INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

PREDICTING INDIVIDUAL JOB MATCH QUALITY

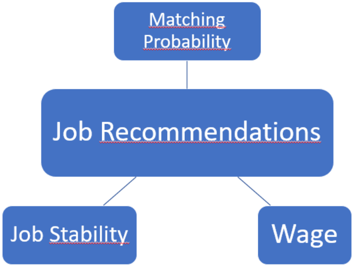
EEA, 28.08.2023

Sabrina Mühlbauer
Enzo Weber

MOTIVATION

- Definition Matching: a job seeker enters into employment by being matched to a certain job category (i.e. occupational field)
- improve matching process in German Employment Agencies (support caseworkers)
- provide a list of job recommendations individually for each job seeker
- find out differences between the application of traditional estimation methods and machine learning algorithms

IN BRIEF



→ provide index for job match quality for each person

→ additional alternatives could support caseworkers and improve the matching (Belot et al. (2019), Blundell et al. (2004))

JOB RECOMMENDATIONS

- **Matching Probability**: predict probability that a person (with specific characteristics) gets employed in a certain job category
  - field experiment shows that having more alternatives and information has a positive effect on labor market success (Altmann et al. (2018))
- Job Match Quality
  - **Job Stability**: probability for being long term employed after starting a new job
  - **Wage**: expected wage if starting a job in a certain occupation
  - job stability and wages are common measures for job match quality (i.e. Caliendo et al. (2013), van den Berg and Vikström (2014) or Nekoei and Weber (2017))

METHODS - OVERVIEW

Common Methods

- Logit
- Multinomial Logit
- OLS

Machine Learning

- Random Forest
- Xgboost
- K-nearest Neighbors
- Support Vector Machines
- Neural Networks

Requirements

- Prediction
- Classification
- Regression

METHODS - SELECTION FOR THE PRESENT PROJECT

- OLS, logit, multinomial logit
  - manual model selection (time-consuming)
  - large classification models can not be estimated with multinomial logit
  - models do not improve much by having more data
- random forest, xgboost
  - does not need (much) hyperparameter tuning
  - fast in computing large and complex models
  - high prediction performance
- neural networks, support vector machines
  - for complex models: hyperparameter tuning is extremely time consuming
  - model gets extremely large: estimation collapses
  - splitting the sample: error rate increases dramatically

METHODS - SELECTION FOR THE PRESENT PROJECT II

- k-Nearest Neighbors
  - the more observations the higher the optimal k
  - the higher k the higher the computing time and the required RAM
- machine learning in labor market research
  - analysis of vacancies by text classification (Amato et al. (2015))
  - matching vacancies to candidates (Bhatia et al. (2020), van Belle et al.(2018), Fang (2015))

$\rightarrow$ **prefer random forest and xgboost**

DATA - INTEGRATED EMPLOYMENT BIOGRAPHIES

Integrated Employment Biographies (IEB)

- contains information from 1975 onwards
- covers all employment biographies in Germany
- administrative, high frequency dataset
- sources: Jobseeker Histories and Employee History
- estimations:
  - use random 10 %-sample
  - observations from 2012 onwards

EXPLANATORY VARIABLES

- gender (male, female)
- federal state (a person lives in)
- nationality (German, EU, Europe without EU, 8 migration countries, remaining nations)
- marital status (single, partnership)
- children (at minimum one child under 15 years, no children)
- education (no school leaving certificate, . . . , university)
- job category of completed vocational training
- job category someone was employed in before starting a new job
- skill level
- age at the start of employment
- number of days in unemployment before starting a new employment

MATCHING PROBABILITY: THE MODEL

Definition

$$P(M_i = j) = f(X_i, Y_i),$$

- $j = 1, \ldots, J$, J is the number of different job categories,
- $i = 1, \ldots, N$, N is the number of observations
- M_i denotes the occupation of observation i
- X is a vector denoting the characteristics of observation i
- Y is a vector denoting the characteristics of jobs of observation i

MATCHING PROBABILITY

Endogenous Variable: **Job Category**

- 144 different occupational groups: 3-digit defined in the German classification of occupations 2010
- consider jobs subject to social security
- observation period: 2012-2018

Sample

- stock of persons having a job subject to social security
- 54,781,854 observations of 2,883,188 different persons
- unbalanced distribution of persons across job categories

MATCHING PROBABILITY: ESTIMATION

- test-train split by year
  - train set: 2012-2017
  - test set: 2018
- best method: random forest
- measure of goodness: classification error rates (= number of wrong predictions/ total number of observations)
- out-of-sample error: 42.20 %
- random forest error is by 18.50 percentage points lower than for OLS
- important variables
  - calculate Gini-based importance
  - most important predictors: last job category someone was employed in, skill level required for previous employment

JOB STABILITY: THE MODEL

Definition:

$$\text{P(duration} > 6 \text{ months)} = f(\mathbf{X}, \mathbf{Y}),$$

- $\mathbf{X}$ is a matrix covering i characteristics of every person
- $\mathbf{Y}$ is a matrix covering j characteristics of jobs of every person
- n is the number of observations (i.e. spells)

JOB STABILITY

Endogenous Variable: **employment duration**

- two categories: short term (< 6 months) and long term employment (> 6 months)
- test-train split by year
  - train set: 2012-2016
  - test set: 2017
- best method: xgboost
- measure of goodness: classification error

JOB STABILITY: ESTIMATION

Two samples

- **occupation duration sample**: being employed in the same occupation (i.e. job category)
  - classification error rate for xgboost: 13.68 %
  - by 21.6 % lower than logit
- **employment duration sample**: being employed without interruption by an unemployment period
  - classification error rate for xgboost: 15.25 %
  - by 36.7 % lower than logit

WAGES

Definition: The Model

$$\ln(\text{wage}) = f(\mathbf{X}, \mathbf{Y}).$$

endogenous variable: **daily wages**

- daily wages are available for full-time employed persons
- imputation of daily wages above the contribution limit

WAGES: ESTIMATION

- best method: xgboost
- measure of goodness: mean squared error
- test-train split
  - train set: 2017
  - test set: 2018

Descriptives

- train set: 1,050,210 observations of 856,636 different persons
- test set: 1,092,315 observations of 869,990 different persons

Results

- MSE log daily wage: 0.0558
- xgboost MSE is by 69% smaller than OLS MSE

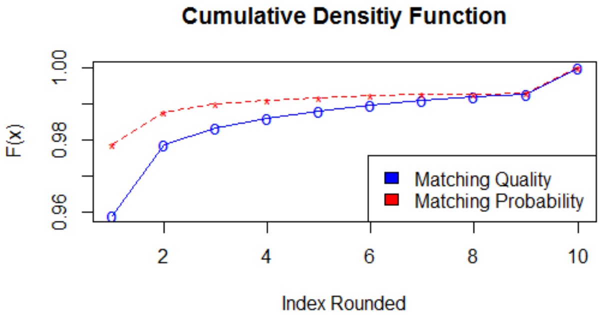
INDEX: JOB MATCH QUALITY

Definition:

$$Q_{rs} = P(\text{duration}_r > 6 \text{ months}|s) * E[\text{wage}_r|s] * P(\text{M}_r = s),$$

- $r = 1, \ldots, N$, N is the number of observations
- $s = 1, \ldots S$, S is the number of job categories
- scale index to a range from 1 to 10
- list of job recommendations for each individual

DIFFERENCE: MATCHING PROBABILITIES VS. INDEX



$\rightarrow$ additional information on job match quality (job stability and wages) leads to a difference in job recommendations

CONCLUSION

- machine learning (ML) can play an important role in labour market matching
- ML should be preferred over common methods in any case
- tree-based methods (random forest and xgboost) work best
- ML results get better, the larger the training data while common methods not: ML finds additional patterns
- Outlook:
  - add information on skills and competencies
  - start a field experiment in German Employment Agencies

THANK YOU FOR YOUR ATTENTION!

Sabrina Mühlbauer

sabrina.muehlbauer@iab.de