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

IAB-DISCUSSION PAPER

Articles on labour market issues

03|2022 Machine Learning for Labour Market Matching

Sabrina Mühlbauer, Enzo Weber

Aktualisierte Fassung vom 08.03.2022

ISSN 2195-2663

Machine Learning for Labour Market Matching

Sabrina Mühlbauer (IAB),
Enzo Weber (IAB)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

List of Figures

List of Tables

Abstract

This paper develops a large-scale application to improve the labour market matching process with model- and algorithm-based statistical methods. We use comprehensive administrative data on employment biographies covering individual and job-related information of workers in Germany. We estimate the probability that a job seeker gets employed in a certain occupational field. For this purpose, we make predictions with common statistical methods (OLS, Logit) and machine learning (ML) methods. The findings suggest that ML performs best regarding the out-of-sample classification error. In terms of the unemployment rate hypothetically, the advantage of ML compared to the common statistical methods would stand for a difference of 0.3 - 1.0 percentage points.

Zusammenfassung

Dieses Papier beschäftigt sich mit einer groSS angelegten Datenanalyse mit dem Ziel den Matchingprozess auf dem Arbeitsmarkt mit algorithmen- und modellbasierten statistischen Methoden zu verbessern. Wir nutzen einen umfassenden administrativen Datensatz zu Arbeitsmarktbiographien von erwerbstätigen Personen in Deutschland. Der Datensatz beinhaltet sowohl personenbezogene als auch berufsbezogene Informationen. Auf Basis unserer Schätzungen berechnen wir, in welchem Berufsfeld eine arbeitslose Person mit der gröSSten Wahrscheinlichkeit Beschäftigung findet. Wir nutzen für unsere Prognosen konventionelle statistische Methoden (OLS, Logit) und maschinelles Lernen. Anhand unserer Ergebnisse zeigt sich, dass für die zugrundeliegende Fragestellung Algorithmen des maschinellen Lernens die besten Ergebnisse liefern. Als GütemaSS hierfür nutzen wir den out-of-sample Prognosefehler. Hinsichtlich der Arbeitslosenquote würde der Vorteil der Methoden des maschinellen Lernens hypothetisch für einen Unterschied von 0,3 - 1,0 Prozentpunkten im Vergleich zu den konventionellen statistischen Methoden stehen.

JEL

C14, C45, C55, J64

Keywords

Labour Market, Machine Learning, Matching, Random Forest

Acknowledgements

We are grateful to Anja Bauer, Bernd Fitzenberger, Hermann Gartner, Max Kunaschk, Johannes Ludsteck, Martin Spindler and Michael Stops as well as the participants of the Machine Learning Workshop of the Friedrich-Alexander University Erlangen-Nürnberg (FAU), the Statistical Week 2021, the Young Researchers Seminar (IAB and FAU) and the Working Group Artificial Intelligence at the IAB for helpful comments and support. Special thanks go to the DIM department of the Institute for Employment Research (IAB) for their effort to provide the data.

1 Introduction

Matching is the key task of labour markets. In Germany, at the moment, the placement officer makes job suggestions based on the information provided to him. In general, matching refers to a job seeker entering into employment by being matched with a concrete job. For example, Mortensen/Pissarides (1994) or Petrongolo/Pissarides (2001) describe approaches that rely on matching theory. They consider relationships between unemployment and vacancies as inputs for a production function of job findings. The matching function shows that, in general, the more job seekers and vacancies are available the higher the number of matches. Additionally, hirings are affected by matching efficiency, i.e. the productivity parameter of the matching function (matching job seekers to available jobs). Turnover models show that a higher matching efficiency leads to a decline in unemployment. Matching efficiency can be improved by optimizing the individual matching probability. For example, Hall/Schulhofer-Wohl (2018) investigate job-finding rates and matching efficiency. Our main goal is to improve the individual matching by using model- or algorithm-based statistical methods.

Our results are calculated with data on employment biographies. This historical data informs about job seekers and their previous occupations. We focus on finding a pattern and thus, an algorithm relates personal and job characteristics of being employed in certain jobs. Ultimately, the algorithm based on representative labour market data can be used to make suggestions to job seekers. With this information (i.e. employment biographies) recommendations for job seekers can be based on occupations in which comparable other job seekers successfully found jobs. We estimate statistical models with data that contains information about the characteristics of an individual and the characteristics of the corresponding job. As machine learning (ML) algorithms have promising abilities for pattern recognition and making predictions we apply different methods. We use Logit, OLS, Random Forest (RF) and k-Nearest-Neighbours (kNN). For every occupation, we predict the probability that a certain job seeker gets employed. A list with any number of job recommendations can be produced for every person individually.

The required data is collected by the German Employment Agency in the Integrated Labour Market Biographies (IEB). This administrative, high frequency and extensive data covers well the hard facts about the characteristics of persons and jobs. For estimation purposes, we use two samples. On the one hand side, we take employment spells subject to social security and on the other side we filter the transitions of unemployment into employment. The focus is on the comparison of the out-of-sample error rates estimated with microeconometric and machine learning methods. We find that using RF instead of OLS can reduce the error rate by 30 percent. For illustrative purposes, we compare the first best job suggestions of RF and OLS. Using RF stands for a difference of 2.9 - 3.6 percentage points in terms of the unemployment rate compared to OLS.

Our study is related to several strands of literature. One general problem is mismatch unem-

ployment (compare Sahin et al. (2014)), which occurs if the qualifications of workers differ from those required for the job. More detailed information about the situation in Germany is described in Bauer (2013). Further, Hutter/Weber (2017) investigate the impact of mismatch on the matching efficiency. In this regard, we provide a method that can be used to generate recommendations for jobs that might be some alternatives for job seekers. Studies that investigate job suggestions are Belot/Kircher/Muller (2019) or Blundell et al. (2004). They show that having more alternatives leads to a "broader" search strategy. Along with increasing the number of job interviews, this can reduce mismatch. In a field experiment, Altmann et al. (2018) also show that having more information is positively linked to labour market success. Big data (see for example Gerunov (2014) and machine learning methods are often used for labour market research. Colombo/Mercorio/Mezzanzianica (2018) use these techniques for analysing which skills are required for different occupations. Another application where ML is used for the investigation of web job vacancies is given by Boselli et al. (2017). Matching vacancies to candidates with machine learning methods is investigated in van Belle/Dehling/Foster (2018), Fang (2015) or Braun (2017). They try to select people who will fit a certain vacancy perfectly by using ML. The algorithm learns from the company's historical placement data to order a list of candidates on a given job offer. The techniques for the ranking procedure are described in Li (2011) and Cao et al. (2007). Another application for ML in an online recruitment system is described in Faliagka et al. (2012). Further, Van Huynh et al. (2010) also show how ML techniques, in this case, neural networks, are used for determining suitable jobs. They investigate the task of IT job prediction (i.e. a classification task) for predicting a job based on job descriptions including for example job requirements, skills or interests.

We contribute by optimise matching probability based on a large administrative dataset. For that purpose, we use microeconometric and ML methods. We demonstrate how this can be put into practice relying on comprehensive data covering the full universe of employment spells in a big labour market. For every person the tool produces a ranked list of recommendations. To reach this, we make predictions based on historical administrative data of unemployed with information about the final job. Finally, obtaining out-of-sample error rates is possible. These error rates are an indicator for the quality of the predictions on unknown data (i.e. a job seeker that should be reintegrated into the labour market). Our motivation is finding an algorithm that can be applied in practice for supporting the placement process. Besides, having additional job suggestions might be an incentive for job seekers to expand their job search strategy.

This paper is structured as follows. In the second section, we describe the general setting and the dataset. The estimation strategy and the estimation methods are described in the third section. In section four, we present and interpret the empirical results. Section five shows some robustness checks and section six concludes the paper.

2 Setting and Data Structure

In Germany, a placement officer recommends a certain job based on his experience and personal conversations. Statistical methods could support this process by giving additional recommendations to job seekers or people who search a new job.

2.1 Defining Matching

The main idea is matching persons with certain characteristics to a certain occupational group. The model is given by

$$P(M_i = j) = f(\mathbf{X}_i, \mathbf{Y}_i), \tag{2.1}$$

with $j = 1, \ldots, J$ where J is the number of different job categories, $i = 1, \ldots, N$, N is the number of observations, **M** is a Nx1-vector containing the occupational group someone is employed in, so M_i denotes the occupation of observation i. Further, $\mathbf{X}_i$ is a 1xS-vector, where S is the number of variables denoting the characteristics of person i. $\mathbf{Y}_i$ is a 1xS-vector, where R is the number of variables denoting the characteristics of jobs of person i. Thus, we estimate the probability that a person gets employed in a certain occupation and assume that the process is a function of personal and job characteristics, both included in the dataset.
A match is considered successful when a person who was unemployed in the previous period (i.e. an unemployment spell is recorded in the previous period), starts a job. We define a match as the occupation in which a person starts an employment. The main idea is providing a list of recommendations of possible occupations. Thus, it is possible to provide alternatives to job seekers. In practice, we use two different samples for estimation (see chapter 2.4). The first one represents transitions from unemployment into employment. The second one covers all employment spells.

