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

Álvaro F. Junquera*

Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball, Institut d'Estudis del Treball, Universitat Autònoma de Barcelona.
Email: alvaro.junquera@uab.cat
*Corresponding author

Christoph Kern

LMU Munich, Munich Center for Machine Learning (MCML), University of Maryland

Abstract

Public employment services (PES) commonly apply profiling models to target labour 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 evaluate the predictive performance of a rule-based profiling procedure currently implemented by the PES of Catalonia, Spain, in comparison to the performance of statistical models in predicting future long-term unemployment (LTU) episodes. Using comprehensive administrative data, we develop logit and machine learning models and evaluate their performance with respect to both discrimination and calibration. Compared to the current rule-based procedure of 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 episodes 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 offices in Spain.

Keywords: algorithmic profiling, unemployment, public employment services, machine learning.

JEL codes: 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 episodes can have serious consequences for the affected individuals. These impacts include economic deprivation through the so-called scarring effects (Filomena 2023), but also 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 labour market problem is especially prevalent in countries like Spain or Greece with annual unemployment rates even doubling the average of the EU in 2023 (Eurostat 2024a). This has involved a high expenditure in passive labour market policies, positioning Spain at the second place of the European ranking with a 1.52 % of the GDP devoted to these programs in 2019. At the same time, comparatively limited funding is used in these countries to support active labour market polices such as job search interventions (DG EMPL 2024). In this situation, 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 via profiling procedures and allocate targeted support to increase labour market prospects. Accurately predicting adverse outcomes early on is a central concern in these efforts since support programs are sought to be allocated as preemptive measures. Given the promise of flexible machine learning models to achieve high prediction performance across various tasks (Caruana & Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014), there is an increased interest by PES in many countries to explore profiling approaches that draw on modern statistical models to improve the efficiency and effectiveness of current 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) either test or already implement machine learning models in their profiling practices.

However, assessing the potentials of statistical profiling in specific application contexts is a nuanced process and requires careful comparisons to the current profiling procedures that are already employed, which commonly include caseworker and rule-based approaches (Loxha and Morgandi 2014). While statistical models might draw on millions of data points to detect risk factors of LTU, caseworker and rule-based procedures may similarly make use of many years of “historical data” and institutional expertise and thus do not necessarily need to lead to inferior outcomes if those approaches are compared on the same grounds. However, such comparisons are complicated as detailed documentation of the specific profiling approaches employed by PES is often not available to the public. To the best of our knowledge, Desiere and Struyven (2021) and Van den Berg et al. (2024) are the few studies that explicitly compare the predictive performance of statistical profiling methods to rule-based and caseworker-based profiling implemented in the respective country (Belgium and Germany).

The contributions of this paper are as follows. First, utilizing a unique data base provided by the public employment office of Catalonia, we are able to compare their currently implemented rule-based profiling procedures with machine learning models in a highly realistic setting. The region of Catalonia is an interesting case of study due to its innovative use of data both to profile jobseekers and to evaluate public policies, something that is not typical in Spain (Junquera 2024). Second, we follow a broader vision of predictive performance in these comparisons, including measures of both statistical discrimination and calibration. This perspective recognizes that the predicted scores under any profiling approach should be an honest reflection of true labour market prospects as the mere reporting of such scores in counselling 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 rule-based approaches currently used in profiling practice on various metrics. We further highlight the need of tailoring 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 the allocation of individuals into labour market programs can be implemented via an allocation system with two stages, involving a decision and a profiling step. In the decision step, the decision-maker must establish an allocation principle, a function that maps individuals into 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 as decision criterion the value of an unobserved variable at the time of decision. In the profiling step, such 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 and decision steps further helps to channel recent critiques in the social policy literature to the emphasis on accuracy made by previous research on statistical 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 go deeper into the similarity of the predictions of different models and their interpretation, taking into account the importance of human discretion when 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 through 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 supporting a posterior action like allocating individuals into treatments, although they can also be used to describe jobseekers more succinctly or as an intervention of information provision (Harmon et al. 2021; Loxha and Morgandi 2014). In-depth reviews of jobseeker profiling models are available at Duell and Moraes (2023), Desiere et al. (2019), or Barnes et al. (2015). Here we focus on strands of the literature related to our research: the profiling performance as a function of the degree of human discretion and the application of these models in public employment offices.

It is common to distinguish three types of profiling models that differ in the degree of human discretion: caseworker-based, rule-based, and statistical profiling (Desiere et al. 2019; Rebollo-Sanz 2018). Caseworker-based procedures allow each counsellor to have her own model, which is often implicit and unknown. Rule-based and statistical profiling establish a common model for all caseworkers and jobseekers. The difference lies in the specification of such function. 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]. Regarding rule-based models, to our knowledge, only Desiere and Struyven (2021) have studied the discrimination ability of a rule-based procedure. This model, deployed in Belgium, attains an accuracy of 0.58 and has a higher false alarm rate for foreign (non-Belgium) individuals compared to Belgian nationals. The authors conclude that their statistical model, in contrast, would allow 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 mimic 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 has been applied at least in Ireland, Norway, Poland, and United Kingdom. In practice, they might be more prevalent. It is common that entry into active labour market programs is governed by specific eligibility criteria (Cronert 2022) and they may be understood as a consequence of implicit rule-based models. Nonetheless, ruled-based approaches have not been sufficiently studied and their specific implementation details are hardly reported by PES.

[1] Along this paper we use "discrimination" with the meaning it has in the biostatistical literature (Austin & Steyerberg, 2012), i.e. as the ability 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.

Concerning statistical models, the literature is more extensive. They have been implemented and publicly assessed by the PES of Austria, Belgium, Denmark, Ireland, or Netherlands (Desiere et al. 2019). Pooling all of them, their prediction accuracies range from 0.61 to 0.85. Academics 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); Finland, with a ROC-AUC of 0.8 (Viljanen and Pahikkala 2020), or 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 articles 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 ALMP (Casanova et al. 2021). Molina Romo et al. (2023) studied the prediction ability of personality traits, personal networks, and job expectancies separately. Their results go in line with the findings of Van den Berg et al. (2024), with expectancies as a remarkable predictor of long-term unemployment in both cases. Our research tries to integrate both perspectives by constructing a long panel of episodes that incorporates information on lagged outcomes, which are possibly related to unobservables (Caliendo et al. 2017).

[3] There is a tool called “Send@” developed by the central public employment office (SEPE) that is closer to a targeting model in the sense of Körtner and Bonoli (2023). Profiling models try to predict a potential outcome after no intervention ($\Pr[Y(0) = y|X]$), whereas targeting models focus on a vector with an element for each potential outcome after going to a certain intervention ($\mathbf{v} = (\Pr[Y(d_1) = y|X], \Pr[Y(d_2) = y|X], \dots, \Pr[Y(d_K) = y|X])$). According to Muñiz (2021), Send@ detects those individuals that had certain covariates values $\boldsymbol{X} = \boldsymbol{x}$ with the highest improvement in labour insertion ($i \in Best_x$). Then, it offers two sorted vectors of conditional probabilities on interventions in which they participated ($\mathbf{v_1} = (\Pr[D = d_1|i \in Best_x], \dots, \Pr[D = d_K|i \in Best_x]$)) and on the occupations of interest of these individuals ($\mathbf{v_2} = \left(\Pr[O = o_1|i \in Best_x], \dots, \Pr\left[O = o_J\middle|i \in Best_x\right]\right)$).

[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.

