O'REILLY®

Deutsche Ausgabe

MLOps

Kernkonzepte im Überblick

Machine-Learning-Prozesse im Unternehmen nachhaltig automatisieren und skalieren

Mark Treveil
und das Dataiku-Team
Übersetzung von Marcus Fraaß

MLOps – Kernkonzepte im Überblick

Machine-Learning-Prozesse im Unternehmen nachhaltig automatisieren und skalieren

Mark Treveil und das Dataiku-Team

Deutsche Übersetzung von Marcus Fraaß

O'REILLY®

Mark Treveil und das Dataiku-Team

Lektorat: Alexandra Follenius
Übersetzung: Marcus Fraaß
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.d-nb.de* abrufbar.

ISBN:
Print 978-3-96009-172-1
PDF 978-3-96010-580-0
ePub 978-3-96010-581-7
mobi 978-3-96010-582-4

1. Auflage

Wieblinger Weg 17
69123 Heidelberg

Authorized German translation of the English edition of *Introducing MLOps: How to Scale Machine Learning in the Enterprise*, ISBN 9781492083290 © 2020 Dataiku. This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

Hinweis:
Dieses Buch wurde auf PEFC-zertifiziertem Papier aus nachhaltiger Waldwirtschaft gedruckt. Der Umwelt zuliebe verzichten wir zusätzlich auf die Einschweißfolie.

PEFC zertifiziert
Das Papier für dieses Buch stammt aus nachhaltig bewirtschafteten Wäldern und kontrollierten Quellen.
PEFC™
PEFC/04-31-0810 www.pefc.de

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: *kommentar@oreilly.de*.

5 4 3 2 1 0

Inhalt

Teil III MLOps-Anwendungsfälle aus der Praxis

Vorwort

Wir haben einen Wendepunkt in der Geschichte des maschinellen Lernens erreicht, an dem die Technologie aus dem theoretischen und dem akademischen Umfeld in die »reale Welt« vorgedrungen ist – d. h. in Unternehmen, die alle möglichen Arten von Dienstleistungen und Produkten für Menschen auf der ganzen Welt anbieten. Dieser Wandel ist nicht nur äußerst spannend, sondern stellt auch eine Herausforderung dar, da er die Komplexität von Machine-Learning-Modellen mit der Komplexität moderner Unternehmen zusammenbringt.

Eine Schwierigkeit, wenn Unternehmen von der experimentellen Nutzung des Machine Learning zur Skalierung in Produktivumgebungen übergehen, stellt die Wartung dar. Wie können Unternehmen von der Verwaltung eines einzigen Modells zum Managen von Dutzenden, Hunderten oder sogar Tausenden übergehen? Hier kommt nicht nur MLOps ins Spiel, sondern auch die bereits erwähnten Komplexitäten, und zwar sowohl auf der technischen als auch auf der geschäftlichen Ebene. Dieses Buch führt Leserinnen und Leser in diese Herausforderungen ein und bietet gleichzeitig praktische Einblicke und Lösungen für die Entwicklung von Fähigkeiten im Bereich MLOps.

An wen sich dieses Buch richtet

Wir haben dieses Buch speziell für Managerinnen und Manager von Analytics- und IT-Operations-Teams geschrieben, also für die Personen, die direkt mit der Aufgabe betraut sind, Machine Learning (ML) in einer Produktivumgebung zu skalieren. Da MLOps ein neues Feld ist, haben wir dieses Buch als Leitfaden für das Erstellen einer erfolgreichen MLOps-Umgebung konzipiert, beginnend bei den organisatorischen bis hin zu den technischen Herausforderungen.

Aufbau des Buchs

Dieses Buch ist in drei Hauptteile gegliedert. Teil I, *Was ist MLOps, und warum wird es benötigt?*, stellt das Thema MLOps grundlegend vor und geht darauf ein, wie (und warum) es sich als Disziplin entwickelt hat, wer beteiligt sein sollte, um MLOps erfolgreich durchzuführen, und welche Bausteine erforderlich sind.

Teil II, *MLOps einsetzen*, orientiert sich im Wesentlichen an dem Lebenszyklus von Machine-Learning-Modellen und umfasst mehrere Kapitel, die sich mit der Entwicklung von Modellen, der Produktionsvorbereitung, dem Deployment in die Produktivumgebung, dem Monitoring und der Governance befassen. Diese Kapitel behandeln nicht nur allgemeine Aspekte, sondern auch solche, die sich speziell auf den Einsatz von MLOps in jeder Phase des Lebenszyklus beziehen, wodurch die in Kapitel 3, *Die Kernkomponenten von MLOps*, behandelten Themen noch genauer ausgeführt werden.

Teil III, *MLOps-Anwendungsfälle aus der Praxis*, enthält konkrete Beispiele, die Ihnen zeigen sollen, wie MLOps heute in Unternehmen aussieht, wie es konzipiert ist und mit welchen Implikationen zu rechnen ist. Obwohl die Firmennamen fiktiv gewählt wurden, basieren die Beispiele auf Erfahrungen, die reale Unternehmen im Zusammenhang mit MLOps und einem im großen Maßstab angelegten Modellmanagement gemacht haben.

Danksagungen

Wir möchten dem gesamten Dataiku-Team für seine Unterstützung bei der Entstehung dieses Buchs danken, angefangen von der Konzeption bis hin zur Fertigstellung. Es war eine echte Teamleistung und ist, wie die meisten Dinge, die wir bei Dataiku machen, das Ergebnis einer intensiven Zusammenarbeit zwischen unzähligen Menschen und Teams.

Danke an alle, die unsere Idee, dieses Buch mit O'Reilly herauszubringen, von Anfang an unterstützt haben. Danke an alle, die beim Schreiben und Herausgeben mitgeholfen haben. Ebenfalls danke an alle, die uns ehrliches Feedback gegeben haben (auch wenn es zur Folge hatte, noch mehr zu schreiben oder umzuschreiben). Danke an alle, die uns intern stets ermutigt haben, und natürlich an alle, die uns geholfen haben, das finale Produkt der weltweiten Öffentlichkeit vorstellen zu können.

TEIL I

Was ist MLOps, und warum wird es benötigt?

KAPITEL 1

Warum jetzt, und was sind die Herausforderungen?

Machine Learning Operations (MLOps) entwickelt sich zusehends zu einer unverzichtbaren Komponente, um Data-Science-Projekte im Unternehmen erfolgreich in den Einsatz zu bringen (siehe Abbildung 1-1). Dabei handelt es sich um Prozesse, die dem Unternehmen und den Verantwortlichen dabei helfen, im Zusammenhang mit Data Science, Machine Learning und KI-Projekten langfristigen Wert zu generieren und Risiken zu reduzieren. Dennoch stellt MLOps ein relativ neues Konzept dar. Warum hat es also scheinbar über Nacht Einzug in das Data-Science-Lexikon erhalten? In diesem einführenden Kapitel wird erläutert, was MLOps auf einer übergeordneten Ebene ist, welche Herausforderungen es mit sich bringt, warum es für eine erfolgreiche Data-Science-Strategie im Unternehmen unverzichtbar geworden ist und, was besonders wichtig ist, warum es gerade jetzt in den Vordergrund rückt.

MLOps im Vergleich zu ModelOps und AIOps

MLOps (oder ModelOps) ist eine relativ neue Fachdisziplin, die seit Ende des Jahres 2018 unter diesen Namen in Erscheinung trat. Die beiden Termini – MLOps und ModelOps – werden zum Zeitpunkt der Erstellung dieses Buchs weitgehend synonym verwendet. Einige argumentieren jedoch, dass ModelOps umfassender als MLOps ist, da es nicht nur um Machine-Learning-(ML)-Modelle geht, sondern um jede Art von Modellen (z.B. auch regelbasierte Modelle). Im Rahmen dieses Buchs werden wir uns speziell mit dem Lebenszyklus von ML-Modellen befassen und daher den Begriff *MLOps* verwenden.

Auch wenn es manchmal mit MLOps verwechselt wird, bezieht sich AIOps hingegen auf ein ganz anderes Thema und bezeichnet den Prozess der Lösung operativer Herausforderungen im Rahmen des Einsatzes von künstlicher Intelligenz (d.h. KI für DevOps). Ein Beispiel wäre eine Form der vorausschauenden Wartung im Zusammenhang mit Netzwerkausfällen, bei der DevOps-Teams auf mögliche Probleme aufmerksam gemacht werden, bevor sie auftreten. Obwohl AIOps für sich genommen wichtig und interessant ist, liegt es außerhalb des Rahmens dieses Buchs.

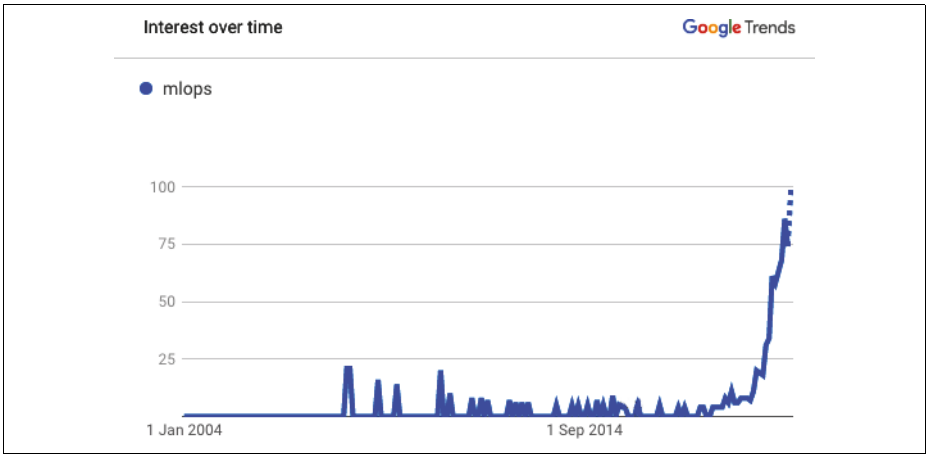

Abbildung 1-1: Darstellung des exponentiell verlaufenden Suchtrends des Begriffs »MLOps« (ohne gleichzeitige Berücksichtigung des Terminus »ModelOps«)

MLOps – Definition und Herausforderungen

Im Kern ist MLOps die Standardisierung und Straffung des Lebenszyklusmanagements von ML-Modellen (siehe Abbildung 1-2). Doch weshalb muss der ML-Lebenszyklus überhaupt gestrafft werden? Oberflächlich betrachtet, könnte man annehmen, dass die Arbeitsschritte, die vom Geschäftsproblem zu einem ML-Modell führen, sehr einfach sind.

Für die meisten traditionellen Unternehmen ist die Entwicklung mehrerer Machine-Learning-Modelle und deren Einsatz in einer Produktivumgebung relativ neu. Bis vor Kurzem war die Anzahl der Modelle vielleicht noch überschaubar, oder es bestand einfach weniger Interesse daran, diese Modelle und ihre Abhängigkeiten auf unternehmensweiter Ebene zu verstehen. Mit der fortschreitenden Automatisierung von Entscheidungsprozessen (d.h. mit einer zunehmenden Verbreitung von Entscheidungen, die ohne menschliches Zutun getroffen werden) rücken Modelle immer stärker in den Fokus, und parallel dazu wird auch das Management von Modellrisiken auf höchster Ebene immer wichtiger.

Insbesondere in Bezug auf die Anforderungen und die genutzten Tools erweist sich das Lebenszyklusmanagement von Machine-Learning-Modellen in einem Unternehmen tatsächlich als durchaus komplex (siehe Abbildung 1-3).

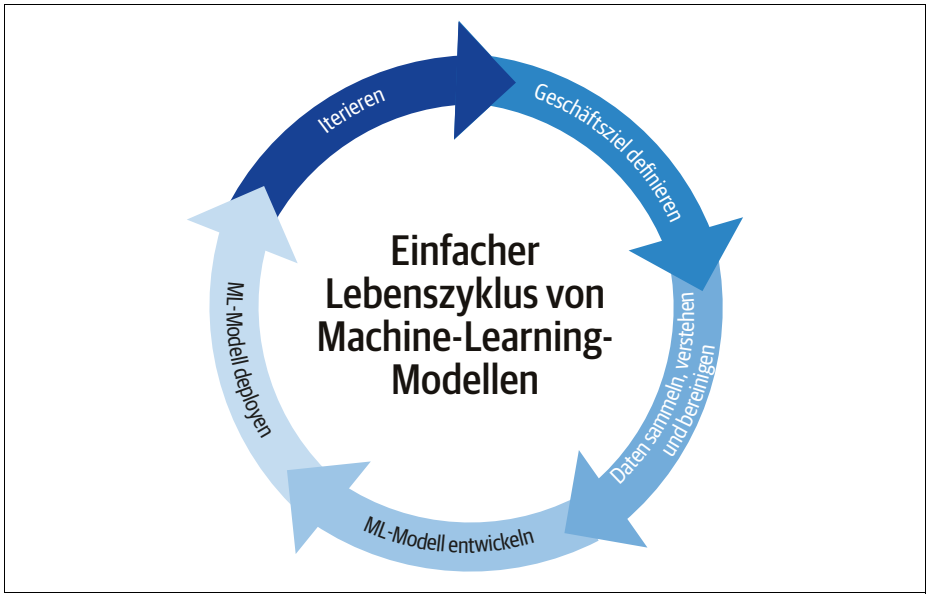

Abbildung 1-2: Eine vereinfachte Darstellung des Lebenszyklus von ML-Modellen, die die Notwendigkeit von MLOps nur unzureichend abbildet, speziell im Vergleich zu Abbildung 1-3

Es gibt drei Hauptgründe dafür, dass das Lebenszyklusmanagement skalierbarer ML-Modelle eine Herausforderung darstellt:

- Es gibt zahlreiche Abhängigkeiten. Nicht nur die Daten ändern sich ständig, sondern auch die geschäftlichen Anforderungen. Neue Informationen müssen kontinuierlich an das Unternehmen zurückgegeben werden, um sicherzustellen, dass der Produktivbetrieb des Modells, auch in Bezug auf die Akkuranz der Produktionsdaten, mit den Erwartungen übereinstimmt und – was von entscheidender Bedeutung ist – dass das ursprüngliche Problem gelöst bzw. die ursprüngliche Zielsetzung erreicht wird.
- Nicht alle sprechen die gleiche Sprache. Auch wenn am ML-Lebenszyklus Mitarbeiter aus Business-, Data-Science- und IT-Teams beteiligt sind, ist es nicht zwingend gegeben, dass diese Teams die gleichen Tools oder – in vielen Fällen – sogar die gleichen grundlegenden Fähigkeiten, die als Kommunikationsbasis dienen, teilen.
- Data Scientists sind keine Softwareentwickler. Die meisten sind auf die Entwicklung und Evaluierung von Modellen spezialisiert, und ihr Know-how liegt nicht zwingend in der Entwicklung von Anwendungen. Obwohl sich dies im Laufe der Zeit ändern könnte, da sich einige Data Scientists auf die Bereitstellung bzw. den Betrieb von Modellen spezialisieren werden, müssen derzeit viele Data Scientists mit verschiedenen Rollen gleichzeitig jonglieren, was es schwierig macht, eine davon vollständig auszufüllen. Die Überforde-

rung von Data Scientists wird insbesondere im Rahmen der Skalierung – wenn es immer mehr Modelle zu verwalten gibt – problematisch. Noch komplexer wird es, wenn man zusätzlich die Fluktuation der Mitarbeitenden in den Datenteams berücksichtigt: Schließlich gibt es nicht wenige Data Scientists, die sich plötzlich dazu gezwungen sehen, Modelle zu verwalten, die sie nicht selbst entwickelt haben.

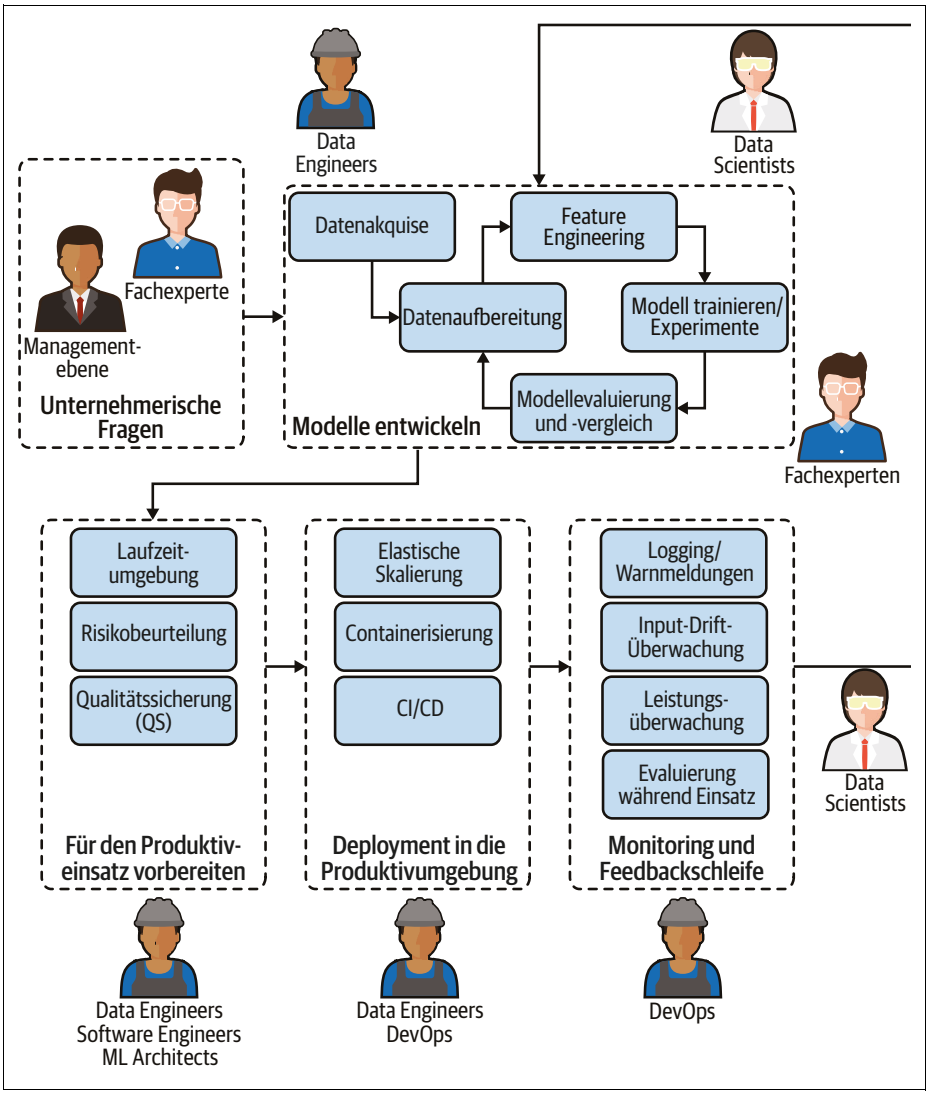

Abbildung 1-3: Ein realistischeres Bild des Lebenszyklus eines ML-Modells in einem modernen Unternehmen, in den viele verschiedene Personen mit völlig unterschiedlichen Fähigkeiten involviert sind, die oft völlig unterschiedliche Tools verwenden

Wenn Ihnen die Definition (oder lediglich die Bezeichnung MLOps) bekannt vorkommt, liegt das vor allem daran, dass sie sich stark an das Konzept, das hinter DevOps steht, anlehnt: DevOps dient dazu, die Prozesse im Rahmen von Softwareänderungen und -aktualisierungen zu straffen. In der Tat haben beide Konzepte ziemlich viel gemeinsam. Zum Beispiel geht es bei beiden darum,

- eine robuste Automatisierung und vertrauensvolle Zusammenarbeit zwischen den Teams zu gewährleisten,
- den Leitgedanken einer kooperativen Zusammenarbeit und einer verbesserten Kommunikation zwischen den Teams zu fördern,
- den Lebenszyklus des Diensts ganzheitlich (Build, Test, Release) zu berücksichtigen und
- den Schwerpunkt auf eine kontinuierliche Auslieferung (*Continuous Delivery*) und hohe Qualitätsanforderungen zu setzen.

Es gibt jedoch einen entscheidenden Unterschied zwischen MLOps und DevOps, der dafür sorgt, dass letzteres Konzept nicht sofort auf Data-Science-Teams übertragbar ist: In der Produktion unterscheidet sich das Deployment von Softwareprogrammen grundlegend vom Deployment von ML-Modellen. Während Softwareprogramme relativ statisch sind (»relativ«, da viele moderne Software-as-a-Service-(SaaS-)Unternehmen bereits über DevOps-Teams verfügen, die recht schnell iterieren und in der Produktion mehrmals am Tag deployen können), ändern sich Daten hingegen ständig, was bedeutet, dass ML-Modelle ständig neu (hinzu-)lernen und sich an neue Eingabedaten anpassen – oder eben nicht. Die dieser Umgebung zugrunde liegende Komplexität – einschließlich der Tatsache, dass ML-Modelle sowohl aus Programmcode als auch aus Daten bestehen – ist der Grund dafür, dass MLOps zu einer neuen und einzigartigen Disziplin heranwächst.

Und was hat es mit DataOps auf sich?

Zusätzlich zur komplexen Gegenüberstellung von MLOps und DevOps müssen wir noch den Begriff DataOps abgrenzen, der im Jahr 2014 von IBM eingeführt wurde. DataOps zielt darauf ab, geschäftsfähige Daten bereitzustellen, die schnell für die Nutzung verfügbar sind, wobei der Datenqualität und der Metadatenverwaltung ein besonderer Stellenwert beigemessen wird. Wenn es beispielsweise eine plötzliche Änderung in den Daten gibt, auf denen ein Modell beruht, würde ein DataOps-System das Businessteam alarmieren, damit es sich sorgfältig mit den neuesten Erkenntnissen befasst, und das Datenteam würde ebenfalls informiert werden, damit es die Änderung untersuchen oder ein Upgrade einer Bibliothek rückgängig machen und die entsprechende Partition neu erstellen kann.

Die Entwicklung von MLOps überschneidet sich daher auf einer gewissen Ebene mit DataOps, obwohl MLOps einen Schritt weitergeht und durch zusätzliche Kernfunktionen (die in Kapitel 3 ausführlicher erläutert werden) eine noch stärkere Robustheit bietet.

Wie bei DevOps und später auch bei DataOps konnten sich Teams bis vor Kurzem ohne vordefinierte und zentralisierte Prozesse behelfen, vor allem weil sie maschinelle Lernmodelle – auf Unternehmensebene – nicht in so großem Maßstab angelegt in die Produktion brachten. Jetzt wendet sich das Blatt, und die Teams suchen zunehmend nach Möglichkeiten, einen mehrstufigen, multidisziplinären und mehrphasigen Prozess mit einer heterogenen Umgebung und einem Rahmen für MLOps-Best-Practices zu formalisieren, was keine kleine Aufgabe darstellt. Teil II des Buchs, *MLOps einsetzen*, wird Ihnen hierzu einen Leitfaden bieten.

MLOps zum Reduzieren von Risiken

MLOps ist wichtig für jedes Team, das auch nur ein Modell im Produktivbetrieb hat, da je nach Modell eine kontinuierliche Leistungsüberwachung und -anpassung erforderlich ist. Indem es einen sicheren und zuverlässigen Betrieb ermöglicht, ist MLOps der Schlüssel zur Eindämmung der Risiken, die durch den Einsatz von ML-Modellen entstehen. Allerdings sind mit dem Einsatz von MLOps auch Kosten verbunden, für jeden Anwendungsfall sollte daher eine angemessene Kosten-Nutzen-Bewertung durchgeführt werden.

Risikobeurteilung

In Bezug auf ML-Modelle gibt es sehr unterschiedliche Risiken. Zum Beispiel sind die Risiken bei der Nutzung eines Empfehlungssystems, das einmal im Monat verwendet wird, um zu entscheiden, welches Marketingangebot an einen Kunden geschickt werden soll, viel geringer als bei einer Reiseplattform, deren Preissetzung und Umsatz von einem ML-Modell abhängen. Daher sollte sich die Analyse bei der Betrachtung von MLOps als Möglichkeit zur Risikominimierung auf folgende Risiken erstrecken:

- Das Risiko, dass das Modell für eine bestimmte Zeitspanne nicht verfügbar ist.
- Das Risiko, dass das Modell für eine bestimmte Beobachtung eine unzutreffende Vorhersage liefert.
- Das Risiko, dass die Genauigkeit oder die Fairness des Modells mit der Zeit abnimmt.
- Das Risiko, dass die zur Wartung des Modells erforderlichen Kompetenzen (d.h. die Fähigkeiten der jeweiligen Data Scientists) nicht mehr zur Verfügung stehen.

Bei Modellen, die weit verbreitet sind und außerhalb des eigenen Unternehmens eingesetzt werden, sind die Risiken in der Regel größer. Wie in Abbildung 1-4 gezeigt, basiert die Risikobeurteilung im Allgemeinen auf zwei Größen: der Eintrittswahrscheinlichkeit und dem Schadensausmaß des unerwünschten Ereignisses. Maßnahmen zur Risikominderung basieren in der Regel auf einer Kombination aus beidem, dem sogenannten Risikograd bzw. -ausmaß des Modells. Die Risikobeur-

teilung sollte zu Beginn eines jeden Projekts durchgeführt und in regelmäßigen Abständen neu bewertet werden, da Modelle auf eine ursprünglich nicht vorgesehene Weise verwendet werden können.

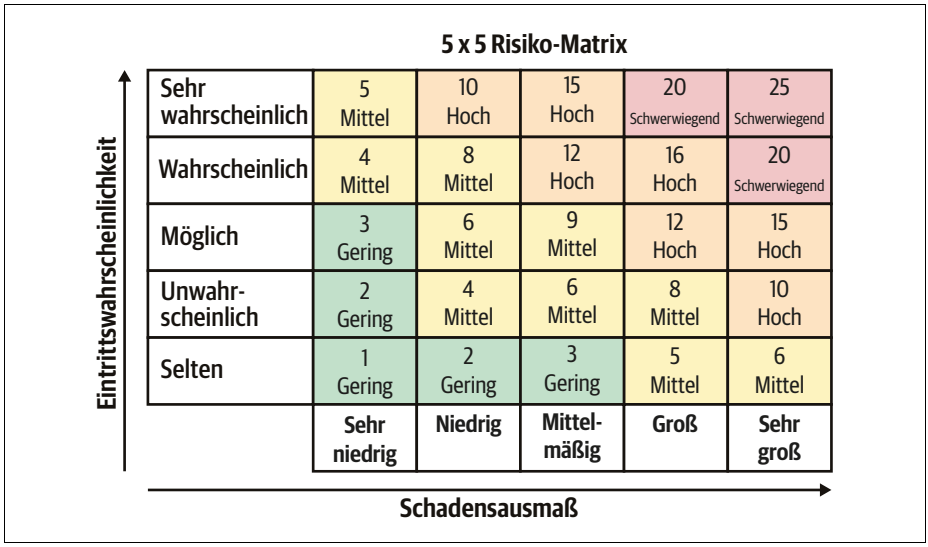

Abbildung 1-4: Eine Tabelle, die Entscheidungsträgern bei der quantitativen Risikobeurteilung hilft und auf Eintrittswahrscheinlichkeit und Schadensausmaß des Ereignisses basiert.

Risikominderung

MLOps trägt vor allem dann entscheidend zur Risikominderung bei, wenn ein zentrales Team (mit einer klaren Berichterstattung über seine Aktivitäten – was nicht bedeutet, dass es in einem Unternehmen nicht mehrere solcher Teams geben kann) mehr als eine Handvoll Modelle im operativen Einsatz hat. An diesem Punkt wird es schwierig, den Gesamtüberblick über die Zustände dieser Modelle ohne eine Form der Standardisierung zu behalten, die es ermöglicht, für jedes dieser Modelle die entsprechenden Maßnahmen zur Risikominderung ergreifen zu können (siehe den Abschnitt »Anpassung der Governance an das Risikoniveau« auf Seite 133).

Es ist aus vielen Gründen riskant, ML-Modelle in die Produktivumgebung zu überführen, ohne dass eine entsprechende MLOps-Infrastruktur vorhanden ist, zumal eine vollständige Bewertung der Leistung bzw. der Güte eines ML-Modells oft nur in der Produktivumgebung erfolgen kann. Warum? Weil Prognosemodelle nur so gut sind wie die Daten, auf denen sie trainiert wurden. Das bedeutet, dass die Trainingsdaten ein gutes Abbild der Daten sein müssen, die in der Produktivumgebung anfallen. Wenn sich die Rahmenbedingungen in der Produktion ändern, wird infolgedessen wahrscheinlich relativ schnell auch die Güte des Modells darunter leiden (siehe Kapitel 5 für Einzelheiten).