2.2 Data Description

The underlying dataset is a random ten percent sample of the Integrated Employment Biographies (IEB) which is described in Schneider (2020). It is an administrative dataset that contains the German workforce covered by the social security system from 1980 onwards and gives information about the characteristics of a certain person as well as the characteristics of their jobs. Using this dataset allows retracing the hard facts in the employment history.

The dataset combines different data sources. For the underlying research question, the Jobseeker Histories (ASU/XASU) and the Employee History (BeH) are required. The dataset must also be large enough because it contains a large number of occupational groups with a different number of observations. For every group a sufficient number of observations is crucial. The distribution of people (and observations) across jobs varies widely. While some occupational categories contain many observations, there are also occupational categories that contain comparatively few observations. Too little information (i.e. observations) would not provide a sufficient basis for an estimate for these groups. We consider workers to be employed if they had a job subject to social security and to be unemployed if someone is registered as unemployed at the public employment service (excluding those temporarily unfit for work).. In general, the basis for our estimations are employment spells. Thus, we obtain the characteristics of jobs and persons at the beginning of each spell.

2.3 Endogenous Variable

The jobs are grouped as defined in the German classification of occupations 2010 (KldB2010)[1]. Paulus/Matthes (2013) give detailed information on the data and the categories. The five breakdown levels are represented in Table 1.

Table 1: Levels of occupations
classification of occupations 2010

group	number of different occupational fields
occupational areas (1-digit)	10
occupational main-groups (2-digit)	37
occupational groups (3-digit)	144
occupational sub-groups (4-digit)	700
occupational types (5-digit)	1288

Source: statistics provided by the Federal Employment Agency

In this project, the estimation results should be able to support the placement process. The differentiation between occupational areas or occupational main groups is not precise enough, because this classification would provide no additional information for placement officers. Thus, it is necessary to define the job categories as precisely as possible, but at the same time, the number of observations per category has to be sufficiently large for estimation purposes. Because of this, we use the different job categories defined in the occupational groups (3-digit) on the one side and the occupational sub-groups (4-digit) on the other side as endogenous variables. More concrete, in the 3-digit case a classification problem with 144 different

[1] more detailed information in: https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/Systematik-Verzeichnisse/Systematik-Verzeichnisse-Nav.html

classes has to be solved, while in the 4-digit model has to deal with 700 different categories. An estimation for the 5-digit is not necessary because this group differentiates occupational types (i.e. qualification level). The decision which qualification level suits to a job seeker is based on formal rules or the individual qualification.
The number of observations for the occupational groups has a wide range. This difference in the size of the classes reflects the labour market situation or the change in the labour market due to technological change, for example. While there are more and more vacancies in some occupational fields, other occupational fields are in decline or die out. To get an overview of the job categories, we sort the table in ascending order according to the number of observations. Table 11 (in Appendix) shows the 3-digit job categories and the corresponding number of transitions as well as the total number of observations (i.e. the number of observations of jobs subject to social security in the original sample).

2.4 Choosing Samples for Estimation

Our estimations are based on two different samples. Both contain information from 1975 until 2018. The first sample, called "flow sample" in the following, is adjusted to the underlying research question. Here we only consider those employment spells (i.e. people who start a job subject to social security) with a preceding unemployment spell. Table 2 shows some descriptives for the 3-digit job categories for the whole dataset and the two subsamples.
There might be a problem with potential discrimination due to the unemployment status. The second sample, called "stock sample" in the following eliminates this problem. It contains all spells of all persons in our sample reporting an employment subject to social security. Therefore the reason why someone starts a new job, especially if someone was unemployed in the preceding period is not of interest. To be more precise, this sample includes also job-to-job transitions, for example due to a change in the occupational group within the same firm or the beginning of a new job in the same or another occupational group. Thus, here we do not distinguish between the employment status in the preceding time period or the incentive why someone is starts a new job. Information about the number of children and the martial status are only available in the Jobseeker Histories. This means, that there is a missing for persons who have no unemployment spell before an employment spell (in this case the information is written one period forward). We are aware that some groups like seasonal workers have a large number of employment spells in the same job category. This overrepresentation should be no problem for the statistical methods. Especially ML algorithms can capture this pattern.

Table 2: Descriptives for the different samples

	complete	stock	flow
no. of obs. total	141,327,440	54,781,854	6,019,817
no. of diff. pers.	3,251,949	2,883,188	2,359,075
median*	162,465.5	121,682.5	11,491.5
mean*	498,905.2	380,429.5	41,804.28
standard deviation*	1,020,538	750,014.1	84,801.27
min*	77	27	6
max*	7,761,145	6,053,898	532,939

* number of observations per job category

Source: own calculations

2.5 Exogenous Variables

In the following, the explanatory variables are described in more detail. Table 3 and Table 4 show the exogenous variables, the description and for factor variables the corresponding proportion of each factor level. First, the gender and the federal state a person lives in are included in the model. This information is available for every person. The number of observations across federal states reflects the difference in population size. Another explanatory variable is the nationality. For non-German, we distinguish between nationalities from asylum seekers[2] and EU and non-EU Europe. Whether a person really has a migrant background cannot be determined from the data.

The variables marital status and children are only captured in the Jobseeker Histories because the information is just relevant for the employment agency if a person becomes unemployed. The Employee History does not cover such information. Thus, we take the information from the previous unemployment sequence (i.e. the information is written one period forward). This may cause an error in some case if the status changes between the beginning of unemployment and the beginning of a new employment. Both variables were not captured in the same way from the beginning and had to be re-encrypted (more information in Schneider (2020)). According to the data description, due to the reclassification process or incorrect reporting some information can be lost or transformed incorrectly by creating the new variables. In the case of the children variable "nonresponse" and "no children" can't always be distinguished. The reference category thus captures both cases. Concerning the marital status, we only distinguish between living alone or in a partnership.

To improve the education variable, which is fraught with missing values and inconsistencies (i.e. a person has a lower educational level in a spell that comes after a spell with a higher one), we rely on the imputation process described in Fitzenberger/Osikominu/Völter (2006). After imputation the variable "education" contains only 2.24 percent missings. A further ex-

[2] Here we are referring to people whose nationality is equal to one of the eight countries of origin with the highest number of asylum seekers in July 2020 (source: https://de.statista.com/statistik/daten/studie/154287/umfrage/hauptherkunftslaender-von-asylbewerbern/). In particular, the countries are Syria, Iraq, Afghanistan, Turkey, Nigeria, Iran, Eritrea, Somalia and Georgia.

planatory variable is the job category in which a person carries vocational training in. We assume that a person who was in vocational training for more than 2 years in the same job category, has completed the vocational training in this category. If this person completes another vocational training afterwards for more than two years, the value changes into the new one. Here, we assume that the new vocational training is preferred because otherwise there would be no incentive for beginning another vocational training. Further, we take into account the age at the beginning of an employment spell for every person and the days in unemployment before finding a new occupation. Unfortunately, there is no information about the subject of the university degree.

Table 3: Explanatory categorical variables

variable	characteristics	proportion
gender	1 = female	39,95 %
	0 = male	60.05 %
federal states	1 = Nordrhein-Westfalen	17.51 %
	2 = Bayern	13.43 %
	3 = Schleswig-Holstein	3.31 %
	4 = Sachsen-Anhalt	4.16 %
	5 = Hessen	5.82 %
	6 = Baden-Württemberg	9.02 %
	7 = Brandenburg	4.08 %
	8 = Mecklenburg-Vorpommern	3.16 %
	9 = Thüringen	3.83 %
	10 = Sachsen	6.60 %
	11 = Niedersachsen	8.90 %
	12 = Bremen	0.83 %
	13 = Berlin	5.51 %
	14 = Hamburg	2.22 %
	15 = Saarland	0.98 %
	16 = Rheinland-Pfalz	4.05 %
nation	1 = German	88,66 %
	2 = EU	3,65 %
	3 = Europe without EU	4.76 %
	4 = 8 migration countries	0,90 %
	5 = remaining nations	2,03 %
marital status	1 = single/ live alone	54.90 %
	2 = partnership/ married	44.80 %
children	1 = at minimum one child under 15 years	26.44 %
	2 = no child/ unknown	73.56 %
education	1 = no school leaving certificate	0.02 %
	2 = grade-/ lower school certificate, intermediate school or equivalent qualification without vocational training	11.91 %
	3 = grade-/ lower school certificate, intermediate school or equivalent qualification with vocational training	65.33 %
	4 = upper secondary school leaving certificate, A-level equivalent, qualification for university without vocational training	1.29%
	5 = upper secondary school leaving certificate, A-level equivalent, qualification for university with vocational training	9.15 %
	6 = University of applied sciences	2.64 %
	7 = University	7.42 %
last skill level	skill level required for the previous employment	
last occupation	job category someone was employed in before starting a new employment	
completed vocational training	job category of vocational training	

Source: IEB; own calculations

Table 4: Explanatory variables

variable	description
age	age at the start of employment
days in unemployment	number of days a person is unemployed before starting a new employment
previous occupations	job categories of preceding employments
skill level	skill levels required for preceding employments

Source: IEB; own calculations

3 Estimation Strategy

For the underlying research question, pattern recognition as well as prediction performance are important. The main goal is finding a pattern that shows which personal and job characteristics are typical for certain occupations. We assume, that the abilities of machine learning (ML) may be able to map the given situation better than traditional methods. Additionally, comparing the results of ML methods with traditional methods is another central question.

