Journal for Labour Market Research

ORIGINAL ARTICLE

Open Access

From rules to forests: rule-based versus statistical models for jobseeker profiling

Álvaro F. Junquera[1*] and Christoph Kern[2,3,4]

Abstract

Public employment services (PES) commonly apply profiling models to target labor market programs to jobseekers at risk of becoming long-term unemployed. Such allocation systems often codify institutional experiences in a set of profiling rules, whose predictive ability, however, is seldomly tested. We systematically compare the predictive performance of a rule-based profiling procedure currently used by the PES in Catalonia, Spain, with the performance of statistical models in predicting future long-term unemployment (LTU) spells. Using comprehensive administrative data, we develop logit and machine learning models and evaluate their performance with respect to both model discrimination and calibration. Compared to the rule-based model used in Catalonia, our machine learning models achieve greater discrimination ability and remarkable improvements in calibration. Particularly, our random forest model is able to accurately forecast LTU spells and outperforms the rule-based model by offering robust predictions that perform well under stress tests. This paper presents the first performance comparison between a complex, currently implemented, rule-based approach and complex statistical profiling models. Our work illustrates the importance of assessing the calibration of profiling models and the potential of statistical tools to assist public employment services.

Keywords Algorithmic profiling, Unemployment, Public employment services, Machine learning

JEL Classification J64, J68, J08

1 Introduction

Preventing long-term unemployment (LTU) remains a central objective of many labor market policies and is one of the main tasks of public employment services (PES). Low employment prospects and prolonged unemployment spells can have serious consequences for the affected individuals, including economic deprivation through the so-called scarring effects (Filomena 2024) or adverse health outcomes in the long run (Picchio and Ubaldi 2022). From a societal perspective, unemployment is associated with high costs for welfare services. In the European Union, this problem is especially prevalent in countries such as Spain or Greece, with annual unemployment rates even doubling the EU average in 2023 (Eurostat 2024a). This has led to high expenditures in passive labor market policies, placing Spain second in the European ranking with 1.52% of its GDP allocated to such programs in 2019. At the same time, comparatively little funding is used in these countries to support active labor market polices, such as job search interventions (DG EMPL 2024). Under these circumstances, an efficient allocation of access to such programs is essential.

Given these manifold challenges, public employment services aim to identify individuals at risk of long-term unemployment using profiling procedures and provide them with targeted support to increase their labor market prospects. Accurately predicting adverse outcomes

*Correspondence:
Álvaro F. Junquera
alvaro.junquera@bsc.es
[1] Centre d'Estudis Sociològics Sobre la Vida Quotidiana i el Treball, Institut d'Estudis del Treball, Universitat Autònoma de Barcelona, Bellaterra, Spain
[2] Munich Center for Machine Learning (MCML), LMU Munich, Munich, Germany
[3] University of Maryland, College Park, USA
[4] University of Mannheim, Mannheim, Germany

early on is a central concern in these efforts since support programs are intended to be used as preemptive measures. As flexible machine learning models promise to achieve high prediction performance across various tasks (Caruana & Niculescu-Mizil 2006; Fernández-Delgado et al. 2014), PES in many countries are increasingly interested in exploring profiling approaches that draw on modern statistical models to improve the efficiency and effectiveness of their procedures (Körtner and Bonoli 2023). Countries such as Belgium (Desiere and Struyven 2021), France (Gallagher and Griffin 2023), New Zealand (Desiere et al. 2019), and Portugal (Troya et al. 2018) are currently testing or have already implemented machine learning models in their profiling practices.

However, assessing the potential of statistical profiling in specific application contexts is a complex process and requires careful comparison to existing procedures, which commonly include caseworker- and rule-based approaches (Loxha and Morgandi 2014). While statistical models can draw on millions of data points to identify risk factors of LTU, caseworker- and rule-based procedures can similarly leverage many years of "historical data" and institutional expertise. Therefore, when compared on the same grounds, these traditional approaches may not necessarily yield inferior outcomes. However, such comparisons are difficult because detailed documentation of the specific profiling approaches used by PES is often not publicly accessible. To the best of our knowledge, Desiere and Struyven (2021) and Van den Berg et al. (2024) are the only studies that explicitly compare the predictive performance of statistical profiling methods to rule-based and caseworker-based profiling methods implemented in the respective countries (Belgium and Germany).

The contributions of this paper are as follows. First, using a unique database provided by the Public Employment Service of Catalonia (Servei Públic d'Ocupació de Catalunya, SOC), we are able to compare their currently implemented rule-based profiling procedures with machine learning models in a highly realistic setting. Catalonia presents an interesting case to study for its innovative use of data both to profile jobseekers and to evaluate public policies, which is not typical in Spain (Junquera 2024). Second, we follow a broader vision of predictive performance in these comparisons, including measures of both model discrimination and calibration. This perspective recognizes that the predicted scores of any profiling approach should be an honest reflection of actual labor market prospects because the mere reporting of such scores in counseling practice as a form of "weak intervention" can have significant consequences. Third, we present results of the first statistical models trained for Catalonia and the first machine learning models for Spain. We show that administrative databases may be used to build models that can considerably outperform the rule-based approaches currently used in profiling practice on various metrics. We further highlight the need to tailor the model evaluation routine to the unique demands of the profiling context by considering stress tests, group-specific performance scores, and model interpretability.

Following Kuppler et al. (2022), we argue that individuals can be allocated to labor market programs via a two-step allocation system, which includes a decision and a profiling step. In the decision step, the decision-maker must establish an allocation principle, a function that maps individuals to treatments according to certain variables. The allocation principle may be formulated according to distributive justice principles such as those presented in Elster (1992). Profiling is only required if the allocation principle includes the value of a variable that is unobserved at decision time as decision criterion. In the profiling step, this value is usually approximated through a predictive model if the criterion is a value in the future or through a descriptive model if the criterion is a latent value at decision time. Human discretion thus does not disappear in an allocation system with statistical profiling, since the selection of an allocation principle may often be guided by normative or political principles. The distinction between the profiling step and the decision step further helps to channel recent critiques in the social policy literature regarding the emphasis on accuracy in previous research on statistical profiling models (Gallagher and Griffin 2023).

In the following, we start by reviewing the literature on jobseeker profiling procedures, paying special attention to rule-based and statistical models. We then present our database and the techniques used to build our prediction models. The next section reports the main results of our research. We then elaborate on the similarity of the predictions of the different models and their interpretation, taking into account the importance of human discretion in choosing a model for decision-making. Lastly, we offer some conclusions with lines of future research.

1.1 Profiling models for jobseekers

In the field of employment services, profiling models are used to sort jobseekers by classes (e.g., low or high risk of long-term unemployment) or scores (e.g., the probability of long-term unemployment) (Körtner and Bonoli 2023). The main goal of these tools is to support a posterior action, such as allocating individuals to treatments, although they can also be used to provide a more concise description of jobseekers or as an intervention of information provision (Harmon et al. 2021; Loxha and Morgandi 2014). Detailed overviews of jobseeker profiling models are available in Duell and Moraes (2023), Desiere et al. (2019), or Barnes et al. (2015). Here we focus on the strands of literature related to our research: profiling performance as a function of the degree of human discretion and the application of profiling models in public employment services.

It is common to distinguish three types of profiling models, which differ in their degree of human discretion: caseworker-based, rule-based, and statistical profiling (Desiere et al. 2019; Rebollo-Sanz 2018). Caseworker-based profiling allows each counsellor to have their own model, which is often implicit and unknown. In rule-based and statistical profiling, all caseworkers (and jobseekers) have a common model. The difference lies in the way the profiling model is constructed. While statistical models learn the parameters from data, rule-based models have parameters whose values are usually determined ad hoc by employment offices or politicians (Rebollo-Sanz 2018). In this article, we concentrate on and contrast the performance of these two last model types.

The performance of profiling models is usually assessed through discrimination metrics.[1] To our knowledge, only Desiere and Struyven (2021) have studied the discrimination ability of rule-based models, more specifically, the rule-based model used in Belgium. Their analyses show that the model attains an accuracy of 0.58 and has a higher false alarm rate for foreign (non-Belgium) individuals than for Belgian nationals. The authors conclude that their statistical model, in contrast, would allow for better accuracy while presenting the same ratio of false alarms generated by the rule-based model. The downside is that they focus on a very simple rule-based model, which does not mirror the more complex structure these functions may have in other PES.[2] According to Desiere et al. (2019) and Loxha and Morgandi (2014), rule-based profiling models have been applied at least in Ireland, Norway, Poland, and United Kingdom. In practice, they may be more prevalent. It is common that entry into active labor market programs is governed by specific eligibility criteria (Cronert 2022), which may be understood as a consequence of implicit rule-based models. Nonetheless, rule-based approaches have not been sufficiently studied, and their specific implementation details are rarely made public by PES.

The literature on statistical profiling models is more extensive. They have been implemented and publicly assessed by the PES in Austria, Belgium, Denmark, Ireland, and the Netherlands (Desiere et al. 2019). Pooling all of them, their prediction accuracies range from 0.61 to 0.85. Researchers have also recently proposed statistical models for Germany, attaining a ROC-AUC of 0.75–0.77 (Bach et al. 2023; Kunaschk and Lang 2022); for Finland, with a ROC-AUC of 0.8 (Viljanen and Pahikkala 2020), and for Slovakia, with an accuracy of 0.918 (Gabrikova et al. 2023). The range of ROC-AUC found in the literature is 0.7–0.8 (see Appendix B for a detailed comparison). Research by Van den Berg et al. (2024) or Arni and Schiprowski (2015) has also shown that classifications made by caseworkers perform substantially worse than a statistical model in terms of sensitivity.

In Spain, only Felgueroso et al. (2018) and Molina Romo et al. (2023) have explored the development of statistical profiling models.[3] Both studies estimated generalized linear models, but they experimented with different sets of predictors. The model of Felgueroso et al. (2018) incorporates classical covariates and indicates age as one of the most important predictors of long-term unemployment.[4] A similar version has already been used with a private provider of active labor market programs (ALMPs) (Casanova et al. 2021). Molina Romo et al. (2023) studied the prediction ability of personality traits, personal networks, and jobseekers' expectancies regarding the probability of finding employment. Their results go in line with the findings of Van den Berg et al. (2024), with expectancies being a remarkable predictor of long-term unemployment in both cases. Our research tries to integrate both perspectives by constructing a long panel

[1] Throughout this paper, we use "discrimination" with the meaning it has in the biostatistical literature (Austin & Steyerberg 2012), i.e., the ability of a model to separate units that will and will not experience the event.

[2] See Appendix A for a graphical representation of one of the rule-based models studied in this article. The allocation system used in Catalonia relies on a more complicated rule-based profiling model than the one used in Belgium because it pursues a different aim. The Catalan system, like the Polish one, creates groups (profiles) directly attached to different interventions (Niklas et al. 2015). In contrast, the Belgian model seems to provide a simpler order, directed to just one intervention: outreaching users (Desiere and Struyven 2021; Ernst et al. 2024). Moreover, Appendix A provides a complete overview of the whole set of rules passed in a PES to regulate access to each of the many programs. Previous expositions of rule-based models in other countries usually offer only a partial view of the system, because they focus on the rules that govern access to only one program (Loxha and Morgandi 2014). The Catalan allocation system is currently under review by a committee of PES technicians, trade unions, employer organizations, and local public administrations (Consell de Direcció del SOC 2023).

[3] There is a tool called Send@, which was developed by the State Public Employment Office (Servicio Público de Empleo Estatal, SEPE) and is closer to a targeting model in the sense of Körtner and Bonoli (2023). Profiling models try to predict a potential outcome under/after no intervention $\left(\Pr\left[Y(0) = y|X\right]\right)$, whereas targeting models focus on a vector with an element for each potential outcome after going to a certain intervention $\left(\mathbf{v} = \left(\Pr\left[Y(d_1) = y|X\right], \Pr\left[Y(d_2) = y|X\right], \ldots, \Pr\left[Y(d_K) = y|X\right]\right)\right)$. According to Muñiz (2021), Send@ detects the individuals who had certain covariate values $\boldsymbol{X} = \boldsymbol{x}$ with the highest improvement in labor insertion $(i \in Best_{\boldsymbol{x}})$. Then, it offers two sorted vectors of conditional probabilities for the interventions in which they participated $\left(\mathbf{v}_1 = \left(\Pr\left[D = d_1|i \in Best_{\boldsymbol{x}}\right], \ldots, \Pr\left[D = d_K|i \in Best_{\boldsymbol{x}}\right]\right)\right)$ and for the occupations of interest to these individuals $(\mathbf{v}_2 = (\Pr[O = o_1|i \in Best_{\boldsymbol{x}}], \ldots, \Pr[O = o_J|i \in Best_{\boldsymbol{x}}]))$.

[4] It is difficult to judge the importance of each covariate, since all of them are categorical (usually with more than two levels) and only average marginal changes for each category are presented.

of spells that incorporates information on lagged outcomes, which are possibly related to unobserved features that predict LTU (Caliendo et al. 2017; Mueller and Spinnewijn 2024).

