KINRW

Machine Learning Operations (MLOps)

Grundlagen, Chancen und Herausforderungen beim MLOps-Einsatz in Unternehmen

Studie 2024

Die Landesregierung Nordrhein-Westfalen

Fraunhofer IAIS

Machine Learning Operations (MLOps)

Grundlagen, Chancen und Herausforderungen beim MLOps-Einsatz in Unternehmen

Studie 2024

Inhaltsverzeichnis

Executive Summary

Bei Machine Learning Operations (MLOps) handelt es sich um ein Paradigma zur Entwicklung und zum Betrieb von Machine-Learning-Anwendungen (ML) im produktiven Einsatz. Durch die Verbindung der jahrzehntelangen Erfahrung und bewährten Praktiken aus der klassischen Softwareentwicklung mit den spezifischen Anforderungen der KI-Entwicklung ergibt sich ein umfassendes Prozessmodell zum effizienten Einsatz von Künstlicher Intelligenz in Unternehmen.

Einer aktuellen Studie der Bitkom[1] zufolge, sehen sich 43 Prozent aller Unternehmen in Deutschland unter den Nachzüglern beim Thema KI. 38 Prozent geben an, den Anschluss verpasst zu haben. Gleichzeitig sehen rund zwei Drittel der Befragten Künstliche Intelligenz als Chance für ihr Unternehmen und wollen diese einsetzen. Unter den häufigsten Hindernissen für die Einführung von KI wird neben den Anforderungen an Datenschutz (85 Prozent der Befragten) »fehlendes technisches Know-how« genannt (84 Prozent).

Die vorliegende Studie befasst sich mit dem Einsatz und der Verbreitung von MLOps-Praktiken in Unternehmen. Sie stellt einerseits die theoretischen Grundlagen dar und arbeitet andererseits existierende Herausforderungen und daraus abgeleitete Handlungsempfehlungen heraus.

Zu diesem Zweck führten die KI-Experten des Fraunhofer IAIS gemeinsam mit der Kompetenzplattform KI.NRW Interviews mit insgesamt 29 Unternehmen, welche sich bereits mit dem Thema Machine Learning Operations beschäftigt haben. Besonderer Fokus wurde dabei auf den Mittelstand gelegt, um den Stand der Entwicklung und Unterstützungsbedarf bei der Einführung von MLOps zu untersuchen. Methodisch orientiert sich die Studie an dem MLOps-Zyklus und formuliert die Fragen entsprechend der Schwerpunkte in den einzelnen Phasen. Der MLOps-Zyklus bezeichnet eine von Beck et al.[2] entwickelte Unterteilung von Machine Learning Operations in sechs Phasen, die ein typisches KI-Projekt durchläuft. Eine detaillierte Beschreibung erfolgt in Kapitel 1.

Die Ergebnisse der Studie zeigen das dynamische Umfeld, in welchem MLOps eingesetzt wird. Je nach Unternehmen unterscheiden sich Tools, Herangehensweise, Vorwissen und damit die konkrete Ausgestaltung der MLOps-Prozesse.

Viele Unternehmen decken weite Teile des MLOps-Zyklus ab, aber nur wenige haben alle Phasen des Zyklus in der Tiefe umgesetzt. Während die meisten von ihnen den Übergang von der Phase Experiment (Exploration) zur Phase Entwicklung (Development) einer produktiven ML-Lösung beherrschen, verzichten viele Unternehmen auf weitergehende Automatisierung im Bereich Continuous Deployment. Das Monitoring von bereits im Betrieb befindlichen KI-Anwendungen beschränkt sich vielfach auf das Feedback der Nutzer*innen im Hinblick auf die Güte der Vorhersagen. Wir gehen davon aus, dass mit fortschreitender Reife und Erfahrung im Einsatz von MLOps ein höherer Grad der Automatisierung umgesetzt werden wird.

Der Einsatz von Cloud-Diensten (hier vor allem Azure), um ML-Anwendungen bereitzustellen und zu entwickeln, ist weit verbreitet. Die Unternehmen schätzen die niedrigen Eintrittsbarrieren und flexible Kosten. Dennoch gibt es auch eine Vielzahl von Unternehmen, die eine On-Premise-Lösung vorziehen. Die häufigsten Gründe dafür sind datenschutzrechtliche Bedenken bzw. regulatorische Einschränkungen sowie hohe Kosten bei einer aktiven Nutzung der Cloud-Dienste beim Einsatz von Deep Learning.

In den Interviews haben wir gesehen, dass schon wenige Mitarbeiter*innen mit ML-Erfahrung ausreichen, um erste Lösungen zu entwickeln. Jedoch wird eine Mindestgröße des Teams benötigt, um einerseits effektiv arbeiten zu können und andererseits viele Modelle in den produktiven Einsatz zu bringen, unabhängig davon, ob es sich um die ML-Teams bei einem kleinen Start-up oder einem Großunternehmen handelt.

Eine große Herausforderung sehen die Unternehmen bei den Daten, beispielsweise in Bezug auf deren Verfügbarkeit und Qualität. Die meisten der Befragten haben das Thema erkannt und arbeiten an einer Lösung. Eine weitere Erkenntnis aus den Interviews ist, dass keines der Unternehmen einen Feature Store als feste Komponente einsetzt, sondern bestenfalls für einzelne Use Cases.

Während manche Unternehmen einen Ende-zu-Ende-Ansatz verfolgen (d. h. ein*e Data-Scientist*in baut die Anwendung vom Experiment zum Betrieb), etablieren vor allem größere Unternehmen ein Konzept, das eine Zweiteilung vorsieht: Fachseiten und Data Science bauen einen Proof of Concept, danach übernehmen Softwareentwickler*innen bzw. die IT-Abteilung.

Insgesamt lässt sich sagen, dass MLOps bei den befragten Unternehmen als wichtig erkannt wird und schon weit verbreitet ist. Die MLOps-Umgebungen und Prozesse befinden sich jedoch zum Teil noch im Aufbau bzw. sind in einigen Phasen des MLOps-Zyklus noch ausbaufähig.

Abschließend möchten wir noch auf eine Besonderheit dieser Studie eingehen. Während der Interviewphase hat OpenAI ChatGPT veröffentlicht und KI damit zu einem dominierenden Thema gemacht. Die rasante Verbreitung und Adaptierung am Markt spiegelt sich auch in den Interviews wider und wird daher in einem dedizierten Abschnitt betrachtet.

Wir bedanken uns an dieser Stelle für das Engagement der beteiligten Unternehmen und deren Bereitschaft, ihre Erkenntnisse mit uns zu teilen.

1 Bitkom Research (2023).
2 Beck, N., Martens, C., Sylla, K.-H., Wegener, D. und Zimmermann, A. (2020).

Die Autoren

Lennard Helmer

Er ist Mitglied des MLOps-Teams am Fraunhofer IAIS. Neben seiner Forschung im Bereich von Machine Learning Operations für vertrauenswürdige KI und Datenpipelines berät er Unternehmen bei der Implementierung von MLOps in ihre Geschäftsprozesse. Darüber hinaus ist er als Dozent für MLOps für die Fraunhofer-Allianz Big Data und Künstliche Intelligenz tätig.

Andreas Kerbel

Er ist KI-Manager bei der Kompetenzplattform KI.NRW. Sein Schwerpunkt liegt auf der Beratung von Unternehmen im Bereich KI. Als Wirtschaftsingenieur arbeitete er an der Schnittstelle zwischen Technologie und Wirtschaft und verfügt über langjährige Erfahrung bei der Beratung und Umsetzung von Projekten im Technologieumfeld.

Claudio Martens

Er ist für das Fraunhofer IAIS im Team MLOps in der Beratung und Umsetzung sowie als Dozent für Schulungen im Bereich MLOps aktiv. Zu seinen Forschungsinteressen gehören Continuous Training, Model Monitoring und Drift Detection sowie Trustworthy AI Engineering und der Transfer dieser Themen in die Thematik LLMOps.

Dr. Christian Temath

Der KI.NRW-Geschäftsführer arbeitet mit seinem Team daran, die Marke »KI made in NRW« zu etablieren und die technologische Souveränität NRWs zu stärken. Als promovierter Wirtschaftsinformatiker verfügt er über langjährige Erfahrung in der Managementberatung im Bereich Technologie sowie in der praktischen Anwendung von KI-Technologien.

Dr. Dennis Wegener

Er leitet das Team Machine Learning Operations am Fraunhofer IAIS. Zu seinen Forschungsinteressen gehört alles, was mit MLOps zusammenhängt. Mit seinem Team führt er Forschung, Beratung und Implementierung von Data-, Trainings- und Deploymentpipelines für Kund*innen durch und unterrichtet MLOps im Rahmen der Fraunhofer-Allianz Big Data und KI.

Alexander Zimmermann

Er ist Co-Lead der Abteilung Enterprise Information Systems am Fraunhofer IAIS. Zu seinen Forschungsinteressen zählt die Entwicklung neuer datengetriebener Geschäftsmodelle zur Integration von KI in Unternehmen. Mit seiner Abteilung untersucht er hierzu von KI unterstützte Ansätze für Industrie und Forschung.

Alexander Zorn

Er arbeitet als MLOps Engineer und Data-Scientist mit dem Schwerpunkt Natural Language Understanding und ist verantwortlich für das »Innovation Briefing Generative KI«. Seit Ende 2020 ist er am Fraunhofer IAIS tätig. Zuvor hat er an der Rheinischen Friedrich-Wilhelms-Universität Bonn Mathematik und Informatik studiert.

Akteure im Überblick

Die Kompetenzplattform KI.NRW

Exzellenz vernetzen. Sichtbarkeit schaffen. Spitzenposition stärken.

Die Kompetenzplattform KI.NRW baut Nordrhein-Westfalen zu einem bundesweit führenden Standort für angewandte Künstliche Intelligenz aus und etabliert das Land in internationalen Netzwerken. Als zentrale Landes-Dachorganisation für Künstliche Intelligenz vereint KI.NRW den Dreiklang aus Spitzenforschung, Innovation und Unternehmertum. Ziel ist es, den Transfer von KI aus der Spitzenforschung in die Wirtschaft zu beschleunigen, eine Leitregion für berufliche Qualifizierung in KI aufzubauen und Impulse im gesellschaftlichen Dialog zu setzen. Dabei stellt KI.NRW den Menschen in den Mittelpunkt einer vertrauenswürdigen KI.

KI.NRW wird gefördert durch die Landesministerien MWIKE und MKW und geleitet von einem der europaweit führenden Forschungsinstitute auf den Gebieten der angewandten Künstlichen Intelligenz und des Maschinellen Lernens, dem Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS in Sankt Augustin.

www.ki.nrw

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme

Intelligent Systems that Work!

Als Teil der größten Organisation für anwendungsorientierte Forschung in Europa ist das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS mit Sitz in Sankt Augustin/Bonn und einem Standort in Dresden eines der führenden Wissenschaftsinstitute auf den Gebieten Künstliche Intelligenz, Maschinelles Lernen und Big Data in Deutschland und Europa. Rund 350 Mitarbeitende unterstützen Unternehmen bei der Optimierung von Produkten, Dienstleistungen und Prozessen sowie bei der Entwicklung neuer digitaler Geschäftsmodelle. Das Fraunhofer IAIS gestaltet die digitale Transformation unserer Arbeits- und Lebenswelt: mit innovativen KI-Anwendungen für Industrie, Gesundheit und Nachhaltigkeit, mit zukunftsweisenden Technologien wie großen KI-Sprachmodellen oder Quantum Machine Learning, mit Angeboten für die Aus- und Weiterbildung oder für die Prüfung von KI-Anwendungen auf Sicherheit und Vertrauenswürdigkeit.

www.iais.fraunhofer.de

1 MLOps als Treiber für den erfolgreichen Einsatz von KI in Unternehmen

Machine-Learning-Modelle in den produktiven Einsatz zu bringen, stellt auch in der Softwareentwicklung erfahrene Unternehmen vor neue Herausforderungen. Denn ML-Artefakte unterscheiden sich von solchen Artefakten, die Bestandteile klassischer Software sind, da sie im Wesentlichen von der Struktur und Qualität der ihnen zugrunde liegenden Daten abhängen. Während des Modelltrainings werden Auszüge von Echtdaten verwendet, um die darin enthaltenen Muster mithilfe statistischer Methoden zu bewahren. Dieser Vorgang wird als Training des Modells bezeichnet und resultiert in einem trainierten Modell, welches dazu verwendet wird, basierend auf den gelernten Mustern, Vorhersagen zu erzeugen, wie beispielsweise eine Klassifizierung.

Sowohl Training als auch Vorhersage bedingen, dass die Echtdaten zunächst in eine maschinenlesbare Form übertragen werden, damit sie durch die verwendeten Algorithmen interpretiert und die Muster in den Daten möglichst gut herausgearbeitet werden können. Diese Feature-Building-Prozesse werden in Pipelines erfasst und sind integraler Teil einer späteren ML-Lösung, da die exakt gleichen Pipelines angewendet werden müssen, um im Betrieb Daten aufbereiten zu können und durch das ML-Modell eine Vorhersage zu erzeugen. Schon bei den ersten Versuchen in der Experimentierphase werden somit Grundlagen für einen späteren produktiven Einsatz eines Modells gelegt.

Der Entwicklungsvorgang von ML-Modellen, geprägt durch einen engen Austausch zwischen den Fachabteilungen und dem Bemühen, die Daten und deren Struktur zu verstehen, wurde in einem Vorgehensmodell festgehalten. CRISP-DM (CRoss Industry Standard Process for Data Mining) beschreibt das gängige Vorgehen und die notwendigen (iterativen) Schritte, welche durchlaufen werden.

CRISP-DM (CRoss Industry Standard Process for Data Mining) beschreibt das gängige Vorgehen und die notwendigen (iterativen) Schritte für die Entwicklung von ML-Modellen.

In der Praxis werden aber nicht nur ML-Modelle benötigt, sondern Anwendungen, die durch definierte und dokumentierte Schnittstellen die Interaktion mit den verschiedenen Funktionen des ML-Modells erlauben und zusammen mit dem Modell die ML-Lösung darstellen. Ein ML-Modell muss daher in eine Software eingebettet werden, die bestimmte Aufgaben übernimmt, wie die Verarbeitung von Anfragen, die Erzeugung eines verständlichen Outputs und das Logging. Diese Software muss konzeptioniert, entwickelt, getestet und überwacht werden – es handelt sich also um klassische Aufgaben, die aus der Softwareentwicklung bekannt sind.