1.2. Current profiling in Catalonia

The public employment office of Catalonia (*Servei Públic d'Ocupació de Catalunya;* SOC) already has an allocation system for jobseekers in place. Its profiling model is a mixture of a caseworker and rule-based procedures, where the latter is used for assisting office workers in allocating individuals to interventions. It includes an allocation principle for the first two interventions of each unemployment episode experienced by an individual[5], which facilitates her placement among a set of job search assistance actions. Still, the office admits that the scores of its profiling step might also assist future decisions (SOC 2016). This 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 readily be evaluated empirically. Let us review the inputs, processing, and outputs of the Q models. They take as input administrative data on labour markets and data collected through a questionnaire administered to the jobseeker. The first set of variables incorporate information on the economic environment, especially unemployment rates by occupation and sector. The second set includes covariates on work experience and on the skills of the individual[6]. Note that it does not consider variables on individual unemployment or inactivity episodes in the past. The processing of the information is made through two functions: a rule-based model that assigns a number to each individual representing how employable they are (Q-S) and a step function that assigns the individual to a certain group (Q-G).

The important issue is that the weight attached to each variable was not assigned through a statistical method, but rather through human intuition. Q-S is a sum of coefficients attached to qualitative variables, whereas Q-G may be understood as a decision tree. Thereby, we end up with

[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.

two outputs: a continuous value on the assessed employability ($S_{it} \in [0,139]$) and a discrete value for the assigned group ($G_{it} \in \{\mathrm{A1, A2, A3, A4, B1, B2, B3, C1, C2, D, Z1, Z2, Z3, E, R6}\}$). Although they are not explicitly framed as predictive models, we argue that we can interpret Q-S and Q-G as intended proxies of the (long-term) unemployment probability.

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 a intermediation claim (*demanda de empleo*), changes of such claim, or the ending of an ALMP. The implementation of the profiling process was analysed 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 have been granted access to administrative data provided by the public employment office of Catalonia (SOC). Four datasets have been matched: the dataset on employment claims (SICAS), on labour contracts (Contrat@), on active labour market programs (Galileu), and on benefits or passive labour market programs offered by the central PES. In a first stage, we have obtained a simple random sample of 25,000 individuals for four focal years (2017, 2018, 2019, and 2022) from the population of individuals registered as unemployed in that year. In a second stage, we have extracted information on selected variables for each sampled individual of each dataset for the time window [2015, 2023].

The next step has been the construction of the dataset of labour market episodes and the dataset of policy episodes. An episode is simply defined as a closed interval of time started at day t and ended at day t' by individual i. The first kind of episodes collects episodes of participation in the labour market, whereas the second registers episodes of participation in active or passive labour market policies. For a given individual, labour market episodes are non-overlapping time intervals, but policy episodes may overlap in time.

We distinguish three types of labour market episodes: employment episodes, unemployment episodes, and inactivity episodes. A new labour contract configures a new employment episode, whilst unemployment and inactivity episodes are defined according to the type of intermediation claim registered[7]. An exhaustive map of types of claims to the distinction of unemployment or inactivity is available in Appendix C. Some of the factors that define episodes of inactivity are temporary inability, permanent inability, prison entry, or family care. Thereby, unemployment or inactivity episodes are defined as the presence or succession of intermediation claims of such type. The dataset of policy episodes distinguishes four types of episodes: participation in adult training, participation in job search assistance or brokering, participation in an employment subsidy, and receipt of a benefit.

Table 1 – Sample size and events of interest by year in which the episode started.

Year	*Unemployment episodes*	*LTU episodes*	*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 %)

The final step has been to compile a dataset of unemployment episodes. Concerning the outcome, following the bulk of the literature on jobseeker profiling (Körtner and Bonoli 2023), our response variable identifies whether an unemployment episode is a long-term unemployment episode. An unemployment episode is defined as long-term if it lasts at least 365 days (Eurostat 2024b). Table

[7] In Spanish, "*demanda de empleo*".

1 summarizes the number of jobseekers and unemployment episodes by year and the prevalence of the event of interest, i.e. long-term unemployment (LTU).

Regarding predictors, our data includes both time-invariant and time-variant covariates. Like in Bach et al. (2023), we have condensed the time-variant information on past labour market and policy episodes 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 with an adaptation to the Catalan setup. A complete list of predictors is available in Appendix D and summary statistics on the socio-demographic features of our sample are presented in Appendix E.

To compare our prediction models with the current profiling approach of the SOC, we use an extra dataset with profiling scores derived from the rule-based `Q` model. The current implementation of `Q` allows that an individual may receive one G_{it}, but more than one S_{it} for the same episode (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 have calculated S_{it} as the average of the score obtained for each occupation of interest.

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 episodes, number of unemployment episodes, days since last unemployment episode…
Inactivity	3	Total duration of inactivity episodes, number of inactivity episodes, mean duration of inactivity episodes.
Benefits	5	Started the unemployment episode during a benefit interval, number of benefit episodes completed, total duration of benefit episodes…
ALMP	9	Total duration of job search assistance episodes, total duration of adult training episodes, total duration of employment subsidy participations…
Socio-demographics	37	Sex, nationality, age, field of education…

2.2. Analytical strategy

2.2.1. Development of models

We build profiling models based on four prediction techniques, covering conventional regression models and tree-based machine learning algorithms: unpenalized logistic regression, penalized logistic regression (Friedman et al. 2010), random forest (Breiman 2001b), and gradient boosting machine (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. The problem is that, however, this functional form is often poorly justified. Machine learning methods are, on the other side, highly flexible regarding the relationship between predictors and the outcome. Nonetheless, that flexibility provokes a lower degree of interpretability.

To estimate all models, we follow the dataset partition that is usually applied in the machine learning literature to avoid overfitting and provide realistic evaluations (Kuhn and Johnson, 2019). Thereby, the data is split into three subsets: training, evaluation, and test data. The training subset is used to tune the internal parameters of the methods (if any) and to estimate the coefficients of the model. The evaluation subset is employed to select the probability threshold to assign the estimated class (i.e., LTU or non-LTU). The final models are assessed with the test subset. The training, evaluation and test subsets are constructed through two partitions. Firstly, following Bach et al. (2023), we assign the observations from years 2017 to 2020 to a training plus evaluation subset and reserve the 2022 data for the test subset. Secondly, we use a stratified random resampling to separate the training (80 % of units) and the evaluation (20 % of units) subsets. We use the outcome as the stratifying variable to guarantee a sufficient presence of events. Lastly, the hyperparameters of each model are tuned in the training subset with respect to ROC-AUC through temporal cross-validation (Hyndman & Athanasopoulos, 2018), departing from the grid of candidates available in Appendix F.

The test subset is further reduced to a *restrictive* test subset. Note that one of the contributions of our article is the comparison of model coefficients estimated by humans (rule-based models) and statistical methods. This requires a test dataset in which the predictions of both the current (`Q`) and the proposed (`K`) models can be compared. To achieve this, we take the episodes already profiled with `Q` in 2022 and predict the score/class they had received in case they had been profiled with our models. We apply two restrictions to this dataset of Q-profiled episodes in 2022 to have a fair and realistic comparison between both profiling approaches.