1.2 Profiling model currently used in Catalonia

The Public Employment Service of Catalonia (SOC) already uses an allocation system for jobseekers. Its profiling model is a mixture of caseworker- and rule-based procedures, the latter being used to assist office workers in allocating individuals to interventions. It includes an allocation principle for the first two interventions of each unemployment spell that an individual experiences,[5] which facilitates their placement among a range of job search assistance interventions. Still, the office admits that the scores of its profiling step might also assist future decisions (SOC 2016). The system uses two sets of variables as decision-relevant criteria: the so-called occupational variables (combined through the Q models) and criticality variables (combined through the C function). Here, we focus on the Q models, since they are the main tool of diagnosis and allocation (SOC 2016). Caseworker-based models are applied for further decisions and to temporally rank the treatments between individuals (SOC 2016). They are not documented and cannot be readily evaluated empirically.

Let us review the inputs, processing, and outputs of the Q models. They take as inputs administrative data on labor markets and data collected through a questionnaire administered to jobseekers. The first set of variables incorporates information on the economic environment, especially unemployment rates by occupation and sector. The second set includes covariates on the individual's work experience and skills.[6] Note that it does not consider variables on individual unemployment or inactivity spells in the past. The input data is processed using two functions that serve different purposes. First, a step function assigns the individual to a certain group (Q-G). This function serves to determine the first intervention of the jobseeker. Second, a rule-based model assigns each individual a number that represents their employability (Q-S). This function is used to support decision-making for subsequent interventions and to monitor progress in employability.

In both cases, the weight of each variable is not based on a statistical method, but on human intuition. Q-S is a sum of coefficients attached to qualitative variables, whereas Q-G may be understood as a decision tree. In this way, we obtain two outputs: a continuous value on the assessed employability ($S_{it} \in [0,139]$) and a discrete value for the assigned group ($G_{it} \in \{A1, A2, A3, A4, B1, B2, B3, C1, C2, D, Z1, Z2, Z3, E, R6\}$). We argue that we can interpret Q-S and Q-G as intended proxies of the (long-term) unemployment probability, although they are not explicitly framed as predictive models.

The design of the profiling process establishes that S_{it} and G_{it} must be calculated for the same person i at different points in time t, with a maximum of once a month (SOC 2016). Such calculations may be triggered by the beginning of an intermediation claim (*demanda de empleo*), changes of such claim, or the termination of an ALMP. The implementation of the profiling process was analyzed by Everis (2017), finding that 45% of caseworkers thought that the efficacy of Q was either low or moderate. Moreover, they also report that caseworkers manually changed the output of Q-G in 20% of cases.

2 Methods

2.1 Data

To train our profiling models, we were granted access to administrative data from the Public Employment Service of Catalonia (SOC). Four datasets have been matched: the dataset on intermediation claims (SICAS), on labor contracts (Contrat@), on active labor market programs (Galileu), and on benefits or passive labor market programs offered by the Spanish PES. The resulting dataset has the unemployment spell as the unit of analysis. First, we obtained unemployment spells of a simple random sample of 25,000 individuals for each of the four focal years (2017, 2018, 2019, and 2022) from the population of individuals registered as unemployed in that year. We then extracted information on selected variables for each sampled individual for the time window [2015, 2023] from each dataset. Thereby, the final sample of individuals includes persons who were selected in the samples of 2017, 2018, 2019, and/or 2022.

In the next step, we constructed the dataset of labor market spells and the dataset of policy spells. A spell of individual i is simply defined as a closed time interval that started at day t and ended at day $t\prime$. The first type of spells collects spells of participation in the labor market, whereas the second covers spells of participation in active or passive labor market policies. For a given individual, labor market spells are non-overlapping time intervals, but policy spells may overlap in time.

We distinguish three types of labor market spells: employment spells, unemployment spells, and inactivity spells. A new labor contract configures a new employment spell, whilst unemployment and inactivity spells

[5] For some cases, it only defines the allocation principle for the first intervention. A graphical representation of the decision functions is available in Appendix A. In any case, these allocation principles are only formulated vaguely and disconnected to justice principles.

[6] The complete list of variables used in Q models is available in Appendix D.

are defined according to the type of intermediation claim registered. An exhaustive map of types of claims to distinguish unemployment and inactivity is available in Appendix C. Some of the factors that define spells of inactivity are temporary inability, permanent inability, prison entry, or family care. Here, spells of unemployment or inactivity are defined as the presence or succession of intermediation claims of such type. The dataset of policy spells distinguishes four types of spells: participation in adult training, participation in job search assistance or intermediation, participation in an employment subsidy, and receipt of a benefit.

The final step was to compile a dataset of unemployment spells. Following the bulk of the literature on jobseeker profiling (Körtner and Bonoli 2023), we defined our outcome variable to indicate whether an unemployment spell is a long-term unemployment spell. An unemployment spell is defined as long-term if it lasts at least 365 days (Eurostat 2024b). Table 1 summarizes the number of jobseekers and unemployment spells by year and the prevalence of the event of interest, i.e., long-term unemployment (LTU). In Sect. 2.2.1, we specify which unemployment spells were used as units for model training/evaluation and which for model testing.[7]

Our data includes both time-invariant and time-variant covariates as predictors. Like in Bach et al. (2023), we condensed the time-variant information on past labor market spells and policy spells into variables that summarize (un) employment histories. Table 2 displays the groups of predictors used in our models with some examples of specific variables. This list of covariates follows the work of Bach et al. (2023) for Germany, adapted to the Catalan setup. A complete list of predictors is available in Appendix D, and summary statistics on the sociodemographic features of our sample are presented in Appendix E.

To compare our prediction models with the SOC's current profiling approach, we used an extra dataset with profiling scores derived from the rule-based Q model. The current implementation of Q allows that an individual receives only one G_{it} but more than one S_{it} for the same spell (i.e., for the same starting date). This is possible because S_{it} is actually defined for each occupation of interest (at most three). To facilitate the comparison with our models, we calculated S_{it} as the average of the score obtained for each occupation of interest.

Table 1 Sample size and events of interest by year in which the spell started

Year	Unemployment spells	LTU spells	Individuals	Individuals with at least one LTU ep
2017	44,852	9,757 (21.8%)	31,524	9,757 (30.95%)
2018	46,548	9,468 (20.3%)	32,639	9,468 (29.01%)
2019	47,648	11,618 (24.4%)	33,656	11,618 (34.52%)
2020	57,473	18,825 (32.8%)	37,096	18,825 (50.75%)
2021	32,985	7,612 (23.1%)	23,355	7,612 (32.59%)
2022	34,922	5,803 (16.6%)	24,952	5,803 (23.26%)
Total	292,725	63,083 (21.55%)	85,398	54,781 (64.15%)

2.2 Analytical strategy

2.2.1 Development of models

We built profiling models based on four prediction techniques, covering conventional regression models and tree-based machine learning algorithms: unpenalized logistic regression (LR), penalized logistic regression (PLR; Friedman et al. 2010), random forest (RF; Breiman 2001b), and gradient boosting machine (GB; Chen & Guestrin 2016). Logistic regression is the most common technique used for jobseeker profiling (Desiere et al. 2019) and is employed as a baseline. We considered the classic linear and additive specification, which ensures a high degree of interpretability. However, the problem is that this functional form is often poorly justified. Machine learning methods are, on the other hand, highly flexible regarding the relationship between predictors and the outcome. Nonetheless, this flexibility leads to a lower degree of interpretability.

To estimate all models, we followed the dataset partition that is usually applied in the machine learning literature to avoid overfitting and provide realistic evaluations (Kuhn and Johnson 2019). The data was split into three sets: training, evaluation, and test data. The training set was used to tune the internal parameters of the methods (if any) and to estimate the coefficients of the model. The evaluation set served to select the probability threshold for assigning the estimated class (i.e., LTU or non-LTU). The test set was then applied to assess the final models. The training, evaluation, and test sets were constructed using two partitions. First, following Bach et al. (2023), we assigned the observations from 2017 to 2020 to the training set and the evaluation set and reserved the data from 2022 for the test set. Second, we applied stratified random resampling to separate the training set (80% of units) from the evaluation set (20% of units). We used the outcome as the stratifying variable to guarantee sufficient presence of events. Last, the hyperparameters of each model were tuned in the training set with respect to ROC-AUC through temporal cross-validation (Hyndman & Athanasopoulos 2018), departing from the grid of candidates available in Appendix F.

[7] There are more than 25,000 individuals per year because unemployment spells do not only come from the individuals originally sampled that year, but also from historical data of persons sampled in other focal years.

Table 2 Groups of predictors

Group	Number of predictors	Predictors (examples)
Employment	18	Days since last employment, days since last full-time employment, occupation of last employment…
Unemployment	5	Total duration of unemployment spells, number of unemployment spells, days since last unemployment spell…
Inactivity	3	Total duration of inactivity spells, number of inactivity spells, mean duration of inactivity spells
Benefits	5	Start of unemployment spell during a benefit interval, number of benefit spells completed, total duration of benefit spells…
ALMP	9	Total duration of job search assistance spells, total duration of adult training spells, total duration of employment subsidy participations…
Sociodemographics	37	Sex, nationality, age, field of education…

The test set was further reduced to a *restrictive* test set. Note that one of the contributions of our article is the comparison of profiling models with coefficients estimated by humans (rule-based models) with those estimated by statistical methods. This requires a test dataset in which the predictions of both the currently used (Q) and the proposed (K) models can be compared. To achieve this, we took the spells already profiled with Q in 2022 and predicted the score/class they had received in case they had been profiled with our models. We then applied two restrictions to this dataset of Q-profiled spells in 2022 to have a fair and realistic comparison between both profiling approaches.

The first filter levels the playing field between the currently used and the proposed models. The reason is that the variable to be predicted is eventually also affected by the (prediction-based) interventions (Coston et al. 2020). That is, if the allocation had followed the recommendations of the Q predictions and the ALMP had positive effects on re-employment, the currently used profiling model would face a "blessed curse": it would register a bad predictive performance when, in the absence of interventions, it might in fact have a good performance. To mitigate this problem, we removed unemployment spells in which the individual participated in at least one ALMP. The second filter focuses on the target groups of the Public Employment Service of Catalonia. Since the SOC refers people who speak neither Catalan nor Spanish to other public administrations to give them other treatments, it would not be reasonable to prioritize a given model just because it is more sensitive to a group of individuals who eventually would not be treated by the office. For that reason, we removed spells related to persons who do not speak Catalan or Spanish. After these two restrictions, we ended up with a so-called *restrictive* test set of our data.[8]

2.2.2 Model validation

To validate the models, we focus on two dimensions of performance: model discrimination and calibration. Discrimination is the usual objective of research on jobseeker profiling and tries to separate high-risk from low-risk individuals. It can be studied through ranking and classification metrics. Calibration focuses on the difference between the proportions of predicted and observed events. It has been called "the Achilles heel of predictive analytics" (Van Calster et al. 2019) since it is often neglected in model evaluations although it can have significant impact in practice. In our context, it is an important dimension for employment services in any of the following scenarios. If caseworkers were to inform jobseekers about their predicted risk in order to influence their job search, such predictions would need to be reliable, as they could trigger important individual decisions. For instance, if a jobseeker were told that they have a 95% probability of becoming LTU in their current region, they might consider moving to another region. However, if this probability were only 55%, they might reconsider the move, thereby avoiding considerable social and economic costs. In addition, calibrated predictions are critical when intervention decisions are made based on predefined thresholds of predicted risk. If ALMPs are only assigned to jobseekers with, for example, a 75% (predicted) probability of becoming LTU or higher, miscalibrated models can differ considerably in the set of jobseekers exceeding this (fixed) threshold. This issue can be exacerbated when risk predictions are miscalibrated across subgroups (Obermeyer et al. 2019).

Concerning model discrimination, all corresponding metrics are functions of four

[8] To use our statistical models for profiling, we also had to ensure that the individual had at least one spell of labor market history in the past. For 2022, our sample included 18,586 unemployment spells (15,087 individuals) that were profiled with Q models. In Catalonia, the PES does not profile all unemployment spells; instead, it usually profiles only if the last profiling was conducted more than one year ago or if certain variables have changed and the last profiling was carried out more than one month ago. After applying these restrictions, we ended up with 11,082 unemployment spells (9,414 individuals).

quantities that compare the actual LTU outcome and the predicted LTU classification: the number of true positives ($TP = \sum_{it} 1(\widehat{Y} = Yes) 1(Y = \widehat{Y})$), the number of true negatives ($TN = \sum_{it} 1(\widehat{Y} = No) 1(Y = \widehat{Y})$), the number of false positives ($FP = \sum_{it} 1(\widehat{Y} = Yes) 1(Y \neq \widehat{Y})$), and the number of false negatives ($FN = \sum_{it} 1(\widehat{Y} = No) 1(Y \neq \widehat{Y})$).