An diesem Punkt treffen zwei Welten aufeinander: auf der einen Seite die eher wissenschaftlich-mathematisch geprägte Arbeitsweise von Data-Science-Teams und auf der anderen Seite die durch lange Erfahrung, strenge Vorgehensmodelle und Automatisierung geprägte Welt der Softwareentwicklung. Hier hat sich DevOps (Development and Operations) als ein dominantes Vorgehensmodell etabliert. DevOps ist ein Paradigma für die Entwicklung und den Betrieb von Software, das dem Konflikt zwischen Entwicklungs- und Operationsteams durch agile Arbeit, übergreifende Verantwortlichkeiten und Prozesse sowie durch die möglichst umfassende Automatisierung manueller Tätigkeiten begegnet.

DevOps (Development and Operations) ist ein Paradigma für die Entwicklung und den Betrieb von Software.

MLOps greift dieses Paradigma auf und erweitert es durch ML-spezifische Tätigkeiten. Diese umfassen beispielsweise die Datenversionierung in Feature Stores, das Modellmanagement in Model Stores und das Monitoring fachlicher ML-Metriken während des Testens und im Betrieb. DevOps liefert die etablierten Best Practices aus der Softwareentwicklung, und in MLOps werden diese auf die besonderen Herausforderungen im Bereich des Maschinellen Lernens angepasst.

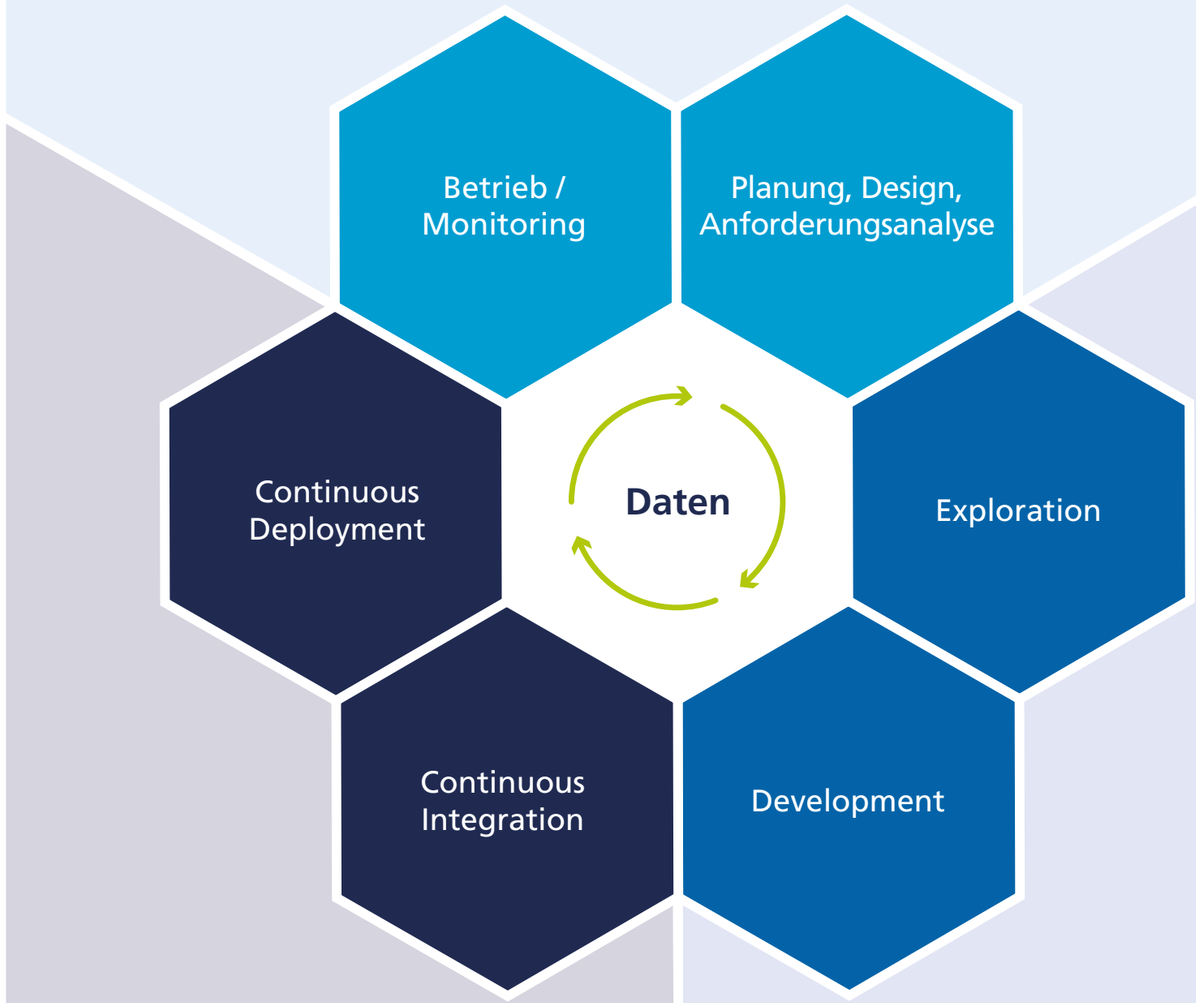


Abb. 1: MLOps-Lebenszyklusmodell nach Beck et al[3]; eigene Darstellung (rechts)

3 Beck, N., Martens, C., Sylla, K.-H., Wegener, D. und Zimmermann, A. (2020).

1.1 Die MLOps-Lebenszyklusphasen

Die Phasen des MLOps-Zyklus fassen Tätigkeiten zusammen, die während eines ML-Anwendungsprojekts notwendig werden. Man unterscheidet zwischen den Phasen »Anforderungsanalyse, Planung und Design«, »Exploration«, »Development«, »Continuous Integration (CI)«, »Continuous Deployment (CD)« und »Betrieb und Monitoring«. Die Daten bilden dabei ein wesentliches verbindendes Element zwischen den Phasen.

Jede Phase umfasst Tätigkeiten aus dem Data-Science-Bereich und der Softwareentwicklung nach DevOps. So wird beispielsweise in der Explorationsphase, in welcher hauptsächlich das Datenverständnis und das Experimentieren mit verschiedenen Algorithmen im Vordergrund steht, ebenfalls darauf geachtet, dass der Code für die Experimente getestet und versioniert wird. Auch die Ergebnisse und Artefakte durchgeführter Experimente werden versioniert und archiviert. Dies soll nach Möglichkeit nicht manuell geschehen, sondern mithilfe spezialisierter Softwaretools, wie MLflow. In Abbildung 1 geben wir eine Übersicht über den MLOps-Zyklus. Eine Auswertung entlang der Phasen des MLOps-Zyklus erfolgt im anschließenden Kapitel 2.

1.2 Einsatz im Unternehmen

Viele KI-Projekte scheitern bereits weit vor der Markteinführung bzw. dem produktiven Einsatz oder leiden unter steigenden Kosten und sich verschiebenden Deadlines. MLOps bietet hierfür passgenaue Lösungen, Vorgehensempfehlungen und geeignete Tools an mit dem Ziel, qualitativ hochwertige ML-Lösungen zügig in den Betrieb zu bringen. Dort sorgen bewährte Monitoring- und Fail-Safe-Prozesse für einen robusten Einsatz. Allerdings ist ein hohes Maß an Expert*innenwissen der Mitarbeitenden gefragt, die die notwendigen Prozesse und Tools einführen und betreiben sollen. Obwohl die Notwendigkeit für einen strukturierten Prozess der Anwendungsentwicklung mit ML-Komponenten offensichtlich erscheint, kann die hohe Eintrittshürde und die damit einhergehende Investition abschreckend wirken. Unternehmen, die MLOps bereits frühzeitig etablieren und bei der Entwicklung und Einführung von KI verwenden, profitieren jedoch von den sich etablierenden Standards und der Möglichkeit, zügig weitere KI-Use-Cases anzugehen und in überschaubaren Zeitspannen in den Betrieb zu bringen. Die Investition in MLOps zahlt sich somit rasch durch Effizienzgewinne, Qualitätsoptimierung und neue Produkte, aber auch durch Vorteile wie bessere Skalierbarkeit, Überwachung und bessere Zusammenarbeit zwischen den Teams aus.

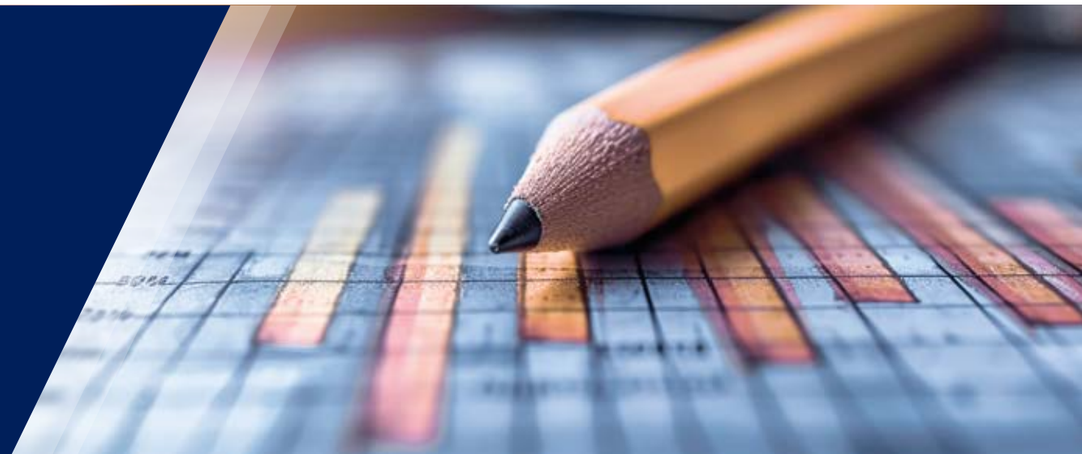
2 Erkenntnisse aus den Interviews entlang der Phasen des MLOps-Lebenszyklus

Um den aktuellen Stand der Entwicklung und den Bedarf an Unterstützung, speziell im Mittelstand, zu ermitteln, führten die KI-Experten des Fraunhofer IAIS gemeinsam mit der Kompetenzplattform KI.NRW zwischen Mai 2022 und Juni 2023 Interviews mit insgesamt 29 Unternehmen, die sich mit dem Thema MLOps befasst haben. Ein Großteil der Interviewfragen wurde in offener Form gestellt, um die Aussagekraft der Gespräche zu erhöhen. Die Ergebnisse der Studie sollen eine Hilfestellung für all diejenigen Unternehmen bieten, die den nächsten Schritt zur Operationalisierung von Machine Learning (MLOps) gehen wollen.

Bei der Auswahl der Teilnehmenden wurden bewusst Unternehmen unterschiedlicher Größe und Branchenzugehörigkeit gewählt. Im Fokus standen vor allem solche, die sich bereits mit der Operationalisierung von Machine Learning befasst haben. Dabei war es zunächst unerheblich, wie lange sich die Unternehmen mit dem Thema bereits befasst haben bzw. wie weit sie in ihrer Entwicklung sind. Denn die dabei entstehenden Herausforderungen und Konzepte liefern wertvolle Hinweise beim Aufbau von ML-Lösungen in Unternehmen.

Die Einbeziehung von Interviewpartner*innen von Start-ups bis hin zu Großunternehmen erlaubt eine Betrachtung über unterschiedliche Organisationsmodelle. So sind die Arbeitsprozesse in großen Unternehmen häufig kleinteiliger organisiert, was einen höheren Spezialisierungsgrad erlaubt, jedoch häufig mit höherer Schnittstellenkomplexität einhergeht.

Die im Rahmen der Studie durchgeführten Interviews folgten in ihrem Aufbau den verschiedenen Phasen des MLOps-Zyklus. Die nachfolgenden Auswertungen zeigen visuell und mit Erläuterungen die Erkenntnisse, welche durch die Interviews gewonnen werden konnten. Ein besonderer Fokus wird dabei auf die verwendeten Verfahren und Tools gelegt, die in den Unternehmen zum Einsatz kommen. Soweit notwendig, wurden zur Aufrechterhaltung der Anonymität und Aussagekraft solche Antworten, die nur einmal genannt wurden, unter die Kategorie »Sonstiges« gefasst.



2.1 Phase 1: Anforderungsanalyse, Planung und Design

Im Rahmen der Projektumsetzung, insbesondere bei der Entwicklung von KI-Anwendungen, nimmt die Anforderungsanalyse eine zentrale Rolle ein. Sie dient dazu, die angestrebten Ergebnisse des Projekts zu erfassen und zu definieren. Dabei stehen verschiedene etablierte Verfahren zur Verfügung, wobei im DevOps-Bereich, und damit auch bei MLOps, eine agile Herangehensweise bevorzugt wird. Ein entscheidender Aspekt der Anforderungsanalyse besteht darin, die relevanten Stakeholder einzubeziehen. Durch ihre Mitwirkung wird sichergestellt, dass das Projekt angemessen ausgestattet wird und die erforderlichen Personen zur Verfügung stehen. Neben den technischen Anforderungen ist es daher von großer Bedeutung, die organisatorischen Rahmenbedingungen zu berücksichtigen. Während dieser Phase werden die Grundlagen für die Entwicklung von KI-Anwendungen gelegt. Eine gründliche Anforderungsanalyse bildet dabei das Fundament für eine strukturierte und gezielte Entwicklung.

Im Rahmen der Interviews wurde in dieser Phase besondere Aufmerksamkeit auf die organisatorische Einbettung der KI-Anwendungsentwicklung in den Unternehmen gelegt sowie darauf, welche Grundlagen bereits existieren.

Im Folgenden werden im Anschluss an die jeweiligen Kernfragen an die Unternehmen die Ergebnisse ausgewertet.

In welchen Anwendungsdomänen bewegen sich die Unternehmen?

Machine Learning hat das Potenzial, in verschiedenen Bereichen große Performancezugewinne zu erreichen. Jedoch ist nicht jedes Feld für jedes Unternehmen gleich interessant, und es ist stark vom jeweiligen Geschäftskontext abhängig, für welche Richtung(en) sich ein Unternehmen entscheidet. Auf die Frage, in welcher Domäne die befragten Unternehmen aktiv sind, haben sich vier große Bereiche herauskristallisiert. Bei der Befragung waren Mehrfachnennungen möglich (s. Abb. 2).

