ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Dumpert, Florian; Beck, Martin

Article — Published Version

Verbesserung der Datengrundlage der Mindestlohnforschung mittels maschineller Lernverfahren

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:
Springer Nature

Suggested Citation: Dumpert, Florian; Beck, Martin (2023) : Verbesserung der Datengrundlage der Mindestlohnforschung mittels maschineller Lernverfahren, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 17, Iss. 1, pp. 5-34, https://doi.org/10.1007/s11943-023-00318-w

This Version is available at:
https://hdl.handle.net/10419/308975

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

AStA Wirtschafts- und Sozialstatistisches Archiv (2023) 17:5–34
https://doi.org/10.1007/s11943-023-00318-w

ORIGINALVERÖFFENTLICHUNG

Verbesserung der Datengrundlage der Mindestlohnforschung mittels maschineller Lernverfahren

Florian Dumpert · Martin Beck

Eingegangen: 24. November 2021 / Angenommen: 1. März 2023 / Online publiziert: 24. April 2023
© Der/die Autor(en) 2023

Zusammenfassung Mit der Einführung des allgemeinen gesetzlichen Mindestlohnes zum 1. Januar 2015 stieg der Bedarf an geeigneten Daten für die Evaluation der Mindestlohnwirkungen. Die Mindestlohnkommission empfahl daher in ihrem ersten Bericht an die Bundesregierung die Verknüpfung der Integrierten Erwerbsbiografien mit der Verdienststrukturerhebung. Die Paneldaten der Integrierten Erwerbsbiografien sollen durch die Verknüpfung mit ansonsten fehlenden Angaben zum Bruttostundenverdienst bzw. zur Mindestlohnbetroffenheit aus der Verdienststrukturerhebung angereichert werden, um so die Analysemöglichkeiten zu verbessern. Sowohl ein deterministisches als auch ein probabilistisches Record Linkage waren aus rechtlichen Gründen nicht umsetzbar. Daher hat das Statistische Bundesamt die Option einer Anreicherung der Integrierten Erwerbsbiografien mit Informationen zur Mindestlohnbetroffenheit aus der Verdienststrukturerhebung auch mittels maschineller Lernverfahren geprüft. Im Fokus standen die Methoden „Random Forest“ und „Boosting“, die keine Verknüpfung oder Weitergabe von Einzeldatensätzen erfordern und daher datenschutzrechtlich handhabbar sind. Über die konkrete Vorgehensweise, speziell die getesteten Modellierungsvarianten für die Teilpopulationen Vollzeit-, Teilzeit- und geringfügig entlohnte Beschäftigte, wird in diesem Beitrag ausführlich berichtet.

Die vorliegende Untersuchung hat gezeigt, dass entsprechende Modelle prinzipiell erlernbar sind und es somit grundsätzlich möglich ist, die Integrierten Erwerbsbiografien um eine Angabe zur Mindestlohnbetroffenheit zu ergänzen. Deutlich erkennbar wurde im Zuge der Untersuchungen aber auch der Zielkonflikt zwischen Sensitivität und positivem Vorhersagewert, das heißt der Konflikt zwischen dem Ziel, möglichst alle vom Mindestlohn betroffenen Beschäftigten als solche zu klassifizieren, und dem Ziel, bei dieser Klassifikation möglichst keine vom Mindestlohn nicht

✉ Florian Dumpert · Martin Beck
Statistisches Bundesamt, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Deutschland
E-Mail: florian.dumpert@destatis.de

betroffenen Beschäftigten, irrtümlich als „vom Mindestlohn betroffen“ zu kennzeichnen. In der Praxis muss diesem Zielkonflikt bei der Auswahl der eingesetzten Modelle in Abhängigkeit vom angestrebten Analysezweck Rechnung getragen werden.

Ob, wann und wie die Resultate der vorliegenden Untersuchung in den Integrierten Erwerbsbiografien konkret umgesetzt werden können, bedarf weiterer Untersuchungen und Festlegungen.

Schlüsselwörter Maschinelles Lernen · Random Forest · Boosting · Verdienststrukturerhebung · Integrierte Erwerbsbiografien · Mindestlohnforschung

JEL-Klassifikation C45 · C52 · J31

Improvement of the data basis of minimum wage research by means of machine learning methods

Abstract With the introduction of the general statutory minimum wage on January 1st, 2015, the need for suitable data for evaluating the effects of the minimum wage increased. In its first report to the Federal Government, the Minimum Wage Commission therefore recommended linking the Integrated Employment Biographies with the Structure of Earnings Survey. The panel data of the Integrated Employment Biographies are to be enriched by linking them with otherwise missing data from the Structure of Earnings Survey on gross hourly earnings or on the information whether an employee is affected by the minimum wage (minimum wage affectedness) in order to make more analyses possible. Deterministic or probabilistic record linkage were not feasible for legal reasons. Therefore, the Federal Statistical Office examined the option of enriching the Integrated Employment Biographies with information on minimum wage affectedness from the Structure of Earnings Survey using also machine learning methods. The focus was on “random forest” and “boosting” methods, which do not require sharing or linking individual data records and are therefore manageable in terms of data protection. The concrete procedure, especially the tested modelling variants for the sub-populations of full-time, part-time and marginally paid employees, is reported in detail in this article.

The present study has shown that models can be learnt and that the Integrated Employment Biographies can therefore be supplemented with an indication of the minimum wage affectedness in principle. However, the conflict between sensitivity and positive predictive value, i.e., the conflict between the goal of classifying all employees affected by the minimum wage as such and the goal of not erroneously recording any employees not affected by the minimum wage in this classification, also became clearly apparent in the course of the investigations. In practice, this conflict of objectives must be taken into account when selecting the models used, depending on the intended purpose of the analysis.

Whether, when and how the results of the present study can be concretely implemented in the Integrated Employment Biographies requires further investigation.

Keywords Machine Learning · Random Forest · Boosting · Structural Earnings Survey · Integrated Employment Biographies · Minimum wage research

1 Einleitung

„Wir wollen einen gesetzlichen, flächendeckenden Mindestlohn in Höhe von mindestens 8,50 € einführen – einheitlich in Ost und West, der auf Vorschlag einer vom BMAS eingesetzten Mindestlohnkommission jährlich angepasst wird." (SPD 2013, S. 19) Mit dieser im April 2013 auf einem Bundesparteitag als Teil ihres „Regierungsprogramms" beschlossenen Forderung ging die SPD in den Wahlkampf für die Bundestagswahl am 22. September 2013. Für Politik und Wissenschaft stellten sich somit unter anderem folgende Fragen:

- Wie viele Beschäftigte wären von der Einführung eines solchen allgemeinen gesetzlichen Mindestlohnes betroffen?
- Welche Auswirkungen hätte ein solcher Mindestlohn beispielsweise auf Arbeitsplätze und Löhne?
- Gibt es geeignete Datengrundlagen, um Analysen zu den beiden erstgenannten Fragen durchführen zu können?

Die Eignung der vorhandenen Datengrundlagen für die Mindestlohnforschung wurde bereits vor und bei der Evaluation der existierenden Branchenmindestlöhne 2011 bis 2013 kritisch hinterfragt (Mindestlohnkommission 2016, S. 24). Die amtlichen Verdienststatistiken spielten in dieser Evaluationsrunde kaum eine Rolle. Mit Blick auf die absehbare Einführung des allgemeinen gesetzlichen Mindestlohnes schloss der Gesetzgeber jedoch durch eine Anpassung des Verdienststatistikgesetzes noch bestehende Erfassungslücken der für das Jahr 2014 anstehenden, nur alle vier Jahre durchzuführenden Verdienststrukturerhebung (VSE). Ziel war es, diese für eine so genannte „Nullpunktmessung" vor der Einführung des Mindestlohnes zu ertüchtigen.

Mit dem Mindestlohngesetz wurden von der großen Koalition die oben genannten Forderungen der SPD umgesetzt und zum 1. Januar 2015 ein allgemeiner gesetzlicher Mindestlohn von 8,50 € pro Arbeitsstunde eingeführt. Auch wurde die Mindestlohnkommission eingesetzt, deren Aufgabe es unter anderem ist, die Auswirkungen des Mindestlohns zu evaluieren und der Bundesregierung hierüber zu berichten. Hierzu benötigt sie geeignete Daten.

Die Mindestlohnkommission hat in ihrem ersten Bericht Daten der Verdienststrukturerhebung 2014 umfassend verwendet. Da die nächste Datenerhebung aber erst für 2018 anstand, empfahl sie die Verknüpfung der VSE mit den Integrierten Erwerbsbiografien (IEB) der Bundesagentur für Arbeit bzw. des Instituts für Arbeitsmarkt- und Berufsforschung (IAB): „Die Mindestlohnkommission empfiehlt als wichtige Maßnahme zur Verbesserung der Datenlage für die wissenschaftliche Evaluation des Mindestlohns eine Verknüpfung der Beschäftigtenstatistik der Bundesagentur für Arbeit bzw. der darauf aufbauenden Integrierten Erwerbsbiografien (IEB) mit der Verdienststrukturerhebung (VSE). [...] Die Beschäftigtenstatistik und die darauf aufbauenden Integrierten Erwerbsbiografien (IEB) enthalten zwar detaillierte Informationen zu den Erwerbsverläufen von Beschäftigten, haben aber eine entscheidende Schwäche: Die Arbeitszeit wird nur sehr grob erfasst, sodass die Berechnung von Stundenlöhnen nicht möglich ist. [...] Dieses Defizit ließe sich durch eine Verknüpfung beider Datensätze beheben. Angesichts der Sensibilität von Sozi-

aldaten muss eine solche Verknüpfung hohen datenschutzrechtlichen Anforderungen genügen." (Mindestlohnkommission 2016, S. 32) Die Integrierten Erwerbsbiografien enthalten vollständige, historisierte und aufbereitete Prozessdaten unterschiedlicher administrativer Datenquellen. Mithilfe der IEB ist es möglich, Erwerbsverläufe – bestehend aus Beschäftigungszeiten, Leistungsempfangszeiten, Arbeitslosen- und Arbeitsuchendenzeiten sowie Maßnahme-Teilnahmen in Kontenform – nachzuvollziehen (Ohne Verfasser 2017; Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (o.J.)).

Die Paneldaten der IEB für 2014, also dem Jahr vor Einführung des gesetzlichen Mindestlohnes, sollen durch die Verknüpfung mit ansonsten fehlenden Angaben zum Bruttostundenverdienst bzw. zur Mindestlohnbetroffenheit aus der VSE angereichert werden, um so die Analysemöglichkeiten zu verbessern. Neben einem deterministischen oder probabilistischen Record Linkage kommen fachlich und methodisch hierfür (nicht nur, aber auch) statistische maschinelle Lernverfahren, wie Support Vector Machines (SVM) und Random Forests, in Frage (Himmelreicher et al. 2017).

Eine Überprüfung dieser Optionen machte deutlich, dass ein deterministisches Record Linkage, d. h. die Verknüpfung von ein spezifisches Beschäftigungsverhältnis betreffenden Datensätzen der VSE und der IEB über ein geeignetes Verknüpfungsmerkmal, rechtlich nicht zulässig ist, da im Verdienststatistikgesetz ein solcher Verwendungszweck nicht vorgesehen und die nachträgliche Einholung der Zustimmung zur Datenverknüpfung faktisch unmöglich war (Himmelreicher et al. 2017, S. 15). Auch ein probabilistisches Record Linkage begegnete juristischen Bedenken. Daher hat das Statistische Bundesamt auch die Option einer Anreicherung der IEB mit Informationen aus der VSE mittels maschineller Lernverfahren, die keine Verknüpfung von Einzeldatensätzen erfordern und daher datenschutzrechtlich handhabbar sind, geprüft. Ziel ist es, einen Ansatz zur Verknüpfung von VSE und IEB zu entwickeln, der vom IAB als Halter der IEB-Daten künftig praktisch umgesetzt werden kann. Über die konkrete Vorgehensweise bei der Überprüfung der vorhandenen Optionen und die Ergebnisse wird im diesem Beitrag berichtet. Ob, wann und wie die Resultate in den IEB konkret umgesetzt werden können, bedarf weiterer Untersuchungen und Festlegungen.

Die Evaluation des Mindestlohns anhand der beiden amtlichen Datenquellen VSE und IEB ist natürlich nur ein denkbarer Weg. Der „Gesamtbericht zur Evaluation des allgemeinen gesetzlichen Mindestlohns nach § 23 Mindestlohngesetz" (Bundesministerium für Arbeit und Soziales 2020) wertet zahlreiche Studien zu den Auswirkungen des Mindestlohns systematisch aus, die sich auch anderer Datenquellen – unter anderem: Mannheimer Unternehmenspanel, Mikrozensus und Sozio-oekonomisches Panel – bedienen.

2 Anreicherung der IEB um die Mindestlohnbetroffenheit

Das Statistische Bundesamt hat statistische (maschinelle Lern-) Verfahren mit Daten der VSE 2014 getestet. Ziel war es, Modelle zu lernen, die die Unterscheidung (d. h. die Klassifikation) der Beschäftigten in „vom Mindestlohn betroffen" und „vom Mindestlohn nicht betroffen" mit hinreichender Verlässlichkeit ermöglichen.

