Dumpert, Florian

Article

Maschinelles Lernen im Statistischen Bundesamt - Ein Überblick über die Historie seit 2015 und aktuelle Entwicklungen

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Dumpert, Florian (2024) : Maschinelles Lernen im Statistischen Bundesamt - Ein Überblick über die Historie seit 2015 und aktuelle Entwicklungen, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 4, pp. 17-28

This Version is available at:
https://hdl.handle.net/10419/302056

MASCHINELLES LERNEN IM STATISTISCHEN BUNDESAMT

Ein Überblick über die Historie seit 2015 und aktuelle Entwicklungen

Florian Dumpert

Schlüsselwörter: Digitalisierung – Automatisierung – methodische Weiterentwicklung – Zusammenarbeit mit der Wissenschaft – Technologie

ZUSAMMENFASSUNG

Maschinelles Lernen ist ein wichtiger Bestandteil der amtlichen Statistik. Der Artikel stellt erstmals umfassend die Historie sowie aktuelle Entwicklungen zum Maschinellen Lernen in der amtlichen Statistik in Deutschland dar. Er zeigt drei charakteristische Phasen auf, von den Anfängen über die quantitative und qualitative Erweiterung hin zur Reifung und Etablierung, und benennt anstehende Herausforderungen. Dabei liegt der Fokus auf der Bundesstatistik und dem Statistischen Bundesamt, ergänzt durch Verweise auf den Statistischen Verbund und internationale Entwicklungen.

Keywords: *digitalisation – automation – methodological development – collaboration with academia – technology*

ABSTRACT

Machine learning is an essential element of official statistics. This article is the first to present a comprehensive account of the history and current developments of machine learning in official statistics in Germany. It looks at three characteristic phases, detailing the beginnings of machine learning, its quantitative and qualitative expansion and its subsequent maturation and establishment, and identifies upcoming challenges. The focus of the article is on federal statistics and the Federal Statistical Office, with references to the German Official Statistics Network and international developments.

Dr. Florian Dumpert

leitet das Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes, das sich beispielsweise mit Verfahren des Maschinellen Lernens und der Imputation befasst. Der Diplom-Mathematiker beschäftigt sich unter anderem mit methodischen Fragestellungen beim Einsatz dieser Verfahren und verantwortet die Weiterentwicklung der Bundesstatistik mittels dieser Verfahren.

1

Einleitung

Dieser Aufsatz beleuchtet das Maschinelle Lernen (ML) in der deutschen amtlichen Statistik und berücksichtigt dabei besonders die Perspektive des Statistischen Bundesamtes. Die Einsatzzwecke von Maschinellem Lernen in der amtlichen Statistik sind breit gestreut: Sie reichen von für sich stehenden Klassifikationen und Regressionen über ML-basiertes Zusammenführen von Datensätzen bis hin zum Einsatz von Maschinellem Lernen, um fehlerhafte und fehlende Werte zu ersetzen. Die Historie des Einsatzes von Maschinellem Lernen im Statistischen Bundesamt beginnt ungefähr 2015 – eine kurze Zeitspanne im Vergleich zur Historie des Statistischen Bundesamtes selbst, das 2023 seinen 75. Geburtstag feierte (Statistisches Bundesamt, 2023a). Das ist auf den ersten Blick nicht besonders überraschend, handelt es sich bei Maschinellem Lernen oder (häufig synonym verwendet) bei Künstlicher Intelligenz (KI) doch um eine „State of the art"-Technologie. Ob sich die Anwendung von Verfahren des Maschinellen Lernens tatsächlich noch als vergleichsweise „jung" bezeichnen lässt, hängt vor allem von der Charakterisierung des Begriffs des Maschinellen Lernens ab.[1] Das Statistische Bundesamt nutzt die folgende, auch graduell erfüllbare Charakterisierung:

Bei Maschinellem Lernen handelt es sich

- häufig (nicht immer!) um nichtparametrische statistische Methoden,
- um Muster und Zusammenhänge zu erkennen,
- mit einem Schwerpunkt – zumindest beim sogenannten überwachten Lernen – auf der Prädiktion (und weniger auf der Erklärung),
- die genutzt werden, um konkrete Fragestellungen zu beantworten, ohne die Lösung explizit vorgegeben zu bekommen,
- somit im Allgemeinen daten- und nicht modellgetrieben sind
- und sich dadurch auszeichnen, dass der Raum der Lösungen (gemeint ist der Hypothesenraum) häufig so groß ist, dass er (annähernd) alle Muster und Zusammenhänge enthält.[2]

Die Charakterisierung erfolgt somit nicht in der Art, dass die Liste möglicher Verfahren (lineare Regression, logistische Regression, Klassifikations- und Regressionsbaum, Support Vector Machine, Neuronales Netz und weitere Verfahren) nach „klassischer Statistik"[3] und Maschinellem Lernen sortiert würde. So kann beispielsweise die lineare Regression – je nach Anwendungsfall und wie sie trainiert (das heißt angepasst) wird – sowohl klassisch als auch im Sinne des Maschinellen Lernens betrachtet werden. Je eher die oben genannten Aspekte erfüllt sind, desto eher wird von Maschinellem Lernen gesprochen.[4]

Die folgenden Kapitel 2, 3 und 4 beschreiben den Einsatz und die Weiterentwicklung von Maschinellem Lernen im Statistischen Bundesamt in den Jahren 2015 bis 2017, 2018 bis 2020 sowie 2021 bis 2023. Kapitel 5 geht auf aktuelle Entwicklungen ein, Kapitel 6 skizziert abschließend Herausforderungen und wie die Zukunft von Maschinellem Lernen in der amtlichen Statistik aussehen könnte.

