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

IAB-DISCUSSION PAPER

Articles on labour market issues

8|2022 Can Algorithms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic

Max Kunaschk, Julia Lang

ISSN 2195-2663

Can Algortihms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic

Max Kunaschk (University of Groningen, IAB)
Julia Lang (IAB)

Contents

Abstract

In this paper, we compare two popular statistical learning techniques, logistic regression and random forest, with respect to their ability to classify jobseekers by their likelihood to become long-term unemployed. We study the performance of the two methods before the COVID-19 pandemic as well as the impact of the pandemic and its associated containment measures on their prediction performance. Our results show that random forest consistently out-performs logistic regression in terms of prediction performance, both, before and after the beginning of the pandemic. During the lockdowns of the first wave, the number of unemployment entries and the fraction of individuals that become long-term unemployed strongly increases, and the prediction performance of both methods declines. Finally, while the composition of the (long-term) unemployed changed at the beginning of the COVID-19 pandemic, we do not find systematic patterns across groups with different levels of labor market attachment or across different sectors of previous employment in terms of declines in prediction performance.

Zusammenfassung

In diesem Beitrag vergleichen wir zwei gängige Machine Learning Methoden, Logistische Regression und Random Forest, im Hinblick darauf, wie geeignet sie sind um Arbeitssuchende nach ihrer Wahrscheinlichkeit, langzeitarbeitslos zu werden, zu klassifizieren. Wir untersuchen die Prognosegüte der beiden Methoden vor der COVID-19-Pandemie sowie die Auswirkungen der Pandemie und der damit verbundenen Eindämmungsmaßnahmen auf ihre Vorhersagekraft. Unsere Ergebnisse zeigen, dass Random Forest Modelle Langzeitarbeitslosigkeit besser vorhersagen können als logistische Regressionsmodelle, sowohl vor als auch nach Beginn der Pandemie. Während des Lockdowns in der ersten Welle der Pandemie nimmt sowohl der Anteil der Personen, die sich arbeitslos melden, als auch der Anteil der Personen, die langzeitarbeitslos werden, stark zu. Gleichzeitig nimmt die Prognosegüte beider Methoden ab. Obwohl sich die Zusammensetzung der (Langzeit-)Arbeitslosen zu Beginn der COVID-19-Pandemie geändert hat, finden sich keine systematischen Unterschiede in der Prognosegüte zwischen arbeitsmarktnäheren und -ferneren Personen oder zwischen Personen, die in unterschiedlichen Branchen tätig waren.

JEL classification

JEL Classification: C53, J64, J68, J71

Keywords

COVID-19, long-term unemployment, machine learning, profiling

Acknowledgements

We thank and Gerard van den Berg, Pascal Heß, Philipp Nickol, Regina Riphahn, and Gesine Stephan, as well as conference participants of the final network workshop of the DFG priority program 1764, the IAAE Annual Conference 2021 and VfS Annual Conference 2021 for many useful comments and suggestions. Special thanks go to the DIM department of the Institute for Employment Research (IAB) for their effort to provide the data. We gratefully acknowledge support of the German Research Foundation (DFG priority program 1764, grant number STE 1424/5-1).

1 Introduction

The spread of the COVID-19 virus and the measures taken to contain the pandemic have hit economies and labor markets around the world hard. In April 2020, the unemployment rate across the OECD increased by 3 percentage points in a matter of weeks and by the end of 2020, 114 million jobs had disappeared globally (OECD 2021). The impact on the labor market was most severe at the beginning of the pandemic because large parts of public life were suddenly shut down. Even though many measures to contain the pandemic are now more targeted, the medium- and long-term consequences of the pandemic for labor markets due to new virus variants and repeated waves of different magnitudes are still difficult to predict.

Such stark, unpredictable changes in labor market conditions could have a negative impact on the predictive power of statistical learning techniques that employment services in various countries around the world use to help caseworkers in their day-to-day work and to aid unemployed jobseekers in finding a new job.[1] In light of the COVID-19 pandemic that led to lockdowns and restrictions in various areas of life and subsequently to stark disruptions in labor markets, the question whether such severe changes in labor market conditions have a negative impact on the performance of machine learning algorithms is an important one to answer, as an increase in misclassification by statistical learning techniques may be costly – to employment services and to jobseekers alike.

In this paper, we investigate the performance – measured as the ROC-AUC Score[2] – and the stability of popular machine learning algorithms before and during the beginning of the COVID-19 pandemic with respect to their ability to predict long-term unemployment, defined as unemployment with a duration of more than six months, among newly registered jobseekers. For our analyses, we use rich German administrative labor market data covering the period from January 2011 until the end of March 2021. We start by comparing the performance of two popular statistical learning techniques, logistic regression and random forest, with respect to different variable input combinations. Next, we use yearly data from 2011 – 2021 to investigate the stability of the prediction performance of both methods during the years before and after the pandemic started. Then, we analyze monthly unemployment entries from January 2018 until September 2020 to investigate in more detail whether prediction performance deteriorates during the lockdowns associated with the first waves of the pandemic. Finally, we discuss the change in the performance of our models for different sub-groups of unemployment entries, differentiating between different industries and recent labor market histories.

We find that the prediction performance of both statistical learning methods is relatively stable during the years and months before the start of the pandemic. Furthermore, our results show

[1] The basic idea behind using statistical methods in this context is to classify individuals that become unemployed into different risk groups on the basis of observable characteristics, e.g., into those at high or low risk of long-term unemployment. For an overview of countries that currently use statistical learning techniques to classify jobseekers, see Desiere/Langenbucher/Struyven (2019).

[2] ROC-AUC stands for „Receiver Operating Characteristic – Area Under the Curve“. The ROC curve plots the fraction of positive outcomes correctly identified (TPR) against the fraction of negative outcomes incorrectly identified as positive (FPR) and shows how these measures change if the classification threshold is varied (see Figure A1 for an example of a ROC curve). The ROC-AUC Score measures overall model performance, potentially ranging from 0 – 1, with higher values indicating better prediction performance. For more details, see Section 4.

that random forest out-performs simple logistic regression, both, before and after the onset of the pandemic, and for all potential variable input combinations we explore.

At the beginning of the pandemic, the number of unemployment entries and the fraction of individuals that become long-term unemployed skyrocketed, increasing by roughly 29 percent and 59 percent at their respective peaks during the first lockdown in Germany. At the same time, the performance of the statistical learning techniques we use declines. The ROC-AUC Score of our best model decreases from 0.69 in 2018 to 0.65 in 2020 and the maximum fraction of individuals we can classify correctly drops by almost 10 percentage points – from just over 0.7 to just over 0.6. Overall, the models tend to perform worse as the fraction of long-term unemployed individuals increases. However, with respect to recent labor market histories and to the last industry a jobseeker had worked in before becoming unemployed, we do not find clear patterns regarding the change in long-term unemployment and the change in prediction performance.

Our paper contributes to the literature in several ways. First, we contribute to the literature that investigates the impact of the COVID-19 pandemic and the associated containment measures on the labor market (Brinca/Duarte/Faria-e-Castro 2021; Bradley/Ruggieri/Spencer 2021; Ke/Hsiao 2022). In particular, we add to other work on (long-term) unemployment in Germany (Bauer/Weber 2021) and around the world (Albanesi/Kim 2021; Mayhew/Anand 2020) by presenting new evidence on the development of unemployment entries and changes in long-term unemployment.

Second, closely related to the last point, our analyses complement other work on the distributional patterns of the labor market shocks caused by the pandemic and its associated containment measures (Brewer/Gardiner, 2020; Fernández-Reino/Sumption/Vargas-Silva 2020). Our analyses show that the composition of new unemployment entries and of long-term unemployed individuals changed considerably after the beginning of the pandemic.

Third, our paper contributes to the literature that focuses on the practical implications of the use of machine learning techniques in the context of (long-term) unemployment (Agrawal/Gans/Goldfarb 2019; Goller et al. 2020; Pope/Snydor 2011; Sansone/Zhu 2021). Specifically, our paper relates to the literature that aims to use statistical learning techniques in the context of the classification of jobseekers (Kern et al. 2021; Mühlbauer/Weber, 2022). Going beyond previous work, we present novel evidence comparing the performance and stability of different machine learning algorithms, both, in times of relatively calm labor markets and in times of a large shock – the COVID-19 pandemic.

The remainder of this paper is structured as follows. Section 2 briefly discusses the development of (long-term) unemployment during the beginning of the pandemic. Further, Section 2 gives an overview of the literature focused on using statistical learning techniques in the context of labor market policy. Section 3 gives an overview of the data we use for our analyses and presents some first descriptive analyses of unemployment entries and durations before and after the beginning of the pandemic. Section 4 discusses the methods we use to predict long-term unemployment and section 5 presents our main results. Section 6 concludes.

2 Related Literature

The Impact of the COVID-19 Pandemic on (Long-Term) Unemployment

The COVID-19 pandemic had a major impact on labor markets worldwide. Overall, the rise in unemployment was large in countries like the U.S. or Canada, where the number of unemployed persons increased rapidly at the beginning of the pandemic. In other countries, especially in those in which job retention schemes were used to retain workers currently not employable, the rise in unemployment was comparatively moderate (OECD 2021).

In Germany, the number of short-time workers reached unprecedented levels at the start of the pandemic (Gartner/Hutter/Weber 2021). However, despite the heavy use of short-time work, entries into unemployment have increased considerably – albeit to a lesser extent than in many other countries – especially in industries that were particularly strongly affected by the lockdowns. Bauer/Weber (2021) show that at the beginning of the pandemic, both, an increase in layoffs and a decrease in the job-finding rate caused unemployment to rise. Buch et al. (2021) find that in the months April to November 2020, jobs were lost primarily in sectors like the hotel, restaurant and tourism industries. In the trade, transport and hotel and restaurant sector, the number of employed persons decreased by almost 400,000 during the first 12 months of the COVID-19 pandemic (Gartner/Hutter/Weber 2021). The business service sector, other service producers and the manufacturing industry also suffered large employment losses (Gartner/Hutter/Weber 2021).

Different groups of employees may be particularly strongly affected by unemployment than before the pandemic. Similarly, the COVID-19 pandemic may affect other groups of workers than during the Great Recession.[3] Furthermore, the risk of becoming long-term unemployed has increased as a result of the pandemic, as many employees that lost their jobs had difficulties finding a new job. At the end of 2020, the number of people unemployed for at least six months in OECD countries had increased by 60 percent compared to one year before (OECD 2021). Similarly, in Germany, the number of persons remaining in unemployment for more than 12 months had increased by 325,000 one year after the beginning of the pandemic (Gartner/Hutter/Weber 2021). This shift toward a higher share of the unemployed experiencing long-term unemployment, combined with a potentially changing composition of the (long-term) unemployed, may have a considerable impact on the predictive quality of statistical models designed to predict those outcomes.

