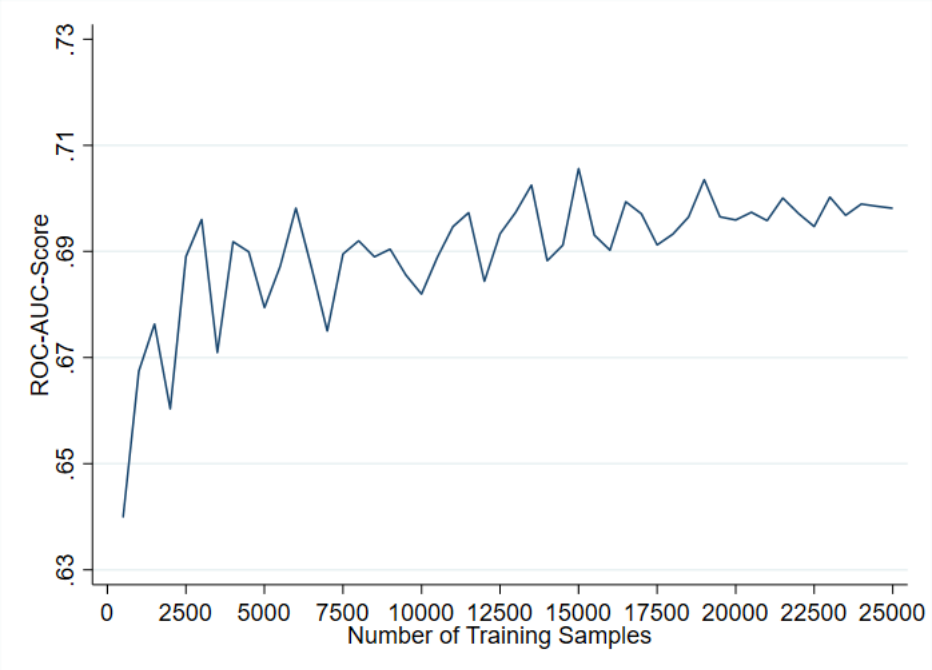
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample for different sets of variables included in the prediction algorithm. The models do not include caseworker assessments. Source: EVA; IEB v.12.01.00.

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



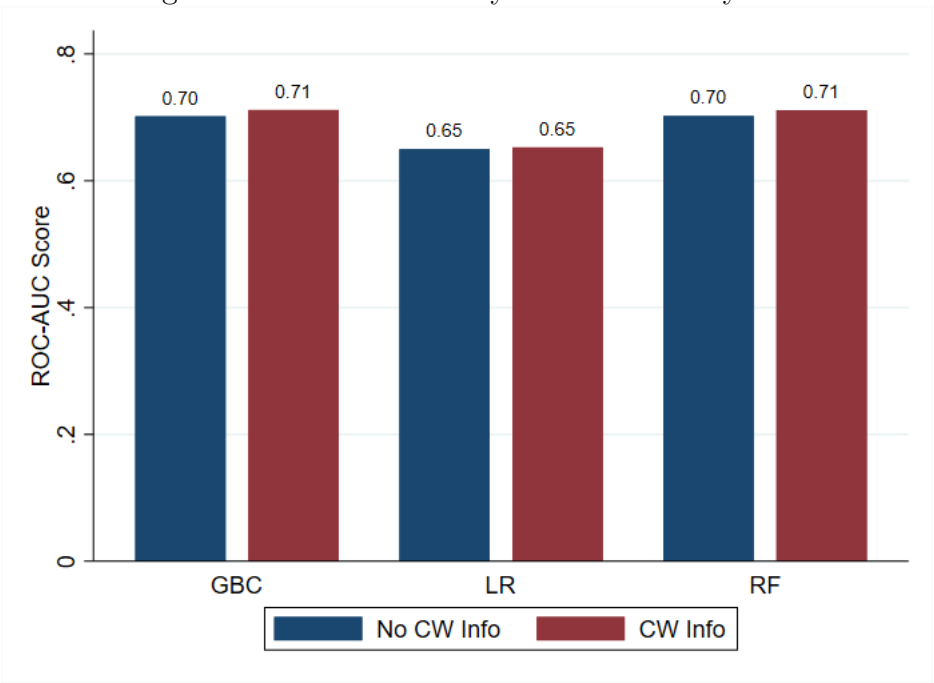
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample for different sets of variables included in the prediction algorithm. All of the models include caseworker assessments. Source: EVA; IEB v.12.01.00.