2

Die Anfänge: 2015 bis 2017

Die Anfangsjahre des Maschinellen Lernens im Statistischen Bundesamt waren geprägt von einem Herantasten und ersten Versuchen des Einsatzes. Vorerfahrungen bestanden kaum, nur wenige Beschäftigte hatten bereits Kontakt mit diesen Verfahren. Das verwundert nicht, denn selbst die meisten neu eingestellten Beschäftigten mit Hochschulabschluss hatten im Rahmen ihrer universitären Curricula zu diesem Zeitpunkt kaum oder keinen Kontakt zu Maschinellem Lernen. So hing es

1 Wobei Charakterisierung hier nicht im Sinne einer exakten „genau dann, wenn"-Beziehung zu verstehen ist.

2 Der letzte Punkt besagt, dass prinzipiell – wenngleich im Einzelfall gegebenenfalls nicht de facto – alle (messbaren) Muster und Zusammenhänge beliebig genau abgebildet werden können, es also keine grundsätzlich nicht darstellbaren Muster und Zusammenhänge gibt.

3 Dieser Begriff wird in diesem Aufsatz nicht näher erläutert, sondern dient als gedanklicher Gegenpol zum Maschinellen Lernen.

4 Eine Quantifizierung oder gar die Angabe eines Schwellenwertes ist müßig.

zunächst von aus Eigeninteresse erworbenen Kenntnissen zu Maschinellem Lernen sowie von der Möglichkeit zur Zusammenarbeit mit Externen ab, wo (das heißt in welcher Fachstatistik) ML-Projekte stattfinden konnten. Diese dienten jedoch nicht ausschließlich dem reinen Erkunden neuer Möglichkeiten. Bereits zu Beginn war die Zielsetzung deutlich: Mittels Maschinellem Lernen sollten Arbeitsschritte im statistischen Produktionsprozess[5] effizienter gestaltet oder Auswertungsmöglichkeiten erweitert werden, um damit die Qualität der Statistiken zu verbessern. Erste Projekte umfassten beispielsweise die Identifikation im statistikrechtlichen Sinne nicht relevanter Einheiten in der Handwerksstatistik (Feuerhake/Dumpert, 2016) oder die Erkennung von Unternehmen, die dem sogenannten Dritten Sektor angehören (Dumpert und andere, 2016). Diese und weitere Beispiele werden auch in Dumpert/Beck (2017) diskutiert.

Charakteristisch für die zunehmende Befassung mit dem Maschinellen Lernen war einerseits ein stetiges Hinzulernen hinsichtlich der Herausforderungen, die der Einsatz von Maschinellem Lernen mit sich brachte. Diese Herausforderungen waren sowohl allgemeiner Art (zum Beispiel der Umgang mit „imbalanced data"-Situationen) als auch der spezifischen Situation im Statistischen Bundesamt geschuldet. Beispielsweise war die Nutzung der Programmiersprache R zu dieser Zeit im Statistischen Bundesamt noch nicht etabliert, sodass Wege gefunden werden mussten, die entwickelten ML-Lösungen überhaupt auf statistischen Einzeldaten einsetzen zu können. Andererseits gab es zu dieser Zeit auch noch keine einheitliche Vorstellung davon, wie der Einsatz von Maschinellem Lernen die amtliche Statistik jenseits der genannten Einzelfälle bereichern könnte. Darüber hinaus mussten ML-bezogene Kooperationen mit der Wissenschaft und der diesbezügliche Austausch mit anderen (statistischen) Ämtern erst aufgebaut werden. Von Anfang an wurde großer Wert darauf gelegt, den Fortschritt und auch die untersuchten und verwendeten Verfahren durch Publikationen und Vorträge sowohl innerhalb der amtlichen Statistik als auch gegenüber der Fachöffentlichkeit transparent zu machen. Dieses Vorgehen erfüllte bereits die Forderung in den aktuellen Debatten zur Regulierung von Künstlicher Intelligenz, den Einsatz solcher Verfahren bei der Erstellung eines Produkts bekanntzugeben.[6]

Somit ist festzuhalten, dass die Umsetzung der ersten Projekte, der „low hanging fruits", stark von einzelnen Personen und deren Engagement (sowohl auf strategischer als auch auf fachlicher Ebene) abhängig war. Wäre es dabei geblieben, wäre die Zukunft des Maschinellen Lernens in der Bundesstatistik schnell infrage gestellt worden.