We assess classification performance through three metrics: accuracy, precision, and sensitivity. Formally, they are defined as the following ratios: $Accuracy = \frac{TP+TN}{TP+TN+FP+TN}$, $Precision = \frac{TP}{TP+FP}$, and $Sensitivity = \frac{TP}{TP+FN}$. The accuracy statistic gives the same weight to correct predictions of events (LTU) and non-events (non-LTU) and gives us a first overall assessment of prediction quality (i.e., the proportion of correct predictions for both outcome classes, LTU and non-LTU, relative to all predictions). However, accuracy scores can be misleading when the prediction target is unbalanced (James et al. 2021). Furthermore, it is reasonable to assume that employment services are more interested in detecting events than non-events, and the sensitivity statistic is calculated for this purpose. It assesses the proportion of true LTU events that are correctly detected by the models' predictions. That is, a model with a high sensitivity has a high capability to capture many high-risk jobseekers. Nonetheless, classifying all unemployment spells as predicted events would lead to perfect sensitivity while being a highly non-efficient solution if treatment is assigned through predictions. Precision informs on the efficiency of predictions by confronting true positives with false positives. It assesses the proportion of predicted LTU events that in fact correspond to true LTU spells. A model with high precision is able to efficiently identify LTU spells without making many false positive predictions.[9]

The outlined classification metrics require the definition of a threshold in order to assign scores to classes. Q is a rule-based profiling approach, and thus the threshold typically would not be defined based on a statistical procedure in practice. For the group profiling (Q-G), we therefore classified the spells that were originally linked to the most intense treatments (individual interventions) as predicted events (SOC 2016).[10]For the score profiling (Q-S), we considered two options. First, we applied the standard procedure for probability models, in which a unit is labeled as "event" if its value is closer (or equally close) to the upper limit of the measure.[11] We called this $\widehat{Y}_{S50}$, since with a probability measure the class is equal to $\widehat{event} = Yes$ if $\widehat{\Pr}[event] \geq 0.5$. Second, we used a stricter function that classifies a score as an event if it fits into the top 25% of possible values of the measure. The transformations from scores to classes follow the functions

$$\hat{Y}_{S25} = \begin{cases} Yes & S \leq 0.25(139) \\ No & S > 0.25(139) \end{cases};$$

$$\hat{Y}_{S50} = \begin{cases} Yes & S \leq 0.5(139) \\ No & S > 0.5(139) \end{cases};$$

$$\hat{Y}_{G} = \begin{cases} Yes & G \in \{C1, C2, D\} \\ No & G \notin \{C1, C2, D\} \end{cases}.$$

Our four statistical profiling techniques output an estimated probability of LTU, which we denote as $\widehat{Y}$. To transform this estimated score to an estimated class, we applied two different classification strategies, A and B, which represent two different rationales.

Classification strategy A interprets probabilities as propensities by understanding binary phenomena as the output of a latent variable model (Long and Freese 2006). The probability threshold is a parameter that *exists* and whose value may be learned. It is denoted by C and is considered a tuning parameter. Specifically, this tuning parameter will be learned in the additional evaluation set. We assume that the SOC is more interested in increasing sensitivity (detecting the true events of interest, i.e., the true LTU spells), but not at any cost. Therefore, the cross-validation will try to maximize the Youden's $J = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$, an equal compromise between specificity and sensitivity. This function reaches its maximum when it simultaneously produces zero false negatives (i.e., perfect sensitivity) and zero false positives (i.e., perfect specificity). Following the taxonomy of Elster (1992), this classification strategy is in line with an *admission* procedure for allocating goods, since it does not establish the number of treatment slots in advance.

Classification strategy B follows the rationale of a limited budget to fund public policies. The logic is that public administrations can only pay for a limited number of services. To fix the number of predicted high-risk individuals, this function classifies as high-risk jobseekers only the individuals whose estimated probability is at

[9] Appendix G reports additional results for two more metrics of classification performance: the Kappa statistic (a chance-corrected version of accuracy defined in Sect. 2.2.3) and the false alarm (false positive) rate ($FP/(FP + TN)$). The false alarm rate is the complementary proportion of another classification metric that will be used later: specificity ($TN/(TN + FP)$). It shows the proportion of non-events that were properly detected.

[10] Appendix A describes the different interventions in detail. Q-G is a descriptive profiling model that outputs an unordered categorical value

Footnote 10 (continued)
which is linked to a specific intervention. We argue that it is possible to quantify the predictive ability of Q-G by binarizing its outcomes into intense (assigned to high-risk jobseekers) versus non-intense (assigned to medium to low-risk jobseekers) interventions, under the assumption that the decision-maker assigns more intense interventions to those jobseekers for which they expect longer durations of the unemployment spell.

[11] Or the opposite if the measure is reversed, as in our case.

least equal to the nineth decile of the predicted probability estimated with the evaluation set ($\widehat{\mathbb{D}}_9^{(eva)}$).[12] If the PES wanted to treat only those at the top of the distribution, it would have to anticipate the value of $\widehat{\mathbb{D}}_9^{(eva)}$ for the respective year in order to allocate the individuals immediately without having to cumulate all the candidates. The reasoning is that the demand for services should not change too much in the short run. In a way, this strategy introduces elements of the decision model into the predictive profiling model. In terms of Elster (1992), this classification strategy fits with a *selection* procedure for allocating goods, because it is a relative allocation based on a ranking of candidates.

In formal terms, $\hat{Y}_A = \begin{cases} Yes & \hat{Y} \geq \hat{C}^{(eva)} \\ No & \hat{Y} < \hat{C}^{(eva)} \end{cases}$; $\hat{Y}_B = \begin{cases} Yes & \hat{Y} \geq \widehat{\mathbb{D}}_9^{(eva)} \\ No & \hat{Y} < \widehat{\mathbb{D}}_9^{(eva)} \end{cases}$.

The ranking metrics we consider are the area under the receiver operating characteristic or c-statistic (ROC-AUC) and the area under the precision-recall curve (PR-AUC). These statistics provide summaries of the discriminatory performance of the models while remaining agnostic regarding the classification threshold. For each possible classification threshold, the ROC curve draws a point that relates the sensitivity (true positive rate) and the false alarm (or false positive) rate produced by such value of the threshold. The PR curve does a similar exercise but plots points that collect the precision and the recall (i.e., sensitivity) for each threshold. By calculating the area, we obtain a panoramic view of the predictive performance regardless of the cutoff that the PES chooses in the future. Note that both metrics can only be computed for profiling functions that output a value measured at the ordinal level.

Regarding calibration, we calculated two statistics following two stringency levels of this dimension. First, we approximated mean calibration through the ratio between the proportion of observed events and the proportion of expected events, denoted as O:E (Van Calster et al. 2019). This metric is also known as calibration-in-the-large, since it gives an aggregated view of how accurate the predicted proportion of events is, but it may hide important deviations at more fine-grained levels. Second, we used moderate calibration to check whether the predicted and the observed proportion of events is equal among units with the same predicted probability. It is assessed through calibration curves summarized with the integrated calibration index proposed by Austin & Steyerberg (2019), denoted as ICI. This statistic is a weighted mean of the absolute difference between the diagonal line of perfect calibration and the calibration curve obtained with a restricted cubic spline of five knots. Although its calculation is more complex, the advantage of this metric is that it measures calibration across the full range of predicted scores and can be easily visualized.

Note that to evaluate our models, we made predictions at the beginning of each unemployment spell (for the test set) or at the moment of the Q prediction (for the restrictive test set, see Sect. 2.2).

2.2.3 Model similarity and interpretation

Even if two models achieve similar classification performance, their unit predictions might differ (Breiman 2001a). This phenomenon has been called model discrepancy (Marx et al. 2020) or model multiplicity (Black et al. 2022). The more discrepancy there is between two models, the higher are the consequences of choosing one model over the other in profiling practice. To measure how prevalent this phenomenon is in our case, we used Cohen's Kappa to approximate the degree of overlap between the predictions of models once agreement by chance is subtracted (Geirhos et al. 2020), similar to the use of this metric to compute the chance-corrected accuracy of model predictions (see Footnote 9). Formally, to compare the predictions of model m to the predictions of model b, we define $Kappa = \frac{v_{obs\,m,b} - v_{exp_{m,b}}}{1 - v_{exp_{m,b}}}$. The first term is the accuracy statistic expressed as $v_{obs\,m,b} = \sum_{it} 1(Y_m = Y_b)/n$, i.e., the number of equal responses out of the total of units. The second term is defined as $v_{exp_{m,b}} = \Pr[Y = Yes]_m \Pr[Y = Yes]_b + \Pr[Y = No]_m \Pr[Y = No]_b$ and collects a binomial process in which the output of model m is statistically independent of the output of model b. The advantage of this metric over the naive accuracy metric is that it considers that models may simply agree by chance.

We further applied the rationale of stress tests in our model evaluation (D'Amour et al. 2022). Stress tests are assessments of model performance using specific inputs designed to evaluate additional criteria of interest. The first test is called shifted performance evaluation and checks the model performance using as input a sample with a different distribution to the one presented by the training sample. We implicitly incorporated this approach by evaluating models with the restrictive test data. The second test is named stratified performance evaluations and analyzes whether performance metrics are similar in certain strata of the population. Given that the SOC (2023) is specially interested in two subpopulations, older jobseekers (> 45 years old) and older female jobseekers, we focused on these groups.

In addition, to facilitate the interpretation of the importance of each predictor in our models, we estimated

[12] In Bach et al. (2023), the quantile is calculated using the test data. This might preclude the implementation of the profiling model because such quantile would have to be calculated for each individual profiling procedure.

permutation-based variable importance scores (Fisher et al. 2019). Specifically, we considered the ROC-AUC as the loss function, and we run ten permutation rounds with a random sample of 10,000 observations to reduce computational burden. This ranking of predictors is especially interesting for jobseeker profiling as it can provide caseworkers with valuable information. Furthermore, we built a shallow decision tree to mimic the predictions of one of the more complex models, random forest, while being more transparent about the learned rules (see Appendix H for details).

To foster transparency and replicability, we publish all the R code necessary to construct both the datasets and the statistical models.[13]

3 Results and discussion

3.1 Performance comparison

In this section, we present the performance of all techniques in predicting LTU in the test sets. In a first step, we focus on a comparison of our statistical profiling models. Next, we assess the performance of the statistical models against the current profiling tools currently used in Catalonia.

The first set of results demonstrates the ranking performance of our models considering the full range of probability thresholds. Table 3 shows the area under the ROC and PR curves and calibration metrics for the four prediction models considered. In line with results for Germany (Bach et al. 2023), tree-based methods do better both in the ROC and in the PR metrics, but the improvements are modest. The gradient boosting model wins in both cases, followed by the random forest, which is reasonable due to the flexibility of these techniques. Considering that a ROC-AUC of 0.5 would simply be a product of chance and that this statistic reaches its maximum at 1, the four models perform remarkably well. Our results for the ROC are slightly superior to the ones found in Belgium (Desiere and Struyven 2021) and Germany (Van den Berg et al. 2024), although these studies define LTU as a six-month interval. Using the same temporal window, the results of Bach et al. (2023) are very close to ours.

Regarding calibration, the gradient boosting (GB) model presents the most reliable probabilities both at the mean and at the whole range. The O:E statistic shows the correspondence between the average probability of LTU computed from the actual test data and from our predictive models, which should ideally be close to 1. Note that all models overestimate the probability of an LTU event, although the GB algorithm comes closest to the actual probability. As a more stringent measure of calibration, the ICI informs on the average error of the predicted probabilities, which means that it should ideally be close to zero. This time the differences between the models are smaller, but the gradient boosting machine wins again. Table 3 shows that the average error in predicting the probability of LTU is 11 percentage points when using this type of prediction model. To our knowledge, we are the first in the literature on jobseeker profiling to measure calibration in this fine-grained way.

A pertinent question for public employment services is whether the adoption of predictive models is worth the effort. To answer this question, Table 4 presents performance metrics for a comparison of the currently used Q-S model with our proposed models using the restrictive test data. Concerning discrimination, the results indicate that all statistical models outperform the rule-based approach, and in this case the random forest model performs best in both ranking metrics. The Q model (Q-S) has a relatively poor performance if we look at the probability of concordance (ROC-AUC) or the precision-recall curve. If we randomly picked one spell from the strata of actual events and another from the strata of actual non-events, using the Q-S model, the probability that the actual LTU spell had a higher predicted probability is 59.3%. In contrast, the random forest achieves a 73.5% concordance probability. The performance gap between the random forest and Q-S is even larger when considering the precision-sensitivity function.[14]

Concerning calibration, the improvements obtained with the best statistical model are even bigger. The ICI column of Table 4 shows that the average error of the random forest model is small (only 3.7 percentage points). For a comparison with the currently used model, we would require multiplying it by six to obtain the average error of the Q-S model. If we consider a softer version of calibration, the gradient boosting model is the winner by generating an almost perfect calibration at the mean (O:E =1.015). Figure 1 shows the calibration curves of each model in the same plot to compare the calibration across the whole support. The profiling model developed by Felgueroso et al. (2018) for all Spain obtained an O:E statistic of 0.999, which is in practice equivalent to the result of our best model.

It is interesting that this time the gradient boosting model performs worse than the random forest in most metrics, although it is still significantly better than the

[13] The code is available at https://osf.io/jye6q/?view_only=3ef06ff290214bfd88f77954d7fb1b73.