The first filter levels the playing field between the current and the proposed models. The reason is that the variable being predicted is eventually also affected by the (prediction-based) interventions. That is, if the allocation had followed the recommendations of the `Q` predictions and the ALMP had positive effects on re-employment, the current profiling 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 bypass this problem, we are going to define A_{ie} as the number of ALMPs in which the individual i participated during the unemployment episode e. Therefore, removing those episodes with $A_{ie} > 0$, we assessed the models with the data $test\backslash\{A_{ie} > 0\}$. The second filter focuses the attention on the target groups of the public employment office of Catalonia. Nowadays, this agency refers people who do not know neither Catalan nor Spanish to other public administrations to give them other treatments. Thereby, 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 episodes related to persons without knowledge neither of Catalan nor Spanish. After these two restrictions, we ended up with a so-called *restrictive* test subset of our data.

2.2.2. Validation of models

To validate the models, we focus on two dimensions of performance: discrimination and calibration. Discrimination is the usual objective of researchers 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 impacts in practice. It is especially important for employment services, since caseworkers can inform jobseekers about their predicted risk as a form of intervention to foster an intensified job search. Such interventions may trigger important individual decisions and thus we need reliable predictions.

Concerning discrimination, firstly, 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 on the discriminatory performance of the models while remaining agnostic (silent) regarding the classification threshold. Note that they can only be computed for profiling functions that output a value measured at the ordinal level.

Secondly, we assess classification performance through three metrics: accuracy, precision, and sensitivity. The accuracy statistic gives the same weight to correct predictions of events (LTU) and non-events (non LTU). It is reasonable to assume that employment offices are more interested in detecting events than non-events, and the sensitivity statistic is calculated for this purpose. Nonetheless, classifying all episodes as predicted events would attain a 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[8].

Unlike ranking performance metrics, these quantities need the establishment of a threshold 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 classified as predicted events those episodes that were originally assigned to the groups linked to the most intense treatments (individual interventions to set the jobs of interest)[9] (SOC, 2016). For the score profiling (Q-S), we consider two options. Firstly, we apply the standard procedure

[8] Appendix G reports additional results on two more metrics of classification performance: the Kappa statistic (a chance-corrected version of accuracy) and the false alarm rate (or rate of false positives).

[9] Appendix A describes in detail the different interventions.

for probability models: label a unit as “event” if its value is closer (or equally close) to the upper limit of the measure[10]. We called this $\hat{Y}_{S50}$, since with a probability measure the class is equal to $\widehat{event} = Yes$ if $\widehat{\Pr}[event] \geq 0.5$. Secondly, 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 followed the functions: $\hat{Y}_{S25} = \begin{cases} Yes & S \leq 0.25(139) \\ No & S > 0.25(139) \end{cases}$; $\hat{Y}_{S50} = \begin{cases} Yes & S \leq 0.5(139) \\ No & S > 0.5(139) \end{cases}$; $\hat{Y}_G = \begin{cases} Yes & G \in \{\mathrm{C1}, \mathrm{C2}, \mathrm{D}\} \\ No & G \notin \{\mathrm{C1}, \mathrm{C2}, \mathrm{D}\} \end{cases}$.

The four techniques employed for our models output an estimated probability of LTU, which we denote $\hat{Y}$. To transform this estimated score to an estimated class, we applied two different policies that represent two different rationales.

Classification policy A interprets probabilities as propensities by understanding binary phenomena as the output of a latent variable model (Long and Freese, 2006). Thereby, the probability threshold is a parameter that *exists* and whose value may be learned. The probability threshold will be denoted by C and will be considered as a tuning parameter. Specifically, this tuning parameter will be learned in the additional evaluation subset. We assume that SOC is more interested in increasing sensitivity (detecting the true events of interest, i.e. the true LTU episodes), but not at any cost. Therefore, the cross-validation will try to maximize the Youden’s J, an equal compromise between specificity and sensitivity. Following the taxonomy of Elster (1992), this classification policy is in line with an *admission* procedure for allocating goods, since it does not establish the number of treatment slots in advance.

Classification policy B follows the rationale of a limited budget to fund public policies. The logic is that public administrations can only pay a finite number of services. To fix the number of predicted high-risk individuals, this function classifies as high-risk jobseeker only those individuals whose estimated probability is at least equal to the nineth decile of the predicted

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

probability estimated with the evaluation subset ($\widehat{\mathbb{D}}_9^{(eva)}$)[11]. If the office wanted to treat only those at the top of the distribution, it would require anticipating the value of $\widehat{\mathbb{D}}_9^{(eva)}$ for the current year to allocate individuals immediately without the need to cumulate all the candidates. The reasoning is that the demand of 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 policy fits with a *selection* procedure for allocating goods, because it is a relative allocation based on a ranking of candidates.

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

Regarding calibration, two statistics are calculated following two stringency levels of this dimension. Firstly, mean calibration is approximated through the ratio of the proportion of observed events divided by the proportion of expected events, denoted as O:E (Van Calster et al., 2019). Secondly, moderate calibration is assessed through flexible calibration curves summarized with the integrated calibration index proposed by Austin & Steyerberg (2019). 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.

Note that to evaluate our models we make predictions at the beginning of each unemployment episode (for the test subset) or at the moment of the Q prediction (for the restrictive test subset, see section 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 deploying one model rather than the other. To measure how

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

prevalent this phenomenon is in our case, we use Cohen's Kappa to approximate the degree of overlap between the predictions of models once agreement by chance is subtracted (Geirhos et al. 2020).

We further apply 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 incorporate this approach by evaluating models with the restrictive test data. The second test is named stratified performance evaluations and analyses whether performance metrics are similar in certain strata of the population. We know that SOC (2023) is specially interested in two subpopulations, older jobseekers (> 45 years old) and older female jobseekers, and thus we focus on these groups.

Lastly, to facilitate the interpretation of the importance of each predictor in our models, we estimate permutation-based variable importance (Fisher et al. 2019). Specifically, we consider 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 in the jobseeker profiling setup as it can provide valuable information for caseworkers.

To foster transparency and replicability, we publish all the R code necessary to construct both the datasets and the statistical models[12].

3. Results and discussion

3.1. Performance comparison

We present the performance of all techniques for predicting LTU in the test subsets. In a first step, we focus on a comparison of our statistical profiling models. Next, we assess the performance of

[12] Code is available at https://osf.io/jye6q/?view_only=3ef06ff290214bfd88f77954d7fb1b73.

the statistical models against the current profiling tools in Catalonia to assess whether changing the profiling scheme may be worthwhile to consider.

The first results evaluate 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 at the ROC and at the PR functions, but the improvements they present are modest. The gradient boosting model wins in both cases, followed by random forests, which is reasonable due to the flexibility of these techniques. Considering that a ROC-AUC of 0.5 would be simply a product of chance and that this statistic reaches its maximum at 1, the four models achieve notable good performance. 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-months interval. Using the same temporal window, the results of Bach et al. (2023) are very close to ours.

Table 3 - Ranking performance of final models in the test subset (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

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. Thereby, it is desirable to be close to 1. Note that all models overestimate the probability of an LTU event, although the GB algorithm most closely approaches the actual probability. Considering a more stringent measure of calibration, the ICI informs on the average error of the predicted probabilities, so it is better to be close to 0. This time the differences between models are smaller, but the gradient boosting machine wins again. Table 3 shows that the average error when predicting the probability of LTU is 11 p. p. when using this type of prediction model. To our knowledge, we are the first in the jobseeker profiling literature to measure calibration in this fine-grained sense.