3

Qualitative und quantitative Erweiterung: 2018 bis 2020

Die erfolgreichen Projekte der Anfangsjahre eröffneten die Möglichkeit, das Thema höherrangig zu behandeln und eine generelle Aussage über den Einsatz von Maschinellem Lernen in der amtlichen Statistik zu treffen. Das Maschinelle Lernen wurde den Digitalisierungsthemen des Statistischen Bundesamtes (und später auch des Statistischen Verbunds[7]) zugeordnet und Teil der Digitalen Agenda: „Eines von vier Leuchtturmprojekten unserer Digitalen Agenda ist die Durchführung eines ‚Proofs of Concept Machine Learning'." (Statistisches Bundesamt, 2019; hier: Seite 15). Der im Jahr 2018 durchgeführte Proof of Concept umfasste folgende Teilbereiche:

- Klärung der Begrifflichkeit,
- Konzept zum Informationsaustausch im Statistischen Bundesamt,
- Umfrage bei Statistikinstitutionen zum Einsatz von Maschinellem Lernen,
- Abfrage im Statistischen Bundesamt zum Einsatz von Maschinellem Lernen,
- Gedanken zur notwendigen Infrastruktur,
- Handlungsempfehlungen.

5 Für eine Erläuterung des allgemeinen statistischen Produktionsprozesses siehe Wirtschaftskommission für Europa der Vereinten Nationen – UNECE (statswiki.unece.org) sowie Blumöhr und andere (2017), hier: Kapitel 3.

6 Beispiele sind selbstverpflichtende Leitlinien für den KI-Einsatz (Denkfabrik Digitale Arbeitsgesellschaft im Bundesministerium für Arbeit und Soziales, 2022), die Stellungnahme des Deutschen Ethikrats (Deutscher Ethikrat, 2023) sowie die europäische KI-Verordnung.

7 Den Statistischen Verbund bilden die Statistischen Ämter des Bundes und der Länder.

Der Informationsaustausch sollte das vorhandene Wissen zu Maschinellem Lernen in Form von Notizen und Dokumenten an einer zentralen Stelle bündeln und für alle Beschäftigten auffindbar und nutzbar machen. Ein zentrales System, das dieses leisten konnte, gab es zu diesem Zeitpunkt noch nicht, es musste auf eine Plattform aus der Softwareentwicklung zurückgegriffen werden. Zudem dienten Veranstaltungsformate für verschiedene Zielgruppen dem Informationsaustausch im Statistischen Bundesamt. Sie stellten die Chancen und die zu bedenkenden Aspekte von Maschinellem Lernen, die Beispielprojekte aus den Jahren 2015 bis 2017 sowie die Ergebnisse der Umfrage bei anderen Statistikinstitutionen vor und zur Diskussion. Themen der Umfrage unter allen europäischen Statistikämtern, vielen deutschen und einigen weiteren ausländischen Statistikproduzenten waren im Wesentlichen der Inhalt des jeweiligen Projekts (Zielsetzung), der aktuelle Status (Idee, Entwicklung, Test, Produktion), das oder die eingesetzte(n) Verfahren und die eingesetzte Software. Ergebnis des Proofs of Concept 2018 war, dass einige Ämter und Institutionen Maschinelles Lernen bereits erprobten, dass hierfür häufig baumbasierte Verfahren (wie Random Forest) eingesetzt wurden, dass aber viele Ansätze noch in der Entwicklungsphase, mithin nicht im Produktivbetrieb waren. Häufig wurden Aufgaben aus dem Bereich der Klassifikation (im Unterschied zur Regression) mittels Maschinellem Lernen bearbeitet.

Bei der Infrastruktur wurde der Fokus auf die Notwendigkeit der Bereitstellung von R gelegt. Alternativen (zum Beispiel MATLAB) und Ergänzungen (zum Beispiel Python) wurden ebenfalls diskutiert. Damit verbunden wurde deutlich, dass Arbeitsplatz-PCs alleine nicht ausreichen, um Maschinelles Lernen im Statistischen Bundesamt zu betreiben. Auch die Notwendigkeit der Nutzung von Grafikprozessoren für den Einsatz von Deep Learning hatte der Proof of Concept bereits ergeben. Der Proof of Concept schloss mit Handlungsempfehlungen für die Leitung des Statistischen Bundesamtes: ein zentrales Referat für das Thema Maschinelles Lernen einrichten, die notwendige Infrastruktur bereitstellen, die Statistischen Ämter der Länder einbeziehen, Innen- und Außenkommunikation praktizieren sowie mit der Wissenschaft zusammenarbeiten (Beck und andere, 2018a; Beck und andere, 2018b).

Zum 1. August 2018 wurde das Referat „Maschinelles Lernen und Imputationsverfahren“ im Statistischen Bundesamt eingerichtet und damit erstmals ein zentrales Kompetenzzentrum zu diesen Themen organisatorisch verankert.[8] Ebenfalls 2018 wurde der erste Platz des Innovationspreises des Statistischen Beirats[9] für die bisherigen Arbeiten im Bereich Maschinelles Lernen vergeben. Mit der Einrichtung des Referats und der Würdigung der bisherigen Arbeiten durch den Statistischen Beirat einher gingen ein personeller Aufwuchs und damit eine quantitative Steigerung der untersuchten Fragestellungen (unter anderem Schmidt, 2020). Neue Anwendungsfälle wurden in gemeinsamen Workshops von Fachreferaten und zentralem Kompetenzzentrum erarbeitet. Damit verbunden zeigte sich erstmals, dass eine Koordinierung der Arbeiten notwendig ist. Darüber hinaus wurde deutlich, dass es offene methodische Fragestellungen bezüglich der Anwendung von Maschinellem Lernen in der amtlichen Statistik gab (und bis heute gibt) und dass die nationale wie internationale Vernetzung ausgebaut werden sollte.