[14] The difference in performance between rule-based and statistical models remains when we retain those who participated in the ALMP after the Q profiling. In Appendix G, we use the full test set with one restriction, which is necessary to observe the actual outcome: having at least one unemployment spell before the Q profiling date in order to estimate the length of the unemployment spell. We find that all models have values of ROC-AUC and PR-AUC very similar to those presented in Table 4.

Table 3 Ranking performance of final models in the test set (2022)

	ROC-AUC	PR-AUC	O:E	ICI
LR	0.742	0.398	0.497	0.170
PLR	0.745	0.396	0.503	0.166
RF	0.758	0.419	0.531	0.147
GB	0.763	0.433	0.603	0.110

Table 4 Ranking performance of the final models in the restrictive test set

	ROC-AUC	PR-AUC	O:E	ICI
Q-S	0.593	0.430	0.609	0.223
LR	0.646	0.480	0.937	0.123
PLR	0.648	0.479	0.943	0.118
RF	0.735	0.603	0.908	0.037
GB	0.696	0.557	1.015	0.067

rule-based model. This suggests that the restrictive test set in which we re-evaluate these techniques may not have the same covariate distribution as the full test set. We checked the first moment of the predictors and found that the highest differences are in the proportions of people whose last job was not temporary (11.4 percentage points more in the test set), who had to commute (9 percentage points more), or who had a tertiary employment spell (7.2 percentage points more).[15] The gradient boosting model estimates its parameters by paying more attention to the units wrongly classified during the learning process. Our results indicate that this model is less robust to shifts in the covariate distribution in our application context. In light of the previous results, we consider the random forest as the best-performing model.

The second set of results take side on the probability threshold to classify an spell as high-risk. Table 5 presents the classification performance of our models in the (full) test set, i.e., with data from 2022. When we use classification strategy A—the classification that uses an optimized threshold—the first three methods (LR, PLR, RF) yield very similar results regarding the discrimination metrics. Accuracy and precision slightly improve with the random forest, although it is the gradient boosting machine that excels in both statistics. We achieve remarkable results for sensitivity, presumably the most important metric for employment services, with a value of 0.793 for the random forest. This high sensitivity is also accompanied by a rise in precision in the case of RF, which is good news in terms of efficiency. Lastly, the results show that the gradient boosting model performs worse in terms of sensitivity, which might indicate overfitting. The RF model achieves a better sensitivity than the statistical profiling model proposed for Spain in Felgueroso et al. (2018), which achieved a sensitivity of 0.682.[16]

Following classification strategy B, the framework that prioritizes the budget, the results are much better in terms of accuracy. However, there is a substantial decrease in sensitivity specifically for the tree-based methods. These techniques correctly predict the outcome classes for 83% of the spells, but the true positives do not represent a remarkable share of these forecasts. Continuing with the budget constraint, a compromise between classification strategy A and B might be to use an alternative outcome variable—the duration of the unemployment spell in days. The ordered nature of this response variable might allow for sorting jobseekers and showing the PES the next candidate to be treated in case there is available funding to do so.

It is interesting to compare the classification performance of our models with the currently used methods. Setting a specific threshold also allows us to assess the discrimination ability of the Q-G profiling model. Table 6 presents the results of such a comparison, this time using data from the restrictive test set. The first panel shows the metrics of the rule-based models. We can see how the quantitative version (Q-S) attains a very high sensitivity when its threshold is located at the middle of its codomain (Q-S50). This is mainly achieved through a very indiscriminate classification of spells, as suggested by a high false alarm rate (see Appendix G). When the classification threshold is located at the top 25% (Q-S25), the rule-based model is more precise, but at the cost of a very low sensitivity. The qualitative version (Q-G) presents a poor 0.204 in sensitivity with an improvement in accuracy against the alternative Q-S50.

The patterns observed for our statistical models are similar to those obtained with the full test set. In a nutshell, we have higher specificity when we rely on thresholds optimized with Youden's J (strategy A) and higher accuracy when we focus on the budget (strategy B). When sensitivity is a high priority, the random forest model under strategy A is the model that performs best. With a sensitivity of 0.860, this algorithm surpasses the discrimination ability of the rule-based Q-S50. Moreover, it improves substantially both in accuracy and in precision. The gradient boosting model may serve as a

[15] Figures showing the quantitative and qualitative variables with the largest differences between the samples are included in Appendix G. We removed the indicators of missingness from these lists.

[16] Note that if we had followed their same procedure, we could have achieved even higher sensitivity. Felgueroso et al. (2018) chose the probability threshold with the test data while measuring sensitivity, whereas we fixed it in a previous step using the evaluation set.

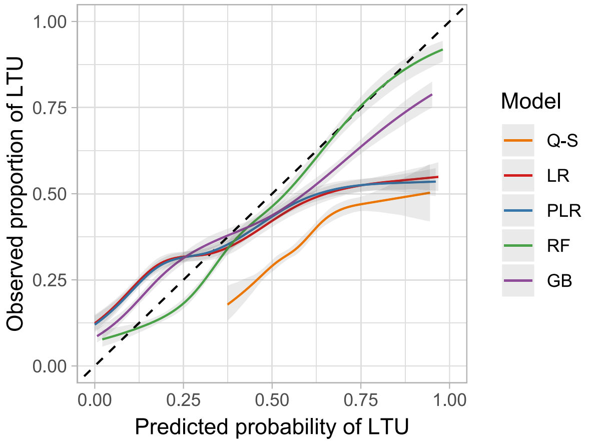

Fig. 1 Calibration curves in the restrictive test set for each model

Table 5 Classification performance in the test set based on different strategies

	Accuracy	Precision	Sensitivity
Strategy A			
LR	0.572	0.251	0.794
PLR	0.580	0.254	0.790
RF	0.595	0.263	0.793
GB	0.667	0.296	0.729
Strategy B			
LR	0.782	0.377	0.483
PLR	0.782	0.378	0.482
RF	0.830	0.481	0.323
GB	0.826	0.469	0.362

compromise between strategy A and B, since it attains a good sensitivity but maintains decent results in accuracy and precision.

With the analysis of calibration and discrimination, we have shown that our random forest model using strategy A (RF-A) outperforms the rule-based model Q-S50 in all metrics. Its added value is especially remarkable in the reliability of its predictions, since the Q-S50 is poorly calibrated. In contrast, our random forest model achieves a remarkable calibration throughout the entire range of probabilities. It also shows an excellent sensitivity (0.860), with improvements in precision and accuracy that may raise the efficiency of treatment assignments.

3.2 Model similarity and interpretation

In this section, we dig into the specific spells flagged by each method and explore how the statistical models utilize the training information. We first measure the degree of model similarity, then analyze the stress tests and interpret the most important predictors of each model.

Table 7 provides the Kappa coefficients of all model comparisons, both for the rule-based and for the statistical models. When comparing the three rule-based models with our alternative statistical classifiers, we see that the agreement between the models is quite low. This might be explained by the fact that the rule-based models mainly represent random classifiers.[17] Therefore, we interpret this disagreement not as a consequence of both approaches approximating different data-generating (sub)processes, but simply as a lack of fit of the rule-based models. If we focus on the statistical models, two results can be highlighted. First, as expected due to the low penalization of the tuned PLR models (see Appendix F), the agreement with the predictions of the LR model is almost perfect for both classification strategies. Second, the agreement of the two big competitors in terms of performance (RF and GB) is in the middle range, especially for classification strategy A. This invites us to review the consequences of choosing one model over the other.

In deciding which model should be deployed, the consequences of model discrepancies may be clarified with so-called stress tests. The first test, the shifted performance evaluation, was carried out when analyzing the differential discrimination and calibration of models with the restrictive test data. Prioritizing sensitivity, the random forest performs best and also attains the highest degree of calibration measured through the entire probability range. The second test, the stratified performance evaluation for older jobseekers and older female jobseekers, is presented in Table 8. Again, the RF performs better than GB in terms of sensitivity for both subpopulations. However, note that this time the simpler models (LR and PLR) do similarly well in predicting events in these groups at the cost of lower precision and lower accuracy. In the end, the model selection should take the cost structure of the SOC into account. GB models offer the lowest sensitivity both for older and for female older jobseekers but have the largest accuracy. Therefore, if the detection of non-events is considered more important, this model could also be implemented.

Lastly, to understand how the statistical models make predictions, we explore the most important predictors used by each method. Figure 2 shows the ranking of the ten most important covariates as measured by the loss in ROC-AUC due to shuffling their values. There is an agreement between the four models that two of the three most important predictors of LTU are the number of days since the last unemployment spell and the total number of unemployment spells experienced in

[17] Appendix G includes a complete table of the Kappa coefficient comparing each model classification with the actual value. The chance-corrected accuracy of Q-S50 is 0.067.

Table 6 Classification performance in the restrictive test set based on different strategies

	Accuracy	Precision	Sensitivity
Q-S25	0.641	0.482	0.090
Q-S50	0.454	0.381	0.852
Q-G	0.605	0.395	0.204
Strategy A			
LR	0.562	0.432	0.729
PLR	0.567	0.436	0.735
RF	0.579	0.452	0.860
GB	0.613	0.472	0.730
Strategy B			
LR	0.653	0.517	0.404
PLR	0.653	0.516	0.406
RF	0.699	0.619	0.402
GB	0.679	0.578	0.367

Note: *N* of the restrictive test set = 11,082

the past. Another variable that ranks high in the four models is age, whether on its own or as a scaling factor of other predictors. This result goes in line with the findings of Felgueroso et al. (2018) for Spain. Looking closer at the tree-based models, the number of days since the last employment spell and the average duration of unemployment spells in the past are also important predictors of LTU. These results are consistent with the findings of Bach et al. (2023), who also reported a high predictive ability of age and labor market histories. Furthermore, McGuinness et al. (2022) developed a model for Ireland and detected that employment histories were also a remarkable set of predictors.

The results of the surrogate tree highlight that it is possible to mimic the predictions of the RF model with good accuracy while presenting a transparent set of rules that may be more accessible to PES caseworkers (see Appendix H).

4 Conclusions

In this article, we have contrasted the performance of rule-based and statistical models for jobseeker profiling. Specifically, we have taken the predictions of the rule-based models currently implemented in Catalonia and compared them with newly developed statistical models for predicting long-term unemployment. Our results show that our statistical models outperform the currently used rule-based profiling approach considerably in terms of both discrimination (ROC-AUC: 0.735 vs. 0.593) and calibration (ICI: 0.037 vs. 0.223). Furthermore, we have seen that machine learning methods achieve higher performance scores than conventional regression models, especially regarding calibration. These are the first machine learning models developed and validated to predict long-term unemployment with Spanish data. We have also shown that, compared with gradient boosting, our random forest model adapts better to covariate shifts and presents better sensitivity for two social groups (older jobseekers and older female jobseekers) that are currently the focus in the operations of the Public Employment Service of Catalonia. Our prediction models have additionally highlighted two important predictor variables that are not utilized in the currently used profiling approach: the number of days since the last unemployment spell and the total number of past unemployment spells.

Our findings corroborate previous results of the profiling literature but also introduce new perspectives. In line with previous research, we confirm the importance of historical data on labor market transitions for accurately predicting long-term unemployment (Gabrikova et al. 2023; McGuinness et al. 2022). Previous literature, however, has highlighted that more flexible methods such as random forests do not make a big difference in performance compared to conventional models such as logistic regression (Bach et al. 2023; Desiere et al. 2019). We argue that this conclusion only holds if we uniquely focus on discrimination. Our dual approach to analyze the prediction performance revealed that machine learning models can improve over regression approaches in terms of calibration, a crucial but overlooked dimension in research on jobseeker profiling. We suggest that calibration be carefully considered in the evaluation of profiling models due to the crucial role of the (predicted) risk scores in the counseling practices of employment services.

Our work also has some limitations that need to be considered. First, compared with related work such as Bach et al. (2023), our set of covariates on past employment spells was limited due to the unavailability of information on the actual end dates of labor contracts. This shortcoming might be tackled in the future by using detailed social security data. Second, our models have been trained with individuals that actively engage with the Public Employment Service of Catalonia. In Spain, registration with PES is not compulsory, which implies that the population of participants in PES may not mirror

Table 7 Kappa coefficients between predictions of different models in the restrictive test set

	Q-S25	Q-S50	Q-G	LR	PLR	RF	GB
Q-S25	1						
Q-S50	0.036	1					
Q-G	0.044	0.051	1				
Strategy A							
LR	0.016	0.014	0.018	1			
PLR	0.015	0.017	0.017	0.925	1		
RF	0.018	0.031	0.025	0.576	0.587	1	
GB	0.018	0.011	0.028	0.703	0.715	0.671	1
Strategy B							
LR	0.014	0.011	0.029	1			
PLR	0.007	0.009	0.029	0.949	1		
RF	0.031	0.025	0.037	0.587	0.586	1	
GB	0.017	0.019	0.046	0.723	0.725	0.732	1

Table 8 Classification performance in two strata of the test set

	Accuracy	Precision	Sensitivity
Older jobseekers			
LR	0.551	0.277	0.826
PLR	0.547	0.276	0.838
RF	0.587	0.295	0.825
GB	0.653	0.328	0.769
Female and older			
LR	0.543	0.279	0.847
PLR	0.537	0.278	0.859
RF	0.603	0.309	0.837
GB	0.665	0.344	0.794

Note: *N* of the older stratum = 20,793. *N* of the female older stratum = 11,452. To simplify the exposition, we only show the results for strategy A

the full population of jobseekers. This implies that some groups, such as young people, may be underrepresented in our training set compared to their presence in the population of jobseekers in general. Relatedly, in our evaluation with the restricted test set, we must consider that we have tested the models with individuals who did not participate in ALMPs to evaluate the models' ability to predict LTU risks under no intervention. However, this introduces the limitation that some of those who were historically treated could be among the most vulnerable individuals, and thus we do not know how the compared models would work for them. Nonetheless, given the considerable performance differences between the statistical and rule-based approaches across all analyses, there is little indication that the relative improvement of the statistical models would not persist (see our subgroup evaluation results). Furthermore, based on our profiling models, we optimized the classification threshold, assuming that false positives and false negatives have the same social costs. There can be sensible arguments for either error to have more significant consequences, and thus the thresholds could be re-optimized with different cost functions.