Statistical Learning and Public Employment Services

While statistical profiling has not yet been used in practice in Germany[4], employment services in many countries around the world have started adopting different statistical learning techniques to aid caseworkers in their day-to-day work assigning jobseekers to different policy measures based on their probability to reach a certain duration of unemployment or to exhaust benefits. In some countries, such as Australia, Austria, the Netherlands, and the US, employment services use

[3] Except for manufacturing, which was also hit hard by the Great Recession.

[4] Instead, soft profiling is carried out by the caseworkers, who classify newly unemployed persons into two categories (close to the labor market and not close to the labor market).

more traditional econometric techniques, such as logistic or probit regression, to classify jobseekers, while employment agencies in countries such as Denmark, Belgium (Flanders), and New Zealand use machine learning techniques to categorize jobseekers (Desiere/Langenbucher/Struyven 2019).

As different statistical models may offer different benefits and drawbacks, policymakers depend on evaluations of different approaches to determine which models are suited for a specific application. However, although employment services in several countries use statistical methods to classify unemployed workers, formal evaluations comparing the performance of different classification methods with respect to their unemployment prediction quality are rare.[5]

A notable exception is a recent working paper by Kern et al. (2021), which compares the predictive performance of different statistical learning techniques using a similar data set to the one we use in this paper. The authors find that machine learning techniques, such as random forest, penalized logistic regression, and gradient boosting, outperform simple unpenalized logistic regression. Our paper adds to their work by focusing more closely on the prediction performance of different techniques using various sets of variables as inputs. Further, we investigate the stability of the prediction models over time and examine how the labor market shock during the initial waves of the COVID-19 pandemic affected the prediction performance of different statistical learning techniques.[6]

3 Data and Descriptive Statistics

For this study, we use a 2- percent random sample of the Integrated Employment Biographies (IEB v.16.00.01; for more information on the data, see, for example, Dorner et al., 2010) of the German Federal Employment Agency. The data provide comprehensive information about employed individuals (excluding self-employment and civil service) and registered unemployed persons in Germany. Our data cover the period from January 2011 – March 2021. Therefore, we can analyze the stability of the predictive performance of our statistical learning techniques for a several years before the start of the COVID-19 pandemic and during roughly the first year of the pandemic.

As caseworkers in German employment agencies classify unemployed persons according to whether they are likely to find a job again within six months (persons close to the labor market) or not (persons not close to the labor market), we use a dummy indicating whether an individual stays unemployed for six months or not as our main target variable (i.e., long-term unemployed, "LTU").[7] As we need an outcome period of at least six months in order to determine whether an

[5] A related strand of literature uses statistical learning in the context of government employment services to identify heterogeneous causal effects of active labor market programs (Knaus/Lechner/Strittmatter 2020, Cockx/Lechner/Bollens, 2020), or on comparing the effectiveness of statistical techniques with caseworker performance (Lechner/Smith, 2007).

[6] A recent working paper by Mühlbauer/Weber (2022) also uses a similar data set as we do to classify jobseekers with respect to potentially suitable jobs. They show that random forest models tend to perform better than simple OLS models.

[7] Note that the legal definition of long-term unemployment in Germany is unemployment of more than twelve months. In a robustness check, we additionally use more than 12 months of unemployment as definition of the outcome instead of 6 months. The results confirm our main findings (see Appendix B). The predictive performance for longer unemployment of 12 months is consistently better than for 6 months and is very similar to the results of Kern et al. (2021), who define long-term unemployment as unemployment lasting more than 12 months.

individual becomes long-term unemployed, we can consider entries into unemployment up until September 2020. For our definition of long-term unemployment, we count all continuous spells of registered unemployment and active labor market program participation up to six weeks. In addition, we count two unemployment spells of the same individual with up to six weeks interruptions as a continuing unemployment spell.

The IEB also contain detailed information on sociodemographic characteristics as well as on individuals` complete employment history, unemployment episodes and episodes with participation in active labor market programs on a daily basis. Thus, we are able to construct a rich set of variables that we use for our prediction exercises, including information on standard sociodemographic characteristics, such as age, gender, and education, information on the recent and on the long-term labor market history (up to 7 years back), and information on the last job held by the individual. In addition, we use the IEB to construct a series of regional control variables, such as average wages and the composition of the workforce, for the district ("Kreis") the unemployed individual lives in. Finally, we collected additional regional information on unemployment rates from the official Statistics of the Federal Employment Agency (Statistik der Bundesagentur für Arbeit 2021).[8]

Table 1 and Table 2 give an overview of a selection of the sociodemographic characteristics and the short-term labor market histories for the full sample (Table 1) and for a subsample of individuals who become long-term unemployed (Table 2). The first column shows the descriptive statistics for our entire sample, columns 2 – 4 show the descriptive statistics for the years 2016, 2018 and 2020 and columns 5 and 6 show the difference of the means between 2016 and 2018 as well as between 2018 and 2020.

[8] For a list of all available variables, see Table A1 in the Appendix and the description in section 5.A.

Table 1: Descriptive Statistics – Full Sample

	(1) 2011 – 2020	(2) 2016	(3) 2018	(4) 2020	(5) Δ2016/2018	(6) Δ2018/2020
UE Duration > 6 Months	0.323	0.301	0.287	0.405	-0.014***	0.118***
Male	0.564	0.576	0.569	0.558	-0.007***	-0.011***
Age	37.85	37.50	37.91	38.27	0.402***	0.364***
German	0.762	0.722	0.686	0.693	-0.037***	0.008***
Vocational Degree	0.458	0.441	0.411	0.402	0.030***	0.009***
Academic Degree	0.060	0.065	0.075	0.085	0.010***	0.010***
Tot. Earnings Last Year	9842.06	9539.93	10286.50	12512.93	746.57***	2226.43***
Tot. Days Employed Last Year	177.08	168.41	169.67	189.80	1.26*	20.12***
Tot. Days Unemployed Last Year	59.13	55.63	55.79	45.86	0.15	-9.93***
Last Wage	43.37	41.70	44.36	50.16	2.66***	5.80***
N	908,844	92,319	86,340	65,583		

Note: This table shows the descriptive statistics for the main sample in the years 2016, 2018, and 2020. For the year 2020, we only include unemployment entries up until September. */**/*** indicate significant differences of mean values between 2016/2018 or 2018/2020 at the 10 percent/5 percent/1 percent level.

Source: IEB 16.00.01.© IAB

Table 2: Descriptive Statistics – Long-term Unemployed

	(1) 2011 – 2020	(2) 2016	(3) 2018	(4) 2020	(5) Δ2016/2018	(6) Δ2018/2020
Male	0.550	0.559	0.559	0.564	-0.000	0.005
Age	40.82	40.54	41.10	40.62	0.566***	-0.488***
German	0.764	0.730	0.709	0.681	-0.021***	-0.029***
Vocational Degree	0.424	0.405	0.388	0.378	-0.017***	-0.009**
Academic Degree	0.053	0.053	0.063	0.078	0.010***	0.016***
Tot. Earnings Last Year	8171.55	7831.47	8401.76	11333.02	570.29***	2931.261***
Tot. Days Employed Last Year	152.44	141.99	143.75	173.21	1.76	29.45***
Tot. Days Unemployed Last Year	74.14	69.69	68.94	53.64	0.75	-15.30***
Last Wage	41.10	39.45	41.98	48.39	2.53***	6.41***
N	292,215	27,824	24,804	26,565		

Note: This table shows the descriptive statistics for the main sample in the years 2016, 2018, and 2020. For the year 2020, we only include unemployment entries up until September. */**/*** indicate significant differences of mean values between 2016/2018 or 2018/2020 at the 10 percent/5 percent/1 percent level.

Source: IEB 16.00.01.© IAB

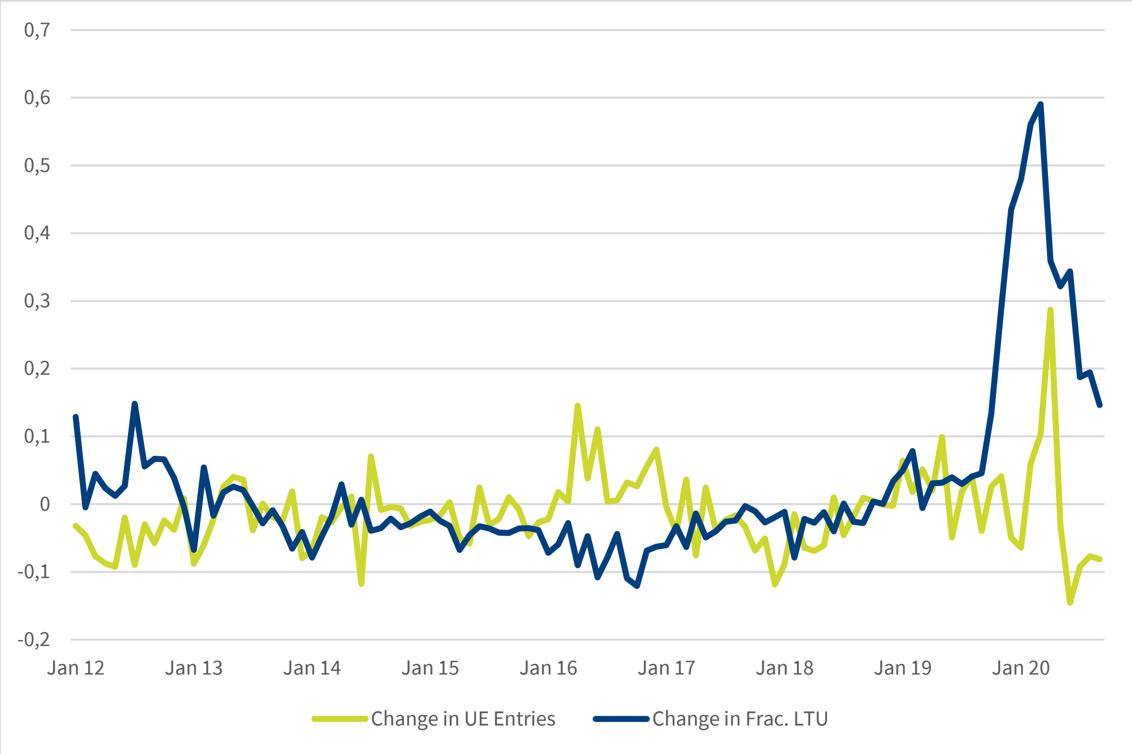
Comparing 2016 and 2018, we see a small decrease of 1.4 percentage points in the fraction of individuals that remain unemployed for at least 6 months. The averages of most other variables are also relatively similar across the two years. Comparing 2018 and 2020, we see a much larger difference in the fraction of individuals that become long-term unemployed: The fraction of individuals remaining unemployed for at least 6 months increases from just under 29 percent to over 40 percent.