Auf nationaler Ebene gründete sich im Jahr 2018 eine Gruppe namens „KI-Labor“, die schließlich im 2020 eingerichteten Arbeitskreis Maschinelles Lernen (AK ML) aufging. Dieses Gremium besteht aus Vertreterinnen und Vertretern des Statistischen Bundesamtes und der Statistischen Ämter der Länder und versteht sich als „Treiber des Prozesses“ im Statistischen Verbund rund um das Thema Maschinelles Lernen. Es übernimmt dort die Steuerung der mit der Thematik verbundenen Aufgaben und Maßnahmen und vernetzt die beteiligten Akteure und Stakeholder.

Besonders wichtig für die internationale Vernetzung war das Machine Learning Project 2019/2020 der High-Level Group for the Modernisation of Official Statistics der Wirtschaftskommission der Vereinten Nationen für Europa (UNECE HLG-MOS). Dieses Projekt hat erstmals einen direkten Austausch auf internationaler Ebene zum Einsatz von Maschinellem Lernen in der amtlichen Statistik ermöglicht (United Nations, 2022). Details zu diesem Projekt und eine entsprechende Einordnung für die deutsche amtliche Statistik liefert Dumpert (2021).

8 Das Thema Imputation war zuvor organisatorisch mit der statistischen Geheimhaltung verbunden.

9 Der Statistische Beirat ist das nach § 4 Bundesstatistikgesetz berufene Gremium, welches das Statistische Bundesamt in statistischen Fachfragen berät und das die Belange der Nutzerinnen und Nutzer der Bundesstatistik vertritt.

4

Reifung und Etablierung: 2021 bis 2023

In den Jahren 2021 bis 2023 hat sich der Umgang mit und der Einsatz von Maschinellem Lernen in der Bundesstatistik (und auch in einigen Statistischen Landesämtern) zunehmend etabliert. In vielen Fällen wird nun Maschinelles Lernen, einschließlich Natural Language Processing[10], erprobt oder im statistischen Produktionsprozess eingesetzt (siehe dazu zum Beispiel Levagin und andere, 2022; Dumpert/Beck, 2023; Moritz und andere, 2024; Weißmann/Herbst, 2024; Limberg, 2024) oder beispielsweise seine Nützlichkeit bei der Erzeugung synthetischer Daten untersucht[11]. Gleichzeitig wurden ausstehende Aufgaben bearbeitet, zu nennen sind hier die methodische Weiterentwicklung sowie Fragen zu Qualitätsaspekten beim Einsatz Maschinellen Lernens in der amtlichen Statistik.

Im Bereich der Methodenfragen wurde das Thema Record Linkage aufgegriffen (Schnell, 2021). Darüber hinaus galt es, Fragen zum Hyperparameter-Tuning[12] zu klären. Hinsichtlich eher allgemeiner Fragen hierzu liefert Bischl und andere (2022) wichtige Einsichten. Für eher spezielle Fragestellungen (zum Beispiel die Visualisierung des Tuning-Prozesses) sei auf Bartz und andere (2023) verwiesen. Außerdem konnten 2022 die Arbeiten in einem Kooperationsprojekt mit der Ludwig-Maximilians-Universität München beginnen. Bearbeitet wurden Fragen zur Evaluation der Prädiktionsgüte von Maschinellen Lernverfahren in komplexen Situationen (Hornung und andere, 2023), zur Interpretierbarkeit von Ergebnissen (zum Beispiel Dandl und andere, 2023), zu Maschinellem Lernen bei besonderen Stichprobendesigns (Nalenz und andere, 2024) und zu Fairness (zum Beispiel Schenk/Kern, 2024). Die Arbeiten von Stock und anderen (2023) zu Federated Learning und von Moritz und anderen (2024) zur Anwendbarkeit von Online Learning in der amtlichen Statistik erweiterten den Blick über die naheliegenden Einsätze hinaus.

Mit Blick auf Qualitätsaspekte beim Einsatz Maschinellen Lernens in der amtlichen Statistik wurde in Zusammenarbeit mit Kolleginnen und Kollegen aus Statistischen Landesämtern in einem mehrstufigen Prozess eine Liste von Dimensionen und Querschnittsaspekten erarbeitet (Saidani und andere, 2023). Diese leitet aus den generischen Ausführungen im Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder (Statistische Ämter des Bundes und der Länder, 2021) die speziellen Anforderungen für Maschinelles Lernen ab. Dem vorausgegangen war ein Workshop mit Teilnehmerinnen und Teilnehmern aus Wissenschaft, amtlicher Statistik und Nutzenden (zum Beispiel Wirtschaftsforschungsinstituten) zu Qualitätsaspekten Maschinellen Lernens (Dumpert und andere, 2023). Darauf aufbauend werden aktuell Fragen nach konkreten Qualitätsrichtlinien und nach erweiterten Möglichkeiten des transparenten Umgangs mit Maschinellem Lernen in der amtlichen Statistik behandelt.