Lastly, while we evaluated prediction performance for sensitive social subgroups, our paper did not engage in a comprehensive fairness evaluation of the developed prediction models. Additional research is needed to carefully understand the fairness implications of the models for the Catalonian context, e.g., by evaluating whether the prediction models result in similar error rates for multiple sensitive (sub)groups of interest. Regarding the deployment of statistical models in PES, researchers could also experiment with different modes of profiling model implementation to foster the acceptance of the tool by both caseworkers and jobseekers. This line of research has been explored by Kern et al. (2022) and Scott et al. (2022), who have offered some possible explanations for how users perceive these models. Despite these potential extensions, our study illustrates the added value of flexible statistical models over rule-based profiling to support PES. Likewise, it highlights the benefits of the machine learning perspective on performance evaluation in terms of studying both the predictive discrimination ability and calibration of (existing and new) profiling models on the same grounds.

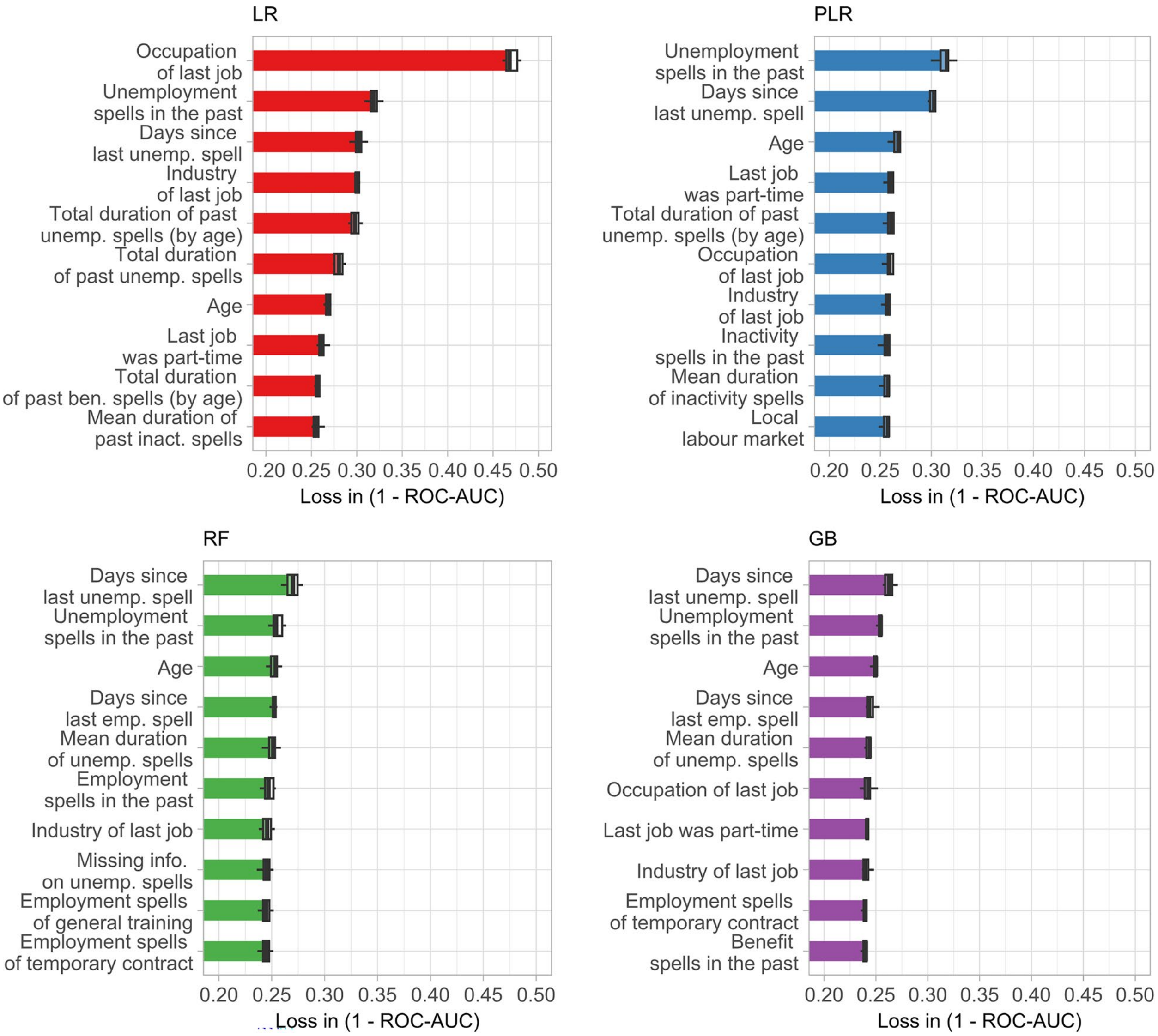

Fig. 2 Top 10 most important variables in the final models. The extension of the bar indicates the permutation statistic, which is a mean across permutation rounds, jointly with the boxplot collecting variability between rounds

Appendix A

Profiling and decision models currently used in the SOC

See Figs. 3, 4 and Table 9

(See figure on next page.)

Fig. 3 Rule-based profiling model Q-G. Note: Own elaboration based on SOC (2016). This decision tree was inferred from the documents provided, so it must be taken with caution. The green line indicates the path if the value is TRUE, the red line if the value is FALSE. It represents the values considered to assign an individual to a treatment, which are the so-called pre-collectives. *Training* indicates if a jobseeker's occupational training is considered enough (1) or if it is not enough (0), denoted as $Tr := \{1, 0\}$. *Experience* indicates if experience is considered enough (1) or if it is not enough (0), denoted as $Exp := \{1, 0\}$. The variable employability for the occupation of interest is represented as $Occu := \{High, Intermediate, Low, \emptyset\}$, and was originally denoted $\{``Viable'', ``Moderado'', ``Enretroceso'', ``Nodefinida''\}$. The variable employability for the sector of interest is represented as $Sector := \{High, Low\}$, and was originally denoted as $\{``No\ enretroceso'', ``Enretroceso''\}$

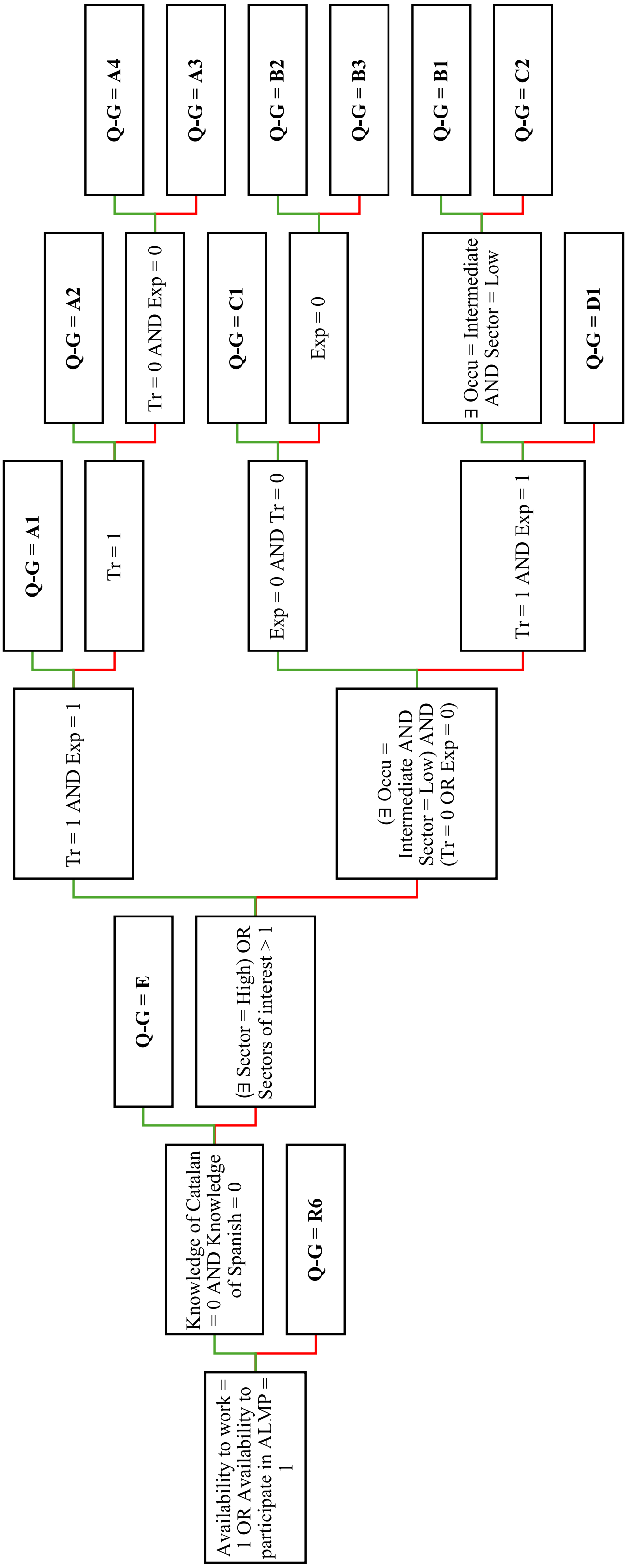
Availability to work = 1 OR Availability to participate in ALMP = 1
Knowledge of Catalan = 0 AND Knowledge of Spanish = 0
Q-G = R6
Q-G = E
(∃ Sector = High) OR Sectors of interest > 1
Tr = 1 AND Exp = 1
(∃ Occu = Intermediate AND Sector = Low) AND (Tr = 0 OR Exp = 0)
Q-G = A1
Tr = 1
Exp = 0 AND Tr = 0
Tr = 1 AND Exp = 1
Q-G = A2
Tr = 0 AND Exp = 0
Q-G = C1
Exp = 0
∃ Occu = Intermediate AND Sector = Low
Q-G = D1
Q-G = A4
Q-G = A3
Q-G = B2
Q-G = B3
Q-G = B1
Q-G = C2

Fig. 3 (See legend on previous page.)