In terms of sociodemographic characteristics, we see significant differences between the years due to the large number of cases, but the differences are rather insignificant in economic terms. While the sociodemographic characteristics are relatively similar in 2016, 2018, and 2020, we see stark differences with respect to the short-term labor market history of new unemployment entries between 2018 and 2020, which is not the case if we compare 2016 with 2018. Individuals becoming unemployed during the first year of the pandemic exhibit much better short-term labor market histories, on average: Individuals entering unemployment in 2020 had higher previous daily wages, spent more time in employment and less time in unemployment than individuals entering in 2018. Thus, the average fraction of individuals that ends up long-term unemployed increased despite a shift towards more positively selected unemployment entries in 2020. If this positively selected group of workers was able to find a new job relatively quickly, this would imply a smaller change in the composition of the long-term unemployed compared to all unemployment entries.

However, Table 2 shows that the group that becomes long-term unemployed is also positively selected after the start of the pandemic. In 2020, long-term unemployed individuals have more recent work experience and higher wages and earnings compared to the pre-pandemic years. Thus, even individuals with a higher labor market attachment had a harder time finding a new job (quickly) during the first year of the pandemic.

Finally, before we turn to our main analyses, we present some general trends in unemployment over the years. Figure 1 shows the change in the number of unemployment entries from 2012 until 2020 and the change in the fraction of individuals that remain unemployed for at least six months after entering unemployment in the respective month compared to the same month of the previous year.

Figure 1: Relative change in the number of unemployment entries and fraction long-term unemployed



Note: This figure shows the change (in percent) in the number of unemployment entries ("UE Entries") and the fraction of individuals that remain unemployed for at least 6 months after entering unemployment ("LTU") relative to the number of UE Entries / the fraction LTU during the same month in the previous year. The year 2020 only includes unemployment entries up until September.

Source: IEB 16.00.01, 2011 – 2020.

From 2012 until late 2019, we do not see major changes in the number of unemployment entries or in the fraction of individuals becoming long-term unemployed. Then, beginning in September 2019, the fraction of individuals that enter unemployment and remains unemployed for at least six months begins to steadily increase, reaching its peak in March 2020.[9] Relative to March 2019, the fraction of individuals that enters unemployment and remains unemployed for at least six months increased by 60 percent in March 2020. The relative changes in the number of unemployment entries occur later and are less pronounced than the changes in the fraction of long-term unemployed individuals. Nevertheless, the shock in the number of unemployment entries was considerable, as the number of unemployment entries was roughly 29 percent higher in April 2020 compared to April 2019.

[9] The first lockdowns in Germany came into effect in late March 2020. Thus, the six-month outcome period of unemployment spells that begin from September 2019 on potentially falls into this first lockdown period. The closer to the actual lockdown, the larger the fraction of individuals that remain unemployed for six months or more becomes.

4 Methodology

We define our outcome variable, long-term unemployment, as a binary variable. Consequently, we treat profiling as a classification problem.[10] As it is one of the most accurate general-purpose techniques available (Biau 2012), we use a random forest classifier as our main classification algorithm (see, Breiman 2001).

Random forest classifiers are based on a collection of tree classifiers that each cast a vote for the most popular class (Breiman 2001). The goal of a tree classifier is to grow a decision tree by recursive binary splitting. In each step, the classification algorithm chooses the variables and the split point to achieve the best fit. The most common criterion used for splitting nodes and pruning the tree is the Gini index, which indicates how mixed the classes are in the two groups created by a split. Then one or both of these groups are split into two more groups. This procedure continues until a stopping rule is applied (Hastie/ Tibshirani/Friedman 2011). Based on the majority vote, the classifier predicts a positive or a negative outcome. The individual trees are based on different random subsamples of the data and only a random subset of variables is used for each tree (Athey/Imbens 2019). Essentially, a random forest can be interpreted as an average of many separate tree classifiers that have all been estimated on a subsample of the data (Athey 2017).[11]

In addition to the random forest classifier, we use a more "traditional" econometric technique – logistic regression – to classify individuals according to their predicted long-term unemployment risk. The main reason for this choice is that logistic regression is one of the classifiers most commonly used in countries that use statistical learning to classify jobseekers into different risk groups (see, Desiere/Langenbucher/Struyven 2019).

For supervised machine learning methods, such as the random forest classifier we use, the data is usually divided into training and test data sets. The training data is split into different folds and in each step all but one of the folds are used to estimate the model. The estimated parameters are used to predict the outcome of interest in the remaining fold. Finally, the parameter with the best average performance across all cross-validation steps is chosen. The final model is then evaluated using the test data, which has not been used to fit the model (Athey/Imbens 2017).

In our application, we use a different sample splitting approach: we use data from one year to estimate our models in a later year.[12] The approach mimics a real-world application, in which an employment agency only has past data available to classify individuals currently registering as unemployed.

[10] Our binary outcome variable is not equally distributed in the data but imbalanced. Only roughly 29-41 percent of all observations belong to the minority class (see Table 1) of long-term unemployed individuals. Such a class imbalance can lead to difficulties classifying positive and negative instances correctly and classifiers may lose their classification ability (Galar et al., 2012). Therefore, we account for the class imbalance in our data by applying class weighted learning, which up-samples observations of the minority class, i.e., the long-term unemployed, during the training process to achieve a balanced class distribution.

[11] We use the Python module scikit-learn, version 0.24.1 (Pedregosa et al. 2011) for all analyses. Rather than letting each classifier vote for a class individually, the scikit-learn implementation of the random forest classifier averages probabilistic predictions of the individual classifiers.

[12] We use five-fold cross-validation to tune the hyper-parameters of the random forest classifier in order to maximize our main performance measure, the ROC-AUC Score (for a detailed explanation of the ROC-AUC Score, see the end of this section). However, this makes little difference in terms of predictive performance in practice.

As a first step, before we turn to our main analyses of the impact of the pandemic on the prediction performance of the classification models, we investigate the performance of the two classifiers with respect to different variable inputs. For this exercise, we use data from 2016 to predict long-term unemployment in 2018 and compare the predictive quality of logistic regression and random forest models regarding different variable inputs.[13] We start by comparing very simple variable inputs that include either only sociodemographic characteristics or only information on the short-term labor market history of the unemployed individual. We then iteratively add more variables to the models and compare how performance develops with a growing number of predictors.

We then proceed to our analyses of the stability of the random forest and the logistic regression classifiers over time. For this exercise, we use the years 2011 – 2020 separately to predict long-term unemployment in year T using data from year T-2 using the preferred variable input identified in the preceding section. Finally, we use data from January 2017 to March 2021 for our monthly analyses, this time using data from the same month in year T-1 to predict long-term unemployment in the respective month in year T.

Finally, we need a measure to compare the performance of the different classification methods. A common performance measure to evaluate classification models is accuracy, which measures the share of correctly classified observations:

$$Accuracy = (TP + TN)/N$$

where TP is the number of correctly positive classified observations, TN is the number of correctly negative classified observations, and N is the total number of observations.

However, if classes are imbalanced, accuracy can be a spurious performance criterion, as simply classifying everyone as the majority class can lead to misleadingly high accuracy rates (Galar et al., 2012). In our application, e.g., in 2018, only around 29 percent of the observations become long-term unemployed. If we were to classify all observations as non-long-term unemployed, we would achieve an accuracy of 71 percent, but we would not identify a single person becoming long-term unemployed correctly. Consequently, relying on accuracy as a performance measure can be misleading.[14]

Therefore, we mainly focus on the ROC-AUC Score to identify our best-performing model. The ROC-AUC Score gives the probability that the classifier correctly identifies two randomly drawn observations, one from the positive and one from the negative class. A ROC-AUC Score of 0.5 is as good as a random guess, whereas a ROC-AUC Score of 1.0 indicates perfect prediction. The ROC-AUC Score is generally preferable to accuracy, as it is independent of the decision threshold and invariant to the a priori probability distribution (Huang/Ling 2005).

[13] We need a two-year lag, as the outcome period of individuals becoming unemployed in the second half of each year spans into the following year.

[14] Furthermore, accuracy depends on threshold chosen to classify observations as positive / negative. In our application, there is quite some variation in accuracy across classification thresholds. We discuss accuracy over classification thresholds in more detail in section 5.B.

5 Results

Variable Input

We start our analyses by comparing which method, random forest or logistic regression, and which variable input combinations achieve the best prediction performance. For this exercise, we use 2016 data to predict long-term unemployment in 2018.

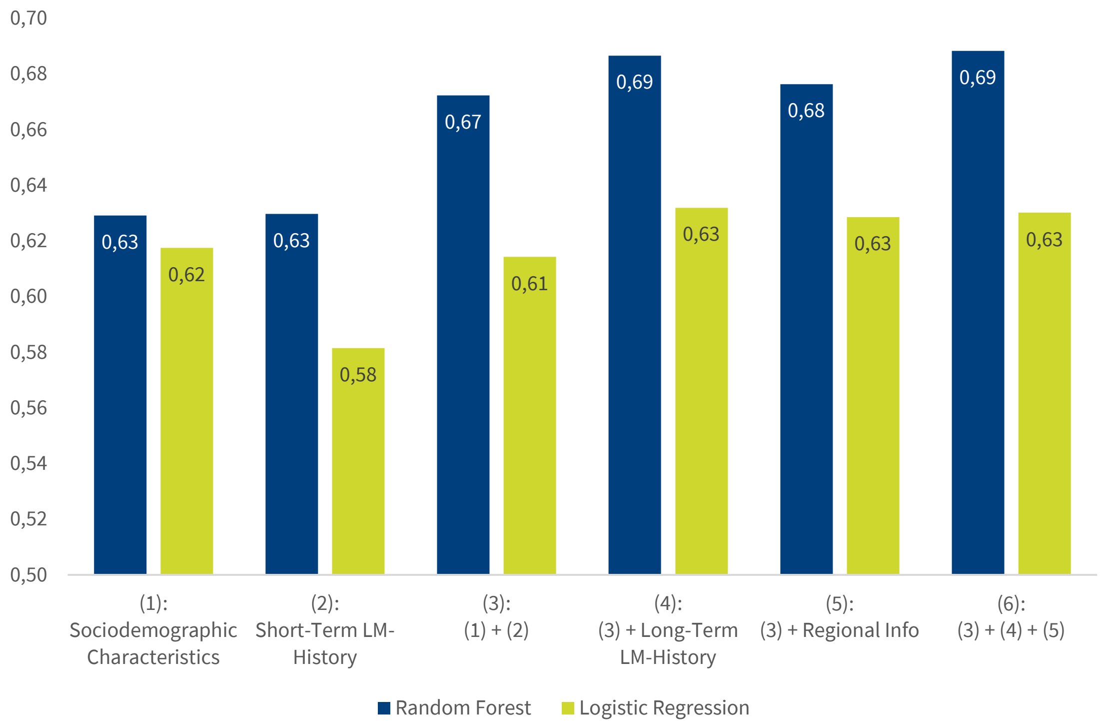
The first specification for our variable input analyses only includes basic sociodemographic characteristics: age, gender, highest schooling and vocational degree, and German nationality. The second specification only includes the short-term labor market histories of the unemployed individuals (e.g., days in (un)employment, income during the last year before entering unemployment, occupation of the last job and corresponding wages). The third specification combines the first two variable sets. The fourth specification adds the long-term labor market history to specification three, while the fifth specification adds regional labor market information to specification three. Finally, the sixth specification includes all available variables.[15]

Comparing the pre-pandemic random forest models to the logistic regression models in Figure 2, we see that the random forest models out-perform the logistic regression models for each possible variable combination. Logistic regression achieves a ROC-AUC Score of 0.58 – 0.63, while the random forest models achieve scores of 0.63 – 0.69.

Taking a closer look at specifications (1) and (2), we see that the random forest models perform very similar, irrespective of whether we include sociodemographic characteristics or the short-term labor market history. The logistic regression models tend to perform better when we use basic sociodemographic characteristics rather than the short-term labor market history. Combining both sets of variables (specification (3)) further increases the predictive power of the random forest models while the performance remains similar to specification (1) when we add (1) + (2) together in the case of the logistic regression. Adding long-term labor market histories up to 7 years back (specification (4)) or regional information (specification (5)) improves both, the random forest and the logistic regression model, where the improvement is more pronounced for specification (4). Finally, specification (6) shows that adding regional labor market information to specification (4) does not further increase the ROC-AUC Score.

[15] Table A1 in the Appendix contains a detailed list of all potential variables.

Figure 2: Out-of-sample ROC-AUC Scores predicting the probability to become LTU (2018)



Note: This figure shows the out-of-sample ROC-AUC Scores for different specifications (see main text for details) of logistic regression and random forest models predicting the probability to remain unemployed for at least 6 months using unemployment entries from 2018. The models were trained using data from 2016.

Source: IEB 16.00.01.

As specification (4) achieves a similar prediction performance as specifications (5) and (6), we opt for specification (4) for the remaining analyses.[16] This has the benefit of reduced computation time, especially for the random forest classifier. Furthermore, specifications (5) and (6) are the only specifications that include information that we derived from an external source (official Statistics of the Federal Employment Agency), while all the other specifications rely exclusively on information we calculated based on our administrative labor market data.

Prediction Stability Over Time

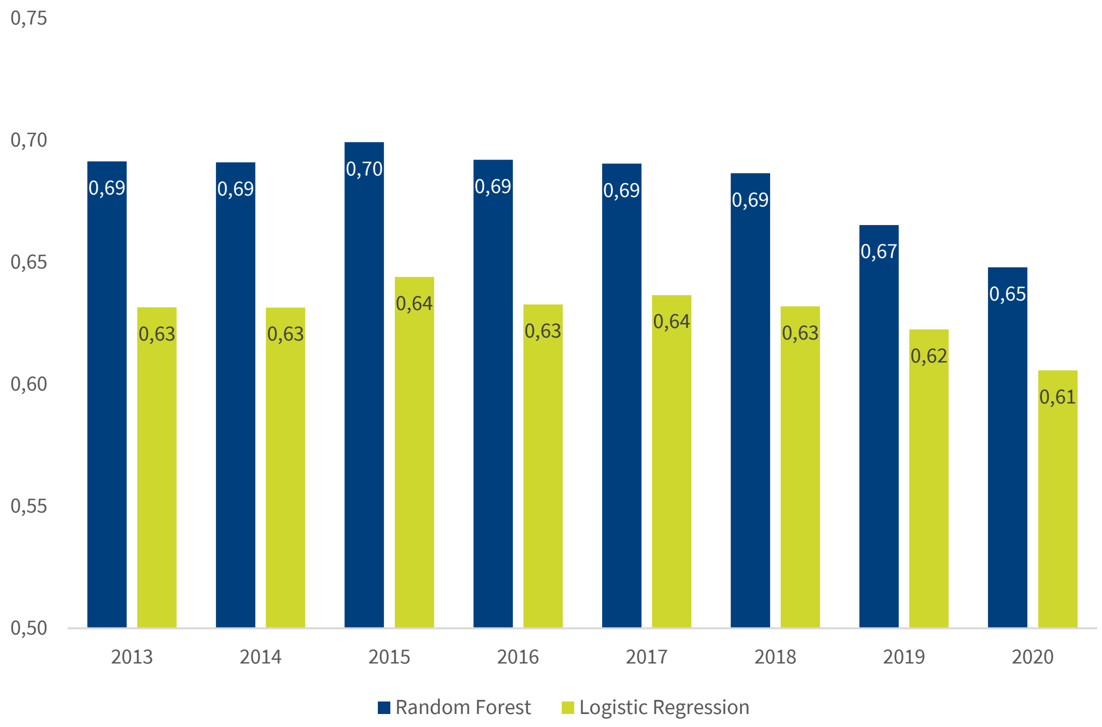
In this section, we analyze the stability of our models over time. We begin by investigating model performance during the years before the COVID-19 pandemic and during the first year after the pandemic had started. We then proceed to investigate model performance more closely, on a monthly level, for the years 2018 – 2020.

Figure 3 compares the yearly ROC-AUC Scores achieved by random forest and by logistic regression models using our main specification. During the pre-pandemic years, the performance of both methods is very stable, albeit at different levels: the ROC-AUC-Score for the random forest model ranges from 0.69 to 0.70 and the ROC-AUC-Score for the logistic regression model ranges from 0.63 – 0.64. In 2019, the ROC-AUC-Scores for both methods begin to decline. A potential reason for this is that the outcome period for individuals that enter unemployment

[16] The same ranking with respect prediction performance of the models holds for the period after the onset of the pandemic (see Figure A2 in the Appendix).

during the last quarter of 2019 already overlaps with the first lockdown that began in March 2020. The decline in model performance continues in 2020, where both methods achieve their lowest scores during the entire observation period, with 0.65 for the random forest model and 0.61 for the logistic regression model.

Figure 3: Yearly out-of-sample ROC-AUC Scores predicting the probability to become LTU



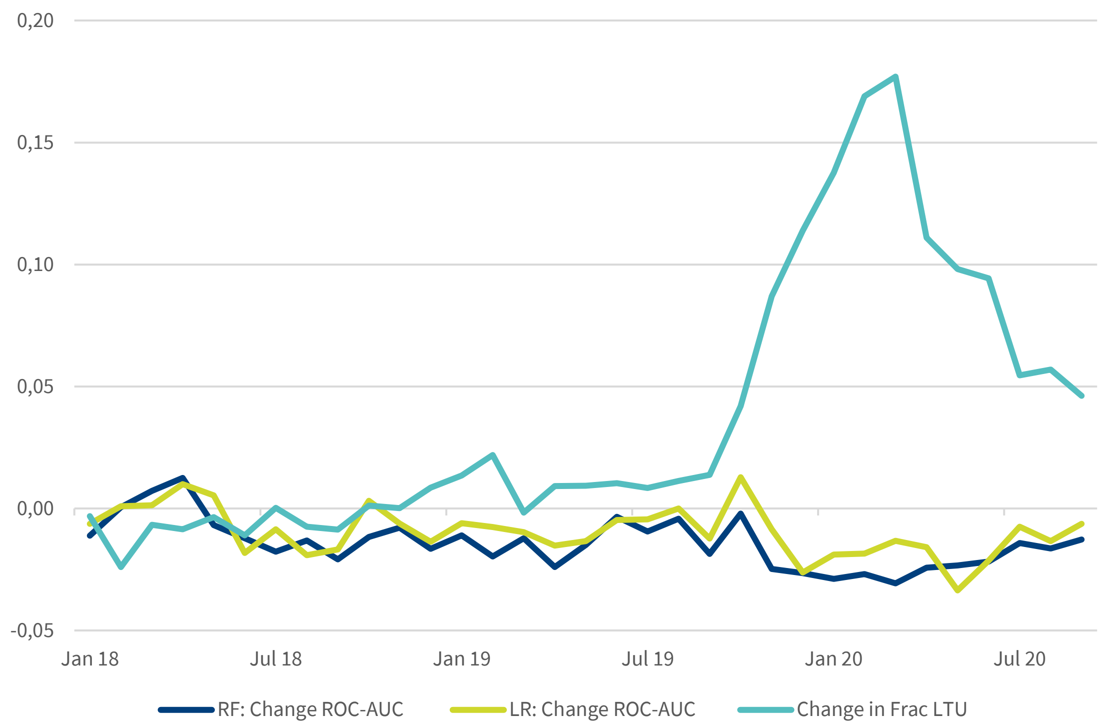
Note: This figure shows the yearly out-of-sample ROC-AUC Scores for the main specification of the random forest model and the logistic regression model predicting the probability to remain unemployed for at least 6 months. Each model was trained using data from year T-2 to predict the probability to become LTU in year T. The predictions in 2020 only include unemployment entries up until September.

Source: IEB 16.00.01, 2011 – 2021.

To investigate changes in model performance during the beginning months of the COVID-19 pandemic and the associated lockdowns in more detail, we next move to our monthly analyses. Figure 4 shows the change in the fraction of individuals that become long-term unemployed and the ROC-AUC Score, relative to previous years' values, on a monthly basis.[17]

[17] For all monthly out-of-sample ROC-AUC Scores from 2012 – 2020, see Figure A3 in the Appendix.

Figure 4: Change in fraction of individuals remaining unemployed for min. six months and monthly out-of-sample ROC-AUC Scores for the main random forest and logit models



Note: This figure shows the monthly change in the fraction of unemployment entries that become long-term unemployed and the change in the monthly out-of-sample ROC-AUC Scores for the main specification of the random forest and logistic regression model, relative to the same month in year T-1, in percentage points.

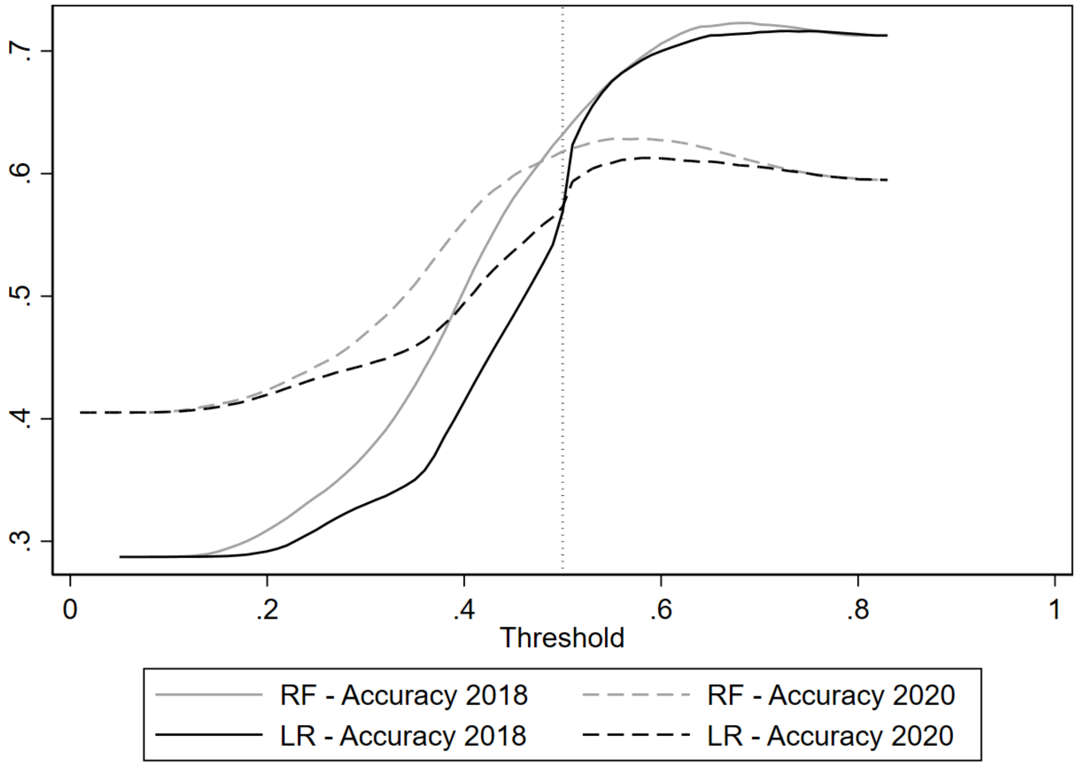
Source: IEB 16.00.01, 2017 – 2021.

In 2018 and during the first half of 2019, the fraction of individuals becoming long-term unemployed and the ROC-AUC Scores of both classifiers do not exhibit substantial changes. At the same time, we observe a negative correlation between the fraction of individuals that become long-term unemployed and the ROC-AUC Scores: when the fraction increases, model performance tends to decrease and vice versa.

From September 2019 on, the fraction of individuals that remained unemployed for at least six months increases continually, reaching its peak in March 2020, when the first lockdown in Germany came into effect. At this time, the fraction of individuals becoming long-term unemployed had increased by almost 18 percentage points. Similar to the pre-pandemic period, prediction performance of both classifiers declines as the fraction of individuals that become long-term unemployed increases. The ROC-AUC Scores decline by more than 3 percentage points, reaching their respective lowest points in March 2020 (random forest) and in May 2020 (logistic regression).

To further illustrate the classification quality of our models, we show the accuracy based on yearly predictions for 2018 and 2020 in addition to the ROC-AUC Scores. As discussed above, accuracy indicates which fraction of individuals we correctly classify as (not) becoming long-term unemployed. Since this performance measure depends on the threshold above which we classify an individual as long-term unemployed, Figure 5: Accuracy for different thresholds for the years 2018 and 2020Figure 5 shows accuracy across different relevant thresholds for the best random forest (RF) and logistic regression (LR) model.

Figure 5: Accuracy for different thresholds for the years 2018 and 2020



Note: This figure shows the out-of-sample accuracy for the main specification of the random forest and logistic regression model for the years 2018 and 2020 over the relevant range of different threshold values.
Source: IEB 16.00.01.

As the share of long-term unemployed is higher in 2020, we classify more people correctly at lower thresholds (i.e., individuals are classified as long-term unemployed even with lower probabilities of becoming long-term unemployed) than in 2018, both in the random forest and logit model. For higher thresholds, it is the other way around and accuracy is higher for individuals entering unemployment in 2018. For a default threshold of 0.5, accuracy in 2018 is slightly higher than for 2020 in the random forest model, whereas for the logistic regression model accuracy hardly changes between 2018 and 2020.

Overall, considering all relevant thresholds, a higher maximum accuracy can be realized for 2018 than for 2020. For the random forest model, the maximum accuracy in 2018 is 72.3 percent (for a threshold of 0.69), while we can correctly classify a maximum of 62.8 percent of the individuals with our model in 2020 (for a threshold of 0.55). Thus, when we compare the maximum number of individuals we can classify correctly within the best random forest model (in terms of maximum accuracy), our model performs worse in 2020, with the proportion of correctly classified individuals falling by almost 10 percentage points.

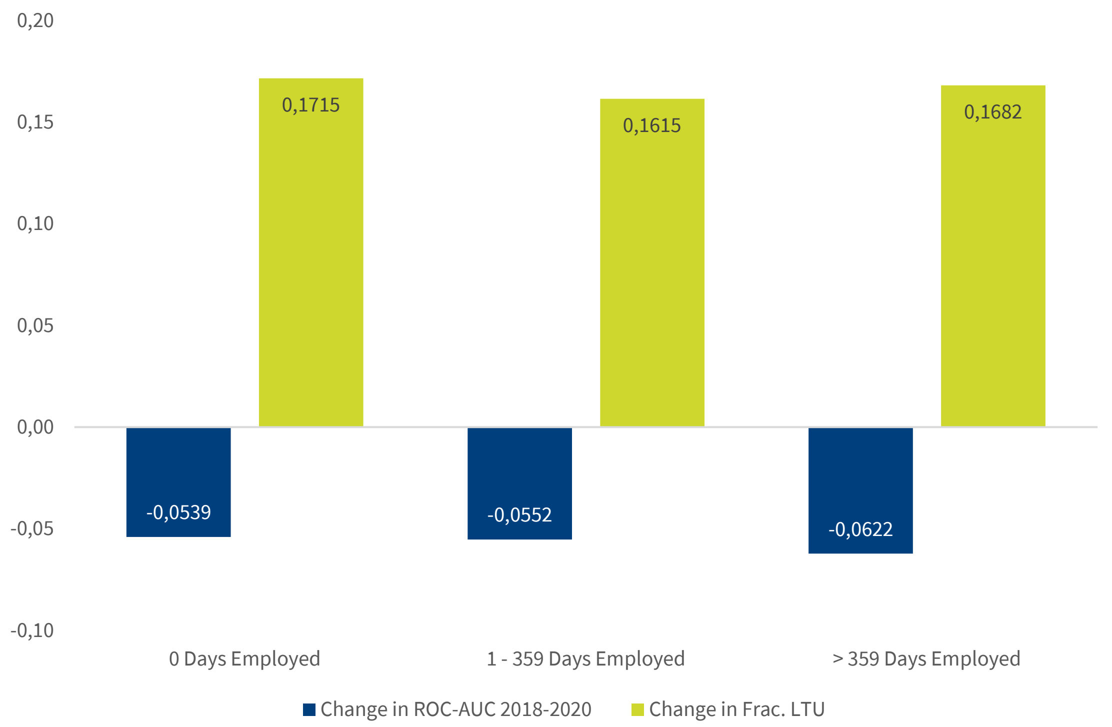
In line with our results on the ROC-AUC Score, we achieve worse performance with the logit model also in terms of accuracy across the entire threshold distribution. The highest accuracy of 61.3 percent can be reached for a threshold of 0.59 in 2020, which is, similar to the random forest model, over 10 percentage points lower than the maximum accuracy in 2018 (71.6 percent for a threshold of 0.73). A potential reason for this deterioration in performance may be that the proportion of jobseekers becoming long-term unemployed increased sharply due to the pandemic. As we estimated the model using data from a period before the pandemic, the distribution of the outcome variable differs in the training and test data.

In addition to a change in the distribution of the outcome, a change in the distribution of the characteristics of jobseekers may also be a cause of the deterioration in model performance. In our descriptive analyses (Table 1 and Table 2), we saw that, during the pandemic, even individuals with a higher labor market attachment were more likely to become (long-term) unemployed than before. If the composition of the unemployed for whom we make the predictions differs from the composition of the individuals we use to estimate the model, the performance of the model may suffer. For certain groups of people that were already affected by (long-term) unemployment to a similar extent before, we would expect performance to deteriorate less than for groups for which there was a sharp increase in (long-term) unemployment and who were hardly affected by (long-term) unemployment before the pandemic.

To examine this aspect in more detail, we calculate the change in the share of long-term unemployed and the change in performance between 2018 and 2020 for specific subgroups. For this analysis, we use pooled data for unemployment entries between February and April 2018 and 2020 because this is the period when the increase in long-term unemployment peaked during the pandemic (cf. Figure 4).

First, we distinguish between individuals with better and worse recent labor market histories by constructing three groups according to the days in employment during the last year before entry into unemployment. Figure 6 shows the differences in the fraction of individuals becoming long-term unemployed and the differences in the ROC-AUC Scores for these three groups between February to April 2018 and February to April 2020.

Figure 6: Change in the fraction remaining unemployed for at least 6 months and change in ROC-AUC Score by recent employment history, 2018 – 2020

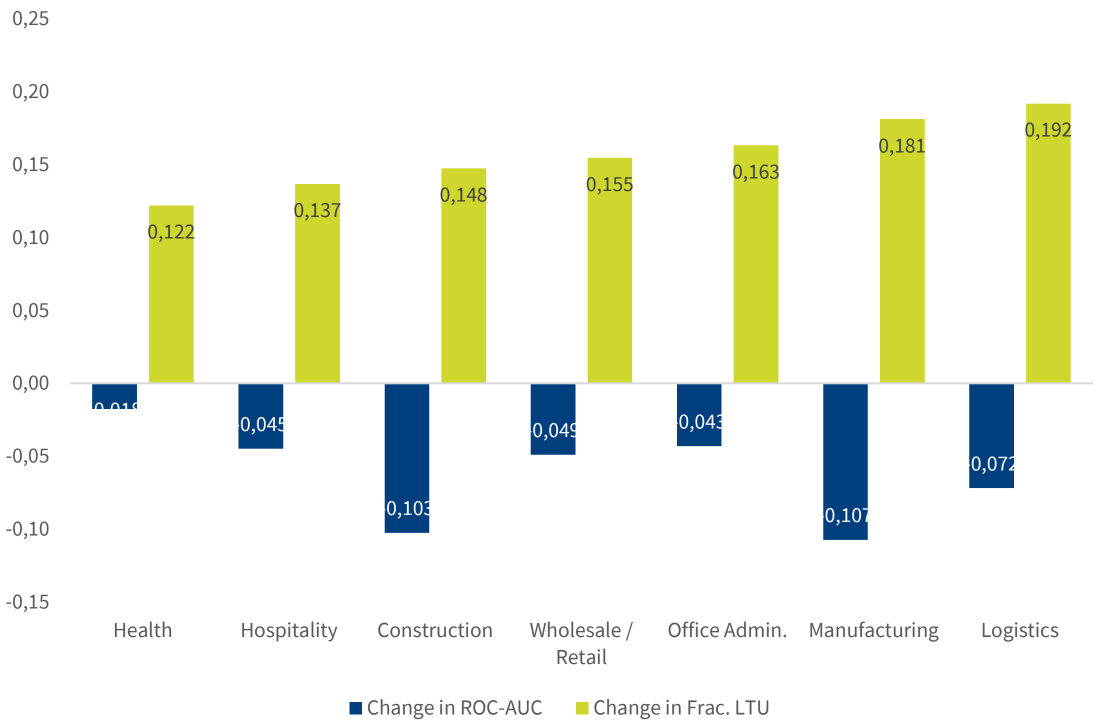


Note: This figure shows, by recent employment history, the percentage point change in the number of unemployment entries and in the out-of-sample ROC-AUC Score from our preferred specification of the random forest model between 2018 and 2020 (see section 5.A). The ROC-AUC Scores were calculated using data from 2016 and 2018 to estimate the models to predict long-term unemployment in 2018 and 2020 respectively. For each year, we pool observations from February – April.
Source: IEB 16.00.01.

Since the descriptive analyses have shown that, in 2020, even individuals with better labor market histories were more likely to become (long-term) unemployed than in 2018, we would expect the increase in long-term unemployment and the decrease in performance to be stronger for this group than for individuals with lower labor market attachment and who were also already more likely to experience long-term unemployment in 2018. However, Figure 6 shows hardly any differences with respect to the changes in performance across the groups. The change in the fraction of jobseekers becoming long-term unemployed varies between 16.2 and 17.2 percentage points and the ROC-AUC Scores decrease by 5.4 to 6.2 percentage points. Hence, there was a large increase in long-term unemployment for all groups, regardless of how severely the group was already affected by long-term unemployment before the pandemic began. Accordingly, the decline in performance is similar for all groups.

The pandemic affected different industries to varying degrees (see Section 2A). If the composition in terms of prior occupations and industries of the (long-term) unemployed has changed to a large degree, this could also affect the predictive performance of our models. Therefore, in Figure 7, we additionally investigate the changes in ROC-AUC Scores and the change in the share of long-term unemployed for selected industries.[18] The analysis refers to the last industry in which a person worked before entering unemployment.

Figure 7: Change in the fraction remaining unemployed for at least 6 months and change in ROC-AUC Score by last sector of employment, 2018 – 2020



Note: This figure shows, by last sector of employment, the percentage point change in the number of unemployment entries and in the out-of-sample ROC-AUC Score from our preferred specification of the random forest model between 2018 and 2020 (see section 5.A). The ROC-AUC Scores were calculated using data from 2016 and 2018 to estimate the models to predict long-term unemployment in 2018 and 2020 respectively. For each year, we pool observations from February – April.
Source: IEB 16.00.01.

We find a large increase in the share of jobseekers that become long-term unemployed across all industries. The increase in the fraction of individuals remaining unemployed for at least six months between 2018 and 2020 is lowest in the health sector, at 12.2 percentage points, and the highest in the logistics sector, at 19 percentage points. Simultaneously, the ROC-AUC Score declined for all industries between 2018 and 2020. However, similar to the subgroup analysis in Figure 6, Figure 7 does not show a clear correlation between the level of increase in long-term unemployment and the level of decrease in the ROC-AUC Score across different industries.

[18] We focus on industries for which we observe at least 1,000 entries into unemployment during February and April 2020 in our data.

6 Conclusion

In this paper, we compared the (changes in) predictive performance of two popular statistical learning techniques – logistic regression and random forest. Our results show that random forest consistently out-performs logistic regression, both, before and after the beginning of the COVID-19 pandemic. With an increasing number of new jobseekers and a growing fraction of those jobseekers becoming long-term unemployed during the pandemic, the predictive performance of both methods declined. Changes in the composition of the long-term unemployed, with respect to recent labor market histories and previous sector of employment, do not seem to be a major explanation for the decrease in prediction performance we observe after the beginning of the pandemic.

Our results illustrate an important facet of statistical learning techniques: Such methods depend on past data and large shocks to either the distribution of characteristics or the distribution of the outcome of interest – or both – can have a negative impact on the predictive performance of such methods. In our application, we use historical labor market data to predict an unprecedented labor market shock and we do indeed observe a decline in predictive performance of algorithm-based predictions. Whether we are able to improve the predictive performance of machine learning techniques during the subsequent waves of the pandemic and the associated lockdowns, however, remains an open question until more recent data become available.

Finally, the initial labor market shock in terms of unemployment was relatively modest in Germany compared to other countries, such as the US and France (Mayhew/Anand 2020). Thus, the predictive performance of statistical models for labor markets in countries where the unemployment rate – and therefore likely also the composition of the unemployed – was more strongly affected by the first lockdowns may decline even more strongly due to those sudden changes in labor market conditions.

References

Agrawal, A., Gans, J. S.; Goldfarb, A. (2019): Artificial intelligence: the ambiguous labor market impact of automating prediction. In: Journal of Economic Perspectives, 33(2), p. 31–50.

Albanesi, S.; Kim, J. (2021): Effects of the COVID-19 recession on the US labor market: Occupation, family; gender. In: Journal of Economic Perspectives, 35(3), p. 3–24.

Athey, S. (2017): Beyond prediction: Using big data for policy problems. In: Science 355, p. 483–485.

Athey, S.; Imbens, G. W. (2017): The state of applied econometrics: Causality and policy evaluation. In: Journal of Economic Perspectives 31.2, p. 3–32.

Athey, S.; Imbens, G. W. (2019): Machine learning methods that economists should know about. Annual Review of Economics 11, p. 685–725.

Brewer, M. and Gardiner, L. (2020): The initial impact of COVID-19 and policy responses on household incomes. In: Oxford Review of Economic Policy, 36 (Supplement_1), p. 187–S199.

Bauer, A. and Weber, E. (2021): COVID-19: how much unemployment was caused by the shutdown in Germany? In: Applied Economics Letters 28(12), p. 1053–1058.

Biau, G. (2012): Analysis of a random forests model. In: The Journal of Machine Learning Research, 13.1, p. 1063-1095.

Breiman, L. (2001): Random forest. In: Machine Learning 45, p. 5–32.

Bradley, J., Ruggieri, A.; Spencer, A. H. (2021): Twin peaks: COVID-19 and the labor market. European Economic Review, 138, 103828.

Brinca, P., Duarte, J. B.; Faria-e-Castro, M. (2021): Measuring labor supply and demand shocks during COVID-19. European Economic Review, 139, 103901.

Buch, T., Hamann, S., Niebuhr, A., Roth, D.; Sieglen, G. (2021): Arbeitsmarkteffekte der Corona-Krise - Sind Berufsgruppen mit niedrigen Einkommen besonders betroffen? * Zeitgespräch. In: Wirtschaftsdienst, 101(1), p. 14–17.

Cockx, Bart, Lechner, Michael and Bollens, Joost (2020): Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium, Papers, arXiv.org, https://EconPapers.repec.org/RePEc:arx:papers:1912.12864.

Desiere, S., Langenbucher, K. and Struyven, L. (2019): Statistical profiling in public employment services: An international comparison.

Dorner, M., Heining, J., Jacobebbinghaus, P. and Seth, S. (2010): The sample of integrated labour market biographies. In: Journal of Contextual Economics: Schmollers Jahrbuch, 130(4), p. 599–608.

Fernández-Reino, M., Sumption, M. and Vargas-Silva, C. (2020): From low-skilled to key workers: the implications of emergencies for immigration policy. In: Oxford Review of Economic Policy, 36 (Supplement_1), p. 382–396.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H.; Herrera, F. (2012): A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-; Hybrid-Based Approaches. In: IEEE

Transactions on Systems, Man; Cybernetics, Part C (Applications and Reviews) 42.4, p. 463–484

Gartner, H., Hutter, C. and Weber, E. (2021): Große Rezession und Corona-Krise: Wie der Arbeitsmarkt zwei sehr unterschiedliche Krisen bewältigt (No. 27/2021): IAB-Kurzbericht.

Goller, D., Lechner, M., Moczall, A.; Wolff, J. (2020): Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. Labour economics, 65, 101855.

Hastie, T., Tibshirani, R.; Friedman, J. (2011): The elements of statistical learning, 2nd edition, Springer, Berlin: Springer series in statistics.

Huang, J.; Ling, C. X. (2005): Using AUC and accuracy in evaluating learning algorithms. In: IEEE Transactions on knowledge and Data Engineering 17.3, p. 299–310.

Ke, X.; Hsiao, C. (2022): Economic impact of the most drastic lockdown during COVID-19 pandemic – The experience of Hubei, China. Journal of Applied Econometrics, 37(1), p. 187–209.

Kern, C., Bach, R.L., Mautner, H. and Kreuter, F. (2021): Fairness in algorithmic profiling: A German case study. arXiv preprint arXiv:2108.04134.

Knaus, M.C., Lechner, M. and Strittmatter, A. (2020): Heterogeneous employment effects of job search programmes: A machine learning approach. In: Journal of Human Resources, 0718-9615R1.

Lechner, M. and Smith, J. (2007): What is the value added by caseworkers? In: Labour economics, 14(2), p. 135–151.

Mayhew, K. and Anand, P. (2020): COVID-19 and the UK labour market. In: Oxford Review of Economic Policy, 36 (Supplement_1), p. 215–224.

Mühlbauer, S. and Weber, E. (2022): Machine Learning for Labour Market Matching (No. 202203). Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Organisation for Economic Co-operation and Development (OECD) (2021): OECD Employment Outlook 2021: Navigating the COVID-19 Crisis and Recovery, OECD Publishing, Paris, https://doi.org/10.1787/5a700c4b-en.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettnehofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M.; Duchesnay, É. (2011): Scikit-learn: Machine Learning in Python. In: The Journal of Machine Learning Research 12, p. 2825–2830.