A pertinent question for the public employment office is whether the adoption of predictive models is really worth the effort. To answer that question, we present in Table 4 performance metrics for a comparison of the current 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 random forest 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 episode 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 episode had higher predicted probability is 59.3 %. Compare this discrimination ability with the 73.5 % concordance probability of the random forest. The performance gap between random forest and Q-S is even larger if we attend to the precision-sensitivity function.

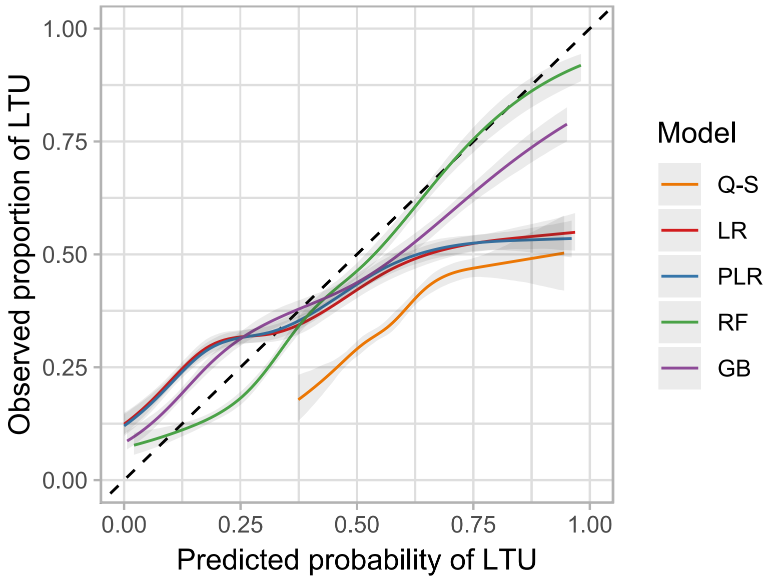
Table 4 - Ranking performance of final models in the restrictive test subset.

	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

Concerning calibration, the improvements obtained with the best statistical model are even bigger. The ICI column of the Table 4 shows that the average error of the random forest model is small (only 3.7 p. p.). If we compared it with the current model, we would require multiplying it by six to obtain the average error of the Q-S model. In case we considered a softer version of calibration, the gradient boosting model would be 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 remarkably better than the rule-based model. This might suggest that the restrictive test subsample in which we are re-evaluating these techniques has not the same covariate distribution as the full test subset. We checked the first moment of the predictors and found that the highest differences were in the proportions of people whose last job was not temporary (11.4 p. p. more in the test subset), required commuting (9 p. p. more), or with a tertiary employment episode (7.2 p.p. more)[13]. The gradient boosting model estimates its parameters 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" model.

Figure 1 – Calibration curves in the restrictive test subset for each model.



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

The second set of results take side on the probability threshold to classify an episode as high-risk. Table 5 presents the classification performance of our models in the (full) test subset, i.e. with data from 2022. Following policy A, the classification that uses an optimized threshold, we can see how the first three methods (LR, PLR, RF) present very similar results on the discrimination metrics. Accuracy and precision slightly improve with the random forest, although it is the gradient boosting machine which 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, notice that the gradient boosting model obtains a worse sensitivity result, which might indicate overfitting. The RF model achieves a better sensitivity than the statistical profiling model proposed for Spain in Felgueroso et al. (2018), who attained a sensitivity of 0.682[14].

Following policy B, the framework that prioritizes the budget, the results are much better in terms of accuracy. On the other side, there is a substantial decrease in sensitivity specifically for the tree-based methods. These techniques correctly predict the outcome classes for 83 % of the episodes, but the true positives do not represent a remarkable share of these forecasts. Continuing with the budget constrain, a compromise between policy A and B might be to use an alternative outcome variable: the duration of the unemployment episode 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.

[14] Notice that if we had followed their same procedure, we could have gotten even higher sensitivity. They chose the probability threshold with the test data while measuring sensitivity, whereas we fixed it in a previous step using the evaluation subset.

Table 5 – Classification performance in the test subset based on different policies.

	Accuracy	Precision	Sensitivity
Policy 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
Policy 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

It is interesting to compare the classification performance of our models with the current methods. Setting a specific threshold also allows to assess the discrimination ability of the Q-G profiling model. Table 6 presents the results of such comparison, this time using data from the restrictive test subset. 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 episodes, 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 subset. In a nutshell, we have higher specificity when based on thresholds optimized with Youden's *J* (policy A) and higher accuracy when focused on the budget (policy B). When sensitivity is placed as a high priority, the random forest model under policy A is the model that performs best in this task. With a sensitivity of 0.860, this algorithm surpasses the discrimination ability of the rule-based Q-S50. Moreover, it improves it substantially both at accuracy and precision. The gradient boosting model may serve as a compromise between policy A and B, since it attains a good sensitivity but maintains decent results in accuracy and precision.

Table 6 – Classification performance in the restrictive test subset based on different policies.

	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
Policy 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
Policy 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 subset = 11,082.

With the analysis of calibration and discrimination, we have shown that our random forest model using policy A (RF-A) outperforms the rule-based model Q-S50 in all the metrics. Its added value is especially remarkable in the reliability of its predictions, since the Q-S50 is poorly calibrated. On the contrary, 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 rise the efficiency of treatment assignments.

3.2. Model similarity and interpretation

In this section, we dig into the specific episodes flagged by each method and explore how the statistical models utilize the training information. We first measure the degree of model similarity, followed by results of stress tests and the interpretation of 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, the agreement between models is quite low. This might be explained by the fact that the rule-based models mainly represent random classifiers[15]. 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 may be highlighted. Firstly, as expected due to the low penalization of the tuned PLR models (see Appendix F), the agreement with the model predictions of LR is almost perfect for both classification policies. Secondly, the two big competitors in terms of performance (RF and GB) have an intermediate agreement, especially for policy A. This invites us to review the consequences of choosing one model instead of the other.

Table 7 – Kappa coefficients between predictions of different models in the restrictive test subset.

	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				
Policy 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
Policy 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

When deciding which model should be deployed, consequences of model discrepancies may be clarified with so-called stress tests. The first test, the shifted performance evaluation, was carried out when analysing the differential discrimination and calibration of models with the restrictive test data. Prioritising sensitivity, the random forest had the best performance and also attained the highest degree of calibration measured through the entire probability range. The second test, the

[15] 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.

stratified performance evaluation for older jobseekers and older female jobseekers, is presented in Table 8. Again, the RF model shows better performance than GB in terms of sensitivity for both subpopulations. However, note that this time simpler models (LR and PLR) do similarly well in predicting events in these groups at the cost of a lower precision and a lower accuracy. In the end, the model selection should take the costs structure of SOC into account. GB models offer the lowest sensitivity both for older and for female older jobseekers but have the largest accuracy. Thereby, if detection of non-events is considered more important, this model could also be implemented.

Table 8 – Classification performance in two strata of the test subset.