Table 9 Rule-based profiling model Q-S

Variable (index L)	Value	λ_L
$occupation_{ito}(L=1)$	$x = \text{level 1}$	$\lambda_1 = 0$
	$x = \text{level 2}$	$\lambda_1 = 5$
	$x = \text{level 3}$	$\lambda_1 = 6$
	$x = \text{level 4}$	$\lambda_1 = 7$
	$x = \text{level 5}$	$\lambda_1 = 10$
$sector_{ito}(L=2)$	$x = \text{emergent}$	$\lambda_2 = 10$
	$x \in \{\text{more than one, normal}\}$	$\lambda_2 = 5$
	$x \in \{\text{declining, missing value}\}$	$\lambda_2 = 0$
$regexperience_{it}(L=3)$	$x > 60\text{months}$	$\lambda_3 = 10$
	$36 \text{ months} < x \leq 60\text{months}$	$\lambda_3 = 7$
	$18 \text{ months} < x \leq 36\text{months}$	$\lambda_3 = 5$
	$6 \text{ months} < x \leq 18\text{months}$	$\lambda_3 = 3$
	$0 \text{ months} \leq x \leq 6\text{months}$	$\lambda_3 = 0$
$irrexperience_{it}(L=4)$	$x > 60\text{months}$	$\lambda_4 = 2$
	$6 \text{ months} < x \leq 60\text{months}$	$\lambda_4 = 1$
	$0 \text{ months} \leq x \leq 6\text{months}$	$\lambda_4 = 0$
$regexperience_{ito}(L=5)$	$x > 60\text{months}$	$\lambda_5 = 10$
	$36 \text{ months} < x \leq 60\text{months}$	$\lambda_5 = 7$
	$18 \text{ months} < x \leq 36\text{months}$	$\lambda_5 = 5$
	$6 \text{ months} < x \leq 18\text{months}$	$\lambda_5 = 3$
	$0 \text{ months} \leq x \leq 6\text{months}$	$\lambda_5 = 0$
$irrexperience_{ito}(L=6)$	$x > 60\text{months}$	$\lambda_6 = 2$
	$6 \text{ months} < x \leq 60\text{months}$	$\lambda_6 = 1$
	$0 \text{ months} \leq x \leq 6\text{months}$	$\lambda_6 = 0$
$education_{it}(L=7)$	$x \in \{0, \text{1nr, missing value, 25Bnr}\}$	$\lambda_7 = 0$
	$x \in \{1, 25B, 24nr, 35Bnr\}$	$\lambda_7 = 2$
	$x \in \{\text{24, 35B, 44nr, 45nr, 55Bnr, 54nr}\}$	$\lambda_7 = 4$
	$x = \text{44, 45, 54, 55B, 55nr, 75Bnr}$	$\lambda_7 = 6$
	$x \in \{\text{6nr, 7nr, 8nr, 55, 75B}\}$	$\lambda_7 = 8$
	$x \in \{6, 74, 75, 8\}$	$\lambda_7 = 10$
$comptrain_{it}(L=8)$	$x \geq 80$	$\lambda_8 = 10$
	$x < 80$	$\lambda_8 = 0$
$driving_{it}(L=9)$	$x = \text{B1}$	$\lambda_9 = 10$
	$x \neq \text{B1}$	$\lambda_9 = 0$
$closed_{ito}(L=10)$	$x = (\text{Yes} \wedge (\text{Has it} \vee \text{Studying it}))$	$\lambda_{10} = 5$
	$x \in \{\text{Recommended} \wedge (Hasit \vee Studyingit), No \wedge (HasISCED2 \vee StudyingISCED2)\}$	$\lambda_{10} = 4$
	$x \in \left\{\text{Recommended} \wedge Doesn'thaveit,\ No \wedge HasnotISCED2\right\}$	$\lambda_{10} = 2$
	$x \in \{Yes \wedge (Doesn'thaveit \vee Unknown),\ \ (\text{Recommended} \vee No) \wedge Unknown\}$	$\lambda_{10} = 0$
$search_{it}(L=11)$	$x = 0$	$\lambda_{11} = 10$
	$x = 1$	$\lambda_{11} = 5$
	$x = 2$	$\lambda_{11} = 3$
	$x = 3$	$\lambda_{11} = 0$
$languages_{it}(L=12)$	$Cat \in \{High, Middle\} \wedge Sp \in \{High, Middle\}$	$\lambda_{12} = 10$
	$Cat \in \{\text{High, Middle}\} \wedge Sp \in \{\text{Basic, Null}\}$	$\lambda_{12} = 5$
	$Cat \in \{\text{Basic, Null}\} \wedge Sp \in \{\text{High, Middle}\}$	$\lambda_{12} = 4$
	$Cat \in \{\text{Basic, Null}\} \wedge Sp \in \{\text{Basic, Null}\}$	$\lambda_{12} = 0$
$digitalskills_{it}(L=13)$	$x = 0$	$\lambda_{13} = 10$
	$x = 1$	$\lambda_{13} = 5$
	$x \in \{2, 3\}$	$\lambda_{13} = 0$

Table 9 (continued)

Variable (index L)	Value	λ_L
$willingness_{it}$ $(L = 14)$	$x = 0$	$\lambda_{14} = 10$
	$x \geq 1$	$\lambda_{14} = 0$
$communication_{it}$ $(L = 15)$	$x =$ Yes	$\lambda_{15} = 10$
	$x =$ No	$\lambda_{15} = 0$
$interpersonal_{it}$ $(L = 16)$	$x =$ Yes	$\lambda_{16} = 10$
	$x =$ No	$\lambda_{16} = 0$

$education_{it}$ in ISCED-11 coding of educational attainment, with "nr" indicating that the credential has not been recognized in Spain and "B" indicating that the credential was obtained in labour market programs. "Cat" denotes the Catalan language and "Sp" the Spanish language. $\vee$ is the Boolean operator for OR, $\wedge$ is the Boolean operator for AND. Appendix D includes a description of each variable

Q–S $= (S_{ito_1}, S_{ito_2}, S_{ito_3}).\tilde{S}_{it} = mean(S_{ito_1}, S_{ito_2}, S_{ito_3})$.
$Sito = \lambda_1 occupation_{ito} + \lambda_2 \sec tor_{ito} + \lambda_3 reg \exp erience_{it} + \lambda_4 irr \exp erience_{it} + \lambda_5 reg \exp erience_{ito} + \lambda_6 irr \exp erience_{ito} + \lambda_7 education_{it} + \lambda_8 comptrain_{it}$
$+ \lambda_9 driving_{it} + \lambda_{10} closed_{ito} + \lambda_{11} search_{it} + \lambda_{12} languages_{it} + \lambda_{13} digitalskills_{it} + \lambda_{14} willingness_{it} + \lambda_{15} communication_{it} + \lambda_{16} \mathrm{interpersonal}_{it}$

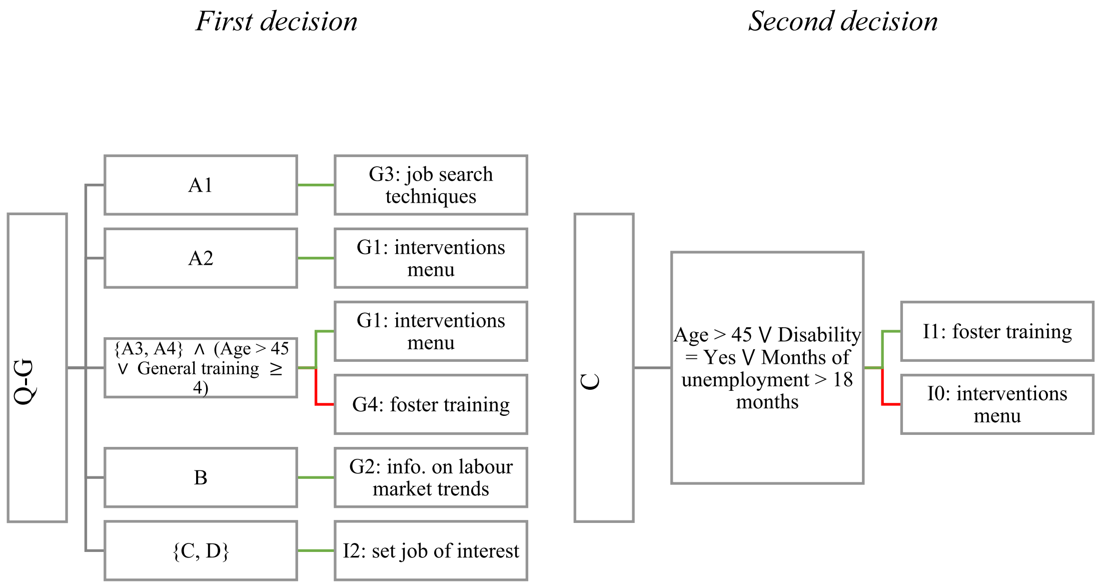

Fig. 4 Decision model for decisions one and two. Source: Own elaboration based on SOC (2016). The green line indicates the path if the value is TRUE, the red line if the value is FALSE. The tree starts with the value of Q-G or C (the so-called "pre-collective") and then links the decision, i.e., the intervention assigned. It also shows whether such intervention is individual (I) or group-based (G) and the main objective of the action. The second decision is only determined by rule-based profiling if the individual was not assigned in Q-G to collectives C, D, or Z. In those cases, the second decision is not regulated by the model. ∨ is the Boolean operator for OR, ∧ is the Boolean operator for AND

The interventions to assign in these decisions may be classified according to three variables (SOC 2016):

1. *Number of participants*: one (individual) or more than one (group).
2. *Main objective*: training on job search techniques, foster adult training, information provision on labor market trends, set job of interest, or present the set of available interventions (the so-called "interventions menu").

3. *Place of implementation*: face to face or remotely.

The intensity of the intervention may be defined according to different criteria. In this article, we have chosen to define individual actions as more intense than group actions, so $d \in \{I0, I1, I2\} > d' \in \{G1, G2, G3, G4\}$

We have maintained the original abbreviations of the group interventions, but we have changed the abbreviations of the individual interventions to avoid confusion. I0 is known as Assessorament Polítiques Actives d'Ocupació (originally abbreviated as APAO), I1 is known as Assessorament Ocupacional (originally abbreviated as AO), and I2 is known as Orientació (originally abbreviated as O).

Note: This table only includes the models that were publicly validated with at least one statistic of discrimination or calibration. Classification metrics are only included if the author recommended or used at least one classification threshold. Note the differences between exit to employment *within* 12 months (at least once in the time interval) and exit to employment *after* 12 months (at the measurement time of month 12). For the results of Gabrikova et al. (2013), although their model uses four categories, we here present the metrics for the category "more than 12 months". (*) Rows with the asterisk indicate that, according to the source, the model has not been yet deployed in public employment services.

Appendix B

Review of performance of jobseeker profiling models

Country	Model	Outcome	ROC-AUC	Sensitivity	Precision	Accuracy	O:E	Source
Austria	Statistical	Labor market integration probability				0.80–0.85		Desiere et al. (2019)
Belgium (Flanders)	Statistical, caseworker-based	Long-term unemployed (> 6 months)	0.76			0.67		Desiere et al. (2019)
Belgium (Flanders)*	Statistical	Long-term unemployed (> 6 months)	0.702			0.702		Desiere & Struyven (2021)
Denmark	Statistical	Long-term (> 26 weeks) unemployed				> 0.60		Desiere et al. (2019)
Finland*	Statistical	Unemployed after 12 months	0.80					Viljanen & Pahikkala (2020)
Germany*	Statistical	Long-term unemployed (> 6 months)	0.7			~ 0.63		Kunaschk & Lang (2022)
Germany*	Statistical	Long-term unemployed (12 months)	0.777	0.29	0.372	0.846		Bach et al. (2023)
Germany*	Statistical	Long-term unemployed (> 6 months)	0.726–0.735	0.8		0.647	1.237	Van den Berg et al. (2024)
Ireland	Statistical	Exit to employment within 12 months				0.70–0.86		Desiere et al. (2019)
Ireland	Statistical	Unemployed after 12 months		0.752		0.777		McGuiness et al. (2022)
Netherlands	Statistical	Long-term unemployed (12 months)				0.7		Desiere et al. (2019)
New Zealand	Statistical, rule-based	Lifetime income support costs, lifetime income support and staff costs	0.63–0.83					Desiere et al. (2019)
Slovakia*	Statistical	Duration of unemployment spell (four categories)		0.7886	0.9147	0.9182		Gabrikova et al. (2023)
Spain*	Statistical	Exit to employment within 12 months				0.682	0.999	Felgueroso et al. (2018)
United Kingdom*	Statistical	Long-term unemployed (12 months)	0.795	0.319	0.333	0.889		Matty (2013)

Appendix C

Classification of intermediation claims

See Table 10

Table 10 Correspondence between causes of intermediation claims and type of spell

Code	*Description of the cause*	*Type*
1	Removal due to placement communicated with prior offer	E
2	Removal due to registration in the general Social Security system	E
3	Removal due to placement in the special self-employed regime	E
4	Removal due to placement communicated without prior offer	E
17	Removal due to the end of a collective dismissal file	U
19	Removal due to call of a seasonal permanent worker	E
25	Removal due to incomplete application	U
30	Suspension without intermediation due to temporary incapacity	I
31	Suspension without intermediation due to maternity/paternity, adoption, or foster care	I
32	Suspension without intermediation due to pregnancy with risk	I
35	Removal due to end of availability	I
36	Removal due to total permanent disability	I
37	Removal due to absolute permanent disability (major disability)	I
38	Removal due to retirement	I
39	Removal due to reaching the minimum retirement age	I
61	Removal due to other causes	I
62	Provisional removal due to untraceable applicant	U
70	Removal due to failure to appear before the managing entity	U
71	Removal due to failure to renew the application	U
73	Removal due to rejecting a suitable job offer	U
75	Removal due to refusal to participate in ALMP	U
100	Voluntary removal	U
102	Removal due to benefit exportation	I
103	Removal due to death	I
104	Suspension due to military service or alternative civilian service	I
105	Removal due to equalization	U
106	Suspension without intermediation due to preventive detention	I
107	Removal due to job placement declaration	E
108	Suspension without intermediation due to deprivation of liberty for fulfilling a sentence of applicants receiving benefits	I
109	Removal due to deprivation of liberty for fulfilling a sentence	I
110	Removal due to non-communication of the renewal of administrative authorization	I
114	Suspension without intermediation due to family obligations	I
120	Suspension without intermediation due to leaving the country	I
121	Suspension without intermediation due to attending training courses	U
122	Suspension with limited intermediation due to collective dismissal file or short-time working arrangements of suspension or reduction of working hours	U
125	Suspension due to cause 125	I
509	Removal due to accumulated benefit payment caused by return to the country of origin	I
530	Suspension due to temporary inability with intermediation	I
531	Suspension due to maternity/paternity, adoption, or foster care with intermediation	I
614	Suspension due to family obligations with intermediation	U
620	Suspension with intermediation due to leaving the country	I
621	Suspension with intermediation due to attending training courses	U
625	Suspension due to assignment to social collaboration work* with intermediation	E
626	Suspension due to deferred coverage with intermediation	E

Table 10 (continued)