Die Domäne »Textverarbeitung« wurde am häufigsten genannt, gefolgt von »Kundenanalyse«, »Bildverarbeitung« und »Kostenmanagement«.

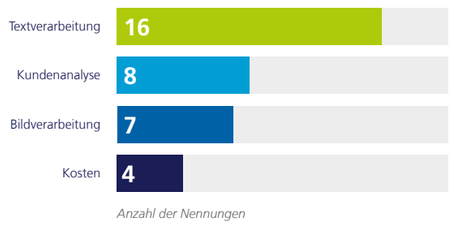


Abb. 2: Domänen der Unternehmensaktivitäten; Mehrfachnennungen möglich

Unter Textverarbeitung wurden insbesondere die Teilbereiche Natural Language Processing (NLP) bzw. Natural Language Understanding (NLU) genannt. Diese umfassen unterschiedlichste Tätigkeiten, wie die Extraktion wichtiger Informationen, Stimmungsanalyse, Textgenerierung oder Klassifizierung. Die hohe Bedeutung bekommt die Domäne durch die Möglichkeiten, die sich hierdurch bei der Reduzierung manueller Tätigkeiten ergeben. Das Lesen, Analysieren und Interpretieren von Texten betrifft eine Vielzahl von berufsbezogenen Tätigkeiten, und die maschinelle Unterstützung verspricht Kostenreduzierung, Effizienzsteigerung und Fehlerminimierung.

Auch die anderen Tätigkeitsfelder bewegen sich in Domänen, die einen hohen Return on Investment versprechen. Bildverarbeitung ermöglicht es Maschinen, visuelle Informationen aus Bildern oder Videos zu verstehen und zu interpretieren, ähnlich wie es mit NLP/NLU bei Text der Fall ist. Durch den Einsatz von Algorithmen und Techniken des Maschinellen Lernens kann Bildverarbeitung typisch menschliche Aufgaben wie Objekt-, Gesichts- und Texterkennung sowie Szenenverständnis automatisieren.

Die Analyse von Kund*innen und Kostenstrukturen erlaubt effizientes Planen und eine sparsame Ressourcenverwendung.

Wer entwickelt KI-Anwendungen im Unternehmen?

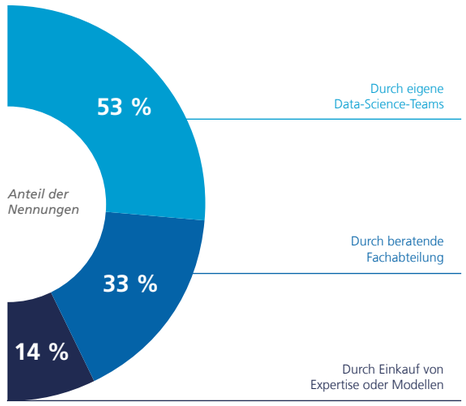


Abb. 3: Entwicklung von KI-Anwendungen

Mehr als die Hälfte der befragten Unternehmen (53 Prozent) verfügen über spezialisierte Teams, die sich mit der Entwicklung von KI-Anwendungen befassen. Sie verfügen über die nötigen Ressourcen und das Fachwissen, um KI-Anwendungen selbstständig zu entwickeln. Einige Unternehmen (33 Prozent) haben ihre Data-Science-Teams in interne Beratungseinheiten ausgegliedert, die unternehmensweit Projekte umsetzen. Seltener kommt es vor, dass Data-Science-Expertise oder Modelle von Externen eingekauft werden (14 Prozent), siehe Abbildung 3.

Wie wird die Entwicklung organisiert?

Im DevOps und MLOps werden agile Herangehensweisen im Projektmanagement bevorzugt. Trotzdem sind auch klassische Ansätze, wie das Wasserfallmodell, weiterhin verbreitet. Die beteiligten Unternehmen wurden gefragt, wie sie bei der Organisation der KI-Entwicklungsprojekte vorgehen. Dabei wurden die drei Möglichkeiten »Klassisch«, »Agil« und »Beides« vorgegeben.

Agiles Projektmanagement wird bei 61 Prozent der befragten Unternehmen präferiert. Auf ein rein »klassisches« Projektmanagement setzen nur 14 Prozent. Jedoch nutzt ein nicht zu vernachlässigender Anteil (25 Prozent) eine eigene Interpretation der beiden Ansätze (s. Abb. 4). In Unternehmen mit bereits seit langem etablierten Entwicklungsprozessen zeigt sich eine Tendenz, die etablierten, klassischen Herangehensweisen durch agile Komponenten zu erweitern, insbesondere im Kontext der Entwicklung von KI-Anwendungen.

Agil:
Hierbei handelt es sich um flexible und iterative Ansätze, wie Scrum oder Kanban, bei denen Teams in kurzen Zyklen arbeiten und sich kontinuierlich an veränderte Anforderungen anpassen können.

Beides:
Hier kommt eine Kombination von klassischen und agilen Methoden zur Anwendung, um die Vorteile beider Ansätze zu nutzen und sie an ihre spezifischen Anforderungen anzupassen.

Klassisch:
Dieser Ansatz bezieht sich auf traditionelle Entwicklungsmodelle, wie den Wasserfallansatz, bei dem Aufgaben sequenziell ausgeführt werden und jede Phase (z. B. Planung, Design, Implementierung) erst abgeschlossen wird, bevor die nächste beginnt.

Abb. 4: Organisation von KI-Projekten

2.2 Phase 2: Exploration

Während es bei der Anforderungsanalyse in Phase 1 um den Aufbau von Verständnis für die Geschäftsprozesse, das Einsatzgebiet und die Problemstellung geht, wird im Rahmen der Exploration geprüft, ob durch Machine Learning eine Lösung mit zufriedenstellender Güte erreicht werden kann und welches Verfahren dafür am geeignetsten ist. Zudem werden in Phase 2 die Grundlagen für ein erfolgreiches Training gelegt, in dem die Daten auf Muster untersucht werden. Die in Phase 1 definierten Businessmetriken werden quantifiziert, und bei dem Projektteam wird ein Verständnis für die Struktur und Zusammensetzung der Daten aufgebaut. Im Rahmen dieser Untersuchungen wird gegebenenfalls eine Vielzahl von Experimenten durchgeführt. Im MLOps-Zyklus wird, um die spätere Nachvollziehbarkeit von Ergebnissen zu verbessern, eine Versionisierung von Experimenten, Modellen und Codes empfohlen.

Nachfolgend werden im Anschluss an die jeweiligen Kernfragen die Ergebnisse ausgewertet.

Welche Rolle spielen Tools und Frameworks im Data-Science-Workspace?

Eine effektive Erprobung und Lösungsentwicklung durch Data-Scientist*innen ist ein wichtiger Faktor für den Erfolg von ML-Projekten. In diesem Kontext spielt ein Data-Science-Workspace, d. h. die technische Arbeitsumgebung mit ihren Tools und Frameworks, eine zentrale Rolle, um die Data-Scientist*innen bei der Durchführung von Experimenten und der Entwicklung von Lösungen zu unterstützen.

Auf die Frage, welches Tooling in der Explorationsphase zur Anwendung kommt, zeigt sich, dass es hier eine große Vielfalt gibt, wobei in den meisten Unternehmen Jupyter Notebook Verwendung findet. In einigen Unternehmen haben Data-Scientist*innen die Möglichkeit, die Tools ihrer Wahl zu verwenden, was zu einer großen Auswahl an genutzten Werkzeugen führt und die hohe Anzahl an verwendeten Tools erklären könnte (s. Abb. 5).

Zu den drei am häufigsten genannten Tools nach Jupyter gehören Azure, VS Code und Databricks. Weitere Tools, die mindestens zweimal erwähnt wurden, sind KNIME, PyCharm, Power BI, RStudio, AWS, Data Lake, Docker, Eigenentwicklungen und Spyder. Unter »Sonstige« wurden noch Data Factory, Google Cloud, BigQuery, Grafana, Informatica (Datenkatalog), ein yolo-basiertes KI-Framework, Kubeflow, Matlab, MLflow, Open-Source-Labeling-Tool, Poetry, PySpark, Snowflake und Vertex AI (GCP) genannt.

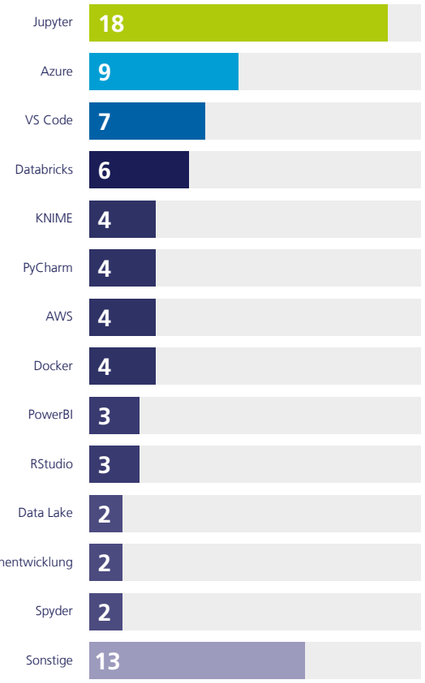


Abb. 5: Explorationstools und Frameworks im Data-Science-Workspace; Mehrfachnennungen möglich

Wie werden ML-Experimente verwaltet?

ML-Experimente werden verwaltet, um den Prozess der Modellentwicklung zu dokumentieren, reproduzierbar zu machen und die Ergebnisse zu verfolgen sowie die Trainingsläufe zu versionieren. Hierbei spricht man auch von Experimenttracking.

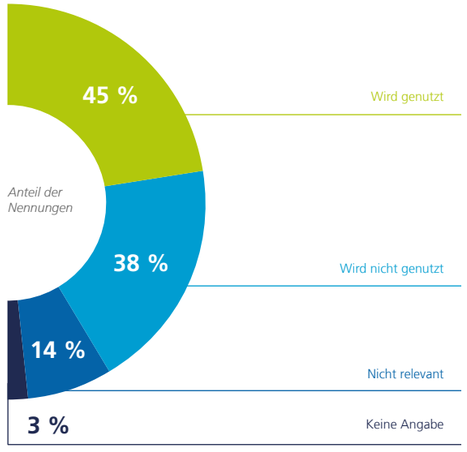


Abb. 6: Experimenttracking in der Explorationsphase

Dieses ermöglicht das systematische Erfassen von Parametern, Metriken und Artefakten während des Experimentierens oder des Modelltrainings und verbessert die Zusammenarbeit, ermöglicht Wiederholbarkeit und unterstützt bei der Optimierung von ML-Modellen.

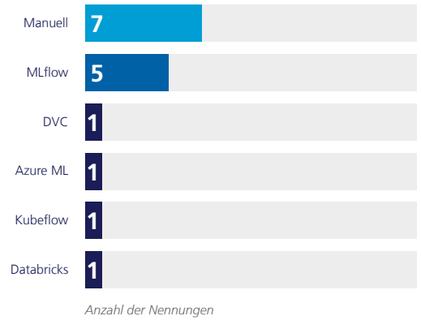


Abb. 7: Für das Experimenttracking genutzte Tools; Mehrfachnennungen möglich

Weniger als 50 Prozent der Unternehmen führen ein Experimenttracking durch. Ein Großteil erledigt dies immer noch manuell (s. Abb. 6).

Das am häufigsten genutzte Tool zur Versionisierung von Experimenten ist MLflow. Weitere verwendete Werkzeuge sind DVC, Azure ML, Kubeflow und Databricks (s. Abb. 7).

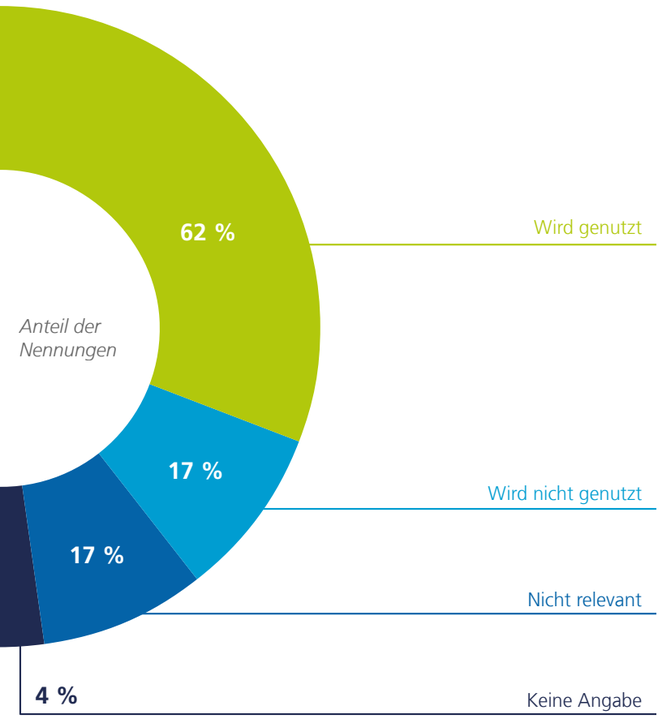


Abb. 8: Modellverwaltung in der Explorationsphase

Wie werden ML-Modelle in der Explorationsphase verwaltet?

Die Modellverwaltung spielt eine unterstützende Rolle in der Explorationsphase. Während das Experimentieren und die Entwicklung von Lösungen in dieser Phase im Vordergrund stehen, kann die Modellverwaltung dabei helfen, den Überblick über verschiedene Modellversionen und die zugehörigen Experimente und Trainingsläufe zu behalten.

Über die Hälfte der Unternehmen verwendet eine Modellverwaltung, die als versionierte Ablage für ihre Machine-Learning-Modelle dient. Die verbreitetsten Werkzeuge zur Umsetzung dieser Modellverwaltung sind Git, gefolgt von MLflow und MinIO/S3 (s. Abb. 8 und 9).

Im Vergleich zur Verfolgung der Experimentdurchläufe, die 45 Prozent der Unternehmen durchführen, versionieren also mehr Unternehmen die Artefakte, die durch die Experimente erzeugt werden. Allerdings erfolgt keine Versionierung des Experimenthintergrunds selbst, der zu diesen Ergebnissen geführt hat.