	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

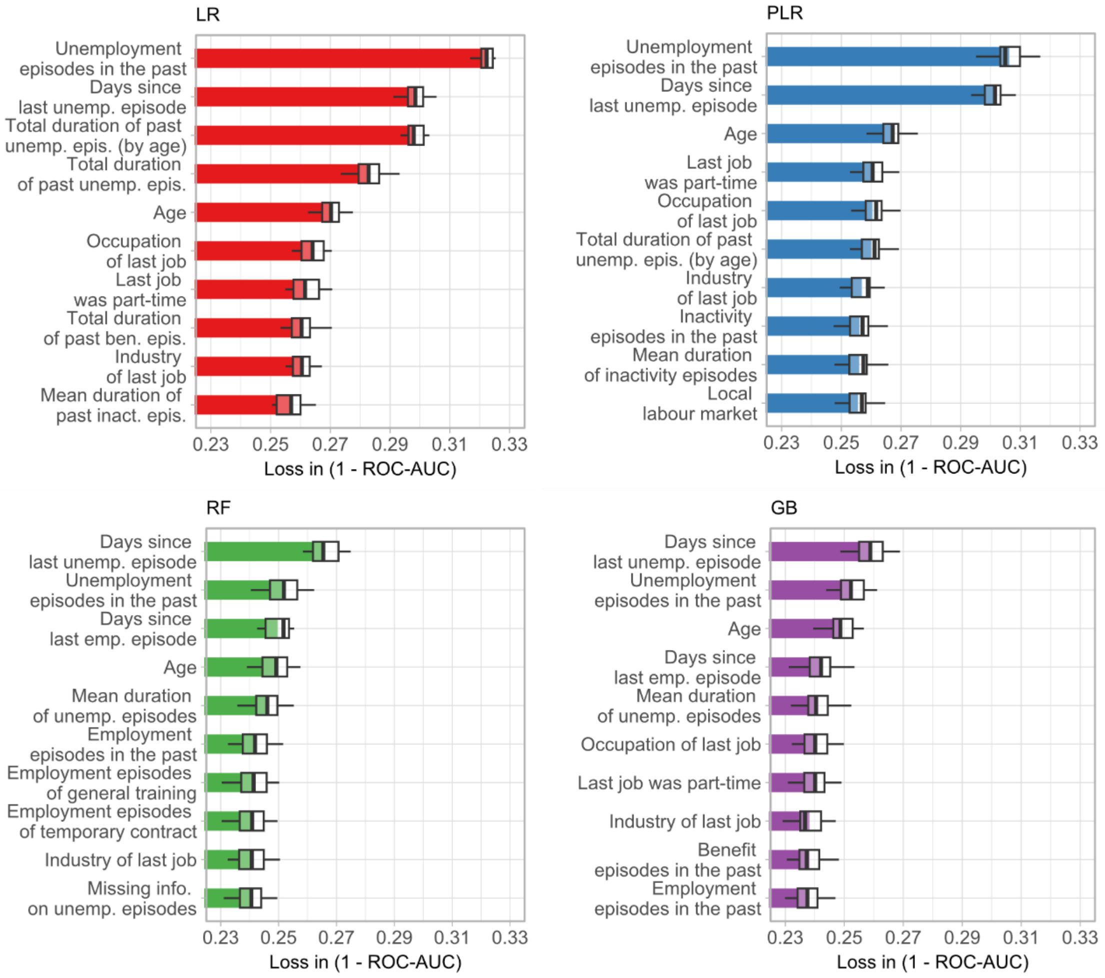
Note: *N* of the older stratum = 20,793. *N* of the female older stratum = 11,452. To simplify the exposition, we only showed the results for policy A.

Lastly, to understand how the statistical models make predictions, we explore the most important predictors used by each method[16]. Figure 2 shows the ranking of the ten most important covariates as measured by the loss in ROC-AUC provoked by shuffling their values. There is an agreement between the four models that the two most important predictors of LTU are the number of days since the last unemployment episode and the total number of unemployment episodes experienced in the past. Another variable that rates 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.

[16] In the Appendix F, we also offer a global surrogate model (a decision tree) to interpret how the estimates of the random forest were produced.

(2018) for Spain. Looking closer at the tree-based models, the number of days since the last employment episode and the average duration of unemployment episodes in the past are also important predictors of LTU. These results are consistent with the findings of Bach et al. (2023), who also indicated the high predictive ability of age and labour market histories. McGuinness et al. (2022) developed a model for Ireland and detected that employment histories were also a remarkable set of predictors.

Figure 2 – Top-10 variable importance of final models.



Note: The extension of the bar indicates the permutation statistic, which is a mean across permutation rounds, joint with the boxplot collecting variability between rounds.

4. Conclusions

In this article, we have contrasted the performance of rule-based versus statistical models for jobseeker profiling. Specifically, we have taken the predictions of the rule-based models currently deployed in Catalonia and compared them with newly developed statistical models for predicting long-term unemployment. Our results show that our statistical models outperform the current rule-based profiling approach considerably both in terms of discrimination (ROC-AUC: 0.735 vs. 0.593) and in terms of 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) targeted in the current operations of the employment services in Catalonia. Our prediction models additionally highlighted two important predictor variables that are not utilized in the current profiling approach: the number of days since the last unemployment episode and the total number of past unemployment episodes.

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 labour market transitions to accurately predict the risk of long-term unemployment (Gabrikova et al. 2023; McGuinness et al. 2022). Previous literature, however, has highlighted that more flexible methods like random forests do not make a big difference in performance compared to conventional models like 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 performance revealed that machine learning models can improve over regression approaches in terms of calibration, a crucial but overlooked dimension in the jobseeker profiling literature. We propose to carefully consider calibration in the evaluation of profiling models due to the crucial role of the (predicted) risk scores in the counselling practices of employment offices.

Our work also presents some limitations that need to be considered. Firstly, compared with related work such as Bach et al. (2023), our set of covariates on past employment episodes was limited due to the unavailability of information on the actual end dates of labour contracts. This shortcoming might be tackled in the future by getting access to detailed social security data. Secondly, our models have been trained with individuals that actively engage with the public employment offices in Catalonia. In Spain, registration with PES is not compulsory, which implies that the population of participants in PES may not mirror the full population of jobseekers. This implies that some groups like young people may be underrepresented in our training set in comparison with their presence in the population of jobseekers in general. Thirdly, 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. Furthermore, 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, by e.g. 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 caseworkers and by jobseekers. This line of research has been explored by Kern et al. (2022) and Scott et al. (2022), who have offered some possible explanations on the perception of uses of these models. Despite these potential extensions, our work illustrates the added value of flexible statistical models versus rule-based profiling to assist PES and also 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.

Declarations

Availability of data and materials: The data that support the findings of this study are available from Servei Públic d'Ocupació de Catalunya (SOC) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

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

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*).

Authors' 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.

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.

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 medical research methodology. 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. Statistics in medicine. 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)

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)

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. Machine learning. 45, 5–32 (2001)(a)

Breiman, L.: Statistical modeling: The two cultures. Statistical science. 16, 199–231 (2001)(b)

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

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 medicine. 17, 230 (2019)

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., Pérez, J.I.G., Jiménez-Martín, S.: El perfilado estadístico como instrumento para la evaluación del impacto del programa Incorpora. In: Cuadernos económicos de ICE (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)

Cronert, A.: The multi-tool nature of active labour market policy and its implications for partisan politics in advanced democracies. Social Policy and Society. 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. Journal of Machine Learning Research. 23, 1–61 (2022)

Desiere, S., Langenbucher, K., Struyven, L.: Statistical Profiling in Public Employment Services. OECD Social, Employment and Migration Working Papers. 224, (2019)

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

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

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

Elster, J.: Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens. Russell Sage Foundation, New York (1992)

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., and Jiménez-Martín, S. (eds.) Perfilado estadístico: un método para diseñar políticas activas de empleo. Fundación Ramón Areces (2018)

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research. 15, 3133–3181 (2014)

Filomena, M.: Unemployment scarring effects: An overview and meta-analysis of empirical studies. Italian Economic Journal. 1–60 (2023)