Code	*Description of the cause*	*Type*
627	Suspension due to deferred call with intermediation	E
700	Registration due to enrolment	U
701	Registration due to coverage of a vacancy (to be phased out)	E
702	Registration due to collective dismissal file	U
703	Registration due to correction of an erroneous removal	U
704	Registration with recovery of a period in a removal situation	U
706	Registration due to initial enrolment	U
707	Registration due to reactivation of suspension	U
708	Registration due to enrolment as employment intermediation	U
709	Registration as a jobseeker for other ALMPs	U
710	Registration for ALMP prior to employment	U
711	Registration due to benefit resumption-compatibility	U

Note: E: part of an employment spell, U: part of an unemployment spell, I: part of an inactivity spell

Appendix D

List of predictors of Q, C, and K

See Tables 11, 12, 13

Table 11 Predictors used in Q models

Group	*Predictor*
Job	Employability of the occupation of interest (5 categories)
	Employability of the sector of interest (3 categories)
General employment experience	Months of experience in regular employment (5 categories)
	Months of experience in irregular employment (4 categories)
Employment experience in the occupation	Months of experience in regular employment in the occupation of interest (5 categories)
	Months of experience in irregular employment in the occupation of interest (4 categories)
General training	Level of education (46 categories)
	Credential of non-formal learning of at least 80 h (2 categories)
	Driving license (2 categories)
Professional training	It is a closed occupation, and he/she has or is enrolled in the credential (2 categories)
	It is an occupation with a recommended credential, and he/she has or is enrolled in the credential (3 categories)
	It is not a closed occupation, and he/she attained or is enrolled in the secondary level of education (3 categories)
Job search	Knowledge and use of job search techniques (4 categories)
Language skills	Knowledge of Catalan or Spanish (4 categories)
Digital skills	ICT abilities (3 categories)
Transversal skills	Willingness to learn (2 categories)
	Proper communication (2 categories)
	Proper interpersonal relation (2 categories)

Source: Own elaboration based on screenshots of the Q software and SOC (2016)

Table 12 Predictors used in the C function

Group	*Predictor*
Used in formal allocation	Age
	He/she has a disability
	Duration of the unemployment spell
Not used in formal allocation	Sex
	He/she receives a benefit
	Geographical mobility
	Availability to work
	Availability to participate in ALMP
	Economic dependence

Source: Own elaboration based on SOC (2016)

Table 13 Predictors used in our statistical models (K)

Group	*Predictor*
PLMP	Unemployment started during a benefit interval
	Number of benefit spells (completed) in the past
	Total duration of previous benefit spells
	Total duration of previous benefit spells, scaled by age
	Mean duration of previous benefit spells
ALMP	Total duration of employment subsidy participations
	Number of JSA/JSM participations in the past
	Number of training participations in the past
	Total durations of JSA/JSM participations in the past
	Total durations of JSA/JSM participations in the past, scaled by age
	Total durations of training participations in the past
	Total durations of training participations in the past, scaled by age
	Mean duration of training participations in the past (in days)
	Mean duration of JSAM participations in the past (in days)
Unemployment	Number of unemployment spells in the past (inside the window)
	Total duration of unemployment spells in the past (until the present spell, not included)
	Total duration of unemployment spells in the past (until the present spell, not included), scaled by age
	Mean duration of unemployment spells until the present (until the present spell, not included)
	Days since last unemployment spell
Inactivity	Total duration of inactivity spells
	Mean duration of inactivity spells until the present
	Number of inactivity spells in the past
Employment	Days since first employment (in the window)
	Days since (the beginning of) the last employment spell
	Days since (the beginning of) the last full-time employment spell
	Occupation of last job by major groups (63 categories)
	Last job was part-time
	Skill level required for last job (11 categories)
	Last job was temporary
	Industry of last job (22 categories)
	Commuted for last job
	Proportion of jobs with commuting in the past
	Number of employment spells without any vocational training held in the past
	Number of occupations held in the past

Table 13 (continued)

Group	*Predictor*
	Number of employment spells in the past
	Number of open-ended contracts in the past
	Number of temporary contracts in the past
	Maximum skill level required for past employment spells (11 categories)
	Maximum skill level required for past employment spells (5 categories)
	Minimum skill level required for past employment spells (11 categories)
Sociodemographics	Sex
	Age in years when the spell started
	Maximum level of education
	Has a disability
	Local labor market (28 categories)
	National group (7 categories)
	Has a credential with field of education = xy (33 binary variables)
Missing blocks	Indicator of missingness on employment spells in the past
	Indicator of missingness on unemployment spells in the past
	Indicator of missingness on local labor market

Note: Qualitative variables that do not indicate the number of categories are binaries, so there are until three possible categories (yes, no, or missing). For the models that use regularization, this list is actually a list of candidate predictors. PLMP: passive labor market policies

Appendix E

Summary statistics

See Tables 14, 15

Table 14 Summary statistics on sociodemographic qualitative variables

	N	%
Sex		
Woman	153,424	52.412
Man	139,301	47.588
Maximum level of education		
0 Less than primary	4,965	1.696
1 Primary	8,137	2.780
24 Lower secondary – General	147,550	50.406
25 Lower secondary – Vocational	41	0.014
34 Upper secondary – General	27,382	9.354
35 Upper secondary – Vocational	40,327	13.776
55 Short-cycle tertiary – Vocational	32,071	10.956
66 Bachelor’s	13,153	4.493
76 Master’s	18,538	6.333
86 Doctoral	561	0.192
Disability		
No	273,733	93.512
Yes	18,992	6.488
National group		
Asia	788	0.269
EU, Northern America, and Oceania	281,292	96.094
Europe not EU	451	0.154
Latin America and the Caribbean	2,088	0.713
Northern Africa	6,352	2.170
Sub-Saharan Africa	1,749	0.597
Missing	5	0.002

Note: The categories related to the local labor market, the field of study and the level of study are not shown to simplify the exposition. The tables are available upon request

Table 15 Summary statistics on sociodemographic quantitative variables

	Mean	*Min*	*Q1*	*Median*	*Q3*	*Max*
Age	46.183	16	40	46	52	64

Appendix F

Tuning parameters

See Tables 16, 17

Table 16 Tuning grids

Model	*Parameter*	*Candidate values*
Penalized logistic regression (PLR)	Amount of regularization (penalty)	**0.001**, 0.01, 0.1, 1, 10, 100, 1000
	Proportion of Lasso penalty (mixture)	0, **1**
Random forest (RF)	Number of predictors (mtry)	sqrt(# predictors), **log2(# predictors)**
	Minimal node size (min_n)	**1**, 5, 10
	Number of trees (trees)	500, **750**
Gradient boosting machine (GB)	Tree depth (tree_depth)	3, 5, **7**
	Number of predictors (mtry)	**sqrt(# predictors)**, log2(# predictors)
	Number of trees (trees)	250, **500**, 750
	Learning rate (learn_rate)	0.01, **0.025**, 0.05
	Proportion of sampled observations (sample_size)	**0.6**, 0.8

Note: In the parameter column, it is shown in parenthesis the name given in the R library {parsnip} to that parameter. The selected value is written in bold in the third column. The unpenalized logistic regression has no internal parameter to tune

Table 17 Probability thresholds selected

Model	*Policy*	*Probability threshold*
Unpenalized logistic regression (LR)	A	0.2425
	B	0.5479
Penalized logistic regression (PLR)	A	0.24
	B	0.5424
Random forest (RF)	A	0.285
	B	0.5374
Gradient boosting machine (GB)	A	0.2675
	B	0.5453

Appendix G

Additional results

See Tables 18, 19, 20, 21 and Fig. 5

Table 18 Ranking performance of final models in the semi-restrictive test set

	ROC-AUC	*PR-AUC*
Q-S	0.584	0.413
LR	0.635	0.456
PLR	0.637	0.454
RF	0.726	0.587
GB	0.680	0.530

Table 19 Kappa statistic of models in the restrictive test set

	Kappa
Q-S25	0.045
Q-S50	0.067
Q-G	0.035
Policy A	
LR	0.172
PLR	0.182
RF	0.236
GB	0.248
Policy B	
LR	0.205
PLR	0.205
RF	0.287
GB	0.238

Note: The Kappa statistic discounts the amount of accuracy generated just by chance. Note that the chance-corrected accuracy of Q-S50 is low ($\kappa_{QS50} = 0.067$) and represents less than one third of the chance-corrected accuracy we could get with the random forest

Table 20 False alarm rates in the restricted test set

	FAR
Q-S25	0.054
Q-S50	0.767
Q-G	0.173
Policy A	
LR	0.531
PLR	0.526
RF	0.576
GB	0.452

Note: FAR = 1 – specificity

Table 21 False alarm rates in two strata of the test set

	FAR
Older jobseekers	
LR	0.514
PLR	0.522
RF	0.469
GB	0.374
Female and older	
LR	0.531
PLR	0.541
RF	0.454
GB	0.367

Note: FAR = 1–specificity

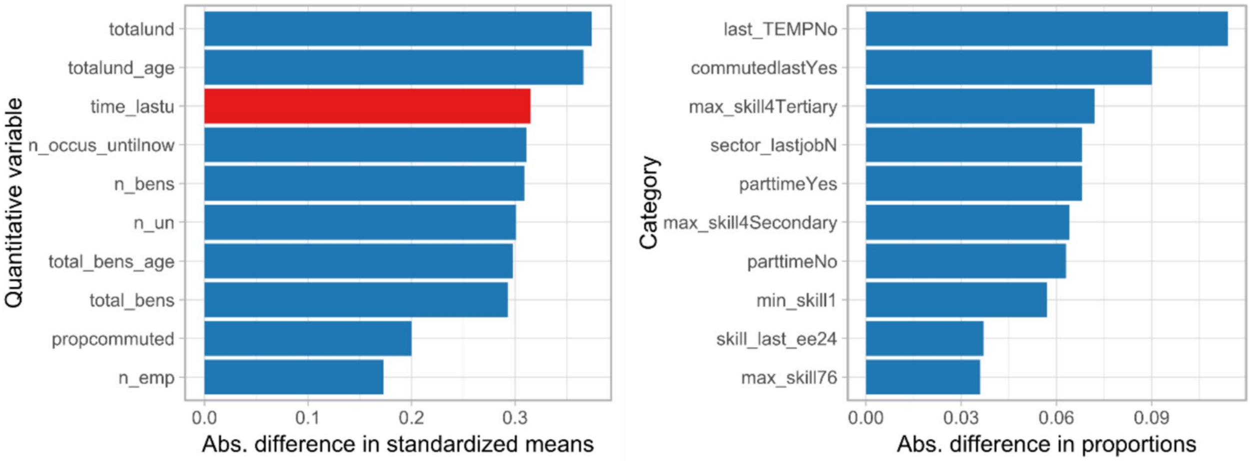

Fig. 5 Top 10 differences in standardized means (left) or proportions (right) between the test set and the restrictive test set. μ_g a summary statistic for the dataset g, shown the difference $\mu_{test} - \mu_{restrictive}$. Therefore, blue bars denote a positive difference, whereas red bars collect a negative difference Source: Own elaboration. Denoting with

Appendix H

Global surrogate model

See Fig. 6

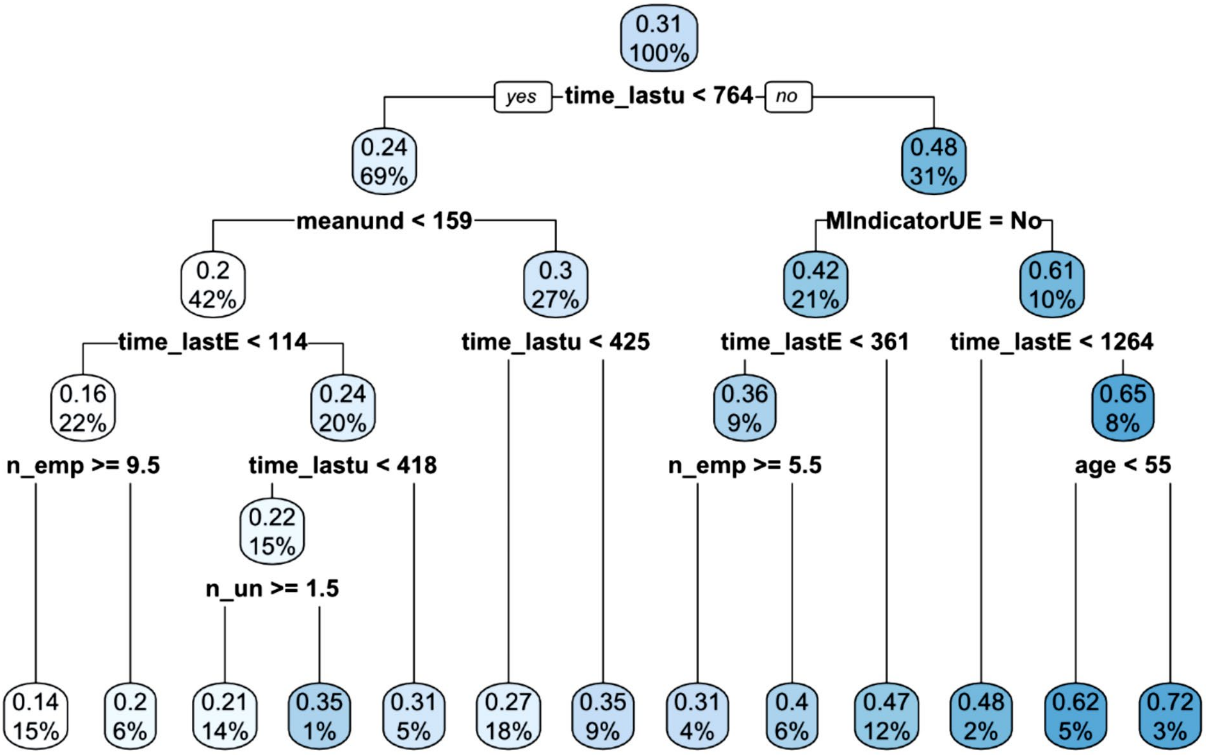

Fig. 6 Graphical representation of the decision tree

As an additional tool to interpretate how our statistical models make predictions, we have estimated a global surrogate model. The following regression tree model tries to forecast the predictions of the random forest model using 80% of the training sample. The tuning parameters were fixed at the following values: the cost-complexity parameter equaled 0.005, the tree depth was 30, and the minimal node size was established at 2. The resulting model has a $R^2_{training} = 0.778$ and a $R^2_{test} = 0.774$, attaining a good approximation to the random forest with a relatively low interaction depth.