3.1 Preparation

Before estimating the model, the dataset is split up in a training sample and test sample which is a common procedure in ML. The training set contains $2/3$ of the full dataset while the remaining observations are in the test set. The training set is used for estimating the model or fitting the parameters while the test set independent from the training set but follows the same probability distribution. In classification tasks, the algorithm determines the optimal combination of variables that generates a good predictive model. As the main goal is making predictions on unknown data the performance of a specified classifier is investigated by making predictions on the test set. Based on the results we make predictions for the training data and the test data. Results from the training set are called in-sample, while those from the test set are called out-of-sample. The empirical methods used for estimating the statistical models are OLS, Logit, k-Nearest-Neighbours and Random Forest. Afterwards, we compare the results and especially take a closer look at the difference between OLS and RF and try to determine their practical benefits (i.e. reduction of the unemployment rate).
Sometimes there are multiple observations from the same person in the dataset. Thus, there are always all observations from the same person either in the training set or in the test set. After having estimated the model, we predict the outcome for the training set and the test set. As a measure of goodness, we use the in-sample and the out-of-sample error, so we calculate

the percentage of wrong predictions. To be more precise, we define the job category with the highest predicted probability as first best job suggestion and check if this outcome coincides with the job category someone really gets employed in. If this is the case, the prediction is defined as right and otherwise wrong.
Note that the underlying classification problem is imbalanced between the classes. This means that the sample size of several classes is significantly higher or lower in comparison to other classes. This may negatively affect the performance because relatively balanced class distributions are assumed for classification problems. Sun/Wong/Kamel (2009) found that imbalanced class distribution might be no problem if the dataset is large enough because it contains also more information about the small class. Hence, pattern recognition is also possible for small classes because distinguishing rare samples from the majority is possible. Nevertheless, in decision trees of random forests, some branches that predict the small classes are probably removed because the new leaf node is labeled with a dominant class.

3.2 OLS and Logit

As multinomial models can not be estimated with OLS or Logit, we estimate a single equation for every job category in both cases. Concretely, 144 equations for the 3-digit and 700 equations for the 4-digit. First, we define a ixj-matrix **M**, where

$$M_{ij} = \begin{cases} 1, & \text{if person } i \text{ is employed in job category } j \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

The OLS model is

$$(M_{ij} = 1) = \mathbf{X}_i\beta + \mathbf{Y}_i\gamma + \epsilon_i \tag{3.2}$$

and the logistic model

$$P(M_{ij} = 1) = \frac{exp(\mathbf{X}_i\beta + \mathbf{Y}_i\gamma + \epsilon_i)}{1 + exp(\mathbf{X}_i\beta + \mathbf{Y}_i\gamma + \epsilon_i)}, \tag{3.3}$$

where β is a Sx1-vector and γ is a Rx1-vector. Thus, in practice J single equations have to be estimated.
After estimating the models we make predictions for every job category in the training sample and test sample. Afterwards comparing the results with the RF results is possible, as a ranked list of occupations can be obtained from all estimation methods. The category with the highest predicted probability is chosen as best the suggestion for a person. It is also possible to compare the predicted category with the real one. Note, that in the case of OLS the outcome values for predictions can not be interpreted as probabilities as in the Logit case.

The resulting values do not lie between 0 and 1. However, the interpretation of coefficients is not of interest in this application and the category with the highest value can be chosen as the suggested job category.

3.3 Random Forest

For constructing a classification model a learning algorithm that obtains the relationship (i.e. the pattern) between the attribute set and the class label is crucial. The model that fits the training data best is generated. Getting good predictions on unseen data is an important point but only possible if the model fits the training data. Hastie/Tibshirani/Friedman (2017) and Breiman (2001) give an overview of random forests and classification. Random forests are constructed by using bagging[3] to build de-correlated trees that are averaged in the end. By definition the random forest classifier is a collection of tree-structured classifiers given by

$$h(\mathbf{z}, \Theta_k), k = 1, \ldots, \tag{3.4}$$

with Θ_k are identically distributed random vectors and $\mathbf{z}$ is the input. In the underlying context, $\mathbf{z}$ contains the variables denoted by $\mathbf{X}_i$ and $\mathbf{Y}_i$. Hence, for classification each tree which is built from a random vector of parameters gives a vote for the most popular class. Finally, the random forest classifies using the majority vote.
The algorithm is described in Liaw/Wiener (2002). In a first step bootstrap samples from the original dataset are drawn. Then, for each of the samples, an unpruned tree for classification is grown by choosing a random sample of predictors and taking the best split. In the last step, the predictions on the new data are computed by aggregating the predictions of the trees (i.e. majority votes for classification). Fitting a machine learning model requires the right choice of hyperparameters. Hyperparameters are second-order parameters that can highly influence the outcome and the predictive performance of a model. Sometimes, hyperparameter settings are chosen a priori, but it can also be advantageous to tune (i.e. try different settings) them before fitting the model on the training data. In random forest the hyperparameters are the number of decision trees being built in the forest and the number of features considered by each tree when splitting a node. Concerning the number of features Bernard/Heutte/Adam (2009) show that the default values that are suggested by the traditional literature are nearly optimal. The most common values are $\sqrt{M}$[4] and $log_2 M + 1$ with M is the number of features in the original dataset. On the other side, Probst/Boulesteix (2018) show that tuning is not recommendable for multiclass classification by comparing the results of estimations with 11 trees up to 2000 trees. We did the same for our model and ob-

[3] For reducing the variance, bagging is used to average many noisy but unbiased models.
[4] In this case the random forests were fit using the R package ranger with 100 trees and $\sqrt{M}$.

tain the same results. The classification error is almost identical for any number of trees. Since the dataset under investigation is very large computational performance is crucial. We use the R package ranger as allows a fast implementation of random forests. Ziegler/Wright (2017) compare several packages that estimate random forests concerning the runtime and the memory usage. They conclude that the ranger package is highly efficient without loosing performance.
In practice, the random forest algorithm estimates a multinomial classification model with 144 and 696 different classes. It is possible to calculate probabilities for each class or to get a vector that contains the category with the highest probability as resulting output. Thus, this vector can be directly compared with the vector of real job categories. For getting more than one job suggestion computing the matrix containing all probabilities is necessary.