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



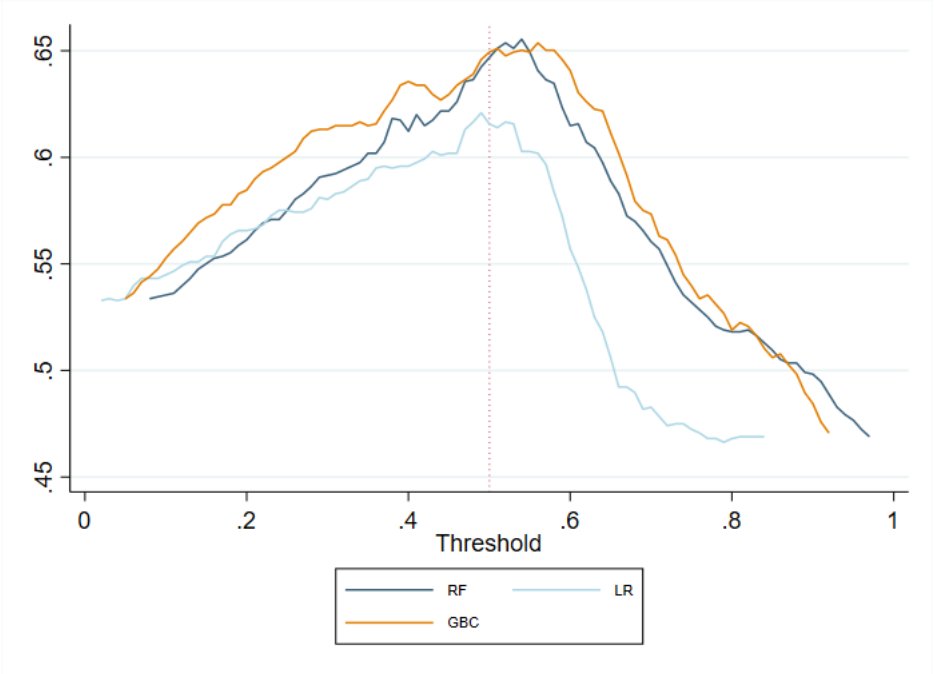
Note: This figure shows the ROC-AUC Scores of the random forest model for the survey sample when we increase the number of training samples available to train the algorithm (the algorithm was trained using a random subset of the admin data from 2011-12, as in the main specification). Source: IEB v.12.01.00.

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



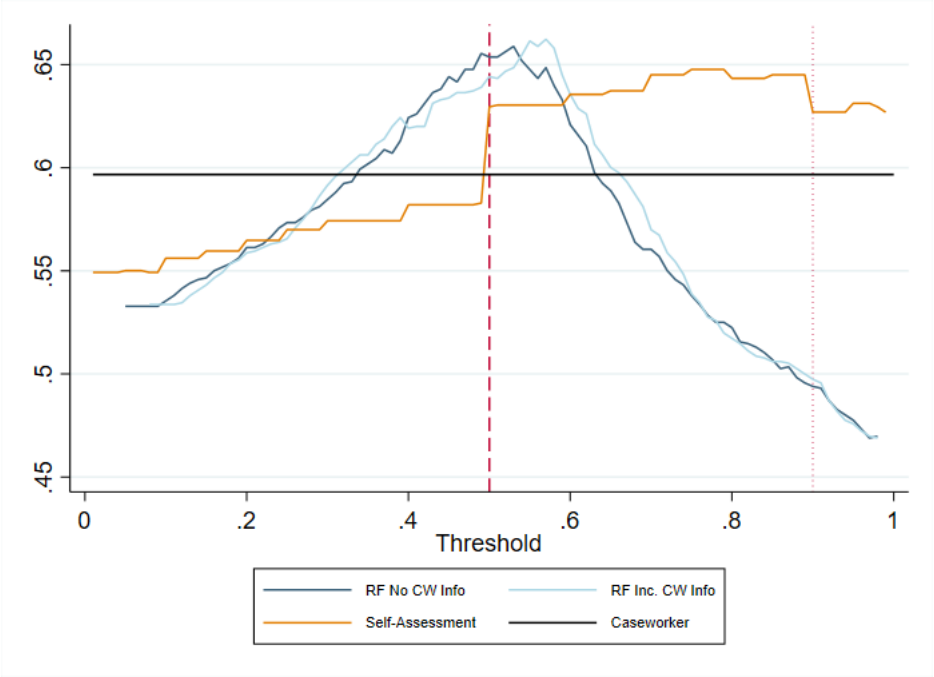
Note: This figure shows the ROC-AUC achieved by different classification methods for the survey sample. GBC = Gradient Boosting Classifier; LR = Logistic Regression; RF = Random Forest. EVA; IEB v.12.01.00.