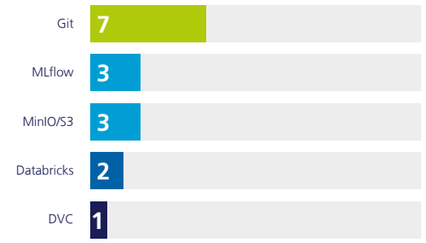


Abb. 9: Für die Modellverwaltung in der Explorationsphase genutzte Tools; Mehrfachnennungen möglich

2.3 Phase 3: Development

In der Entwicklungsphase sollen die Erkenntnisse aus der Explorationsphase genutzt werden, um Modelle zu entwerfen, zu trainieren und zu optimieren. In der vorherigen Explorationsphase liegt der Fokus auf Datenanalyse, -verständnis und -vorbereitung sowie der Wahl der Modellierungstechnik, während die Phase 3 darauf abzielt, qualitativ hochwertigen und wartbaren Code zur Umsetzung einer KI-Anwendung bereitzustellen. Im Folgenden werden nach den jeweiligen Interviewfragen die Ergebnisse ausgewertet.

Wie sieht das Tooling in der Development-Phase aus?

Bei der Analyse, ob für die Entwicklung ein einheitliches Tool genutzt wird, wurde ein heterogenes Bild festgestellt. VS Code ist zwar das am häufigsten verwendete Tool. Jedoch liegen auch alle anderen genannten Tools nur geringfügig hinter der Erstnennung (s. Abb. 10). Unter »Sonstige« wurden auch die folgenden Programmiersprachen genannt: Python, C#, R sowie diverse Libraries, auf die wir an dieser Stelle nicht genauer eingehen.

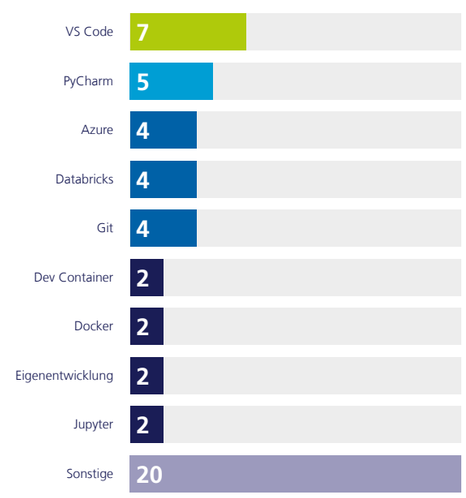


Abb. 10: In der Entwicklung eingesetzte Development-Tools; Mehrfachnennungen möglich

Welche Herausforderungen existieren beim Übergang von der Explorations- zur Development-Phase?

Bei den Herausforderungen, mit denen Unternehmen beim Übergang von der Explorationsphase zur Entwicklung konfrontiert sind, geben knapp ein Viertel der Befragten an, dass keine klare Trennung zwischen diesen beiden Phasen existiert. Häufig auftretende Schwierigkeiten beziehen sich auf Datenoperationen, wie beispielsweise Dateninkonsistenz, die Stabilität der Datenpipeline oder die Datenübertragung zwischen verschiedenen Umgebungen. Zudem werden häufig organisatorische Probleme genannt, die sich auf unklare Vorgehensweisen oder hierarchische Strukturen beziehen. Zusätzliche Herausforderungen ergeben sich in Bezug auf die entstehenden Kosten während der Entwicklung sowie Integrationsprobleme (s. Abb. 11).

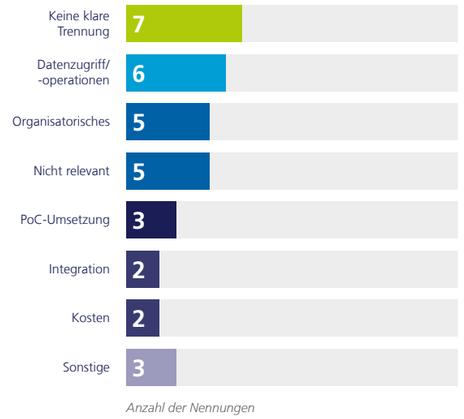


Abb. 11: Herausforderungen beim Übergang von der Explorations- zur Development-Phase; Mehrfachnennungen möglich

2.4 Phase 4: Continuous Integration (CI)

Continuous Integration bezeichnet einen Prozess, durch den unter Verwendung von Pipelines und Automatisierung kontinuierlich Code in eine gemeinsame Codebasis eingebracht werden kann. Automatisiert werden beispielsweise die Durchführung von Tests, das Sicherstellen der Codequalität sowie die Prüfung der Einhaltung vereinbarter Richtlinien. Dadurch wird die Zusammenarbeit im Team erleichtert, da automatisierte Prozesse eine einheitliche Qualität garantieren.

Im Folgenden werden im Anschluss an die jeweiligen Kernfragen an die Unternehmen die Ergebnisse ausgewertet.

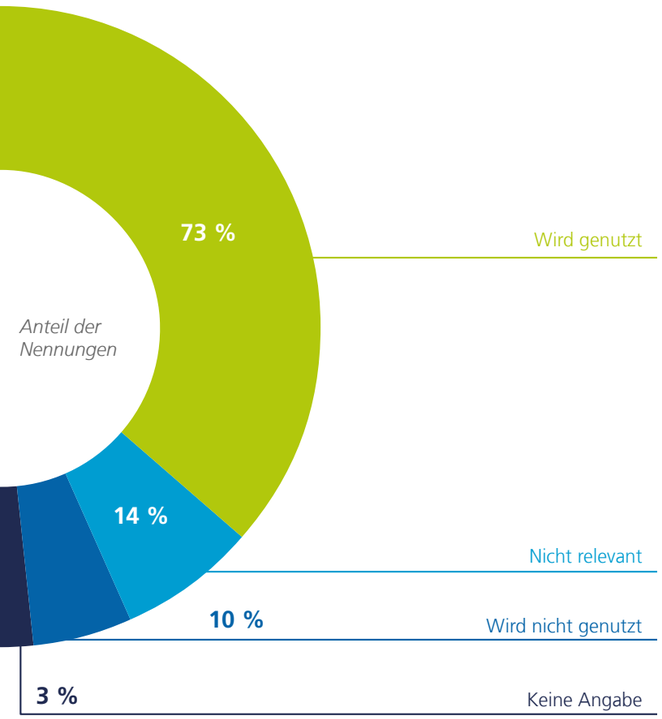


Abb. 12: Nutzung von Continuous Integration

Wie wird Continuous Integration umgesetzt?

Lediglich 10 Prozent der Unternehmen nutzen kein CI. 17 Prozent machen keine Angabe bzw. für diesen Prozentsatz war Continuous Integration nicht relevant. Die verbleibenden 73 Prozent der Unternehmen haben CI-Prozesse etabliert (s. Abb. 12).

Die in diese CI-Prozesse integrierten ML-spezifischen Anteile reichen von der Modellintegration in Anwendungen, über das Training von Modellen und die Durchführung von Funktionstests bis hin zum Retraining. Einige Unternehmen nutzen das CI-Tooling zur Steuerung und Durchführung kompletter Machine-Learning-Pipelines (s. Abb. 13).

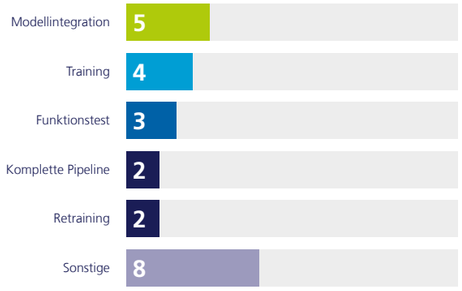


Abb. 13: In CI-Prozesse integrierte ML-spezifische Anteile; Mehrfachnennungen möglich

Welche Tools werden für Continuous Integration genutzt?

Arbeiten die Unternehmen in der Cloud, dominiert das Angebot von Azure. On-Premise wird am häufigsten GitLab CI eingesetzt. Zudem kommt im Rahmen der CI-Pipeline häufig Docker zum Einsatz, eine Technologie, die besonders geeignet ist, um gekapselte Umgebungen bereitstellen zu können (s. Abb. 14).

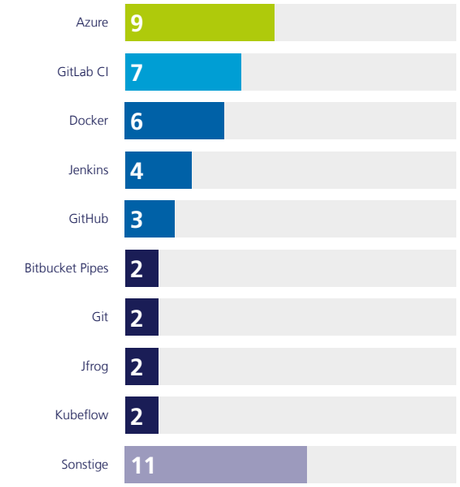


Abb. 14: Für Continuous Integration verwendetes Tooling; Mehrfachnennungen möglich

2.5 Phase 5: Continuous Deployment (CD)

Continuous Deployment unterstützt Entwicklungsteams dabei, Artefakte regelmäßig und unter hoher Qualität in Produktivumgebungen einzubringen. Die weitgehende Automatisierung der damit in Zusammenhang stehenden Prozesse minimiert Fehler und ermöglicht eine höhere Taktung, als es mit manuellen Eingriffen möglich wäre. Unter Continuous Delivery wird ein vergleichbares Vorgehen verstanden, unter der Einschränkung, dass das automatisierte Einspielen in eine Produktivumgebung fehlt. Stattdessen werden die dazu notwendigen Artefakte lediglich bereitgestellt, da ein direkter Zugriff auf Produktivumgebungen nicht immer möglich ist. Beides umfasst aber die weitgehende Automatisierung der Erstellung von notwendigen Artefakten und wird somit häufig nicht trennscharf verwendet.

Es folgen die Kernfragen an die Unternehmen sowie die Auswertung der Ergebnisse.

Wird Continuous Deployment angewendet?

Von den befragten Unternehmen sagen 66 Prozent, dass sie CD-Prozesse implementiert haben. Damit verwendet ein Großteil dieser Unternehmen ebenfalls Continuous Integration (s. Abb. 15).

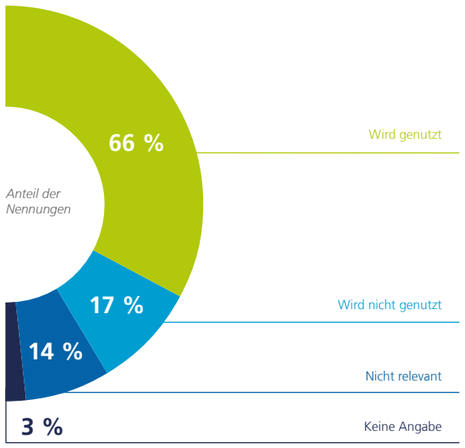


Abb. 15: Anwendung von Continuous Integration

2.6 Phase 6: Betrieb und Monitoring

In der Betriebs- und Monitoringphase werden Aspekte betrachtet, die den produktiven Einsatz von ML-Anwendungen begleiten. Von Interesse sind dabei die eingesetzten Tools und die Frage, welche Metriken im Monitoring beobachtet werden. Im Vergleich zu klassischer Software reagieren Modelle empfindlich auf Änderungen in der Produktivumgebung und insbesondere auf Änderungen in den Daten, welche als Input verarbeitet werden sollen. Daher wird neben dem Tooling ebenfalls untersucht, ob das Unternehmen in der Lage ist, einen Drift der Daten zu erkennen. Als Drift wird eine Verteilungsänderung bzw. Ausprägung der gewöhnlich als Input einkommenden Daten verstanden. Im Falle eines Datendrifts greifen die gelernten Muster nicht mehr ausreichend, und das Modell wird schlechter bei der Interpretation von betroffenen Daten. Von einem Concept-Drift wird gesprochen, wenn sich die Daten so grundsätzlich von den Trainingsdaten unterscheiden, dass ein sinnvoller Einsatz des trainierten Modells nicht mehr möglich ist.

Nachfolgend finden sich die gestellten Kernfragen und die Auswertung der Ergebnisse.

Betrieb

Wie viele Modelle sind bei den verschiedenen Unternehmen in Betrieb?

Die Anzahl der in Betrieb genommenen ML-Modelle innerhalb eines Unternehmens kann verschiedene Herausforderungen mit sich bringen. Es kann den Ressourcenbedarf, die Komplexität der Modellverwaltung und den Aufwand für Wartung und Aktualisierungen erhöhen. Eine sorgfältige Betrachtung und Planung dieser Aspekte sind wichtig, um potenziellen Herausforderungen erfolgreich begegnen zu können.

Von den befragten Unternehmen betreiben über 76 Prozent Modelle in einer produktiven Umgebung. Sechs der insgesamt 29 befragten Unternehmen betreiben ML-Anwendungen im großen Maßstab, mit mehr als 50 oder mehr als 100 Modellen im Einsatz. Ein Drittel der Befragten hat weniger als zehn Modelle im Betrieb, und etwas weniger als ein Drittel betreiben eine niedrige zweistellige Anzahl (s. Abb. 16).

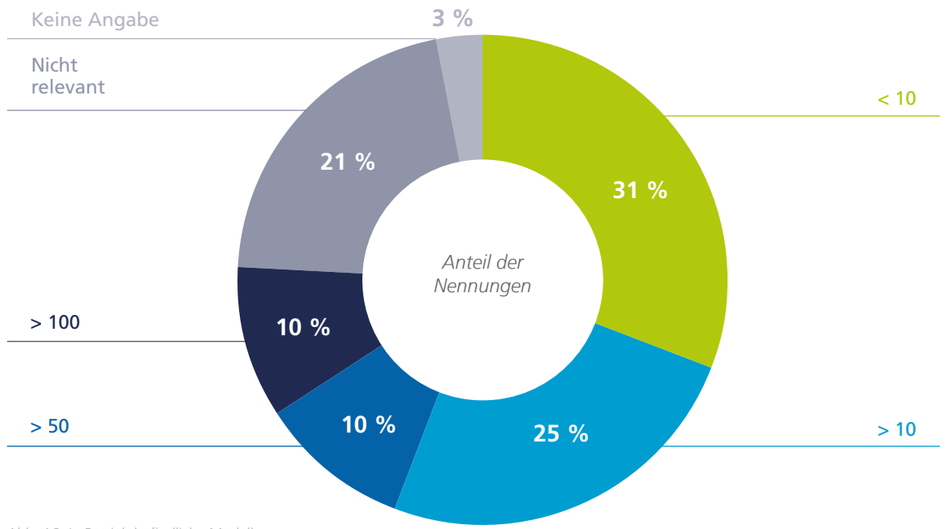


Abb. 16: In Betrieb befindliche Modelle

Welches Tooling wird verwendet, um ML-Anwendungen produktiv zu setzen?