Ein ganz anderer Ansatz, Maschinelles Lernen als gewöhnliches Hilfsmittel zur (Teil-)Automatisierung von statistischen Produktionsprozessen oder zur Datenanalyse zu etablieren, waren die Hackathons in den Jahren 2021 und 2022. Sie standen unter den Aufgabenstellungen „Wie können wir zukünftig die Effektivität von nationaler Klimaschutzpolitik datenbasiert validieren?“ beziehungsweise „Welches innovative Produkt kann die amtliche Statistik entwickeln, um in Zukunft schneller relevante Daten in Krisen bereitzustellen?“. Der Hackathon 2021 wurde als interne Veranstaltung des Statistischen Bundesamtes durchgeführt, der Hackathon 2022 war auch für Teilnehmerinnen und Teilnehmer aus statistischen Ämtern anderer Staaten offen. Anhand der Kriterien Relevanz, Innovationscharakter, Design, Automatisierung und Universalität hat eine Jury die Lösungen ausgewertet. Wesentliches Ziel der beiden Hackathons war es, Freiraum und Gelegenheit für eine intensive und interdisziplinäre Zusammenarbeit außerhalb der normalen Arbeitsroutinen zu schaffen. Gleichzeitig wurde auch das Vernetzen innerhalb des Statistischen Bundesamtes sowie mit Kolleginnen und Kollegen aus anderen statistischen Ämtern gefördert. Obwohl die Hackathons auch Wettbewerbscharakter hatten, waren es letztlich doch die übergreifenden Aspekte, die dieses Format för-

10 Natural Language Processing bezeichnet die Verarbeitung natürlicher Sprache durch Computer. Dabei wird natürliche Sprache (zum Beispiel Texte) in durch statistische Verfahren verarbeitbare Daten transformiert und mit geeigneten Verfahren, zum Beispiel des Maschinellen Lernens, weiterverarbeitet.

11 Dies geschieht beispielsweise im 2023 begonnenen Forschungsprojekt „Anonymität bei integrierten und georeferenzierten Daten“ (AnigeD).

12 Die Ermittlung geeigneter Detailspezifikationen der zu trainierenden ML-Verfahren wird Hyperparameter-Tuning genannt.

derlich für eine weitere Verbreitung von Maschinellem Lernen im Bewusstsein der Beschäftigten erscheinen ließen.[13]

Die Weiterentwicklung betrifft aber nicht nur methodische Fragen und Aspekte der Arbeitskultur im Statistischen Bundesamt, sondern in gleichem Maße die technologischen Möglichkeiten, die dem Statistischen Bundesamt zur Verfügung stehen. Während in den Anfangsjahren des Maschinellen Lernens noch auf Arbeitsplatzrechnern – also im besten Sinne lokal – gearbeitet werden musste, hat sich dies über die Zeit wesentlich verändert. Heute gibt es rechtliche und technische Möglichkeiten, Server- und (private) Cloudlösungen zu nutzen, was sinnvoll und notwendig für einen gewinnbringenden Einsatz von Maschinellem Lernen in der amtlichen Statistik ist.

13 Auch 2023 fand ein Hackathon im Statistischen Bundesamt statt, dieser jedoch ohne expliziten Bezug zu Künstlicher Intelligenz/Maschinellem Lernen. Teams des Statistischen Bundesamtes waren darüber hinaus sehr erfolgreich beim European Big Data Hackathon 2023 (Statistisches Bundesamt, 2023b, hier: Seite 14).

5

Internationale Entwicklungen: 2024

Bereits im Frühjahr 2023 hat das Statistische Amt der Europäischen Union (Eurostat) die nationalen statistischen Institute aufgerufen, sich auf einen Grant zum Thema Künstliche Intelligenz und Maschinelles Lernen in der amtlichen Statistik (AIML4OS) zu bewerben. Das Statistische Bundesamt schloss sich seinerzeit einem Konsortium unter der Leitung Irlands an, das schließlich aus 15 beteiligten Staaten bestand und im März 2024 auch den finalen Zuschlag für die Förderung erhielt. Die Arbeiten im 4-Jahres-Grant begannen am 2. April 2024. Teilfinanziert durch den Grant wird bis 2028 an Querschnitts- und anwendungsspezifischen Themen gearbeitet. Einen Überblick über die Arbeitspakete gibt ↘ Übersicht 1.

Im Frühjahr 2024 hat das Statistische Bundesamt die als wissenschaftliche Fachtagung konzipierte „Conference on Foundations and Advances of Machine Lear-

Übersicht 1

Arbeitspakete am AIML4OS-Grant zum Thema Künstliche Intelligenz und Maschinelles Lernen in der amtlichen Statistik

Nr.	Titel	Leitung	Beteiligung des Statistischen Bundesamtes
1	Project management and coordination	Irland	
2	Communication and community engagement	Italien	
3	ESS AI/ML lab: Technical infrastructure and organisational setup	Frankreich	
4	AI/ML state-of-play and ecosystem monitoring	Deutschland	✓
5	Standards, methodological and implementation frameworks	Niederlande und Deutschland	✓
6	Knowledge repository and training material	Polen	
7	Use Case: AI/ML on earth observation data, satellite imagery	Niederlande	
8	Use Case: Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing	Deutschland	✓
9	Use Case: Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on imputation	Spanien	✓
10	Use Case: From text to code – Experiences and potential of the use of AI/ML for classifying and coding	Deutschland	✓
11	Use Case: Applying ML for estimating firm-level supply chain networks	Niederlande	
12	Use Case: Large language models	Schweden	
13	Use Case: Generation of synthetic data in official statistics: techniques and applications	Italien und Österreich	✓

ning in Official Statistics“ veranstaltet. [14] Ziel war, den Austausch mit der Wissenschaft, anderen Behörden und anderen statistischen Ämtern national wie international zu stärken, Einblicke in die Herangehensweisen und Projekte anderer zu erhalten und hinsichtlich methodischer und technologischer Entwicklungen am Puls der Zeit zu sein. Inspiriert durch über 40 Vorträge haben rund 150 Teilnehmerinnen und Teilnehmer aus insgesamt 19 Staaten aktuelle Herausforderungen und Ansätze zu deren Lösungen präsentiert und diskutiert. Das Themenspektrum reichte dabei von