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. Journal of Machine Learning Research. 20, 1–81 (2019)

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 33, 1 (2010)

Gabrikova, B., Svabova, L., Kramarova, K.: Machine learning ensemble modelling for predicting unemployment duration. Applied Sciences. 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. Advances in Neural Information Processing Systems. 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)

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, (2022)

Körtner, J., Bonoli, G.: Predictive algorithms in the delivery of public employment services. In: Handbook of Labour Market Policy in Advanced Democracies. pp. 387–398. Edward Elgar Publishing (2023)

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

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. Frontiers in sociology. 7, 883999 (2022)

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

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)

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)

Picchio, M., Ubaldi, M.: Unemployment and health: A meta-analysis. Journal of Economic Surveys. (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., and Jiménez-Martín, S. (eds.) Perfilado estadístico: un método para diseñar políticas activas de empleo. Fundación Ramón Areces (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)

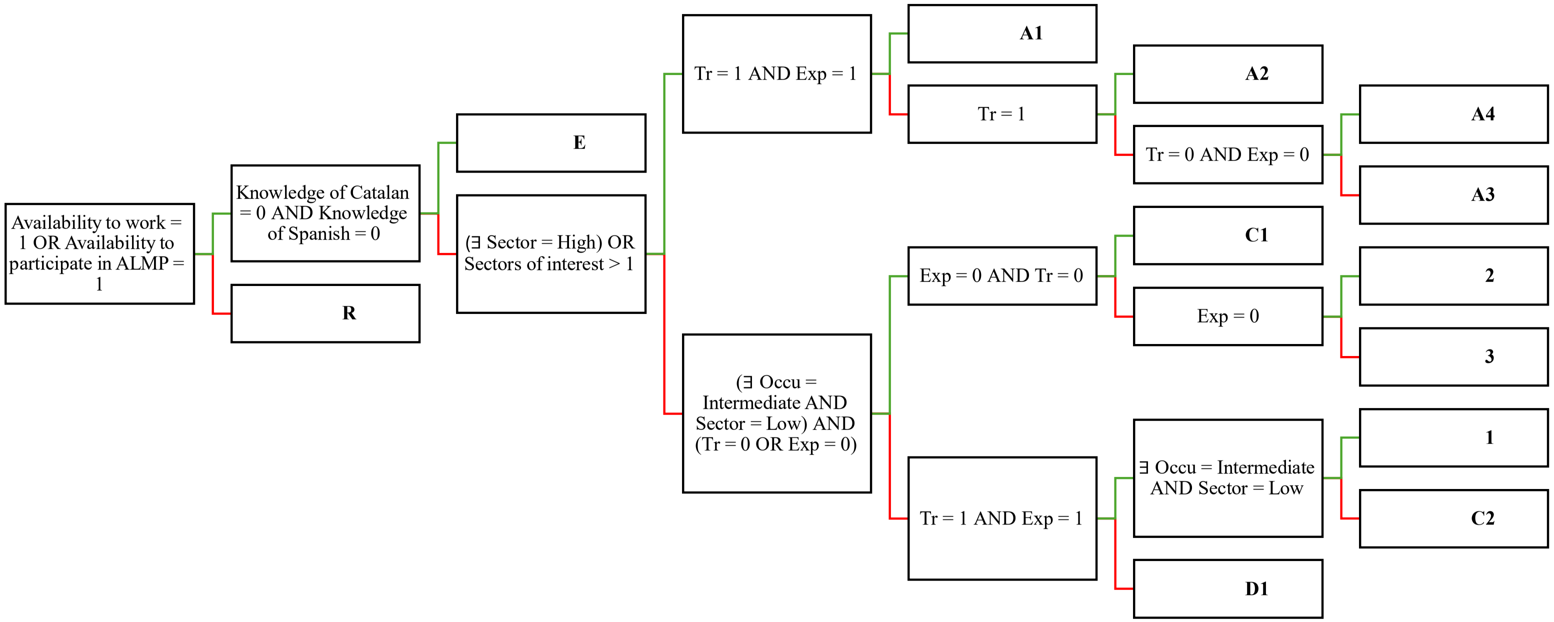
SOC: Pla d'acció per a persones en situació d'atur majors de 45 anys i de llarga durada 2023-2026, https://serveiocupacio.gencat.cat/web/.content/01_SOC/01_Qui-som-i-que-fem/Estrategia-per-locupacio/Pla-de-Desenvolupament-de-Politiques-dOcupacio-de-Catalunya-PDPO/Pla_accio_persones_atur_majors_45_CDSOC_20_07_2023.pdf,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

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

Appendix

Appendix A. Current profiling and decision models of SOC

Figure A1 – Rule-based profiling model Q-G


Availability to work = 1 OR Availability to participate in ALMP = 1
Knowledge of Catalan = 0 AND Knowledge of Spanish = 0
R
E
(∃ Sector = High) OR Sectors of interest > 1
Tr = 1 AND Exp = 1
(∃ Occu = Intermediate AND Sector = Low) AND (Tr = 0 OR Exp = 0)
A1
Tr = 1
Exp = 0 AND Tr = 0
Tr = 1 AND Exp = 1
A2
Tr = 0 AND Exp = 0
C1
Exp = 0
∃ Occu = Intermediate AND Sector = Low
D1
A4
A3
2
3
1
C2

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 that training is enough (1) or not enough (0), denoted as $Tr \coloneqq \{1,0\}$. *Experience* indicates if experience is enough (1) or not enough (0), denoted as $Exp \coloneqq \{1,0\}$. The variable employability of the occupation of interest is represented as $Occu \coloneqq \{High, Intermediate, Low, \emptyset\}$, and was originally denoted $\{"Viable", "Moderado", "En\ retroceso", "No\ definida"\}$. The variable employability of the sector of interest is represented as $Sector \coloneqq \{High, Low\}$, and was originally denoted as $\{"No\ en\ retroceso", "En\ retroceso"\}$.

Table A2 – Rule-based profiling model Q-S

Q-S = $(S_{ito_1}, S_{ito_2}, S_{ito_3})$. $\tilde{S}_{it} = \text{mean}\left(S_{ito_1}, S_{ito_2}, S_{ito_3}\right)$.