3.4 K-Nearest-Neighbours

As a second machine learning method, the k-nearest-neighbor algorithm is applied. Originally, Fix/Hodges (1989) started investigating the kNN classifier. It is an unstructured, memory-based and non-parametric classification method that can be used for classification and regression. Hastie/Tibshirani/Friedman (2017) give an introduction to the kNN classifier. The main idea is finding k training points, where $k = 1, \ldots, N$, for a given query point x_0. The k training points that are closest in distance to x_0 are used for classification by using the majority vote among the k neighbours. The most important hyperparameters for kNN are k and the distance metric. Under the assumption that the features are real-valued and by using the Euclidean distance, d_i is given by

$$d_i = ||x_i - x_0||. \tag{3.5}$$

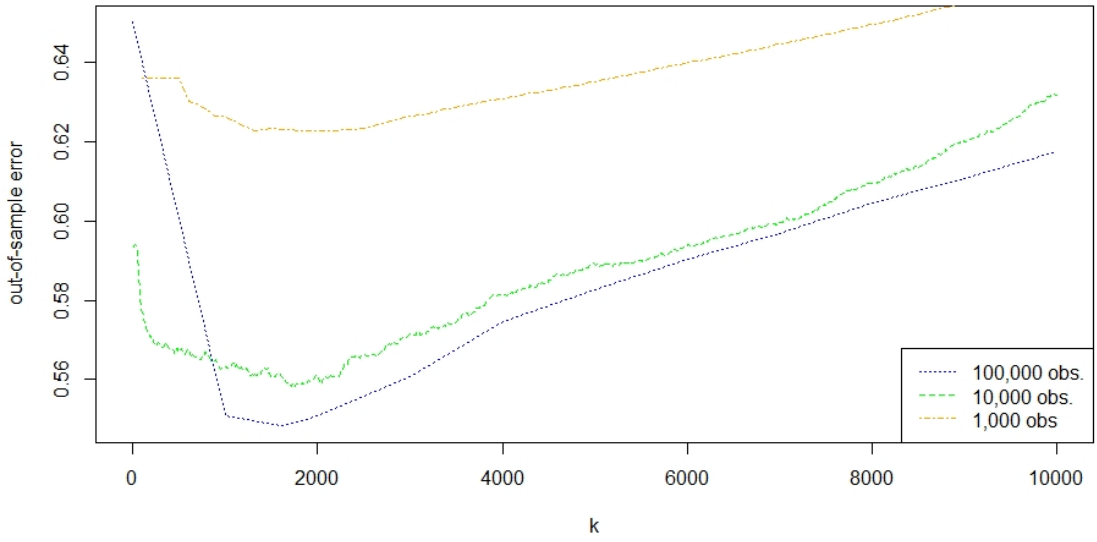
Standardizing the training data is a common procedure. Since the algorithm relies on distances the features should have the same units.
The parameter k is the number of neighbours that are taken into account for classification. An object is classified by choosing the plurality vote of its k neighbours. This value is thus the core of the algorithm. If k is too low, the model becomes very specific and consequently, very sensitive for noise. It results in an overfit model that leads to high accuracy on the training set, but poor performance for new data. On the other side, if k is too large, the model is underfit. The model becomes too general and fails to accurately predict both samples. Since there is no general formula to determine a suitable value for k, in practice, hyperparameter tuning is used to do this. Naturally, there is a default value for each hyperparameter, but achieving optimal performance of the kNN algorithm requires manual selection. To do this, a search space for hyperparameters, a data-driven optimization algorithm and an evaluation method are selected.

There is the possibility of placing weights on the classifier. Typically, equal weight is placed on each class of the k nearest neighbours to x_0.
Since every data point has to be stored, the algorithm becomes very slow with an increasing number of data points. We thus consider subsamples that contain 25 percent of the training set (i.e. about 1,000,000 observations) and take the average of the classification errors. In a first step, finding the right value for k is important. As a rule of thumb, one could use $k = \sqrt{N}$, but a more precise specification can reduce the classification error dramatically. In this application, the standard hyperparameter tuning is not possible, due to the large dataset and the huge number of possible k-values. Instead, we estimate the model with different k's and calculate the classification error. As the out-of-sample error is of interest, we choose the value that minimizes the out-of-sample prediction. Nevertheless, due to technical restrictions this procedure is only possible up to 100,000 observations. Thus, we calculate the optimal k for random samples of 1000, 10,000 and 100,000 observations and use this information for estimating the optimal k for an estimation with 1,000,000 observations. Figure 1 shows the out-of-sample errors for different k's for every number of observations. Here we obtain a reduction of the out-of-sample error with an increasing number of observations and that the courses of the graphs look very similar. One can also see, that choosing the right k is important for optimizing the prediction quality. Figure 2 shows the best value for k in relation to the number of observations. Here one could suspect a linear relationship. Under the assumption of a linear relationship, the right k-value for 1,000,000 observations would be 15,924. Using a very small k leads to very small in-sample errors, but large out-of-sample errors, which might be evidence for overfitting. Tuning k with focus on optimizing the out-of-sample error comes along with an drastically increase in the in-sample error. Therefore, there is no evidence for overfitting if k is chosen correctly.

4 Empirical Results

We use both, the "stock sample" and the "flow sample" for our estimations. The predicted probability for every job category allows us to create a ranked list with job suggestions for every individual. For comparison purposes, we only consider the category with the highest probability as prediction. The measure of goodness is the in-sample and the out-of-sample error that denotes the percentage of wrong predictions.

Figure 1: Out-of-sample errors for different k's
Out-of-sample error rate in decimal numbers (0-1)

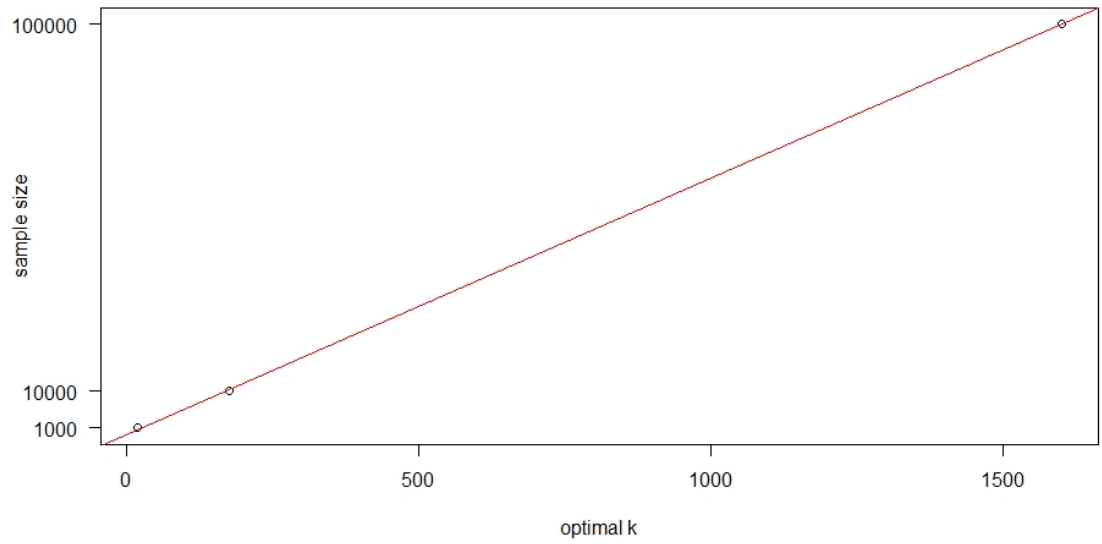


Source: own calculations

4.1 Results

We start with the 3-digit job categories (i.e. 144 different classes) as endogenous variable. The model is estimated with OLS and Random Forest (RF). We estimate the model based on the training sample. If a person is chosen to be part of the training sample, we include all spells of this person. Afterwards we evaluate our model by making predictions based on the training and the test sample and calculate the error rates.
We begin with the "stock sample" (compare section 2.4). Thus, we make predictions for unemployed persons based on people from the whole job market with similar characteristics. In order to limit the computational burden, we draw a random sample of 5 percent for training the model. The test-sample has the same size as the training sample. Table 5 shows the results for OLS and RF. The error rate for RF is 14.40 percent lower. Further, we look at predictions only for observations that are part of the "flow sample". To disentangle the proportion of transitions, we extract the corresponding group of persons and calculate the error rate only for this group. In this case error rate of RF is 30.25 percent lower.
In a second step, we use the "flow sample" (compare section 2.4). In practice, this means that we consider only jobs categories in which an unemployed job seeker got employed. In the "flow sample", the training set has 3,974,727 observations of 1,556,969 persons while the test set has 2,045,090 observations of 802,086 persons. Table 6 shows the results for the new

Figure 2: Relationship between k and the number of observations



Source: own calculations

Table 5: Prediction errors based on the "stock sample" for OLS and RF

method	in-sample error	out-of-sample error
OLS	19.07 %	19.03 %
OLS flow	55.53 %	55.48 %
RF	7.36 %	16.29 %
RF flow	14.61 %	38.70 %

Source: own calculations ©IAB

more specific sample. Again, the RF error rate is 6.7 percent lower than the OLS error rate. In comparison to the "stock sample", there is an increase in the error rate. If we look at the transitions in the RF model, the error rate is 17.1 percent lower if we take the large sample that covers all persons subject to social security. This shows us, that using information based on the biographies from people from the whole job market is the best way for making suggestions for job seekers.