- mathematisch-statistischen Fragestellungen (unter anderem zu Fehler und Unsicherheit, Einfluss des Stichprobendesigns und Resamplingverfahren) über
- Technologie, Standards und Qualität (einschließlich Fairness und Reproduzierbarkeit sowie Fragestellungen rund um das Thema Prozessverbesserung) bis hin zu
- konkreten Anwendungsfällen Maschinellen Lernens innerhalb und außerhalb der Statistikproduktion (unter anderem Datenintegration, Textklassifikation und Natural Language Processing, Codierung statistischer Einheiten auf Basis textueller Beschreibungen, Nutzung von Large Language Models, Datenvalidierung und Imputation).

6

Herausforderungen und ein Blick in eine mögliche Zukunft

Dass Maschinelles Lernen in der amtlichen Statistik auch weiterhin eine wichtige Rolle spielen wird, steht außer Frage. Anforderungen aus dem Bundesstatistikgesetz [15] und dem Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder [16] liefern Gründe dafür, insbesondere das daraus abgeleitete Bestreben, die Produkte der amtlichen Statistik – auch wissenschaftlich – so verwertbar wie möglich für ihre Nutzerinnen und Nutzer zu erstellen (Qualität als „fitness for use“). Auch die Kommission Zukunft Statistik [17] empfiehlt, „lernende Algorithmen in der Entwicklung, Produktion und Evaluation eigener Produkte explorativ und transparent“ einzusetzen (Kommission Zukunft Statistik, 2024; hier: Seite 6). Darüber hinaus trifft der Fachkräftemangel auch den öffentlichen Dienst und die amtliche Statistik konkurriert mit anderen Aufgabenbereichen der öffentlichen Verwaltung um Finanzmittel. Bei der daher notwendigen (Teil-)Automatisierung von Produktionsschritten, gegebenenfalls einhergehend mit einer Digitalisierung und Standardisierung der Prozessschritte und der eingesetzten Werkzeuge, spielt Maschinelles Lernen notwendigerweise eine wichtige Rolle.

Bei allen Erfolgen, die hier bereits erzielt wurden, bleiben aber auch offene Fragen. Neben noch ungeklärten methodischen Fragestellungen und der konkreten Umsetzung des Qualitätsbegriffs und der Qualitätssicherung (siehe Kapitel 4) stellt die Integration der einzusetzenden ML-Lösungen in die IT-Systeme eine Herausforderung für die Zukunft dar. Die Einführung von MLOps [18], das längst Industriestandard ist, steht für die Bundesstatistik noch aus, das Gleiche gilt für ein entsprechendes Management der mannigfaltigen Datenbestände der deutschen amtlichen Statistik. Aus diesen Datenbeständen speisen sich die für Maschinelles Lernen erforderlichen Trainings-, Validierungs- und Testdaten und in diese Bestände gehen durch ML-Verfahren bearbeitete Daten ein. Insbesondere mit Blick auf die Qualitätsdimension der Reproduzierbarkeit erscheinen diese beiden Punkte unabdingbar. Außerdem ist gerade bei den technologischen Möglichkeiten nie ein finaler Zustand erreicht. Durch die Weiterentwicklung auf methodischer und technologischer Seite ist die Infrastruktur der Informationstechnik immer wieder zu aktualisieren, zu verändern und anzupassen – sofern der ent-

14 Die Tagungsdokumentation mit Abstracts und Präsentationsfolien steht unter www.destatis.de zur Verfügung.

15 „Sie [die Bundesstatistik] gewinnt die Daten unter Verwendung wissenschaftlicher Erkenntnisse und unter Einsatz der jeweils sachgerechten Methoden und Informationstechniken.“ (§ 1 Satz 3 Bundesstatistikgesetz)

16 Statistische Ämter des Bundes und der Länder (2021). Hier insbesondere: Qualitätsgrundsätze G07 (solide Methodik), G08 (geeignete statistische Verfahren), G10.2 (Produktivitätspotenzial der Informationstechnologie).

17 Die Kommission Zukunft Statistik (KomZS) wurde vom Statistischen Bundesamt eingerichtet und mit der Erarbeitung von Empfehlungen für eine vorausschauende Programmplanung und eines Zielbilds der amtlichen Statistik für das Jahr 2030 beauftragt. Der Abschlussbericht der Kommission (Kommission Zukunft Statistik, 2024) wurde am 16. Januar 2024 der Leitung des Statistischen Bundesamtes überreicht.

18 Unter MLOps (machine learning operations) versteht man „Praktiken und Prozesse, die darauf abzielen, Modelle für maschinelles Lernen zuverlässig und effizient zu entwickeln, produktiv bereitzustellen, zu verwalten, zu überwachen und zu warten“ (Saidani und andere, 2023; hier: Seite 294).

sprechende Fortschritt für das Statistische Bundesamt nutzbar gemacht werden soll. Weitere erleichternde und hemmende Faktoren von Maschinellem Lernen in der amtlichen Statistik hat kürzlich Karanka (2023) herausgearbeitet.

Bereits heute haben die Kolleginnen und Kollegen in den Fach- und Querschnittsbereichen des Statistischen Bundesamtes die Möglichkeit, sich zum Thema Maschinelles Lernen fortzubilden. Ergänzend ist zu beobachten, dass ML-Themen mittlerweile auch Einzug in viele universitäre Curricula gefunden haben. Daher wird Maschinelles Lernen – zusätzlich begünstigt durch die fortschreitende einfache Nutzbarkeit der ML-Techniken im Standardumfeld des Computerarbeitsplatzes – perspektivisch ein gewöhnliches Werkzeug für viele Beschäftigte in der Statistikproduktion werden. Das aktuell bestehende zentrale Kompetenzzentrum könnte mittelfristig seine Ressourcen stärker zur Weiterentwicklung von Maschinellem Lernen in der amtlichen Statistik einsetzen. Das heißt, es könnte verstärkt methodische Fragestellungen bearbeiten, generische Komponenten entwickeln, die Einhaltung von Standards mit Bezug zu Maschinellem Lernen definieren und überwachen, die fachliche und technologische Integration von Maschinellem Lernen weiter vorantreiben und außerordentliche Fragestellungen lösen.

Obwohl derzeit Teil der öffentlichen Debatte, hat dieser Aufsatz das Thema generative KI nicht behandelt. Zwar gibt es erste Untersuchungen, inwiefern auch diese Klasse von Verfahren (hier vor allem Large Language Models) im Statistischen Bundesamt eingesetzt werden könnte, jedoch zeigten diese innerhalb der Kernprozesse der klassischen Statistikproduktion bislang keinen nennenswerten Vorteil. Es ist für die Zukunft jedoch angezeigt, generative KI für die amtliche Statistik weiter zu beobachten und zu bewerten.

LITERATURVERZEICHNIS

Bartz, Eva/Bartz-Beielstein, Thomas/Zaefferer, Martin/Mersmann, Olaf. *Hyperparameter Tuning for Machine and Deep Learning with R.* Singapur 2023. DOI: 10.1007/978-981-19-5170-1

Beck, Martin/Dumpert, Florian/Feuerhake, Jörg. *Machine Learning in Official Statistics*. 2018a. [Zugriff am 18. Juni 2024]. Verfügbar unter: arxiv.org

Beck, Martin/Dumpert, Florian/Feuerhake, Jörg. *Proof of concept machine learning – Abschlussbericht.* 2018b. [Zugriff am 18. Juni 2024]. Verfügbar unter: www.destatis.de

Bischl, Bernd/Binder, Martin/Lang, Michel/Pielok, Tobias/Richter, Jakob/Coors, Stefan/Thomas, Janek/Ullmann, Theresa/Becker, Marc/Boulesteix, Anne-Laure/Deng, Difan/Lindauer Marius. *Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges*. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Jahrgang 13. Ausgabe 2/2023, Seite e1484 ff. DOI: 10.1002/widm.1484

Blumöhr, Torsten/Teichmann, Corina/Noack, Anke. *Standardisierung der Prozesse: 14 Jahre AG SteP*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2017, Seite 58 ff.

Dandl, Susanne/Casalicchio, Giuseppe/Bischl, Bernd/Bothmann, Ludwig. *Interpretable Regional Descriptors: Hyperbox-Based Local Explanations*. In: Koutra, Danai/Plant, Claudia/Gomez Rodriguez, Manuel/Baralis, Elena/Bonchi, Francesco (Herausgeber). Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023. Cham 2023, Seite 479 ff. DOI: 10.1007/978-3-031-43418-1_29

Denkfabrik Digitale Arbeitsgesellschaft im Bundesministerium für Arbeit und Soziales. *Selbstverpflichtende Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung*. 2022. [Zugriff am 18. Juni 2024]. Verfügbar unter: www.bmas.de

Deutscher Ethikrat. *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. 2023. [Zugriff am 18. Juni 2024]. Verfügbar unter: www.ethikrat.org

Dumpert, Florian. *Machine Learning in der amtlichen Statistik – Ergebnisse und Bewertung eines internationalen Projekts*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2021, Seite 53 ff.

Dumpert, Florian/Beck, Martin. *Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken*. In: AStA Wirtschafts- und Sozialstatistisches Archiv. Jahrgang 11. Ausgabe 2/2017, Seite 83 ff. DOI: 10.1007/s11943-017-0208-6

Dumpert, Florian/Beck, Martin. *Verbesserung der Datengrundlage der Mindestlohnforschung mittels maschineller Lernverfahren*. In: AStA Wirtschafts- und Sozialstatistisches Archiv. Jahrgang 17. Ausgabe 1/2023, Seite 5 ff. DOI: 10.1007/s11943-023-00318-w

LITERATURVERZEICHNIS

Dumpert, Florian/von Eschwege, Katja/Beck, Martin. *Einsatz von Support Vector Machines bei der Sektorzuordnung von Unternehmen*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2016, Seite 87 ff.

Dumpert, Florian/Wichert, Sebastian/Augustin, Thomas/Storfinger Nina. *Editorial Issue 3+4/2023*. In: AStA Wirtschafts- und Sozialstatistisches Archiv. Jahrgang 17. Ausgabe 3–4/2023, Seite 191 ff. DOI: 10.1007/s11943-023-00334-w

Feuerhake, Jörg/Dumpert, Florian. *Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2016, Seite 79 ff.

Hornung, Roman/Nalenz, Malte/Schneider, Lennart/Bender, Andreas/Bothmann, Ludwig/Bischl, Bernd/Augustin, Thomas/Boulesteix, Anne-Laure. *Evaluating machine learning models in non-standard settings: An overview and new findings*. 2023. [Zugriff am 19. Juni 2024]. Verfügbar unter: arxiv.org/abs/2310.15108

Karanka, Joni. *Facilitators and Blockers of ML Adoption in Official Statistics*. 2023. [Zugriff am 19. Juni 2024]. Verfügbar unter: unece.org

Kommission Zukunft Statistik. *Bericht und Empfehlungen der Kommission Zukunft Statistik. Version 1.0*. 2024. [Zugriff am 19. Juni 2024]. Verfügbar unter: www.destatis.de

Levagin, Bogdan/Lange, Kerstin/Walprecht, Sylvana/Gerls, Fabian/Kühnhenrich, Daniel. *Vereinfachtes Verfahren zur interaktiven Schätzung des Erfüllungsaufwands mittels maschinellen Lernens*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2022, Seite 53 ff.

Limberg, Heiko. *Potenziale von Clustering-Algorithmen für die Plausibilisierung im Außenhandel*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2024, Seite 54 ff.

Moritz, Steffen/Dumpert, Florian/Jung, Christian/Bartz-Beielstein, Thomas/Bartz, Eva. *Practical Applications of Online Machine Learning*. In: Bartz, Eva/Bartz-Beielstein, Thomas (Herausgeber). Online Machine Learning. Singapur 2024. Seite 71 ff. DOI: 10.1007/978-981-99-7007-0_7

Moritz, Steffen/Wiynck, Frederik/Wiebels, Johannes. *Prognose der Abgabequote von Einkommensteuererklärungen bei Rentnerinnen und Rentnern*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2024, Seite 83 ff.

Nalenz, Malte/Rodemann, Julian/Augustin, Thomas. *Learning de-biased regression trees and forests from complex samples*. In: Machine Learning. Ausgabe 113. 2024, Seite 3379 ff. DOI: 10.1007/s10994-023-06439-1

Saidani, Younes/Dumpert, Florian/Borgs, Christian/Brand, Alexander/Nickl, Andreas/Rittmann, Alexandra/Rohde, Johannes/Salwiczek, Christian/Storfinger, Nina/Straub, Selina. *Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik*. In: AStA Wirtschafts- und Sozialstatistisches Archiv. Jahrgang 17. Ausgabe 3–4/2023, Seite 253 ff. DOI: 10.1007/s11943-023-00329-7

LITERATURVERZEICHNIS

Schenk, Patrick Oliver/Kern, Christoph. *Connecting Algorithmic Fairness to Quality Dimensions in Machine Learning in Official Statistics and Survey Production*. 2024. [Zugriff am 19. Juni 2024]. Verfügbar unter: https://arxiv.org/abs/2402.09328

Schmidt, Elena. *Korrektur des Tätigkeitsschlüssels der Bundesagentur für Arbeit mithilfe maschineller Lernverfahren*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2020, Seite 37 ff.

Schnell, Rainer. *Expertise Maschinelles Lernen für Record Linkage – Endbericht*. 2021. Internes Dokument.

Statistische Ämter des Bundes und der Länder. *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder (Version 1.21)*. 2021. [Zugriff am 19. Juni 2024]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Digitale Agenda des Statistischen Bundesamtes*. 2019. [Zugriff am 8. Juli 2024]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *WISTA Wirtschaft und Statistik. Ausgabe zu 75 Jahre Statistisches Bundesamt*. 2023a. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *European Big Data Hackathon 2023: Erfolge für Teams des Statistischen Bundesamtes*. 2023b. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2023. Kurznachrichten, Seite 14.

Stock, Joshua/Hauke, Oliver/Weißmann, Julius/Federrath Hannes. *The Applicability of Federated Learning to Official Statistics*. In: Quaresma, Paulo/Camacho, David/Yin, Hujun/Gonçalves, Teresa/Julian, Vicente, Tallón-Ballesteros, Antonio J. (Herausgeber). Intelligent Data Engineering and Automated Learning – IDEAL 2023. Cham 2023. Seite 70 ff. DOI: 10.1007/978-3-031-48232-8_8

United Nations (Vereinte Nationen). *Machine Learning for Official Statistics*. 2021. [Zugriff am 18. Juni 2024]. Verfügbar unter: unece.org

Weißmann, Julius/Herbst, Tim. *Maschinelles Lernen im Basisregister für Unternehmen*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2024, Seite 67 ff.

RECHTSGRUNDLAGEN

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 14 des Gesetzes vom 8. Mai 2024 (BGBl. I Nr. 152) geändert worden ist.

Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24004-4, ISSN 1619-2907