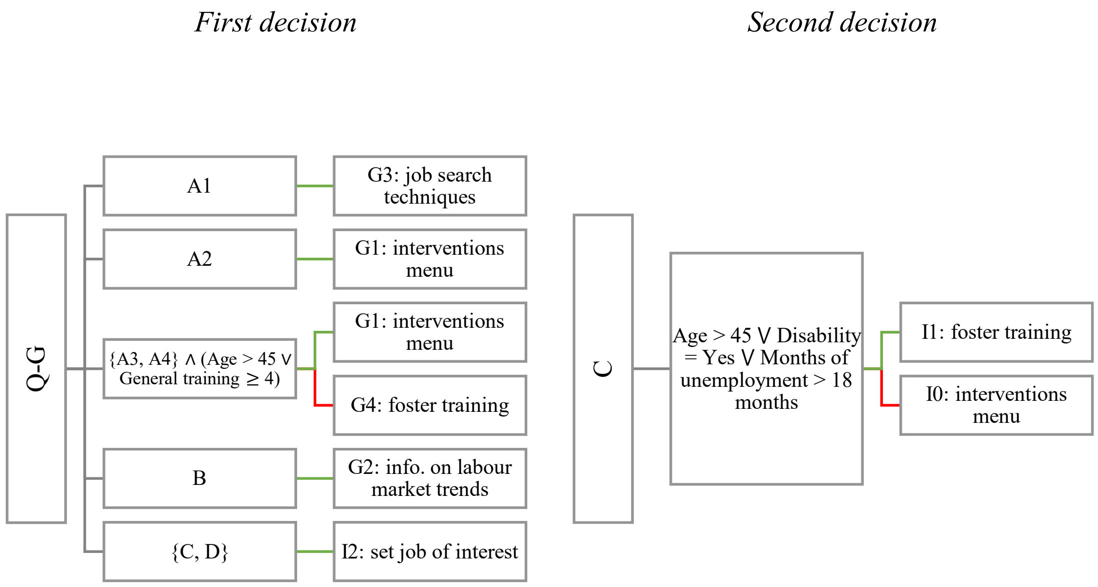
$$\begin{aligned} S_{ito} = {} & \lambda_1 occupation_{ito} + \lambda_2 sector_{ito} + \lambda_3 regexperience_{it} + \lambda_4 irrexperience_{it} \\ & + \lambda_5 regexperience_{ito} + \lambda_6 irrexperience_{ito} + \lambda_7 education_{it} \\ & + \lambda_8 comptrain_{it} + \lambda_9 driving_{it} + \lambda_{10} closed_{ito} + \lambda_{11} search_{it} \\ & + \lambda_{12} languages_{it} + \lambda_{13} digitalskills_{it} + \lambda_{14} willingness_{it} \\ & + \lambda_{15} communication_{it} + \lambda_{16} interpersonal_{it} \end{aligned}$$

Variable (index L)	*Value*	λ_L
$occupation_{ito}$	$x =$ level 1	$\lambda_1 = 0$
$(L = 1)$	$x =$ level 2	$\lambda_1 = 5$
	$x =$ level 3	$\lambda_1 = 6$
	$x =$ level 4	$\lambda_1 = 7$
	$x =$ level 5	$\lambda_1 = 10$
$sector_{ito}$	$x =$ emergent	$\lambda_2 = 10$
$(L = 2)$	$x \in$ {more than one, normal}	$\lambda_2 = 5$
	$x \in$ {declining, missing value}	$\lambda_2 = 0$
$regexperience_{it}$	$x > 60$ months	$\lambda_3 = 10$
$(L = 3)$	36 months $< x \leq 60$ months	$\lambda_3 = 7$
	18 months $< x \leq 36$ months	$\lambda_3 = 5$
	6 months $< x \leq 18$ months	$\lambda_3 = 3$
	0 months $\leq x \leq 6$ months	$\lambda_3 = 0$
$irrexperience_{it}$	$x > 60$ months	$\lambda_4 = 2$
$(L = 4)$	6 months $< x \leq 60$ months	$\lambda_4 = 1$
	0 months $\leq x \leq 6$ months	$\lambda_4 = 0$
$regexperience_{ito}$	$x > 60$ months	$\lambda_5 = 10$
$(L = 5)$	36 months $< x \leq 60$ months	$\lambda_5 = 7$
	18 months $< x \leq 36$ months	$\lambda_5 = 5$
	6 months $< x \leq 18$ months	$\lambda_5 = 3$
	0 months $\leq x \leq 6$ months	$\lambda_5 = 0$
$irrexperience_{ito}$	$x > 60$ months	$\lambda_6 = 2$
$(L = 6)$	6 months $< x \leq 60$ months	$\lambda_6 = 1$
	0 months $\leq x \leq 6$ months	$\lambda_6 = 0$
$education_{it}$	$x \in$ {0, 1nr, missing value, 25Bnr}	$\lambda_7 = 0$
$(L = 7)$	$x \in$ {1, 25B, 24nr, 35Bnr}	$\lambda_7 = 2$
	$x \in \{24, 35\text{B}, 44\text{nr}, 45\text{nr}, 55\text{Bnr}, 54\text{nr}\}$	$\lambda_7 = 4$
	$x =$ {44, 45, 54, 55B, 55nr, 75Bnr}	$\lambda_7 = 6$
	$x \in$ {6nr, 7nr, 8nr, 55, 75B}	$\lambda_7 = 8$
	$x \in$ {6, 74, 75, 8}	$\lambda_7 = 10$
$comptrain_{it}$	$x \geq 80$	$\lambda_8 = 10$
$(L = 8)$	$x < 80$	$\lambda_8 = 0$
$driving_{it}$	$x =$ B1	$\lambda_9 = 10$
$(L = 9)$	$x \neq$ B1	$\lambda_9 = 0$

Variable (index L)	*Value*	λ_L
$closed_{ito}$ $(L = 10)$	$x =$ (Yes ∧ (Has it ∨ Studying it))	$\lambda_{10} = 5$
	$x \in$ {Recommended ∧ (Has it ∨ Studying it), No ∧ (Has ISCED 2 ∨ Studying ISCED 2)}	$\lambda_{10} = 4$
	$x \in$ {Recommended ∧ Doesn't have it, No ∧ Has not ISCED 2}	$\lambda_{10} = 2$
	$x \in$ {Yes ∧ (Doesn't have it ∨ Unknown), (Recommended ∨ No) ∧ 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} ∧ $Sp \in$ {High, Middle}	$\lambda_{12} = 10$
	$Cat \in$ {High, Middle} ∧ $Sp \in$ {Basic, Null}	$\lambda_{12} = 5$
	$Cat \in$ {Basic, Null} ∧ $Sp \in$ {High, Middle}	$\lambda_{12} = 4$
	$Cat \in$ {Basic, Null} ∧ $Sp \in$ {Basic, Null}	$\lambda_{12} = 0$
$digitalskills_{it}$ $(L = 13)$	$x = 0$	$\lambda_{13} = 10$
	$x = 1$	$\lambda_{13} = 5$
	$x \in \{2, 3\}$	$\lambda_{13} = 0$
$willingness_{it}$ $(L = 14)$	$x = 0$	$\lambda_{14} = 10$
	$x \geq 1$	$\lambda_{14} = 0$
$communication_{it}$ $(L = 15)$	$x =$ 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 the framework of labour market policies. "Cat" denotes the Catalan language and "Sp" the Spanish language. ∨ is the Boolean operator for OR, ∧ is the Boolean operator for AND. Appendix D includes a description of each variable.

Figure A3 – 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 if such intervention is individual (I) or group-based (G) and the main objective of the action. The second decision is only determined by the rule-based profiling if 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 labour 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 remote.

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 $d \in \{\mathrm{I0}, \mathrm{I1}, \mathrm{I2}\} > d' \in \{\mathrm{G1}, \mathrm{G2}, \mathrm{G3}, \mathrm{G4}\}$.

We have maintained the original abbreviations of the group interventions, but we have changed the abbreviations of the individual interventions to avoid confusions. 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).

Appendix B. Review on performance of jobseeker profiling models

Country	Model	Outcome	ROC-AUC	Sensitivity	Precision	Accuracy	O:E	Source
Austria	Statistical	Labour 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 episode (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)

Note: This table only includes those 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, here we 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.

References

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)

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

Appendix C. Classification of intermediation claims

Table C1 – Correspondence between causes of intermediation claims and type of episode

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
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 ALMP	U
710	Registration for ALMP prior to employment	U
711	Registration due to benefit resumption-compatibility	U

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

Appendix D. List of predictors of Q, C and K

Table D1 – 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 hours (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 D2 – Predictors used in the C function.

Group	*Predictor*
Used in formal allocation	Age
	He/she has a disability
	Duration of the unemployment episode
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 D3 – Predictors used in our statistical models (Κ).

Group	*Predictor*
PLMP	Started unemployment during a benefit interval
	Number of benefit episodes (completed) in the past
	Total duration of previous benefit episodes
	Total duration of previous benefit episodes, scaled by age
	Mean duration of previous benefit episodes
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 episodes in the past (inside the window)
	Total duration of unemployment episodes in the past (until the present episode, not included)
	Total duration of unemployment episodes in the past (until the present episode, not included), scaled by age
	Mean duration of unemployment episodes until the present (until the present episode, not included)
	Days since last unemployment episode
Inactivity	Total duration of inactivity episodes
	Mean duration of inactivity episodes until the present
	Number of inactivity episodes in the past
Employment	Days since first employment (in the window)
	Days since (the beginning of) the last employment episode
	Days since (the beginning of) the last full-time employment episode
	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 episodes without any vocational training held in the past
	Number of occupations held in the past
	Number of employment episodes 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 episodes (11 categories)

	Maximum skill level required for past employment episodes (5 categories)
	Minimum skill level required for past employment episodes (11 categories)
Socio-demographics	Sex
	Age in years when the episode started
	Maximum level of education
	Has a disability
	Local labour 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 episodes in the past
	Indicator of missingness on unemployment episodes in the past
	Indicator of missingness on local labour 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 Labour Market Policies.

Appendix E. Summary statistics

Table E1 – 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 labour market, the field of study and the level of study are not shown to simplify the exposition. The tables are available upon request.

Table E2 – Summary statistics on sociodemographic quantitative variables.

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

Appendix F. Tuning parameters

Table F1 – 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.** , 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 F2 – 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

Table G1 – Kappa statistic of models in the restrictive test subset

	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 ($\kappa_{\mathrm{QS50}} = 0.067$) and represents less than one third of the chance-corrected accuracy we could get with the random forest.

Table G2 – False alarm rates in the restricted test subset.

	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

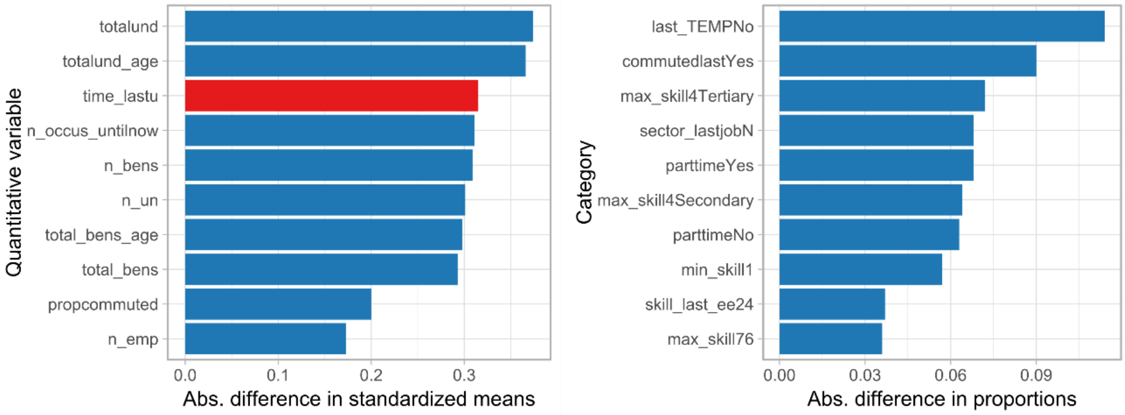
Note: FAR = 1 – specificity.

Table G3 – False alarm rates in two strata of the test subset.

	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.

Figure G1 – Top-10 differences in standardized means (left) or proportions (right) between the test subset and the restrictive test subset.



Source: Own elaboration. Denoting with μ_g a summary statistic for the dataset g, it is shown the difference $\mu_{test} - \mu_{restrictive}$. Therefore, blue bars denote a positive difference, whereas red bars collect a negative difference.

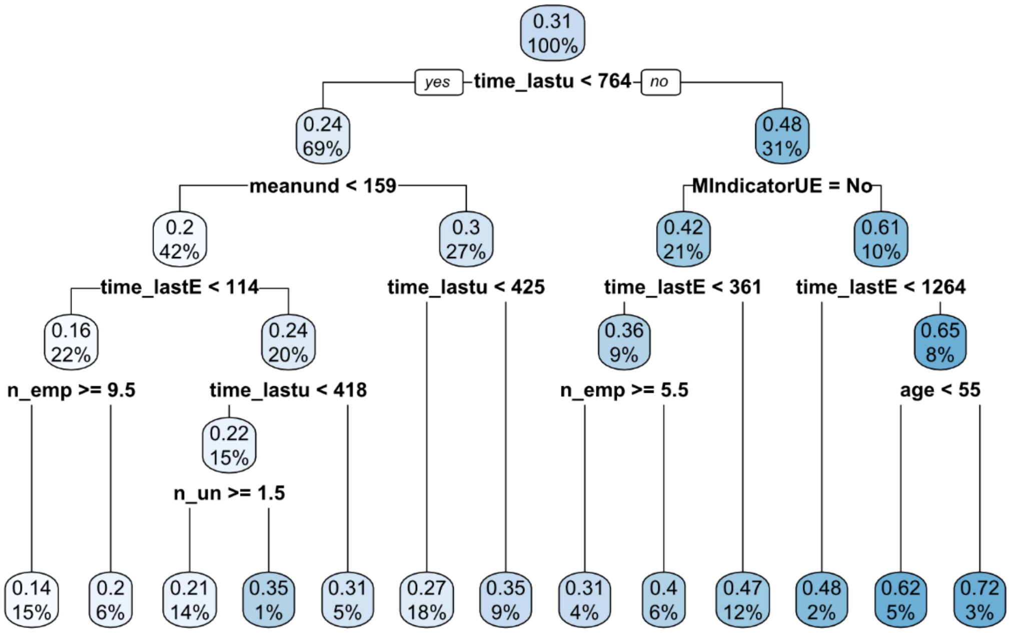
Appendix H. Global surrogate model

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 equalled 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 episode (time_lastu), the number of days since the beginning of the last employment episode (time_lastE), the mean duration of unemployment episodes until the present (meanund), the number of employment episodes in the past (n_emp), the number of unemployment episodes in the past (n_un), the age (age), and the indicator of missingness on unemployment episodes 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 experience a long-term unemployment episode and below the percentage of the sample that fits in each partition. Starting the partition from above, we see that the combinations of value that predict LTU with probability equal to 0.72 is: having the last unemployment episodes at least 764 days ago, having the last employment episode 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.

Figure H1 – Graphical representation of the decision tree


0.31
100%
yes
time_lastu < 764
no
0.24
69%
0.48
31%
meanund < 159
MIndicatorUE = No
0.2
42%
0.3
27%
0.42
21%
0.61
10%
time_lastE < 114
time_lastu < 425
time_lastE < 361
time_lastE < 1264
0.16
22%
0.24
20%
0.36
9%
0.65
8%
n_emp >= 9.5
time_lastu < 418
n_emp >= 5.5
age < 55
0.22
15%
n_un >= 1.5
0.14
15%
0.2
6%
0.21
14%
0.35
1%
0.31
5%
0.27
18%
0.35
9%
0.31
4%
0.4
6%
0.47
12%
0.48
2%
0.62
5%
0.72
3%