Table 6: Prediction errors based on the "flow sample" for RF and OLS

method	in-sample error	out-of-sample error
OLS	49.91 %	50.03 %
RF	14.32 %	46.69 %

Source: own calculations

4.2 Finding the Optimal Training Sample in Practice

Since there was a change in the German classification system of occupations from 2011 on, there might be a structural break in this point of time. The goal of the new classification (i.e. KldB2010) was to map the structure of occupations in Germany better. Further, there is a higher comparability to the international classification of occupations (ISCO-08). To check for this, we restrict the sample and consider only employment spells beginning after 1st January 2012. To be more precise, we take a subsample of the original data containing spells with concrete job recordings after the change in the coding scheme. Additionally, in contrast to further training data that contains observations from all years that occur in the sample, we adjust the splitting of the dataset to the underlying research question. In practice, there is no information from further years for creating a list of recommendations. To map the setting to this situation, we take employment spells beginning from 2012 until 2017 for the training sample and use the observations from 2018 as test sample. As a robustness check, we took several other periods of time for a test-train split. We find that using data from 2012 on minimizes the error rate which is consistent with the point of time where the structural break occurs[5]. Figure 3 shows the different classification errors according to the different time periods. Naturally, this procedure for finding the optimal time period for the training sample could be repeated every year. This would guarantee that the training sample is always chosen in a way of minimizing the out-of-sample error for the preceding year.
We start the calculations again based on the "stock sample". Table 7 shows the results.

Table 7: Error rates based on the "stock sample" for OLS and RF
training sample with observations from 2012-2017 and test sample with observations from 2018

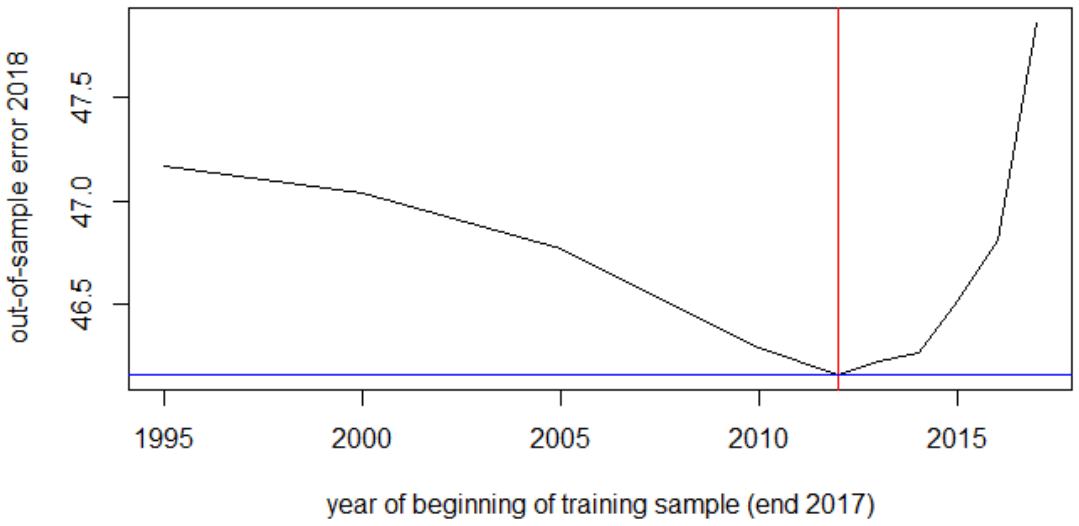
method	in-sample error	out-of-sample error
OLS	17.56 %	16.14 %
OLS flow	58.24 %	60.70 %
RF	5.70 %	15.32 %
RF flows	13.67 %	42.20 %

Source: own calculations

Here we obtain a reduction of the error rate by 5.2 percent for RF compared to OLS. If we look only at transitions the difference increases to 30.5 percent. Compared to the "stock sample" results in the previous section, most of the other error rates are a little bit lower. The results for RF are contrary. Here the error rate increases by 8.3 percent compared to prior estimations for the "stock sample".
In the "flow sample", the training sample contains 1,435,474 observations while the test sample contains 740,266 observations. Table 8 gives on overview of all results based on this sample. Compared to results where the structural break is not taken into account, the out-of-

[5] We obtain the same results when taking the period up to 2016 as training sample and the years 2017 and 2018 as test data. Thus, in general, taking data from 2012 on is the best choice when the year of starting a new employment is taken into account

Figure 3: Out-of-sample errors for different time periods



Source: own calculations

Table 8: Error rates based on the "flow sample" for OLS and RF
training sample with observations from 2012-2017 and test sample with observations from 2018

method	in-sample error	out-of-sample error
OLS	51.08 %	47.25 %
RF	13.38 %	46.06 %

Source: own calculations

sample rate again is reduced by 1.35 percent for random forest.
Finally, we clearly see that RF always produces better results than OLS. The difference between the in-sample and out-of-sample error is small if we look at all OLS results. In contrast, the RF algorithm produces a larger difference, because of very good in-sample results. This shows, that the training data is mapped much better by the RF algorithm. The job suggestions produced by statistical methods often are identical with the job category someone really gets employed in. Thus, statistical methods could produce useful job suggestions in practice. Concluding, for the underlying research question machine learning methods should be preferred to common statistical methods.

4.3 An Illustrative Scenario: Hypothetical Effects of Using RF instead of Logit

For illustrative purposes, we calculate the potential consequences in terms of unemployment rate of using the RF instead of common satistical methods. We take the best result for RF (compare table 7) and the common methods (compare table 8[6]). Evidently, such calculations can only be executed under assumptions in a hypothetical scenario. An extreme assumption would be that in a given month an unemployed finds a job if she receives the top 1 recommendation and otherwise not. In the three-digit case, the share of correct recommendations from RF is 18 percent higher than from Logit. I.e., the job-finding rate would also be 18 percent higher. Equilibrium unemployment in the standard case of two labour market states (employed and unemployed) with job finding rate f and separation rate s results as $u = \frac{s}{f+s}$. We use a standard calibration for the German labour market with monthly $f = 0.08$ and $s = 0.006$ (compare Hartmann et al. (2021)). Note that since neither Logit nor RF so far have been used in reality (and thus are not part of the data generating process), this serves only as a reference point by assumption. Then, increasing f by 18 percent reduces u by 1.0 percentage points.

Of course, not receiving the top 1 recommendation does not necessarily imply an unsuccessful job search. Instead, as a measure of the proximity of the two recommendations from RF and Logit, we can use their probability difference. This follows the idea that the sharper two jobs can be discriminated by the RF model, the more likely it will be that the job search will be more successful with a correct recommendation. In detail, both for the RF and the Logit top 1, we sum the probabilities from the RF estimation (compare equation 3.4) for the cases where RF top 1 is true and Logit top 1 is not. This sum for Logit top 1 is 35 percent of the sum for RF top 1. Hence, the 18 percent from above are downscaled, multiplying by $1 - 0.7$. Logically, also the u-difference would be smaller, namely 0.3 percentage points. Naturally, the calculated difference is lower than above because incorrect recommendations are punished only in so far as the probability is lower than the RF top 1.

Note that the reduction refers to a pure comparison of two methods in order to illustrate the differences under assumptions. It does not imply that data-based matching per se would reduce the unemployment rate by the same amount. This question would lend itself to experimental evaluation.

[6] In this case the Logit error rates are a little bit higher than the OLS error rates. Since probabilities are required for the calculation, we use the Logit results for the comparison with RF.

5 Robustness Checks and Discrimination

5.1 Check List of Suggestions

As the empirical results can be used to advise unemployed people not only the first best suggestion is of interest. The placement officer should be able to make a couple of suggestions and so the unemployed person can choose which job is best suited. This makes the decision process more flexible. The placement officer can consider the clients preferences. Maybe the person wants to find employment which is close to the residence while for others the resulting job suggestions may give them some new perspectives. An unemployed person could also get an impulse for retraining. As not only the first best job suggestion (TOP 1) is relevant for a job seeker, we also considered the top three categories (TOP 3). Table 9 shows the corresponding prediction errors. The OLS and Logit error is 12.71 percent and thus, about 37 percentage points lower than the first best case. In random forest the error is only 0.11 percent and 20 percentage points lower compared to the model with one category. This implies that the three categories with the highest probabilities almost always contain the true category.

Furthermore, we obtain that the out-of-sample performance is almost equal for all estimation methods. The random forest application performs a little bit better. It appears quite differently for the in-sample case. Here, the random forest application performs much better. In the 3-digit case, the error is under 1 percent. Concluding, the results for the 3-digit estimation show that the machine learning algorithm performs much better in-sample (i.e. the given situation is mapped better) while the out-of-sample performance does not differ much from the other methods. Therefore, also in this case the in-sample error of RF is much lower compared to the other methods. The RF out-of-sample error differs not that much from the other results, but still by about 7 percent.

