Leitfaden

bitkom

Machine Learning Operations

Ein Guide für MLOps-Einsteiger

Herausgeber
Bitkom e. V.
Albrechtstraße 10
10117 Berlin
T 030 27576-0
bitkom@bitkom.org
www.bitkom.org

Ansprechpartner
Kai Beerlink | Referent Künstliche Intelligenz
T 030 27576-278 | k.beerlink@bitkom.org

Verantwortliches Bitkom-Gremium
AK Artificial Intelligence

Autoren
Alexander Albrecht, Benjamin Feldmann | Bakdata
Dominik Kreuzberger | IBM
Mathis Börner | SAP
Morten Hesebeck-Brinckmann | Senacor
Burkhard Hilchenbach | Software AG

Inhalt

1 Einführung

Was dieser Guide bietet

1.1 Relevanz & Zielstellung

Maschinelles Lernen (auch »Machine Learning«, ML) hat in den letzten Jahren in Wirtschaft und Gesellschaft zunehmend an Bedeutung gewonnen. Seien es Finanzen, Gesundheit, Einzelhandel oder Produktion – ML bzw. spezifische ML-Modelle werden in vielen Bereichen eingesetzt, um Prozesse zu automatisieren, Empfehlungen zu geben und Entscheidungen zu treffen. In Zukunft wird die Bedeutung von ML weiter zunehmen, da Unternehmen und Regierungen immer mehr Daten sammeln, mit denen sie arbeiten können. Es wird dazu beitragen, die Produktivität und Effizienz von Unternehmen zu steigern und neue Geschäftsmodelle zu entwickeln. Auch in immer mehr Bereichen des täglichen Lebens wird es Einzug halten. So werden wir in Zukunft in der Bildung, im Verkehr oder der öffentlichen Verwaltung mit ML, eingebettet in Produkten und Prozessen, in Berührung kommen bzw. tun das heute schon. Ein wichtiger Bestandteil der Anwendung von ML in der Praxis ist dabei das Management des Lebenszyklus von ML-Modellen, auch bekannt als MLOps.

79 %

der Bevölkerung sind überzeugt, dass KI die Wettbewerbsfähigkeit der deutschen Wirtschaft stärken wird. Quelle: ↗ Bitkom-Studie 2023

MLOps umfasst die Entwicklung, Validierung, Bereitstellung und Überwachung von ML-Modellen.

Durch die Einführung von MLOps-Methoden und -Tools können Unternehmen sicherstellen, dass ihre ML-Modelle sowohl zuverlässig als auch skalierbar sind und die erwarteten Ergebnisse liefern. Es ermöglicht die schnelle Weiterentwicklung und Verbesserung von Modellen und kann bei der Einhaltung von Compliance-Anforderungen unterstützen. Es dient der Zusammenarbeit zwischen Data Scientists, Softwareentwicklerinnen und -entwicklern und Development and Operations (DevOps) Teams, und hilft dabei, dass die Modelle in Produktionsumgebungen reibungslos funktionieren und schnell als auch sicher bereitgestellt werden können.

Mit diesem Leitfaden möchten wir MLOps-Neulingen erklären, in welchen Situationen sich der Einsatz von MLOps lohnen kann. Es wird genau dargestellt, was MLOps eigentlich ist und aufgezeigt, womit man sich tiefergehend auseinandersetzen sollte, wenn man die Einführung in Betracht zieht. Zuletzt werden außerdem Erfolgsfaktoren für die bestmögliche Implementierung von MLOps betrachtet.

1.2 Diagnose – brauchen wir MLOps?

MLOps ist ein junges Gebiet mit einem schnellen Entwicklungstempo und wenig allgemein etablierten Techniken und Best Practices. Der Einstieg kann daher schwierig sein. Umso wichtiger ist es, im ersten Schritt zu evaluieren, ob MLOps überhaupt zur Lösung meines Problems beitragen kann. Um dies zu erleichtern, möchten wir fünf typische Herausforderungen vorstellen, die klare Anzeichen dafür sind, dass es in einer Organisation an MLOps-Methoden, -Tools und -Prozessen mangelt. Eine solche Analyse kann dabei helfen, spezifische Problempunkte in der eigenen Organisation zu identifizieren und realistische Erwartungen bezüglich des Beitrags von MLOps zur Lösung dieser zu formulieren. So können bei der Einführung die richtigen Schwerpunkte gesetzt und nach der Einführung die erzielten Erfolge bewertet werden.