Unternehmen setzen beim Betrieb von ML-Anwendungen auf Cloud-Lösungen, insbesondere Azure wird als Anbieter genannt. Die Containerisierung von produktiven Anwendungen setzt sich auch im Bereich ML durch, und Technologien wie Docker und Kubernetes stellen gängige Methoden dar. Selten wurde von den Unternehmen der Einsatz von dedizierten MLOps-Plattformen genannt, wie z. B. DataDog, welches nur bei einem der befragten Unternehmen zum Einsatz kommt (s. Abb. 17).

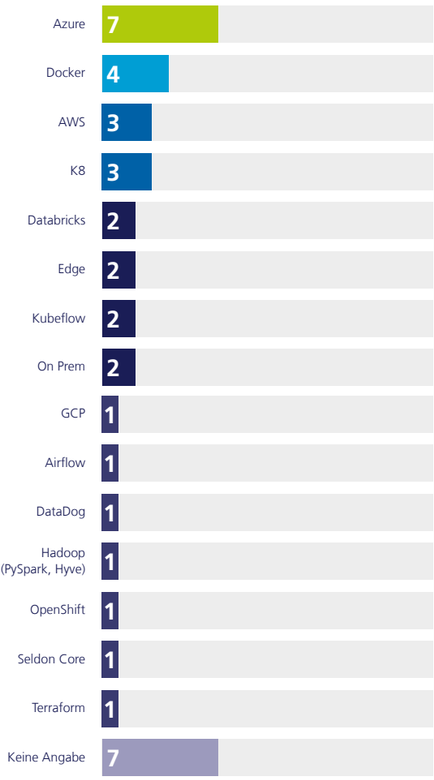


Abb. 17: Für den Betrieb verwendetes Tooling, um ML-Anwendungen produktiv zu setzen; Mehrfachnennungen möglich

Monitoring

Welches Tooling wird verwendet, um ML-Anwendungen zu überwachen?

Automatisiertes Monitoring hat sich bei den befragten Unternehmen noch nicht in der Breite durchgesetzt. Ein Großteil verlässt sich bei der Überwachung der Anwendung auf das Feedback von Endnutzer*innen und Entwickler*innen. Unternehmen, die sich beim Betrieb der ML-Modelle auf Cloud-Anbieter verlassen, greifen in der Regel auf deren angebotene Tools zurück. Hier zeigt sich wieder die deutliche Dominanz, die Azure in der befragten Gruppe innehat. Cloud unabhängige Unternehmen setzen in der Regel auf Open-Source-Lösungen, wobei neben Azure Tools auch Prometheus und Grafana dominieren (s. Abb. 18).

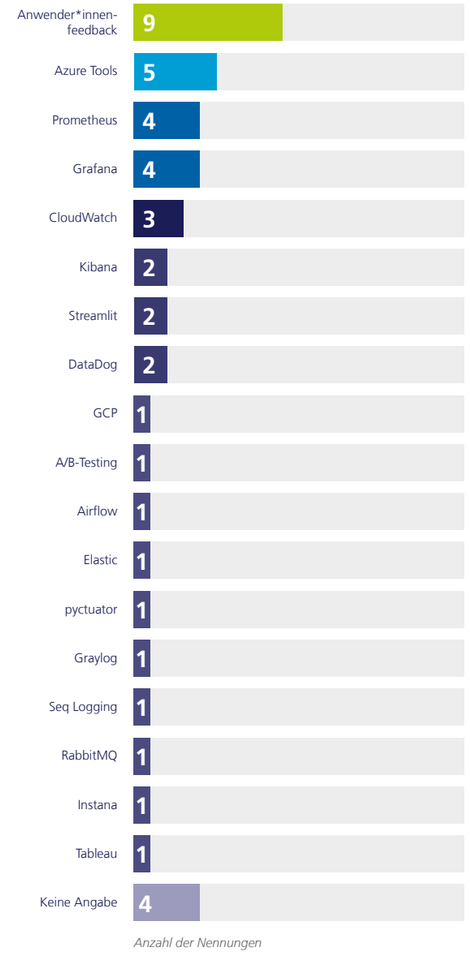


Abb. 18: Für das Monitoring verwendete Tools zur Überwachung von ML-Anwendungen; Mehrfachnennungen möglich (rechts)

Können Data- und Concept-Drifts erkannt werden?

Auf die Frage, ob die Unternehmen in der Lage sind, einen Data- oder Concept-Drift zu erkennen, antworten lediglich 17 Prozent mit »Ja, wird erkannt«. 14 Prozent können das Auftreten eines Drifts händisch erkennen, wobei sie die Daten prüfen, sobald Nutzer*innen eine Verschlechterung der Leistung melden. 7 Prozent der Unternehmen planen, sich mit dem Thema auseinanderzusetzen und die notwendigen Fähigkeiten aufzubauen, während 52 Prozent der Befragten mit »Nein, wird nicht erkannt« antworten (s. Abb. 19).

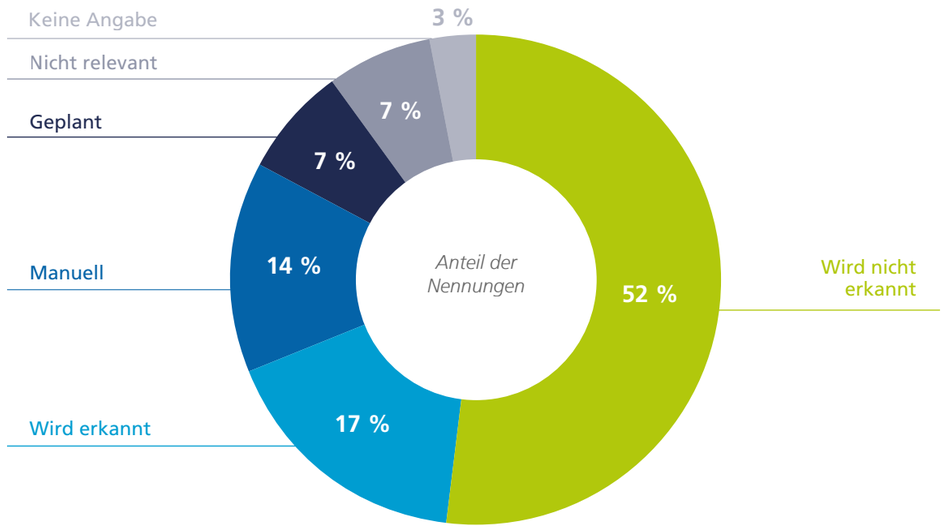


Abb. 19: Erkennung von Data- und Concept-Drifts

3 Erkenntnisse aus weitergehenden Fragestellungen

Im Folgenden widmen wir uns einigen übergreifenden Fragestellungen, die über die statistische Auswertung der Interviewfragen zu den einzelnen Phasen des MLOps-Zyklus hinausgehen. Je Kapitel werden dabei nennenswerte Trends und Erkenntnisse beschrieben, welche im Rahmen der Interviews gewonnen werden konnten.

3.1 Welche Herausforderungen sehen die Unternehmen selbst, und was steht auf deren Roadmap?

Herausforderungen

Die größte Herausforderung aktuell wird von den Unternehmen im Bereich Daten gesehen, genauer in deren Qualität und Management. Dies umfasst neben der reinen Verfügbarkeit von Daten auch die Einhaltung des Datenschutzes je nach Use Case bis hin zum Fehlen oder derzeitigen Aufbau einer übergreifenden Datenstrategie im Unternehmen.

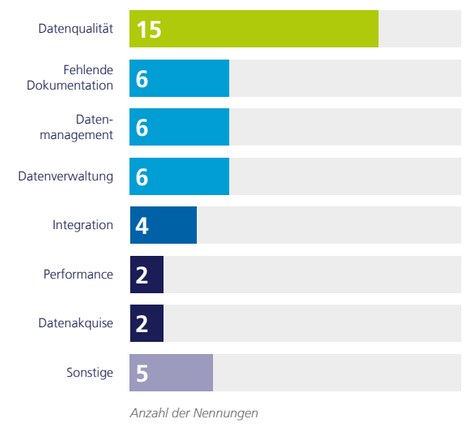


Abb. 20: Von den Unternehmen genannte Herausforderungen; Mehrfachnennungen möglich

Eine Einzelangabe, die heraussticht, aber gut mit den Angaben aus dem Bereich »Herausforderungen« korreliert, ist das Thema Daten. Viele Unternehmen wollen sich hier mit dem Datenmanagement, Datenstrategien, der Datenqualität und -verfügbarkeit beschäftigen.

Neben den Herausforderungen, die sich durch die Aufbereitung, Versionisierung und Bereitstellung der Daten ergeben, nennen viele Unternehmen auch solche, die sich durch unternehmensinterne Prozesse ergeben. Der Aufbau von Abteilungs-, Team- und Rollenstrukturen ist ebenso im Fokus wie eine bessere Einbindung der Fachseite oder der Aufbau einer skalierbaren technischen Infrastruktur mit hohem Automatisierungsgrad.

Zusätzlich haben ML-Lösungen Probleme in der Selbstvermarktung. Sowohl unternehmensintern als auch von externen Parteien werden neu entwickelte ML-Lösungen nicht umfassend akzeptiert. Die geringe Akzeptanz wird dabei vor allem mit fehlendem Vertrauen in und Verständnis für die ML-Lösungen begründet. Bei den äußeren Einflüssen werden die sehr volatile und schnelllebige Lage im KI-Bereich sowie Regulierungen, wie der EU AI Act, und die schwierige Situation am Personalmarkt angeführt.

Roadmap

Gefragt nach den geplanten nächsten Schritten und ihrer Roadmap gab ein Großteil der befragten Unternehmen die technische Skalierung an. Hierunter fallen vor allem die Automatisierung und Standardisierung von Entwicklungs- und Betriebsprozessen als auch die Identifizierung geeigneter ML-Tools, die solche Prozesse unterstützen, bis hin zu kompletten ML-Plattformen.

Flankiert wird dies für viele Unternehmen durch die fachliche Skalierung, d. h. auf der einen Seite Know-how-Aufbau im Unternehmen und auf der anderen Seite eine Vergrößerung der Anzahl an ML-Use-Cases. Wichtig wird hier auch für viele Unternehmen die Erhöhung des Business-Value, was dafürspricht, dass die ML-Lösungen aus der Forschung in den operativen Betrieb übergehen werden müssen.

Als zukünftige KI-Themen entlang der Roadmaps wurden von den befragten Unternehmen auch Themengebiete genannt, die sich aktuell noch in der Forschung befinden, wie z. B. die nachhaltige KI-Entwicklung und Erklärbarkeit sowie langfristige

Ansätze (Digitalisierungsroadmaps oder »AI-First«-Ansätze). Insbesondere zum Ende der Interviewphase gelangten die Themen Large Language Models (LLMs) und Generative KI mehr in den Fokus.

3.2 Auf welchem Stand sind die Unternehmen hinsichtlich des MLOps-Reifegrads?

Im Folgenden wird der MLOps-Reifegrad der befragten Unternehmen bewertet. Dieser basiert darauf, wie gut die Unternehmen im Hinblick auf die sechs Phasen des MLOps-Zyklus und die Datenbasis aufgestellt sind. Im Detail flossen unter anderem die folgenden Kriterien in die Bewertung ein:

- die Strukturierung innerhalb der Unternehmen hinsichtlich Organisation und Prozesse
- die Standards bzw. das Tooling innerhalb der einzelnen Phasen
- der Grad der erreichten Automatisierung für die Phasen CI, CD und Betrieb

Falls einzelne Phasen für die Unternehmen nicht relevant waren oder sie keine Angaben zu bestimmten Phasen gemacht haben, wurden diese Antworten aus der Wertung genommen.

Die Ergebnisse der Bewertung des MLOps-Reifegrads haben wir in einer Fünf-Sterne-Skala zusammengefasst (s. Abb. 21).

Es zeigt sich, dass die meisten Unternehmen eine Bewertung von 4 oder 4,5 Sternen bekommen haben. Dies ist insoweit nicht verwunderlich, da explizit Unternehmen angesprochen wurden, die sich bereits mit MLOps beschäftigen. Die Unternehmen mit mehr als 4 Sternen haben meist einen soliden Ansatz zur Umsetzung von MLOps, ein geeignetes Rollenmodell und ein gutes technisches Setup.

Erfahrungsgemäß sind Unternehmen hinsichtlich der Umsetzung des MLOps-Zyklus ab einer Bewertung von 3 Sternen arbeitsfähig. Das bedeutet, dass sie grundsätzlich in der Lage sind, KI- bzw. ML-Modelle selbst zu entwickeln, diese in den Betrieb zu überführen und zu warten. Aufgrund des Reifegrads kann es vorkommen, dass einige Arbeiten noch manuell durchzuführen sind. Dennoch zeigt sich, dass nicht alle Phasen perfekt umgesetzt sein müssen, um produktiv arbeiten zu können.

Abb. 21: MLOps-Maturity-Modell: Bewertung des Reifegrads anhand einer Fünf-Sterne-Bewertungsskala

Bei der Betrachtung des Reifegrads pro Phase, gemittelt über die Bewertungen der Unternehmen, ergibt sich ein differenziertes Bild (s. Abb. 22).

In der Phase »Betrieb und Monitoring« bzw. beim Thema »Daten« war die Bewertung der Unternehmen deutlich schlechter als für die anderen Phasen.

Das Thema »Daten« hat die schlechteste Bewertung erhalten, d. h., hier bestehen die größten Herausforderungen bzw. die Umsetzung ist dort noch nicht so »perfekt« wie für andere Phasen. Dies deckt sich mit der Einschätzung der Unternehmen selbst, die das Thema »Daten« als eine für sie bekannte Herausforderung genannt haben (siehe Abschnitt 3.1). An dieser Stelle kommt auch das Thema »Legacy« zum Tragen, da hier große Herausforderungen bei der Integration und dem Management von Datenbeständen aus älteren Systemen und Lösungen bestehen, welche nur in sehr zeitaufwändigen Prozessen angegangen werden können.