Table 9: Prediction errors taking the top three categories into account for OLS and RF
training sample with observations from 2012-2017 and test sample with observations from 2018

method	in-sample error	out-of-sample error
OLS	12.71 %	12.79 %
RF	0.11 %	11.20 %

Source: own calculations

5.2 Extending the Number of Job Categories

Another modification is to extend the endogenous variable from the 3-digit to the 4-digit. Thus, the number of different classes increased from 144 to 696. Table 10 shows the results. Making predictions for the first best category leads to an almost similar in-sample and out-of-sample error of about 58 percent. Comparing this with the 3-digit results shows that the error rate increases by about 8 percentage points when having almost five times more different categories. The error rate for both samples for the three best categories is given by about 27 percent which means that the rate is about 14 percentage points higher than in the 3-digit case. Due to capacity problems estimating the 4-digit model on the whole training set with random forest is not possible. Thus, we take ten random samples of 30 percent[7]. After having finished the forecasts on the reaming sub-sample and the whole test sample, we compute the in-sample error and the out-of-sample error in the same way as in the 3-digit case. Note, that some of the categories with a small number of observations drop out when only using a 30 percent sample[8]. This leads to a reduction of job categories, in this case on average 30 categories drop out. We also obtain almost similar results for all sub-sample estimations. This is an indication that taking a 30 percent sample is sufficient to answer the research question for the largest part of jobs. In order to map occupations that do not have many entries, it would nevertheless be advantageous to be able to estimate the whole model. The in-sample error for the first best category is given by 27 percent and the out-of-sample error is given by 57 percent. Looking at three occupations reduces the in-sample error to 2.7 percent and the out-of-sample error to 24.8 percent. Thus, compared to the 3-digit estimation the in-sample error increases by 5 percentage points while the out-of-sample error increases by 10 percentage points. For the top 3, the in-sample error is 2 percentage points lower, while the out-of-sample error is 13 percentage points lower than in the 4-digit case.
In summary, OLS performs very consistently for both samples. Neither in the case of 3- or 4-digit nor comparing only the first best or more category there is a big difference between the error rates. Nevertheless, the error rates lie above the random forest results in every case.

Table 10: Prediction errors (TOP 1 and TOP 3) for the 4-digit job categories for OLS and RF
training sample with observations from 2012-2017 and test sample with observations from 2018

sample no.	in-sample error	out-of-sample error
RF TOP 1	26.986 %	56.872 %
RF TOP 3	2.679 %	24.795 %
OLS TOP 1	58.25 %	58.47 %
OLS TOP 3	26.68 %	26.85 %

Source: own calculations

The in-sample error of the random forest prediction of the first best category is in the 4-digit case by 5.5 percent larger than in the 3-digit case, the out-of-sample error is by 9.7 percent.

[7] 30 percent is the maximum amount of data where the machine is able to compute
[8] Table 12 and Table 13 (in Appendix) show the single results in detail.

Keeping in mind that the number of categories increases by 556, the random forest application is also able to handle a large number of categories efficiently.

5.3 Additional Methods

Another representative for common methods is given by logit. Here we also estimate a singe equation for every category. Afterwards a ranked list can be created by comparing the predicted probabilities. Here, the results (i.e. the list with suggestions) are almost identical with the OLS results. Although OLS does not produce probabilities, these results show that in this application it is as good as logit.
Another promising machine learning method is the k-Nearest-Neighbours (kNN) algorithm. We obtain an out-of-sample error of 54.92 percent for $k = 2000$. Since we find that the optimal k is 16,000 and that there is a linear relationship for the underlying dataset (see section 3.4) the out-of-sample error could be reduced to 47.80 percent. RF still performs better and the model is more efficient to compute than the kNN model. Comparing the results with corresponding OLS results shows, that the error rate is decreased by 2.3 percentage points[9]. Thus, we obtain also for kNN that machine learning performs better than microeconometric methods. Compared to RF the algorithm needs much more computational power and is more time consuming. Thus, we prefer RF for our estimations.

5.4 Discrimination

It is well known that algorithms could also discriminate certain persons or certain groups of persons. Naturally, human decision making takes place on individual basis. In contrast, the algorithm works based on the underlying data. Kleinberg et al. (2020) show that algorithms have also potential to detect and prevent discrimination. In practice, data-based recommendations could be used by placement officers in addition to their own expertise and individual impressions of the person to be matched.
Evidently, recommendations based on realised data may perpetuate certain labour market patterns that may be undesired. This can be related to characteristics such as gender, nationality or region. However, we argue that these characteristics should be included in the estimation approaches in order to be able to control their effect: Once their influence is estimated, practitioners can decide whether to use them or to switch them off. In contrast, if one neglected the variables already during estimation, their explaining power is likely to be partly taken over by other correlated variables. Then, controlling the effects and deciding

[9] We use the "flow sample" and the test-train split equal to section 4.1.

about their usage would be impossible.
For reasons of transparency, we illustrate the influence of the gender, nationality and region characteristics. For that purpose, we simulate artificial variables by randomly drawing from the realised distribution. I.e., estimation is performed with the true data, but recommendations are generated based on the artificial data. First, we determine the out-of-sample errors based on artificial data and compare them with the errors from Table 6. We find that the results get clearly worse if the gender variable is pertubed. The simulation of the other variables also causes a decrease in the classification accuracy, but in this case the difference is very small. Thus, regarding the aim to minimize the classification error, keeping the variables in the model is important. As argued above, this has the further advantage that in practice, one can decide if a certain variable should be eliminated for example due to discrimination or not.
A potential concern may be that recommendations based on realised data only replicate the past behaviour of placement officers. However, the universe of labour market data we use to train the algorithms contains a variety of job findings that were not mediated by the employment service. Indeed, Kubis (2011) finds based on the German Job Vacancy Survey that search via the public employment agency is the decisive search channel only in 7.4 percent of all hires.

6 Conclusion

In this paper, we have shown that applying empirical methods, especially machine learning algorithms, can play an important role in improving the matching on the labour market. Using data that covers past matches and the corresponding information about the characteristics of the persons and the jobs for estimating different models leads to recommendations for job seekers. Thus, job suggestions from two sources, on the one hand, the placement officer who can use his experience and gets an overview of the motivation and the skills and on the other hand the data-driven algorithm that makes suggestions based on a sample that covers the whole population could have a positive effect on the unemployment rate. Using administrative data on employment biographies covering the full universe of employment spells allows us to build statistical models that map very well the current situation on the labour market. For estimating these models, we use OLS, Logit, k-Nearest-Neighbours (kNN) and Random Forest (RF). We use two different samples for our estimations. The "stock sample" that covers all persons with jobs subject to social security and the "flow sample" that contains only transitions from unemployment to employment. We find, that the "stock sample" minimizes the error rates more than using the "flow sample". Furthermore, the performance of RF is best in every modification. Especially, the in-sample error differs highly. Based on the

"stock sample" the error rates for transitions are about 30 percent lower for RF than for OLS. This clearly shows, that RF should be preferred compared to the other methods.
Another advantage is the adaptability of the RF algorithm. Estimating the model with different modifications shows that RF always performs best. The ability of the algorithm to find interactions itself for example eliminates the need for a time-consuming manual modification of the model. Thus, the model can easily be adapted to different datasets without loosing predictive power.
For clarifying the relevance of the results, we take a closer look at the difference between RF and OLS. Thus, we try to figure out the importance of our results for the unemployment rate. Naturally, our calculations have to rely on assumptions in order to determine an impact on a theoretical basis. Thus, they are for illustrative purposes only. In terms of the unemployment rate hypothetically, the advantage of RF compared to common methods would stand for a difference of 0.3 - 1.0 percentage points.
In future research, the value of concrete recommendations in practice would be assessable by randomised experiments. For such approaches, we would further develop the underlying models. Particularly, while ML algorithms already bring clear progress using a data set of standard properties, incorporating further data, for example on specific competencies, may represent a promising path for concrete practical usefulness.
Indeed, one possibility for increasing matching quality (i.e. the model) could be information about skills. The Federal Employment Agency also collected data on skills based on self-assessment. If this data can be tapped for empirical purposes, the skill-data can be combined with the IEB and then one could investigate how far the matching process can be further improved. The importance of having information about skills is also discussed in Chakravarty/R./Lindsay (2020) or Rentzsch/Steneva (2020).

References

Altmann, Steffen; Falk, Armin; Jäger, Simon; Zimmermann, Florian (2018): Learning about Job Search: A Field Experiment with Job Seekers in Germany. In: Journal of Public Economics, Vol. 164, p. 33–49.

Bauer, Anja (2013): Mismathch unemployment: Evidence from Germany, 2000-2010. In: IAB-Discussion Paper.

Belot, Michele; Kircher, Philipp; Muller, Paul (2019): Providing Advice to Jobseekers at Low Cost: An Experimental Study on Online Advice. In: The Review of Economic Studies, Vol. 86, p. 1411–1447.