> Wir schaffen es nur schwer, die Entwicklungen unserer Data Scientists in Produkte zu verwandeln, obwohl die Performance der Modelle gut ist und Anwendungsfälle vorhanden sind.

Der Anwendungsfall ist gut verstanden, die Performance des Modells zufriedenstellend – nichtdestotrotz schlägt die Überführung der Entwicklung ins Produkt fehl. Wenn ML-Projekte häufig nach dem Proof-of-Concept scheitern, stellt sich für Unternehmen die Frage, welche Barrieren beseitigt werden müssen, um ML-Modelle erfolgreich in eigenen Produkten nutzen zu können.

Unterschiedliche Gründe kommen für dieses Dilemma in Frage. Data Scientists optimieren Modelle und deren Performance, besitzen aber nicht unbedingt das Know-how, diese auch produktiv zu betreiben. Um ML in einem Produkt erfolgreich zu implementieren, wird zusätzlich Expertise in den Bereichen Software- und Data-Engineering benötigt. Denn das Ziel bei ML-Projekten ist immer die Entwicklung und der Betrieb einer Software bei der – verglichen mit klassischer Softwareentwicklung – Modelle und Daten eine zusätzliche Komplexität darstellen. Fehlen aber praxiserprobte Best-Practices aus der Softwareentwicklung, insbesondere DevOps-Praktiken, die auf ML-Systeme übertragen werden können oder die Anbindung der Datensysteme, z. B. Data Lakes, die ein regelmäßiges Neu- und Nachtrainieren von Modellen ermöglichen, kann dies die produktive Implementierung von ML erschweren. MLOps kann dazu beitragen, diese notwendigen Faktoren besser zusammenbringen und einen ganzheitlichen Blick auf ML zu haben.

> Neue Modelle oder Modellupdates bringen wir zu langsam in unsere Produkte.

Bei ML ist eine kontinuierliche Anpassung von Modellen möglich und oft auch gewollt. Gibt es bei diesem Update-Prozess Verzögerungen, kostet dies nicht nur Zeit, sondern auch Geld. Die Koordination des ML-Lifecycles durch die Zusammenarbeit zwischen Data Scientists und DevOps-Engineers ist dabei zentral für einen effizienten Ablauf.

MLOps kann hier helfen, den Übergang von der Entwicklung zur Produktion effektiv zu verkürzen, z. B. durch den Einsatz von Techniken wie Continuous Integration oder Continuous Deployment. Im gleichen Zuge können oft auch ML-Plattformen ebenso zur Beschleunigung der Prozesse beitragen, da sie das Aufsetzen einer MLOps-Infrastruktur für Data Scientists deutlich vereinfachen.

> Bei der Integration von Modellen in Produkte und Prozesse müssen wir immer wieder dieselben technischen Probleme lösen.

Werden ML-Modelle in Produkte und Prozesse gebracht, wiederholen sich oft händische Arbeitsschritte und Abläufe. Dies bindet unnötig Ressourcen und ist zudem fehleranfällig.

Durch MLOps und die damit mögliche Automatisierung und Standardisierung kann dies effizienter gestaltet werden. Auch wird ML erst durch Automatisierung gut wiederhol- und reproduzierbar. Ebenso wird die Wiederverwendbarkeit von Code dadurch verbessert.

Wird zudem die Infrastruktur automatisiert über Code bereitgestellt, können ML-Modelle effizient in unterschiedlichen Umgebungen in der Cloud oder On-Premises entwickelt und betrieben werden. Automatisierte Prozesse sind somit der Schlüssel für skalierbares ML und können mithilfe von MLOps realisiert werden.

> Viele Entwicklungen werden gestoppt, weil die Betriebskosten als zu hoch und die Compliance als zu unsicher eingeschätzt werden.

Damit Maschinelles Lernen Compliance-konform eingesetzt werden kann, müssen das Verhalten und die Performance von Modellen reproduzierbar, protokollierbar und evaluierbar sein. Fehlen dafür die Methoden, ist der vorschriftsmäßige Einsatz von Modellen nur schwer überprüfbar. Auch lassen sich die Betriebskosten dann nur unzureichend abschätzen.

MLOps ermöglicht kontinuierliches Logging, Monitoring und Auditing und lässt so rechtzeitig erkennen, ob Modelle an Performance verlieren, bevor das zum Problem wird. Mit diesen automatisierten Evaluationsverfahren lässt sich zudem der Erfolg von Modellen besser messen und auch besser ins Verhältnis zu anfallenden Betriebskosten setzen.

> Wir haben große Probleme neuen Kunden Zugang zu unseren ML-Produkten zu geben, da wir Probleme mit der Skalierung der Rechnerressourcen und der Trennung der Daten von verschiedenen Kunden haben.

Für ML müssen auf Abruf enorme Mengen an Rechnerressourcen mit spezialisierter Hardware zur Verfügung gestellt werden, um Modelle auf großen Datenmengen trainieren zu können. Ist dies nicht möglich, kann das zur Herausforderung werden, wenn man neue Kunden im B2B-Geschäft gewinnen möchte.

Cloud-Anbieter bieten dafür skalierbare ML-Lösungen an. Es gibt inzwischen auch Technologien, die vergleichbare Lösungen in der Private Cloud ermöglichen. Mittels moderner Infrastructure-as-Code-Technologien lässt sich damit auch die Trennung der Daten von verschiedenen Kunden automatisiert erreichen. MLOps hilft dabei, diese Aspekte zu berücksichtigen.

2 MLOps

2.1 DevOps als Grundlage für MLOps

DevOps entstand aus der Erkenntnis, dass die Vorteile des agilen Software-Engineerings nicht erreicht werden können, wenn die nachfolgenden Prozesse und Strukturen zwischen Entwicklung (Dev) und Betrieb (Ops) langsam und schwerfällig sind.

Traditionell agieren Entwicklung und Betrieb separat; als Hauptschnittstelle gibt es die Übergabe produktionsreifer Software von der Entwicklung an den Betrieb. DevOps ersetzt diese Trennung durch einen singulären Prozess, der Entwicklungs- und Betriebsaufgaben umfasst und es ermöglicht, die Vorteile agiler Bereitstellung von Entwicklungsergebnissen durch gleichermaßen agiles Ausrollen und Betreiben auch auszuschöpfen. Dieser DevOps-Ablauf wird auch oft als Endlosschleife visualisiert.

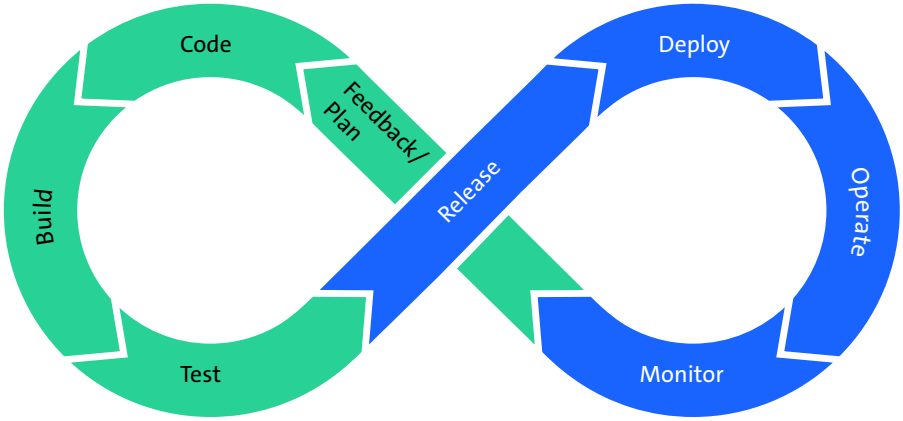

Abbildung 1: DEVOPS-Endlosschleife (eigene Darstellung)

DevOps ist mehr als eine reine Methodik und befasst sich mit sozialen und technischen Problemen in Organisationen, die Software entwickeln. Mit dem Ziel die Lücke zwischen Entwicklung und Betrieb zu beseitigen, legt es den Schwerpunkt auf Zusammenarbeit, Kommunikation und Wissensaustausch. Dabei überschneiden sich die Rollen in Organisationen immer mehr. Entwicklungsteams haben Verantwortung in Bereichen, die früher der Qualitätssicherung und dem Betrieb vorbehalten waren.

DevOps fördert zudem die Automatisierung durch die Taktik der kontinuierlichen Integration, der kontinuierlichen Bereitstellung und des kontinuierlichen Einsatzes (CI / CD) und ermöglicht so schnelle, häufige und zuverlässige Veröffentlichungen. Darüber hinaus soll es kontinuierliche Tests, Qualitätssicherung, kontinuierliche Überwachung, Protokollierung und Feedbackschleifen gewährleisten.

DevOps legt großen Wert auf die stetige Verbesserung von Software. Bei MLOps geht es analog um die kontinuierliche Verbesserung von ML-Modellen. Es gibt viele DevOps- und MLOps-Tools mit den entsprechenden Zielen. Allerdings muss klar sein, dass die Konzepte vor allem auch eine bestimmte Betrachtung von Prozessen und Gestaltung von Arbeitsweisen in Organisationen bedeuten.

Sowohl in der Industrie als auch in der Wissenschaft hat man mit DevOps eine Fülle von Erfahrungen im Software-Engineering gesammelt. Diese Erfahrungen werden nun, rund 10 Jahre nach dem Aufkommen von DevOps, zur Automatisierung und Operationalisierung von ML genutzt – und dies führt nun zu MLOps.

2.2 MLOps im Detail

MLOps übernimmt Ansätze und Prinzipien von DevOps und entwickelt zusätzlich eigene, die benötigt werden, um ML in die Produktion zu bekommen. Es umfasst Best-Practices, Konzepte und eine Entwicklungskultur für die durchgängige Konzeption, Implementierung, Überwachung, Bereitstellung und Skalierbarkeit von ML-Produkten. Vor allem handelt es sich um eine technische Praxis, die drei Disziplinen miteinander verbindet:

- Maschinelles Lernen
- Software-Engineering (insbesondere DevOps)
- und Data Engineering.

> Ein zentraler (aber nicht der einzige) Aspekt von MLOps ist die schnelle und zuverlässige Überführung von ML-Modellen aus der Entwicklung in den Betrieb.

MLOps zielt, vergleichbar zu DevOps, auf Basis dieser Kombination darauf ab, ML-Systeme in die Produktion zu überführen, indem die Lücke zwischen Entwicklung (Dev) und Betrieb (Ops) überbrückt wird. Das Objekt von DevOps ist ‚Software' im Allgemeinen. MLOps geht auf die Besonderheiten und Anforderungen von ML-Systemen ein. Es gibt unterschiedliche Ausführungen des Einsatzes solcher Systeme; normalerweise sind dabei jedoch die folgenden Bestandteile involviert:

1. **ML-Modell:** Ein in Entwicklung trainiertes ML-Modell (Training von Modellen in Produktion kommt auch vor), sowie eine Laufzeitumgebung für das Modell, einschließlich, aber nicht beschränkt auf eine Inference Engine, welche neue Aussagen aus einer gegebenen Wissensbasis ableitet.

2. **Datenhandling:** Softwarekomponenten zum Einlesen und Vorbereiten der Daten, da sich die Rohdaten fast nie in einer für das Modell verarbeitbaren Form befinden. Oft gibt es auch Komponenten zur Nachbereitung der Ergebnisse.

3. **Betriebskomponenten:** Diese Gruppe umfasst alle weiteren Komponenten, die zum Betrieb des ML-Modells erforderlich sind. Diese können ausführbare Softwarekomponenten (z. B. in Python) oder auch deklarative Beschreibungen (z. B. in YAML) sein, welche von entsprechenden Engines umgesetzt werden (ein zentrales Muster bei DevOps). Prominentestes Beispiel sind ML-Workflows, die die erforderlichen Schritte beim Betrieb des Modells orchestrieren, Monitoring konfigurieren, usw.

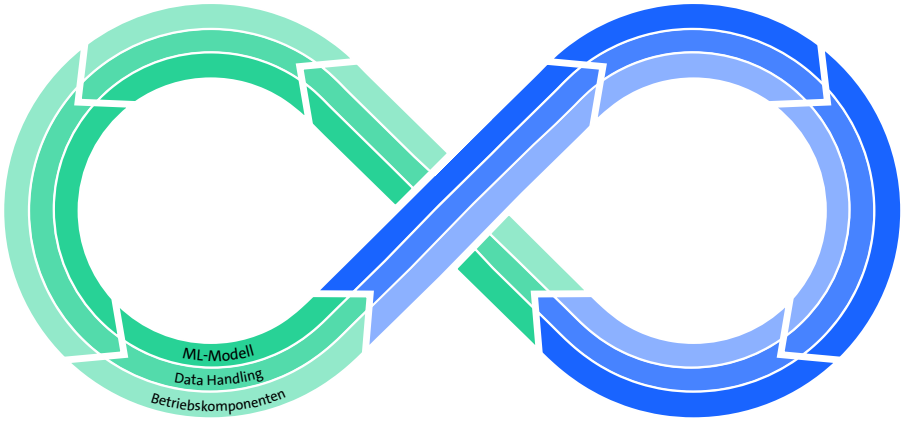

Abbildung 2: MLOps-Endlosschleife (vereinfacht) mit den wichtigsten involvierten Artefakten (eigene Darstellung)

MLOps orchestriert dabei das Zusammenspiel dieser Komponenten, um die schnelle und zuverlässige Entwicklung von ML-Produkten zu erleichtern. Dabei folgt es diesen Prinzipien, mit denen man sich bei der Einführung von MLOps intensiver auseinandersetzen muss: CI / CD-Automatisierung, Workflow-Orchestrierung, Reproduzierbarkeit; Versionierung von Daten, Modellen und Code; Zusammenarbeit; kontinuierliches ML-Training und -Bewertung; Verfolgung und Protokollierung von ML-Metadaten; kontinuierliche Überwachung und Feedback-Schleifen.

3 Erfolgsfaktoren bei der Einführung von MLOps

Probleme und Ziele definieren

Am Anfang einer erfolgreichen Einführung von MLOps steht eine vorangegangene Problemanalyse. Auf einer solchen basierend werden Ziele definiert, die anschließend durch MLOps erreicht werden sollen. Typische Zieldefinitionen sind beispielsweise die Veröffentlichung von ML-Produkten, eine Verringerung von Time to Market, die Reduzierung technischer Probleme oder auch die Klärung von Compliance-Fragen.

ML(Ops) Reifegrad einschätzen

Um den Weg von Problem zum Ziel mithilfe von MLOps erfolgreich zu bewältigen, kann es helfen, die zum Start des Prozesses gegebene ML(Ops) Reife (Maturity Level) festzustellen. Gibt es bereits hilfreiche Dokumentation? Sind Logs vorhanden, die genutzt werden können? Je nach Reifegrad bieten sich unterschiedliche Schritte für die Einführung von MLOps an. Hier ist die individuelle Ausgangsituation des Unternehmens entscheidend. Sie muss mit den definierten Zielen abgeglichen werden, um die für ihre Erreichung notwendige Schwerpunktsetzung bei der Einführung von MLOps zu gewährleisten.

Kollaboration in den Mittelpunkt rücken

Für die erfolgreiche Umsetzung von MLOPs ist die Zusammenarbeit verschiedener Rollen unabdingbar, da fachübergreifendes Know-how benötigt wird. Bei der Datenbeschaffung und Säuberung können Data Engineers helfen, Data Scientists können basierend auf diesen Daten die Modelle trainieren und die ML-Pipelines mit Hilfe von Software-Engineers optimieren. DevOps ermöglichen einen reibungslosen Betrieb in Produktion durch CI / CD sowie Monitoring der betriebenen Modelle. Die Business-Domäne fokussiert sich auf das Definieren der Geschäftsziele und die Kommunikation zu den Stakeholdern.

Eine Aufgabenteilung hinsichtlich der Spezifizierung der Bereiche ist oft notwendig, um einen Übergang von Entwicklung nach Produktion effektiv zu verkürzen, z. B. durch den Einsatz von Techniken wie Continuous Integration oder Continuous Deployment (DevOps).

Unterschiedliche Nutzeranforderungen abfragen

Ein weiterer Erfolgsfaktor für die Einführung von MLOps ist das Matching von Nutzeranforderungen auf Systeme und Tools sowie der Zugang zu benötigten Ressourcen. Es gibt keine One-Fits-All-Lösung, wenn es um Programmiersprachen, ML-Libraries, Tools und Infrastruktur-Ressourcen geht. Unterschiedliche Teammitglieder haben unterschiedliche Expertisen und Anforderungen. Ebenfalls spielen nicht nur technologische Aspekte eine Rolle, sondern auch bestehende Geschäftsbeziehungen mit Toolsets oder Providern. Infrastruktur-Anforderungen für das Betreiben der Data- und Feature-Engineering-Pipelines wiederum sind abhängig vom Datenvolumen und der Datenlatenzzeit. Es sollte daher eine gemeinsame Evaluierung der Anforderungen stattfinden, nach der die Teammitglieder die Entscheidung über benötigte Ressourcen und Tools gemeinsam treffen.

Einbettung und Schnittstellen schaffen

Für erfolgreiche MLOps sind bestehende Data Governance und DevOps-Prozesse essenziell. Sowohl DevOps als auch MLOps sind Software Development-Strategien, die sich auf die Kollaboration zwischen Entwicklerinnen und Entwicklern, Operations und Data Science fokussieren. Auch wenn sich DevOps auf Application Development bezieht, und MLOps auf ML, können bestehende DevOps-Prozesse in der Organisation vorteilhaft sein, da MLOps auf ihnen aufsetzt. Ebenfalls ist ML auf den Zugang zu Daten in ausreichender Qualität und Menge angewiesen. Einem Unternehmen kann es daher helfen, bestehende Data Governance-Prozesse haben, die Datenqualität und Zugangsrechte gewährleisten.

4 Fazit & Ausblick

Um langfristig unternehmerischen Erfolg mit Machine Learning und MLOps zu haben, bedarf es eines ganzheitlichen Blickes. MLOps ist nicht als Tool zu verstehen, das nach der Einführung ohne weiteren Aufwand läuft, sondern beinhaltet Arbeitsweisen und Prozesse. Diese müssen langfristig eingehalten werden, damit eine effektive dauerhafte Operationalisierung der ML-Modelle gewährleistet ist. Wenn dem Rechnung getragen wird, wird der Entwicklungsprozess dauerhaft beschleunigt und die Development-, Business- und Data-Domänen können sich durch wiederverwendbare Prozesse und Strukturen auf rollenspezifische Aufgaben fokussieren. Dadurch können ML-Modelle effizient in Produktion gebracht und langfristig in die IT-Infrastruktur eines Unternehmens integriert werden.

Bitkom vertritt mehr als 2.000 Mitgliedsunternehmen aus der digitalen Wirtschaft. Sie erzielen allein mit IT- und Telekommunikationsleistungen jährlich Umsätze von 190 Milliarden Euro, darunter Exporte in Höhe von 50 Milliarden Euro. Die Bitkom-Mitglieder beschäftigen in Deutschland mehr als 2 Millionen Mitarbeiterinnen und Mitarbeiter. Zu den Mitgliedern zählen mehr als 1.000 Mittelständler, über 500 Startups und nahezu alle Global Player. Sie bieten Software, IT-Services, Telekommunikations- oder Internetdienste an, stellen Geräte und Bauteile her, sind im Bereich der digitalen Medien tätig oder in anderer Weise Teil der digitalen Wirtschaft. 80 Prozent der Unternehmen haben ihren Hauptsitz in Deutschland, jeweils 8 Prozent kommen aus Europa und den USA, 4 Prozent aus anderen Regionen. Bitkom fördert und treibt die digitale Transformation der deutschen Wirtschaft und setzt sich für eine breite gesellschaftliche Teilhabe an den digitalen Entwicklungen ein. Ziel ist es, Deutschland zu einem weltweit führenden Digitalstandort zu machen.

Bitkom e.V.

Albrechtstraße 10
10117 Berlin
T 030 27576-0
bitkom@bitkom.org

bitkom.org

bitkom