Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Preparing automated decision-making in public employment services

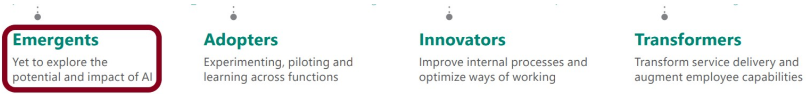
Open projects and challenges (Switzerland)

February 23, 2023

Martin Gasser
State Secretariat for Economic Affairs SECO
Swiss Unemployment Insurance
martin.gasser@seco.admin.ch

(SEMI-)AUTOMATED DECISION-MAKING (ADM)

- ADM = fully-automated or semi-automated ("human-in-the-loop") decision-making
- PES = Public Employment Services
- Currently, Swiss PES are not using any ADM
- New data protection law allows the use of ADM, if those affected recognize the decision as such and have recourse
- We have to prepare for potential ADM uses



How 213 Public Organizations Benefit from AI

POTENTIAL USES OF ADM IN PES

Figure 3. AI has the potential to improve ALMP provision across PES activities

"Traditional" **administrative data** in PES and linked registers

+

Big Data (click data in PES online tools, vacancy search behaviour, free text in vacancy posts and CVs, etc.)

+

AI algorithms

AI tools in PES

- **Profiling** to segment jobseekers, identify employment barriers, predict their labour market outcomes, target ALMPs, suggest labour market integration pathways.
- **Identifying vacancies** (companies with a high potential for recruitment) proactively.
- **Matching** jobseekers and vacancies (identifying matches using semantic matching, selecting best fit).
- **Mapping jobseekers' distance to occupations** and gaps in competencies, analysing expected skills by employers and career choices of workers for recommender systems in career services.
- **Providing information and counselling** via chatbots.
- **Detecting fraud** and **assuring quality** in processing applications.

Note: AI – artificial intelligence, ALMP – active labour market policy, PES – public (and private) employment services.

OECD (2022)

GUIDELINES UNDER DISCUSSION

1. **Technology and risk assessement**: required pre-development with relevant stakeholders, users and developers
2. **Privacy impact assessement**: legally required previous to any development
3. **Data quality**: Data are contextualized together with stakeholders and PES (e.g. data quality, expressiveness, and proxy outcomes)
4. **Sufficient precision**: necessary accuracy/performance is defined with stakeholders and independently evaluated (e.g. on test data)
5. **Non-discrimination**: statistical measure(s) of discrimination are defined with stakeholders and regularly evaluated
6. **Transparency and reproducibility**: automated decisions are recognizable as such, researchers can study the model (no black box)
7. **Interpretability and explainability**: model class as a whole should be interpretable, individual decisions can be reliably explained

Based on existing guidelines from the Swiss government and the Swiss Competence network for data science.

OVERVIEW GUIDELINES

Prior to development
- Technology and risk assessement
- Privacy impact assessment
- Data quality assessement

During development
- Restriction to interpretable models
- Restriction to explainable output
- Incorporate domain knowledge

After development
- Test-set based assessement of non-discrimination and accuracy
- Transparency and reproducibility

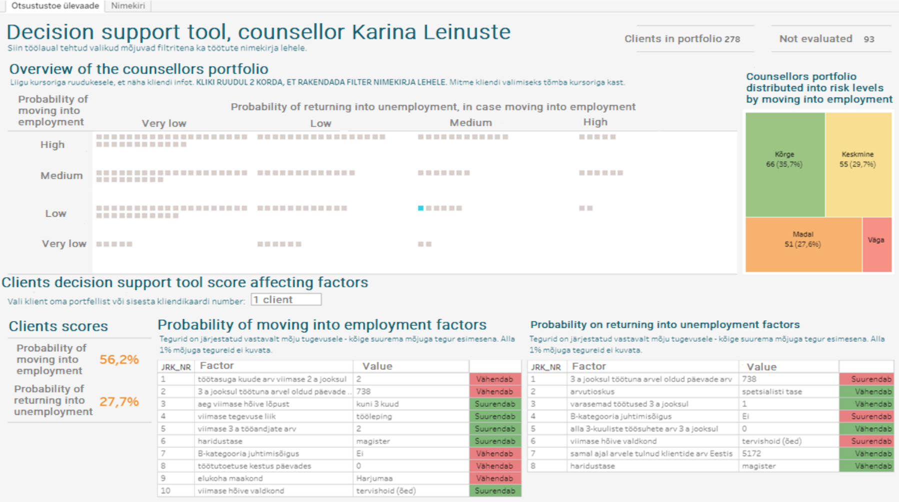
- Non-discrimination and legitimacy
- Training model users, ensuring sufficient resources for override and opt-out