Bernard, Simon; Heutte, Laurend; Adam, Sébastian (2009): Influence of Hyperparameters on Random Forest Accuracy. In: International Workshop on Multiple Classifier Systems, p. 171–180.

Blundell, Richard; Costa Dias, Monica; Meghir, Costas; Van Reenen, John (2004): Evaluating the Employment Impact of a Mandatory Job Search Program. In: Journal of the European Economic Association, Vol. 2, p. 569–606.

Boselli, Roberto; Cesarini, Mirko; Mercorio, Fabio; Mezzanzanica, Mario (2017): Using Machine Learning for Labour Market Intelligence. In: In Machine Learning and Knowledge Discovery in Databases - European Conference, ECKML PKDD 2017, Skopje, Macedonia, p. 330–342.

Braun, H. (2017): Applying Learning-to-Rank to Human Resourcing's Job-Candidate Matching Problem: A Case Study. Master's thesis, Radboud Universiteit.

Breiman, Leo (2001): Random Forests. In: Machine Learing, p. 5–32.

Cao, Zhe; Qin, Tao; Liu, Tie-Yan; Tsai, Ming-Feng; Li, Hang (2007): Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th international conference on Machine learning, p. 129–136.

Chakravarty, Surajeet; R., Kaplan Todd; Lindsay, Luke (2020): Increasing Employment Through the Partial Release of Information.

Colombo, Emilio; Mercorio, Fabio; Mezzanzianica, Mario (2018): Applying machine learning tools on web vacancies for labour market and sill analysis. In: Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence.

Faliagka, Evanthia; Ramantas, Kostas; Tsakalidis, Athanasios; Tzimas, Giannis (2012): Application of machine learning algorithms to an online recruitment system. In: Proc. International Conference on Internet and Web Applications and Services, Citeseer, p. 215–220.

Fang, M. (2015): Learning to Rank Candidates for Job Offers using Field Relevance Models. Master's thesis, University of Groningen and Saarland University.

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2006): Imputation Rules to Improve the Education Variable in the IAB Employment subsample. In: Schmollers Jahrbuch, Jg. 126.

Fix, Evelyn; Hodges, Joseph Lawson (1989): Discriminatory analysis. Nonparametric discrimination: Consistency properties. In: International Statistical Review/Revue Internationale de Statistique, Vol. 57, No. 3, p. 238–247.

Gerunov, Anton (2014): Big Data Approaches to Modeling the Labour Market.

Hall, Robert E.; Schulhofer-Wohl, Sam (2018): Measuring Job-Finding Rates and Matching Efficiency with Heterogeneous Job-Seekers. In: American Econonmic Journal: Macroeconomics, Vol. 10, No. 1.

Hartmann, Michael; Klaus, Anton; Beckmann, Ralf; Stephani, Jens (2021): Monatsbericht zum Arbeits- und Ausbildungsmarkt. In: Berichte: Blickpunkt Arbeitsmarkt.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2017): The Elements of Statistical Learning – Data Mining, Inference, and Prediction.

Hutter, Christian; Weber, Enzo (2017): Mismatch and the Forecasting Performance of Matching Functions. In: Oxford Bulletin of Economics and Statistics, Vol. 79, p. 101–123.

Kleinberg, Jon; Ludwig, Jens; Mullainathan, Sendhil; Sunstein, Cass R (2020): Algorithms as discrimination detectors. In: Proceedings of the National Academy of Sciences, Vol. 117, No. 48, p. 30 096–30 100.

Kubis, Alexander (2011): IAB-Stellenerhebung 1/2021: Betriebe sind bei Neueinstellungen weiter zurückhaltend. In: IAB Forum.

Li, Hang (2011): A Short Introduction to Learning to Rank. In: IEICE TRANSACTIONS on Information and Systems, Vol. 94, No. 10, p. 1854–1862.

Liaw, Andy; Wiener, Matthew (2002): Classification and Regression by randomForest. In: R News, Vol. 2/3, p. 18–22.

Mortensen, Dale T.; Pissarides, C.A. (1994): Job Creation and Job Destruction in the Theory of Unemployment. In: Review of Economic Studies 61, p. 397–415.

Paulus, Wiebke; Matthes, Britta (2013): Klassifikation der Berufe - Struktur, Codierung und Umsteigeschlüssel. In: FDZ Methodenreport.

Petrongolo, B.; Pissarides, C. A. (2001): Looking into the Black Box: A Survey of the Matching Function. In: Journal of Economic Literature XXXIX.

Probst, Philipp; Boulesteix, Anne-Laure (2018): To Tune or Not to Tune the Number of Trees in Random Forest. In: Journal of Machine Learning Research, , No. 18, p. 1–18.

Rentzsch, Robert; Steneva, Mila (2020): Skills-Matching and Skills Intelligence through curated and data-driven ontologies. In: Proceedings of the DELFI Workshops 2020.

Sahin, Aysegül; Song, Joseph; Topa, Giorgio; Violante, Giovanni L. (2014): Mismatch Unemployment. In: American Economic Review, Vol. 104, No. 11.

Schneider, Andreas (2020): IEB Integrierte Erwerbsbiografien. In: DIM Datenreport.

Sun, Yanmin; Wong, Andrew K. C.; Kamel, Mohamed S. (2009): Classification of Imbalanced Data: A Review. In: International Journal of Pattern Recognition and Artificial Intelligence, Vol. 23, No. 4, p. 687–719.

van Belle, Agnes; Dehling, Eike; Foster, Daniel (2018): Improving Candidate to Job Matching with Machine Learning.

Van Huynh, Tin; Van Nguyen, Kiet; Nguyen, Ngan Luu-Thuy; Nguyen, Anh Gia-Tuan (2010): Job Prediction: From Deep Neural Network Models to Applications. In: IEEE RIVF 2020 Conference.

Ziegler, Andreas; Wright, Marvin N. (2017): ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. In: Journal of Statistical Software, Vol. 77, p. 1–17.

Appendix

7 Appendix

Table 11: Occupational categories
number of transitions (trans.) and the total number of observations of jobs subject to social security (obs. comp.)

3-digit no.	occupation	trans.	obs. comp.
13	Junior non-commissioned officers	6	77
14	Armed forces personnel in other ranks	10	121
12	Senior non-commissioned officers and higher	18	260
411	Occupations in mathematics and statistics	218	3,442
116	Occupations in vini-and viticulture	355	3,710
11	Commissioned officers	458	2,156
931	Occupations in product- and industrial design	488	4,629
421	Occupations in geology, geography and meteorology	521	5,209
912	Occupations in the humanities	536	4,310
114	Occupations in fishing	543	5,260
947	Technical and management occupations in museums and exhibitions	548	6,523
936	Occupations in musical instrument making	554	8,875
625	Sales occupations (retail) selling books, art, antiques, musical instruments, recordings or sheet music	575	8,672
824	Occupations in funeral services	655	10,819
822	Occupations providing nutritional advice or health counselling, and occupations in wellness	731	9,311
423	Occupations in environmental protection management and environmental protection consulting	741	6,758
933	Occupations in artisan craftwork and fine arts	750	9,700
712	Legislators and senior officials of special interest organisations	845	22,848
934	Artisans designing ceramics and glassware	905	14,869
532	Occupations in police and criminal investigation, jurisdiction and the penal institution	1,015	11,334
943	Presenters and entertainers	1,194	13,449
922	Occupations in public relations	1,394	14,240
815	Occupations in veterinary medicine and non-medical animal health practitioners	1,404	13,999
816	Occupations in psychology and non-medical psychotherapy	1,703	18,268
113	Occupations in horsekeeping	1,765	17,621
523	Aircraft pilots	1,796	22,194
211	Occupations in underground and surface mining and blasting engineering	2,041	75,535
833	Occupations in theology and church community work	2,126	46,217
941	Musicians, singers and conductors Occupations in environmental protection engineering	2,152	34,336

Table 11 continued...