Tab. 1 Zur Verfügung stehende Merkmale aus dem gemeinsamen Merkmalskranz von VSE und IEB

Merkmal	Anmerkung
ANZ_BESCH	Anzahl der Beschäftigten des Betriebs
LAND	Sitz des Betriebs (Bundesland)[a]
WZ_Stich	Wirtschaftszweig-2-Steller des Betriebs zum Zeitpunkt der Stichprobenziehung[b]
GESCHLECHT	1 = männlich, 2 = weiblich
ALTER	Alter in Jahren
ZUGEHOERIGKEIT	Dauer der Unternehmenszugehörigkeit in Jahren
TAETIGKEITSSCHLUESSEL2	Tätigkeitsschlüssel 2010, Stelle 6, Höchster allgemeinbildender Schulabschluss (1 = Ohne Schulabschluss, 2 = Haupt-/Volksschulabschluss, 3 = Mittlere Reife oder gleichwertiger Abschluss, 4 = Abitur, 9 = Abschluss unbekannt)
TAETIGKEITSSCHLUESSEL3	Tätigkeitsschlüssel 2010, Stelle 7, Höchster beruflicher Ausbildungsabschluss (1 = Ohne beruflichen Ausbildungsabschluss, 2 = Abschluss einer anerkannten Berufsausbildung, 3 = Meister-/Techniker- oder gleichwertiger Fachschulabschluss, 4 = Bachelor, 5 = Diplom/Magister/Master/Staatsexamen, 6 = Promotion, 9 = Abschluss unbekannt)
TAETIGKEITSSCHLUESSEL4	Tätigkeitsschlüssel 2010, Stelle 8, Arbeitnehmerüberlassung (1 = nein, 2 = ja)
TAETIGKEITSSCHLUESSEL5	Tätigkeitsschlüssel 2010, Stelle 9, Vertragsform (1 = Vollzeit unbefristet, 2 = Teilzeit unbefristet, 3 = Vollzeit befristet, 4 = Teilzeit befristet)
MVERDIENSTGESAMT	Bruttomonatsverdienst insgesamt (voller Betrag)
MVERDIENSTDAVONSV	Sozialversicherung insgesamt (voller Betrag)

[a] Siehe https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Glossar/bundeslaender.html
[b] Siehe https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/klassifikation-wz-2008.html

Diese (eingeschränktere) Information kann an Stelle des eigentlich idealerweise gewünschten Bruttostundenverdienstes auf den vollständigen Datenbestand der IEB (oder potenziell auch auf andere Datenbestände) übertragen werden, indem das Klassifikationsmodell dort zur Ausführung kommt. Es handelt sich also um die Übertragung eines (binären) Musters, das anhand der VSE trainiert wurde, auf die IEB, also eine Zuschätzung von Informationen an die IEB. Als erklärende Variablen für die Modelle werden daher, um die Anwendung auf den IEB zu ermöglichen, nur Merkmale verwendet, die sowohl in der VSE als auch in den IEB vorliegen, und zwar jeweils für Vollzeit-, Teilzeit- und geringfügig entlohnte Beschäftigte. Dieser Kranz besteht aus den in Tab. 1 gelisteten Merkmalen.

Im Paneldatensatz der IEB (bzw. der daraus gewonnenen Stichprobe SIAB) können sodann diejenigen, die 2014 weniger als 8,50 € pro Stunde verdienten, in den Jahren 2015 ff. der Gruppe der Mindestlohnbetroffenen zugeordnet werden, sodass eine Analyse verschiedener relevanter Fragestellungen zu den Mindestlohnwirkungen möglich wird. Wie üblich in derartigen Fällen ist die bedingte Unabhängigkeitsannahme (conditional independence assumption) nicht überprüfbar. Fragestellungen, die auf den Zusammenhang zwischen Mindestlohnbetroffenheit und solchen Variablen abstellen, die nur in den IEB vorhanden sind und nicht zu den erklärenden

Variablen bei der Modellbildung gehören, können somit durch dieses Verfahren im Allgemeinen nicht valide untersucht werden.

Dieser Aufsatz beleuchtet im Folgenden weniger die Interpretation dieser Zuschätzung als Imputation von Werten in die IEB als vielmehr die Fragestellung, wie gut maschinelle Lernverfahren das der Mindestlohnbetroffenheit zugrundeliegende Muster anhand des gemeinsamen Merkmalskranzes rein datenbasiert, d. h. ohne explizite Modellierung, lernen können. Insbesondere arbeitet die Herangehensweise nicht mit Wahrscheinlichkeiten für das gemeinsame Auftreten von Ausprägungen von Merkmalen (aus den drei Gruppen „nur in den IEB vorhanden“, „nur in der VSE vorhanden“, „aus dem gemeinsamen Merkmalskranz“). Stattdessen werden lediglich die aus den statistischen (maschinellen Lern-) Verfahren resultierenden Scores genutzt. Anders als für die valide Bestimmung von Wahrscheinlichkeiten wird hierfür die Annahme der bedingten stochastischen Unabhängigkeit von Teilgruppen der vorhandenen Merkmale nicht benötigt.

3 Datenbasis: Die Verdienststrukturerhebung 2014

In diesem Kapitel wird die Verdienststrukturerhebung 2014 (zum Teil wörtlich) auf Grundlage des zugehörigen Qualitätsberichtes (Statistisches Bundesamt 2016) eingehender dargestellt, um der Leserschaft die Grundzüge sowie die Vor- und Nachteile dieser Erhebung zu verdeutlichen.

In der Verdienststrukturerhebung werden Daten zu Verdiensten erfasst. Diese Daten werden nach Wirtschaftszweigen (WZ) und persönlichen Angaben über die Beschäftigten wie Geschlecht, Geburtsjahr, die Dauer der Unternehmenszugehörigkeit, Beruf und Ausbildungsabschluss gegliedert. Darüber hinaus enthält die VSE Merkmale über das Beschäftigungsverhältnis: Anzahl der bezahlten Arbeitsstunden, Angaben zu Tarifvertrag, Leistungsgruppe, Art der Beschäftigung und den Umfang des Urlaubsanspruchs. Die Verdienststrukturerhebung ermöglicht damit Aussagen über die Verteilung der Verdienste der Beschäftigten sowie über den Einfluss wichtiger Faktoren, die die individuelle Verdiensthöhe bestimmen. Entscheidend für das vorliegend beschriebene Projekt war, dass die Bruttomonatsverdienste sowie die monatlich bezahlten Arbeitsstunden erfasst werden. Folglich können für alle Beschäftigten Bruttostundenverdienste berechnet werden.

Die Verdienststrukturerhebung umfasst Haupt- und Nebenbeschäftigungen; selbständige Tätigkeiten werden nicht erfasst. Es werden nur solche Beschäftigungen erfasst, die über den gesamten Berichtsmonat April bestanden und für die im Berichtsmonat eine Lohnzahlung stattfand. Nicht monatsscharf begonnene oder beendete sowie ruhende Beschäftigungsverhältnisse werden nicht erfasst. Der Berichtszeitraum ist für einige Merkmale das Kalenderjahr und für die meisten Merkmale der Monat April. Saisonale Beschäftigungen werden nicht repräsentativ erfasst. Seit dem Berichtsjahr 2006 findet die VSE regelmäßig alle vier Jahre statt. Für das vorliegende Projekt wurde die Verdienststrukturerhebung 2014 herangezogen. Die Abgrenzung der Wirtschaftszweige basiert auf der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008) (Statistisches Bundesamt 2008).

Die VSE 2014 wurde als Stichprobenerhebung in Betrieben mit sozialversicherungspflichtig Beschäftigten der Wirtschaftsabschnitte A bis S ohne O (Öffentliche Verwaltung, Verteidigung; Sozialversicherung) sowie dem Staat zuzuordnenden Einheiten von P (Erziehung und Unterricht) durchgeführt. Die Grundgesamtheit der Stichprobe umfasste alle Beschäftigungsverhältnisse in diesen Betrieben. Die Stichprobe wurde über ein zweistufiges Auswahlverfahren realisiert, in der ersten Stufe mit einer Betriebsauswahl, in der zweiten Stufe mit einer Auswahl von Beschäftigungsverhältnissen. In die Auswahlgrundlage der ersten Stufe wurden alle Betriebe mit sozialversicherungspflichtig Beschäftigten einbezogen, geschichtet nach dem Bundesland des Betriebs, dem Wirtschaftszweig des Betriebs (84 WZ-2-Steller) und der Zahl der Beschäftigten des Betriebs (sieben Größenklassen). Der nominale Stichprobenumfang betrug 60.000 Betriebe, der Auswahlsatz im Durchschnitt aller Betriebe 3,4 %. Die Auswahlgrundlage der zweiten Stufe umfasste alle Beschäftigten eines in der ersten Stufe ausgewählten Betriebs. Für jede der sieben Beschäftigtengrößenklassen wurde ein fester Auswahlsatz vorgegeben. Der Auswahlsatz nahm mit wachsender Größenklasse ab. Praktisch umgesetzt wurde die Auswahl über eine Startzahl und einen festen Auswahlabstand, die dem Betrieb vorgegeben wurden. Für die zweite Stufe konnten die Auskunftspflichtigen wählen, ob sie die Auswahl selbst durchführen oder den Statistischen Ämtern der Länder überlassen. In letzterem Fall waren die Angaben aller Beschäftigten zu übermitteln, die zufallsgesteuerte Auswahl und Löschung überzähliger Datensätze nahm das statistische Amt vor. Realisiert wurde ein Stichprobenumfang von rund 0,8 Mio. Beschäftigungsverhältnissen (hochgerechnet 32,1 Mio. Beschäftigungsverhältnisse).

Für Beschäftigungsverhältnisse der Wirtschaftsabschnitte O (vollständig) und P (überwiegend) liegt ergänzend eine Sekundärnutzung einer Stichprobe der Datensätze der Personalstandstatistik vor. Die so gewonnenen Angaben über Beschäftigte im öffentlichen Dienst wurden im Rahmen dieses Projektes nicht einbezogen.

Die Hochrechnung der VSE-Ergebnisse erfolgte gebunden mithilfe eines verallgemeinerten Regressionsschätzers (Generalised regression estimator – GREG). Die für GREG relevanten Hilfsvariablen waren die Anzahlen der sozialversicherungspflichtig Beschäftigten und der geringfügig entlohnten Beschäftigten des Betriebs und die Zahl der Betriebe mit abhängig Beschäftigten im April 2014 laut Verwaltungsdatenspeicher der Statistischen Ämter.

Ab 2022 geht die Verdienststrukturerhebung in der neuen Verdiensterhebung auf, einer monatlichen Verdienststatistik, die aber ähnlich tief die Struktur der Verdienste abbildet und darüber hinaus auch einen Panelcharakter aufweisen wird (Statistische Ämter des Bundes und der Länder 2020; De la Croix 2021).

4 Aufbau der Untersuchung

4.1 Überblick

Bereits frühere, unveröffentlichte Untersuchungen zeigten, dass eine Trennung des Datenbestandes nach Vollzeit- (VZ), Teilzeit- (TZ) und geringfügig entlohnten Beschäftigten (GF) mit dem Ziel, separate Modelle zu lernen, sinnvoll ist. Dieser

Ansatz wurde während der hier dargestellten Untersuchungen auch weiterverfolgt. Insgesamt standen dabei Daten von 490.707 Vollzeitbeschäftigten, 162.646 Teilzeitbeschäftigten und 110.010 geringfügig entlohnten Beschäftigten zur Verfügung.

Als statistische (maschinelle Lern-) Verfahren wurden in allen drei Fällen (VZ, TZ, GF) neben dem Naive-Bayes-Klassifizierer und der logistischen Regression, die als Benchmark dienten, verschiedene Random-Forest- und Boosting-Ansätze sowie Support Vector Machines untersucht. Weitere Variationen ergaben sich durch verschiedene Strategien (upsampling, downsampling, SMOTE, ROSE) zum Umgang mit dem Ungleichgewicht im Auftreten der beiden Klassen („vom Mindestlohn betroffen" und „vom Mindestlohn nicht betroffen") im Datenbestand (imbalanced data) sowie durch die Berücksichtigung von Zusatzinformationen auf Aggregatsebene.

Sobald ein Modell trainiert ist, kann es auf Validierungs- oder Testdaten angewendet werden. Für einen Beschäftigten liefert ein Modell dann einen Score (zwischen 0 und 1), der angibt, ob das Modell dem Beschäftigten eher die Eigenschaft „vom Mindestlohn betroffen" oder eher die Eigenschaft „vom Mindestlohn nicht betroffen" zuweist. Da am Ende eine binäre Entscheidung erforderlich ist, muss ein Schwellenwert festgelegt werden, ab dem einem Beschäftigten die Eigenschaft „vom Mindestlohn betroffen" zugewiesen wird. Dieser Klassifikationsschwellenwert kann, muss aber nicht 0,5 betragen. Aus der Variation des Schwellenwertes ergeben sich Konsequenzen: Würde der Schwellenwert beispielsweise – als Extremfall auf der einen Seite – auf 0,0 festgelegt, so würden (unabhängig vom Modell und dessen Output) alle Beschäftigten als „vom Mindestlohn betroffen" klassifiziert. Analoges gilt – als Extremfall auf der anderen Seite – für einen Schwellenwert von 1,0 und „vom Mindestlohn nicht betroffen". Je höher also der Schwellenwert gewählt wird, desto weniger Beschäftigte werden als „vom Mindestlohn betroffen" klassifiziert. Die Variation des Schwellenwertes ist damit eine weitere Möglichkeit zum Umgang mit imbalanced data und erlaubt die Ausrichtung der Klassifikation zugunsten oder zulasten einer Klasse.