Das Thema »Betrieb und Monitoring« wurde von den Unternehmen eher unkritisch gesehen – dort wird viel auf Basis von Anwender*innenfeedback umgesetzt – oder (noch) nicht als Herausforderung erkannt. Im nächsten Abschnitt (3.3) werden wir auf dieses und andere Themen genauer eingehen.

Generell lässt sich sagen, dass der Reifegrad der Unternehmen nicht von der Firmengröße o. ä. abhängt, sondern vom individuellen Aufwand bzw. der Investition, die für die Umsetzung des MLOps-Zyklus getätigt wird. Die Größe der Teams bei den befragten Unternehmen, die für die Umsetzung von Data-Science- und ML-Lösungen zuständig sind, variiert zwischen drei und 200 Mitarbeitenden. Es lässt sich erkennen, dass schon wenige Mitarbeiter*innen mit ML-Erfahrung ausreichen, um erste Lösungen zu entwickeln. Jedoch wird eine Mindestgröße des Teams (bspw. > 10) benötigt, um eine größere Anzahl an Lösungen in den produktiven Einsatz zu bringen.

Zwei verschiedene Strategien stechen bei der Abdeckung des MLOps-Zyklus heraus: Einige Unternehmen verfolgen einen Ende-zu-Ende-Ansatz, d. h. Data-Scientist*innen bauen die Anwendung vom Experiment zum Deployment bzw. begleiten diese bis hin zum Betrieb. Andere Unternehmen, vor allem größere, etablieren ein Konzept, das eine (harte) Zweiteilung vorsieht: Fachseiten und Data Science bauen insbesondere innerhalb der Explorationsphase einen Proof of Concept, danach übernimmt eine andere Abteilung – entweder sind dafür Softwareentwickler*innen ab der Development-Phase zuständig oder die IT-Abteilung.

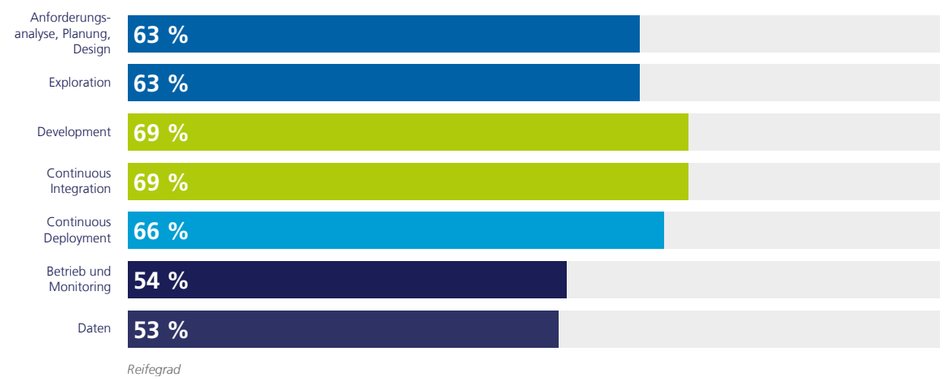


Abb. 22: Betrachtung des Reifegrads über die MLOps-Phasen und in Bezug auf die Daten

3.3 Wo weichen die Unternehmen von Empfehlungen in der Literatur ab?

Gängige MLOps-Praktiken und Empfehlungen sind in diversen Veröffentlichungen beschrieben worden, z. B. von Google[4] oder Microsoft[5]. Im Folgenden analysieren wir, welche Abweichungen sich dazu aus den Ergebnissen der Befragung ableiten lassen.

Vollautomatisierung und Continuous Applications

Die Auswertung der Interviews zeigt, dass fast alle Unternehmen eine manuelle Kontrolle vor dem tatsächlichen Deployment einer ML-Anwendung durchführen und dass nur bei wenigen von ihnen Methoden wie Continuous Training oder Continual Learning angewendet werden.

Ein Hauptgrund, warum beim Betrieb von Machine-Learning-Anwendungen die genannten Methoden noch nicht sehr weit verbreitet sind, könnte in den Herausforderungen liegen, die mit diesen Ansätzen verbunden sind. Continuous Training beinhaltet das regelmäßige Aktualisieren von Machine-Learning-Modellen, um mit sich ändernden Daten Schritt zu halten. Continual Learning geht noch einen Schritt weiter und ermöglicht es den Modellen, während des Betriebs aus neuen Daten zu lernen und sich anzupassen. Diese Ansätze erfordern eine solide Infrastruktur für Datenmanagement, Modellversionierung sowie eine kontinuierliche Integration und Bereitstellung. Zudem müssen die beteiligten Teams über das erforderliche Fachwissen und die Ressourcen verfügen, um diese Prozesse effektiv zu implementieren und zu überwachen. Da Machine-Learning-Anwendungen in vielen Unternehmen noch relativ neu sind, könnte es an Erfahrung und etablierten Best Practices fehlen. Die Nutzung von Continuous Training und Continual Learning erfordert außerdem eine sorgfältige Planung und Risikobewertung, da unkontrollierte Aktualisierungen zu unerwünschten Auswirkungen führen könnten. Obwohl diese Methoden vielversprechend sind, müssen Unternehmen möglicherweise noch Zeit investieren, um die erforderlichen Voraussetzungen zu schaffen und ihre Implementierung zu optimieren, bevor sie weiterverbreitet werden.

Trotz einer toolgestützten Deploymentpipeline kann es zudem sinnvoll sein, eine manuelle Freigabe der Deployments für Machine-Learning-Anwendungen durchzuführen. Dies liegt daran, dass Machine-Learning-Modelle in der Regel komplexe und hochdimensionale Systeme sind, die eine besondere Sorgfalt erfordern. Eine manuelle Freigabe bietet die Möglichkeit, die Ergebnisse der Modelle zu überprüfen, um sicherzustellen, dass sie den gewünschten Anforderungen entsprechen und korrekt funktionieren. Dies ist besonders wichtig, da Machine-Learning-Modelle auf großen Datenmengen trainiert werden und Fehler in deployten Modellen zu erheblichen Auswirkungen führen können. Durch die manuelle Überprüfung kann sichergestellt werden, dass das Deployment den relevanten rechtlichen, ethischen oder geschäftlichen Anforderungen entspricht. Obwohl eine manuelle Freigabe den Prozess verlangsamt, trägt sie zur Sicherheit und Qualität der bereitgestellten Anwendungen bei und ermöglicht es, potenzielle Probleme frühzeitig zu erkennen und zu beheben.

Trennung von Entwicklung und Betrieb der ML-Lösung

Die Auswertung der Studie zeigt, dass bei vielen Unternehmen eine strikte organisatorische Trennung zwischen der Entwicklung von ML-Lösungen und deren Betrieb besteht. Dies ist überraschend, da diese Trennung der Teams beim Übergang von der Entwicklung in den Betrieb einen Widerspruch zu den klassischen DevOps-Prinzipien darstellt. DevOps zielt darauf ab, die Kluft zwischen Entwicklung und Betrieb zu überbrücken und eine nahtlose Zusammenarbeit zwischen den Teams zu fördern. Dies beinhaltet den kontinuierlichen Austausch von Wissen, die Automatisierung von Prozessen und die gemeinsame Verantwortung für die Qualität und Stabilität der Anwendungen.

Das Gleiche gilt auch für den MLOps-Ansatz. Insbesondere erfordern die Entwicklung und Bereitstellung von Machine-Learning-Anwendungen spezifische Fachkenntnisse und Expertise. Eine strikte Trennung könnte zu einer Fragmentierung des Wissens und zu Engpässen bei der Zusammenarbeit führen, was die Effizienz und Effektivität beeinträchtigen könnte. Daher ist es wichtig, eine Balance zu finden, indem man die Teams in bestimmten Bereichen zusammenarbeiten lässt und gleichzeitig sicherstellt, dass ein kontinuierlicher Wissensaustausch und eine gemeinsame Verantwortung gewährleistet sind.

Notebooks-to-Code

Die Auswertung der Interviews zeigt, dass Unternehmen den Code von Jupyter Notebooks nicht direkt (mittels Tools) in produktionsreifen Code überführen (s. Abb. 23). Dessen direkte Umwandlung wird vermutlich nicht verbreitet genutzt, weil der dort verwendete Code nicht unter der gleichen Zielsetzung erstellt wurde wie Code, der produktiv genutzt werden soll. Jupyter Notebooks sind hauptsächlich für explorative Datenanalysen, Ad-hoc-Analysen und Prototyping gedacht, während produktionsfähiger Code in der Regel modular, effizient und gut dokumentiert sein muss. Die automatische Umwandlung von Notebooks kann zu inkonsistentem Code führen, der schwierig zu warten und zu debuggen ist. Es ist oft besser, ihn aus dem Notebook zu extrahieren und in einer geeigneten Entwicklungsumgebung neu zu strukturieren und zu optimieren.

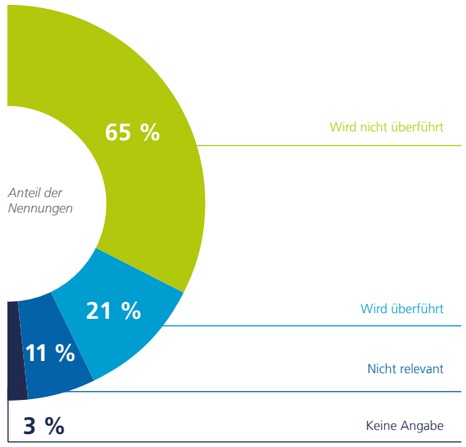


Abb. 23: Überführung von Jupyter Notebooks direkt in Produktionscode

Fachliches vs. technisches Monitoring

Bei unserer Umfrage gaben 86 Prozent der Unternehmen an, dass eine Überwachung von ML-Modellen auf Basis von IT-basierten Metriken im produktiven Einsatz durchgeführt wird. Gefragt nach den überwachten fachlichen Metriken (s. Abb. 24) sagen 62 Prozent der Befragten, dass keine festgelegten ML-Metriken im Betrieb überwacht werden. 21 Prozent der Unternehmen überwachen klassische, statistische Metriken, welche zur Performanceüberwachung von ML-Modellen verwendet werden. Beispiele sind Precision, Recall oder der f1-Score. 17 Prozent geben an, dass sie (zusätzlich) dedizierte Businessmetriken überwachen, die mit der Verwendung der Modelle in Zusammenhang stehen. Dies zeigt, dass das fachliche Monitoring von Machine-Learning-Modellen im produktiven Einsatz noch sehr unterrepräsentiert ist.

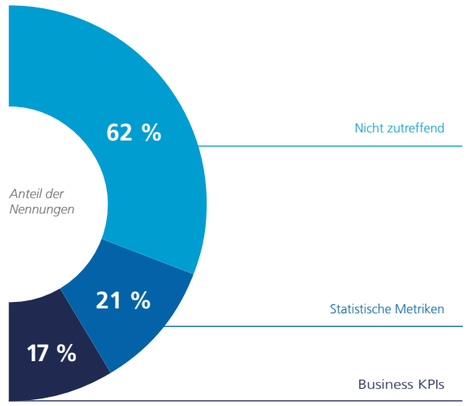


Abb. 24: Überwachte Metriken

Das fachliche Monitoring könnte aus mehreren Gründen weniger stark ausgeprägt sein als das technische Monitoring:

1. Interpretation der Ergebnisse:
Das fachliche Monitoring erfordert die Interpretation der Ergebnisse des Modells im Hinblick auf die Geschäftsziele. Es kann schwierig sein, die Vorhersagen des Modells in Bezug auf die tatsächlichen Ergebnisse zu bewerten und zu verstehen, insbesondere wenn es sich um komplexe Modelle handelt. Machine-Learning-Modelle können sehr komplex sein und basieren auf statistischen und mathematischen Algorithmen. Das fachliche Monitoring erfordert daher ein tiefes Verständnis des

4 Google (2023).
5 Microsoft (2023).

Modells und seiner Auswirkungen auf die Geschäftsprozesse. Daher kann es schwierig sein, Expert*innen in den bzw. für die Betriebsteams zu finden, die sowohl über das erforderliche Fachwissen als auch über das technische Verständnis verfügen.

2. Datenabhängigkeit:
Machine-Learning-Modelle sind stark von den Eingangsdaten abhängig. Das fachliche Monitoring erfordert die Überwachung der Qualität und Relevanz der Daten, die für das Modell verwendet werden. Hier ist zu klären, ab wann geeignete Daten für eine Validierung vorhanden sind, wie mit Ausreißern umgegangen wird und welche Abweichungen von den Testmetriken als kritisch angesehen werden. Der Aufbau eines geeigneten fachlichen Monitorings wird dadurch komplex und bedeutet einen nicht unwesentlichen Mehraufwand.

3. Kulturelle Aspekte:
In einigen Unternehmen liegt der Fokus eher auf dem technischen Monitoring, da es einfacher zu quantifizieren und von der IT zu automatisieren ist. Das fachliche Monitoring erfordert möglicherweise eine stärkere Zusammenarbeit zwischen der IT bzw. den Fachbereichen und der Data-Science-Abteilung, was kulturelle Veränderungen und eine klare Kommunikation erfordert. Zudem muss geklärt werden, wer für die Überwachung der fachlichen Metriken zuständig ist – die Data-Science- oder die IT-Abteilung.

4. Risikoaverse Auswahl umgesetzter Use Cases:
Für viele Unternehmen stellt das Thema Machine Learning ein relativ neues Feld dar. Bei der Auswahl der Use Cases wird dabei häufig auf möglichst risikoarme Lösungen gesetzt, die als Assistenzsysteme für die menschlichen Nutzer*innen fungieren. Solange die bereitgestellten Lösungen seitens der Kund*innen genutzt werden, wird häufig von ausreichender Güte des Modells ausgegangen und »Human in the Loop« als Monitoringmethode akzeptiert.

Es ist wichtig zu beachten, dass das fachliche Monitoring von Machine-Learning-Modellen dennoch von großer Bedeutung ist, um sicherzustellen, dass die Modelle korrekt arbeiten, den Geschäftsanforderungen entsprechen und keine unerwünschten Auswirkungen haben.

Nicht-Vorhandensein von Feature Stores

Feature Stores sind sinnvoll für ML-Lösungen, da sie eine zentrale und konsistente Datenquelle für Features bieten, die von verschiedenen Modellen genutzt werden können. Zudem ermöglichen sie eine einfache Wiederverwendung von Features, was die Entwicklung und Wartung von ML-Modellen erleichtert.

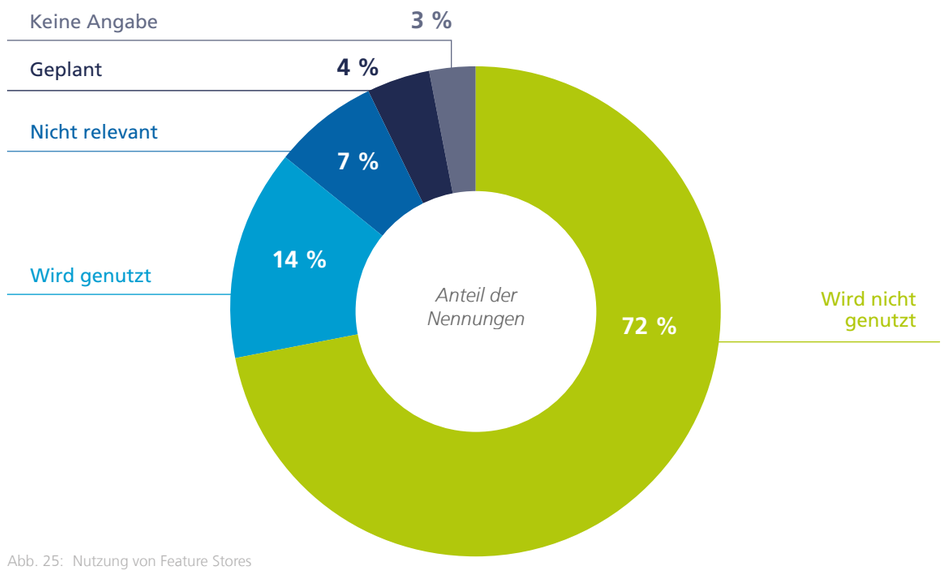


Abb. 25: Nutzung von Feature Stores

Die Auswertung der Interviews zeigt, dass Feature Stores bei den befragten Unternehmen bisher kaum genutzt werden (s. Abb. 25). Dafür könnte es mehrere Gründe geben:

1. Mangelnde Sensibilisierung:
Viele Unternehmen sind sich möglicherweise nicht bewusst, dass Feature Stores eine effektive Möglichkeit bieten, Features zu organisieren, zu speichern und für Machine-Learning-Modelle zugänglich zu machen. Es gibt möglicherweise einen Mangel an Sensibilisierung und Verständnis für die Vorteile eines Feature Stores.

2. Komplexität der Implementierung:
Die Implementierung eines Feature Stores erfordert oft eine Anpassung der bestehenden Dateninfrastruktur und Datenpipelines. Dies kann technische Herausforderungen und zusätzliche Kosten mit sich bringen. Unternehmen könnten zögern, in solche Änderungen zu investieren, insbesondere wenn sie bereits andere funktionierende Lösungen haben.

3. Mangelnde Standardisierung:
Es gibt derzeit keine einheitlichen Standards für Feature Stores, was zu Unsicherheit führen kann. Unternehmen könnten zögern, in ein bestimmtes Feature-Store-Framework zu investieren, wenn sie befürchten, dass es in Zukunft möglicherweise nicht mehr unterstützt oder von anderen Lösungen überholt wird.

4. Organisationsstruktur:
Unternehmen, die noch über keine datengetriebene Kultur verfügen, könnten Schwierigkeiten haben, die erforderlichen Veränderungen zur Nutzung von Feature Stores in der Organisation vorzunehmen. Insbesondere das Vorhandensein von siloartigen Strukturen zwischen den Abteilungen kann die Einführung eines Feature Stores erschweren. Zudem könnte ein weiterer Grund das Fehlen einer übergeordneten Organisation bestehender und neuer Use Cases sein. Ein Feature Store bringt am meisten Vorteile, wenn Features in neuen Lösungen wiederverwendet werden können. Aber das setzt eine Struktur voraus, die solche Potenziale aufdeckt.

5. Skalierung und Wartung:
Ein Feature Store muss in der Lage sein, große Mengen an Features zu verwalten und effizient abzurufen. Skalierbarkeit und Wartbarkeit können Herausforderungen darstellen, insbesondere wenn Unternehmen mit einer Vielzahl von Machine-Learning-Modellen arbeiten.

Obwohl Feature Stores noch nicht weit verbreitet sind, gewinnen sie zunehmend an Bedeutung, weil Unternehmen die Vorteile der zentralen Verwaltung und Wiederverwendung von Features erkennen. Es ist zu erwarten, dass sich die Verbreitung von Feature Stores in Zukunft weiter erhöhen wird.

3.4 Welche Methoden und Tools haben sich bei den Unternehmen etabliert?

Cloud- und On-Premise-Umgebungen

Grundsätzlich lässt sich sagen, dass sich im Hinblick auf die Nutzung von Cloud- und On-Premise-Lösungen ein gemischtes Bild ergibt (s. Abb. 26). Ein Viertel der befragten Unternehmen (24 Prozent) verfolgt eine Cloud-First-Strategie. Dies bedeutet, dass das Unternehmen so viele Prozesse und Services über einen Cloud-Provider durchführt wie möglich. Der größte Anteil (41 Prozent) jedoch verfolgt eine hybride Strategie. Hierbei wird bewusst ein Teil der Prozesse On-Premise (d. h. auf eigener Infrastruktur) gehostet und der Rest ausgelagert. Insbesondere die Datenhaltung wird auf eigenen Servern betrieben, während für die einfache Verfügbarkeit von performanten Rechnern auf Cloud-Anbieter

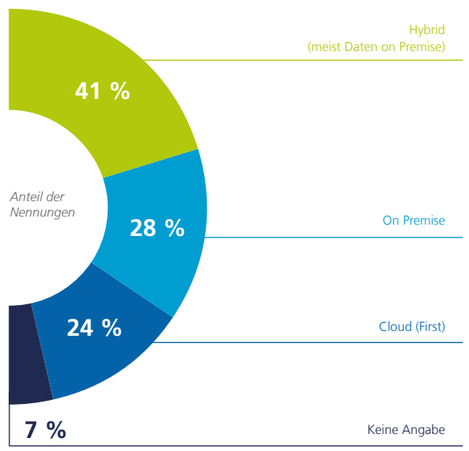


Abb. 26: Cloud-Strategien der Unternehmen

Tooling. Neben den Phasen und den Best Practices haben sich auch die aus der DevOps-Welt bekannten Tools bei den Unternehmen durchgesetzt. Bei der Versionisierung von Code und Artefakten wird in den häufigsten Fällen auf Git zurückgegriffen. Interessanterweise dominiert dieses auch bei der ML-spezifischen Modellverwaltung und liegt hinsichtlich der Nutzung weit vor anderen Open-Source-Tools, wie z. B. MLflow. Bei CI/CD wird auf Tools wie die Azure Toolsuite, GitLab und GitHub zurückgegriffen. Und im Betrieb dominieren neben der Cloud auch Docker und Kubernetes. Beim Monitoring hat sich noch kein Tooling durchgesetzt und es wird hauptsächlich auf den Faktor Mensch zurückgegriffen. Dies bedeutet, dass Fehler hauptsächlich von Anwender*innen und Entwickler*innen erkannt und gemeldet werden.

Data-Science-spezifische Aufgaben

Speziell auf die Bedarfe von Data-Scientist*innen zugeschnittene Tools haben sich noch nicht über alle Phasen bei den Unternehmen durchsetzen können. In der Exploration wird Jupyter eingesetzt, danach ergibt sich ein breites Bild an Tools, welche von ETL-Tools (Extract-Transform-Load-Tools zum Datenzugriff), über IDEs bis hin zu Eigenentwicklungen reichen. Als ein spezifisch für die Data-Science-Welt entwickeltes Tool hat sich insbesondere MLflow für das Experimenttracking durchgesetzt.

zurückgegriffen wird. 28 Prozent der Unternehmen betreiben sowohl die Entwicklung als auch das Hosting der Anwendung vollständig auf ihrer eigenen Infrastruktur (s. Abb. 26).

Für die Nutzung von Cloud-Lösungen wurden meist niedrige Eintrittshürden sowie eine flexible Infrastruktur und flexible Kosten als Vorteile genannt. Als Gründe für die Nutzung von On-Premise-Lösungen hingegen nannten die Unternehmen vor allem Bedenken hinsichtlich des Datenschutzes, regulatorische Einschränkungen sowie hohe Kosten bei einer aktiven Nutzung der Cloud-Dienste – insbesondere beim Einsatz von Deep Learning auf GPU-basierten Cloud-Ressourcen.

Etablierte Standards in der Entwicklung

Die Professionalisierung der Entwicklung und des Betriebs von Anwendungen mit ML-Komponenten zeigt sich bei den befragten Unternehmen auch im

Ausblick

Trotz der Menge neuer Anbieter, die spezifische Plattformen und Umgebungen für die Entwicklung von ML-Anwendungen bereithalten, konnten wir nicht feststellen, dass sich diese bereits umfassend bei den befragten Unternehmen etabliert haben. Keines von ihnen nutzt Ende-zu-Ende-MLOps-Lösungen, welche die notwendigen Tools und die benötigte Infrastruktur innerhalb einer Software anbieten. Durch das Zusammenrücken von Data Science und Softwareentwicklung haben sich im Bereich MLOps bekannte und etablierte Tools aus der Softwareentwicklung durchgesetzt und dominieren auch bei ML-spezifischen Aufgaben, wie z. B. der Modellverwaltung. Einzige Ausnahme bildet der Einsatz von Cloud-Tools, insbesondere Azure. Unternehmen, die bereits auf die Cloud setzen, verwenden die angebotenen Tools auch für die Entwicklung von ML-Anwendungen und deren Betrieb.

4 Welchen Einfluss hatte der Hype um Generative KI auf die Studie?

Die ersten Interviews wurden zwischen Sommer und Herbst 2022 durchgeführt. Zu diesem Zeitpunkt hatte der Hype um Generative Sprachmodelle noch nicht begonnen. Die Veröffentlichung von ChatGPT, ein Sprachmodell entwickelt von OpenAI, fand am 30. November 2022 statt. Seitdem ist dieses Thema in den Medien sehr präsent. Dies hat unter anderem auch die Studie beeinflusst. Der Großteil der Interviews wurde im ersten Halbjahr 2023 durchgeführt, und währenddessen wurde das Thema Generative KI von den Interviewpartner*innen immer wieder angesprochen. Die Geschwindigkeit, mit der die Technologie Bekanntheit erlangte bzw. adaptiert wurde, ist bemerkenswert. ChatGPT hat in kurzer Zeit beeindruckende Nutzungszahlen erreicht und wurde zu einem Thema, mit dem sich IT- und Data-Science-Abteilungen beschäftigen müssen.

Auffällig war hierbei die differenzierte Betrachtung von ChatGPT seitens der Data-Science-Teams. Mehrere Interviewpartner*innen berichteten einerseits von hoher (positiver) Aufmerksamkeit für das Thema Künstliche Intelligenz, andererseits aber auch von überhöhten Erwartungen im Hinblick auf den Einsatz Generativer KI-Lösungen. So waren die Data-Science-Teams im Interviewzeitraum noch stark damit beschäftigt, »Aufklärungsarbeit« in den Unternehmen zu leisten und eine Vielzahl von Anwendungsbereichen zu validieren. Gleichwohl waren sich die Teilnehmenden einig, dass die großen Sprachmodelle einen enormen Einfluss im Bereich Data Science haben werden.

Die Risiken beim Einsatz von ChatGPT sind dabei ein Thema. Datenschutzprobleme, die durch die Interaktion mit ChatGPT entstehen, sind zwar präsent, können aber nicht von der Verwendung abschrecken. Unternehmen müssen sich darüber bewusst sein, dass ihre Daten für das weitere Training der Modelle verwendet werden und damit potenziell offenliegen. Aber nicht nur der Datenschutz ist ein Risiko. Generative Modelle, wie ChatGPT, sind bekannt dafür, zu halluzinieren, also Fakten zu erfinden, und nicht immer ist das für die Nutzer*innen transparent. Aktuelle Studien[6] zeigen, dass Mitarbeitende, die Technologien, wie ChatGPT, bei ihrer täglichen Arbeit verwenden, dazu tendieren, diesen Glauben zu schenken und auch offensichtlich falsche Aussagen akzeptieren. Gleichzeitig sind die Vorteile groß. Die gleiche Studie konnte nachweisen, dass Mitarbeitende große Performancesteigerungen erreichen konnten. Die schnelle Akzeptanz und die Hinnahme der existierenden Risiken deuten darauf hin, dass Generative KI einen großen Einfluss auf die Implementierung von KI-Strategien in Unternehmen haben wird.

Es gibt Rückmeldungen, die darauf hinweisen, dass beispielsweise das Prototyping mit Generativer KI, Large Language Models, ChatGPT und ähnlichen Technologien oft zeitsparender ist. Jedoch haben erste Erkenntnisse gezeigt, dass die Integration in die Unternehmensprozesse aufgrund von Datenschutzfragen, Qualitätskontrolle und anderen Aspekten länger dauern kann.

Im Verlauf des Jahres 2023 ist eine Vielzahl von Angeboten rund um das Thema »Generative KI« entstanden, die die hohe Nachfrage nach verlässlichen Informationsformaten adressieren. Unter anderem bietet das Fraunhofer IAIS das Format »Innovation Briefing« an, das explizit auf die Möglichkeiten rund um den Einsatz Generativer KI-Modelle eingeht, sowie zahlreiche Schulungen im Themenfeld »Generative KI«, »Prompting« und »Foundation Models«.

6 Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K. R. (2023).

5 Ergebnisse, Herausforderungen und Empfehlungen für eine erfolgreiche Umsetzung von MLOps

Zusammenfassende Ergebnisse aus der Studie zum Einsatz von MLOps

Viele der befragten Unternehmen haben sich bereits mit MLOps beschäftigt, jedoch befinden sie sich auf unterschiedlichen Stufen oder in verschiedenen Phasen des Umsetzungsprozesses. Einige von ihnen haben bereits Organisationsstrukturen und eine robuste Infrastruktur für das Management von Machine-Learning-Lösungen aufgebaut und verfügen über automatisierte Pipelines zur Bereitstellung und Aktualisierung von ML-Lösungen. Andere Unternehmen haben gerade erst begonnen, sich mit MLOps zu beschäftigen und befinden sich noch in der Planungs- oder Experimentierphase.