As shown in Figure H1, the tree incorporates seven covariates: the number of days since last unemployment spell (time_lastu), the number of days since the beginning of the last employment spell (time_lastE), the mean duration of unemployment spells until the present (meanund), the number of employment spells in the past (n_emp), the number of unemployment spells in the past (n_un), the age (age), and the indicator of missingness on unemployment spells in the past (MIndicatorUE). Note that all the predictors selected by the global surrogate model were also highlighted as remarkably important by the permutation-based variable importance statistic.

To interpret Figure H1, we must consider that each node shows the probability of experiencing a long-term unemployment spell and below the percentage of the sample that fits in each partition. Starting the partition from above, we see that the combination of value that predict LTU with a probability equal to 0.72 is having the last unemployment spells at least 764 days ago, having the last employment spell at least 1,264 days ago, and being older than 55 years. This profile is in line with the literature and fits with 3% of our sample.

Acknowledgements
We thank the Public Employment Service of Catalonia and specially Josep Márquez for the data access and their help during the whole research process.

Author contributions
AFJ conceptualized and wrote the manuscript and conducted the data analysis. CK co-conceptualized the study and contributed to the writing and editing of the manuscript. Both authors have read and approved the final manuscript.

Funding
This research has been funded by the Ministry of Universities of Spain (FPU grant FPU19/05291) and Fundación Banco Sabadell (*Ayudas a la Investigación 2023*).

Availability of data and materials
The data for this study are drawn from the Servei Públic d'Ocupació de Catalunya (SOC) but were used under license and hence are not publicly available.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 October 2024 Accepted: 19 May 2025
Published online: 17 October 2025

References

Arni, P., Schiprowski, A.: Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung - Teilprojekt 2: Pilotprojekt Jobchancen-Barometer. IZA Research Report. (2015)

Austin, P.C., Steyerberg, E.W.: Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med. Res. Methodol. **12**, 1–8 (2012)

Austin, P.C., Steyerberg, E.W.: The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat. Med. **38**, 4051–4065 (2019)

Bach, R.L., Kern, C., Mautner, H., Kreuter, F.: The impact of modeling decisions in statistical profiling. Data & Policy. **5**, 32 (2023)

Barnes, S.-A., Wright, S., Irving, P., Deganis, I.: Identification of latest trends and current developments in methods to profile jobseekers in European public employment services: final report, http://ec.europa.eu/social/BlobServlet?docId=14173&langId=en, (2015)

Black, E., Raghavan, M., Barocas, S.: Model multiplicity: Opportunities, concerns, and solutions. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 850–863 (2022)

Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001a)

Breiman, L.: Statistical modeling: the two cultures. Stat. Sci. **16**, 199–231 (2001b)

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

Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning. pp. 161–168 (2006)

Casanova, J., Felgueroso, F., García Pérez, J.I., Jiménez-Martín, S.: El perfilado estadístico como instrumento para la evaluación del impacto del programa Incorpora. CICE. **102**, 189–220 (2021)

Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)

Consell de Direcció del SOC: Pla de desenvolupament de les polítiques d'ocupació de Catalunya 2023–2025, (2023)

Coston, A., Mishler, A., Kennedy, E.H., Chouldechova, A.: Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 582–593. ACM, Barcelona Spain (2020)

Cronert, A.: The multi-tool nature of active labour market policy and its implications for partisan politics in advanced democracies. Soc. Policy Soc. **21**, 210–226 (2022)

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Sculley, D.: Underspecification presents challenges for credibility in modern machine learning. J. Mach. Learn. Res. **23**, 1–61 (2022)

Desiere, S., Struyven, L.: Using artificial intelligence to classify jobseekers: the accuracy-equity trade-off. J. Soc. Policy **50**, 367–385 (2021)

Desiere, S., Langenbucher, K., Struyven, L.: Statistical profiling in public employment services. OECD Social, Employment and Migration Working Paper, Paris (2019)

Duell, N.H., Moraes, G.: Statistical Profiling - Lessons from OECD Countries, https://documents1.worldbank.org/curated/en/099011924113033615/pdf/P17655315c128d03218bbd1af6608050c69.pdf, (2023)

DG EMPL: LMP expenditure by type of action [LMP_EXPSUMM], (2024)

Elster, J.: Local justice: how institutions allocate scarce goods and necessary burdens. Russell Sage Foundation, New York (1992)

Ernst, S., Mueller, A.I., Spinnewijn, J.: Risk Scores for Long-Term Unemployment and the Assignment to Job Search Counseling. AEA Papers and Proceedings. 114, 572-576 (2024). https://doi.org/10.1257/pandp.20241092

Eurostat: Thematic glossaries, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Thematic_glossaries, (2024)(a)

Eurostat: Total unemployment rate [tps00203], https://doi.org/10.2908/TPS00203. (2024)(b)

Everis: Evaluación de implementación y de impacto de los Servicios de Orientación Profesional, https://serveiocupacio.gencat.cat/web/.content/01_SOC/09_Transparencia-i-bon-govern/Avaluacio-i-estudis/Avalua_Serveis_OP_-SOC_CAT.pdf, (2017)

Felgueroso, F., García-Pérez, J.I., Jiménez-Martín, S., Gorjón, L., García, M.: Herramienta de perfilado de parados: modelización y resultados preliminares. In: Felgueroso, F., García-Pérez, J.I., Jiménez-Martín, S. (eds.) Perfilado estadístico: un método para diseñar políticas activas de empleo. Fundación Ramón Areces, Madrid (2018)

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**, 3133–3181 (2014)

Filomena, M.: Unemployment scarring effects: an overview and meta-analysis of empirical studies. Italian Econ. J. **10**, 459–518 (2024)

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**, 1–81 (2019)

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)

Gabrikova, B., Svabova, L., Kramarova, K.: Machine learning ensemble modelling for predicting unemployment duration. Appl. Sci. **13** (2023). https://doi.org/10.3390/app131810146

Gabrikova, B., Svabova, L., Kramarova, K.: Machine learning ensemble modelling for predicting unemployment duration. Appl. Sci. **13**, 10146 (2023)

Gallagher, P., Griffin, R.: (in) Accuracy in Algorithmic Profiling of the Unemployed–An Exploratory Review of Reporting Standards. Social Policy and Society. 1–14 (2023)

Geirhos, R., Meding, K., Wichmann, F.A.: Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. Adv. Neural. Inf. Process. Syst. **33**, 13890–13902 (2020)

Harmon, N.A., Mahlstedt, R., Rasmussen, M., Rasmussen, M.: Helping the Unemployed Through Statistical Prediction?, https://www.econstor.eu/bitstream/10419/240564/1/phd-216.pdf, (2021)

Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts. (2018)

James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning: with applications in R. Springer, New York (2021)

Junquera, Á.F.: Efectos de la asistencia en la búsqueda de empleo: una revisión sistemática para España. Gestión y Análisis De Políticas Públicas. **35**, 7–25 (2024)

Kern, C., Gerdon, F., Bach, R.L., Keusch, F., Kreuter, F.: Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. Patterns. **3**, 100591 (2022)

Körtner, J., Bonoli, G.: Predictive algorithms in the delivery of public employment services. In: Clegg, D., Durazzi, N. (eds.) Handbook of labour market policy in advanced democracies, pp. 387–398. Edward Elgar Publishing, Cheltenham (2023)

Kuhn, M., Johnson, K.: Feature engineering and selection: a practical approach for predictive models. Chapman and Hall/CRC, Boca Raton (2019)

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

Kuppler, M., Kern, C., Bach, R.L., Kreuter, F.: From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. Front. Sociol. **7**, 883999 (2022)

Long, J.S., Freese, J.: Regression models for categorical dependent variables using Stata. Stata press, College Station (2006)

Loxha, A., Morgandi, M.: Profiling the unemployed: A review of OECD experiences and implications for emerging economies. Social Protection and Labor discussion paper, SP 1424, https://documents1.worldbank.org/curated/en/678701468149695960/pdf/910510WP014240Box385327B0PUBLIC0.pdf, (2014)

Marx, C., Calmon, F., Ustun, B.: Predictive multiplicity in classification. In: International Conference on Machine Learning. pp. 6765–6774. PMLR (2020)

Matty, S.: Predicting likelihood of long-term unemployment: The development of a UK jobseekers' classification instrument. Department for Work and Pensions Working paper. (2013)

McGuinness, S., Redmond, P., Kelly, E., Maragkou, K. Predicting the probability of long-term unemployment and recalibrating Ireland's statistical profiling model, (2022)

Molina Romo, O., Junquera, Á.F., Verd Pericàs, J.M., Sánchez Martínez, R., Úbeda Molla, P., Galobardes, M., Miró Martín, S.: PerfilaSP - Anàlisi de variables sociològiques i psicològiques com a predictores d'ocupabilitat en el perfilatge dels Serveis d'Ocupació de Catalunya, https://eapc.gencat.cat/web/.content/home/recerca/Convocatories_de_recerca/subvencions_a_la_realitzacio_de_treballs_de_recerca/2021/treballs-complets/2021_TR_ocupabilitat.pdf, (2023)

Mueller, A., Spinnewijn, J.: The nature of long-term unemployment: predictability, heterogeneity and selection. National Bureau of Economic Research, Cambridge (2024)

Muñiz, F.: Send@: Digitalización y uso masivo de datos para ayudar a encontrar trabajo - SEPE x JOBMadrid'20, https://www.youtube.com/watch?v=TNKVWL0pFRU, (2021)

Niklas, J., Sztandar-Sztanderska, K., Szymielewicz, K.: Profiling the unemployed in Poland: social and political implications of algorithmic decision making, (2015)

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**, 447–453 (2019). https://doi.org/10.1126/science.aax2342

Picchio, M., Ubaldi, M.: Unemployment and health: a meta-analysis. J. Econ. Surv. **38**, 1437–1472 (2022). https://doi.org/10.1111/joes.12588

Rebollo-Sanz, Y.F.: El modelo de perfilado estadístico: una herramienta eficiente para caracterizar a los demandantes de empleo. In: Felgueroso, F., García-Pérez, J.I., Jiménez-Martín, S. (eds.) Perfilado estadístico: un método para diseñar políticas activas de empleo. Fundación Ramón Areces, Madrid (2018)

Scott, K.M., Wang, S.M., Miceli, M., Delobelle, P., Sztandar-Sztanderska, K., Berendt, B.: Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 2138–2148 (2022)

SOC: Manual del model operatiu de les oficines de treball, (2016)

SOC: Pla d'acció per a persones en situació d'atur majors de 45 anys i de llarga durada 2023–2026, https://serveiocupacio.gencat.cat/web/.content/01_SOC/01_Qui-som-i-que-fem/Estrategia-per-locupacio/Pla-de-Desenvolupament-de-Politiques-dOcupacio-de-Catalunya-PDPO/Pla_accio_persones_atur_majors_45_CDSOC_20_07_2023.pdf, (2023)

Troya, I.M., Chen, R., Moraes, L.O., Bajaj, P., Kupersmith, J., Ghani, R., Zejnilovic, L.: Predicting, explaining, and understanding risk of long-term unemployment. In: 32nd Conference on Neural Information Processing Systems (2018)

Van Calster, B., McLernon, D.J., Smeden, M., Wynants, L., Steyerberg, E.W.: Topic group 'evaluating diagnostic tests and prediction models' of the STRATOS initiative: calibration: the Achilles heel of predictive analytics. BMC Med. **17**, 230 (2019)

Van den Berg, G.J., Kunaschk, M., Lang, J., Stephan, G., Uhlendorff, A. Predicting Re-Employment: Machine Learning Versus Assessments by Unemployed Workers and by Their Caseworkers. Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research]. (2024)

Viljanen, M., Pahikkala, T.: Predicting unemployment with machine learning based on registry data. In: Dalpiaz, F., Zdravkovic, J., Loucopoulos, P. (eds.) Research challenges in information science, pp. 352–368. Springer International Publishing, New York (2020)

Publisher's Note