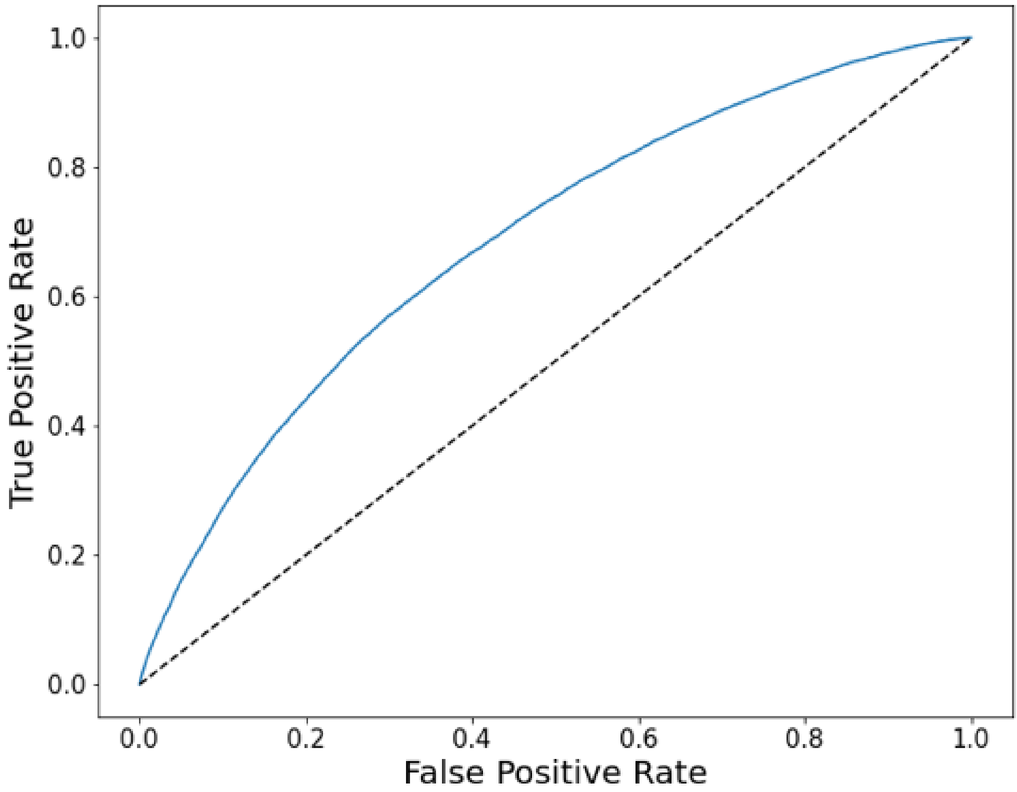
Pope, D. G. and Sydnor, J. R. (2011): Implementing anti-discrimination policies in statistical profiling models. In: American Economic Journal: Economic Policy, 3(3), p. 206–31.

Sansone, D. and Zhu, A. (2021): Using Machine Learning to Create an Early Warning System for Welfare Recipients (No. 14377). Institute of Labor Economics (IZA).

Statistik der Bundesagentur für Arbeit (2021): Tabellen: Arbeitslosenquoten – Zeitreihe (Monats- und Jahreszahlen).

Appendix

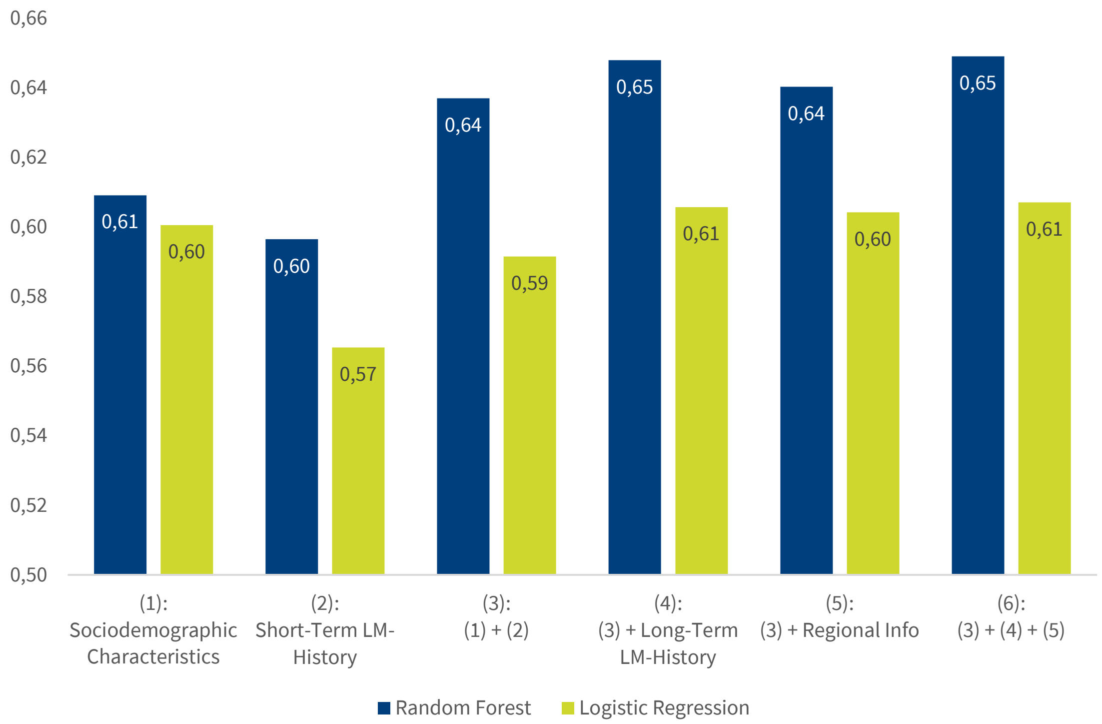
Figure A1: Example of a ROC Curve



Note: This figure shows the ROC-AUC curve for the main specification of the random forest model predicting the probability to become long-term unemployed for unemployment entries in 2018 using a model trained with 2016 data.

Source: IEB 16.00.01.

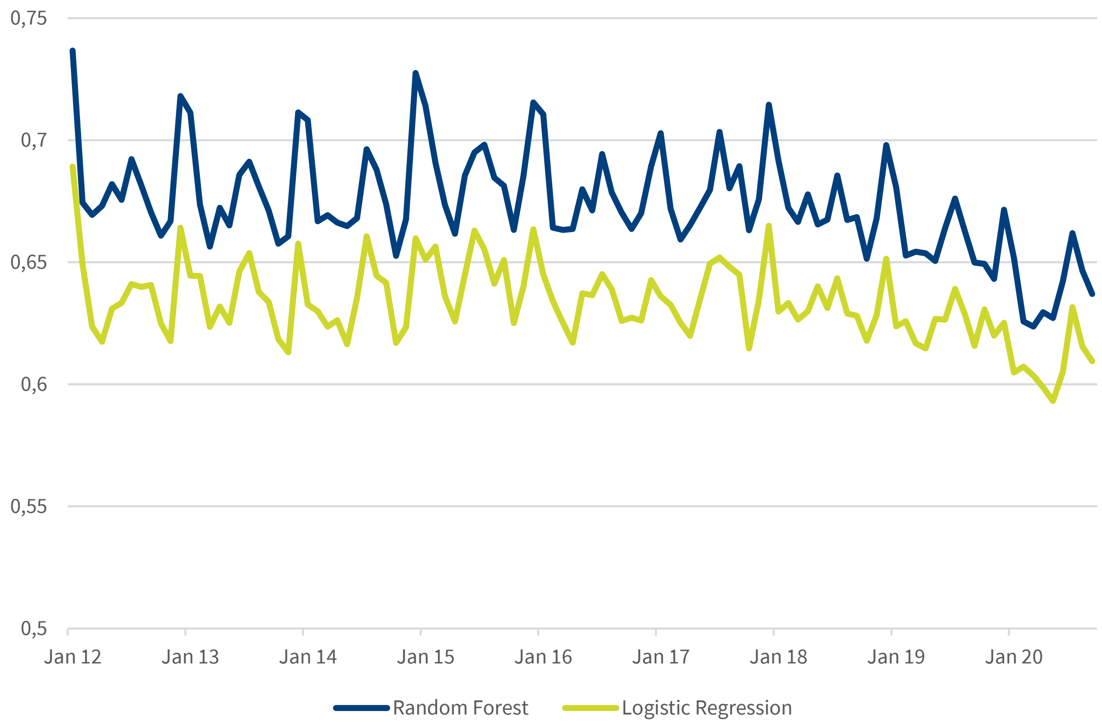
Figure A2: Out-of-sample ROC-AUC Scores predicting the probability to become LTU (2020)



Note: This figure shows the out-of-sample ROC-AUC Scores for different specifications (see main text for details) of logistic regression and random forest models predicting the probability to remain unemployed for at least 6 months using unemployment entries from 2020. The models were trained using data from 2018.

Source: IEB 16.00.01.

Figure A3: Monthly out-of-sample ROC-AUC Scores for the main random forest and logistic regression



Note: This figure shows the monthly out-of-sample ROC-AUC Scores for the main logistic regression and random forest models predicting the probability to remain unemployed for at least 6 months. Each model was trained using data from the same month in year T-1 to predict the probability to become LTU in the respective month in year T. The graph includes predictions for unemployment entries from Jan. 2012 – Sept 2020.

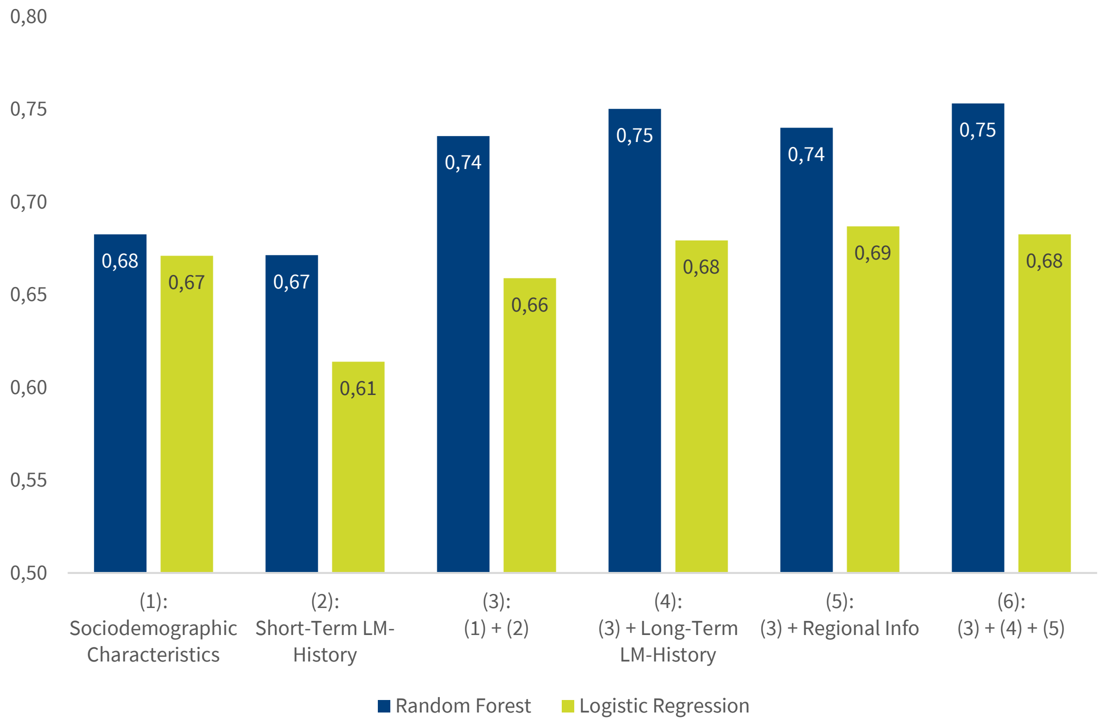
Source: IEB 16.00.01.

Table A1: List of Available Features

Feature Group	Feature
Sociodemographic Characteristics	Sex
	Age
	German
	School Degree (6 Dummies)
	Vocational Degree (4 Dummies)
Short-Term Labor Market History	Tot. Earnings 1 Year Before Unemployment
	Tot. Earnings 1 Year Before Unemployment (Regular Employment)
	Tot. Earnings 1 Year Before Unemployment (Marginal Employment)
	Tot. Earnings 1 Year Before Unemployment (Apprenticeship)
	Tot. Days Employed 1 Year Before Unemployment
	Tot. Days Employed 1 Year Before Unemployment (Regular Employment)
	Tot. Days Employed 1 Year Before Unemployment (Marginal Employment)
	Tot. Days Employed 1 Year Before Unemployment (Apprenticeship)
	Daily Wage Last Job
Long-Term Labor Market History	Tot. Earnings 7 Years Before Unemployment
	Tot. Earnings 7 Years Before Unemployment (Regular Employment)
	Tot. Earnings 7 Years Before Unemployment (Marginal Employment)
	Tot. Earnings 7 Years Before Unemployment (Apprenticeship)
	Tot. Days Employed 7 Years Before Unemployment
	Tot. Days Employed 7 Years Before Unemployment (Regular Employment)
	Tot. Days Employed 7 Years Before Unemployment (Marginal Employment)
	Tot. Days Employed 7 Years Before Unemployment (Apprenticeship)
	Industry Last Job (21 Dummies)
	Occupation Last Job (37 Dummies)
Regional Characteristics District (=Kreis) Level	Overall Unemployment Rate
	Long-Term Unemployment Rate
	Unemployment Rate (Ages 15-25)
	Unemployment Rate (Ages 55-64)
	Unemployment Rate Men
	Unemployment Rate Women
	Mean Wage
	Mean Wage (Regular Employees)
	Mean Wage (Marginal Employees)
	Mean Wage (Apprentices)
	Mean Worker Age
	Mean Worker Age (Regular Employees)
	Mean Worker Age (Marginal Employees)
	Mean Worker Age (Apprentices)
	Fraction Regular Employees
	Fraction Marginal Employees
	Fraction Apprentices
	Fraction of Workers Below 25 Years
	Fraction of Workers Above 50 Years

Note: This table shows the list of available features.
Source: IEB 16.00.01.© IAB

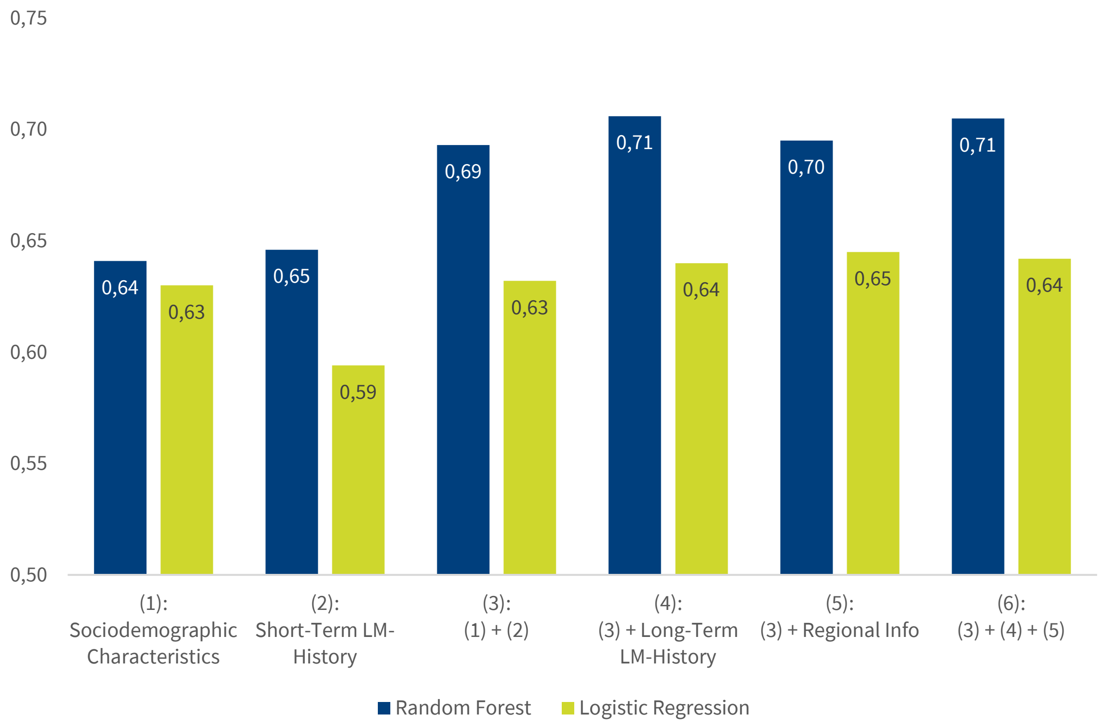
Figure B1: Out-of-sample ROC-AUC Scores predicting the probability to become 12M-LTU (2018)



Note: This figure shows the out-of-sample ROC-AUC Scores for different specifications (see main text for details) of logistic regression and random forest models predicting the probability to remain unemployed for at least 12 months using unemployment entries from 2018. The models were trained using data from 2016.

Source: IEB 16.00.01.

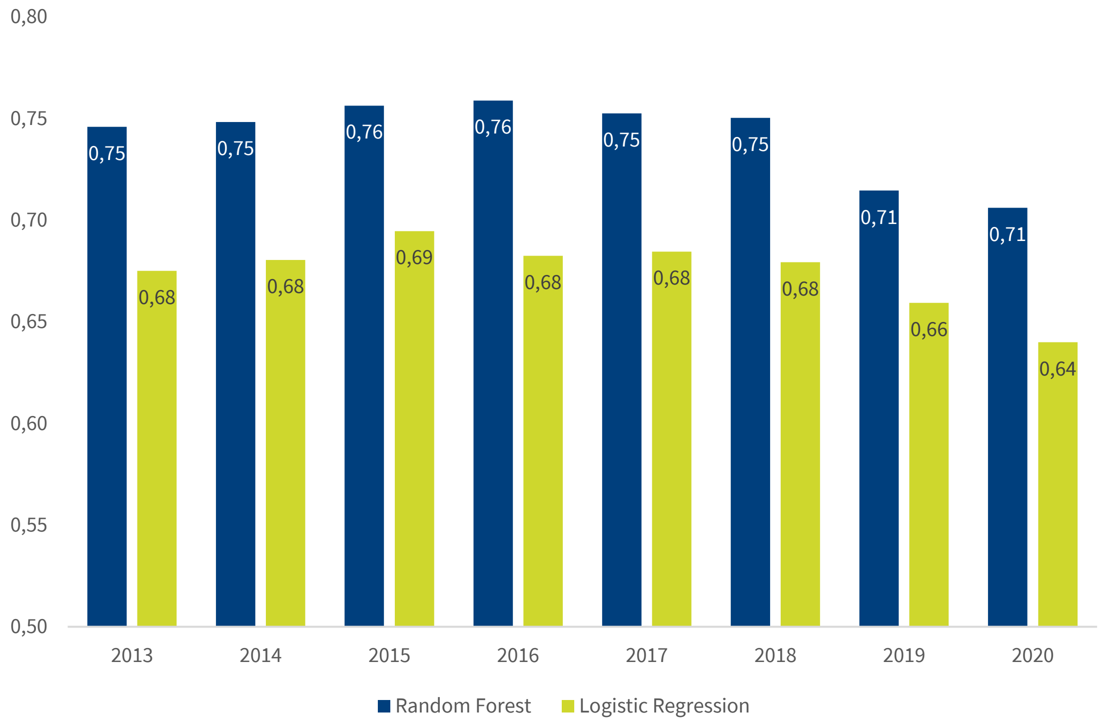
Figure B2: Out-of-sample ROC-AUC Scores predicting the probability to become 12M-LTU (2020)



Note: This figure shows the out-of-sample ROC-AUC Scores for different specifications (see main text for details) of logistic regression and random forest models predicting the probability to remain unemployed for at least 12 months using unemployment entries from 2020. The models were trained using data from 2018. The data from both years only include unemployment entries up until March.

Source: IEB 16.00.01.

Figure B3: Yearly out-of-sample ROC-AUC Scores predicting the probability to become 12M-LTU



Note: This figure shows the yearly out-of-sample ROC-AUC Scores for the main specification of the random forest model and the logistic regression model predicting the probability to remain unemployed for at least 12 months. Each model was trained using data from year T-2 to predict the outcome of interest in year T. The predictions in 2020 only include unemployment entries up until March.

Source: IEB 16.00.01, 2011 – 2021.

Figures

Tables

Imprint

IAB-Discussion Paper 8|2022

Date of publication
9 May 2022

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

Download of this IAB-Discussion Paper
https://doku.iab.de/discussionpapers/2022/dp0822.pdf

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

Website
www.iab.de

ISSN
2195-2663

DOI
10.48720/IAB.DP.2208

Corresponding author

Max Kunaschk
Email: max.kunaschk@iab.de
Julia Lang
Email: julia.lang@iab.de