Der Stand der Unternehmen in Bezug auf MLOps hängt von verschiedenen Faktoren ab, wie z. B. der Reife des Unternehmens in Bezug auf Daten- und Analytics-Kapazitäten, der Verfügbarkeit von Ressourcen und dem Grad der Veränderungsbereitschaft innerhalb der Organisation. Es ist wichtig zu beachten, dass MLOps ein kontinuierlicher Prozess ist und Unternehmen mit ihrem individuellen Entwicklungs- und Reifegrad arbeiten müssen, um das volle Potenzial von MLOps auszuschöpfen.

Die grundsätzliche Methodik von MLOps, d. h. die Umsetzung des Prozessmodells mit seinen verschiedenen Phasen, wird häufig angewendet. Allerdings gibt es organisatorische Unterschiede in der Umsetzung. In einigen Unternehmen begleiten Data-Scientist*innen den gesamten MLOps-Zyklus – von der Modellentwicklung, über das Training bis hin zur Bereitstellung und Überwachung im produktiven Einsatz. In anderen Unternehmen gibt es eine strikte Trennung hinsichtlich der Verantwortung der IT-Abteilungen: entweder ab der Development-Phase, in der die Erkenntnisse der Explorationsphase durch andere Teams neu (nach-) gebaut werden, oder ab der Deployment-Phase, ab der die IT für die Überführung und den Betrieb des Modells in der Produktivumgebung verantwortlich ist. Darüber hinaus gibt es eine breite Palette von Ansätzen in Bezug auf das Pooling von Ressourcen. Viele Unternehmen nutzen Cloud-Lösungen, um Skalierbarkeit und Flexibilität zu gewährleisten, während andere intern gehostete Infrastrukturen bevorzugen. Diese unterschiedlichen organisatorischen Ansätze spiegeln die Vielfalt der Unternehmenskulturen und Prioritäten wider, zeigen jedoch auch eine Tendenz zur Nutzung von Cloud-Lösungen im MLOps-Bereich.

Darüber hinaus sind die Werkzeuge in diesem Bereich noch sehr vielfältig. Es existieren zahlreiche Tools und Technologien, die verwendet werden, um den MLOps-Zyklus zu unterstützen. Beim Funktionsumfang der verschiedenen Tools bestehen untereinander jedoch große Schnittmengen. Dabei besteht eine Tendenz hin zur Nutzung von Cloud-Lösungen, da diese neben der bereits erwähnten Skalierbarkeit und Flexibilität auch eine vereinfachte Infrastruktur bieten und über das häufig vereinheitlichte User Interface das Gefühl einer übersichtlichen All-in-one-Lösung aufkommt. Die Wahl der richtigen Werkzeuge hängt von den individuellen Anforderungen und Präferenzen eines Unternehmens ab.

Herausforderungen und Potenziale

Selbst Unternehmen, die bereits viele Phasen des MLOps-Zyklus erfolgreich umgesetzt haben, sehen sich weiterhin vor Herausforderungen gestellt. Eine Vielzahl dieser Herausforderungen konnten wir direkt aus den Interviews herausarbeiten. Die aktuell Größte ist das Datenmanagement, einschließlich Datenverfügbarkeit, -qualität und -schutz sowie der Entwicklung einer übergreifenden Datenstrategie. Viele Unternehmen planen hierzu Maßnahmen zur Verbesserung. Darüber hinaus nennen die Befragten auch interne Prozessherausforderungen, wie den Aufbau von Strukturen, die Fachseitenintegration und eine skalierbare technische Infrastruktur. Weiter wurden auch mangelndes Vertrauen, fehlendes Verständnis und äußere Einflüsse, wie die volatile KI-Landschaft, Regulierungen und der schwierige Personalmarkt, als Herausforderungen aufgeführt. Viele sehen sich zudem mit der fachlichen Skalierung konfrontiert, was einerseits den Aufbau von Know-how im Unternehmen, aber andererseits auch die Zunahme von ML-Use-Cases bedeutet. Zuletzt sehen sich Unternehmen auch in Forschungsthemen vor Herausforderungen gestellt, wie nachhaltige KI-Entwicklung, Erklärbarkeit und langfristige Ansätze (Digitalisierungs-Roadmaps und »AI-First«). Aber auch das Thema LLMs wurde gegen Ende der Interviewphase vermehrt genannt.

Zusätzlich zu den genannten und herausgearbeiteten Herausforderungen sehen wir weitere Potenziale, die von Unternehmen ausgeschöpft werden können. Eines ist das Experimenttracking, bei dem es darum geht, alle Experimente und Iterationen während des Modellentwicklungsprozesses zu verfolgen und zu dokumentieren. Viele Unternehmen haben dies noch nicht umgesetzt. Dies würde jedoch helfen, den Überblick über den Fortschritt bei der Exploration zu behalten und die Reproduzierbarkeit zu gewährleisten.

Ein weiteres wichtiges Thema ist das Monitoring und die Detektion von Drifts im produktiven Einsatz, also Veränderungen in den Eingabedaten oder der Leistung des Modells. Es ist entscheidend, dass Unternehmen ihre Modelle kontinuierlich überwachen, sowohl fachliche als auch IT-Metriken, um sicherzustellen, dass sie weiterhin präzise und zuverlässig arbeiten. Die Erkennung von Drifts ermöglicht es, schnell darauf zu reagieren und das Modell anzupassen. Auch hier gibt es für die Unternehmen bei der Umsetzung noch viel Potenzial.

Diese Herausforderungen und Potenziale zeigen, dass MLOps ein fortlaufender Prozess ist, der kontinuierliche Anpassungen und Verbesserungen erfordert, um erfolgreich zu sein.

Empfehlungen für den Einstieg in und die Professionalisierung von MLOps

Dass sich der Einsatz von MLOps für Unternehmen durchaus lohnen kann, verdeutlicht diese Studie. Wir empfehlen Unternehmer*innen daher, sich intensiv mit MLOps auseinanderzusetzen und das Thema schrittweise weiterzuentwickeln. Dabei sind die Aspekte Datenverwaltung und Datenmanagement mit hoher Priorität zu behandeln und insbesondere die Fachbereiche in die Verantwortung zu nehmen. Das Gleiche gilt auch für die Auswahl und die Bewertung der Werthaltigkeit von ML-Use-Castes.

Die wichtigsten Empfehlungen finden Sie nachfolgend im Überblick:

Bereiten Sie Ihr Unternehmen auf Veränderungen vor, die mit der Einführung von MLOps einhergehen, und managen Sie den Wandel aktiv.

Stellen Sie sicher, dass Experimente und Modelle reproduzierbar sind, um die Nachvollziehbarkeit und Qualitätssicherung lückenlos zu gewährleisten.

Automatisieren Sie den ML-Lebenszyklus, wo immer möglich, um manuelle Fehler zu reduzieren und die Effizienz zu steigern.

Fördern Sie eine Kultur der Zusammenarbeit zwischen Data-Scientist*innen, Entwickler*innen und IT-Operations, um Silos zu vermeiden.

Nutzen Sie Continuous-Integration- und Continuous-Deployment-Praktiken, um die Bereitstellung von ML-Modellen zu beschleunigen und zu standardisieren.

Stellen Sie sicher, dass Ihre Machine-Learning-Modelle ethische Standards einhalten und vertrauenswürdig sind.

Bilden Sie Ihr Team in den Grundlagen von Machine Learning und MLOps weiter, um ein gemeinsames Verständnis zu schaffen.

Wählen Sie die richtigen Tools und Plattformen, die zu den Anforderungen Ihres Unternehmens passen und die MLOps-Prozesse unterstützen.

Prüfen Sie sorgfältig den möglichen Einsatz von neuen Technologien wie LLMs und Generativer KI, die zusätzliche Herausforderungen an MLOps stellen.

Widmen Sie den Themen Datenmanagement und Datenintegration genügend Aufmerksamkeit und Zeit, da diese unerlässlich für die Qualität, Robustheit und Zuverlässigkeit der später trainierten Modelle sind und insbesondere für den erfolgreichen produktiven Einsatz.

Investieren Sie in eine skalierbare Infrastruktur, die das Wachstum von ML-Anwendungen unterstützen kann.

Etablieren Sie technische und fachliche Feedbackschleifen, um Modelle kontinuierlich zu verbessern, auf sich ändernde Daten reagieren zu können und sicherzustellen, dass die Modelle den Geschäftsanforderungen entsprechend Wert liefern.

Implementieren Sie ein umfassendes Monitoring und Logging, um die Leistung der Modelle und die Systemgesundheit zu überwachen.

Ausblick

MLOps als Paradigma muss sich beständig weiterentwickeln. Mit der zunehmenden Regulierung, insbesondere des Betriebs und Einsatzes von Maschine-Learning-Anwendungen werden Fragestellungen in Bezug auf die Vertrauenswürdigkeit von ML-Lösungen zunehmend dringender, und Entwickler*innenteams müssen sich bei der Entwicklung neuer Lösungen und der Wartung bereits im Betrieb befindlicher Modelle damit beschäftigen, wie diese sichergestellt werden kann. Im Unterschied zum nachgelagerten Auditing könnten angepasste MLOps-Praktiken dabei unterstützen, die notwendigen Tätigkeiten und Dokumentationspflichten bereits während der Entwicklung und möglichst automatisiert zu berücksichtigen.

Mit der zunehmenden Verbreitung und dem Einsatz von Generativer KI für ML-Lösungen stellt sich die Frage, wie diese betrieben und entwickelt werden können und ob MLOps diese Prozesse ausreichend berücksichtigt. Es wird bereits von angepassten LLMOps (Large Language Model Operations) gesprochen, wobei noch nicht eindeutig ist, ob es sich hierbei um die Entstehung eines neuen Paradigmas handelt oder lediglich um eine spezialisierte Form der bisherigen MLOps. Das Fraunhofer IAIS wird sich fortlaufend mit der Entwicklung und dem Betrieb dieser Technologien auseinandersetzen und prüfen, inwieweit die bisherigen Best Practices im Bereich MLOps angepasst werden sollten.

6 Publikationsempfehlungen und Schulungsangebote

Innovation Briefing Generative KI
Mit diesem kompakten Briefingformat für Führungskräfte und alle, die unternehmensrelevantes Überblickswissen erhalten wollen, verschaffen sich die Teilnehmenden einen Überblick über das Potenzial der großen KI-Sprachmodelle. Die Veranstaltung findet wahlweise online oder in Präsenz statt.

Kompakteinstieg Machine Learning Operations (MLOps)
Durch die interaktive Onlineschulung erfahren Teilnehmende die wichtigsten Grundlagen und erhalten einen kompakten Überblick über die Herausforderungen und Lösungsansätze zum produktiven Einsatz von ML-Anwendungen in Unternehmen.

Kompakteinstieg Prompting für Generative KI
Diese Schulung richtet sich an Teilnehmende aus Management und sämtlichen Fachbereichen, die einen kompakten Einstieg in das Thema GPT, Prompting und generell große Sprachmodelle erhalten möchten.

Whitepaper Zukunftssichere Lösungen für Maschinelles Lernen
In der Publikation zu den »Machine Learning Operations (MLOps) – Prozesse für Entwicklung, Integration und Betrieb« geben wir Orientierungshilfen, wie der Übergang von ML-Lösungen aus Forschung und Entwicklung in das strukturierte Tagesgeschäft eines Unternehmens gelingen kann.

7 Quellenverzeichnis

Beck, N., Martens, C., Sylla, K.-H., Wegener, D. und Zimmermann, A. (2020): Zukunftssichere Lösungen für Maschinelles Lernen, Sankt Augustin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS.

Bitkom Research (Hrsg.) (2023): Deutsche Wirtschaft drückt bei Künstlicher Intelligenz aufs Tempo, verfügbar unter: *https://www.bitkom.org/Presse/Presseinformation/Deutsche-Wirtschaft-drueckt-bei-Kuenstlicher-Intelligenz-aufs-Tempo#_* [zuletzt aufgerufen: 19.12.2023].

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K. R. (2023): Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality, Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, verfügbar unter: *https://ssrn.com/abstract=4573321* or *http://dx.doi.org/10.2139/ssrn.4573321* [zuletzt aufgerufen: 19.12.2023].

Google (Hrsg.) (2023): MLOps: Continuous Delivery und Pipelines zur Automatisierung im maschinellen Lernen, verfügbar unter: *https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=de* [zuletzt aufgerufen: 19.12.2023].

Microsoft (Hrsg.) (2023): Machine Learning Operations-Framework (MLOps) zum Hochskalieren des Machine Learning-Lebenszyklus mit Azure Machine Learning, verfügbar unter: *https://learn.microsoft.com/de-de/azure/architecture/ai-ml/guide/mlops-technical-paper* [zuletzt aufgerufen: 19.12.2023].

8 Impressum

Herausgeber
Kompetenzplattform KI.NRW
Geschäftsführung Dr. Christian Temath
c/o Fraunhofer-Institut für Intelligente
Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.ki.nrw | www.iais.fraunhofer.de

Kontakt
Andreas Kerbel
KI-Manager KI.NRW
Telefon 02241 14-2980
andreas.kerbel@iais.fraunhofer.de

Redaktion und Lektorat
Claudia Könsgen, KI.NRW
Mirco Lange, KI.NRW

Grafik und Layout
Sina Bolder, Fraunhofer IAIS
Jessica Schmitz, KI.NRW

Bildquellen
© monsitj – stock.adobe.com / KI NRW – Cover
© Lucky Ai – stock.adobe.com – Seite 4-5
© peopleimages.com – stock.adobe.com – Seite 6
© Daniel – stock.adobe.com – Seite 13
© 18042011 – stock.adobe.com – Seite 17
© killykoon – stock.adobe.com – Seite 18
© visoot – stock.adobe.com – Seite 21
© Gorodenkoff – stock.adobe.com – Seite 22-23
© bornmedia – stock.adobe.com – Seite 29
© Yi_Studio – stock.adobe.com – Seite 34
© Planetz – stock.adobe.com – Seite 36-37
© sandsun – stock.adobe.com – Seite 39
© chayantorn – stock.adobe.com – Seite 42-43

Stand
April 2024, 1. Auflage

Sankt Augustin 2024

www.ki.nrw