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



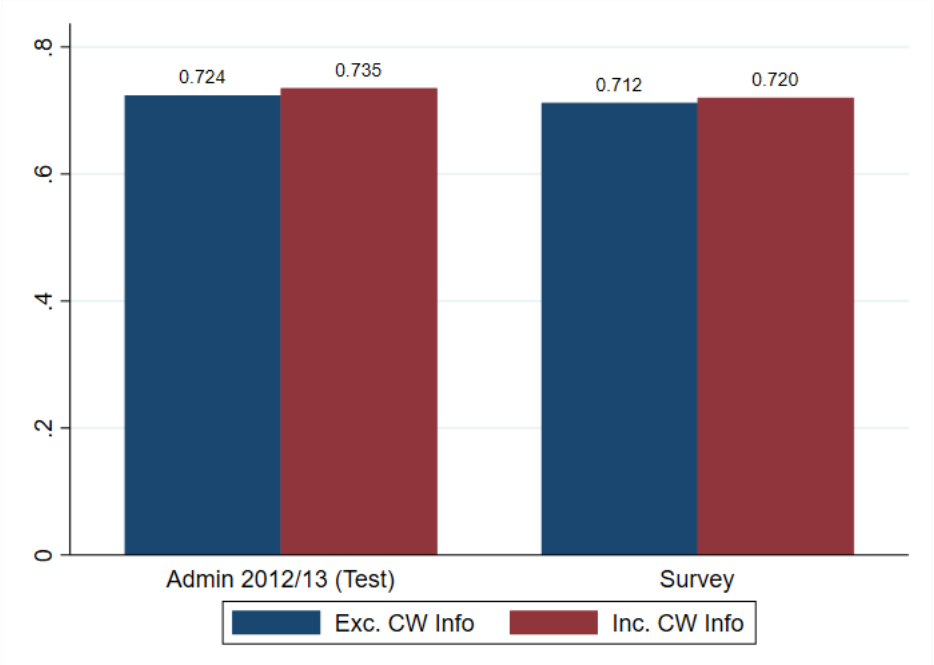
Note: This figure shows the accuracies of different classification methods across all relevant classification thresholds for the survey sample. GBC = Gradient Boosting Classifier; LR = Logistic Regression; RF = Random Forest. Source: EVA; IEB v.12.01.00.

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



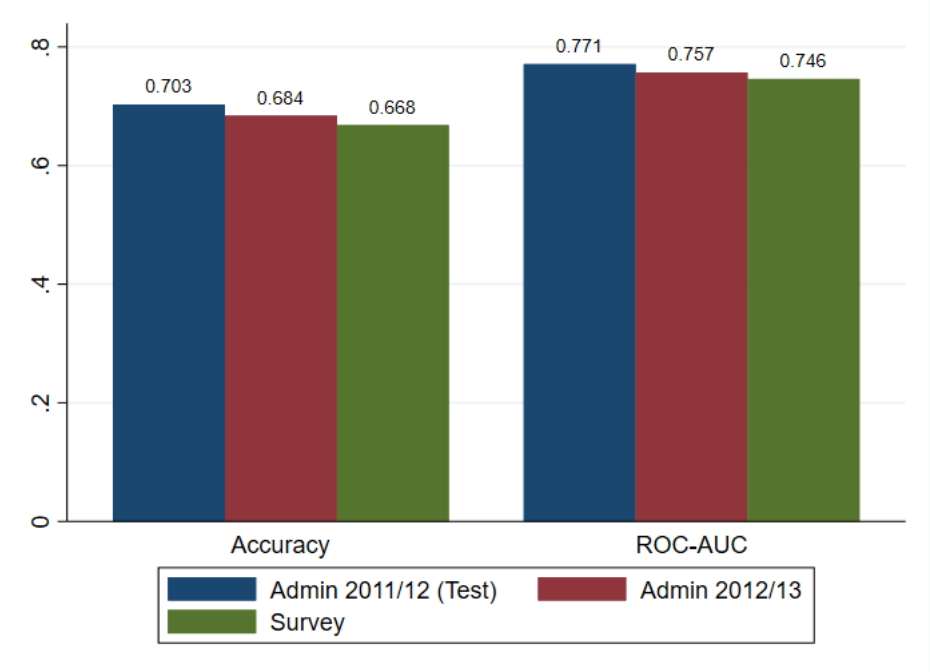
Note: This figure shows the accuracies of different classification methods across all relevant classification thresholds for the survey sample. Instead of using data from the year before (2011/12) to train the algorithm, we used data from 2012/13 as training input. Source: EVA; IEB v.12.01.00.

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	Tot. Days Employed Last 1-7 Years
	Tot. Days Employed Last 1-7 Years (Marg. Empl.)
	Tot. Days Employed Lifetime
	Tot. Days Employed Lifetime (Reg. Empl.)
	Tot. Days Employed Lifetime (Marg. Empl.)
	Tot. Earnings Last 1-7 Years
	Tot. Earnings Last 1-7 Years (Reg. Empl.)
	Tot. Earnings Lifetime (Lifetime)
	Tot. Earnings Last 1-7 Years (Marg. Empl.)
Unemployment History	Amount UE Benefits at Interview
	Tot. Days Receiving UE Ben. Last 1-7 Years
	Tot. Days Receiving UE Ben. Lifetime
	Tot. Amount UE Ben. Last 1-7 Years
	Tot. Amount UE Ben. Lifetime
	Tot. Days Registr. UE Last 1-7 Years
	Tot. Days Registr. UE Lifetime
	Tot. Days Labor Market Program Last 1-7 years
	Tot. Days Labor Market Program Lifetime
LHG History	LHG at interview
	Tot. Days LHG Lifetime
	Tot. Days LHG Last 1-7 Years
Caseworker Profile	Integration within 6 / 12 / >12 Months

Note: This table shows the predictors used to train the main random forest models used to predict reintegration into the labor market within six months. Source: 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 contin. (no CW)	0.011***		
	(0.0006)		
RF contin. (incl CW)		0.011***	
		(0.0006)	
Self-Assessment contin.			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. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Source: EVA; IEB v.12.01.00.

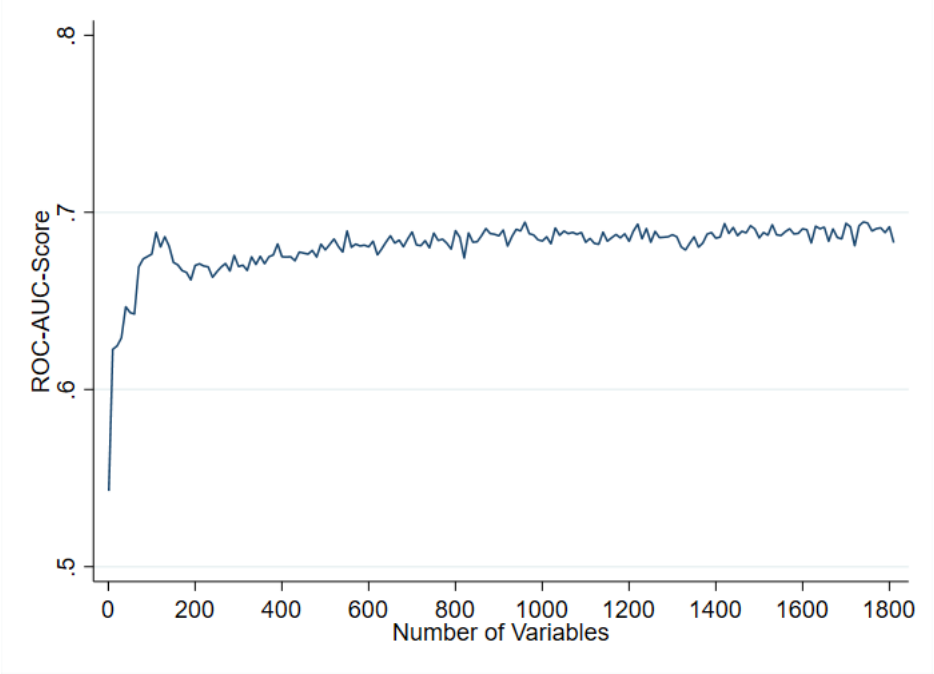
Table A 5: Explanatory power of different predictors - combined models (continuous predictors)

	(1)	(2)	(3)
VARIABLES	RE6	RE6	RE6
RF contin. (no CW)	0.011***	0.009***	0.0086***
	(0.0006)	(0.0007)	(0.0008)
Self-Assessment contin.		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 a continuous variable, 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.

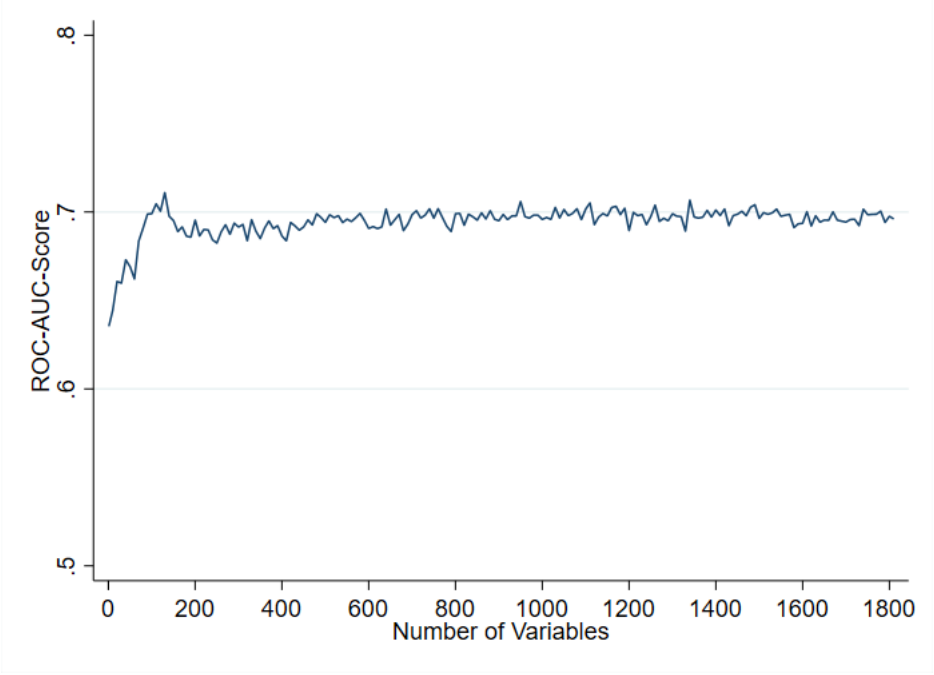
Appendix B. Additional Figures and Tables