422	Occupations in environmental protection engineering	2,303	33,398
522	Drivers of vehicles in railway traffic	2,309	35,707
514	Service occupations in passenger traffic	2,388	26,313
511	Technical occupations in railway, aircraft and ship operation	2,489	52,430
524	Ships officers and masters	2,568	36,065
512	Occupations in the inspection and maintenance of traffic infrastructure	2,589	25,331
935	Artisans working with metal	2,742	48,610
214	Occupations in industrial ceramic-making and -processing	2,885	59,890
946	Occupations in stage, costume and prop design	2,999	10,986
634	Occupations in event organisation and management	3,085	34,743
613	Occupations in real estate and facility management	3,179	36,361
533	Occupations in occupational health and safety administration, public health authority, and disinfection	3,204	34,543
291	Occupations in beverage production	3,311	40,516
414	Occupations in physics	3,523	71,764
112	Occupations in animal husbandry	3,594	45,479
312	Occupations in surveying and cartography	3,771	50,632
261	Occupations in mechatronics, automation and control technology	3,781	41,964
923	Occupations in publishing and media management	3,862	71,509
911	Occupations in philology	4,030	74,267
944	Occupations in theatre, film and television productions	4,123	23,346
433	Occupations in IT-network engineering, IT-coordination, IT-administration and IT-organisation	4,199	61,955
283	Occupations in leather- and fur-making and -processing	4,544	120,743
913	Occupations in economics	4,762	46,193
233	Occupations in photography and photographic technology	4,871	90,776
243	Occupations in treatment of metal surfaces	4,955	88,927
711	Managing directors and executive board members	5,056	68,874
432	Occupations in IT-system-analysis, IT-application-consulting and IT-sales	5,189	68,948
733	Occupations in media, documentation and information services	5,813	63,418
812	Laboratory occupations in medicine	5,900	107,984
213	Occupations in industrial glass-making and -processing	5,911	87,687
115	Occupations in animal care	6,024	69,927
434	Occupations in software development and programming	6,175	76,895
914	Occupations in economics	6,417	91,538
515	Occupations in traffic surveillance and control	6,477	107,314
281	Occupations in textile making	7,004	173,895
412	Occupations in biology	7,678	108,594
624	Sales occupations (retail) selling drugstore products, pharmaceuticals, medical supplies and healthcare goods	8,656	163,979
845	Driving, flying and sports instructors at educational institutions other than schools	9,113	122,790
842	Teachers for occupation-specific subjects at vocational schools and in-company instructors in vocational training	9,175	97,971
924	Occupations in editorial work and journalism	9,941	213,791
231	Technical occupations in paper-making and -processing and packaging	10,380	176,488
122	Occupations in floristry	10,489	139,359
932	Occupations in interior design, visual marketing, and interior decoration	11,200	147,306

Table 11 continued...

241	Occupations in metal-making	11,783	213,270
818	Occupations in pharmacy	12,491	180,538
245	Occupations in precision mechanics and tool making	12,604	299,508
825	Technical occupations in medicine, orthopaedic and rehabilitation	12,904	241,159
715	Occupations in human resources management and personnel service	12,949	124,624
731	Occupations in legal services, jurisdiction, and other officers of the court	13,083	164,694
631	Occupations in tourism and the sports (and fitness) industry	13,755	191,814
222	Occupations in colour coating and varnishing	14,008	179,903
723	Occupations in tax consultancy	14,193	245,535
844	Teachers at educational institutions other than schools (except driving, flying and sports instructors)	14,769	132,286
516	Management assistants in transport and logistics	15,130	239,075
843	Teachers and researcher at universities and colleges	15,548	303,146
234	Occupations in printing technology, print finishing, and book binding	15,811	318,150
232	Occupations in technical media design	17,426	271,166
282	Occupations in the production of clothing and other textile products	18,106	400,700
117	Occupations in forestry, hunting and landscape preservation	18,685	128,607
262	Technical occupations in energy technologies	18,933	200,060
272	Draftspersons, technical designers, and model makers	20,103	381,590
814	Occupations in human medicine and dentistry	20,299	232,594
945	Occupations in event technology, cinematography, and sound engineering	20,612	168,029
331	Floor layers	20,675	160,952
343	Occupations in building services and waste disposal	22,819	254,916
311	Occupations in construction scheduling and supervision, and architecture	26,379	342,979
341	Occupations in building services engineering	26,391	353,622
942	Actors, dancers, athletes and related occupations	26,841	253,592
841	Teachers in schools of general education	26,854	189,484
212	Conditioning and processing of natural stone and minerals, production of building materials	27,900	201,234
732	Occupations in public administration	28,235	377,345
413	Occupations in chemistry	28,235	521,959
721	Occupations in insurance and financial services	29,539	884,677
623	Sales occupations (retail) selling foodstuffs	31,231	302,481
722	Occupations in accounting, controlling and auditing	31,730	556,684
817	Occupations in non-medical therapy and alternative medicine	31,886	354,357
622	Sales occupations (retail trade) selling clothing, electronic devices, furniture, motor vehicles and other durables	32,546	355,842
632	Occupations in hotels	37,625	393,951
273	Technical occupations in production planning and scheduling	38,236	633,787
821	Occupations in geriatric care	39,400	409,794
823	Occupations in body care	40,979	657,375
431	Occupations in computer science	42,568	575,724
271	Occupations in technical research and development	43,260	593,950
612	Trading occupations	43,663	935,017
221	Occupations in plastic- and rubber-making and -processing	46,516	564,279

Table 11 continued...

111	Occupations in farming	48,258	316,228
525	Drivers and operators of construction and transportation vehicles and equipment	49,616	396,185
322	Occupations in civil engineering	51,935	433,222
611	Occupations in purchasing and sales	57,665	738,450
263	Occupations in electrical engineering	57,949	1,069,966
921	Occupations in advertising and marketing	59,381	450,172
832	Occupations in housekeeping and consumer counselling	61,436	636,524
333	Occupations in the interior construction and dry walling, insulation, carpentry, glazing, roller shutter and jalousie installation	63,813	522,652
342	Occupations in plumping, sanitation, heating, ventilating, and air conditioning	67,320	833,165
811	Doctors receptionists and assistants	68,470	1,213,193
292	Occupations in the production of foodstuffs, confectionery and tobacco products	75,211	967,994
233	Occupations in photography and photographic technology	76,780	90,776
713	Occupations in business organisation and strategy	87,772	1,230,525
813	Occupations in nursing, emergency medical services and obstetrics	98,193	1,503,772
531	Occupations in physical security, personal protection, fire protection and workplace safety	112,333	1,059,851
252	Technical occupations in the automotive, aeronautic, aerospace and ship building industries	115,268	1,727,693
244	Occupations in metal constructing and welding	124,952	1,177,706
242	Occupations in metalworking	129,373	1,400,857
332	Painters and varnishers, plasterers, occupations in the waterproofing of buildings, preservation of structures and wooden building components	130,860	869,263
251	Occupations in machine-building and -operating	143,800	1,932,378
121	Occupations in gardening	145,079	765,348
293	Cooking occupations	188,256	1,801,060
633	Gastronomy occupations	188,412	2,296,617
831	Occupations in education and social work, and pedagogic specialists in social care work	213,960	2,141,165
541	Occupations in cleaning services	271,847	4,006,181
521	Driver of vehicles in road traffic	287,743	2,760,959
321	Occupations in building constructions	359,606	2,538,077
621	Occupations in retail trade (without product specialisation)	378,624	5,299,499
513	Occupations in warehousing and logistics, in postal and other delivery services, and in cargo handling	487,495	4,805,261
714	Office clerks and secretaries	532,939	7,761,145

Source: own calculations

Table 12: TOP 1 prediction errors for RF for the 4-digit subsamples

sample no.	in-sample error	out-of-sample error
1	27.05 %	57.02 %
2	27.05 %	56.95 %
3	26.91 %	56.78 %
4	27.02 %	56.89 %
5	27.07 %	56.83 %
6	26.89 %	56.64 %
7	26.89 %	56.96 %
8	27.03 %	56.94 %
9	26.98 %	56.91 %
10	26.97 %	56.80 %
average	26.986 %	56.872 %

Source: own calculations

Table 13: TOP 3 prediction errors for RF for the 4-digit subsamples

sample no.	in-sample error	out-of-sample error
1	2.69 %	24.87 %
2	2.67 %	24.81 %
3	2.70 %	24.73 %
4	2.66 %	24.73 %
5	2.69 %	24.78 %
6	2.69 %	24.72 %
7	2.64 %	24.83 %
8	2.69 %	24.84 %
9	2.69 %	24.83 %
10	2.67 %	24.81 %
average	2.679 %	24.795 %

Source: own calculations

Imprint

IAB-Discussion Paper 03|2022

Publication Date
2 February 2022

Publisher
Institute for Employment Research
of the Federal Employment Agency
Regensburger Strasse 104
90478 Nürnberg
Germany

Download
https://doku.iab.de/discussionpapers/2022/dp0322.pdf

All publications in the series "IAB-Discusssion Paper" can be downloaded from
https://www.iab.de/en/publikationen/discussionpaper.aspx

Website
www.iab.de/en

ISSN
2195-2663

DOI
10.48720/IAB.DP.2203

Corresponding author
Sabrina Mühlbauer
Telefon +49 (911) 179 9743
E-Mail Sabrina.Muehlbauer@iab.de
Enzo Weber
Telefon +49 (911) 179 7643
E-Mail Enzo.Weber@iab.de