- Note: PES is Switzerland are organized regionally
  - regional authorities have large room for maneuvre
  - any ADM will be used differently according to region
  - meaning, language and quality of data vary by region

CHALLENGES

- There are templates for technology and risk assessments, transparency rules, and privacy impact assessments; as well as established measures of accuracy
- Explainability is a practical issue (you know it when you use it)
- However, non-discrimination and interpretability are active and contentious areas of research
- Moreover, these areas of research are often highly technical. But in practice, we would have to discuss these matters with non-technical stakeholders
- Technical and ethical trade-offs have to be resolved beforehands because any ADM will fail on some criteria

EXPLAINABILITY



Otsustustoe ülevaade | Nimekiri

Decision support tool, counsellor Karina Leinuste

Siin töölaual tehtud valikud mõjuvad filtritena ka töötute nimekirja lehele.

Clients in portfolio 278

Not evaluated 93

Overview of the counsellors portfolio

Liigu kursoriga ruudukesele, et näha kliendi infot. KLIKI RUUDUL 2 KORDA, ET RAKENDADA FILTER NIMEKIRJA LEHELE. Mitme kliendi valimiseks tõmba kursoriga kast.

Counsellors portfolio distributed into risk levels by moving into employment

Clients decision support tool score affecting factors

Vali klient oma portfellist või sisesta kliendikaardi number: 1 client

Clients scores

Probability of moving into employment	56,2%
Probability of returning into unemployment	27,7%

Probability of moving into employment factors

Tegurid on järjestatud vastavalt mõju tugevusele - kõige suurema mõjuga tegur esimesena. Alla 1% mõjuga tegureid ei kuvata.

JRK_NR	Factor	Value	
1	töötasuga kuude arv viimase 2 a jooksul	2	Vähendab
2	3 a jooksul töötuna arvel oldud päevade ..	738	Vähendab
3	aeg viimase hõive lõpust	kuni 3 kuud	Suurendab
4	viimase tegevuse liik	tööleping	Suurendab
5	viimase 3 a tööandjate arv	2	Suurendab
6	haridustase	magister	Suurendab
7	B-kategooria juhtimisõigus	Ei	Vähendab
8	töötutoetuse kestus päevades	0	Vähendab
9	elukoha maakond	Harjumaa	Vähendab
10	viimase hõive valdkond	tervishoid (õed)	Suurendab

Probability on returning into unemployment factors

Tegurid on järjestatud vastavalt mõju tugevusele - kõige suurema mõjuga tegur esimesena. Alla 1% mõjuga tegureid ei kuvata.

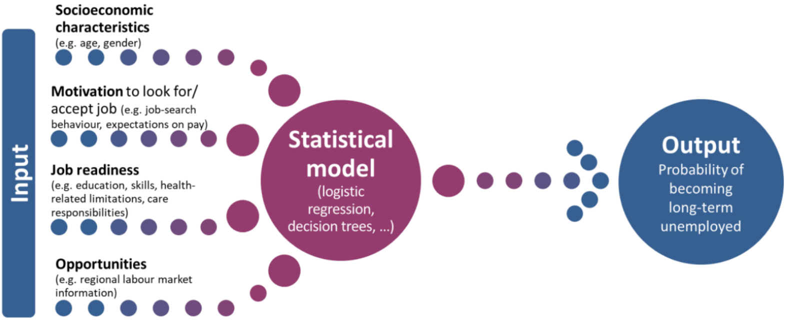
JRK_NR	Factor	Value	
1	3 a jooksul töötuna arvel oldud päevade arv	738	Suurendab
2	arvutioskus	spetsialisti tase	Vähendab
3	varasemad töötused 3 a jooksul	1	Vähendab
4	B-kategooria juhtimisõigus	Ei	Suurendab
5	alla 3-kuuliste töösuhete arv 3 a jooksul	0	Vähendab
6	viimase hõive valdkond	tervishoid (õed)	Suurendab
7	samal ajal arvele tulnud klientide arv Eestis	5172	Vähendab
8	haridustase	magister	Vähendab

Presentation Estonia (OECD 2021)

USE CASES

- **Matching.** Implement a match-making engine on our job platform
    - There seem to be ready-made software solutions already used in e.g. the *WCC Employment Platform* used in Belgium, Germany, Austria
    - Might test such a platform for skill-based matching
    - In case of explicit, rule-based matching, only moderate requirements necessary
- **Profiling (risk assessment).** e.g. predicting long-term unemployment based on labour market and individual data
    - Non-discrimination and explainability are more important for profiling/targeting than for recommender tools

USE CASE: PROFILING



Desiere, S., K. Langenbucher and L. Struyven (2019), "Statistical profiling in public employment services: An international comparison", OECD Social, Employment and Migration Working Papers, No. 224.