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



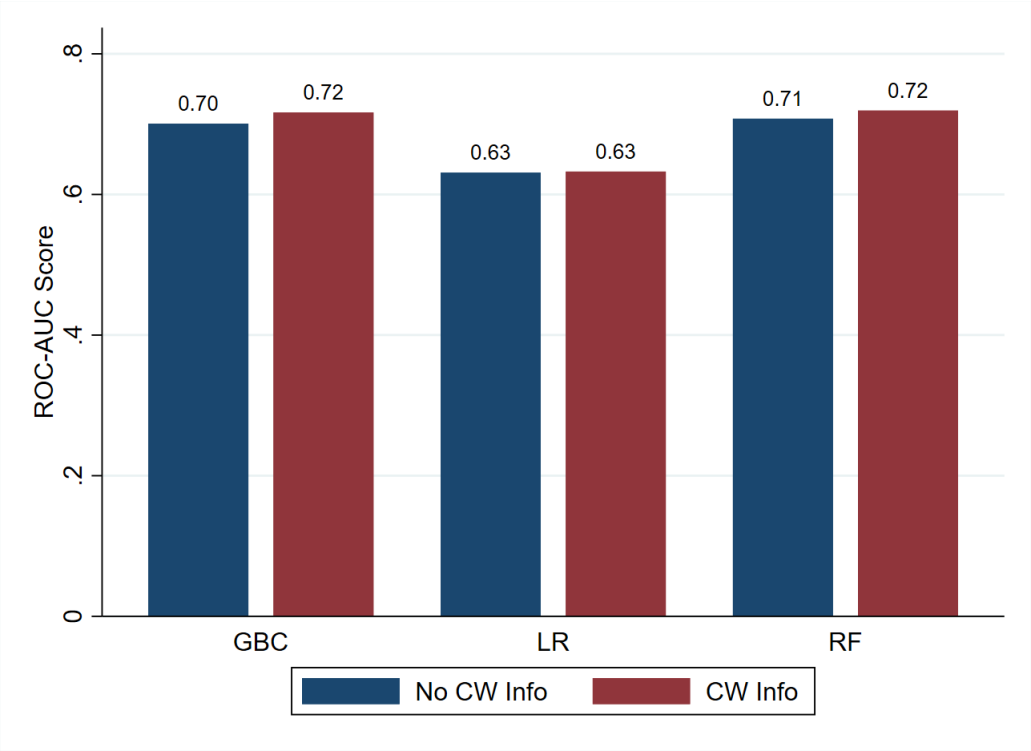
Note: This figure shows the ROC-AUC Scores of the random forest model for the admin sample when we increase the number of variables available for prediction (no additional model tuning). Source: IEB v.12.01.00.

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



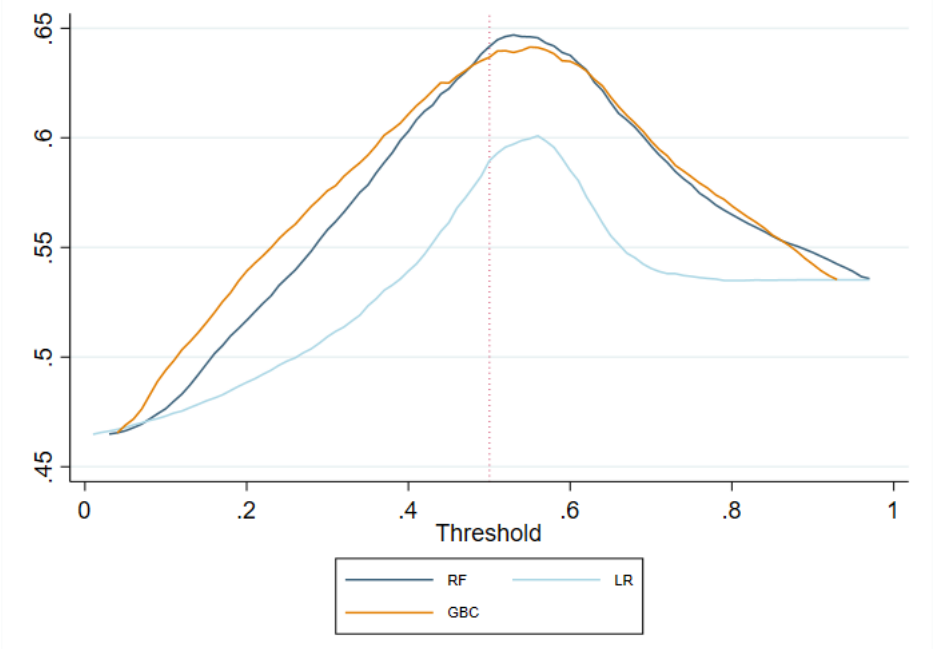
Note: This figure shows the ROC-AUC Scores of the random forest model for the admin sample when we increase the number of variables available for prediction (no additional model tuning). Source: IEB v.12.01.00.

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 = Logistic 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 = Logistic 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.