Verschiedene Methoden in verschiedenen Variationen liefern dabei verschiedene Ergebnisse für die drei wichtigsten Zielgrößen (Gesamtgenauigkeit, Sensitivität, positiver Vorhersagewert, siehe auch Abschn. 4.6) und lassen somit Raum für Schwerpunktsetzungen bei der Modellauswahl. Tab. 2 zeigt im Überblick die geprüften Kombinationen aus 15 Methoden, fünf Samplingverfahren und zwei Varianten hinsichtlich der Menge der erklärenden Variablen. Es wurden 150 ($=15\times5\times2$) verschiedene Modelle trainiert; jedes dieser Modelle wurde für neun (0,1; 0,2; ...; 0,9) Klassifikationsschwellenwerte ausgewertet. Insgesamt liegen damit 1350 Auswertungen vor.

Das Vorgehen der Untersuchung und die denkbare spätere Anwendung der Modelle lässt sich grob wie folgt – und separat für jede der drei Gruppen (VZ, TZ, GF) – skizzieren:

1. Die Menge der einschlägigen Beobachtungen in der VSE 2014 wird disjunkt zerlegt in einen Übungsdatensatz und einen Testdatensatz. Details hierzu finden sich in Abschn. 4.5.
2. Der Übungsdatensatz wird mehrfach in einen Trainings- und einen Validierungsdatensatz aufgeteilt (siehe ebenfalls Abschn. 4.5).

Tab. 2 Kombinationen der Spezifikationen der gelernten Modelle

Statistische Methoden *(vgl. Abschn. 4.2)*	Samplingverfahren *(vgl. Abschn. 4.3)*	Zusatzinformationen *(vgl. Abschn. 4.4)*
– Logistische Regression (`logreg`) – Naive Bayes – Boosting (C50) mit bis zu 100 Zyklen und minimal 4 Beobachtungen in einem Blatt (`C50_100_4`) – Boosting (C50) mit bis zu 100 Zyklen und minimal 50 Beobachtungen in einem Blatt (`C50_100`) – Boosting (C50) mit bis zu 40 Zyklen und minimal 4 Beobachtungen in einem Blatt (`C50_40_4`) – Boosting (C50) mit bis zu 40 Zyklen und minimal 50 Beobachtungen in einem Blatt (`C50_40`) – Random Forest (ranger) mit mtry = 3 und minimal 100 Beobachtungen in einem Blatt (`ranger100sqrt`) – Random Forest (ranger) mit mtry = 3 und minimal 50 Beobachtungen in einem Blatt (`ranger50sqrt`) – Random Forest (ranger) mit mtry = 3 und minimal 25 Beobachtungen in einem Blatt (`ranger25sqrt`) – Random Forest (ranger) mit mtry = 3 und minimal 4 Beobachtungen in einem Blatt (`ranger4sqrt`) – Random Forest (ranger) mit mtry = 10 und minimal 100 Beobachtungen in einem Blatt (`ranger100hoch`) – Random Forest (ranger) mit mtry = 10 und minimal 50 Beobachtungen in einem Blatt (`ranger50hoch`) – Random Forest (ranger) mit mtry = 10 und minimal 25 Beobachtungen in einem Blatt (`ranger25hoch`) – Random Forest (ranger) mit mtry = 10 und minimal 4 Beobachtungen in einem Blatt (`ranger4hoch`) – SVM (liquidSVM)	– Upsampling – Downsampling – SMOTE – ROSE – unveränderter Datensatz	– Mindestlohnanteil berücksichtigt – Mindestlohnanteil nicht berücksichtigt

3. Auf den Trainingsdatensätzen werden verschiedene Modelle und verschiedene Modellvarianten (siehe Abschn. 4.2 und 4.3 und 4.4) trainiert, auf den zugehörigen Validierungsdatensätzen werden die resultierenden Modelle auf ihre Güte hin überprüft (siehe Abschn. 4.6).
4. Auf Basis dieser Ergebnisse kann nun das IAB eines oder mehrere Modelle zur späteren Anwendung auf die IEB auswählen.
5. Ein eigens zurückgelegter Testdatensatz steht zur Verfügung, um die Güte der in Schritt 4 ausgewählten Modelle noch einmal anhand bislang nicht gesehenen Datenmaterials zu schätzen (vgl. hierzu Abschn. 4.5 und Kap. 9). Dieser Schritt simuliert die Anwendung der ausgewählten Modelle auf die IEB und liefert somit eine Schätzung für die erwartete Genauigkeit der Modelle dort. Anschließend werden die ausgewählten Modelle auf allen zur Verfügung stehenden Daten (d. h. Übungsdatensatz und Testdatensatz) noch einmal neutrainiert und zur Verfügung gestellt.

6. Die ausgewählten Modelle werden auf die Daten der Beschäftigten in den IEB angewandt und reichern diese dadurch um die Mindestlohnbetroffenheit im Jahr 2014 an. Neben den ausgewählten Modellen wird hierzu je „Wirtschaftszweig-2-Steller × Bundesland × Größenklasse des Betriebs"-Schicht der durchschnittliche Anteil der vom Mindestlohn betroffenen Beschäftigten der Schicht in der VSE bereitgestellt (siehe Abschn. 4.4).
7. Das IAB selbst, aber auch Forscherinnen und Forscher können auf Basis der angereicherten IEB und im Bewusstsein, dass die Mindestlohneigenschaft eine geschätzte Größe ist, Untersuchungen zur Wirkung des Mindestlohns durchführen.

4.2 Einbezogene statistische Methoden

Auf eine eingehende Beschreibung der logistischen Regression (Cox 1958) wird mit Verweis auf die Literatur, beispielsweise Fahrmeir et al. (2009), verzichtet. Ebenfalls verzichtet wird auf die Darstellung der Methoden Naive Bayes (Hastie et al. 2009) und Support Vector Machines (Boser et al. 1992; Hastie et al. 2009). Naive Bayes schnitt in beinahe allen Fällen deutlich schlechter ab als die zweite Benchmarkmethode, die logistische Regression, und wird daher im Folgenden nicht weiter betrachtet oder ausgewiesen. Support Vector Machines werden aus zweierlei Gründen von der weiteren Betrachtung ausgeschlossen. Einerseits schnitten sie im Allgemeinen nicht besser als die baumbasierten Verfahren Random Forest und Boosting ab (in keinem Fall waren sie so deutlich besser, dass sie zwingend in der weiteren Auswertung berücksichtigt werden müssten), andererseits stehen Modelle basierend auf Support Vector Machines zur Weitergabe sowieso nicht zur Verfügung. Letzteres liegt darin begründet, dass die ein SVM-Modell konstituierenden Support-Vektoren (also Daten einzelner Beschäftigter) bei Weitergabe den Schutzbereich der amtlichen Statistik unzulässigerweise verlassen würden. Die skizzenhafte Methodenbeschreibung behandelt also nur Random Forests (Breiman 2001; Hastie et al. 2009) und Boosting (Freund und Schapire 1996; Hastie et al. 2009).

Sowohl Random Forest als auch Boosting sind (zumindest in ihrer Anwendung hier) baumbasierte Verfahren und Ensemble-Methoden, d. h. bei beiden Ansätzen werden hier mehrere Klassifikationsbäume (stets 500 bei dem für Random Forest eingesetzten R-Paket „ranger"; maximal 40 bzw. maximal 100 bei dem für das Boosting eingesetzten R-Paket „C50") gelernt. Die Klassifikation selbst, d. h. die Zuordnung eines Beschäftigten zu einer der beiden Klassen „vom Mindestlohn betroffen" oder „vom Mindestlohn nicht betroffen" erfolgt durch Berücksichtigung der Klassifikationsergebnisse der jeweils zugrundeliegenden Bäume. Bei Random Forest hat dabei jeder Baum das gleiche Gewicht, beim Boosting hängt das Gewicht von der Klassifikationsgüte des jeweiligen zugrundeliegenden Baumes ab. Ein weiterer Unterschied zwischen Random Forest und Boosting besteht darin, dass bei ersterem eher weit verzweigte Bäume gelernt werden, bei letzterem eher Baumstümpfe. Die wesentliche Unterscheidung zwischen den beiden Ensemblemethoden besteht jedoch darin, dass beim Random Forest die Bäume unabhängig voneinander gelernt werden; beim Boosting hingegen aufeinander aufbauend, jeweils fokussierend auf diejenigen Beschäftigten, die der vorangehende Baum falsch klassifiziert hat. Wesentlicher Tuningparameter von Random Forests ist „mtry", also die Anzahl an je Split in

den zugrundeliegenden Bäumen zufällig ausgewählten erklärenden Variablen (für Details siehe Breiman et al. 1984; Breiman 2001). Literatur und Voruntersuchungen ergaben, dass im vorliegenden Fall kein Feintuning dieser Größe erforderlich ist, sondern sich auf zwei Fälle beschränkt werden kann: vergleichsweise viele (nämlich mtry = 10, gekennzeichnet durch das Suffix „hoch“) und vergleichsweise wenige (nämlich mtry = 3, gekennzeichnet durch das Suffix „sqrt“) Merkmale. Letzteres orientiert an der aus der Literatur (vgl. James et al. 2013) bekannten Daumenregel, bei Klassifikation je Split ungefähr die Wurzel der insgesamt vorhandenen erklärenden Variablen heranzuziehen.

Im Hinblick auf den bisherigen Einsatz von Random Forest in der deutschen amtlichen Statistik sei auf Feuerhake und Dumpert (2016) und Dumpert und Beck (2017) verwiesen, für Boosting auf Schmidt (2020). Methodische Details finden sich beispielsweise in James et al. (2013) und Hastie et al. (2009).

Beim Trainieren dieser baumbasierten Modelle wurden die aus dem Design der VSE resultierenden Hochrechnungsfaktoren in Form von übergebenen Gewichten der Datenpunkte berücksichtigt, was einer Änderung der Verlustfunktion (hin zu einer die Fehler bei verschiedenen Datenpunkten unterschiedlich gewichtenden Verlustfunktion) entspricht; ebenso bei der Validierung (siehe Abschn. 4.6).

Um sicherzustellen, dass bei den baumbasierten Verfahren Boosting und Random Forest keine Probleme mit der statistischen Geheimhaltung bei der Weitergabe der Modelle auftreten, wurden die Mindestanzahlen an Beschäftigten in den Blättern der jeweils zugrundeliegenden Bäume gesondert betrachtet. Hieraus entstanden unterschiedliche Modellrechnungen: für Random Forest mit mindestens 100 (ranger100x), mindestens 50 (ranger50x), mindestens 25 (ranger25x) und mindestens 4 (ranger4x) Beschäftigten je Blatt, für Boosting mit mindestens 50 (kein Suffix) und mit mindestens 4 (Suffix „_4“) Beschäftigten. Dabei zeigte sich in keinem Fall ein durch die Änderung dieser Mindestanzahl hervorgerufener bedeutsamer Unterschied in den Gütemaßen. Dies bestätigt diesbezügliche Erfahrungen aus einem weiteren auf Basis der Verdienststrukturerhebung 2014 durchgeführten Projekt (Schmidt 2020).

4.3 Einbezogene Samplingverfahren

In allen drei Teilgruppen (Vollzeit-, Teilzeit-, geringfügig entlohnte Beschäftigte) liegt – mehr oder weniger stark ausgeprägt – ein imbalanced-data-Problem vor, also das Phänomen, dass eine der beiden Klassen, hier die vom Mindestlohn betroffenen Beschäftigungsverhältnisse, deutlich unterrepräsentiert ist. Um zu vermeiden, dass eine statistische Methode die Minderheitenklasse gleichsam „opfert“, um gute Ergebnisse auf der ja dominierenden Mehrheitsklasse zu erzielen, werden verschiedene Ansätze erprobt, die die Trainingsdaten derart variieren, dass eine solche „Opferung“ der Minderheitenklasse nicht mehr optimal ist. Untersucht wurden upsampling, downsampling – siehe für beide im Rahmen einer allgemeinen Einführung in den Umgang mit imbalanced data beispielsweise Fernández et al. (2018) oder He und Ma (2013) –, SMOTE (Chawla et al. 2002) und ROSE (Lunardon et al. 2014).

Upsampling gleicht die beiden Klassen in ihrem Vorkommen in den Trainingsdaten dadurch aus, dass der Minderheitenklasse zugehörige Einheiten unter Einsatz

eines Zufallsmechanismus vervielfacht werden. Downsampling erreicht das gleiche Ziel durch eine zufällige Auswahl von der Mehrheitsklasse zugehörigen Einheiten. Sowohl Up- als auch Downsampling berücksichtigen keine über die Klassenzugehörigkeit hinausgehenden Informationen des Datensatzes, insbesondere keine Zusammenhänge zwischen den im Datensatz enthaltenen Einheiten. SMOTE und ROSE hingegen leisten dies. SMOTE erzeugt synthetische Einheiten mittels Interpolation von nahe beieinander liegenden realen Einheiten der Minderheitenklasse (zugrundeliegend ist ein Nächste-Nachbarn-Ansatz). ROSE erzeugt synthetische Einheiten auf Basis von auf die Klassenzugehörigkeit bedingten Kerndichteschätzungen der Trainingsdaten und stellt insofern eine Weiterentwicklung von SMOTE dar.

In der vorliegenden Untersuchung wurden nur Anpassungen der Trainingsdaten – sowie später die Variation des Schwellenwertes bei der Klassifikation – vorgenommen. Spezielle Algorithmen oder über die Variation des Schwellenwertes hinausgehendes explizites kostensensitives Lernen – vgl. Kap. 4 und 6 in Fernández et al. (2018) – wurden nicht untersucht.

4.4 Einbezogene Zusatzinformationen

Brauchbare statistische (maschinelle Lern-) Verfahren besitzen die Eigenschaft, mit zunehmendem Umfang an neue Information tragendem Datenmaterial den zugrundeliegenden datengenerierenden Prozess, zumindest den funktionalen Zusammenhang zwischen erklärenden Variablen und zu erklärender Variable besser zu schätzen. In der Regel handelt es sich bei dieser Eigenschaft um asymptotische Aussagen (d. h. Aussagen für den Grenzwert, dass unendlich viele Trainingsdaten zur Verfügung stehen). In der Praxis liegt diese Situation natürlich nicht vor. Oben genannte Stichprobenumfänge (490.707 Vollzeitbeschäftigte, 162.646 Teilzeitbeschäftigte und 110.010 geringfügig entlohnte Beschäftigte) sind groß, aber endlich; es gibt mithin keine Garantie, dass die statistischen Verfahren bereits alle in den Daten erhaltenen Informationen hinreichend gut erfassen können. Es lohnt sich deshalb, den Verfahren bei komplexeren Zusammenhängen „auf die Sprünge zu helfen“, sie also durch die Vorgabe eines Zusammenhangs zu unterstützen. Dies wurde im vorliegenden Fall auch getan und zwar in Form einer Zusatzinformation derart, dass allen Beschäftigten einer Schicht der Stichprobe (WZ-2-Steller × Bundesland × Größenklasse des Betriebs) als zusätzliche erklärende Variable beigegeben wurde, wie hoch der Anteil der vom Mindestlohn betroffenen Beschäftigten an allen Beschäftigten der jeweiligen Schicht in der VSE ist. Der jeweils einschlägige Anteil wurde aus dem Gesamtbestand der VSE geschätzt und den Beobachtungen aller Teildatensätze (siehe Abschn. 4.5) zugespielt. Auch steht diese Information – wie eine externe Datenquelle – mitgeliefert zur Verfügung, wenn ein Modell beispielsweise auf den IEB zum Einsatz kommt. Die Ergebnisse werden in Kap. 7 präsentiert.

4.5 Aufteilung der Datensätze

Wie insbesondere im Bereich des maschinellen Lernens üblich, erfolgte in allen drei Gruppen (VZ, TZ, GF) jeweils eine zufällige Einteilung der vorhandenen Beobachtungen in Trainingsdaten, Validierungsdaten und Testdaten. In allen drei Fällen

wurden 45.000 Beobachtungen als Testdatensatz zurückgelegt (siehe hierzu auch Abschn. 4.1 und Kap. 9). Die jeweils verbleibenden Fälle – die „Übungsdaten“ – wurden einem zehnfachen Subsampling-Verfahren im Verhältnis von 70 % zu 30 % unterworfen, d. h. zehnmal wurden 70 % der Übungsdaten zufällig und ohne Zurücklegen dem Trainingsdatensatz zugeordnet, die verbleibenden 30 % dem korrespondierenden Validierungsdatensatz. Es gab mithin (für alle drei Gruppen VZ, TZ und GF) jeweils zehn Trainings- und zehn Validierungsdatensätze. Aufgrund des Konstrukts des Subsamplingansatzes sind dabei zwar die jeweils zusammengehörigen Trainings- und Validierungsdatensätze disjunkt, nicht jedoch die Trainingsdatensätze untereinander oder die Validierungsdatensätze untereinander. Auf den Trainingsdatensätzen wurden dann die einzelnen statistischen (maschinellen Lern-) Verfahren in den jeweiligen Kombinationen (siehe Tab. 2) trainiert, d. h. die prädiktiven Modelle gelernt. Auf den korrespondierenden Validierungsdatensätzen wurden die prädiktiven Modelle anschließend angewendet und die so geleistete Klassifikation der Beobachtungen anhand verschiedener Gütemaße evaluiert. Somit stehen je oben genannter Kombination zehn Einschätzungen der Klassifikationsgüte zur Verfügung, mithin nicht nur eine Punktschätzung, sondern auch eine Aussage zur Streuung und damit zur Stabilität einer Kombination.

Zum Vorgehen ist des Weiteren zu bemerken, dass eine große Zahl an zurückgelegten Datenpunkten für das Testen (im vorliegenden Fall 45.000) potenziell zu Lasten der Modellbildung geht, da dadurch der Umfang der „Übungsdaten“ (und somit der Umfang der Trainingsdaten) kleiner ausfällt. Sie soll jedoch beim Testen verlässliche Schätzungen für die Güte eines Modells sicherstellen. Darüber hinaus zeigten Voruntersuchungen – insbesondere auch bei den Teilzeit- und den geringfügig entlohnten Beschäftigten –, dass eine Verminderung der Testdatensätze auf 25.000 Beobachtungen trotz korrespondierender Zunahme der „Übungsdaten“ keine deutliche Verbesserung der Modelle auf den Validierungsdatensätzen mit sich bringt.

4.6 Eingesetzte Gütemaße

Um die auf den Trainingsdaten gelernten binären Klassifikationsmodelle hinsichtlich ihrer Güte bewerten zu können, werden diese auf die Validierungsdaten (und später auch auf die Testdaten) angewendet. Die Validierungsdaten enthalten allerdings bereits die Information, ob ein Beschäftigter vom Mindestlohn betroffen ist oder nicht. Somit stehen für die Validierungsdaten sowohl die Einschätzung des Modells als auch die angenommenermaßen wahre Information selbst zur Verfügung. Das ermöglicht einen Vergleich von geschätzter und wahrer Einstufung bezüglich der Mindestlohnbetroffenheit und erlaubt vier Fälle: (1) Schätzung und Wahrheit stimmen darin überein, dass ein Beschäftigter vom Mindestlohn betroffen ist („true positive“, TP); (2) Schätzung und Wahrheit stimmen darin überein, dass ein Beschäftigter nicht vom Mindestlohn betroffen ist („true negative“, TN); (3) das Modell schätzt irrtümlich, dass ein Beschäftigter vom Mindestlohn betroffen ist („false positive“, FP); (4) das Modell schätzt irrtümlich, dass ein Beschäftigter nicht vom Mindestlohn betroffen ist („false negative“, FN). Aus diesen vier Anzahlen (TP, TN, FP, FN) lassen sich nun die Kennzahlen bestimmen, die für eine Einschätzung der Güte der Modelle herangezogen werden sollen. Positiv und negativ stellen dabei keine

Wertungen dar; es handelt sich lediglich um Labels, um die eine von der anderen Klasse abgrenzen zu können.

Grundlegendes Gütemaß bei Klassifikation ist die *Accuracy*, also die Gesamtgenauigkeit, die ein Klassifikator erreicht. Die Gesamtgenauigkeit ist der Anteil aller korrekt klassifizierten Beschäftigten an allen klassifizierten Beschäftigten, also

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Selbstverständlich kann die Accuracy im Fall von imbalanced data irreführend sein, da schon ein naiver Klassifikator, der schlicht alle Beobachtungen der Mehrheitsklasse zuordnet – und damit inhaltlich sinnlos ist –, eine Accuracy in Höhe des Anteils der Mehrheitsklasse erzielt. Die Betrachtung weiterer Gütemaße ist daher angeraten.

Dies sind zunächst die *Sensitivität (Recall)* bzw. die *Spezifität*, also die Anteile der „korrekterweise als vom Mindestlohn betroffen klassifizierten Beschäftigten“ an „den vom Mindestlohn betroffenen Beschäftigten“,

$$Sensitivitaet = \frac{TP}{TP + FN},$$

bzw. den „korrekterweise als vom Mindestlohn nicht betroffen klassifizierten Beschäftigten“ an den „vom Mindestlohn nicht betroffenen Beschäftigten“,

$$Spezifitaet = \frac{TN}{TN + FP}.$$

Darüber hinaus ist zu prüfen, wie vertrauenswürdig eine Zuordnung zu einer der beiden Klassen ist. Dies leisten positiver *Vorhersagewert (Precision)* bzw. *negativer Vorhersagewert*, also die Anteile der „als korrekterweise vom Mindestlohn betroffen klassifizierten Beschäftigten“ an den „als vom Mindestlohn betroffen klassifizierten Beschäftigten“,

$$Vorhersagewert_pos = \frac{TP}{TP + FP},$$

bzw. „als korrekterweise vom Mindestlohn nicht betroffen klassifizierten Beschäftigten“ an den „als vom Mindestlohn nicht betroffen klassifizierten Beschäftigten“,

$$Vorhersagewert_neg = \frac{TN}{TN + FN}.$$

Auf welches dieser Gütemaße ein größeres Gewicht bei der Auswahl des Klassifikators gelegt werden soll, ist weniger eine statistische Frage als eher eine Entscheidung des diesen Klassifikator schließlich anwendenden Fachexperten. Accuracy, Sensitivität und positiver Vorhersagewert werden in den Kap. 5 und 6 zur vergleichenden Bewertung der alternativ gelernten Modelle herangezogen und in den entsprechenden Abbildungen dargestellt.

Ergänzend in Boxplots ausgewiesene Gütemaße sind die *Balanced Accuracy* als das arithmetische Mittel aus Sensitivität und Spezifität, also

$$BalancedAccuracy = \frac{Sensitivitaet + Spezifitaet}{2},$$

der *G-Wert* als das geometrische Mittel von Sensitivität und Spezifität, also

$$GValue = \sqrt{Sensitivitaet \cdot Spezifitaet},$$

die beiden *F-Maße* (eines je Klasse) als das harmonische Mittel der Sensitivität und des positiven Vorhersagewertes, also

$$F_{\text{pos}} = \frac{2 \cdot Sensitivitaet \cdot Vorhersagewert_pos}{Sensitivitaet + Vorhersagewert_pos},$$

bzw. als das harmonische Mittel der Spezifität und des negativen Vorhersagewertes, also

$$F_{\text{neg}} = \frac{2 \cdot Spezifitaet \cdot Vorhersagewert_neg}{Spezifitaet + Vorhersagewert_neg},$$

sowie der *Kappa-Wert* als Maß der zufallskorrigierten Übereinstimmung der wahren Klassenzugehörigkeiten und der Vorhersagen durch den Klassifikator (Landis und Koch 1977).

Für alle genannten Gütemaße gilt: Je näher an 1, desto besser ist die Leistung des Klassifikators auf dem Validierungsdatensatz. Beim Kappa-Wert stufen Landis und Koch (1977) Werte zwischen 0,21 und 0,40 als „fair“, Werte zwischen 0,41 und 0,60 als „moderate“, Werte zwischen 0,61 und 0,8 als „substantial“ und Werte zwischen 0,81 und 1,00 als „almost perfect“ ein.

Eine Übersicht über gängige Gütemaße bei binärer Klassifikation unter besonderer Berücksichtigung von imbalanced data bietet Luque et al. (2019).

Bei den Auswertungen wurden die Beobachtungen des jeweiligen Validierungsdatensatzes nicht einfach durchgezählt, sondern mit ihren Hochrechnungsfaktoren aus der Verdienststrukturerhebung gewichtet. Anders als der Ausweis der ebenfalls bestimmten ungewichteten Gütemaße erlaubt dieses Vorgehen eine realistische Einschätzung der Güte eines Klassifikators jenseits der VSE, die ja eine Stichprobe von Beschäftigungsverhältnissen ist.

5 Ergebnisse

5.1 Ergebnisse für die Vollzeitbeschäftigten

Bei den Vollzeitbeschäftigten liegt die Benchmark hinsichtlich der zu erreichenden Gesamtgenauigkeit (Accuracy) bei 90,1 % (das entspricht dem hochgerechneten Anteil an vom Mindestlohn nicht betroffenen Vollzeitbeschäftigten). Das Modell mit

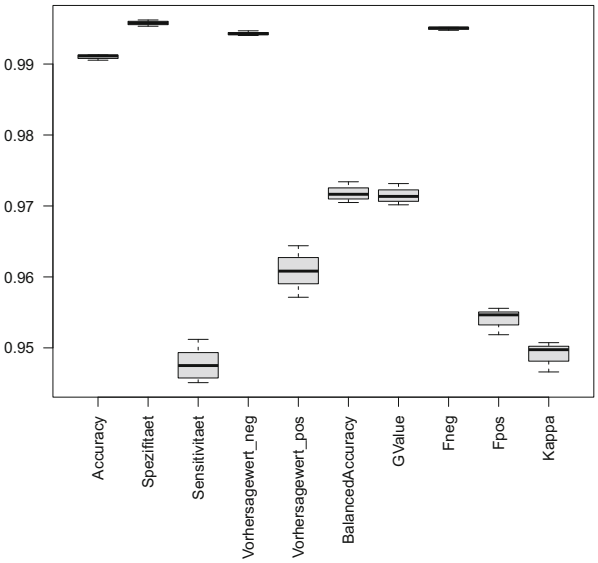


Abb. 1 *Gütemaße eines Boosting-Verfahrens für Vollzeitbeschäftigte.* Die Boxplots geben die Gütemaße des Modells C50_100_4 (Boosting-Verfahren mit maximal 100 Zyklen und minimal vier Beschäftigten in den Blättern der zugrunde liegenden Bäume) wieder. Die Zielgröße war dabei binär („vom Mindestlohn betroffen“ (pos), „vom Mindestlohn nicht betroffen“ (neg)), der Schwellenwert betrug 0,5, es wurde kein besonderes Samplingverfahren zur Kompensation der imbalanced-data-Situation herangezogen und die Zusatzinformation über den Mindestlohnanteil in der Schicht (vgl. Abschn. 4.4) wurde als zusätzliche erklärende Variable genutzt. Die Boxplots entstehen dabei, weil das Modell nicht nur einmal gelernt und validiert wurde sondern insgesamt zehnmal gemäß dem in Abschn. 4.5 beschriebenen Subsampling-Ansatz. Je kleiner der Boxplot, desto stabiler verhält sich das zugehörige Gütemaß über die zehn Subsamples

der höchsten mittleren Genauigkeit (ein Boosting-Verfahren mit maximal 100 Zyklen und einer Mindestanzahl von vier Datenpunkten in den Blättern der zugrundeliegenden Bäume, Schwellenwert 0,5, kein besonderes Sampling, Nutzung der Zusatzinformation) erreicht eine Accuracy von 99,1 % bei einer Sensitivität von 94,8 % und einem positiven Vorhersagewert von 96,1 % (siehe Abb. 1).

Die Ergebnisse für andere Verfahren und Samplingansätze sind vergleichbar (siehe Abb. 2).

5.2 Ergebnisse für die Teilzeitbeschäftigten

Ähnlich wie bei den Vollzeitbeschäftigten liegt auch bei den Teilzeitbeschäftigten ein Ungleichgewicht der Klassen vor: nur ca. 11 % der Teilzeitbeschäftigten haben einen Bruttostundenverdienst von unter 8,50 € und sind somit vom Mindestlohn betroffen. Das hier beste Modell bzgl. der mittleren Accuracy ist wiederum ein Boosting-Ansatz (maximal 100 Zyklen, eine Mindestanzahl von 50 Datenpunkten in den Blättern der zugrundeliegenden Bäume, Schwellenwert 0,7, upsampling, mit Zusatzinformation) und erreicht im Mittel eine Genauigkeit von 92,4 %, eine Sensitivität von 48,8 % und einen positiven Vorhersagewert von 71,1 % (Abb. 3).

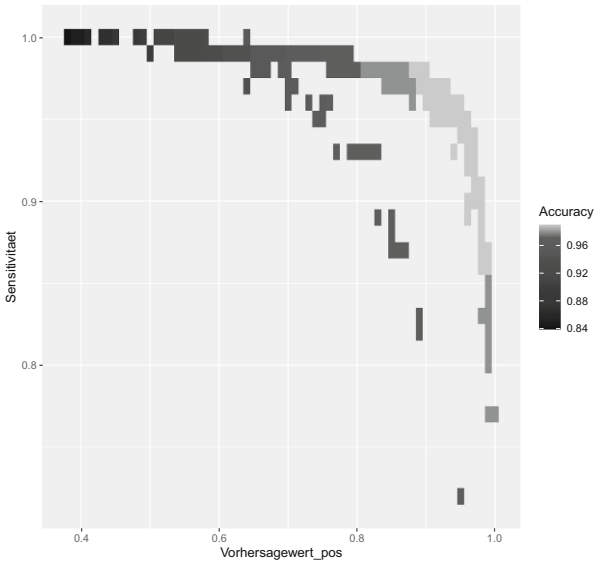


Abb. 2 *Spektrum der Ergebnisse für Vollzeitbeschäftigte.* Die Abbildung zeigt die drei wichtigsten Gütemaße für die Vollzeitbeschäftigten: den positiven Vorhersagewert für „vom Mindestlohn betroffen" auf der Abszisse, die Sensitivität für „vom Mindestlohn betroffen" auf der Ordinate und die Accuracy farblich codiert. Die Rechtecke stellen dabei – hier wie auch in allen Abbildungen gleichen Typs – in Graustufen und Koordinate Mittelwerte der Gütemaße über die zehn Subsamples (siehe Abschn. 4.5) dar; ein ideales Modell befände sich oben rechts, also bei (1 | 1), und wäre durch ein *helles Rechteck* repräsentiert. Zu erkennen ist, dass sehr viele der gelernten Modelle leistungsfähig bzgl. der drei wichtigsten Gütemaße sind und dass somit für die Gruppe der Vollzeitbeschäftigten kein Zielkonflikt zwischen positivem Vorhersagewert und Sensitivität besteht. Potenzielle Nutzer müssen also keinen Kompromiss bezüglich dieser Gütemaße eingehen

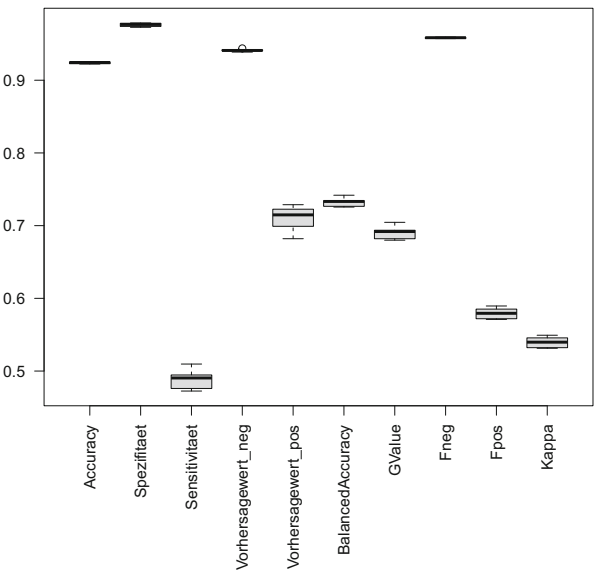


Abb. 3 *Gütemaße eines Boosting-Verfahrens für Teilzeitbeschäftigte.* Die Boxplots geben die Gütemaße des Modells C50_100 (Boosting-Verfahren mit maximal 100 Zyklen und minimal 50 Beschäftigten in den Blättern der zugrunde liegenden Bäume) wieder. Der Schwellenwert betrug 0,7, es wurde ein Upsamplingverfahren zur Kompensation der imbalanced-data-Situation herangezogen und die Zusatzinformation über den Mindestlohnanteil in der Schicht (vgl. Abschn. 4.4) wurde als zusätzliche erklärende Variable genutzt

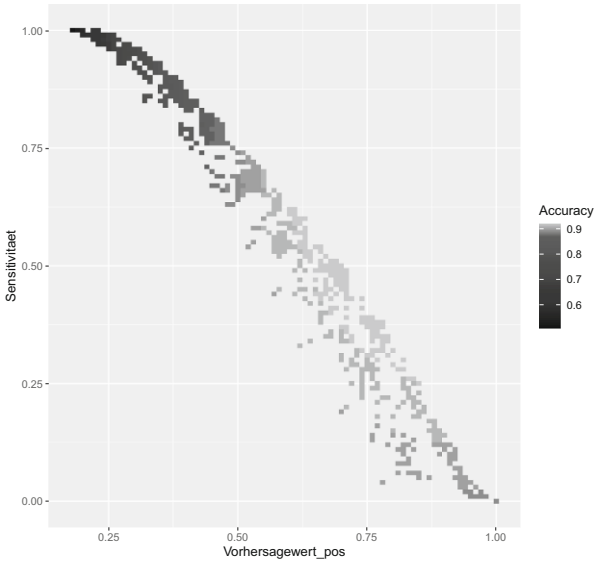


Abb. 4 *Spektrum der Ergebnisse für Teilzeitbeschäftigte.* Die Abbildung zeigt die drei wichtigsten Gütemaße für die Teilzeitbeschäftigten: den positiven Vorhersagewert für „vom Mindestlohn betroffen" auf der Abszisse, die Sensitivität für „vom Mindestlohn betroffen" auf der Ordinate und die Accuracy durch Graustufen codiert. Deutlich zu erkennen ist der Zielkonflikt insbesondere zwischen positivem Vorhersagewert und Sensitivität: Je weiter oben ein Modell in dieser Abbildung zu finden ist (hohe Sensitivität), desto weiter links ist es (geringer positiver Vorhersagewert) und umgekehrt. Die Skalierung der Achsen zeigt, dass der Zielkonflikt bei den Teilzeitbeschäftigten deutlich ausgeprägter ist als bei den Vollzeitbeschäftigten. Potenzielle Nutzer müssen bei der Auswahl eines oder mehrerer Modelle für ihre Zwecke also einen Kompromiss eingehen und können sich zu diesem Zweck an dieser Abbildung orientieren, da sie veranschaulicht, wie viel positiver Vorhersagewert geopfert werden muss, um eine höhere Sensitivität zu erlangen

Im Vergleich zu den Vollzeitbeschäftigten sind die Resultate bei den Teilzeitbeschäftigten weniger gut und weniger eindeutig. Sie lassen Raum für Variationen im Hinblick auf den Zielkonflikt zwischen den Größen Accuracy, Sensitivität und positiver Vorhersagewert (siehe Abb. 4). Für die weitere Diskussion der Ergebnisse sei auf Kap. 10 verwiesen.

5.3 Ergebnisse für die geringfügig entlohnten Beschäftigten

Anders als bei den Teilzeit- und Vollzeitbeschäftigten sind die Klassen nicht ganz so ungleich verteilt: ca. 40 % der geringfügig entlohnten Beschäftigten haben einen Bruttostundenverdienst von unter 8,50 € und sind somit vom Mindestlohn betroffen. Das hier beste Modell bzgl. der mittleren Accuracy ist ein Random-Forest-Ansatz (mtry = 3, Mindestanzahl von vier Datenpunkten in den Blättern der zugrundeliegenden Bäume, Schwellenwert 0,5, kein besonderes Sampling) und erreicht im Mittel eine Genauigkeit von 76,7 %, eine Sensitivität von 66,1 % und einen positiven Vorhersagewert von 72,3 % (Abb. 5).

Im Vergleich zu den Vollzeitbeschäftigten (und ähnlich wie bei den Teilzeitbeschäftigten) sind die Resultate bei den geringfügig entlohnten Beschäftigten weniger

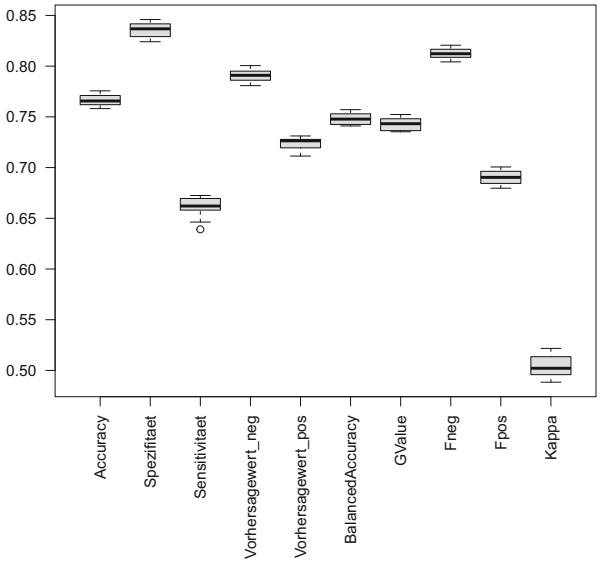


Abb. 5 *Gütemaße eines Random Forest für geringfügig entlohnte Beschäftigte.* Die Boxplots geben die Gütemaße des Modells ranger4sqrt (Random Forest mit mtry = 3 und minimal vier Beschäftigten in den Blättern der zugrunde liegenden Bäume) wieder. Der Schwellenwert betrug 0,5, es wurde kein besonderes Verfahren zur Kompensation der imbalanced-data-Situation herangezogen und die Zusatzinformation über den Mindestlohnanteil in der Schicht (vgl. Abschn. 4.4) wurde als zusätzliche erklärende Variable genutzt

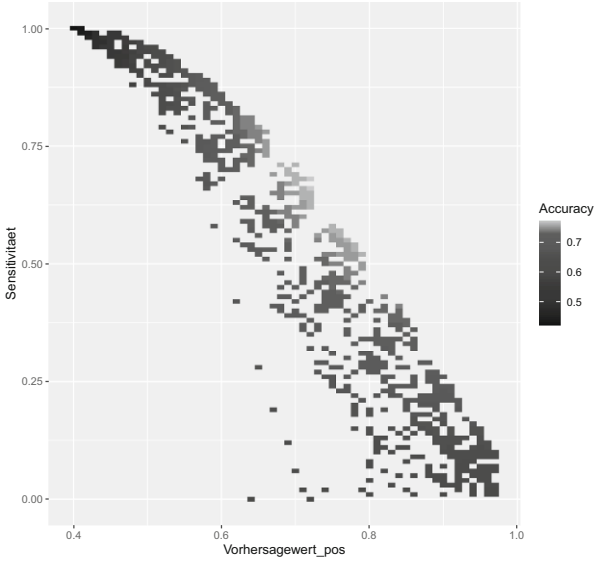


Abb. 6 *Spektrum der Ergebnisse für geringfügig entlohnte Beschäftigte.* Die Abbildung zeigt die drei wichtigsten Gütemaße für die geringfügig entlohnten Beschäftigten: den positiven Vorhersagewert für „vom Mindestlohn betroffen" auf der Abszisse, die Sensitivität für „vom Mindestlohn betroffen" auf der Ordinate und die Accuracy durch Graustufen codiert. Deutlich zu erkennen ist auch hier der Zielkonflikt insbesondere zwischen positivem Vorhersagewert und Sensitivität: Je weiter oben ein Modell in dieser Abbildung zu finden ist (hohe Sensitivität), desto weiter links ist es (geringer positiver Vorhersagewert) und umgekehrt. Die Skalierung der Achsen zeigt, dass der Zielkonflikt bei den geringfügig entlohnten Beschäftigten (ebenso wie bei den Teilzeitbeschäftigten) deutlich ausgeprägter ist als bei den Vollzeitbeschäftigten. Potenzielle Nutzer müssen bei der Auswahl eines oder mehrerer Modelle für ihre Zwecke also einen Kompromiss eingehen und können sich zu diesem Zweck an dieser Abbildung orientieren, da sie veranschaulicht, wie viel positiver Vorhersagewert geopfert werden muss, um eine höhere Sensitivität zu erlangen

gut und weniger eindeutig. Sie lassen Raum für Variationen im Hinblick auf den Zielkonflikt zwischen den Größen Accuracy, Sensitivität und positiver Vorhersagewert (siehe Abb. 6). Für die weitere Diskussion der Ergebnisse sei auf Kap. 10 verwiesen.

6 Betrachtung von weniger kritischen Missklassifikationen

Sowohl bei den Teilzeitbeschäftigten als auch bei den geringfügig entlohnten Beschäftigten tritt ein Zielkonflikt zwischen positivem Vorhersagewert und Sensitivität auf. Diesem Zielkonflikt ist zunächst natürlich inhaltlich zu begegnen, indem ein Modell gewählt wird, das beispielsweise Mindestanforderungen an Gesamtgenauigkeit und Sensitivität stellt und anschließend bzgl. des positiven Vorhersagewertes optimal ist. Aus Arbeiten des Statistischen Bundesamtes im Nachgang der Einführung des gesetzlichen Mindestlohns können jedoch zusätzliche Erkenntnisse gezogen werden. Die Verdiensterhebung 2015 lieferte nämlich folgende Einsichten in die Wirkungen des zum 1. Januar 2015 eingeführten allgemeinen gesetzlichen Mindestlohns: (1) Der Mindestlohn wirkt deutlich am unteren Rand der Verteilung der Bruttostundenverdienste; (2) der Mindestlohn hat die Verdienstverteilung mutmaßlich nur im Bereich bis etwa 10 € beeinflusst (Frentzen und Günther 2017).

Natürlich sind im Rahmen der vorliegenden Untersuchung vornehmlich Beschäftigte von Interesse, deren Bruttostundenverdienst vor Einführung des Mindestlohns unter 8,50 € lag. Beschäftigte mit einem Bruttostundenverdienst von mehr als 8,50 € sollten idealerweise nicht als „vom Mindestlohn betroffen“ klassifiziert werden. Ausweislich der Ergebnisse der Verdiensterhebung 2015 sind jedoch auch Beschäftigungsverhältnisse mit einem Bruttostundenverdienst zwischen 8,50 und 10,00 € indirekt vom Mindestlohn betroffen (Spillover-Effekte). Deren (falsche) Zuordnung zur Klasse „vom Mindestlohn betroffen“ ist für Analysen der Mindestlohnwirkungen dadurch weniger kritisch als Missklassifikationen (genauer: false positives), bei welchen der tatsächliche Bruttostundenverdienst des Beschäftigten sogar oberhalb von 10 € liegt. Wie hoch der Anteil der in diesem Sinne weniger problematischen Missklassifikationen an allen irrtümlich als „vom Mindestlohn betroffen“ klassifizierten Beschäftigten ist, wurde daher eigens untersucht (siehe Abb. 7 und 8).

Unter Berücksichtigung der so herausgearbeiteten Zielkonflikte zwischen Accuracy, Sensitivität, positivem Vorhersagewert und Anteil der weniger kritischen Missklassifikationen ist es nun möglich, einen Kompromiss zu finden, indem Accuracy, Sensitivität und positiver Vorhersagewert auf Mindestwerte gesetzt werden und dann der Anteil der weniger kritischen Missklassifikationen maximiert wird.

Im Fall der Teilzeitbeschäftigten könnte dieser Kompromiss – rein exemplarisch – lauten: Accuracy mindestens 90 %, Sensitivität mindestens 60 % („mindestens drei von fünf vom Mindestlohn betroffenen Teilzeitbeschäftigten werden gefunden“), positiver Vorhersagewert mindestens 60 % („in mindestens drei von fünf Fällen ist die Klassifikation als „vom Mindestlohn betroffen“ korrekt) und möglichst hoher Anteil weniger kritischer Missklassifikationen. Ein solches Modell, ein Boosting-

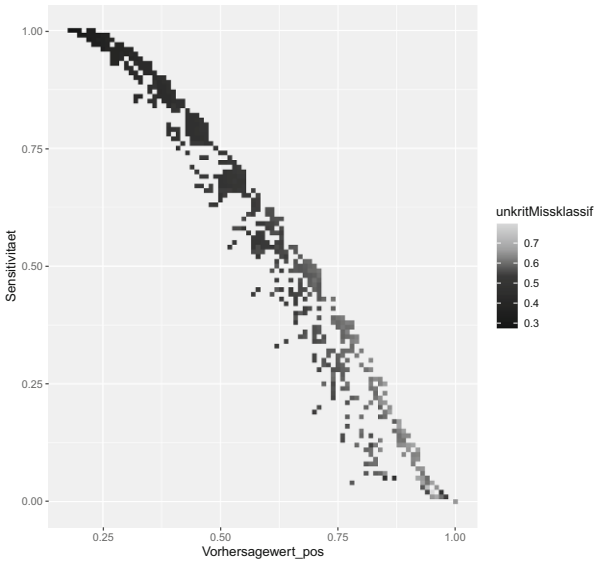
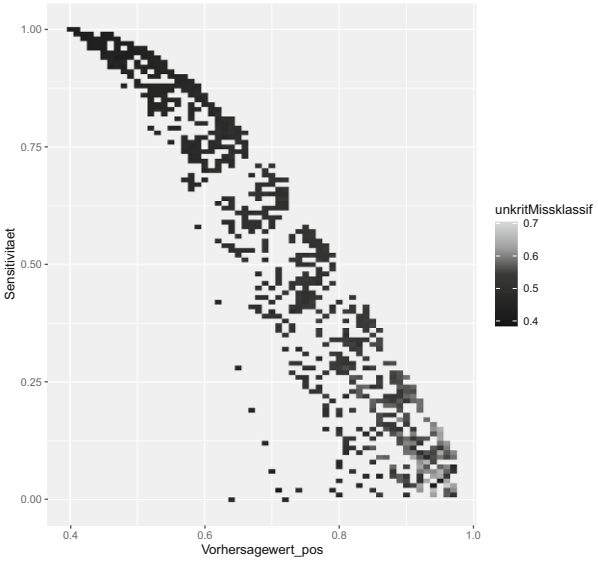


Abb. 7 *Spektrum der Ergebnisse für Teilzeitbeschäftigte unter besonderer Berücksichtigung der unkritischen Missklassifikationen.* Die Abbildung zeigt zwei Gütemaße für die Teilzeitbeschäftigten (den Vorhersagewert für „vom Mindestlohn betroffen“ auf der Abszisse, die Sensitivität für „vom Mindestlohn betroffen“ auf der Ordinate) sowie durch Graustufen codiert den Anteil derjenigen Beschäftigten im Validierungsdatensatz, die durch das Modell irrtümlicherweise der Klasse „vom Mindestlohn betroffen“ zugewiesen wurden, gleichzeitig aber einen Bruttostundenverdienst von weniger als 10 € hatten, an allen irrtümlich der Klasse „vom Mindestlohn betroffen“ zugewiesenen Beschäftigten, mithin den Anteil der weniger kritischen Missklassifikationen. Mit abnehmendem positiven Vorhersagewert (weiter links) sinkt tendenziell auch der Anteil der weniger kritischen Missklassifikationen (*dunklere Rechtecke*); je höher also der positive Vorhersagewert, desto höher der Anteil der weniger kritischen Missklassifikationen

Abb. 8 *Spektrum der Ergebnisse für geringfügig entlohnte Beschäftigte unter besonderer Berücksichtigung der unkritischen Missklassifikationen.* Auch im Fall der geringfügig entlohnten Beschäftigten zeigt sich, dass mit abnehmendem positiven Vorhersagewert (weiter links) tendenziell auch der Anteil der weniger kritischen Missklassifikationen sinkt (*dunklere Rechtecke*); je höher also der positive Vorhersagewert, desto höher der Anteil der weniger kritischen Missklassifikationen



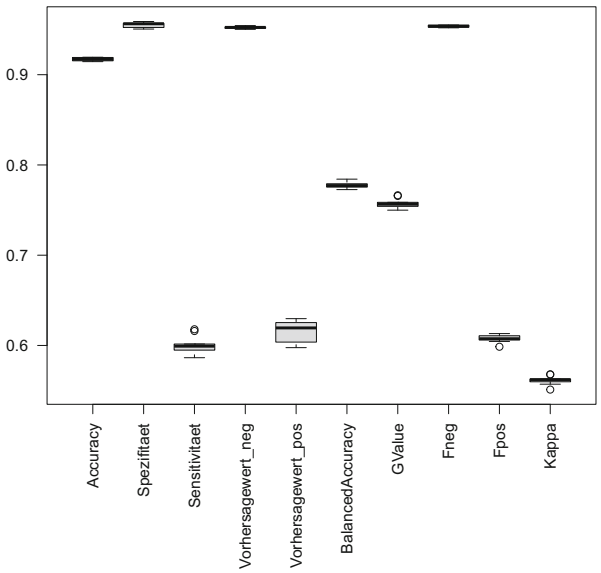


Abb. 9 *Gütemaße eines Boosting-Verfahrens für Teilzeitbeschäftigte.* Die Boxplots geben die Gütemaße des Modells C50_40_4 (Boosting-Verfahren mit maximal 40 Zyklen und minimal vier Beschäftigten in den Blättern der zugrunde liegenden Bäume) wieder. Der Schwellenwert betrug 0,5, es wurde ein Upsampling-Verfahren zur Kompensation der imbalanced-data-Situation herangezogen und die Zusatzinformation über den Mindestlohnanteil in der Schicht (vgl. Abschn. 4.4) wurde als zusätzliche erklärende Variable genutzt

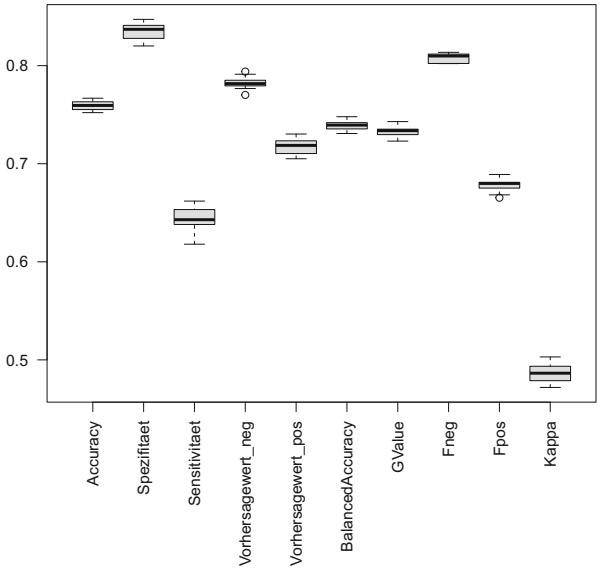


Abb. 10 *Gütemaße eines Boosting-Verfahrens für geringfügig entlohnte Beschäftigte.* Die Boxplots geben die Gütemaße des Modells C50_100 (Boosting-Verfahren mit maximal 100 Zyklen und minimal 50 Beschäftigten in den Blättern der zugrunde liegenden Bäume) wieder. Der Schwellenwert betrug 0,5, es wurde kein besonderes Verfahren zur Kompensation der imbalanced-data-Situation herangezogen und die Zusatzinformation über den Mindestlohnanteil in der Schicht (vgl. Abschn. 4.4) wurde als zusätzliche erklärende Variable genutzt

Verfahren, existiert (siehe Abb. 9) und bietet einen Anteil von 56,2 % an weniger kritischen Missklassifikationen in obigem Sinne.

Eine analoge Suche nach geeigneten Modellen liefert bei den geringfügig entlohnten Beschäftigten einen Anteil weniger kritischer Missklassifikationen in Höhe von 52,5 % (für ein Boosting-Verfahren, siehe Abb. 10).

7 Bedeutung des Mindestlohnanteils je Schicht

Die folgende Auswertung soll einen oben bereits genannten Aspekt noch einmal näher beleuchten, nämlich den, dass die explizite Bereitstellung der Zusatzinformation zum durchschnittlichen Anteil der vom Mindestlohn betroffenen Beschäftigten einer Schicht (WZ-2-Steller × Bundesland × Größenklasse des Betriebs) die Klassifikation tatsächlich verbessert. Alle Komponenten dieser zusätzlichen Variable „ml_anteil" sind bereits in der Verdienststrukturerhebung enthalten und könnten somit vom statistischen (maschinellen Lern-) Verfahren selbst auch in ihrer Interaktion verarbeitet werden, wenn dafür hinreichend viele Datenpunkte zur Verfügung stünden. Dass Letzteres nicht der Fall ist, zeigen die Abb. 11 für die Teilzeitbeschäftigten und 12 für die geringfügig entlohnten Beschäftigten, jeweils für den Schwellenwert 0,5 und für alle erprobten Methoden. Hier nicht abgedruckte Abbildungen für die Sensitivitäten und die positiven Vorhersagewerte liefern ebenfalls keine Evidenz gegen den Einbezug der Zusatzinformation. Für die Teilzeitbeschäftigten sind die Resultate – bezugnehmend auf oben skizzierte Modellauswahlen – unter Verwendung eines Upsamplingansatzes dargestellt, für die geringfügig entlohnten Beschäftigten ohne besonderes Samplingverfahren. Als Gesamtbewertung ist festzuhalten, dass sich die explizite Einbeziehung der Zusatzinformation lohnt, da sich das Klassifikationsergebnis bzgl. der Gütemaße Accuracy, Sensitivität und positiver Vorhersagewert dadurch in fast allen Fällen verbessert und besonders bei den geringfügig entlohnten Beschäftigten darüber hinaus auch stabiler wird.

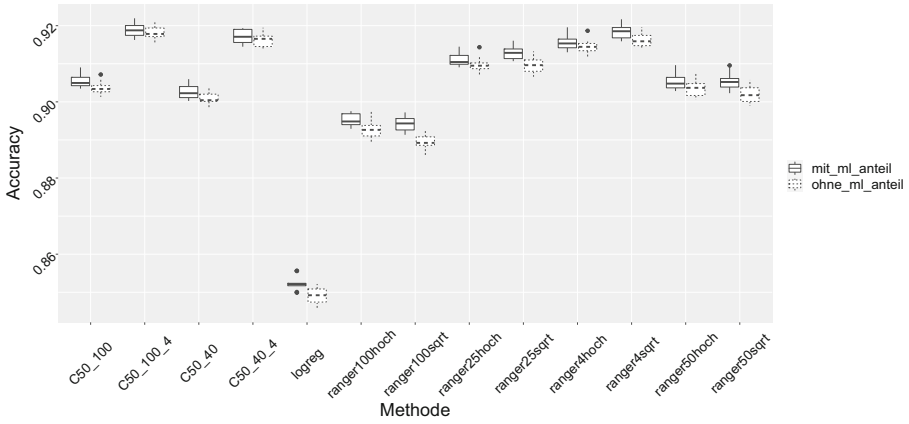


Abb. 11 *Einfluss der expliziten Einbeziehung des Mindestlohnanteils je Schicht auf die Accuracy für Teilzeitbeschäftigte.* Die Abbildung zeigt die Accuracy für die Gruppe der Teilzeitbeschäftigten aufgeschlüsselt nach den verschiedenen erprobten statistischen (maschinellen Lern-) Verfahren sowie bezüglich der Unterscheidung, ob die Zusatzinformation (der Mindestlohnanteil in der jeweiligen Schicht) explizit als erklärenden Variable aufgenommen wurde oder nicht. Es zeigt sich stets eine Verbesserung bei explizitem Einbezug der Zusatzinformation, teils sind die Boxplots sogar deutlich gegeneinander verschoben

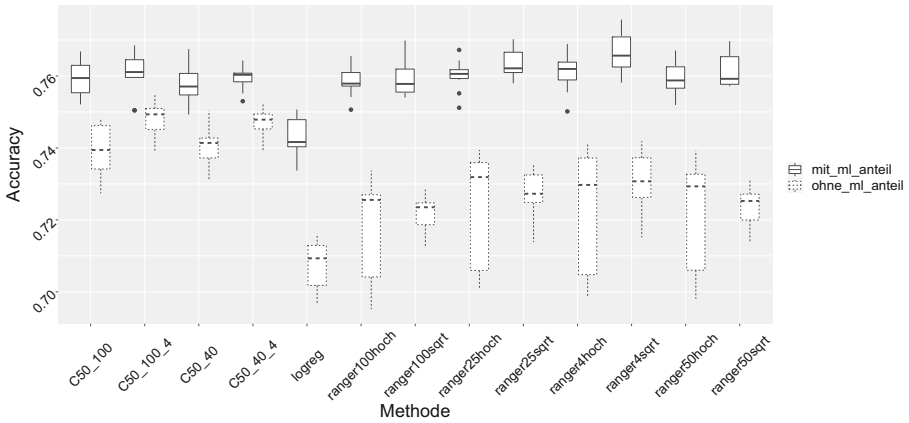


Abb. 12 *Einfluss der expliziten Einbeziehung des Mindestlohnanteils je Schicht auf die Accuracy für geringfügig entlohnte Beschäftigte.* Die Abbildung zeigt analog die Accuracy für die Gruppe der geringfügig entlohnten Beschäftigten. Stets zeigt sich hier eine Verbesserung bei explizitem Einbezug der Zusatzinformation, fast immer sind die Boxplots sogar deutlich gegeneinander verschoben. Darüber hinaus ist in der Gruppe der geringfügig entlohnten Beschäftigten häufig eine größere Streuung der Resultate zu erkennen, wenn die Zusatzinformation nicht explizit als erklärende Variable eingeht, was erneut für das Explizitmachen spricht

8 Performanz in Abhängigkeit vom konkreten Mindestlohn

Die Klassifikation in „vom Mindestlohn betroffen“ und „vom Mindestlohn nicht betroffen“ wurde in vorliegender Untersuchung anhand des Bruttostundenverdienstes und dort anhand des Werts 8,50 € vorgenommen. Eine weitere Sonderauswertung untersucht nun die Frage, ob sich das Klassifikationsproblem vereinfacht oder erschwert hätte, wenn ein anderer Trennwert herangezogen worden wäre, beispielsweise 8,00 € oder 9,00 €. Exemplarisch werden die Resultate dieser Untersuchung hier für Teilzeitbeschäftigte unter Verwendung eines Upsampling-Ansatzes und für den Schwellenwert 0,5 sowie unter Einbeziehung der Zusatzinformation gezeigt. Für alle eingesetzten statistischen (maschinellen Lern-) Verfahren zeigt sich bezüglich der Accuracy das gleiche monotone Muster: Je geringer der Wert, der die beiden Klassen „vom Mindestlohn betroffen“ und „vom Mindestlohn nicht betroffen“ trennt, desto besser (gemessen anhand der Accuracy) sind die Methoden in der Lage, die Trennung vorzunehmen (Abb. 13a). Dabei ist jedoch zu beachten, dass mit kleinerem klassentrennenden Wert das Ungleichgewicht der Klassen deutlich zunimmt. Die – hier nicht abgedruckten – Grafiken für andere Szenarien, insbesondere für andere Schwellenwerte (0,1; ...; 0,9) zeigen das gleiche Muster.

Eine analoge Auswertung für die geringfügig entlohnten Beschäftigten liefert vergleichbare Ergebnisse (siehe Abb. 14a, b).

Ein nützlicher Nebeneffekt dieser Sonderauswertung besteht darin, dass die Modelle für die Bruttostundenverdienste von 7,50 € (aus Gründen der Übersichtlichkeit in den Abb. 13a, b sowie 14a, b nicht explizit dargestellt) und 8,00 € ebenso trainiert wurden wie die für den Bruttostundenverdienst von 8,50 €. Es ist daher denkbar,

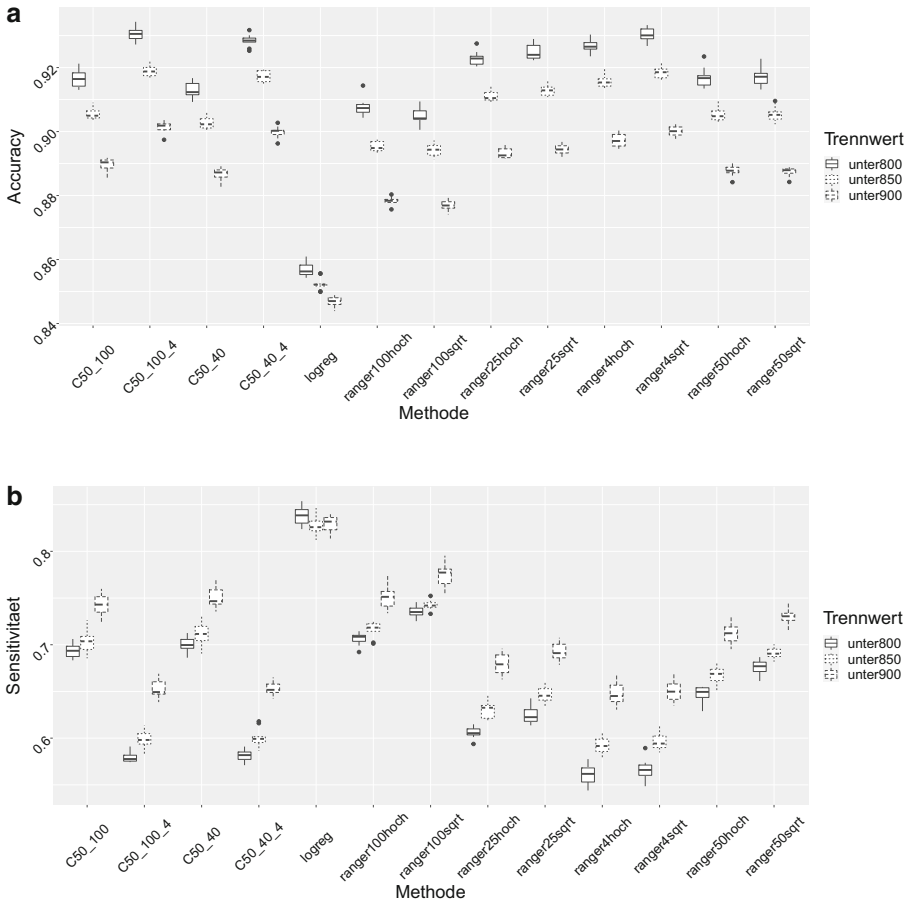


Abb. 13 **a** *Einfluss des Mindestlohn-Bruttostundenverdienstes auf die Accuracy für Teilzeitbeschäftigte.* **b** *Einfluss des Mindestlohn-Bruttostundenverdienstes auf die Sensitivität (für „vom Mindestlohn betroffen") für Teilzeitbeschäftigte.* Die Boxplots in **a** und **b** zeigen die Gütemaße Accuracy (**a**) und Sensitivität (**b**) sortiert nach den erprobten statistischen (maschinellen Lern-) Verfahren und in Abhängigkeit vom klassentrennenden Bruttostundenverdienst (8,00 €, 8,50 € und 9,00 €) für die Klassifikation von Teilzeitbeschäftigten, wobei ein Schwellenwert von 0,5 betrachtet und die Zusatzinformation zum Mindestlohnanteil sowie ein Upsampling-Verfahren genutzt wird

zusätzlich zum Modell für den Bruttostundenverdienst von 8,50 € noch ein weiteres Modell, z. B. jenes für 7,50 €, auf die vorliegenden Daten anzuwenden. Liefern beide Modelle das Ergebnis „vom Mindestlohn betroffen", sind damit „besonders vom Mindestlohn betroffene" Beschäftigte identifiziert, bei denen die Einführung des Mindestlohnes zu einem erheblichen Anstieg des Bruttostundenverdienstes führte. Diese Gruppe dürfte für die Wissenschaft von besonderer Bedeutung sein.

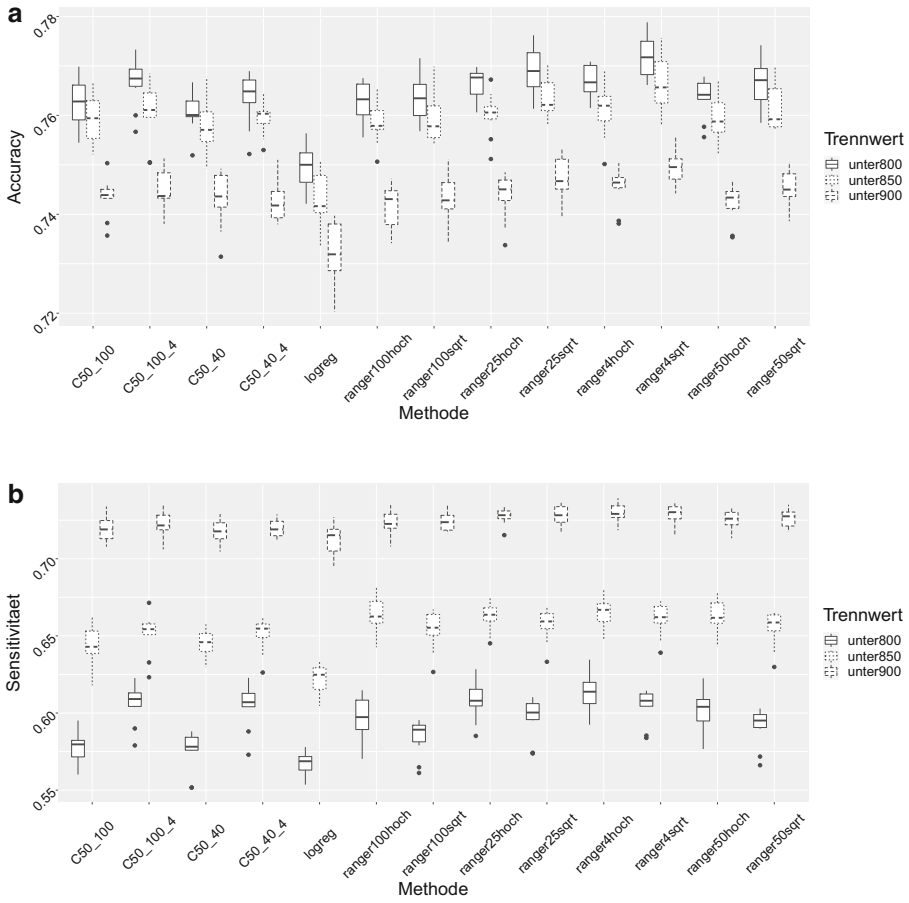


Abb. 14 **a** *Einfluss des Mindestlohn-Bruttostundenverdienstes auf die Accuracy für geringfügig entlohnte Beschäftigte.* **b** *Einfluss des Mindestlohn-Bruttostundenverdienstes auf die Sensitivität (für „vom Mindestlohn betroffen") für geringfügig entlohnte Beschäftigte.* Die Boxplots in **a** und **b** entsprechen bezüglich Inhalt und Interpretation denen der Abb. 13a und b, zeigen die Ergebnisse jedoch für die Gruppe der geringfügig entlohnten Beschäftigten

9 Bedeutung der zurückgelegten Testdaten

Wie in Abschn. 4.5 erläutert, steht für Vollzeit-, Teilzeit- und geringfügig entlohnte Beschäftigte noch jeweils ein Testdatensatz im Umfang von 45.000 Beschäftigten zur Verfügung. Damit ist es möglich, nach der Entscheidung für ein Modell auf Grundlage der in den vorangegangenen Kapiteln diskutierten Aspekte, die Güte des gewählten Modells durch den zukünftigen Nutzer noch einmal anhand eines bislang in der Untersuchung nicht genutzten Datensatzes unverzerrt und aufgrund des großen Umfangs der Testdaten auch sehr verlässlich einzuschätzen. Somit stehen einerseits die hier ausgewiesenen Ergebnisse für die Modellauswahl zur Verfügung, andererseits jedoch zusätzlich die bislang nicht besehenen Testdaten für eine davon

unabhängige Schätzung der Güte des gewählten Modells bzw. der gewählten Modelle. Vor Einsatz eines Modells im Echtbetrieb (z. B. auf den IEB) würde dieses – dann auf allen vorhandenen Daten – noch einmal trainiert.

10 Bewertung, Nutzungsvorschlag und weiterer Untersuchungsbedarf

Bewertung Vorliegende Untersuchung hat gezeigt, dass Modelle, die auf Basis des gemeinsamen Merkmalskranzes von Verdienststrukturerhebung des Statistischen Bundesamtes und Integrierten Erwerbsbiographien der Bundesagentur für Arbeit Beschäftigte der Klasse „vom Mindestlohn betroffen" bzw. „vom Mindestlohn nicht betroffen" (zum Zeitpunkt April 2014, also im Jahr vor der Einführung des gesetzlichen Mindestlohns in Deutschland) zuordnen, prinzipiell erlernbar sind. Für Vollzeit-, Teilzeit- und geringfügig entlohnte Beschäftigte ergaben sich dabei unterschiedliche Resultate. Während bei der Gruppe der Vollzeitbeschäftigten sehr gute Ergebnisse hinsichtlich der wichtigsten Gütemaße erzielt werden konnten, ist das Bild bei Teilzeitbeschäftigten und geringfügig entlohnten Beschäftigten weniger gut und weniger eindeutig. Erzielt wurden bei Letzteren Sensitivitäten und positive Vorhersagewerte deutlich unter denen, die bei der Gruppe der Vollzeitbeschäftigten erreicht werden konnten (vgl. Abb. 2, 4 und 6). Vor Anwendung von denkbaren Kompromissmodellen auf die IEB ist daher fachlich zu bewerten, ob Sensitivitäten und positive Vorhersagewerte in den erzielten Größenordnungen akzeptabel sind. Dass die Klassifikationsmodelle einen Mehrwert gegenüber rein zufälligen Zuweisungen der Mindestlohnbetroffenheit bieten, wird beispielsweise anhand der in diesem Aufsatz ebenfalls ausgewiesenen Kappa-Werte deutlich. Erkennbar wurde im Zuge der Untersuchungen aber der Zielkonflikt zwischen Sensitivität und positivem Vorhersagewert, das heißt der Konflikt zwischen dem Ziel, möglichst alle vom Mindestlohn betroffenen Beschäftigten als solche zu klassifizieren, und dem Ziel, bei dieser Klassifikation möglichst keine vom Mindestlohn nicht betroffenen Beschäftigten irrtümlich als „vom Mindestlohn betroffen" zu kennzeichnen. Dieser Zielkonflikt in Verbindung mit den Analysen zur Mindestlohnwirksamkeit bei Bruttostundenverdiensten von mehr als 8,50 € (siehe Kap. 6) bietet potenziellen Nutzern der gelernten Modelle im Gegenzug jedoch Spielraum für Schwerpunktsetzungen.

Nutzungsvorschlag Für Analysen könnte nun beispielsweise einmal ein Modell mit hohem positiven Vorhersagewert genutzt werden (zu dem Preis, dass weniger vom Mindestlohn betroffene Beschäftigte entdeckt werden) und einmal ein Modell mit hoher Sensitivität (zu dem Preis, dass auch einige Beschäftigte als vom Mindestlohn betroffen klassifiziert werden, die es tatsächlich jedoch gar nicht sind). Würde eine Analyse (z. B. zur Frage, wie viele Beschäftigte den Arbeitgeber infolge der Mindestlohneinführung wechselten, oder zur Frage, wie viele Beschäftigte bei Einführung des Mindestlohns arbeitslos wurden) also zweimal (einmal je Einsatz eines Klassifikationsmodells) durchgeführt, ergäbe sich ein Band möglicher Auswirkungen in Form einer Ober- und einer Untergrenze. Alternativ oder ergänzend kann aber auch ein Kompromissmodell gewählt werden, das vertretbare Gesamtgenauigkeit, Sensitivität und positiven Vorhersagewert aufweist. Die Herangehensweise, mehrere

Modelle anzuwenden und daraus Ober- und Untergrenzen für mögliche Auswirkungen ermitteln zu können, lässt sich verallgemeinern, indem für alle Nutzer der IEB mehrere Klassifikationsergebnisse je Beschäftigtem bezüglich der Mindestlohnbetroffenheit basierend auf der Anwendung mehrerer Modelle bereitgestellt werden. Die Metainformationen zu den zugrundeliegenden Modellen wären entsprechend zu hinterlegen. Der Wissenschaft bietet sich auf diese Weise die Gelegenheit, selbst zwischen verschiedenen bereitgestellten Modellvarianten wählen und die eigenen Resultate dementsprechend bewerten zu können. Werden die Ergebnisse mehrerer Modelle mit ähnlichen Resultaten im Hinblick auf die Gütemaße bereitgestellt, ergibt sich auf diese Weise außerdem eine Möglichkeit, die Unsicherheit der Klassifikationsergebnisse auch für die Wissenschaft transparent in den IEB abzubilden.

Weiterer Untersuchungsbedarf Ob, wann und wie die Resultate der vorliegenden Untersuchung in den IEB konkret umgesetzt werden können, bedarf weiterer Festlegungen und Untersuchungen. Beispielsweise ist die weitestgehende Übereinstimmung der erklärenden Variablen tiefgehend und auf Basis von Analyseergebnissen aus beiden Datenquellen (VSE und IEB) nochmals eingehend zu prüfen und sicherzustellen. Zwar wäre es darüber hinaus wünschenswert, auch die Klassifikationsergebnisse selbst, ganz im Sinne eines überwachten Lernansatzes, anhand der IEB prüfen zu können. Dies ist jedoch – strukturell bedingt – nicht direkt möglich. Wäre dies der Fall, hätte es der vorliegenden Untersuchung gar nicht erst bedurft.

Danksagung Die Autoren danken Georg Dautfest für seine Unterstützung im Projekt und bei der Erstellung des Aufsatzes.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf http://creativecommons.org/licenses/by/4.0/deed.de.

Literatur

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, S 144–152 https://doi.org/10.1145/130385.130401

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall/CRC, Boca Raton

Bundesministerium für Arbeit und Soziales (2020) (Hg.) Gesamtbericht zur Evaluation des allgemeinen gesetzlichen Mindestlohns nach § 23 Mindestlohngesetz, Forschungsbericht 558, Dezember 2020. https://www.bmas.de/SharedDocs/Downloads/DE/Publikationen/Forschungsberichte/fb-558-gesamtbericht-zur-evaluation-des-gesetzlichen-mindestlohns.pdf?__blob=publicationFile&v=2. Zugegriffen: 4. Jan. 2023

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artifi Intell Res 16:321–357. https://doi.org/10.1613/jair.953

Cox DR (1958) The regression analysis of binary sequences. J Royal Stat Soc Ser B 20:215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

De la Croix M (2021) Die Novellierung des Verdienststatistikgesetzes ab 2021 – Aktuellere und umfassendere Daten durch das neue System der Verdienststatistiken. Stat Monatsh Baden-Württemb 2021(3):43–45

Dumpert F, Beck M (2017) Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken. AStA Wirtsch Sozialstat Arch 11:81–106. https://doi.org/10.1007/s11943-017-0208-6

Fahrmeir L, Kneib T, Lang S (2009) Regression, 2. Aufl. Springer, Heidelberg

Fernández A, Garciá S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer, Cham

Feuerhake J, Dumpert F (2016) Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken. WISTA Wirtschaft Stat 2016(2):79–94

Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (o.J.) Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB). https://fdz.iab.de/unsere-datenprodukte/personen-und-haushaltsdaten/siab/. Zugegriffen: 4. Jan. 2023

Frentzen K, Günther R (2017) Verdiensterhebung 2015 – Abschlussbericht einer Erhebung über die Wirkung des gesetzlichen Mindestlohns auf die Verdienste und Arbeitszeiten der abhängig Beschäftigten. https://www.destatis.de/DE/Themen/Arbeit/Verdienste/Mindestloehne/Publikationen/Downloads-Mindestloehne/verdiensterhebung-mindestlohn-5611112159004.pdf. Zugegriffen: 4. Jan. 2023

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on Machine Learning, S 1–9

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2. Aufl. Springer, New York

He H, Ma Y (2013) Imbalanced learning. John Wiley & Sons, Hoboken

Himmelreicher R, vom Berge P, Fitzenberger B, Günther R, Müller D (2017) Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE). RatSWD Working Papers, Bd. 262

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Lunardon N, Menardi G, Torelli N (2014) ROSE: a package for binary imbalanced learning. R J 6:82–92. https://doi.org/10.32614/rj-2014-008

Luque A, Carrasco A, Martín A, de las Heras A (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 91:216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mindestlohnkommission (2016) Erster Bericht zu den Auswirkungen des gesetzlichen Mindestlohns. Bericht der Mindestlohnkommission an die Bundesregierung nach § 9 Abs. 4 Mindestlohngesetz. Mindestlohnkommission, Berlin

Ohne Verfasser (2017) Integrierte Erwerbsbiografien (IRB). https://www.iab-forum.de/glossar/integrierte-erwerbsbiographien-ieb/. Zugegriffen: 4. Jan. 2023

Schmidt E (2020) Korrektur des Tätigkeitsschlüssels der Bundesagentur für Arbeit mithilfe maschineller Lernverfahren. WISTA Wirtschaft Stat 2020(6):37–47

SPD (2013) Das WIR entscheidet. Das Regierungsprogramm 2013–2017. https://www.spd.de/fileadmin/Dokumente/Beschluesse/Bundesparteitag/20130415_regierungsprogramm_2013_2017.pdf. Zugegriffen: 4. Jan. 2023

Statistische Ämter des Bundes und der Länder (2020) Die neue Verdiensterhebung. https://erhebungsportal.estatistik.de/Erhebungsportal/#6gI4hNRqAWUsOvgS/unterstuetzte-statistiken/preise-verdienste-arbeitskosten/verdienste-und-arbeitskosten/verdiensterhebung/weiterfuehrende-informationen. Zugegriffen: 4. Jan. 2023

Statistisches Bundesamt (2008) Klassifikation der Wirtschaftszweige. https://www.destatis.de/static/DE/dokumente/klassifikation-wz-2008-3100100089004.pdf. Zugegriffen: 4. Jan. 2023

Statistisches Bundesamt (2016) Qualitätsbericht Verdienststrukturerhebung 2014. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Verdienste/verdienststrukturerhebung-2014.pdf?__blob=publicationFile. Zugegriffen: 4. Jan. 2023

Hinweis des Verlags Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.