USE CASE: NON-DISCRIMINATION IN RISK PROFILING

- Three standard observational definitions of group fairness, which are are mutually incompatible[1]
- Auditing can be based on a hold-out test set. But stakeholders would have to first decide on
    1. a (smallish) set of protected attributes and their mode of interaction (intersectionality)
    2. an appropriate definition of non-discrimination
    3. a measure of discrimination
    4. an " acceptable" threshold for discrimination
- Statistically, there are well established procedures to measure discrimination with risk classes. When dealing with risk scores, there remain many open questions

[1] For a good introduction: https://fairmlbook.org/. Other definitions include individual and causal fairness.

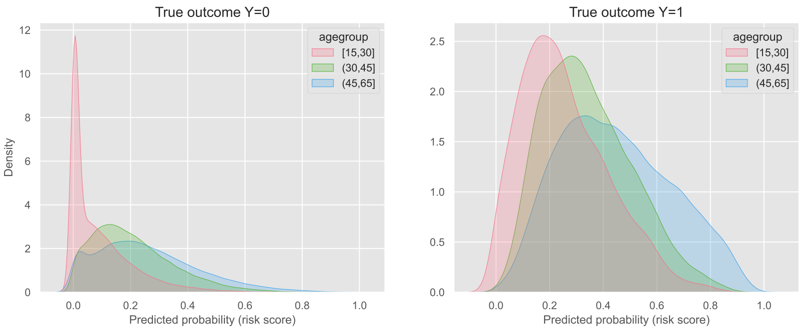
USE CASE: NON-DISCRIMINATION IN RISK PROFILING

As a dry run, we trained an XGBoost model on a full data set (years 2014-2018) with 78 predictors and kept 2019 as test set. Accuracy was 0.78 (AUC).

- Assume stakeholders choose age as a protected attribute. The model was trained without access to age
- Assume stakeholders choose separation as a criterion: All age groups should have equal error rates any decision thresholds
- Assume stakeholders choose expected risk difference as a measure and are willing to accept a value ≤ 0.1.

Then, the proposed model would fail the non-discrimination audit.

USE CASE: NON-DISCRIMINATION IN RISK PROFILING



Expected risk differences of younger and older jobseekers relative to the middle-aged group: $0.116, 0.005, 0.086, 0.104$.

OPEN QUESTIONS

- Do stakeholders understand/accept technical definitions of non-discrimination that rely on statistical independence?
- How do we navigate conflicting definitions of discrimination in practice? We lack real-world best practice cases
- How do we deal with multiple protected attributes, each with an appropriate definition of fairness? There is little research
- Should we test for full non-discrimination or measure discrimination. There is surprisingly little research on measuring discrimination in an interpretable way
- Can we really expect a model to be fully fair and, if not, how would we determine "acceptable levels" for a measure?

A CAVEAT

- Even if the ADM output were non-discriminatory and explainable, it does not follow that it is fair or that it is *legitimate* to use the ADM at all[2]
- A major challenge in all ADM remains to make it useful to and accepted by practitioners and those affected
    - Two early attempts (2005 and 2015) at targeting/profiling failed due to being rejected by users (PES caseworkers)

[2] cf. fairmlbook.org/legitimacy

The three "standard" definitions of observational group fairness:

Name	General $\hat{Y}$	Special case $\hat{Y} \in \{0,1\}$
Independence	$A \perp\!\!\!\perp \hat{Y}$	**Demographic parity** $P(\hat{Y}{=}1 \mid A{=}a) = P(\hat{Y}{=}1 \mid A{=}b)$ for all a, b
Separation	$A \perp\!\!\!\perp \hat{Y} \mid Y$	**Error rate parity** $P(\hat{Y}{=}y \mid Y{=}1{-}y, A{=}a) = P(\hat{Y}{=}y \mid Y{=}1{-}y, A{=}b)$ for all $y \in \{0,1\}$ and a, b
Sufficiency	$A \perp\!\!\!\perp Y \mid \hat{Y}$	**Predictive parity** $P(Y{=}y \mid \hat{Y}{=}y, A{=}a) = P(Y{=}y \mid \hat{Y}{=}y, A{=}b)$ for all $y \in \{0,1\}$ and a, b

Legend: A: protected attribute, Y: observed outcome, $\hat{Y}$: predictions

Relative risk estimates in case of risk groups: