Fraunhofer

MARC HANUSSEK | HARALD PAPP | MATTHIAS BLOHM | MAXIMILIEN KINTZ
ARTHUR GRIGORJAN | DAVID BRANDT | CHRISTOPH HENNEBOLD | MICHAEL OBERLE

CLOUDBASIERTE KI-PLATTFORMEN

CHANCEN UND GRENZEN VON DIENSTEN FÜR MACHINE LEARNING AS A SERVICE

HRSG.: WILHELM BAUER | THOMAS BAUERNHANSL | MARCO HUBER | WERNER KRAUS
MATTHIAS PEISSNER | THOMAS RENNER

KI Fortschrittszentrum LERNENDE SYSTEME

Marc Hanussek, Harald Papp, Matthias Blohm, Maximilien Kintz
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

Arthur Grigorjan, David Brandt, Christoph Hennebold, Michael Oberle
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

CLOUDBASIERTE KI-PLATTFORMEN

Chancen und Grenzen von Diensten für Machine Learning as a Service

Herausgeber
Wilhelm Bauer, Thomas Bauernhansl, Marco Huber, Werner Kraus, Matthias Peissner, Thomas Renner

Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S·TEC

VORWORT

Künstliche Intelligenz (KI) ist eine der zentralen Technologien für die Zukunft. Ihre Einführung und der Einsatz fordern Unternehmen im besonderen Maß heraus. Es gilt, das Potenzial zu erkennen und dieses wirtschaftlich nutzbar zu machen. Lassen Sie sich dabei durch Europas größte Forschungskooperation auf dem Gebiet der KI, Cyber Valley, begleiten.

Mit dem KI-Fortschrittszentrum von Fraunhofer IAO und Fraunhofer IPA unterstützen wir Unternehmen dabei, das Potenzial von KI nutzbringend einzusetzen. An der Schnittstelle zwischen anwendungsorientierter Wirtschaft und exzellenter Forschung des Cyber-Valley-Konsortiums entwickeln wir innovative KI-Anwendungen für die Praxis und treiben damit die Kommerzialisierung von KI voran. Erklärtes Ziel ist dabei, menschzentrierte KI-Lösungen zu entwickeln.Denn nur wenn Menschen mit einer neuen Technologie intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann ihr Potenzial optimal ausgeschöpft werden.

Die Studienreihe »Lernende Systeme« des KI-Fortschrittszentrums gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Dabei werden übergreifende Themen wie Zuverlässigkeit, Erklärbarkeit (xAI), cloudbasierte Plattformen, Technologien und Einführungsstrategien diskutiert. Zudem werden einzelne Anwendungsbereiche in der Wissensarbeit, Bauwirtschaft, Produktion und dem Kundenservice im Detail beleuchtet.

Die vorliegende Studie vergleicht die vier größten cloudbasierten Machine-Learning-Plattformen und gibt anhand vier realitätsnaher Anwendungsfälle Empfehlungen über deren Eignung und Verwendung.

Wir wünschen Ihnen eine spannende Lektüre, und freuen uns, wenn wir in Zukunft auch Sie mit unserer Expertise auf Ihrem Weg zur menschzentrierten KI unterstützen dürfen.

Wilhelm Bauer, Matthias Peissner, Thomas Renner
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

Thomas Bauernhansl, Marco Huber, Werner Kraus
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

INHALT

ABBILDUNGEN

TABELLEN

1 MANAGEMENT SUMMARY

Heutige Unternehmen unterliegen einem stetigen Wandel. Aktuell wird dieser besonders durch volatile Märkte und einen damit einhergehenden steigenden Kostendruck verursacht. Kleine und mittelständische Unternehmen werden von dieser schnellen Entwicklung besonders gefordert, da ihnen oftmals die Expertise fehlt, moderne Technologien zeitnah einzuführen. Künstliche Intelligenz (KI), oder genauer der sich in den letzten Jahren rasant entwickelnde Teilbereich des Maschinellen Lernens (ML), birgt in vielen Anwendungsbereichen über Branchengrenzen hinweg das Potenzial, effizientere Vorgehensweisen zu etablieren und bestehende Geschäftsmodelle zu erweitern.

Oft scheitern Einführungen von KI bereits in der Anfangsphase, da nötiges Know-how oder anderweitige Ressourcen nicht ausreichend zur Verfügung stehen. Dieses Problem wurde von Lösungsanbietern früh erkannt, was dazu führte, dass mittlerweile mehrere Lösungen am Markt existieren, die Anwender*innen bei der Umsetzung von KI-Anwendungsfällen unterstützen sollen. Sie versprechen gute Ergebnisse, teilweise ohne dabei tiefes Wissen im Bereich des maschinellen Lernens zu erfordern.

In dieser Studie wurden die vier größten Cloud-Anbieter Amazon Web Services, Google Cloud Platform, IBM Cloud sowie Microsoft Azure bzw. deren cloudbasierte KI-Plattformen (Machine Learning as a Service, MLaaS) evaluiert. Dabei wurden zunächst allgemeine technische sowie geschäftliche Kriterien definiert, die Anwender*innen bei der Auswahl einer solchen Lösung unterstützen sollen. Die Plattformen wurden auf Basis einer Teilmenge dieser Kriterien bewertet. Aufgrund der Heterogenität der einzelnen Plattformen genügt ein allgemeiner Vergleich nicht, weswegen vier gängige KI-Anwendungsfälle untersucht wurden:

- Binäre Klassifikation auf tabellarischen Daten
- Textklassifikation
- Bildklassifikation
- Zeitreihenanalyse

Für jeden Anwendungsfall wurde ein realitätsnaher und öffentlich verfügbarer Datensatz verwendet und ein Satz an Bewertungskriterien definiert. Die Anwendungsfälle wurden jeweils auf allen vier Plattformen prototypisch umgesetzt.

Die Umsetzung zeigt, dass alle Anbieter anwendungsfallspezifische Stärken und Schwächen aufweisen und die optimale Wahl stark von individuellen Anforderungen abhängt. Für die umgesetzten Anwendungsfälle können jedoch Empfehlungen ausgesprochen werden:

Anwendungsfall	Anbieterempfehlung
Binäre Klassifikation auf tabellarischen Daten	IBM Watson, Microsoft Azure
Textklassifikation	Microsoft Azure
Bildklassifikation	Google Cloud Platform, Microsoft Azure
Zeitreihenanalyse	IBM Watson, AWS, Microsoft Azure

Tabelle 1: Die vier evaluierten Anwendungsgebiete und die empfohlenen Lösungsanbieter.

Unabhängig von konkreten Anwendungsfällen zeigt diese Studie die Vor- und Nachteile von MLaaS-Lösungen auf. Einerseits geben die Lösungen einem breiten Kreis von Anwender*innen vielfältige mächtige Werkzeuge an die Hand, unterstützt durch fast unbegrenzte Rechenkapazität und einen komfortablen Zugriff über den Webbrowser. An fehlender Funktionalität auf den Plattformen sollten tatsächlich die wenigsten KI-Vorhaben scheitern. Andererseits erschwert die Vielzahl von Teildiensten, Möglichkeiten und Bezeichnungen sowie deren Beziehungen untereinander einen schnellen Einstieg. Damit wird deutlich, dass die Plattformen einem schnellen Wandel unterliegen, der wiederum Folge eines harten Wettbewerbs auf dem zukunftsträchtigen MLaaS-Markt ist. Wie lange heute verfügbare Lösungen oder Schnittstellen zukünftig unverändert nutzbar bleiben, ist unklar.

KI-Expert*innen können die meisten Anwendungsfälle bedenkenlos ohne die Benutzung von KI-Plattformen durchführen. Eine Einbindung solcher Dienste ist insbesondere dann sinnvoll, wenn ausgesprochen leistungsfähige Infrastruktur vonnöten ist oder sich ein Anwendungsfall über mehrere, miteinander verbundene Teildienste einer KI-Plattform erstreckt, sodass Synergieeffekte genutzt werden können. KI-Einsteiger*innen, die einerseits durch anwendungsfreundliche und automatisierte Module angesprochen werden, können andererseits durch den unüberschaubaren Funktionsumfang der Plattformen überfordert werden.

2 EINLEITUNG

Zunächst wird auf den Hintergrund und die Motivation dieser Studie eingegangen. Es wird außerdem gezeigt, welche Fragen diverser Zielgruppen beantwortet werden und auf Begriffsdefinitionen eingegangen. Abschließend werden der Aufbau sowie die Methodik der Studie beleuchtet.

2.1 Hintergrund und Motivation

Vor allem kleine und mittlere Unternehmen stehen in Zeiten fortschreitender Digitalisierung vor der Frage, wie Methoden der Künstlichen Intelligenz die Wertschöpfung im Betrieb steigern können. Der Wettbewerb am Arbeitsmarkt um spezialisierte Fachkräfte ist allerdings groß und die Umsetzung entsprechender Projekte teuer. In der Folge droht vielen KI–Ideen, trotz möglicherweise großen Potenzials, die frühe Auflösung.

Abhilfe versprechen cloudbasierte KI-Plattformen von Unternehmen wie Amazon, Google, IBM oder Microsoft. Auch kleinere Anbieter beteiligen sich mit spezialisierten Dienstleistungen und Nischenprodukten am Markt. Grundidee ist, gebräuchliche Methoden und Werkzeuge eines KI-Projekts intuitiv und Arbeitsabläufe von KI-Entwickler*innen automatisiert in der Cloud zur Verfügung zu stellen. Dadurch soll einerseits die Anschaffung teurer Hardware entfallen und anderseits die Nutzung der Plattform durch Fachfremde ermöglicht werden.

Während der Markt für Infrastructure as a Service, wozu auch KI-Dienstleistungen in der Cloud gehören, weltweit rasant steigt (siehe Abbildung 1) und zukünftig weiterhin hohe Wachstumsraten erwartet werden, erfolgt der KI-Einsatz in der deutschen Wirtschaft verhalten (siehe Abbildung 2).

Umsatz mit Infrastructure as a Service (IaaS) weltweit von 2010 bis 2019 und Prognose bis 2022 (in Milliarden US-Dollar)

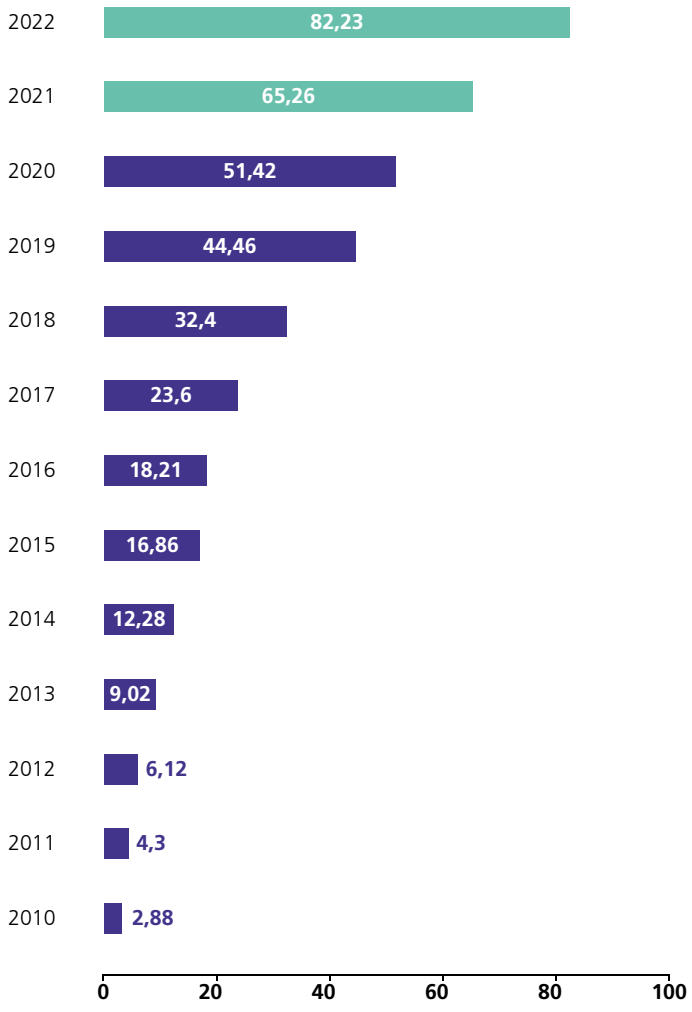


Abbildung 1: Umsatz im Markt für Infrastructure as a Service[1].

1 *Quelle: https://de.statista.com/statistik/daten/studie/307025/umfrage/umsatz-mit-infrastructure-as-a-service-weltweit-seit-2010/*

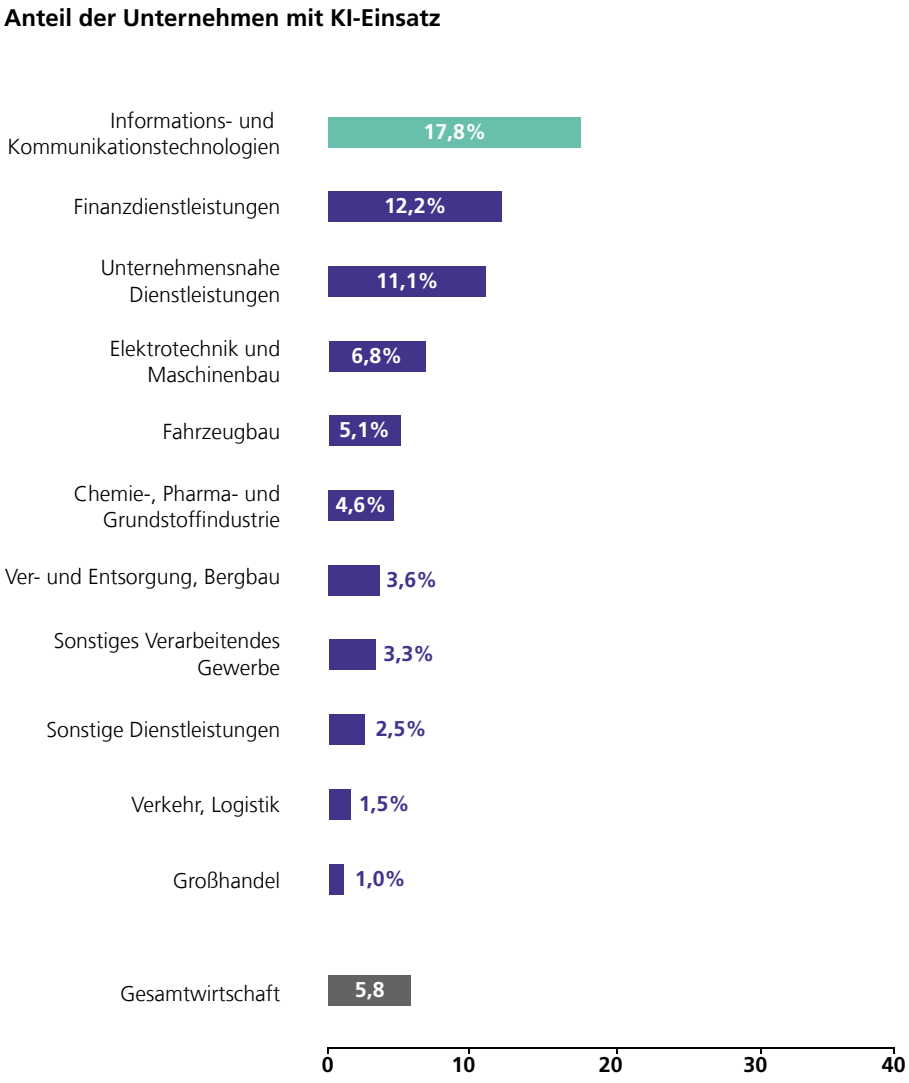


Abbildung 2: Anteil der Unternehmen mit KI-Einsatz (in Prozent aller Unternehmen[2]).

Die Vielzahl der Dienste und die Spezifizität der Lösungen macht eine Auswahl gerade für Fachfremde schwer. Es fehlt an Übersicht und an Bewertungskriterien, die es ermöglichen, KI-Plattformen unabhängig von Anbietern, aber auf das individuelle Problem angepasst, miteinander zu vergleichen.

2 Quelle: [4].

2.2 Zielgruppen und Zielstellungen

Diese Studie soll daher Entscheiderinnen und Entscheidern wie CIOs oder CTOs eine Hilfestellung und Zusammenfassung bieten, damit sie einordnen können, inwiefern sich cloudbasierte KI-Plattformen überhaupt für das eigene Unternehmen eignen und welche Plattform den individuellen Ansprüchen gerecht wird. Zum anderen soll die Studie den nötigen Grad an Expertise zur Nutzung der cloudbasierten Plattformen offenlegen. Das Personalressourcen-Management sowie technische Akteur*innen im Unternehmen wie IT-Beauftragte sollen damit leichter eruieren können, welche konkreten Dienste der cloudbasierten KI-Plattformen sich für welche KI-Use-Cases anbieten, besonders in Abhängigkeit von der Erfahrung der Arbeitskräfte.

Weiterhin untersucht diese Studie den Einfluss cloudbasierter KI-Plattformen auf die Arbeitsweise im Bereich des Maschinellen Lernens. Auch erfahrene, spezialisierte Arbeitskräfte wie Data Scientists oder Analyst*innen sollen so evaluieren können, wie sich deren Arbeitsweise durch cloudbasierte KI-Plattformen unterstützen lässt, wo Mehrwert durch Automatisierung der Arbeitsprozesse generiert werden kann und welche Aspekte doch weiterhin von Hand gelöst werden sollten. Tabelle 2 bietet eine Zusammenfassung der Zielgruppen und Zielstellungen.

Zielgruppe	Fragestellungen	Kapitel
Manager*innen, Entscheider*innen, CIOs, CTOs	Welche geschäftlichen Kriterien sind bei der Auswahl einer cloudbasierten KI-Plattform zu beachten?	4.1.3
	Welche cloudbasierten KI-Plattformen existieren auf dem Markt?	3
	Welche cloudbasierte KI-Plattform bietet sich aus strategischer Sicht für mein Unternehmen an?	4, 6
Mittleres Management, IT-Beauftragte, Personalressourcenzuständige	Welche Dienste der cloudbasierten KI-Plattformen unterstützen meine konkreten KI-Use-Cases?	5.4
	Welche technischen Kriterien existieren für die Auswahl des passgenauen Dienstes?	4.1.1, 4.1.2
	Welche Mitarbeiter *innen-Expertise muss ich für ein Projekt einplanen?	5.4
Data Analysts, Data Scientists, Machine Learning Engineers	Welchen Teil meines Workflows kann ich durch cloudbasierte KI-Plattformen automatisieren?	5.4
	Welche konkreten Dienste bieten mir Baseline-Modelle für die Erstellung von ML-Modellen?	5.4
	Welche Aspekte der Dienste eignen sich nicht für mich?	5

Tabelle 2: Zielgruppen der Studie und deren Fragestellungen.

2.3 Definition cloudbasierter KI-Plattformen und Abgrenzung

Unter cloudbasierten KI-Plattformen verstehen wir in dieser Studie Machine-Learning-as-a-Service-Lösungen (MLaaS-Lösungen). Wie auch bei anderen »as a Service«-Begriffen wie Software as a Service oder Infrastructure as a Service geht es bei MLaaS-Lösungen darum, Dienstleistungen, die für die Benutzung von Machine Learning notwendig oder sinnvoll sind, bei Bedarf zu vermieten. Konkret heißt das, dass die MLaaS-Lösungen Datenspeicher, Rechenkapazität, Algorithmen, Schnittstellen und vieles mehr den Endnutzer*innen über die Cloud zur Verfügung stellen. Insbesondere muss ein User weder lokal Software installieren noch Server oder andere IT-Ressourcen betreiben. Nutzer*innen können im Idealfall zu jeder Zeit uneingeschränkt auf die Ressourcen des Dienstes zugreifen.

Machine-Learning-Softwarebibliotheken wie TensorFlow[3] oder scikit-learn[4], aber auch Software mit KI-Funktionalitäten wie Knime[5] oder RapidMiner[6] zählen nicht zu MLaaS-Lösungen, da man sie lokal installieren oder betreiben muss.

Eine weitere Anforderung, die wir an die zu betrachtenden Dienste stellen, ist die Eignung für möglichst viele Benutzergruppen, insbesondere solche, die nur über Grundlagenkenntnisse des Machine Learnings oder des Programmierens verfügen. Eine solche Eignung kann zum Beispiel über grafische Oberflächen oder hochgradig automatisierte Teildienste erreicht werden und schließt andererseits rein codebasierte Plattformen aus. Tabelle 3 erläutert Fachbegriffe, die öfter in dieser Studie vorkommen werden.

3 *https://www.tensorflow.org/*

4 *https://scikit-learn.org/stable/*

5 *https://www.knime.com/*

6 *https://rapidminer.com/*

Begriff	Erklärung
Accuracy	Anteil korrekter Vorhersagen an allen Vorhersagen
F1-Score	Harmonisches Mittel von Precision und Recall, F1=2*(Precision*Recall)/(Precision+Recall)
Named Entity Recognition	Verfahren, das Elemente wie Namen, Orte etc. in Texten markiert
Precision	Anteil korrekt erkannter, relevanter Entitäten zu allen Entitäten
Recall	Anteil korrekt erkannter, relevanter Entitäten zur Gesamtheit relevanter Entitäten
REST-Schnittstelle	HTTP-Schnittstelle eines RESTful Web Service
RESTful Web Service	RESTful Web Services ermöglichen es den anfragenden Systemen, auf textuelle Darstellungen von Web-Ressourcen zuzugreifen und diese zu manipulieren, indem sie einen einheitlichen und vordefinierten Satz von zustandslosen Operationen verwenden.
Service/Dienst	Der Begriff Service bezieht sich auf eine Softwarefunktionalität mit einem Nutzen, die von verschiedenen Anwender*innen für unterschiedliche Zwecke wiederverwendet werden kann.
Testdatensatz	Im Idealfall sollte ein Machine-Learning-Modell an Proben evaluiert werden, die nicht zur Erstellung oder Feinabstimmung des Modells verwendet wurden, sodass sie ein unvoreingenommenes Bild von der Leistungsfähigkeit des Modells vermitteln.
Trainingsdatensatz	Ein Trainingsdatensatz ist ein Datensatz mit Beispielen, der während des Lernprozesses eines Machine Learning Modells verwendet wird und dazu dient, dessen Parameter (z. B. Gewichte innerhalb eines neuronalen Netzes), z. B. eines Klassifikators, anzupassen.
Validierungsdatensatz	Ein Validierungsdatensatz ist ein Beispieldatensatz, der zur Abstimmung der Hyperparameter (d. h. der Architektur) eines Machine-Learning-Modells verwendet wird. Ein Beispiel für einen Hyperparameter für künstliche neuronale Netze enthält die Anzahl der Neuronen in jeder Schicht.

Tabelle 3: Begriffsdefinitionen.

2.4 Aufbau der Studie

Die Studie ist in vier inhaltliche Teile gegliedert:

Kapitel 3 beschreibt, wie die Auswahl der betrachteten Plattformen zustande kam und stellt die vier Plattformen vor (Amazon Web Services, Google Cloud Platform, IBM Watson, Microsoft Azure). Zudem werden alternative Lösungen angesprochen, die jedoch nicht in derselben Tiefe betrachtet wurden. Es wird auf potenzielle Risikofaktoren eingegangen, die bei der Nutzung von Machine Learning as a Service zu berücksichtigen sind.

Kapitel 4 stellt einen allgemeinen Plattformvergleich auf und beschreibt technische sowie geschäftliche Kriterien, die hier relevant sind. Eine Teilmenge dieser Kriterien wird verwendet, um die Plattformen zu bewerten.

Kapitel 5 adressiert den Use-Case-spezifischen Plattformvergleich. Eingangs werden die Anwendungsfälle anhand von User Storys geschildert, um zu verdeutlichen, wie die Problemstellungen zustande kommen könnten. Des Weiteren werden die Bewertungskriterien der jeweiligen Anwendungsfälle beschrieben und der gesetzte Betrachtungsrahmen erläutert. Abschließend wird bei der Use-Case-spezifischen Bewertung der Versuchsablauf der einzelnen Lösungsumsetzungen beschrieben und nach den zuvor erwähnten Kriterien bewertet.

Kapitel 6 fasst die gesammelten Erkenntnisse in einem Fazit zusammen und gibt Leserinnen und Lesern Empfehlungen, die für eine Entscheidungsfindung oder Umsetzung relevant sind.

Die in dieser Studie verwendeten Abbildungen sind, sofern keine Quelle angegeben ist, entweder selbst erstellt worden oder sie wurden in Form eines Screenshots aufgenommen. Diese wurden in Q3–Q42020 aufgenommen.

2.5 Methodik der Studie

Diese Studie unternimmt eine qualitative Bewertung der MLaaS-Lösungen der vier großen Cloud-Anbieter (AWS, Google Cloud Platform, IBM Cloud und Microsoft Azure). Dabei wurde ein Vergleich auf zwei Ebenen durchgeführt:

- Allgemeiner Vergleich der Plattformen
- Anwendungsfallspezifischer Vergleich der Plattformen

Im allgemeinen Vergleich wurden Kriterien definiert, die diverse Aspekte der primären MLaaS-Lösungen der jeweiligen Anbieter widerspiegeln, um diese dann einander gegenüberzustellen. Da diverse Kriterien subjektiv und individuell zu beurteilen sind, wurden die Plattformen nach einer Teilmenge der Kriterien bewertet. Diese allgemeinen Kriterien sind jedoch nicht ausreichend, um einen gerechten Vergleich zu ermöglichen, da die Anbieter heterogene, anwendungsfall-spezifische Lösungsdienste bieten. Aus diesem Grund wurden vier gängige Anwendungs-gebiete für KI ausgewählt und prototypisch umgesetzt:

- Binäre Klassifikation auf tabellarischen Daten
- Textklassifikation
- Bildklassifikation
- Zeitreihenanalyse

In der Beschreibung der Anwendungsfälle wird je ein Szenario geschildert, das zur jeweiligen Umsetzung führen könnte. Dabei werden beteiligte Personen und deren Expertise im Themen-feld der KI beschrieben. Grundsätzlich wurde stets ein Lösungsweg gewählt, der möglichst geringe Vorkenntnisse der Umsetzenden verlangt. Bei der Beschreibung der Lösungsumsetzun-gen wurde dokumentiert, wie die Umsetzung durchgeführt wurde und welche Aspekte dabei wahrgenommen wurden. Für jeden Use Case wurden anwendungsfallspezifische Kriterien definiert, die von Relevanz sind. Der anwendungsfallspezifische Vergleich schließt mit einer Ge-genüberstellung der Lösungen ab, wobei die bereits erwähnten Kriterien angewendet werden.

3 DIE BETRACHTETEN PLATTFORMEN

In diesem Kapitel werden die vier betrachteten Plattformen vorgestellt, auch eine Open-Source-Alternative wird angesprochen. Abschließend wird erläutert, welche Risikofaktoren bei der Einführung und Benutzung von MLaaS-Lösungen eine Rolle spielen könnten.

3.1 Auswahl der vorgestellten Plattformen

Die im Rahmen dieser Studie betrachteten Plattformen von Google Cloud, Amazon Web Services, Microsoft Azure und IBM Cloud wurden unter einer Vielzahl an Anbietern im Bereich Cloud Computing zur näheren Betrachtung ausgewählt. Gründe hierfür sind zum einen der bereitgestellte Funktionsumfang der jeweiligen Plattform, da innerhalb dieser Studie unterschiedliche Anwendungsfälle mit verschiedensten Anforderungen an Nutzer*innen betrachtet werden. Es wurden daher Plattformen gewählt, deren Funktionsumfang die Bereiche Text- und Bildverarbeitung sowie AutoML abdeckt. Zur Einschätzung des Funktionsumfangs der verschiedenen Plattformen möchten wir an dieser Stelle auf die von Gartner erstellte Studie [9] verweisen. Sie ordnet die vier genannten Plattformen in die Gruppe der sogenannten Leader ein, wobei ein Leader die drei Schlüsselbereiche Sprachverarbeitung, Bilderkennung und AutoML bedient.

Zum anderen wurde die Auswahl anhand des Marktanteils am Gesamtmarkt getroffen. Dabei hat sich, Stand Q4 2019, für die vier Anbieter ein weltweit kumulierter Marktanteil von 65 Prozent ergeben. Die zwei größten Plattformen bilden dabei Amazon Web Services und Microsoft Azure mit zusammen 51 Prozent des Marktanteils (siehe Abbildung 3). Die ausgewählten Plattformen zeichnen sich außerdem dadurch aus, dass sie ein breites Spektrum an Lösungen für KI-Dienste anbieten und durch ihre tiefe Verwurzelung in den Alltag vielen Menschen bekannt sind. Eine allgemeine Zusammenfassung findet sich in Tabelle 4.

Amazon ist die Nummer 1 in der Cloud

Weltweiter Marktanteil von Cloud-Infrastruktur-Dienstleistern in Q4 2019*

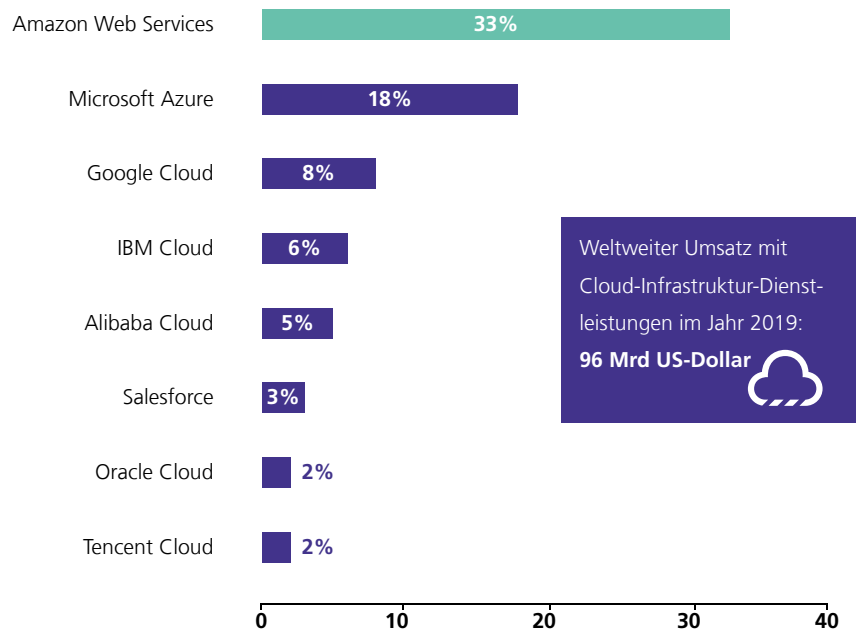


Abbildung 3: Marktanteil nach Cloud-Infrastruktur-Dienstleistern im Q4 2019[7].

**Inklusive IaaS, PaaS, und Private-Cloud-Dienstleistungen*
Quelle: Synergy Research Group.

7 Quelle: https://de.statista.com/infografik/20802/weltweiter-marktanteil-von-cloud-infrastruktur-dienstleistern/

Tabelle 4: Allgemeine Plattforminformationen.

	Amazon Web Services	Google Cloud Platform	IBM Watson	Microsoft Azure
Umsatz in Mrd US-Dollar (Q4 19)	31,7	7,7	5,8	17,3
Marktanteil in % (Q4 19)	33	8	6	18
Gründung der Mutter-plattform	2006 AWS	2008 GCP	2014 als Bluemix 2017 Rebranding zu IBM Cloud	2010 Windows Azure 2014 Rebranding zu Microsoft Azure
Gründung der MLaaS-Lösung	2017 AWS SageMaker	2018 Google AI-Plattform	2018[8] IBM Watson Studio	2014 Azure Machine Learning
URL	https://aws.amazon.com/de/sagemaker/	https://cloud.google.com/ai-platform	https://www.ibm.com/de-de/cloud/watson-studio	https://azure.microsoft.com/de-de/services/machine-learning/

3.2 Amazon Web Services

Mit Amazon Web Services (AWS) stieg der Amazon-Konzern im Jahr 2006 in die Cloudbranche ein. Das angebotene Dienstleistungsspektrum umfasst neben Software as a Service (SaaS) und Platform as a Service (PaaS) auch Infrastructure as a Service (IaaS). AWS erzielt konstant hohe jährliche Wachstumsraten, wodurch Amazon der marktführende Anbieter von Cloud-Diensten weltweit ist.

In dieser Studie standen primär die AWS-Dienste SageMaker, Rekognition und Comprehend bei der Umsetzung der unterschiedlichen Anwendungsfälle im Vordergrund. Amazon SageMaker ist die Machine-Learning-Plattform von AWS, die 2017 veröffentlicht wurde, um Data-Science-Prozesse zu unterstützen. Dabei werden die vier Hauptphasen Datenkennzeichnung (Labeling), Entwicklung (Build), Training (Train & Tune) sowie die Bereitstellung und Verwaltung (Deploy & Manage), die über mehrere Dienste hinweg angeboten werden, abgedeckt. Der folgende Abschnitt geht näher auf einige dieser Dienste ein.

8 *https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&htmlfid=899/ENUSLP18-0548*

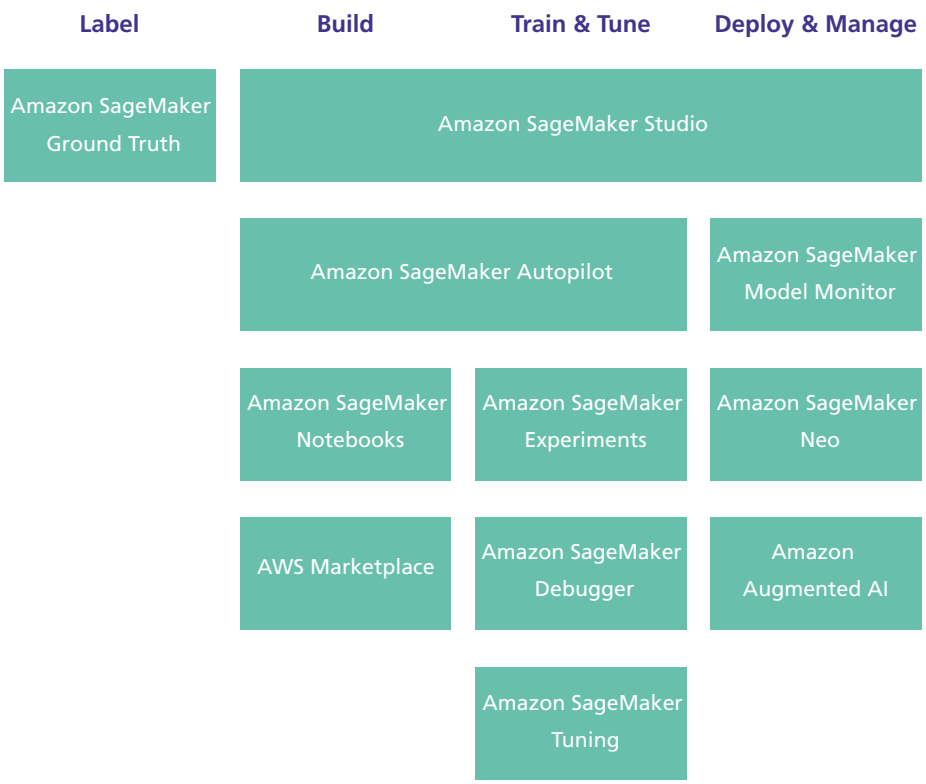


Abbildung 4: AWS – Amazon SageMakers Teilkomponenten und deren Einordnung im Prozess.

Die zu verarbeitenden Daten können mithilfe von SageMaker Ground Truth gekennzeichnet werden (Labeling). SageMaker Studio unterstützt die Entwickler+innen in den nächsten drei Phasen. Dabei handelt es sich um eine integrierte webbasierte Entwicklungsumgebung. Mit diesem Tool wird für Unternehmen und Entwickler*innen eine Entwicklungsumgebung bereitgestellt, in der zugeschnittene Modelle erstellt werden können. Benutzer*innen erhalten vollständigen Zugriff auf alle Schritte, die zum Erstellen, Trainieren und Bereitstellen von Modellen erforderlich sind.

SageMaker Notebooks ermöglicht es Teams, kollaborativ an Jupyter Notebooks zu arbeiten. Die unterliegenden Rechenressourcen können hierbei nach Bedarf skaliert werden. SageMaker Autopilot war nach Angaben von AWS die erste automatisierte Machine-Learning-Lösung in der Branche, in der Modelle automatisch erstellt und trainiert werden können. Dieser Ansatz zeichnet sich durch seine einfache Handhabung aus und benötigt keine Erfahrungen mit maschinellen Lernverfahren. Auch für erfahrene Entwickler*innen ist der Autopilot interessant, um schnell Ergebnisse zu erzielen und auf diesen aufzubauen. Durch die Integration in SageMaker Studio ist auch hierbei eine kollaborative Arbeit im Team möglich.

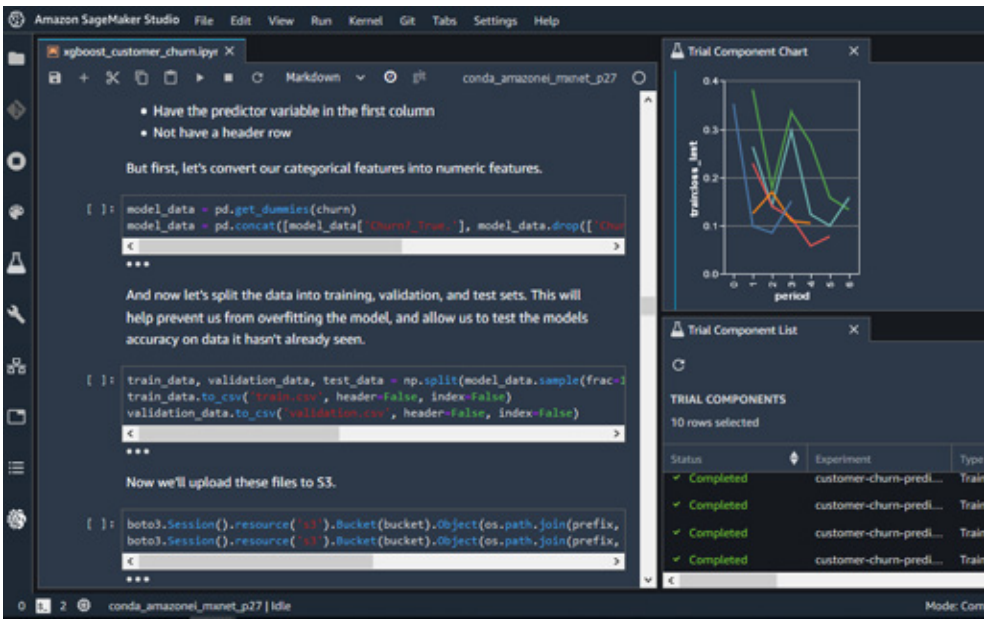


Abbildung 5: Ausschnitt aus SageMaker Studio in der ein Studio Notebook und deren Ergebnis zu sehen ist[9].

Amazon Rekognition ist ein Dienst, der Analysen von Bild- und Videomaterial ermöglicht. Dabei können bereits vortrainierte Modelle genutzt werden, die diverse Objektklassen in Bilder und Videos erkennen oder durch die Nutzung eines eigenen Datensatzes ein eigenes Modell trainiert werden. Amazon Comprehend ermöglicht die Analyse von Texten (Natural Language Processing, NLP). Der Dienst identifiziert die Sprache von Texten, extrahiert Schlüsselwörter, Orte, Personen, Marken oder Ereignisse, erkennt die Stimmung eines Textes, analysiert Text mittels Tokenisierung und Parts of Speech (PoS) und organisiert automatisch die Erfassung von Textdateien, geordnet nach Themenfeldern.

3.3 Google Cloud Platform

Mit der Google Cloud Platform (GCP) stieg der Google-Konzern 2008 in die Cloudsparte ein. Sie bietet Funktionalitäten zur Entwicklung, Erprobung, Bereitstellung und Verwaltung von Anwendungen und Diensten. GCP bietet Dienstleistungsmodelle wie Software as a Service (SaaS), Platform as a Service (PaaS) und auch Infrastructure as a Service (IaaS).

9 *https://aws.amazon.com/de/sagemaker/*

Die AI Platform ist Teil der GCP und setzt sich als Ziel, Anwender*innen bei der Erstellung von KI-Anwendungen zu unterstützen und diese sowohl in der GCP als auch lokal verfügbar zu machen. Über die Plattform können Entwickler*innen, Data Scientists und Data Engineers im Bereich maschinelles Lernen ihre ML-Projekte von der Konzeption über die Umsetzung bis hin zur Bereitstellung durchführen.

Mit der Unterstützung der Open-Source-Plattform Kubeflow ist es möglich, portierbare ML-Pipelines zu erstellen, welche lokal oder in der Google Cloud ausgeführt werden können, ohne eine notwendige Codeanpassung vorzunehmen. Mit AI Platform Notebooks wird ein Service bereitgestellt, mit dem Nutzer*innen eine sichere JupyterLab-Umgebung erhalten. Es können Instanzen erstellt werden, in denen die verbreitetsten Frameworks bereits vorinstalliert sind. Dies ermöglicht eine kollaborative Zusammenarbeit am Quellcode in fortgeschrittenen ML-Projekten. Mit dem Dienst AI Platform Training können Benutzer*innen ihre Trainingsprozesse auf Rechenressourcen in der Cloud ausführen. Die trainierten Modelle können dann durch den Dienst AI Platform Prediction für Vorhersagen bereitgestellt werden.

In dieser Studie standen die Dienste AutoML Vision, AutoML Table und AutoML Natural Language im Vordergrund. Mit AutoML Vision bietet die GCP ein fast voll automatisiertes Werkzeug zur Erstellung von Modellen zur Bilderkennung bzw. Objekterkennung. Die Anwender*innen liefern hierbei nur den gekennzeichneten Datensatz und starten das Training, ohne eine Modellarchitektur oder ähnliches konzipieren zu müssen. Die erzeugten Modelle können direkt in der Cloud bereitgestellt oder für die Installation auf eigener Hardware exportiert werden.

Möchten Anwender*innen aus strukturierten Daten automatisch Modelle entwickeln und bereitstellen, kann dafür AutoML Tables genutzt werden. Der Dienst kümmert sich beim sogenannten Feature Engineering der Daten um fehlende Werte, Ausreißer und andere häufige Datenprobleme. Auch hier wird eine grafische Oberfläche genutzt, die die Benutzer*innen durch den gesamten ML-Lebenszyklus begleitet. Darin werden unter anderem Funktionen angeboten, welche die Nachvollziehbarkeit von Eingabedaten sicherstellen. AutoML Tables testet mehrere Modellarchitekturen, um damit bessere Ergebnisse zu ermöglichen.

Mit AutoML Natural Language lassen sich, Modelle generieren, die sich für die Klassifizierung, Named Entity Recognition sowie Sentimentanalysen in Textdokumenten eignen.

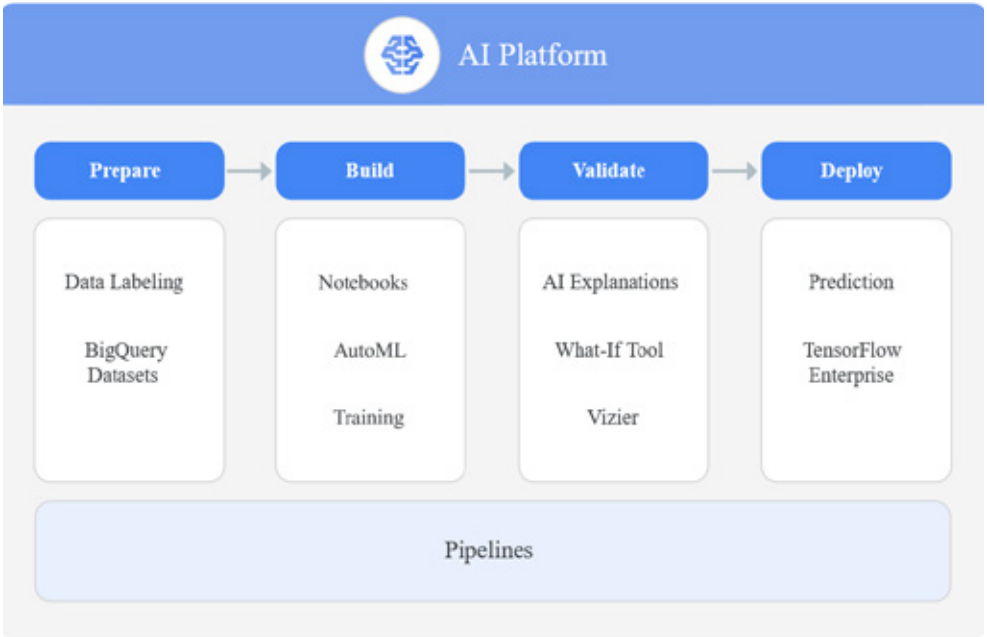


Abbildung 6: GCP AI Platform – Teillösungen für die jeweiligen Phasen des Entwicklungsprozesses[10].

Die GCP bietet zum Einstieg für Neukund*innen ein Startguthaben in Höhe von 300 US-Dollar[11]. Dieses kann für beliebige GCP-Produkte verwendet werden und verfällt nach 90 Tagen.

3.4 IBM Watson

Watson ist eine cloudbasierte KI-Plattform, betrieben von dem 2014 gegründeten IBM-Geschäftsbereich Watson Group[12]. Ursprünglich als ein KI-Programm (benannt nach dem ersten Präsidenten von IBM, Thomas J. Watson) zur Beantwortung von Fragen in natürlicher Sprache [7] entwickelt, trat es medienwirksam im Jahr 2011 bei der Quizsendung »Jeopardy!« gegen menschliche Kontrahenten an[13]. 2020 verfügt Watson über eine Fülle verschiedener Dienste, die die meisten Anwendungsgebiete der KI (Bild, Sprache, Text, Datenanalyse) abdecken, sowie über 30 Programmierschnittstellen, worüber die Dienste ansprechbar sind. IBM selbst bevorzugt in diesem Kontext den Begriff »cognitive«. Dazu bemerkte Steve Abrams, Direktor der IBM-Watson-Plattform, dass mit dem Wort »cognitive« gemeint ist, dass Watson lernen, verstehen, argumentieren und interagieren kann.

10 *Quelle: https://cloud.google.com/ai-platform*

11 *https://cloud.google.com/free*

12 *https://www-03.ibm.com/press/us/en/pressrelease/42869.wss*

13 *https://www.nytimes.com/2011/02/17/us/17deepblue.html*

Watson Machine Learning ist ein Produkt aus der Watson-Suite auf IBMs Cloudplattform, das als KI-Lifecycle Management Tool fungiert. IBM ermöglicht eine integrierte Nutzung mit den anderen Produkten der KI-Lifecycle Management Tools, wie Watson Studio und Watson Open-Scale. Neben der IBM-Cloud lässt sich Watson Machine Learning im sogenannten Local-Dienst (der allerdings ein eigenes Preismodell hat) auch auf einer beliebigen anderen Cloud oder auf eigenen Servern implementieren.

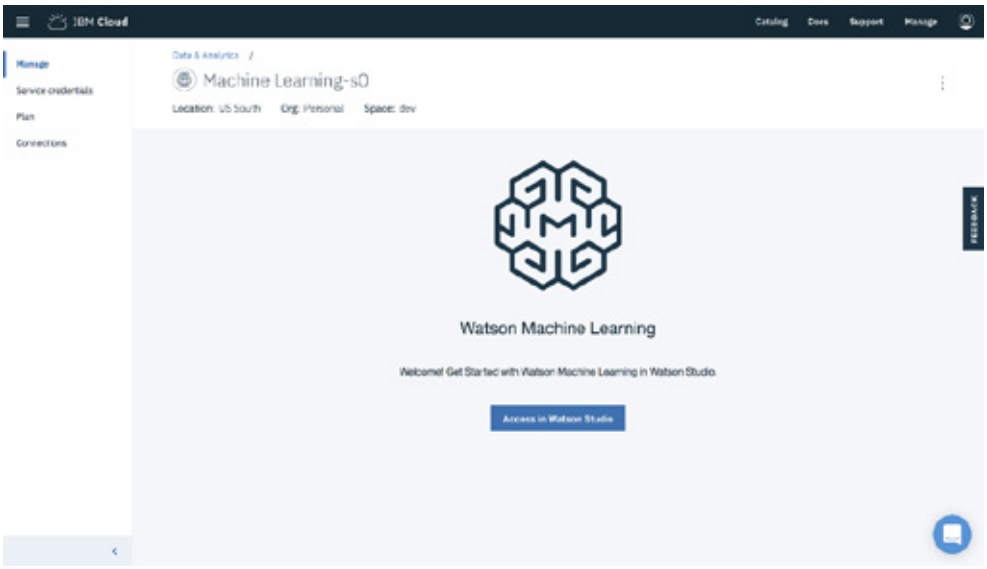


Abbildung 7: IBM Watson Studio mit Watson Machine Learning auf der IBM-Cloud[14].

Data Scientists und Analyst*innen soll geholfen werden, indem sich Funktionen und Modelle aus diversen Open-Source-Bibliotheken (Apache Spark MLib, scikit-learn, TensorFlow) sowie Python-Funktionen über wenige Klicks aufsetzen lassen. Diese können dann u. a. über zur Verfügung gestellte REST-API-Endpunkte genutzt oder trainiert werden. Unter AutoAI ist außerdem ein Modul integriert, das Datenaufbereitung, Modellerstellung, Feature-Engineering und Hyperparameteroptimierung automatisiert und damit auch unerfahreneren Nutzer*innen den Einstieg erleichtert, da innerhalb kurzer Zeit Modelle mit passablen Ergebnissen erstellt werden können.

14 *Quelle: https://wml-dashboard-artifacts-stage1.s3-api.us-geo.objectstorage.softlayer.net/img-service-broker/ml-screen-1.png*

Bei einer Nutzung in der IBM-Cloud werden für Watson Machine Learning drei Kostenmodelle angeboten: Während »Lite« kostenfrei ein begrenztes Modell- und Aufrufkontingent bereitstellt, werden unter »Standard« und »Professional« unterschiedliche kostenpflichtige Bereitstellungsmodelle angeboten. Bei der ersten Anmeldung ist ein Startguthaben von 200 US-Dollar verfügbar[15], das einen Monat gültig ist.

3.5 Microsoft Azure

Microsoft Azure ist einer von Microsofts Cloud-Computing-Diensten. Er wurde 2010 unter dem Namen Windows Azure eingeführt und bietet Funktionalitäten zur Entwicklung, Erprobung, Bereitstellung und Verwaltung von Anwendungen und Diensten. Zu den Dienstleistungsmodellen gehören neben Software as a Service und Platform as a Service auch Infrastructure as a Service. Azure erzielt seit einiger Zeit hohe jährliche Wachstumsraten, der Umsatzunterschied zum Branchenprimus AWS wird kleiner.

In unserer Studie steht Azure Machine Learning im Fokus. Dieser Dienst wurde im Juli 2014 im Rahmen eines Public Previews öffentlich eingeführt und ist nach eigenen Aussagen ein »Machine-Learning-Dienst für Unternehmen zur schnelleren Erstellung und Bereitstellung von Modellen«[16]. Der Dienst richtet sich an Entwickler*innen und Data Scientists und verspricht die Verringerung der Time to Market, eine Optimierung der Zusammenarbeit, eine Unterstützung des gesamten ML-Lebenszyklus, Interpretierbarkeit von ML-Modellen und eine Unterstützung von Open-Source-Frameworks wie Pytorch oder Python.

Unterschiedliches Vorwissen der Benutzer*innen wird durch ein dreistufiges Konzept adressiert. Erstens gibt es ein AutoML-Modul, das keinen Code benötigt und Features sowie Modelle und Hyperparameter automatisiert erstellt bzw. optimiert. Zweitens gibt es den No-Code-Designer, der es ermöglicht, Machine-Learning-Pipelines per Drag-and-drop zu erstellen. Drittens erlaubt der Code-First-Ansatz die Benutzung integrierter Jupyter-Notebooks. Über die Ressource Azure Machine Learning Studio werden Nutzer*innen alle Tools zum Erstellen, Trainieren und Deployment von Machine-Learning-Modellen zur Verfügung gestellt.

15 *https://www.ibm.com/de-de/cloud/free*

16 *https://azure.microsoft.com/de-de/services/machine-learning/*

Der Dienst verfolgt ein nutzungsbasiertes Abrechnungskonzept, man startet mit einem temporären Guthaben von 200 US-Dollar[17]. Es gibt zwei Preisstufen, Basic und Enterprise. Sowohl das AutoML Modul als auch der No-Code-Designer sind nur in der höherpreisigen Enterprise Edition verfügbar. Ein detaillierter Vergleich der Funktionalitäten findet sich auf der Azure-Website[18].

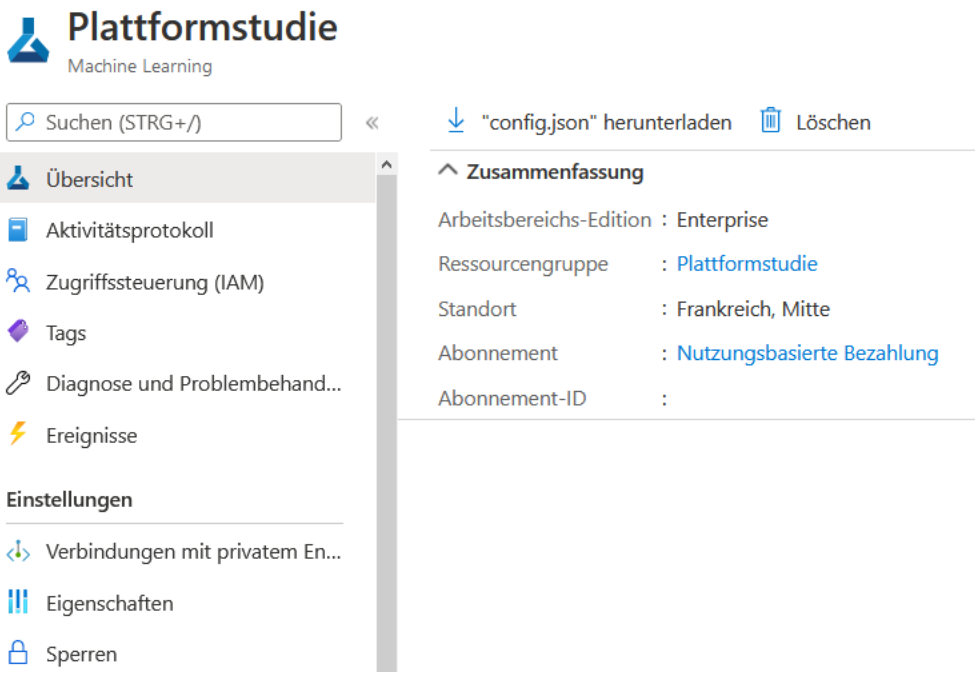


Abbildung 8: Ausschnitt aus dem Azure-Portal, der eine Übersicht der von uns erstellten Machine-Learning-Instanz »Plattformstudie« zeigt[19].

3.6 Weitere Anbieter

Neben den großen vier KI-Cloudplattformen existieren auch Alternativen. Diese weisen in der Regel zwar nicht vergleichbar breite Nutzungspotenziale auf, etablieren sich aber als leistungsfähige Nischenprodukte, die in klar abgesteckten Use Cases gute Ergebnisse erzielen können.

17 *https://azure.microsoft.com/de-de/free/*

18 *https://azure.microsoft.com/de-de/pricing/details/machine-learning/*

19 *Quelle: Microsoft Azure Portal.*

Open-Source-Alternative – H2O

H2O (eine Software des Unternehmens H2O.ai, ehemals 0xdata) vertritt die Firmenphilosophie »demokratisiere KI für alle« und zählt unter den Open-Source-KI-Cloudplattformen zu den visionärsten [9]. Ebenso bietet es eine vergleichsweise breite Kompatibilität und Integration zu gängigen Softwarelösungen aus dem Big-Data-Kontext sowie Schnittstellen zu den meisten etablierten ML-Programmiersprachen (siehe Abbildung 9).

Abbildung 9: H2O im Kontext anderer Softwarelösungen und -plattformen. Integrierte Softwarelösungen und Technologie, die mit oder aus H2O nutzbar sind[20].

Auf technischer Ebene basiert die H2O-Plattform u. a. auf dem Hadoop Distributed File System und unterstützt dabei auch in-memory, verteiltes und skalierbares Machine Learning. Die Auswahl an Algorithmen ist groß und umfasst zahlreiche Methoden des (un-) überwachten Lernens, sodass Problemstellungen mit strukturierten Daten oder aus dem Bereich der Bilderkennung sowie State-of-the-art-Methoden (Embeddings) für komplexere Sprachverarbeitungs-Tasks (NLP) adäquat lösbar sind.

Zusätzlich verfügt die H2O-Plattform über ein (u. a. über Web-GUI nutzbares) AutoML-Modul und ist daher auch zugänglich für Zielgruppen mit nur grundlegender ML-Expertise. Eine tiefreichende Dokumentation über alle kostenlosen Dienste und Funktionen ist einsehbar[21]. Daneben bieten zahlreiche Tutorials zu verschiedenen Use Cases einen einfachen Einstieg in H2O.

20 *Quelle: https://www.h2o.ai/products/h2o/*

21 *http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html*

Neben den frei verfügbaren Open-Source-Lösungen bietet H2O auch kostenpflichtige maßgeschneiderte Unterstützung für Unternehmen. Diese reicht von H2O Wave (einem Framework zu Erstellung von KI-unterstützten Web-Apps) über Enterprise Puddle (Nutzung von H2O-Software über andere Cloudplattformen, wie AWS, GoogleCloud, Microsoft Azure oder IBM Cloud) bis hin zu einem eigenen AutoML-Modul: Driverless AI. Kostenmodelle der einzelnen Angebote sind nicht öffentlich einsehbar. Für konkrete Angebote muss man sich an den Vertrieb wenden.

Deutsche Anbieter

Auch deutsche Anbieter agieren mit vielfältigen Ideen und Produkten am MLaaS-Markt. Keines bietet jedoch eine vergleichbare Breite an Funktionen, wie die in dieser Studie untersuchten US-amerikanischen Cloud-Dienstleister. Vielmehr handelt es sich bei deutschen Anbietern um oftmals hoch spezialisierte Nischenanbieter, die in ihren jeweiligen Bereichen durchaus mit »Big Tech« konkurrieren können. Allerdings ist es nicht möglich, eine vollständige oder repräsentative Liste solcher Dienste anzuführen, weswegen wir an dieser Stelle darauf verzichten, einzelne Unternehmen hervorzuheben.

3.7 Rahmenbedingungen bei der Nutzung von MLaaS-Lösungen

Dieser Abschnitt geht in prägnanter Weise auf Rahmenbedingungen und Risikofaktoren sowie zu ergreifende Maßnahmen bei der Benutzung von Cloud-Diensten ein. Detailliertere Betrachtungen insbesondere rechtlicher und informationssicherheitstechnischer Aspekte finden sich in den zahlreichen Referenzen dieses Abschnitts.

Mit der DSGVO ist eine EU-Verordnung in Kraft getreten, die die Verarbeitung personenbezogener Daten regelt. Im Zusammenhang mit Cloud Computing ist die Weitergabe solcher Daten durch Unternehmen an Cloud-Dienstleister typisch. Bei dieser Weitergabe kann es sich um eine Auftragsdatenverarbeitung handeln, wobei »die datenschutzrechtliche Verantwortlichkeit uneingeschränkt beim Cloud-Nutzer als Auftraggeber« [3] verbleibt. Das bedeutet, dass Cloud-Nutzer*innen jederzeit die Kontrolle über die betreffenden Daten haben müssen, um geltenden Datenschutzbestimmungen zu entsprechen. Nach Art. 13 DSGVO muss unter anderem sichergestellt sein, dass die Daten jederzeit gelöscht werden können oder deren Verarbeitung eingeschränkt werden kann. Zudem ist nach Art. 6 DSGVO die Weitergabe personenbezogener Daten an Dritte ohne Zustimmung der betroffenen Personen im Allgemeinen nicht zulässig [6]. Ein deutsches Unternehmen steht somit vor der Frage, wie die Daten, die bei ausländischen Cloud-Anbietern hochgeladen werden, gespeichert und verarbeitet werden. Dabei sind politische, technische und organisatorische Herausforderungen zu beachten.

Politische Rahmenbedingungen

Mit dem im März 2018 in Kraft getretenen CLOUD Act (Clarifying Lawful Overseas Use of Data Act) steht der DSGVO ein konkurrierendes US-amerikanisches Gesetz gegenüber. Der CLOUD Act regelt den Zugriff der US-Behörden auf Daten, die außerhalb der USA gespeichert werden [1] und ist die Folge eines jahrelangen Rechtsstreits zwischen der US-Regierung und Microsoft, bei dem die US-Regierung im Zuge einer Strafverfolgung versuchte, auf E-Mails, die auf Servern in Irland gespeichert waren, zuzugreifen[22] [23]. US-Behörden haben mit dem CLOUD Act ein Werkzeug, um auch auf Daten auf EU-Servern zuzugreifen, wenn der Besitzer der Daten US-Bürger *in ist oder das Unternehmen, dem die Server gehören, seinen Sitz in den USA hat [1, 11]. Die DSGVO versucht, den Widerspruch zwischen US-Ansprüchen und EU-Schutzrechten mit Art. 48 aufzuheben und gestattet die Weitergabe von Daten nur im Rahmen weiterer internationaler Abkommen. Allerdings müssen die Fälle detailliert geprüft werden [6, 11].

Technische Risikofaktoren

Ein weiteres Risiko stellt neben dem CLOUD Act die Sicherheit der Cloud im Allgemeinen dar. Zu möglichen Bedrohungen gehören [3, 13–15]:

- Cloud-Ausfälle und Datenverluste aufgrund technischer Störungen
- Sicherheitsvorfälle
- Attacken von externen Angreifern
- Entstehung einer Schatten-IT durch unkontrollierte Rechtevergabe für die Cloud-Anwendungen
- Identitätsdiebstahl bzw. Missbrauch von Accounts
- Verletzung geltender Vorgaben und Richtlinien (z. B. Datenschutzanforderungen)

Sicherheitsvorfälle können große Dimensionen annehmen, erheblichen Schaden verursachen und sind oft auf technische Probleme oder menschliches Versagen zurückzuführen. 2019 etwa kam es aufgrund einer nutzerseitigen Fehlkonfiguration bei einem Cloud-Anbieter zu einem Sicherheitsvorfall, wobei sensible Daten von Netflix, Ford und der TD Bank wie Benutzernamen und Passwörter öffentlich wurden [19][24] [25] [26].

22 *https://www.spiegel.de/netzwelt/netzpolitik/microsoft-rechtsstreit-erledigt-supreme-court-beruft-sich-auf-cloud-act-a-1203475.html*

23 *https://www.handelsblatt.com/unternehmen/it-medien/cloud-act-us-gericht-legt-streit-um-microsoft-mails-in-irland-zu-den-akten/21185490.html*

24 *https://www.hardwareluxx.de/index.php/news/allgemein/wirtschaft/50115-aws-s3-server-leakt-firmendaten-von-netflix-ford-td-bank-u-v-m.html*

25 *https://www.zdnet.de/88363915/aws-cloudserver-gibt-daten-von-fortune-100-unternehmen-preis/*

26 *https://divvycloud.com/leaky-s3-bucket-exposes-ford-netflix-tdbank-data/*

Organisatorische Risikofaktoren

Unternehmen sind verpflichtet, Aufsichtsbehörden nach einem Datenverstoß zu benachrichtigen, weswegen Verfehlungen schnell erkannt werden müssen. Dabei ist es wichtig zu wissen, welche und wie viele Daten betroffen sind und was die möglichen Auswirkungen des Verstoßes sind. Es sollte zudem ein Konzept zur Schadensbegrenzung vorliegen. Dies erfordert gute Betriebs- und Sicherheitskontrollen in Unternehmen für den Umgang mit sensiblen Daten [5].

Maßnahmen

Um den genannten Herausforderungen gerecht zu werden existiert eine Vielzahl an Maßnahmen. Schon jetzt hat die DSGVO einen Einfluss auf Unternehmen bei der Wahl der Cloud-Dienste und den jeweiligen Speicherort [13]. Zu den Vorkehrungen der Sicherheitsorganisation zählen [13]:

- Prüfung von Cloud-Zertifikaten
- Datenschutzaudits bei Cloud-Diensten (bei fehlender Zertifizierung)
- Cloud Policy für die Nutzung von Cloud-Lösungen und Zugangsgeräten
- Vorkehrungen für das Einhalten von Compliance
- Definieren von Personen, die Cloud-Dienste im Sinne eines Vendor-Managements steuern
- Prüfung der Transparenzberichte (Zugriffe staatlicher Stellen)
- Neue Security-Stellen
- Detailliertes Konzept für das Vendor-Management
- Joiner-Mover-Leaver-Prozess (Behandlung von Daten nach Wechsel oder Austritt von Mitarbeiter*innen)

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) empfiehlt grundsätzlich, bei der Wahl des Cloud-Dienstes zu überprüfen, welche Zugriffe durch den Cloud-Dienst oder Dritte erlaubt werden und welchem geltenden Recht bzw. welchen rechtlichen Rahmenbedingungen ein Vertrag unterliegt [3]. Um die Auswirkungen von technischem oder menschlichem Versagen bei der Cloud-Anwendung sowie bei Angriffen von außen zu minimieren, weisen der Bitkom-Leitfaden und das BSI auf die Verschlüsselung der Daten hin [2, 3]. Die Daten weisen dann keinen Personenbezug mehr auf und es wird somit auch der Datenschutz in Drittländern gewährleistet.

Verschlüsselung und Anonymisierung

Um die Kontrolle über die Daten zu behalten, dürfen zu keinem Zeitpunkt andere Parteien Zugriff darauf haben. Dabei bedarf es in drei Bereichen besonderer Vorkehrungen [10]:

- Auf Transportwegen
- Am Speicherort in der Cloud
- Am Speicherort bei den Benutzer*innen

Die Transportwege werden standardmäßig seitens der Anbieter verschlüsselt, wobei die angewendeten Verfahren im Einzelnen überprüft werden sollten. Die Verschlüsselung der Daten in der Cloud erhöht die Sicherheit, hat aber den Nachteil, dass das Schlüsselmanagement beim Anbieter liegt und er somit theoretisch in der Lage ist, die Daten wieder zu entschlüsseln [10]. Die großen Cloud-Anbieter bieten verschiedene Möglichkeiten, die Daten zu verschlüsseln und zu verwalten. Google bietet zum Beispiel neben der standardmäßigen Verschlüsselung, die innerhalb der Cloud stattfindet, die Option an, eigene Sicherheitsschlüssel zu verwenden, die über eine Schnittstelle bereitgestellt werden können[27]. Amazon ermöglicht es ebenfalls, eigene Schlüssel zu generieren und zu verwalten, allerdings innerhalb der Cloud-Umgebung[28]. Microsoft bietet mit der Azure Key Vault die Verwaltung und Steuerung des Zugriffs auf Sicherheitsschlüssel, wobei die Anwendungen und Microsoft nie auf die Schlüssel zugreifen können[29]. IBM gibt an, dass Schlüssel manuell verwaltet werden können, indem eigene Sicherheitsschlüssel beim Speichern der Daten bereitgestellt werden[30].

Die größte Sicherheit bieten Verfahren, mit denen die Daten vor dem Transport in die Cloud bereits lokal verschlüsselt werden und das Schlüsselmanagement somit bei den Benutzer*innen liegt. Zusatzsoftware wie Boxcryptor[31] oder Cryptomator[32] können bei einigen Cloud-Anbietern dafür eingesetzt werden [8, 10].

27 *https://cloud.google.com/security/encryption-at-rest?hl=de*

28 *https://aws.amazon.com/de/cloudhsm/*

29 *https://docs.microsoft.com/de-de/microsoft-365/compliance/office-365-encryption-in-the-microsoft-cloud-overview?view=o365-worldwide*

30 *https://cloud.ibm.com/docs/key-protect?topic=key-protect-integrate-cos&locale=de*

31 *https://www.boxcryptor.com/de/*

32 *https://cryptomator.org/de/*

Haltung der Cloud-Dienstleister

Die Cloud-Dienstleister sind sich der Herausforderungen in Bezug auf Datensicherheit und Datenschutz bewusst. Sie geben an, jede Anfrage der US-Regierung werde gründlich geprüft und ggf. öffentlich gemacht. Zudem werden teilweise Leitfäden für eine DSGVO-konforme Verarbeitung der Daten angeboten.[33] [34] [35]

Fazit

Die Verwendung von Cloud-Diensten kann Risiken in Bezug auf Datenschutz und Datensicherheit bergen. Cloud-Dienstleister bieten Unterstützung bei einer datenschutzkonformen und sicheren Verarbeitung in ihren Anwendungen in Form von Aktionsplänen, DSGVO-Leitfäden, Transparenz- und Verschlüsselungstechniken an, um Anwender*innen zu entlasten. Es besteht zudem die Möglichkeit, die Daten durch Zusatzsoftware noch vor der Cloud zu verschlüsseln, um eine maximale Unabhängigkeit von den Cloud-Diensten zu erlangen.

33 *https://docs.microsoft.com/de-de/microsoft-365/compliance/office-365-encryption-in-the-microsoft-cloud-overview?view=o365-worldwide*

34 *https://aws.amazon.com/de/compliance/cloud-act/*

35 *https://www.ibm.com/blogs/policy/cloud-act/*

4 ALLGEMEINER PLATTFORM-VERGLEICH

Die in dieser Studie betrachteten cloudbasierten KI-Plattformen bieten ein breites Spektrum an Diensten und Lösungen für verschiedenste Problemstellungen – ein möglichst umfassendes Auftreten scheint in diesem Marktsegment ein zentrales Bestreben der Anbieter zu sein. Aufgrund der angebotenen Fülle an Funktionen kann die Suche nach der passgenauen Lösung für ein Unternehmen eine ressourcenintensive Aufgabe sein. Noch bevor man allerdings auf der Ebene eines konkreten Dienstes vergleicht, gilt es anhand allgemeinerer Aspekte Gemeinsamkeiten und Unterschiede aufzuzeigen. Das durch die Plattformanbieter weit aufgespannte Feld muss anhand dezidierter Kriterien vergleichbar gemacht werden.

4.1 Kriterien

Aus Literatur [17] und einschlägiger Erfahrung wurde eine Liste von Kriterien erstellt, die einen Vergleich cloudbasierter KI-Plattformen ermöglichen soll. Dadurch werden die verschiedenen Eigenschaften der Plattformen möglichst greifbar gemacht. Während die Kriterien mancher Aspekte eine klar quantitative Ausprägung haben, also in konkreten Zahlen darstellbar sind, sind andere Aspekte ausschließlich durch Kriterien auf einer subjektiven, qualitativen Ebene erfassbar. Dieser Kritik sind sich die Autoren bewusst. Sie haben allerdings entschieden, dass eine möglichst ganzheitliche Darstellung der Plattformen im Vordergrund stehen soll. Vor diesem Hintergrund sollen Bewertungen entsprechender Dimensionen von Leserinnen und Lesern rezipiert werden.

Zur inhaltlichen Aufbereitung wurden die Kriterien für den allgemeinen Plattformvergleich in drei Kategorien eingeteilt: Funktionsumfang, technische und geschäftliche Kriterien. Dabei beschränken sich die technischen Kriterien auf Entwickler*innen, die geschäftlichen Kriterien beziehen sich auf strategische Entscheidungen durch das Management.

In der hier durchgeführten Bewertung der Plattformen (Abschnitt 4.2) geht allerdings nur eine kurierte Untermenge der kompletten Liste ein. Das geschieht zum einen, damit eine gewisse Übersicht gewahrt wird, zum anderen, weil diverse Kriterien subjektiv und individuell zu beur-

teilen sind. Dies trifft insbesondere auf die technischen Kriterien zu, weshalb diese nur in einem Fall in 4.2 behandelt werden. Dennoch wird die vollständige Liste zum Zweck der Hilfestellung für eigene Evaluationen bei der Auswahl der cloudbasierten KI-Plattform präsentiert.

4.1.1 Funktionsumfang

Anomalieerkennung
In der Anomalieerkennung geht es um die Identifikation einzelner Datenpunkte, die aufgrund auffälliger Merkmale aus der Masse der Daten herausstechen. Oftmals, aber nicht notwendigerweise handelt es sich bei solchen abnormalen Datenpunkten um Instanzen, die auf unerwünschte Weise vom Normalzustand abweichen. Es ist daher ratsam, sie zu erkennen (Betrugsfälle, medizinische Auffälligkeiten).

AutoML
Bei der Erstellung von ML-Modellen ist typischerweise ein iteratives Vorgehen, verbunden mit tiefem Fachwissen nötig. Daher werden in den letzten Jahren zunehmend automatisierte Verfahren marktreif, die Data Scientists bei der Modellauswahl und dem Fine Tuning unterstützen. In der Idealform übernimmt AutoML nahezu sämtliche Aufgaben eines Data Scientists in einem KI-Projekt, bisher jedoch in der Regel mit etwas schlechteren Ergebnissen.

Chatbot
Chatbots sind digitale Assistenten, die mit Menschen in natürlicher Sprache kommunizieren. Sie werden von Unternehmen gerne zur Bearbeitung von Kundenanfragen verwendet. Chatbots basieren nicht notwendigerweise auf maschinellem Lernen, jedoch hat ihre Entwicklung und Qualität durch die Verwendung von KI einen Schub erhalten.

Clustering
Clustering oder auch Clusteranalysen sind Verfahren zur Erkennung von Ähnlichkeiten in Datensätzen. Ähnliche Datenpunkte sollen dabei zu sogenannten Clustern zusammengefasst werden. Die Attribute bzw. deren Ausprägungen, die einzelne Datenpunkte ähnlich erscheinen lassen, werden dabei von geeigneten Algorithmen erkannt. Potenzielle Anwendungsfälle können die Identifikation von Kundengruppen im Marketing oder Auffinden von Communities in sozialen Netzwerken sein (Social Network Analysis).

Echtzeitanalyse von Videos
Hierbei geht es um die Erkennung und Nachverfolgung von Objekten in Videodateien oder -streams. Durch die KI-basierte Analyse von Videos kann etwa ein Videoarchiv erstellt werden, das das spätere Auffinden von Informationen aus Videos vereinfacht. Denkbar ist ein Einsatz solcher Echtzeitanalysen, beispielsweise in der Produktion bei der Erkennung von defekten Bauteilen.

Erkennung deutscher Sprache (Speech to Text)
Dieses Kriterium beschreibt die Erkennung deutscher Sprache beim Umwandeln von natürlicher Sprache zu Text.

Explainable AI
Explainable AI vereint Ansätze und Verfahren, die für mehr Transparenz, Nachvollziehbarkeit und Erklärbarkeit bei maschinellen Lernverfahren sorgen. Das Ziel ist es dabei, den Trainingsprozess und/oder die Vorhersage einzelner Instanzen verständlich zu machen. Die übergeordnete Absicht ist größere Akzeptanz in breiten gesellschaftlichen Gruppen, die Einhaltung gesetzlicher Vorgaben und die sichere Benutzung von ML-Modellen auch durch Nicht-Expert*innen.

Named Entity Recognition (NER)
Diese Verfahren identifizieren und lokalisieren Entitäten, also zum Beispiel Namen, Orte oder Datumsangaben, in unstrukturierten Texten. Die erkannten Entitäten können anschließend strukturiert, etwa in Form einer Tabelle, aufbereitet werden und somit Arbeitsabläufe beschleunigen.

Objekterkennung in Bildern
In diesem Teilbereich der Computer Vision geht es um die Erkennung von Menschen, Gebäuden, Fahrzeugen etc. in digitalen Bildern. Typischerweise werden die erkannten Objekte durch ein Rechteck eingerahmt und die zugeschriebene Klasse hinzugefügt. Automatische Objekterkennung kann Anwender*innen dabei helfen, den Inhalt von Bildern ohne menschliches Zutun zu erfassen.

Recommender Systems
Empfehlungsdienste beabsichtigen, möglichst optimale Empfehlungen an Benutzer*innen auszusprechen. Genauer geht es um eine quantifizierbare Vorhersage des Interesses von Benutzer*innen an weiteren Objekten. Bekannt ist dieses Vorgehen zum Beispiel aus Video-Streamingdiensten, die, basierend auf in der Vergangenheit konsumierten Videos, Nutzer*innen spezielle Empfehlungen unterbreiten.

Sentiment Analysis
Verfahren der Stimmungserkennung werten Texte oder Audiomaterial aus mit dem Ziel, Stimmungen zu identifizieren und ggf. zu quantifizieren. Ein typischer Anwendungsfall ist das Klassifizieren von Bewertungen im E-Commerce als positiv, neutral oder negativ.

Texterkennung in Bildern
Texterkennung in Bildern, auch Optical Character Recognition (OCR) genannt, ist eine sehr häufig auftretende Aufgabe im Unternehmensalltag. Typischerweise erfordern Scans, aber auch andere digitale, als Bilder auftretende Unterlagen, die Überführung in bearbeitbaren oder durchsuchbaren Text. Die resultierenden Textinformationen können anschließend mit Methoden des Text Minings weiterverarbeitet werden. So lässt sich OCR zum Beispiel gut mit NER verbinden.

Übersetzung
Dieses Kriterium beschreibt Funktionen zur automatischen Übersetzung von Texten.

4.1.2 Weitere technische Kriterien

Skalierbarkeit der Hardwareressourcen
Eine Skalierbarkeit der Hardware seitens der MLaaS-Lösung ist nötig, wenn in einem KI-Projekt mehr Datenpunkte oder mehr eingehende Merkmale erforderlich werden. Auch wenn sich die Aufgabe verschiebt oder erweitert und dann mehr Rechenleistung benötigt, muss der Dienst skalierbar sein, etwa wenn Machine Learning auch für unstrukturierte Daten statt nur für strukturierte Daten benutzt werden soll oder wenn sich die Anzahl der zu verarbeitenden Predict-Aufrufe erhöht. Die Frage dabei ist, wie einfach, schnell und kostengünstig eine solche Skalierbarkeit durchführbar ist.

Angemessener Schutz gegen Bedrohungen der Informationssicherheit
Informationssicherheitsbedrohungen können wirtschaftliche Schäden, Reputations- und Vertrauensverlust oder rechtliche Auseinandersetzungen nach sich ziehen. Daher ist es von äußerster Wichtigkeit, dass Cloud-Dienste sich und damit ihre Kund*innen gegen solche Bedrohungen schützen. Gleichzeitig ist es selbst für IT-affine Mitarbeiterinnen und Mitarbeiter schwer zu beurteilen, wie effektiv einzelne Strategien sind. Zu den Maßnahmen gehören Zugangskontrollen, das Updaten verwendeter Software, Erstellen von Sicherungskopien, Benutzung von Firewalls und Antiviren-Software, Verschlüsselung sensibler Daten, Protokollierung und vieles mehr. Aus Sicht der Kund*innen von MLaaS-Lösungen ist insbesondere der Schutz von personenbezogenen Daten, Kundendaten und geistigem Eigentum wichtig.

Einfache und aufwandsarme Zusammenarbeit mit anderen IT-Lösungen
In der Regel wird der Cloud-Dienst nicht isoliert, sondern im direkten oder indirekten Zusammenspiel mit anderen IT-Lösungen eingesetzt. Ein häufig auftretender Fall ist der, dass Daten aus einem Bestandssystem extrahiert und in den Cloud-Dienst hochgeladen werden müssen. Gibt es für derartige Aufgaben Standard-Schnittstellen? Wie einfach funktioniert die Zusammenarbeit mit anderen IT-Lösungen generell?

Open-Source-Einbindung
Open-Source-Technologien können die Beschleunigung von zielführenden KI-Lösungen im Produktionsumfeld erheblich erleichtern. KI-Plattformen unterstützen idealerweise eine flexible Einbindung dieser Technologien.

Updatezyklen
Prüfung der Release-Strategie der Plattformen. Einerseits ist dies ein Indikator, welchen Schwerpunkt neue Funktionalitäten bei den Plattformen haben. Andererseits ist dies ebenfalls für die Sicherheit der Plattform wichtig.

4.1.3 Geschäftliche Kriterien

Einfluss auf Datenhaltung
Für viele Unternehmen stellt die Datenhaltung in der Cloud, insbesondere personenbezogener Daten, eine große, wenn nicht unüberwindbare Hürde dar. Während es für manche deutsche Unternehmen ausreicht, ihre Daten in Deutschland oder Europa und nicht im außereuropäischen Ausland zu speichern, haben andere weitergehende Sicherheitsbedürfnisse. So stellt für sie der CLOUD Act einen inakzeptablen Eingriff in ihre Datensouveränität dar. Dieses Kriterium bewertet also, ob Nutzer*innen der betreffenden Plattformen den Standort ihrer Daten wählen können.

Preistransparenz und -sicherheit
Dieses Kriterium umfasst Fragen nach der Transparenz der anfallenden Kosten bzw. der Kostenstruktur, der Möglichkeit ihrer Berechnung im Vorhinein und der Möglichkeit des Setzens einer Preisgrenze.

Support-Umfang
Gerade Nicht-Expert*innen sind in größerem Maße abhängig von schnellem, zielgerichtetem und evtl. deutschsprachigem Support. Wie sehen hier die Leistungsversprechen der vier Dienste aus?

Unterstützung von IT-Governance-Anforderungen
Unter IT-Governance versteht man generell Konzepte, die dafür sorgen, dass IT die Unternehmensziele sowie -strategie unterstützt. Zu Methoden der IT-Governance gehören technische Maßnahmen wie rollenbasierte Autorisierungskonzepte, Löschen von Benutzerkonten und Daten nach Beendigung des Geschäftsverhältnisses oder das Vorhandensein von Geheimhaltungsklauseln ggf. mit Konventionalstrafe.

Time to Market
Im Unternehmenskontext ist die Zeit vom Beginn eines Projekts bis hin zum Produktivgehen ein entscheidender Faktor. Time to Market beschreibt im Kontext dieser Studie, in welcher Größenordnung diese Zeitspanne liegt, bis eine KI-Anwendung alle Arbeitsprozesse durchlaufen hat und aktiv im/vom Unternehmen genutzt werden kann.

Total Cost of Ownership (TCO)
TCO, oder die Gesamtkosten des Betriebs, sind definiert als die Summe aller anfallenden Kosten von Investitionsgütern. Die Betrachtung geht insbesondere auf Kosten ein, die über den Einkaufspreis hinausgehen. Da die meisten MLaaS-Anbieter über nutzungsbasierte Preismodelle verfügen, ist kein eigentlicher Einkaufspreis zu entrichten. Vielmehr sind in unserer Betrachtung TCO solche Kosten, die durch Nutzung der Dienste entstehen, also Nutzungs- bzw. Lizenzgebühren, Supportkosten, Mietkosten von Hardware etc.

Return on Investment
Der Return on Investment, kurz ROI, gibt vereinfacht gesagt an, ob sich die Investition in MLaaS-Lösungen bezahlt macht. Man setzt die durch Nutzung der MLaaS-Lösungen erzielten Gewinne (in unserem Fall insbesondere durch eingesparte Kosten) zum eingesetzten Kapital, d. h. den zu entrichtenden Nutzungsgebühren, ins Verhältnis.

Wertschöpfung im Unternehmen
Beim Einsatz neuer Technologien, so auch von Machine Learning durch beispielsweise MLaaS-Lösungen, stellt sich die Frage, ob dadurch Produkte, Dienstleistungen oder Einsichten produziert, angeboten oder erlangt werden können, die einen höheren Geldwert als ihre Pendants ohne Einsatz dieser Technologie besitzen. Wertschöpfung ist somit das Ziel produktiver Tätigkeit und für jedes Unternehmen relevant.

Problemorientierung
Problemstellungen, denen KMUs häufig gegenüberstehen, erfordern sehr spezifische, angepasste und einzelfallorientierte Vorgehen. Zusätzlich liegen meist eher zu wenig als zu viele Daten vor. Dieses Kriterium soll beschreiben, ob trotz widriger Verhältnisse ein Problem zufriedenstellend gelöst werden kann. Dabei muss zufriedenstellend weiter unterteilt werden.

Reife, Zuverlässigkeit und Fehlertoleranz
Der MLaaS-Markt verändert sich rasant. Features werden abgeschaltet, verändert oder hinzugeschaltet. Vor diesem Hintergrund stellt sich die Frage, wie ausgereift, zuverlässig und fehlertolerant die MLaaS-Lösung ist. Kann man zum Beispiel die Anwendung nach einem Ausfall ohne Datenverlust neu starten?

Konformität mit gesetzlichen und regulatorischen Anforderungen
Gerade für Unternehmen aus stark regulierten Branchen wie der Finanzbranche ist die Einhaltung gesetzlicher Vorgaben entscheidend. Ist die rechtskonforme Nutzung der MLaaS-Lösungen gegeben und falls ja, bieten die Dienstleister vielleicht sogar branchen- und länderspezifische Konzepte, die es Unternehmen erleichtern, mit ihrer ML-Anwendung in die Cloud zu gehen? Anderhesrum ist es wichtig zu betrachten, welchen gesetzlichen Anforderungen der Cloud-Dienst selbst unterworfen ist und wie diese im Einklang mit hiesigen Regularien stehen.

Möglichkeit des Plattformwechsels/Projektumzugs (Vendor Lock-in)
Ist es möglich, KI-Modelle, Code bzw. alle zum Ausführen nötigen Komponenten einer KI-Anwendung in einem solchen Format zu exportieren, dass eine beliebige Weiterverarbeitung, sei sie proprietär, sei sie auf einer anderen Plattform, möglich ist? In diesem Zusammenhang wird oft von einem sogenannten Vendor-Lock-in gesprochen, also der Tatsache, dass Kund*innen von MLaaS-Lösungen bzw. von deren Eigenheiten in einem Maße abhängig sind oder werden, dass ein Wechsel zu einem konkurrierenden Dienst nicht oder nur schwer möglich ist.

4.2 Vergleichender Überblick

Die folgenden Tabellen enthalten eine Teilmenge obiger Kriterien und sollen einen Überblick über die allgemeinen Unterschiede der KI-Plattformen geben. Die Merkmale, die den Funktionsumfang betreffen, bewerten wir genau dann positiv, wenn sie ohne das Schreiben oder Ausführen von Code verfügbar sind.

Technische Kriterien	Amazon Web Services	Google Cloud Platform	IBM Watson	Microsoft Azure
Anomalieerkennung	☑	☐	☑	☑
AutoML	☑	☑	☑	☑
Chatbot	☑	☑	☑	☑
Clustering	☐	☐	☑	☑
Echtzeitanalyse von Videos	☑	☑	Mit Einschränkungen[36]	☑
Explainable AI	Mit Einschränkungen[37]	☑	☑	☑
Named Entity Recognition	☑	☑	☑	☑
Objekterkennung in Bildern	☑	☑	☑	☑
Recommender Systems	☑	☑	☐	☐
Sentiment Analysis	☑	☑	☑	☑
Skalierbarkeit der Hardwareressourcen (bereitgestellte Schnittstelle)	☑	☑	☑	☑
Texterkennung in Bildern	☑	☑	☑	☑
Transkription deutscher Sprache (Speech to Text)	☑	☑	☑	☑
Übersetzung	☑	☑	☑	☑

Tabelle 5: Überblick anhand ausgewählter technischer Kriterien.

36 Laut Beschreibung nur bzw. vor allem für forensische Zwecke verfügbar: https://www.ibm.com/watson/media/video-content-analysis

37 Teilweise als externer Dienst verfügbar.

Tabelle 6: Überblick anhand ausgewählter geschäftlicher Kriterien.

Geschäftliche Kriterien	Amazon Web Services	Google Cloud Platform	IBM Watson	Microsoft Azure
Einfluss auf Datenhaltung	☑ (Europa)	☑ (Europa)	☑ (Europa)	☑ (Europa)
Preistransparenz und -sicherheit	Budgetbenachrichtigung: vorhanden Preisrechner: vorhanden[38]	Budgetbenachrichtigung: vorhanden Preisrechner: vorhanden[39]	Budgetbenachrichtigung: vorhanden Preisrechner: vorhanden[40]	Budgetbenachrichtigung: vorhanden Preisrechner: vorhanden[41]
Support-Umfang (fastest reponse time)	3 Supportstufen (Entwickler, Business, Enterprise)[42] (< 15 Minuten)	4 Supportstufen (Basic, Entwicklung, Produktion, Premiumsupport)[43] (< 15 Minuten)	4 Supportstufen (Gratis, Basic, Advanced, Premium)[44] (< 1 Stunde)	4 Supportstufen (Basic, Entwickler, Standard, Professional Direct)[45] (< 1 Stunde)
Time to Market (direkte Bereitstellung möglich?)	☑ [46]	☑	☑	☑
Unterstützung von IT-Governance-Anforderungen (Rollenkonzept)	☑ [47]	☑ [48]	☑ [49]	☑ [50]

38 *https://calculator.aws/#/*
39 *https://cloud.google.com/products/calculator?hl=de*
40 *https://cloud.ibm.com/estimator/review*
41 *https://azure.microsoft.com/de-de/pricing/details/machine-learning/*
42 *https://aws.amazon.com/de/premiumsupport/plans/*
43 *https://cloud.google.com/support#support-plans*
44 *https://www.ibm.com/de-de/cloud/support*
45 *https://azure.microsoft.com/de-de/support/plans/*
46 *https://aws.amazon.com/de/sagemaker/autopilot/*
47 *https://aws.amazon.com/de/iam/details/manage-roles/*
48 *https://cloud.google.com/solutions/best-practices-for-iam-and-billing-in-higher-education?hl=de*
49 *https://cloud.ibm.com/docs/Db2onCloud?topic=Db2onCloud-iam*
50 *https://docs.microsoft.com/de-de/azure/role-based-access-control/overview*

5 USE-CASE-SPEZIFISCHER PLATTFORMVERGLEICH

Noch wichtiger als ein allgemeiner Vergleich der cloudbasierten KI-Plattformen ist ein Vergleich auf Ebene der Teildienste. Eine Evaluation und ein Vergleich aller verfügbaren Dienste würden allerdings aufgrund der hohen Anzahl verfügbarer Module den Rahmen dieser Studie sprengen. Daher sollen exemplarisch vier stellvertretende KI-Use-Cases durch den jeweils passenden Dienst gelöst werden. Die KI-Use-Cases stammen aus den größten in der Wirtschaft gängigen Anwendungsgebieten des Machine Learnings [16]:

1. Strukturierte Daten: Use Cases, die Datensätze liefern, die in tabellarischer Form zur Verfügung stehen
2. Sprachverarbeitung (Natural Language Processing): Use Cases, die mit der Verarbeitung natürlicher Sprache zu tun haben und damit Datensätze liefern, die z. B. aus Worten oder Sätzen bestehen
3. Bilderkennung: Use Cases, die Datensätze in Form von Bildern liefern.
4. Zeitreihenanalyse: Use Cases, die Datensätze liefern, die eine zeitlich geordnete Struktur aufweisen

Zwar lassen sich Fragestellungen einer Kategorie auch als Fragestellungen anderer Kategorien auffassen bzw. transformieren (z. B. werden Daten einer Zeitreihe auch in tabellarischer Form festgehalten und sind damit strukturierte Daten), dennoch soll hier eine Betrachtung aus Nutzerperspektive erfolgen.

5.1 Betrachtungsrahmen

Zu jedem Use Case gehört ein Datensatz, anhand dessen ein Problem mit entsprechenden KI-Methoden gelöst werden soll. Es ist zu betonen, dass die ausgewählten Datensätze wie auch die vorgestellten Szenarien illustrativen Zwecken dienen. Die in diesem Kapitel beschriebenen Methoden und Ergebnisse können bei gleichem grundlegendem Charakter vollständig auf weitere Problemstellungen übertragen werden.

Um eine möglichst hohe Realitätsnähe zu gewährleisten, muss der Datensatz gewissen Anforderungen entsprechen:

1. Der Datensatz wurde tatächlich erhoben. Es handelt sich nicht um synthetische Daten.
2. Der Datensatz muss frei verfügbar sein.
3. Die Anzahl der Datenpunkte beschränkt sich auf KMU-übliche Größenordnungen.
4. Die Datenqualität entspricht branchenüblichen Standards.
5. Die Sprache soll Deutsch sein.

Anforderungen 3, 4 und 5 können durch verschiedene Methoden erfüllt werden: Die dritte Anforderung wird durch Zufallsziehung einer Stichprobe (Sampling) sichergestellt. Aus einem beliebig großen Datensatz werden so viele Punkte gezogen, wie in der Branche üblich (basierend auf subjektiver Erfahrung). Dabei wird die unterliegende Verteilung des Datensatzes erhalten. So können branchenspezifische Kardinalitäten simuliert werden.

Frei verfügbare Datensätze haben außerdem im Vergleich zum Branchenstandard eine überdurchschnittliche Qualität bzgl. der Vollständigkeit der Datenspalten. Anforderung 5, die hauptsächlich den Text-Use-Case betrifft, wurde durch Wahl eines Beispieldatensatzes mit deutschen Texten berücksichtigt.

5.2 Beschreibung der Use Cases

Binäre Klassifikation auf tabellarischen Daten

Der in diesem Use Case betrachtete Datensatz ist thematisch in der Hotelbranche angesiedelt. Er wurde gemäß Abschnitt 5.1 leicht verändert und stammt von Kaggle[51]. Der Datensatz besteht aus 14.295 vergangenen Hotelreservierungen und insgesamt 28 numerischen sowie kategorischen Variablen (z. B. is_repeated_guest, *children*, ***country***). Die Zielvariable heißt is_canceled. Ziel ist es, ein KI-Modell zu trainieren, das aus den 28 buchungsspezifischen Variablen is_canceled vorhersagen kann. Es handelt sich also um eine zweiwertige überwachte Klassifikationsaufgabe. Solche zweiwertigen, auf strukturierten Daten basierende Problemstellungen kommen sinnverwandt in vielen Unternehmen und Branchen vor, zum Beispiel in der Betrugserkennung in der Versicherungsbranche oder bei Kreditvergaben durch Banken.

51 *kaggle.com/jessemostipak/hotel-booking-demand*

is_canceled	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
1	2015	September	37	10
0	2016	May	19	2
0	2017	February	8	22
1	2017	August	31	3
0	2017	July	27	4
	2017	May		16
1	2017		14	7
0	2017	March	13	30
0	2017	May	18	2
1	2016	June	23	1
1	2017	June	23	4
0	2016	January	5	29
1	2017	April		27
0	2017		20	17
0	2015	December	52	24
1	2017	March	12	20
0	2015	October		1
	2017	August	33	14
1	2016	April	18	28
0	2016	May	23	30

Tabelle 7: Ausschnitt des Stadthotel-Datensatzes.

Auf Geschäftsebene könnte diese Aufgabe dem Bestreben von Hotelbesitzer*innen entsprechen, auf Basis historischer Daten zukünftige Stornierungen frühzeitig zu erkennen. Somit ließen sich Kosten sparen und die Zimmerbelegung effizienter gestalten.

Textklassifikation

Für das Problem der Textklassifikation existieren für die deutsche Sprache nur sehr wenige frei zugängliche Datensätze, die den in 5.1 genannten Anforderungen genügen.

Der für diesen Use Case aufgrund seiner Popularität ausgewählte und beispielhaft verwendete Datensatz befasst sich mit der Einordnung von Nachrichtentexten in verschiedene Sparten und steht auf GitHub zur freien Verfügung[52]. Der hier erzeugte und für die Studie verwendete

52 https://github.com/tblock/10kGNAD

Datensatz wurde ursprünglich dem One Million Posts Corpus [18] entnommen, der unter einer Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[53] veröffentlicht wurde.

Jede Datenreihe besteht aus einem Text, der den eigentlichen Artikel beinhaltet, sowie einer der in Tabelle 8 aufgelisteten Zielkategorien, die während des Trainings und der Evaluation als Zielvariable dient. Bei diesem 9-Klassen-Problem wird jedem Text genau eine der Kategorien zugewiesen.

Tabelle 8: Anzahl der Texte pro Kategorie.

Kategorie	Anzahl Texte
Web	1678
Panorama	1677
International	1511
Wirtschaft	1411
Sport	1201
Inland	1014
Etat	668
Wissenschaft	573
Kultur	539

Es wurden keine Veränderungen an den Texten oder sonstigen Daten vorgenommen. Es erfolgte lediglich die Aufbereitung für die verschiedenen Plattformen inkl. eigener (stratifizierter) Aufteilungen für Training (80 Prozent), Validierung (10 Prozent) und Test (10 Prozent), die für alle Plattformen (soweit möglich) einheitlich gewählt wurde.

Die Lösung des genannten Klassifikationsproblems für Texte könnte eine Nachrichtenagentur dabei unterstützen, große Mengen neuer bzw. noch nicht vollständig kategorisierter Artikel automatisch einzuordnen, was gegenüber manueller Zuordnung eine enorme Zeitersparnis bedeuten würde. Generell steht der gewählte Datensatz allerdings lediglich exemplarisch für sämtliche Aufgaben, bei denen Textdokumente in bestimmte Kategorien eingeteilt werden müssen, sei es die in jedem Unternehmen täglich anfallenden Aufgaben in Form von E-Mails, Briefen und gescannten Dokumenten oder die automatische Einordung spezieller Fachtexte wie Gutachten oder Berichte aus der Versicherungsbranche.

53 *http://creativecommons.org/licenses/by-nc-sa/4.0/*

Ursprünglich war für den vorliegenden Text-Use-Case ein Problem aus dem Bereich der Named Entity Recognition (NER) geplant anstelle einer Textklassifikation. Erste Versuche in diesem Bereich zeigten jedoch bereits zahlreiche Schwierigkeiten auf. Während Microsoft Azure derzeit überhaupt keine Easy-to-use-Funktionen zum Training eigener NER-Modelle anbietet, verlangen die anderen Plattformen alle ein eigenes, nicht standardisiertes Eingangsformat der Trainings- und Testdaten, das einiges an manueller Vorverarbeitung und Konvertierung erfordert. Nach erfolgreichem Einlesen der Daten verweigern die Tools anschließend zum Teil den Start eines Trainings, wenn die im Datensatz enthaltenen Entitäten nicht in ausreichend großer und ausgewogener Zahl repräsentiert sind, was bei dem gewählten Datensatz nicht der Fall war. Aus diesen Gründen wurde statt der geplanten NER eine Textklassifikationsaufgabe gewählt, mit der die Dienste gegenwärtig einfacher und standardisierter anzuwenden sind.

Bildklassifikation

Der in diesem Use Case betrachtete Datensatz wurde in [12] referenziert und ist auf GitHub[54] zu finden. Der Datensatz enthält Bilder von magnetischen Kacheln und wurde für eine sogenannte »Pixelwise Instance Segmentation« erzeugt. Hierbei handelt es sich um eine Objekterkennung, bei der nicht nur ein Begrenzungskasten (Bounding Box) um ein Objekt gezogen wird (hier ein Fehlertyp auf der Kachel), sondern bei der jeder einzelne Pixel, der zu einem Objekt gehört, als solcher markiert bzw. erkannt wird. Für unseren Anwendungsfall nutzen wir lediglich die Information des in dem Bild vorhandenen Fehlertyps und erzeugen ein Modell, das Bilder klassifiziert. Es handelt sich hierbei also um eine Single-Label-Klassifizierung. Es ist anzumerken, dass die Bilder nicht alle dieselbe Auflösung bzw. dasselbe Seitenverhältnis aufweisen, was von den Lösungswerkzeugen berücksichtigt werden muss.

Der Datensatz besteht aus 1344 Bildern und wird in sechs Klassen bzw. Fehlertypen aufgeteilt:

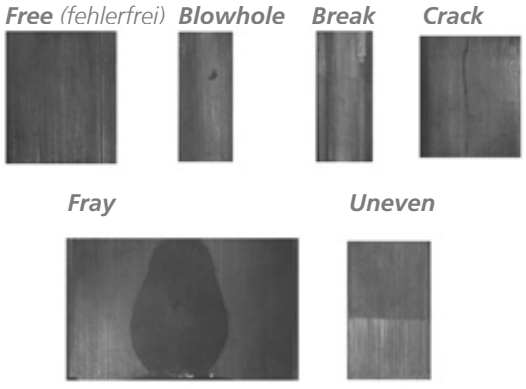


Abbildung 10: Auszüge aus dem Datensatz zur Bilderkennung[55].

54 *https://github.com/abin24/Magnetic-tile-defect-datasets.*

55 *Quelle: https://github.com/abin24/Magnetic-tile-defect-datasets.*

Auf Geschäftsebene könnte diese Aufgabe dem Bestreben eines produzierenden Unternehmens entsprechen, auf Basis von Bilddaten zukünftig die Qualitätsprüfung zu automatisieren. Somit ließen sich Kosten sparen und es könnte eine präzisere Erkennung von Ausschuss erzielt werden.

Zeitreihenanalyse

Für Unternehmen des produzierenden Gewerbes ist das Erkennen von Verschleiß bei Werkzeugmaschinen von großer Bedeutung. Die frühzeitige Vorhersage der Werkzeugabnutzung erlaubt es Unternehmen, die entsprechende Maschine rechtzeitig zu warten und abgenutzte Teile zu ersetzen, bevor es zu Produktionsausfällen kommt. So kann der Produktionsprozess durch die möglichst optimale Planung von Wartungsintervallen angepasst werden, um Stehzeiten zu minimieren.

Im vorgestellten Anwendungsfall wird der Produktionsprozess einer Fräsmaschine betrachtet. Dabei soll der Werkzeugverschleiß des Fräsers durch Anwendung intelligenter Verfahren des maschinellen Lernens möglichst optimal vorhergesagt werden. Bei dem verwendeten Datensatz handelt es sich um einen Datensatz[56], der von UC Berkeley zu diesem Zweck veröffentlicht wurde. In Abbildung 11 ist die Entwicklung des Werkzeugverschleißes in Abhängigkeit des Fräsdurchlaufs dargestellt. Der Datensatz besteht aus 16 Fräsdurchläufen, wobei je zwei Durchläufe die gleiche Parametrisierung in Form kategorischer Features aufweisen. Ein Durchlauf besteht dabei aus sechs verschiedenen kontinuierlichen Sensormesswerten mit jeweils 9000 Messpunkten, wobei es sich bei diesen Messwerten um den gemessenen elektrischen Strom und Vibrationsdaten handelt, die während des Fräsvorgangs anfallen.

56 *https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/*

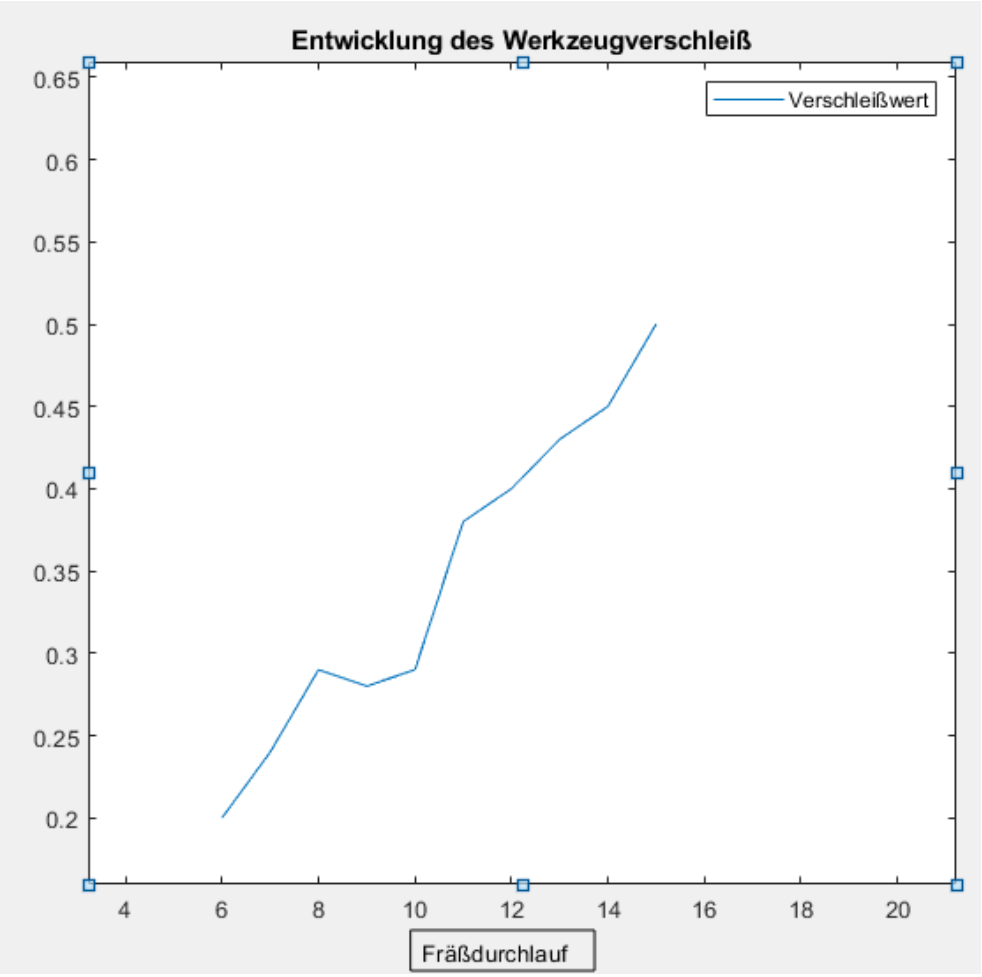


Abbildung 11: Entwicklung des Werkzeugverschleißes am Fräskopf in Abhängigkeit des Fräsdurchlaufs.

Der für mehrere Testläufe erfasste Verschleißwert am Fräskopf dient der Planung zukünftiger Wartungsarbeiten und stellt einen lohnenswerten Ansatzpunkt bei der Überwachung von Fräsprozessen dar. Da dieser Wert allerdings erst durch manuelle Analyse und Vermessung verfügbar ist, wird versucht, anhand der erfassten Sensordaten den Verschleißwert durch Anwendung eines Regressionsverfahrens vorherzusagen.

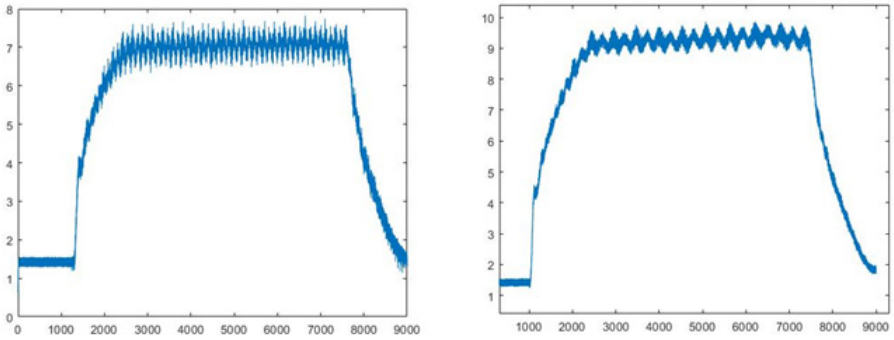


Abbildung 12: Vergleich der gemessenen Stromstärke bei auftretendem Verschleiß am Fräskopf.

In Abbildung 12 ist ein Beispiel für diese Sensordaten dargestellt. Dabei handelt es ich um die Stromstärke (gemessen in Ampere) über zwei unterschiedliche Fräsdurchläufe hinweg. Links ist die Stromstärke bei Verwendung eines neuen Fräskopfs abgebildet, rechts die Stromstärke nach sechs Fräsdurchläufen. Es ist erkennbar, dass sich der Wertebereich auf der y-Achse nach oben entwickelt, was unter Umständen ein Indiz für Verschleiß am Fräskopf ist.

5.3 Kriterien

Da sich die vier Anwendungsgebiete sehr voneinander unterscheiden, ist es nicht möglich, einen allgemeinen Kriteriensatz anzuwenden. Jeder Use Case erhält einen auf ihn angepassten Kriteriensatz. Im Folgenden wird die Problemkategorie des Use Cases genau spezifiziert.

5.3.1 Binäre Klassifikation auf tabellarischen Daten

Die Kriterien für diesen Use Case orientieren sich an den unterstellten Vorkenntnissen der Benutzer*innen. Anfänger*innen verfügen über Grundkenntnisse in ML und möchten basierend auf Vorschlägen vom System ggf. einzelne Parameter einstellen können. Folglich nutzen sie die AutoML-Komponente der Cloud-Dienste. Fortgeschrittene User kennen die wesentlichen Konzepte und Begriffe im ML-Lebenszyklus und möchten diese möglichst einfach einstellen können. Diese User benutzen den Designer- bzw. Visual-Coding-Ansatz. Aus dieser Zweiteilung ergeben sich folgende spezifische Kriterien:

AutoML

- Unterstützte Datentypen für Fileupload
  Unterstützte Datentypen für den Fileupload sollen aufgelistet werden.

- Richtige Erkennung und Parsing der Feature-Skalen
  Wie viele Features werden hinsichtlich ihres Datentyps richtig erkannt? Werden zum Beispiel Fließkommazahlen als Zahlen und nicht als String erkannt?

- Wahlmöglichkeit der Zielfunktion (Evaluationsmetrik)
  *Gibt an, ob die Benutzer*innen je nach Aufgabe die gewünschte Zielfunktion wählen können oder ob diese vorgegeben wird.*

- Ergebnisdarstellung und Evaluationsmetriken
  Auflistung aller zur Verfügung stehenden Visualisierungmöglichkeiten und Standard-Evaluationsmetriken (inkl. Werten), wie z. B. die ROC-Kurve.

- Erklärungsansätze/XAI
  Gibt an, ob die Ergebnisse aus Machine-Learning-Sicht erklärt werden, d. h. welche Features welche Relevanz haben.

- Exportierbarkeit/Deployment vom ML-Modell
  Auflistung aller möglichen Deployment-Arten und Exportmöglichkeiten der ML-Modelle.

- Iterierbarkeit des ML-Prozesses
  Gibt an, wie einfach der ML-Prozess iterierbar und anpassbar ist. Konkret handelt es sich um eine eindeutige Auswahl aus den drei Ausprägungen »beliebiger Punkt«, »nur Zurück« und »an den Anfang«, wobei nur die bestmögliche Ausprägung (die eben genannte Aufzählung ist als absteigend zu verstehen) angeführt wird.

Designer/Visual Coding

- Methoden der Datenbereinigung
  Auflistung aller bereitgestellten Methoden für den Umgang mit Datenqualitätsproblemen, wie fehlerhaften Werten oder doppelten Werten.

- Wahlmöglichkeit von Imputation-Methoden bzw. Umgang mit fehlenden Daten
  Auflistung aller bereitgestellter Methoden für den Umgang mit fehlenden Werten inkl. Imputation-Methoden.

- Benutzerdefinierte Zielfunktion
  Gibt an, ob es möglich ist, eine benutzerdefinierte Zielfunktion zu verwenden.

Man beachte, dass die zu AutoML gehörigen Kriterien gleichermaßen für den Designer/Visual-Coding-Ansatz relevant sind, um Dopplungen zu vermeiden, aber nur für den AutoML-Ansatz bewertet werden.

5.3.2 Textklassifikation

Die in diesem Use Case betrachteten Anwender*innen sind IT-affin und können gebräuchliche strukturierte Datenformate konvertieren und auslesen. Außerdem verfügen sie über rudimentäres Grundlagenwissen zu Machine Learning, haben jedoch kein tieferes Know-how zur Verarbeitung natürlichsprachlicher Texte bzw. deren Auf- und Vorbereitung für Machine-Learning-Verfahren. Sie nutzen daher, wo immer möglich, zunächst den AutoML-Ansatz und wählen nur unumgängliche Parameter selbst. Sie bevorzugen weiterhin das Training und die Evaluierung seiner Modelle über graphische Oberflächen anstelle der Ausführung von (selbst zu schreibendem) Programmcode. Aus dieser Konstellation ergeben sich für den Use Case Textklassifikation die folgenden Kriterien:

- Einfaches Eingangsdatenformat
  Die Daten liegen als CSV-Datei vor und sollen möglichst direkt verwendbar sein.

- Unterstützung deutscher Texte
  Der Umgang mit deutschen Fließtexten sollte in allen Schritten der Vorverarbeitung und des Trainings unterstützt werden.

- Verfügbarkeit und automatische Wahl geeigneter Schritte und Algorithmen
  *Den Anwender*innen sollte nach Möglichkeit eine optimale vorkonfigurierte Pipeline mit allen nötigen Schritten zur Bearbeitung des Problems vorgelegt werden.*

- Qualität der Ergebnisse
  Die erzeugten Modelle sollten möglichst gute Ergebnisse auf den Testdaten erzielen (gemessen an Accuracy und F1-Score).

- Benutzerfreundlichkeit während Training und Evaluation
  Oberflächen und Eingabemasken sind klar strukturiert und einfach zu bedienen, etwaige Fehlermeldungen sind aussagekräftig und Hilfestellungen verfügbar.

- Leichte Wiederverwendbarkeit/Integrierbarkeit der trainierten Modelle
  *Den Anwender*innen sollte eine einfache Möglichkeit gegeben werden, die trainierten Modelle mit weiteren Daten zu testen bzw. sie in andere Systeme einzubinden.*

- Erforderliche Programmierkenntnisse
  *Die Anwender*innen sollten möglichst wenig eigenen Programmcode schreiben müssen und stattdessen über Oberflächen durch den Prozess geführt werden.*

5.3.3 Bildklassifikation

Der in diesem Use Case betrachteten Nutzer*innen weisen eine geringe bis mittlere IT-Affinität auf und haben kein tieferes Know-how im Bereich der Bilderkennung sowie Deep Learning. Sie entscheiden sich damit in aller Regel für die Nutzung von AutoML-Werkzeugen. Die erstellten Modelle sollen exportierbar sein, damit sie ebenfalls auf eigener bzw. anderer Infrastruktur verwendet werden können. Dadurch ergeben sich für den Use Case Bildklassifikation die folgenden Kriterien:

- Verfügbarkeit und automatische Wahl geeigneter Schritte und Algorithmen
  *Den Anwender*innen sollte nach Möglichkeit eine optimale vorkonfigurierte Pipeline mit allen nötigen Schritten zur Bearbeitung des Problems vorgelegt werden.*

- Qualität der Ergebnisse
  Die erzeugten Modelle sollten möglichst gute Ergebnisse auf den Testdaten erzielen.

- Benutzerfreundlichkeit während Training und Evaluation
  Oberflächen und Eingabemasken sind klar strukturiert und einfach zu bedienen. Etwaige Fehlermeldungen sind aussagekräftig und Hilfestellungen sind verfügbar.

- Leichte Wiederverwendbarkeit/Integrierbarkeit der trainierten Modelle
  *Den Anwender*innen sollte eine einfache Möglichkeit gegeben werden, die trainierten Modelle mit weiteren Daten zu testen bzw. sie in andere Systeme einzubinden.*

- Erforderliche Programmiererkenntnisse
  *Die Anwender*innen sollten möglichst wenig eigenen Programmcode schreiben müssen und stattdessen über Oberflächen durch den Prozess geführt werden.*

5.3.4 Zeitreihenanalyse

Die in diesem Anwendungsfall betrachtete Nutzer*innen weisen eine geringe bis mittlere IT-Affinität auf und haben nur rudimentäre Kenntnisse bei der Analyse von Zeitreihendaten zur Regressionsanalyse. In aller Regel wird ein auf AutoML-Werkzeugen basierter Ansatz gewählt. Dies hat für Anwender*innen den Vorteil, dass sie sich nicht im Detail mit der Datenvorverarbeitung und dem Modelltraining auseinandersetzen müssen. Für diesen Anwendungsfall können so automatisierte Lösungen generieren werden, was für die Nutzer*innen und den gesamten Prozess eine große Zeitersparnis darstellt. Für den Anwendungsfall ergeben sich daher die folgenden Kriterien:

- Verfügbarkeit und automatische Wahl geeigneter Schritte und Algorithmen
  *Den Anwender*innen sollte nach Möglichkeit eine optimale vorkonfigurierte Pipeline mit allen nötigen Schritten zur Bearbeitung des Problems vorgelegt werden.*

- Qualität der Ergebnisse
  Die erzeugten Modelle sollten möglichst gute Ergebnisse auf den Testdaten erzielen.

- Benutzerfreundlichkeit während Training und Evaluation
  Oberflächen und Eingabemasken sind klar strukturiert und einfach zu bedienen, etwaige Fehlermeldungen sind aussagekräftig und Hilfestellungen sind verfügbar.

- Leichte Wiederverwendbarkeit/Integrierbarkeit der trainierten Modelle
  *Den Anwender*innen sollte eine einfache Möglichkeit gegeben werden, die trainierten Modelle mit weiteren Daten zu testen bzw. sie in andere Systeme einzubinden.*

- Erforderliche Programmierkenntnisse
  *Anwender*innen sollten möglichst wenig eigenen Programmcode schreiben müssen und stattdessen über Oberflächen durch den Prozess geführt werden.*

5.4 Bewertung der Plattformen

In diesem Abschnitt werden die Versuchsabläufe der einzelnen Use Cases auf den vier Plattformen geschildert und nach den vorher beschriebenen spezifischen Kriterien bewertet. Die folgende Tabelle gibt dabei einen Überblick über die schwerpunktmäßig genutzten Teildienste.

Dienste	Anwendungsfälle			
	5.4.1 Binäre Klassifikation auf tabellarischen Daten	5.4.2 Textklassifikation	5.4.3 Bildklassifikation	5.4.4 Zeitreihenanalyse
Amazon	Amazon SageMaker Autopilot	Amazon Comprehend	Amazon Rekognition	Amazon SageMaker Autopilot
Google	Google AutoML Tables	Google AutoML Natural Language	Google AutoML Vision	Google AutoML Tables
IBM	IBM Watson Studio AutoAI	IBM Watson Natural Language Classifier	IBM Watson Visual Recognition	IBM Watson Studio AutoAI
Microsoft	Microsoft Azure Machine Learning Automated ML	Microsoft Azure Machine Learning Automated ML	Microsoft Cognitive Services (Custom Vision)	Microsoft Azure Machine Learning Automated ML

Tabelle 9: Je Anwendungsfall schwerpunktmäßig genutzter Teildienst der MLaaS-Plattform.

5.4.1 Binäre Klassifikation auf tabellarischen Daten

Allgemeine Beschreibung des Versuchsablaufs

Als primäre Metrik soll Accuracy, gemessen durch 10-fache Kreuzvalidierung, verwendet werden. Da die Aufgabe aus Machine-Learning-Sicht weder besonders komplex noch rechenintensiv ist, wird ökonomisch vorgegangen. Das heißt es wird, sofern möglich, die schwächste Hardwarekonfiguration mit einer geringen Laufzeit (zum Beispiel 0,5 Stunden) ausgewählt. Im Folgenden werden positive wie negative Auffälligkeiten bei der Durchführung des Experiments beschrieben.

In dem vorliegenden Datensatz wurde der Schritt des Data Cleanings schon unternommen. Um eine wirklichkeitsgetreuere Datenqualität zu erhalten, wird ein künstliches Rauschen hinzugefügt. Dabei werden 5 Prozent der Daten aus der Datensatzstichprobe entfernt (z. B. zur Simulation menschlicher Fehler bei der Bearbeitung, Bugs bei Datenerfassung …). Außerdem wird auf 10 Prozent der Variablen (Spalten), aber mindestens auf einer, ein normalverteiltes Rauschen ($\mu=\bar{x}$_Variable, $\sigma = 0.1\sigma$_Variable) addiert (z. B. zur Simulation falsch angebrachter Sensoren).

Amazon Web Services (AWS): Auch bei AWS gilt es, zuerst den richtigen Dienst auszuwählen. Für den Use Case mit strukturierten Daten bietet sich dabei SageMaker an. Innerhalb des Dienstes sind wiederum viele Funktionen anwählbar (Ground Truth, Notebooks …). Um die AutoML-Funktionalität zu erreichen, muss SageMaker Studio ausgewählt werden. Vor der ersten Nutzung des Studios muss dieses eingerichtet werden. Der komplizierte Prozess soll hier nicht weiter erörtert werden. Allerdings kann Studio nicht in jeder beliebigen Region ausgeführt werden: Für europäische Regionen kommt nur Irland infrage. Zwar sind alle nötigen Informationen,

um zu einer funktionierenden Amazon-SageMaker-Studio-Instanz zu gelangen, in der AWS-Dokumentation enthalten, es ist aber umständlich, sie herauszusuchen, da wichtige Informationen auf verschiedenen Seiten der Dokumentation stehen.

Startet man SageMaker Studio, wird eine JupyterLab-Session im Browser gestartet. Auf den ersten Blick ist die AutoML-Funktionalität im (ausschließlich englischen) Menü unter dem Namen »Autopilot« zu sehen.

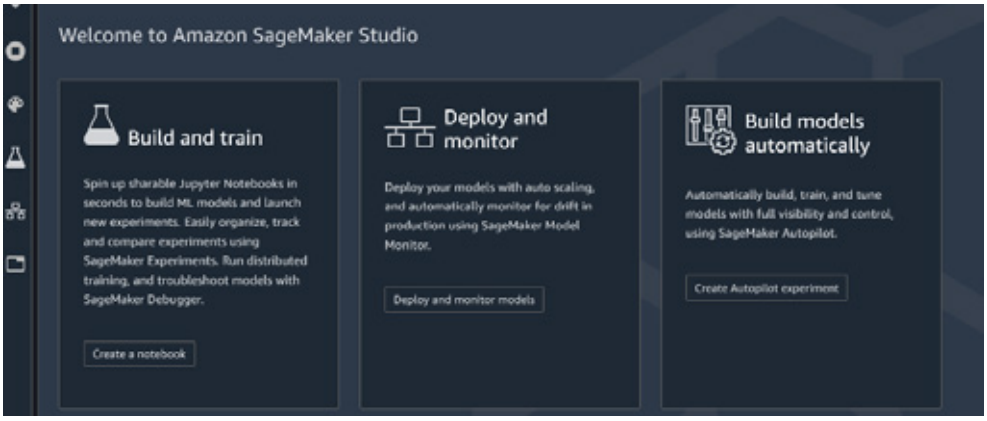


Abbildung 13: Übersicht des Amazon SageMaker Studios[57].

Nach einem Klick darauf wird ein sogenanntes Experiment gestartet. Der Name des Experiments muss ebenso wie ein S3-Bucket, in dem die zu bearbeitenden Daten liegen und das Modell abgelegt werden soll (Input und Output Data Location), spezifiziert werden. S3-Buckets sind u. a. der Storage-Dienst von AWS (der Link zu der Dokumentation ist im Menü angeführt). Auf die Erstellung soll nicht eingegangen werden. Fügt man allerdings Daten in einen Bucket, während der Experiment-Erstellungsdialog offen ist, muss dieser neu geladen werden, um die neue Datei sichtbar zu machen. Die Zielvariable muss von Hand eingegeben werden. Auf Tippfehler wird man (allerdings zu einem späteren Zeitpunkt) hingewiesen.

Weiterhin kann der Problemtyp (Binäre Klassifikation, Regression, Multiklassen-Klassifikation) händisch definiert oder automatisch erkannt werden (Auto). In einem weiteren Reiter »Advanced Settings« können Einstellungen zu den Nutzerrollen (bei Account-Nutzung von mehr als einer Person) durchgeführt werden. Anschließend kann das Experiment gestartet werden. Weitere Informationen bzw. Einstellungsmöglichkeiten zu Kosten und Leistungsressourcen, Zielfunktion oder Kreuzvalidierung sind nicht vorhanden. Ist das Experiment gestartet, wird der Fortschritt angezeigt. Dieser gliedert sich in die Stufen »Analyzing Data«, »Feature Engineering«, »Model Tuning« und »Completed«. Mit einem grünen Häkchen werden die Schritte als

57 *Quelle: Amazon SageMaker Studio.*

beendet angezeigt. Während die ersten beiden Punkte kein weiteres Feedback liefern, sieht man bei »Model Tuning« jedes Modell, das momentan trainiert wird, inklusive Güte (bis Ende November 2020 ohne weitere Details »Objective« genannt, ab Dezember 2020 »Objective: F1«).

Abbildung 14: Ansicht aller vom Autopilot durchgeführten Jobs, inkl. deren Abschneiden[58].

Aus derselben Ansicht können zwei Notebooks geöffnet werden. Das Notebook »candidate generation« liefert detaillierte Informationen zu den Modellen wie Modelltyp, Hyperparameteroptimierungsmethoden, Zielfunktion und genutzte Hardware-Ressourcen. Diese müssen allerdings aus Python-Code herausgelesen werden und bieten somit nur erfahreneren Nutzer*innen weiterführende Einstellungsmöglichkeiten. Das Notebook »data exploration« liefert Informationen zu den Statistiken des Eingabe-Datensatzes. In beiden Notebooks sind hilfreiche Tipps
vermerkt, an welcher Stelle man worauf achten muss und zu welchen weiteren Handlungen geraten wird. Der komplette Prozess dauerte 76 Minuten, das Experiment kann auch von Hand gestoppt werden. Alle trainierten Modelle sind einsehbar. Per Rechtsklick können auch Modelldetails eingesehen und das Modell verfügbar gemacht werden.

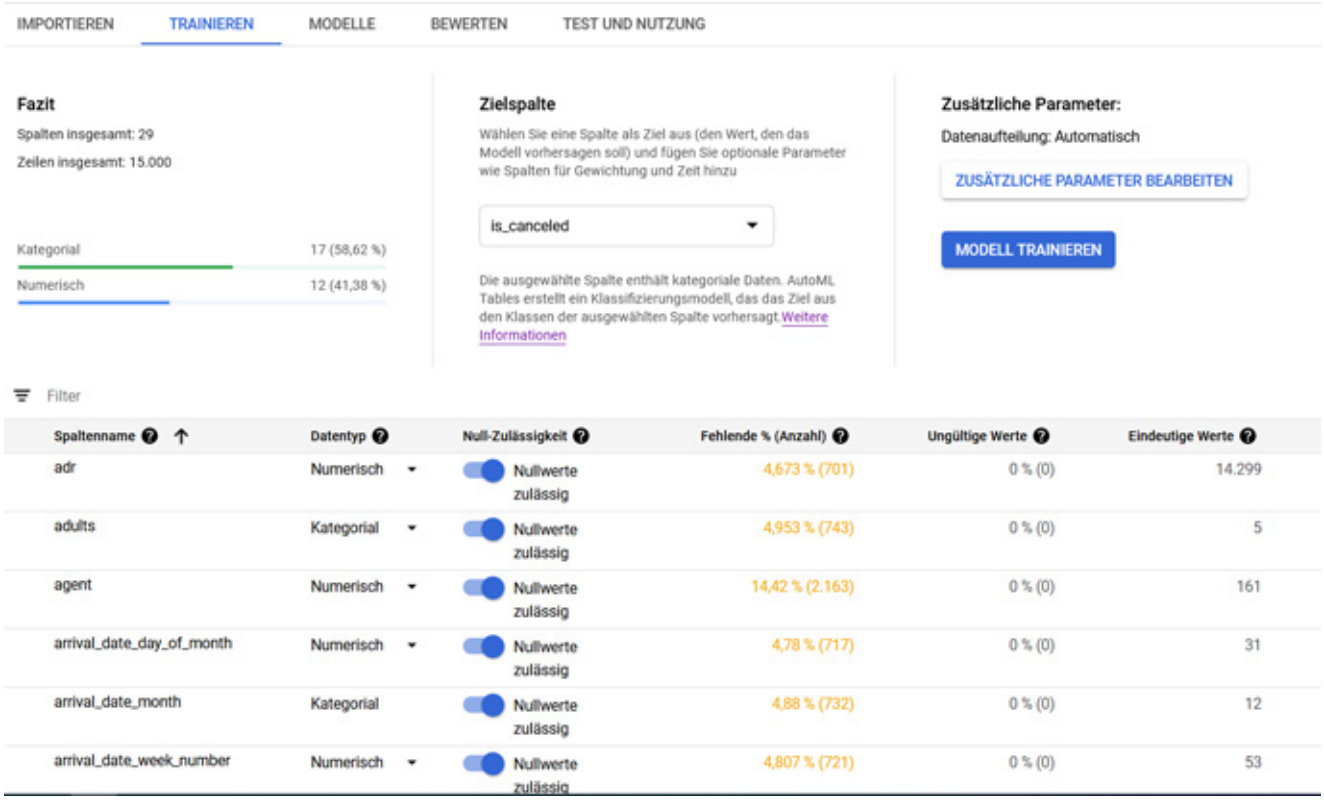
Google Cloud Plattform (GCP): Voraussetzung zur Nutzung von GCP für diesen Use Case ist, dass schon ein Projekt innerhalb des Dienstes erstellt wurde. Auf die Erstellung eines Projekts soll nicht weiter eingegangen werden. In einem ersten Schritt muss innerhalb des Projekts zunächst der Dienst angewählt werden, der das vorliegende Problem der strukturierten Daten am besten löst. Alle verfügbaren Dienste sind nach Themenfeldern (»Tools«, »Big Data« …) sortiert. Unter dem Themenfeld »Künstliche Intelligenz« ist der Dienst »Tabellen (beta)« zu finden. Die Nutzer*innen werden ab diesem Zeitpunkt durch den Arbeitsprozess (organisiert in verschiedenen Reitern) anhand der Menüführung gelotst.

58 *Quelle: Amazon SageMaker Autopilot.*

Zunächst muss ein neuer Datensatz erstellt werden. Dazu muss Name und Region (weltweit, Europäische Union) angegeben werden. Letzteres bestimmt über die Verfügbarkeit des Datensatzes sowie die damit einhergehenden rechtlichen Implikationen. Nach Erstellung des (leeren) Datensatzes müssen konkrete Daten damit assoziiert werden. Es können verschiedene Import-Modi (BigQuery oder CSV) ausgewählt werden, unter anderem auch der CSV-Import vom eigenen Computer. In jedem Fall muss ein CloudStorage Bucket vorhanden sein, also ein weiterer Dienst der GCP. Über die Import-Menüführung wird man automatisch zum richtigen Dienst geleitet. Auf die konkrete Erstellung eines Buckets soll nicht weiter eingegangen werden, allerdings ist zu unterstreichen, dass er in derselben Region wie der Datensatz liegen muss. Das kann zu einiger Verwirrung führen.

Der Import erfordert je nach Größe der Datei einige Zeit, über die Fertigstellung wird man per Mail informiert. Für den hier präsentierten Use Case dauerte der Import zwei Minuten. Anschließend wird man direkt auf eine Seite geführt, auf der der Datensatz zusammengefasst zu sehen ist (Abbildung 15).

Abbildung 15: Ausschnitt aus Google AutoML Tables. Im Reiter »Trainieren« wird eine Zusammenfassung des hochgeladenen Datensatzes dargestellt. Variablentypen werden automatisch erkannt, können aber auch manuell angepasst werden. Nullwerte in Spalten können ebenfalls entfernt werden[59].



59 Quelle: GCP AutoML Tables.

Für das Modelltraining muss eine Zielvariable ausgewählt werden. Dass dazu keine Nullwerte in der Zielspalte enthalten sein dürfen, wird nicht erwähnt. Ansonsten werden verschiedene Aspekte durch Tooltipps (Fragezeichensymbole in Abbildung 15) oder Links zur Dokumentation vertiefend erklärt. Zusätzliche Parameter wie die unterschiedliche Gewichtung der Zeilen sowie ein benutzerdefinierter Train-Test-Split, können spezifiziert werden. Der Problemtyp wird anhand der Zielvariable automatisch erkannt. Für das Training kann ein Modellname angegeben werden sowie ein Trainingsbudget in ganzzahligen Knotenstunden zwischen 1 und 72. Ein weiterführender Link zur Preisliste informiert über die Kosten pro Knotenstunde. Der Trainingsaufwand bleibt immer innerhalb des Budgets und stoppt automatisch, sollte erkannt werden, dass sich das Modell nicht mehr verbessert. Ebenfalls können konkret Spalten ausgewählt werden, die mitmodelliert werden sollen, sowie die Optimierungsfunktion. Kreuzvalidierung kann nicht ausgewählt werden. Das Training kostete eine Knotenstunde, über den Abschluss wird erneut per Mail informiert.

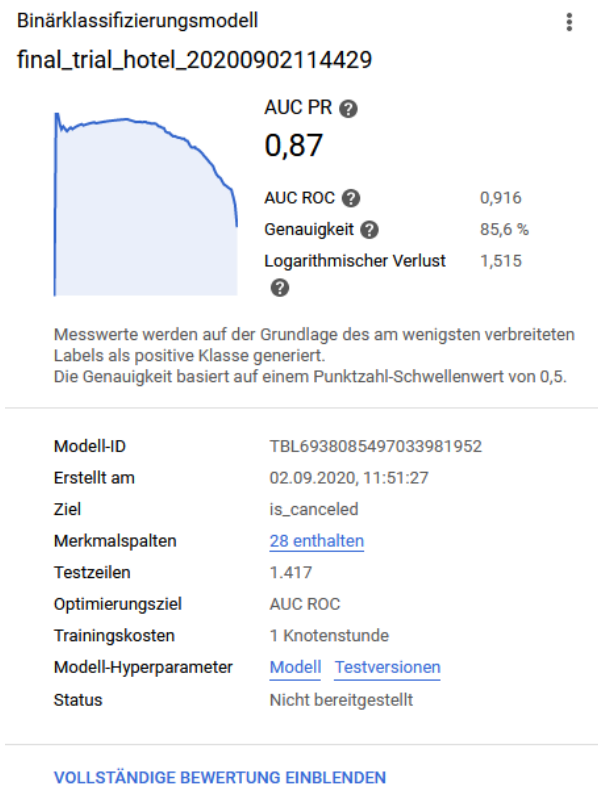


Abbildung 16: Zusammenfassende Darstellung des trainierten Modells[60].

60 *Quelle: GCP AutoML Tables.*

Im »Bewerten«-Reiter wird man über die Güte des Modells anhand verschiedener Metriken informiert. Schaubilder der ROC-Kurven geben zusätzliche Informationen. Auch ein Schaubild über die Merkmalswichtigkeit (vorteilhaft für Erklärbarkeit des Modells) ist einsehbar.

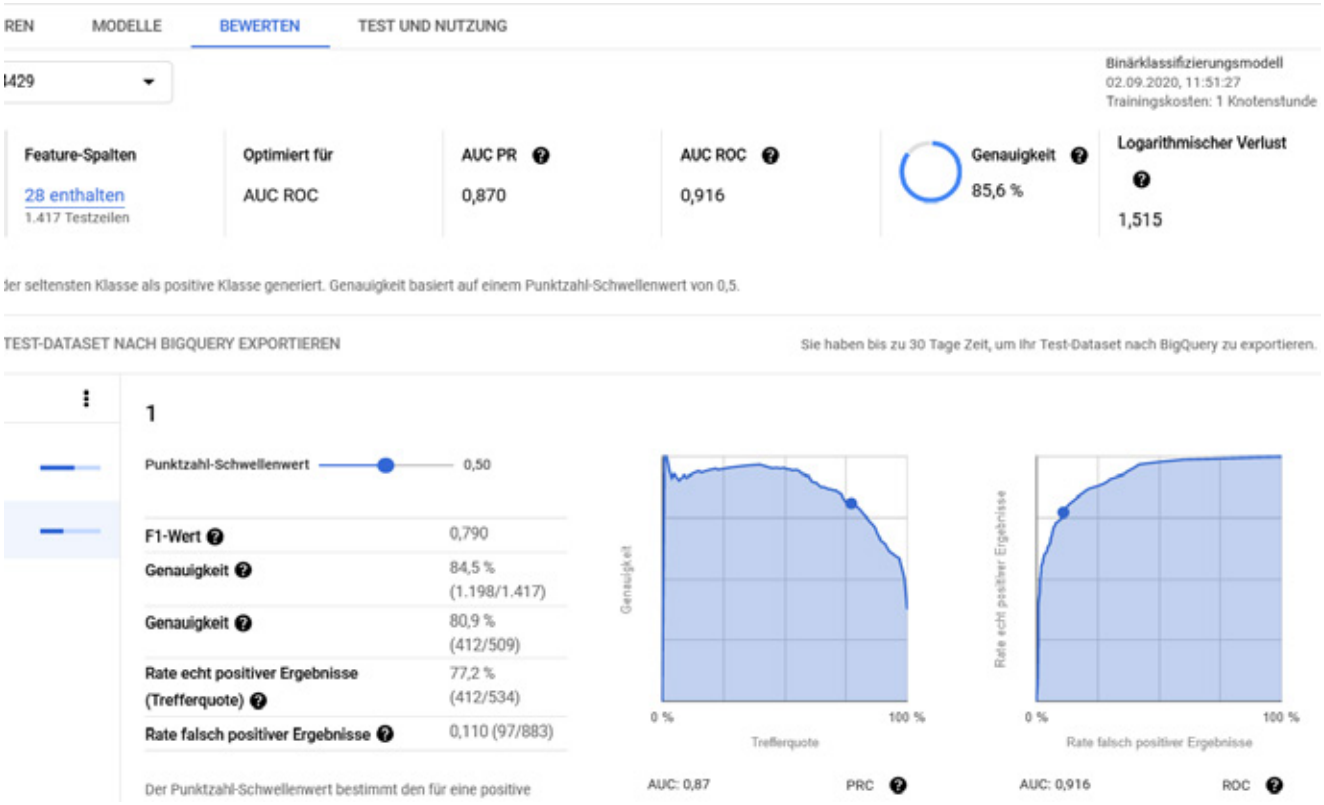


Abbildung 17: Ausschnitt aus einer detaillierten Ansicht des trainierten Modells[61].

61 Quelle: GCP AutoML Tables.

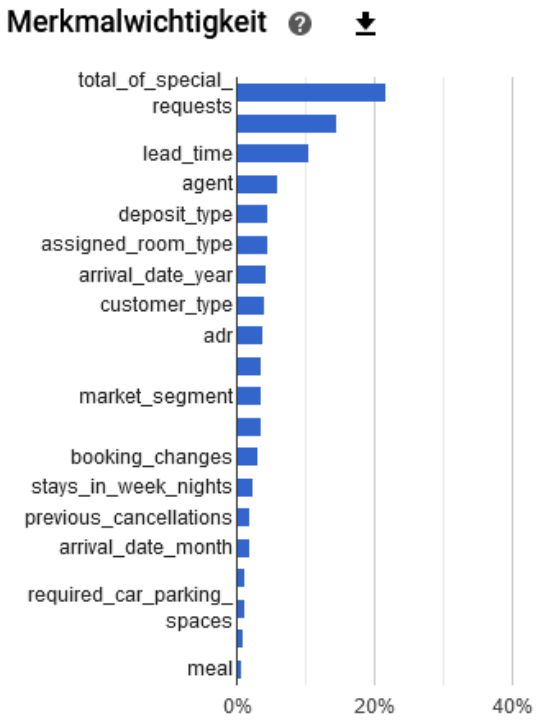


Abbildung 18: Darstellung der XAI-Komponenten »Merkmalswichtigkeit«. Dabei wird der Beitrag einzelner Variablen aus dem Datensatz zur Inferenz erörtert[62].

Das fertige Modell kann für Onlinevorhersagen genutzt werden: Variablenwerte können händisch eingetragen und dann geschätzt werden. Alternativ kann das Modell auf der GCP bereitgestellt werden. So wird eine REST-Schnittstelle bereitgestellt, mit der in Echtzeit das Modell genutzt werden kann. Über zusätzlich anfallende Kosten wird man erneut per Link informiert. Ebenfalls kann man eine weitere CSV-Datei angeben oder eine BigQuery für eine Batch-Vorhersage. Das Modell kann in Form eines (TensorFlow-) Docker-Containers exportiert werden.

IBM Watson Machine Learning: Zunächst wurde der automatisierte Assistent AutoAI genutzt. Bei der Erstellung eines solchen Experiments fällt auf, dass die verwendeten Rechenressourcen nicht konfigurierbar sind. Nachdem das Experiment angelegt ist, kann es mit dem Dateiupload losgehen; hier wird nur ein einziges Format, CSV, unterstützt. Die Benutzeroberfläche ist sehr übersichtlich und verständlich (viele Tooltipps). Fehlermeldungen sind auf Englisch formuliert. Nachdem die Datenquelle hochgeladen wurde, kann bei Bedarf sofort mit dem Training begonnen werden. Dies ist Automated Machine Learning par excellence und gerade für Anfänger*innen, die schnell und unkompliziert starten wollen, von Vorteil. Auch wenn fast alle Feature-Skalen richtig erkannt werden, fehlt die Möglichkeit, sie manuell anzupassen bzw. zu korrigieren. Die Anzahl der Folds der voreingestellten Kreuzvalidierung sind auf drei festgelegt

62 *Quelle: GCP AutoML Tables.*

und nicht änderbar. Das Training, insbesondere dessen Fortschritt, wird ansprechend dargestellt, siehe Abbildung 20. Parallel wird eine Pipeline-Bestenliste aktualisiert und ein Protokoll, das man auch herunterladen kann, geführt. Nach Abschluss des Trainings, das 24 Minuten dauerte, stellt sich heraus, dass der LGBM-Klassifikator mit einer Erstellzeit von nur vier Sekunden die beste Pipeline darstellt. Jede Pipeline kann ausgewählt und näher untersucht werden, es stehen jeweils einige Möglichkeiten der Ergebnisvisualisierung zur Verfügung. Mit wenigen Mausklicks kann das Experiment mit anderen Einstellungen wiederholt werden. Anschließend kann das Ergebnis als Modell oder als Jupyter Notebook gespeichert werden. Im ersten Fall wird die Bereitstellung des Modells als Web-Service angeboten. Im zweiten Fall kann man zwischen einer Verwendung in IBM Watson und anderweitiger Verwendung wählen, den Code also etwa als Jupyter Notebook oder .py-Datei herunterladen.

Neben dem AutoAI-Modul ist der Visual-Coding-Ansatz von Relevanz: Der »Data Refinery-Ablauf« ermöglicht laut Beschreibung die Bereinigung, Anpassung und Erweiterung der Daten. Darüber hinaus bietet er sehr umfangreiche Möglichkeit der explorativen Datenanalyse. Doch erst das Modul »Modeler-Datenfluss« ermöglicht das Erstellen eines ML-Arbeitsablaufs inkl. Modelltraining. Auch dieses Modul bietet umfangreiche Bearbeitungsmöglichkeiten der Daten. Insgesamt bietet das Modul über 100 Operatoren (u. a. aus den Bereichen Import, Feldoperationen, Modellierung, Grafik, Export) mit schier endlosen Konfigurationsmöglichkeiten. Die Anzahl der Möglichkeiten ist umfassend, die angesprochene Zielgruppe sollte nahezu alle Ideen umsetzen können.

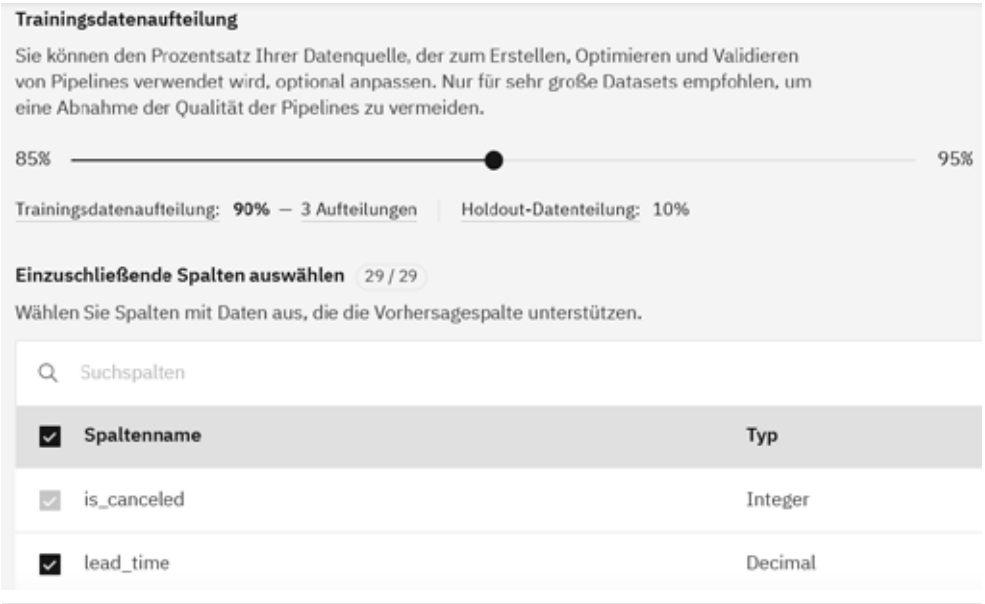


Abbildung 19: Experimenteinstellungen eines AutoAI-Experiments[63].

63 *Quelle: IBM Watson AutoAI.*

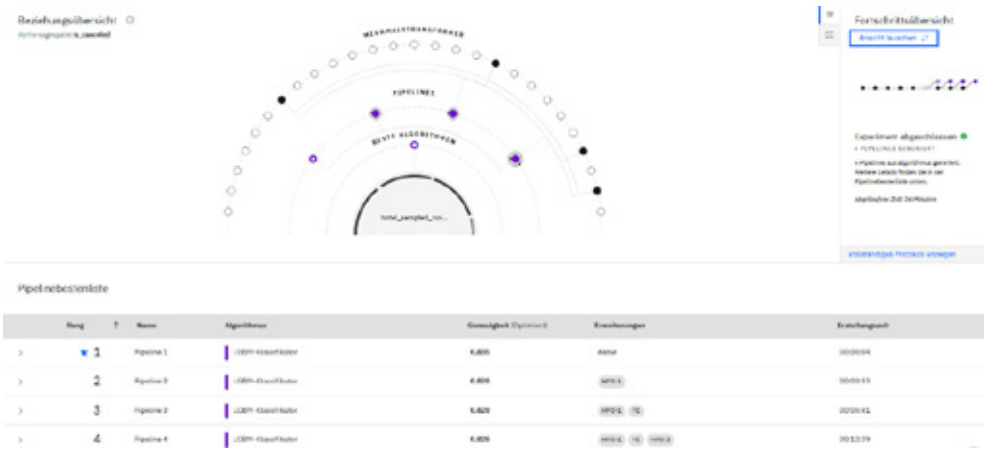


Abbildung 20: Visualisierung der Pipeline-Ergebnisse bei IBM Watson Machine Learning AutoAI[64].

Name	Tool	Hardwarekonfiguration	Sprache
Default Spark 2.4 & Python 3.7	Notebook	2 Executors: 1 vCPU and 4 GB RAM, Driver: 1 vCPU and 4 GB RAM	Python 3.7 with Spark
Default Spark 3.0 & R 3.6	Notebook	2 Executors: 1 vCPU and 4 GB RAM, Driver: 1 vCPU and 4 GB RAM	R 3.6 with Spark
Default Spark 3.0 & Scala 2.12	Notebook	2 Executors: 1 vCPU and 4 GB RAM, Driver: 1 vCPU and 4 GB RAM	Scala 2.12 with Spark
Default Spark 3.0 & Python 3.7	Notebook	2 Executors: 1 vCPU and 4 GB RAM, Driver: 1 vCPU and 4 GB RAM	Python 3.7 with Spark
Default Python 3.7 S	Notebook	4 vCPU and 16 GB RAM	Python 3.7
Default Python 3.7 XS + DO	Notebook	2 vCPU and 8 GB RAM	Python 3.7
Default Python 3.7 XS	Notebook	2 vCPU and 8 GB RAM	Python 3.7

Abbildung 21: Breite Auswahl sogenannter Umgebungs-definitionen[65].

Microsoft Azure Machine Learning: Microsoft leitet die User übersichtlich, deutschsprachig und mit vielen Tooltipps durch die einzelnen Schritte. Beim Fileupload wird kein Excel-/.xlsx-Format unterstützt, Abhilfe verschafft jedoch zum Beispiel das leicht zu erzeugende CSV-Format. Die Kosten des zu erstellenden Computeclusters werden transparent und direkt sichtbar angezeigt. Die erste Ausführung im Experiment schlug fehl, es erschien eine englische Fehlermeldung, die auf eine Nichtverträglichkeit von Kreuzvalidierung und Deep Learning hinwies. Eine direkte Behebung der Fehlerursache, also zum Beispiel das Abschalten von Deep Learning, scheint nicht möglich zu sein. Stattdessen musste eine neue Ausführung erstellt werden. Nach Abschluss des Trainings, das 38 Minuten dauerte, wird eine Zusammenfassung der Ergebnisse

64 *Quelle: IBM Watson AutoAI.*

65 *Quelle: IBM Watson.*

angezeigt. Ein sogenannter »Datenintegritätsschutz« ist »eine Folge von Überprüfungen für die Eingabedaten, um sicherzustellen, dass qualitativ hochwertige Daten zum Trainieren des Modells verwendet werden«.

Das AutoML-Modul trainiert mehrere Modelle, wobei das beste (in unserem Fall LightGBM mit nur 57 Sekunden Trainingszeit) vorausgewählt wird. Unter »Erklärungen« wird dargestellt, welche Features sich auf das Modell auswirken und warum. Was den anschließenden Export des Modells bzw. dessen Bereitstellung angeht, ist anzumerken, dass ad hoc nur Azure Container Instances (Bereitstellung über Rest-Endpunkt, schlüsselbasierte Authentifizierung optional) funktionierte. Laut Dokumentation sind mehrere andere Bereitstellungsarten möglich, diese erfordern jedoch die vorherige Vorbereitung bzw. das Ausführen von Code.

Die Fortbewegung innerhalb eines AutoML-Projekts ist zufriedenstellend, Nutzer*innen navigieren navigiert vorwärts und rückwärts und dabei werden die vorgenommenen Einstellungen beibehalten. Da die Erstellung eines solchen Experiments recht schlank gehalten ist, ist es nicht unbedingt nötig, an beliebige Stellen des Experiments springen zu können. Wünschenswert jedoch wäre die Möglichkeit, bei abgeschlossenen Experimenten einzelne Einstellungen verändern zu können und das Experiment daraufhin zu wiederholen. Dies ist so nicht möglich, es muss ein neues Experiment angelegt werden.

Der Designer/Visual Coding differenziert sich durch Einfachheit und viele vorkonfigurierte Funktionen (78 an der Zahl). Der Funktionsumfang der bereitgestellten Module umfasst viele der Funktionen, die man per Code und bekannten ML-Libraries wie scikit-learn verwenden würde. Interessant ist der Baustein »Tune Model Hyperparameters«, mit dem sich Hyperparameter bei der Modellerstellung optimieren lassen. Insgesamt sei angemerkt, dass die vorhandenen Beispielprojekte (wie zum Beispiel »Binary Classification« oder »Text Classification« im Designer auf eigene Problemstellungen leicht anpassbar sind, sodass bei Bedarf nicht von Grund auf ein Arbeitsablauf neu erstellt werden muss.

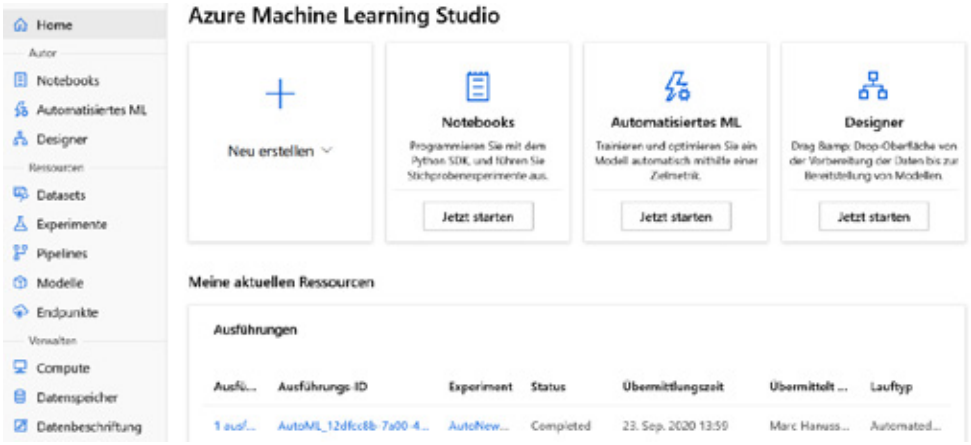


Abbildung 22: Microsoft Azure Machine Learning Studio[66].

Algorithmusname	Erklärt	Genaui... ↓	Stichprob... ⓘ
MaxAbsScaler, LightGBM	Erklärung anzeigen	0.84897	100.00 %
StackEnsemble		0.84855	100.00 %
VotingEnsemble		0.84855	100.00 %

Abbildung 23: Auszug aus erstellten Modellen samt Genauigkeit und Dauer der Erstellung[67].

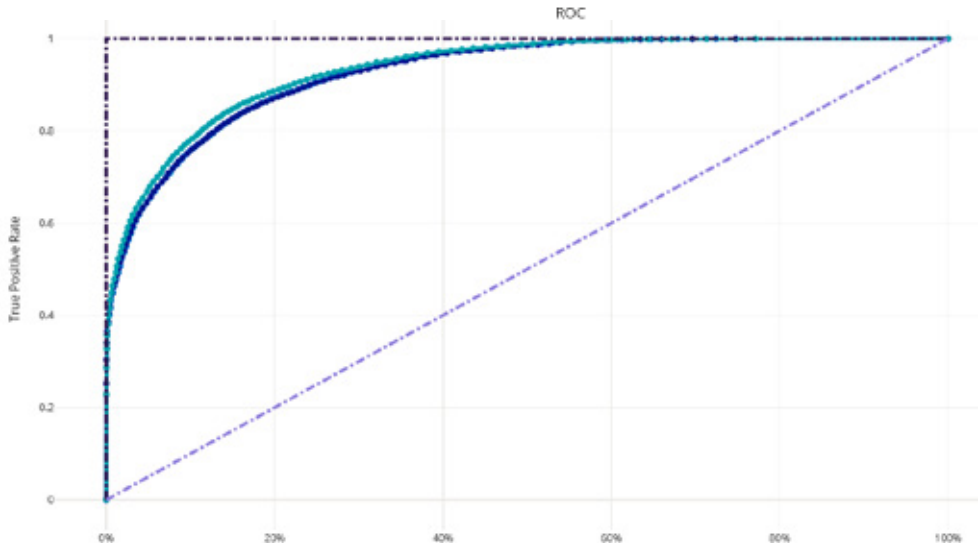


Abbildung 24: Automatisch erzeugte ROC-Kurve verschiedener Klassifikatoren[68].

66 *Quelle: Microsoft Azure Machine Learning Studio.*

67 *Quelle: Microsoft Azure Machine Learning.*

68 *Quelle: Microsoft Azure Machine Learning.*

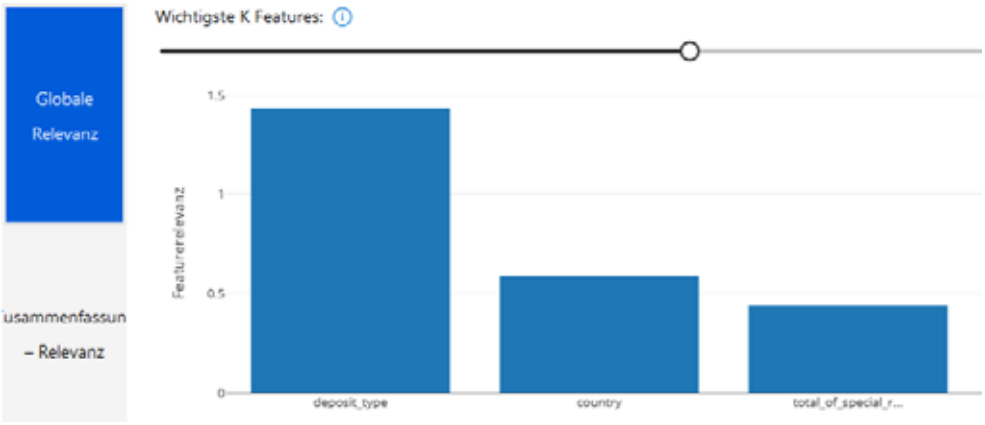


Abbildung 25: Erklärungsansatz zu einem erstellten Modell, hier globale Feature Importance[69].

	Amazon Web Services (SageMaker)	Google Cloud Platform	IBM Watson Machine Learning	Microsoft Azure Machine Learning
AutoML Unterstützte Datentypen für Fileupload	CSV	CSV, BigQuery	CSV	Durch Trennzeichen getrennt (z. B. CSV, TSV), Parquet, JSON Lines und Nur-Text
Richtige Erkennung und Parsing der Feature-Skalen	K.A. in der Oberfläche	21/28	27/28	25/28
Wahlmöglichkeit der Zielfunktion	Nein	Ja	Ja	Ja
Ergebnisdarstellung und Evaluations-metriken	Ergebnisdarstellung: Metrikdiagramm	Metrikdiagramm, ROC-Kurve, PR-Kurve	Metrikdiagramm, ROC-Kurve, Konfusionsmatrix, Genauigkeitsrückrufkurve	Precision-Recall, ROC-Kurve, Calibration Curve, Lift Curve, Gain Curve, Confusion Matrix
	Evaluationsmetriken: F1-Score, »error«	Accuracy, ROC AUC, F1-Score, PR AUC, Loga-rithmischer Verlust, TP-Ra-te, FP-Rate	Accuracy, ROC AUC, F1-Score	Accuracy, AUC gewichtet, F1-Score Makro
Erklärungsansätze/ XAI	Nein	Ja	Ja	Ja

69 *Quelle: Microsoft Azure Machine Learning.*

	Amazon Web Services (SageMaker)	Google Cloud Platform	IBM Watson Machine Learning	Microsoft Azure Machine Learning
Exportierbarkeit des ML-Modells	Download der Modelle, automatisch generierter Code (Notebooks)	Download eines Docker-Containers mit TensorFlow-Code	Automatisch generierter Code	Download des ML-Modells als .pkl-Datei
Deployment des ML-Modells	Web-Service	Batch-Vorhersage, Web-Service	Web-Service	Azure Kubernetes Service, Azure Container Instances
Iterierbarkeit des ML-Prozesses	Nein	Beliebiger Punkt	Beliebiger Punkt	Nur Zurück
Designer/Visual Coding Methoden der Datenbereinigung			Beschneiden unerwünschter Feature-Ausprägungen und anschließendes Ersetzen mittels aller gängigen Methoden, Entfernen identischer Zeilen	Beschneiden unerwünschter Feature-Ausprägungen und anschließendes Ersetzen mittels aller gängigen Methoden, Entfernen identischer Zeilen
Wahlmöglichkeit von Imputation-Methoden bzw. Umgang mit fehlenden Daten		Zeilen mit Nullwerten können ausgeschlossen werden	Ersetzen fehlender Daten mittels aller gängigen Methoden	Ersetzen fehlender Daten mittels aller gängigen Methoden
Benutzerdefinierte Zielfunktion		Nein	Ja	Nein

Tabelle 10: Zusammenfassung der Ergebnisse.

Zur besseren Vergleichbarkeit werden die Metriken der Modelle auf den verschiedenen Plattformen in einer gesonderten Tabelle präsentiert. So vorhanden, sollen bei jedem Modell die Standard-Metriken für binäre Klassifikationsprobleme (Accuracy, zu Deutsch Genauigkeit und der F1-Score) zum Vergleich herangezogen werden. Eine genaue Beschreibung der Metriken ist in Tabelle 3 zu finden. Fehlt ein Wert, so wurde diese Metrik von der Plattform nicht (ohne größeren Aufwand im Quellcode zu suchen) angeboten.

Dienst	Accuracy	F1-Score
Amazon	k.A.	0.84
Google	0.86	0.79
IBM	0.84	0.81
Microsoft	0.85	0.84

Tabelle 11: Zusammenfassung der ML-Modellgüte.

5.4.2 Textklassifikation

Allgemeine Beschreibung des Versuchsablaufs
Als primäre Metrik dient Accuracy der korrekt erkannten Zielklassen, gemessen anhand eines einheitlichen Train-Test-Splits von 9:1. Wo möglich, wurden weitere 10 Prozent des Datensatzes (entnommen aus dem Trainingsdatensatz) als Validierungsdatensatz spezifiziert. Bei Tools, in denen sich der Validierungsdatensatz nicht manuell spezifizieren ließ, fand lediglich die Aufteilung in Train/Test statt. Da die Anwender*innen aus dem Use Case über keine tieferen Kenntnisse zu Machine Learning und Textanalyse verfügen, wurden im Regelfall (sofern nicht im Folgenden anders beschrieben) alle Tools in ihrer voreingestellten Standardkonfiguration verwendet.

Amazon Web Services (AWS): Innerhalb der bei AWS zur Verfügung stehenden Dienste war der passende für Textklassifikation in Form von Amazon Comprehend[70] schnell zu finden. Es können verschiedene NLP-Analysejobs durchgeführt werden. Das Training eigener Modelle wird derzeit allerdings nur für Textklassifikation und Entity Recognition unterstützt. Entscheidet man sich auf der übersichtlichen Web-Oberfläche für das Training eines eigenen Text-Klassifikators, werden die auszuwählenden Optionen entsprechend für diese Aufgabe dynamisch angepasst: Im Gegensatz zu den anderen Plattformen muss hier erstmals explizit die Sprache der Textinhalte ausgewählt werden, die in diesem Fall auf Deutsch gesetzt wurde.

Amazon bietet zwei verschiedene, aber fest definierte CSV-Formate jeweils für Multi-Class oder für Multi-Label-Probleme an. Hier wurde der Multi-Class-Modus gewählt, d. h. mehrere Klassen, aber jedem Text wird nur genau eine davon zugewiesen. Alternativ kann auch ein mit Amazon SageMaker erzeugter »Ground Truth« als Input dienen. In diesem Versuch wurde jedoch aufgrund der vorliegenden CSV-Datensätze die erste Eingabemethode gewählt. In jedem Fall muss ein in Amazon Comprehend zu verwendender Datensatz vor Verwendung zunächst manuell in einem Amazon S3-Bucket abgelegt worden sein. Eine Upload-Möglichkeit direkt aus Amazon Comprehend heraus ist nicht vorhanden. Alle an dieser Stelle ausgewählten Daten fließen als Trainingsdaten ein. Eine manuelle Aufteilungsmöglichkeit in Trainings- und Validierungsmenge wird Anwender*innen auch hier nicht zur Verfügung gestellt.

Da ansonsten keine weiteren Trainingsparameter einstellbar sind, kann auch bei Amazon Comprehend von einem reinen AutoML-Ansatz gesprochen werden. Das Training dauerte in diesem Fall 142 Minuten und konnte erst nach mehreren erfolglosen Versuchen vollständig durchgeführt werden. Im Nachhinein ist dies wohl auf die nicht zu 100 Prozent den Anforderungen genügende Datenformatierung zurückzuführen. Es war jedoch aus den fehlgeschlagenen Versuchen in keiner Weise herauszulesen, da diese lediglich mit einer nicht weiter erklärten Fehlermeldung

70 https://aws.amazon.com/de/comprehend

abgebrochen wurden. Nach erfolgreichem Training haben Anwender*innen die Möglichkeit, sich die resultierenden Metriken auf dem (nicht selbst spezifizierten) Validierungsset, wie in Abbildung 26 dargestellt, anzuzeigen oder die Ergebnisse in Form einer Wahrheitsmatrix aus einem Output-Bucket herunterzuladen. Das eigentliche Modell steht nach dem Training sofort als AWS-Service mit zugehöriger API-Schnittstelle für das Durchführen von Jobs bereit. Über die Web-Oberfläche konnte anschließend auch ein solcher Job zur Vorhersage auf dem Testdatensatz durchgeführt werden, von dem wiederum die vorhergesagten Labels in ein Output-Bucket geschrieben wurden. Nach dem Download wurde mittels eines eigens erstellten Python-Scripts die Accuracy berechnet, da diese von Analysis-Jobs in Amazon Comprehend nicht automatisch generiert wird. Umfang des Skripts und benötigter Programmieraufwand fielen hier allerdings sehr gering aus und sollten auch von IT-affinen Personen ohne Machine-Learning-Hintergrund zu bewerkstelligen sein.

Für eine Live-Vorhersage auf Texten lassen sich für ein trainiertes Modell in Amazon Comprehend über die Web-Oberfläche weiterhin ein Endpunkt oder mehrere Endpunkte definieren, die pro eingegangene Anfrage bepreist werden.[71]

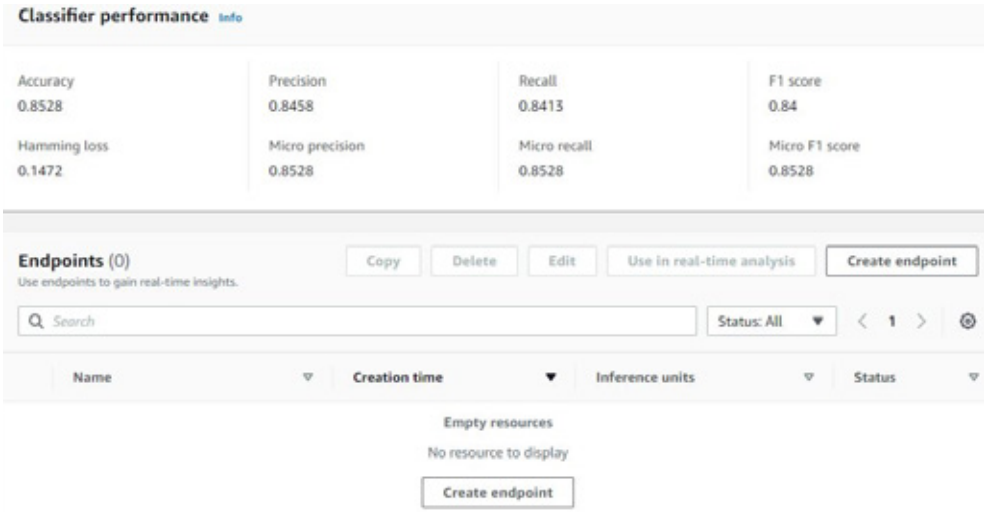


Abbildung 26: Ergebnisdarstellung der Textklassifikation mit Amazon Comprehend[72].

Google Cloud Plattform (GCP): Google stellt mit Cloud AutoML Natural Language ein Tool zur Verfügung, das sich explizit für das Training von eigenen Modellen zur Textklassifikation eignet. Hier ist zu bemerken, dass die Anwender*innen von Anfang an durch einen klaren Dialog mit Auswahlmöglichkeiten zur Bearbeitung des Problems geführt werden. Neben AutoML für

71 *https://aws.amazon.com/de/comprehend/pricing/*

72 *Quelle: Amazon Comprehend.*

Textklassifizierung existieren auch Tools für Named Entity Recognition und Sentiment Analysis. Für die Textklassifizierung selbst haben Anwender*innen wiederum die Wahl, ob sie ein Modell mit einer einzigen oder mehreren Zielklassen pro Text trainieren möchten. In diesem Use Case wurde die Problemstellung mit einem einzigen Ziellabel gewählt, da alle Texte einer einzigen festen Kategorie zugeordnet sind.

Im nächsten Schritt kann der Datensatz hochgeladen werden. Auch bei Google ist wieder ein spezielles CSV-Datenformat vorgegeben, im Gegensatz zu den anderen Plattformen lassen sich hier jedoch von Anfang an einzelne Datenreihen auch direkt über eine zusätzliche Spalte als zugehörig zu Train-, Validation- oder Testdatensatz markieren. Dies hat zur Folge, dass das optimierte Modell zum Schluss bereits automatisch einmal final auf den Testdaten ausgewertet wird. Die Ergebnisse auf dem Testdatensatz werden anschließend in der in Abbildung 27 aufgeführten Modellübersicht visualisiert. Hier wird unter anderem auch eine Wahrheitsmatrix mit Informationen zur Performance auf den einzelnen Zielklassen abgebildet.

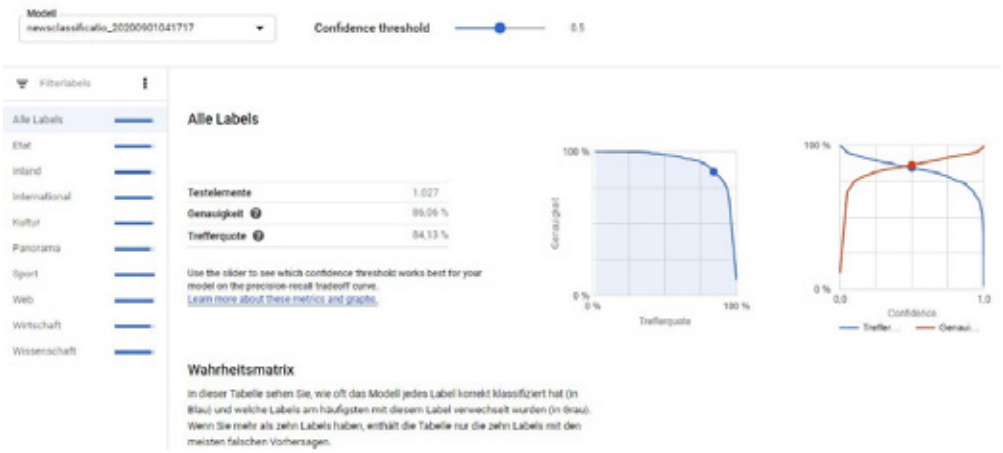


Abbildung 27: Trainings- und Bewertungsübersicht bei Google Cloud AutoML Natural Language[73].

Google bietet die Möglichkeit, einen hochgeladenen Datensatz noch einmal anzupassen, allerdings kann lediglich die Labelzuordnung verändert werden, die Texte selbst sind nicht editierbar. Es existieren keinerlei Konfigurationsmöglichkeiten für das Training. Lediglich der Name des Modells kann hier spezifiziert werden, und ob das Modell im Anschluss an das Training unmittelbar bereitgestellt werden soll. Letzteres ist zwingend erforderlich, um weitere Vorhersagen oder Modellbewertungen auf anderen Daten als den bereits zu Beginn als Train, Validation und Testset spezifizierten Datensätzen durchzuführen. Die Bereitstellung eines Modells bringt allerdings weitere laufende Kosten mit sich[74] und wurde in diesem Versuch nicht durchgeführt, da die Ergebnisse für das Testset bereits zur Verfügung standen.

73 *Quelle: Google Cloud AutoML Natural Language.*

74 *https://cloud.google.com/natural-language/automl/pricing*

IBM Watson Machine Learning: Der Text-Use-Case wurde mithilfe des IBM Cloud Pak for Data[75] in Watson Studio realisiert. In dieser Plattform ist es möglich, weitere Watson Services wie den Natural Language Classifier[76], der für diesen Anwendungsfall als der passende Kandidat erschien, direkt zu integrieren und damit eine Art AutoML-Training durchzuführen. IBM Watson erlaubt einen direkten Upload von lokalen Datensätzen in ein neues Projekt. Das Dateiformat ist hier allerdings auf ein festes und durch IBM spezifiziertes CSV-Format[77] beschränkt, das auch mit seiner Beschränkung auf 1024 Zeichen pro Textsample eine gewisse Einschränkung darstellt und zusätzliche manuelle Vorverarbeitung der Daten erforderte.

Positiv ist, dass die hochgeladenen Texte in einer direkten Übersicht einsehbar sind und sich die einzelnen Texte und Labels für jedes Datensample noch manuell in der Plattform nachbearbeiten lassen. Invalide Datenreihen, die die Vorgaben nicht erfüllen, werden wie in Abbildung 28 illustriert mit Fehlermeldungen angezeigt und lassen sich so auch nachträglich noch korrigieren.

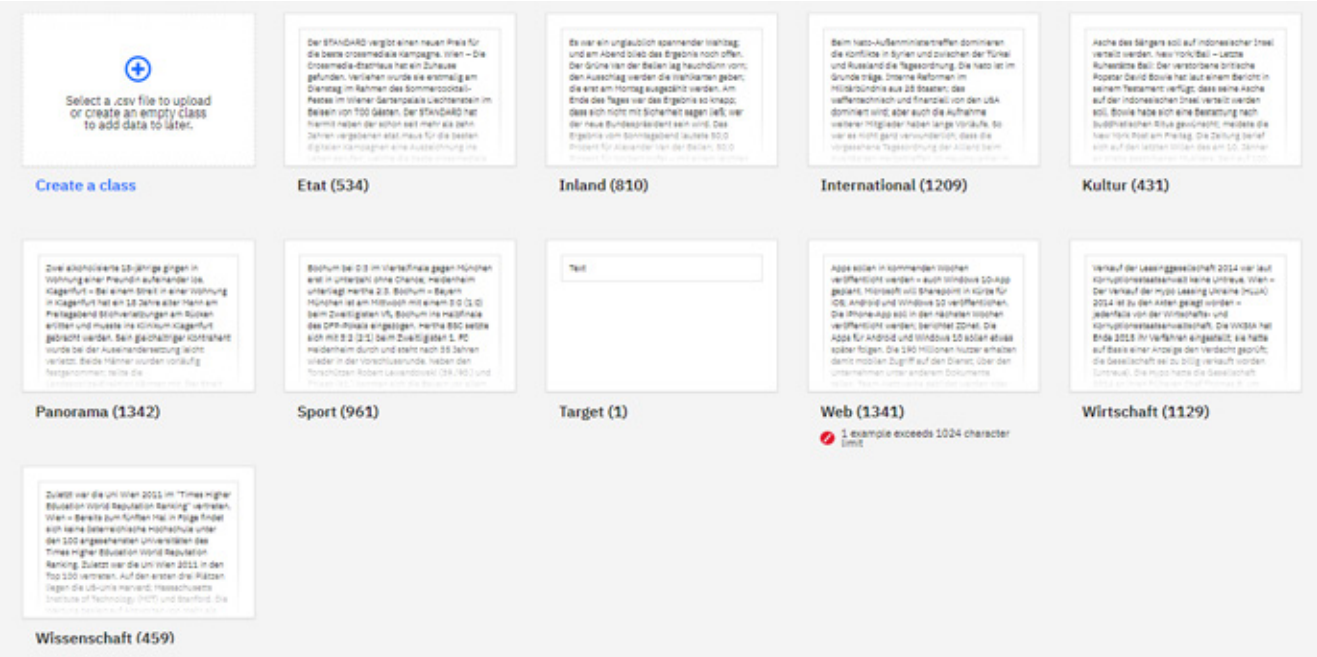


Abbildung 28: Datenübersicht für IBM Watson mit dem Natural Language Classifier[78].

75 *https://www.ibm.com/de-de/products/cloud-pak-for-data*

76 *https://www.ibm.com/watson/services/natural-language-classifier*

77 *https://eu-gb.dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/nlc-prepare.html*

78 *Quelle: IBM Watson Natural Language Classifier.*

Auf der anderen Seite lässt der Natural Language Classifier an dieser Stelle keinerlei weitere Konfiguration des Trainingsablaufs zu, weder Trainingszeit noch Zielmetrik oder andere Parameter können von den Anwender*innen angepasst werden. Auch der eigentliche Trainingsablauf erfolgt in einer noch eher experimentell erscheinenden Blackbox, die in den ersten Versuchen auch bei jedem Anlauf in einem Abbruch mit nichtssagenden Fehlermeldungen endete, ohne dass den Anwender*innen weitere Hilfestellungen geboten wurden. Erst mit der mehrmaligen Erstellung eines neuen Projekts und einer neuen Instanz des NL-Classifiers konnte das Experiment fehlerfrei durchlaufen und lieferte ein trainiertes Klassifikationsmodell. Dafür wurden jedoch in der Weboberfläche keinerlei Statistiken oder weitere Informationen wie Trainingsdauer oder Performance auf einem Validierungsdatensatz angezeigt. Stattdessen steht das Modell allerdings direkt als Service zur Verfügung und kann entweder über die Web-Oberfläche oder aufgelistete API-Aufrufe für Vorhersagen genutzt werden.

Da das User Interface zum Testen des Modells lediglich eine maximale Größe von 30 Texten in einer CSV auf einmal zulässt, wurde ähnlich zu Microsoft Azure auch hier auf einen programmatischen Ansatz gesetzt, in dem mithilfe eines einfachen API-Aufrufs aus einem Python Script die Vorhersagen auf dem Modell für die einzelnen Texte aus dem Testdatensatz generiert wurden. Die vorhergesagten Textlabels konnten wiederum mit dem eigens erstellten Auswertungs-Script in Python zu einer Accuracy-Metrik aggregiert werden.

Microsoft Azure Machine Learning: Für das Training eines Textklassifikations-Modells wurden die in Microsoft Azure zur Verfügung stehenden AutoML-Funktionen verwendet. Die Suche nach der geeigneten Anwendung verglichen mit den anderen Plattformen war am zeitaufwendigsten. Zwar beinhalten die AutoML-Tools die Möglichkeit, einzelne Spalten in einem tabellarischen Datensatz als Textmerkmale zu deklarieren, und damit den geforderten Use Case abzudecken, allerdings wurde die erforderliche Komponente eher zufällig entdeckt. Stattdessen gelangen neue Anwender*innen auf der Suche nach speziellen Funktionen für Textverarbeitung schnell zu dem »gewöhnlichen« Machine Learning Designer, der bereits eine einfache Beispielpipeline für Textklassifikation aufweist. Diese liefert allerdings in ihrer Originalversion spürbar schlechtere Ergebnisse und muss manuell angepasst und optimiert werden. Durch die weniger intuitive Web-Oberfläche von Microsoft Azure wird der Einstieg in die Plattform generell zusätzlich erschwert.

Problemlos hingegen erfolgte der Upload der Trainings- und Testdaten. Azure unterstützt hier diverse strukturierte Formate wie JSON, TSV oder das hier bei den Daten bereits vorliegende CSV-Format. Außerdem müssen die Daten nicht erst über eine weitere externe Oberfläche in Datei-Stores eingepflegt werden, sondern können direkt in der ML-Umgebung in einen solchen

hochgeladen werden. Die Daten werden den Anwender*innen nach dem Upload in einer Übersicht präsentiert, in der sich noch diverse Einstellungen vornehmen lassen. Weiterhin können einzelne Spalten ausgelassen und die jeweiligen Datentypen festgelegt werden.

Beim Anlegen eines neuen AutoML-Experiments konnte der Datensatz anschließend als Trainings- und Validierungsinput ausgewählt werden. Da im Gegensatz zum Use Case mit tabellarischen Daten hier keine Kreuzvalidierung vorgesehen war, konnte auch das in 5.4.1 erwähnte Deep-Learning-Feature für die Textklassifikation aktiviert werden. Weiter spezifizieren ließ sich hier die Größe des Validierungssets als Prozentangabe des gesamten Trainingssets, nicht jedoch, welche Daten im Detail hierfür verwendet werden sollten.

Von essenzieller Bedeutung sind die sogenannten »Featurisierungseinstellungen«, die standardmäßig ausgeblendet sind und sich erst über einen unscheinbaren Link öffnen. Hier musste die Text-Spalte des Datensatzes explizit als Feature-Typ »Text« deklariert werden, um die für längere Fließtexte vorgesehene Feature-Aufbereitung zu inkludieren. In Microsoft Azure werden außerdem weitere und vergleichsweise viele Kontrollparameter aufgelistet, mit denen Anwender*innen den Trainingsablauf steuern können, darunter maximale Trainingszeit und Optionen für Erklärbarkeit des besten Modells. Die weiteren Werte wurden jedoch für dieses Experiment in ihren Standardeinstellungen belassen. Der darauffolgende AutoML-Trainingsdurchlauf dauerte 97 Minuten und lieferte analog zu 5.4.1 eine Liste an trainierten Modellen. Für das beste Modell werden weitere Informationen zur Erklärbarkeit und zur Bedeutung einzelner Inputfeatures angeboten.

Für die Evaluierung des finalen besten Modells auf dem Testdatensatz waren generell zwei Methoden möglich: Bereitstellung des Modells als Azure Service oder Download des Modells und manuelle Auswertung über Code. Obwohl die Anwender*innen aus dem Use Case generell die Nutzung grafischer Oberflächen bevorzugen, haben sie sich für den zweiten Ansatz entschieden, da die Vorhersage auf dem Testdatenset hier als einmaliger Task geplant war, der keine längere Bereitstellung erfordert. Abgesehen davon wurde die Möglichkeit zum direkten Download eines Modells und dessen anschließende Offlinenutzung lediglich bei Microsoft Azure entdeckt. Da dieser aber für andere Anwendungsfälle in datenschutztechnischer Hinsicht einen großen Vorteil darstellen könnte, wurde hier dieser Weg gewählt. Des Weiteren beinhaltet der Download eines Modells hier bereits direkt für dieses Modell angepassten Beispielcode in Python, der sich nach Installation der Azure-Bibliothek für Python problemlos für lokale Vorhersagen auf dem Modell verwenden ließ.

Für die Aggregation der vorhergesagten Labels zu der geforderten Accuracy-Metrik musste allerdings erneut auf ein weiteres selbst erstelltes Python-Skript zurückgegriffen werden. Abbildung 29 zeigt die Ergebnisübersicht nach Abschluss des AutoML-Trainings mit den besten Modellen und deren Performance auf dem Validierungsdatensatz.

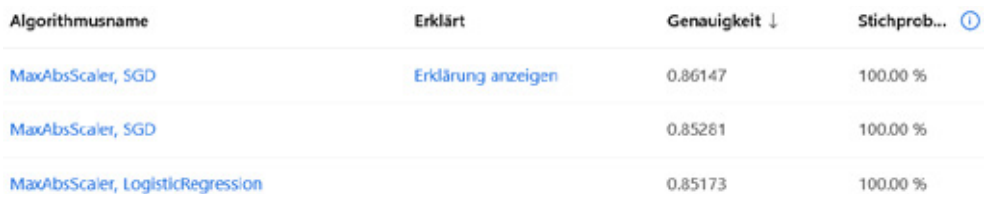


Abbildung 29: Microsoft-Azure-Ergebnisübersicht für den Use Case Textklassifikation[79].

Zusammenfassung: Die von den einzelnen Plattformen erreichten Accuracy Scores für den Testdatensatz sind in Tabelle 12 aufgelistet. Amazon liegt hier mit einigen Prozentpunkten vorn. Insgesamt fällt auf, dass die über alle Plattformen erreichten Werte denen des tabellarischen Datensatzes stark ähneln bzw. im selben Bereich liegen, obwohl mit dem Fließtextdatensatz eine (vermeintlich) ganz andere Problemstellung gewählt wurde.

Tabelle 12: Erreichte Performance der einzelnen Plattformen auf dem Testdatensatz für Textklassifikation.

Dienst	Accuracy
Amazon	0.87
Google	0.86
IBM	0.84
Microsoft	0.84

Tabelle 13 bietet eine Übersicht über die verschiedenen Plattformen hinsichtlich der in 5.3.2 definierten Use-Case-spezifischen Kriterien.

Während Microsoft über das flexibelste Eingangsdatenformat verfügt, geben Amazon und Google klare Definitionen für die CSV-Dateien vor und erlauben keinerlei Abweichung. IBM beschränkt hierbei zusätzlich noch die maximale Länge der Texte. Obwohl angesichts der Ergebnisqualität scheinbar alle Plattformen mit deutschen Fließtexten umgehen können, lässt lediglich Amazon eine explizite Festlegung auf deutsche Texte bei der Trainingskonfiguration zu und verwendet dementsprechend vermutlich ein spezielles deutsches Sprachmodell. Eine komplette Abwicklung des gesamten Use Cases über die Web-Oberfläche konnte (abgesehen von der Konvertierung der Eingangsdaten) am ehesten Google bieten. Hierzu sei angemerkt, dass bei den anderen Plattformen ebenso weitere Dienste zur Verfügung stehen, mit denen sich sowohl Inputdaten als auch Ausgangsmetriken im geforderten Format mit Sicherheit hätten

79 *Quelle: Microsoft Azure Machine Learning.*

erzeugen lassen. Hierfür hätten allerdings wiederum weitere Dienste eingerichtet und evtl. kostenpflichtig lizensiert werden müssen, während bei Google die Durchführung innerhalb eines festen Produkts und ohne die manuelle Integration weiterer Dienste erfolgen konnte.

Bezüglich automatischer Modelloptimierung bieten alle Plattformen die Möglichkeit, eine vordefinierte Pipeline mit Textinput zu nutzen, die auch von fachfremden Nutzer*innen bedient werden kann. Am meisten Freiheit bzw. Kontrolle über den Trainingsablauf bietet hierbei Microsoft, wo abgesehen von der erwähnten Einstellung der Featurisierung für Textinput allerdings auch keine weiteren Änderungen seitens der Anwender*innen erfolgen müssen, um ein gutes Ergebnis zu erzielen. Bei der Qualität der Ergebnisse führt wie bereits in Tabelle 12 aufgezeigt Amazon mit dem höchsten Accuracy-Wert auf dem Testdatensatz. Dies mag, muss aber nicht zwangsweise mit der vorhandenen Möglichkeit zur Spezialisierung auf deutsche Texte zusammenhängen.

Hinsichtlich der Benutzerfreundlichkeit punktet wiederum vor allem Google. Während es bei IBM und Microsoft zunächst vergleichsweise schwerer fiel, die für Textklassifikation geeigneten Dienste zu identifizieren bzw. einzurichten und zu integrieren, bieten Amazon und Google gut dokumentierte Ready-to-use-Produkte an, die durch ihre simple Menüführung den Einstieg und die Wahl der richtigen Einstellung erleichtern. Im Fehlerfall werden Anwender*innen bei Google in Form von aussagekräftigen Hinweisen die beste Hilfestellung geboten, während die anderen Produkte während der Experimente den Trainingsvorgang lediglich mit einem nichtssagenden Fehlercode abbrachen.

Mit dem Angebot, ein trainiertes Modell entweder als Service bereitzustellen oder aber herunterzuladen und offline zu nutzen, bietet Microsoft momentan die beste Möglichkeit zur Integration der KI-Komponenten in andere Projekte und Umgebungen. Amazon stellt das trainierte Modell immerhin direkt als Service zur Verfügung, sodass weitere Vorhersagen und Auswertungen sowohl über die Web-Oberfläche als auch über API erfolgen können. Dies ist auch bei IBM der Fall, hier allerdings wieder mit der Einschränkung, dass über die Oberfläche höchstens 30 Texte auf einmal verarbeitet werden können, für größere Auswertungen muss die API verwendet werden. Das Schlusslicht bildet hier Google, wo ein einmal trainiertes und ausgewertetes Modell für weitere Nutzung (kostenpflichtig) als Service bereitgestellt werden muss, damit überhaupt weitere Vorhersagen darauf ausgeführt werden können.

	Amazon Web Services	Google Cloud Platform	IBM Watson	Microsoft Azure
Datenformat	2	2	1	3
Deutscher Text	4	3	3	3
Automatisierung	3	3	3	4
Qualität	3	3	3	3
Benutzerfreundlichkeit	3	4	2	2
Wiederverwendbarkeit	3	2	2	4
Erforderliche Programmierkenntnisse	3	4	2	3

Tabelle 13: Übersicht Erfüllung der Use-Case-Kriterien für Textklassifikation bei den einzelnen Plattformen[80].

5.4.3 Bildklassifikation

Allgemeine Beschreibung des Versuchsablaufs

Bei der Umsetzung des Versuchs soll mit möglichst geringem Aufwand ein Model generiert werden, das imstande ist, Bilder des bereits in 5.2 vorgestellten Datensatzes zu klassifizieren. Im Regelfall (sofern nicht im Folgenden anders beschrieben) wurden alle Tools in ihrer voreingestellten Standardkonfiguration verwendet. Als Metrik dient Accuracy sowie Recall. Im Folgenden werden positive wie negative Auffälligkeiten bei der Durchführung des Experiments beschrieben.

AWS Amazon Rekognition ist ein Service, der es Nutzer*innen ermöglicht, Objekte in Bildern sowie Videos zu erkennen. Hierbei können bereits von AWS trainierte neuronale Netze verwendet werden, die unterschiedliche Objekte in Bildern und Videos erkennen (Szenenerkennung, Gesichtserkennung, Texterkennung).

Der für diesen Use Case relevante Service ist Amazon Rekognition Custom Labels, denn er erlaubt es Anwender*innen, Modelle zu trainieren und dabei eigene Trainingsdaten sowie Labels zu verwenden. Der Service ermöglicht Klassifizierung sowie Objekterkennung. Durch die Vergabe eines Titels wird ein neues Projekt[81] erzeugt. Anzumerken ist, dass die Benutzeroberfläche größtenteils in englischer Sprache verfasst ist. Bevor das Training gestartet werden kann, muss ein Dataset erzeugt werden. Hierzu ist es notwendig, die Trainingsdaten in einen S3-Bucket (AWS Simple Storage Service) hochzuladen. Die Benutzeroberfläche von Amazon Rekognition erlaubt den direkten Upload von Bildern, ohne in die S3-Serviceoberfläche wechseln zu müssen. Jedoch ist es nur möglich, maximal 30 Bilder in einem Zug hochzuladen. Außerdem

80 *0 = nicht erfüllt, 1 = ansatzweise erfüllt, 2 = teilweise erfüllt, 3 = größtenteils erfüllt, 4 = vollständig erfüllt.*

81 *https://eu-west-1.console.aws.amazon.com/rekognition/custom-labels#/create/project*

müssen die Dimensionen von mindestens 64 x 64 Pixeln sowie maximal 4096 x 4096 Pixeln eingehalten werden und die Bilder im JPG- oder PNG-Format vorliegen. Aus diesem Grund empfehlen wir, einen S3-Bucket zu erstellen und die Bilder mithilfe der AWS-Befehlszeilen-Schnittstelle (AWS-CLI) hochzuladen. Für jede Klasse sollte hierbei ein eigener Ordner erzeugt werden, da dies das Data Labeling beim Erstellen des Datasets erheblich erleichtert. Da es hierbei diverse Lösungsmöglichkeiten gibt, könnten unerfahrene Nutzer*innen Probleme bekommen.

Sobald der S3-Bucket erstellt und befüllt ist, kann das Dataset erstellt werden. Hierbei wählen wir den Punkt »Import images from Amazon S3 bucket« an und geben den Pfad unseres S3-Buckets an. Da die Bilder bereits in Ordnern nach Klassen sortiert wurden, wird ein Haken bei »Automatically attach a label to my images based on the folder they are stored in« gesetzt. Dies erspart es, die Bilder manuell im Nachgang einzelnen Klassen zuzuordnen. Sollte dies dennoch notwendig sein, können Bilder in der Ansicht des Datasets gelabelt werden. Nachdem das Projekt und das Dataset erzeugt wurden, kann das Training gestartet werden. Bei Bedarf kann hierbei explizit ein Testdatensatz angegeben werden. Wir empfehlen jedoch die Option »Split training dataset«, wodurch 20 Prozent des Datasets als Test-Dataset verwendet werden. Nachdem das Training beendet ist kann die Genauigkeit (F1-Score, Average precision, Overall recall) des Modells eingesehen werden.

Abbildung 30: Amazon Rekognition – Vorhersagegüte des trainierten Modells[82].

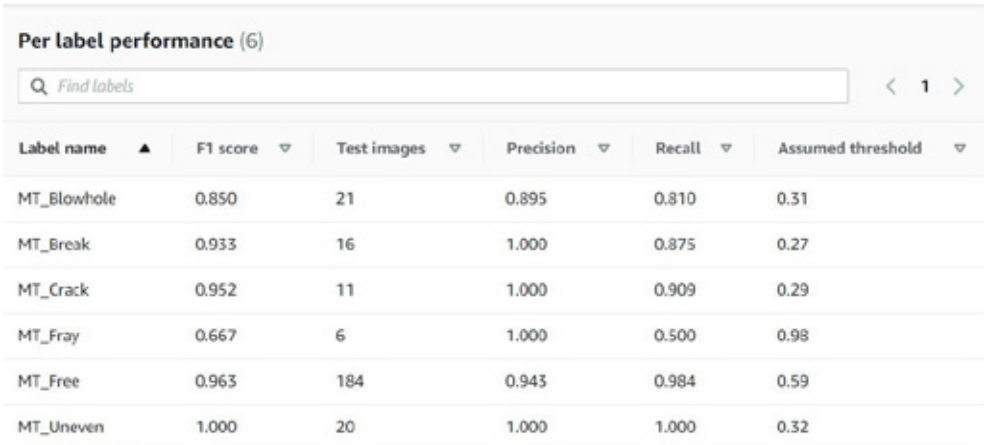


Abbildung 31: Amazon Rekognition – Vorhersagegüte des Modells für einzelne Klassen[83].

82 *Quelle: Amazon Rekognition.*

83 *Quelle: Amazon Rekognition.*

Soll das Modell per REST-API zur Verfügung gestellt werden, muss das Modell per AWS-CLI gestartet und gegebenenfalls gestoppt werden. Amazon Rekognition bietet leider keinerlei Möglichkeit, das erzeugte Modell zu exportieren.

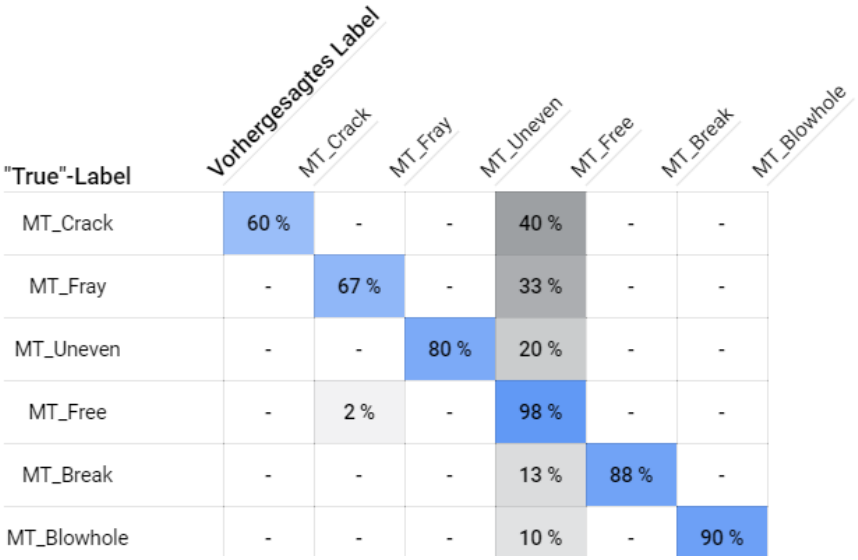
Google Cloud Platform Google bietet für das Erstellen eigener Modelle zur Bilderkennung den Service AutoML Vision an. Er ist auf der Google Cloud Platform über die Suchleiste schnell zu finden. Im ersten Schritt ist es notwendig, ein Dataset anzulegen. Hierbei hat man die Wahl zwischen Klassifizierung mit einem Label, Klassifizierung mit mehreren Labels oder Objekterkennung. In der nächsten Ansicht muss der Datensatz von den Nutzer*innen hochgeladen werden. Es ist sinnvoll, alle Bilder in einer ZIP-Datei mit Unterordnern (Klassen) vorzubereiten, dadurch sind die Bilder bereits nach dem Import den korrekten Klassen zugeordnet. Des Weiteren muss ein Speicher angelegt werden. Dieser Schritt kann aus derselben Ansicht durchgeführt werden. Nachdem die Bilder hochgeladen und importiert sind, wird die Verteilung über die einzelnen Klassen hinweg dargestellt und den Nutzer*innen wird die Möglichkeit gegeben, die Bilder manuell anderen Klassen zuzuordnen.

Beim Starten des Trainings steht Nutzer*innen die Wahl offen zwischen »Cloud hosted« und »Edge«. Um einen späteren Export des Modells zu ermöglichen, wird hier »Edge« ausgewählt. Im nächsten Schritt haben Nutzer*innen die Wahl zwischen drei Optionen, wobei sie sich zwischen der Verarbeitungsdauer einer Klassifizierung und der Genauigkeit des Vorhersagemodells entscheiden müssen. Im Experiment wurde die höchste Genauigkeit ausgewählt. Nutzer*innen können außerdem angeben, wie lange das Model höchstens trainiert werden soll. AutoML Vision gibt an dieser Stelle ebenfalls eine Empfehlung für eine Mindesttrainingsdauer, die abhängig vom genutzten Datensatz ist.

Sobald das Training beendet ist, werden die Nutzer*innen per E-Mail benachrichtigt. Die Oberfläche gibt ausreichend Informationen über die Leistung des Modells. Unter dem Bewerten-Reiter sind detaillierte Kennzahlen einzusehen.

Abbildung 32: GCP AutoML Vision – Statistiken zum trainierten Modell[84].

84 Quelle: GCP AutoML Vision.



"True"-Label	MT_Crack	MT_Fray	MT_Uneven	MT_Free	MT_Break	MT_Blowhole
MT_Crack	60 %	-	-	40 %	-	-
MT_Fray	-	67 %	-	33 %	-	-
MT_Uneven	-	-	80 %	20 %	-	-
MT_Free	-	2 %	-	98 %	-	-
MT_Break	-	-	-	13 %	88 %	-
MT_Blowhole	-	-	-	10 %	-	90 %

Abbildung 33: GCP AutoML Vision – Confusion-Matrix des trainierten Modells[85].

Unter dem Reiter »Test und Nutzung« kann das Modell in der GCP bereitgestellt werden. Die hierbei geltenden Preise sind über einen Link aufrufbar. Da bei der Erstellung des Modells die Option »Edge« gewählt wurde, ist das Modell in diversen Formaten exportierbar und kann somit auf eigenen Geräten installiert werden.

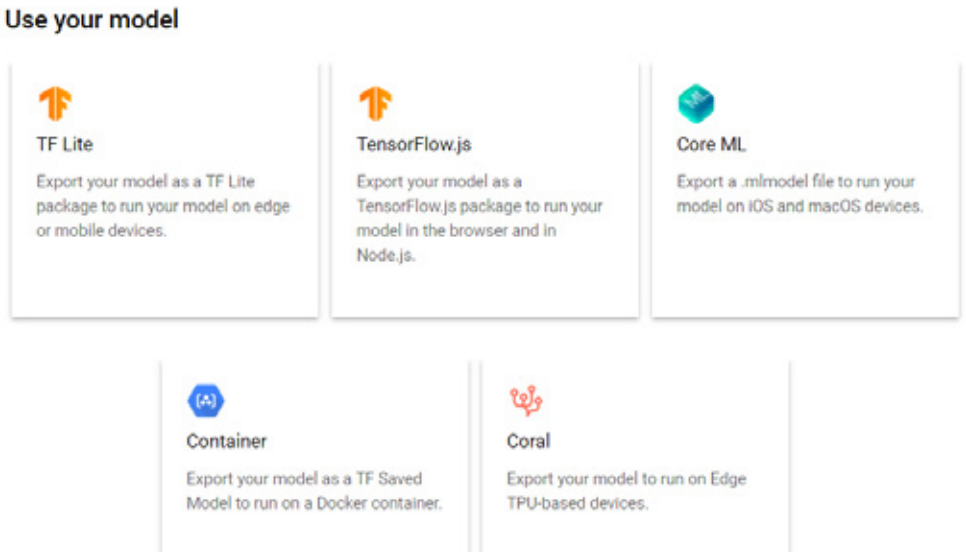


Abbildung 34: GCP AutoML Vision – Möglichkeiten, das trainierte Modell zu exportieren[86].

85 Quelle: GCP AutoML Vision.

86 Quelle: GCP AutoML Vision.

IBM Die IBM Cloud bietet für Bildklassifizierung und Objekterkennung den »Visual Recognition Service«. Um diesen zu nutzen, muss zunächst ein neues Projekt im Watson Studio Service erzeugt werden. Dabei muss ein Speicherservice angegeben werden, der an dieser Stelle erzeugt werden kann. In der Projektansicht wird dem Projekt ein Visual-Recognition-Modell hinzugefügt. Wir erhalten an dieser Stelle eine Benachrichtigung darüber, dass mit dem Projekt ein »Watson Visual Recognition Service« assoziiert sein muss. Durch einen Link gelangt man zu der Oberfläche, auf der man den Service erzeugen und verknüpfen kann. Wie auch die anderen Plattformen bietet die IBM-Lösung die Möglichkeit sowohl zur Bildklassifizierung und Objekterkennung als auch zur Benutzung vortrainierter Modelle zur Erkennung von alltäglichen Gegenständen. Die Oberfläche ist an einigen Stellen in deutscher und an anderen in englischer Sprache verfasst. Der bis hierhin beschriebene Prozess wurde als wenig intuitiv wahrgenommen, da Nutzer*innen nicht auf den ersten Blick ersichtlich wird, was der nächste notwendige Schritt ist.

Durch die Auswahl von »Classify Model« gelangen wir in die Ansicht des Vision Recognition Service. Dieser gestaltet sich sehr übersichtlich. Der Upload erlaubt JPG-, PNG- sowie ZIP-Dateien. Wir empfehlen, die Dateien der einzelnen Klassen in separaten ZIP-Dateien vorzubereiten und hochzuladen, um das Labeling zu vereinfachen. IBM Vision Recognition ist die einzige Lösung in unserem Vergleich, die es erlaubt, Bilder einer Negative-Klasse zuzuordnen. Diese soll Beispielbilder enthalten, die kein Merkmal der sonstigen Klassen aufweisen. In unserem Aufbau werden die Bilder, die sonst der Klasse Free zugeordnet werden, nun der Negative-Klasse zugewiesen. Nach dem erfolgreichen Upload aller Bilder wird das Training gestartet.

Abbildung 35: IBM Visual Recognition – hochgeladener Datensatz[87].

Nachdem das Training beendet ist, sind Informationen zum erzeugten Modell zu sehen, es werden jedoch keine Aussagen über dessen Leistung gemacht. Die Oberfläche bietet die Möglichkeit, Bilder aus dem Validierungsdatensatz zur Überprüfung des Modells hochzuladen.

87 *Quelle: IBM Visual Recognition.*

Das Modell kann über eine REST-Schnittstelle bereitgestellt werden. IBM bietet an dieser Stelle außerdem an, das Modell im von Apple entwickelten Format für die Core-ML-Bibliothek herunterzuladen.

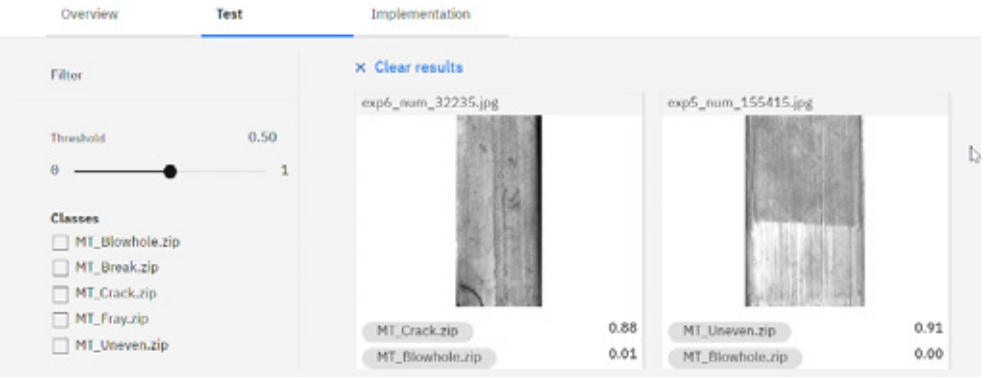


Abbildung 36: IBM Visual Recognition – Testen des Modells über die grafische Oberfläche[88].

Microsoft Azure Zum Erstellen von individuellen Bilderkennungswerkzeugen bietet Microsoft den Custom-Vision-Service, der zur Gruppe der Cognitive Services der Azure-Plattform gehört. Unterstützt wird Bildklassifizierung (Single- und Multi-Label) sowie Objekterkennung. Der Dienst ist über eine eigene URL[89] zu erreichen und benötigt zur Nutzung einen Azure-Account. Die grafische Oberfläche ist nur in englischer Sprache verfügbar, jedoch steht die Dokumentation in deutscher sowie englischer Sprache zu Verfügung[90]. Beim Erstellen eines Projekts muss eine »Domäne« gewählt werden (in unserem Fall »General«). Wenn ein späterer Export des Modells gewünscht ist, so ist die Auswahl »General (compact)« zu wählen. Wir wählen »Classification« und »Multiclass (Single tag per image)« aus. Bei der Auswahl der »Export Capabilities« empfehlen wir »Basic Platforms« auszuwählen, um den Export zu geläufigen Technologien wie TensorFlow, ONNX, Docker u. a. zu ermöglichen. Bei der ersten Nutzung müssen Azure-Ressourcen erzeugt werden, was simpel im selben Fenster zu bewerkstelligen ist. Microsofts Oberfläche erweist sich als sehr intuitiv, sodass selbst Nutzer*innen, die nicht besonders technikaffin sind, sich schnell zurechtfinden sollten. Nachdem das Projekt erstellt ist, müssen die Bilddateien hochgeladen werden. Beim Upload der Bilddateien sind mehrere Dateien anwählbar und können gleichzeitig mit einem Label versehen werden. Unterstützt werden Dateien in den Formaten JPG, PNG sowie BMP mit einer maximalen Dateigröße von 6 MB. Die Label können durch die intuitive Oberfläche auch im Nachgang hinzugefügt oder geändert werden. Beim Starten des Trainings ergeben sich zwei Möglichkeiten, »Quick Training« sowie »Advanced Training«. Bei der Wahl des »Advanced Training« bietet sich die Möglichkeit, die Dauer des Trai-

88 *Quelle: IBM Visual Recognition.*

89 *https://www.customvision.ai*

90 *https://docs.microsoft.com/de-de/azure/*

nings (1–24 Stunden) auszuwählen. Sollten die Trainingserfolge früher stagnieren, wird das Training frühzeitig beendet, sodass bei der Nutzung keine unnötigen Kosten erzeugt werden. Nutzer*innen können sich über die Fertigstellung des Trainings per Mail benachrichtigen lassen. Die Leistung des trainierten Modells kann nach der Trainingsphase im »Performance«-Reiter eingesehen werden, wo die Kennzahlen Precision, Recall und AP (Average Precision) angegeben werden. Durch den »Probability Threshold«-Regler wird sichtbar, wie sich Precision und Recall in Abhängigkeit zum Threshold ändern.

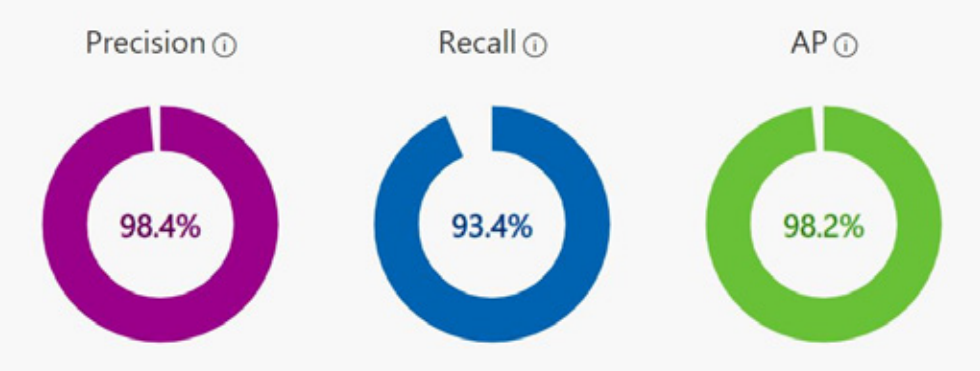


Abbildung 37: Azure Custom Vision – Vorhersagegüte des trainierten Modells[91].

Im selben Reiter ist auch »Performance per Tag« zu sehen, dies gibt an, wie die einzelnen Tags bzw. Klassen verteilt sind und erkannt werden. Hierbei ist positiv anzumerken, dass Azure Warnzeichen gibt, wenn das Dataset beispielsweise zu viele oder zu wenige Bilder einer Klasse enthält und somit nicht balanciert ist. Warum diese Information nicht bereits vor dem Start des Trainings gegeben wird, ist unklar. Nutzer*innen sollten in diesem Fall überlegen, ob nicht weitere Daten eingespeist werden können, oder wie eine gleichmäßige Verteilung erzielt werden kann. In unserem Fall ist die Free-Klasse logischerweise überrepräsentiert, da es sich hierbei um Bilder ohne jeglichen Fehler handelt. Es werden mindestens 50 Bilder je Klasse empfohlen.

91 *Quelle: Microsoft Azure Custom Vision.*

Tag	Precision	Recall	A.P.	Image count
MT_Uneven	100.0%	95.0%	99.8%	97
MT_Crack	100.0%	100.0%	100.0%	52
MT_Blowhole	100.0%	90.5%	95.2%	104
MT_Free	97.9%	99.5%	99.7%	926
MT_Break	93.3%	87.5%	87.6%	77
MT_Fray	85.7%	100.0%	83.7%	29

Abbildung 38: Azure Custom Vision – Hinweis zur Verteilung der einzelnen Klassen[92].

Um einen schnellen Test des Modells zu ermöglichen, bietet die Oberfläche unter »Quick Test« die Möglichkeit, ein Bild aus dem Validierungsdatenset hochzuladen und zu klassifizieren. Wie auch die anderen Plattformen bietet Azure die Möglichkeit, das Modell über eine REST-Schnittstelle bereitzustellen. Das Modell kann, wie eingangs erwähnt, für verschiedene Technologien exportiert werden, sodass es auf eigener Hardware genutzt werden kann.

Abbildung 39: Azure Custom Vision Service – Möglichkeiten für den Export des trainierten Modells[93].

92 *Quelle: Microsoft Azure Custom Vision.*

93 *Quelle: Microsoft Azure Custom Vision.*

Zusammenfassung der Ergebnisse

Tabelle 14: Bewertung der Anbieter für den Anwendungsfall der Bilderkennung.

	AWS	Google	IBM	Microsoft
Verfügbarkeit und automatische Wahl geeigneter Schritte und Algorithmen	4	4	4	4
Benutzerfreundlichkeit während Training und Evaluation	3	4	3	4
Leichte Wiederverwendbarkeit/ Integrierbarkeit der trainierten Modelle	0-Modellexport nicht möglich	4-Modellexport möglich für mehrere Technologien	2-Modellexport möglich für CoreML	4-Modellexport möglich für mehrere Technologien
Erforderliche Programmierkenntnisse	4	4	4	4

Folgende Tabelle zeigt die Leistung der erzeugten Modelle an. Es sei erneut angemerkt, dass die Ergebnisse, wenn nicht anders beschrieben, mit den Standardeinstellungen der jeweiligen Werkzeuge erzielt wurden.

Tabelle 15: Leistung der trainierten Modelle für die Bildklassifizierung.

Dienst	Accuracy	Recall
Amazon	0.97	0.84
Google	0.94	0.92
IBM	k. A.	k. A.
Microsoft	0.98	0.93

5.4.4 Zeitreihenanalyse

Ausgehend von der Anwendungsfallbeschreibung zur Zeitreihenanalyse in Abschnitt 5.2 wird deutlich, dass gerade kleine und mittelständische Betriebe (KMU) meist nicht ausreichend Mitarbeiterinnen und Mitarbeiter mit tiefer gehenden Kenntnissen im Bereich Data Science zur Verfügung haben. Es wird daher oftmals versucht, diese Lücke durch IT-affine Mitarbeiter*innen zu schließen, die sich ihre Kenntnisse meist neben ihrer eigentlichen Tätigkeit angeeignet haben und versuchen, diese Aufgaben zu übernehmen. In diesem Fall ist es daher oftmals ratsam, die von den Plattformen angebotenen AutoML-Werkzeuge zu verwenden, um die Mitarbeiter*innen so gut wie möglich zu entlasten.

AWS AWS bietet mit SageMaker Autopilot seinen Anwendern die Möglichkeit, Regressionsprobleme, wie den in diesem Anwendungsfall betrachteten Werkzeugverschleiß, automatisiert berechnen zu lassen. Bevor Nutzer*innen allerdings zum Modelltraining und dessen Validierung kommen, müssen die entsprechenden Trainingsdaten bereitgestellt werden. Dies erfolgt zum Beispiel, wie bei AWS üblich, über den Upload in ein dafür bereitgestelltes S3-Bucket.

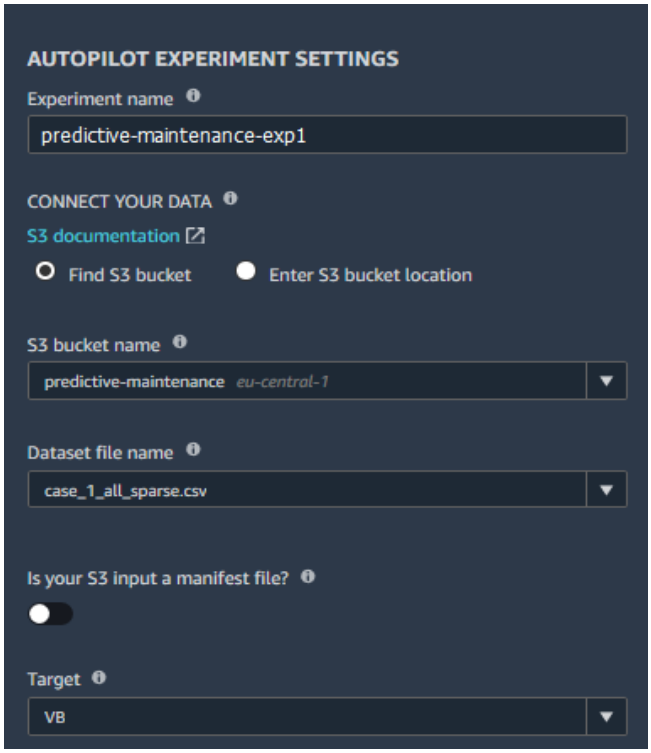


Abbildung 40: Konfigurationsseite des Experiments[94].

Wenn der zu verwendende Datensatz in der Plattform bereitgestellt wurde, steht der Erstellung eines neuen AutoML-Experiments nichts mehr im Weg. Dazu müssen Benutzer*innen SageMaker Studio öffnen und dort ein neues Autopilot-Experiment anlegen. Die Eingabe eines aussagekräftigen Namens erlaubt die einfache Unterscheidung einer Vielzahl von Experimenten. Nach Wahl des S3-Buckets, der den Datensatz beinhaltet, haben die Nutzer*innen die Möglichkeit, diesen im Feld »Dataset file name« auszuwählen. Es erfolgt eine Vorabanalyse des Datensatzes und die anschließende Möglichkeit, das Experiment im Detail zu konfigurieren.

94 *Quelle: AWS SageMaker Autopilot.*

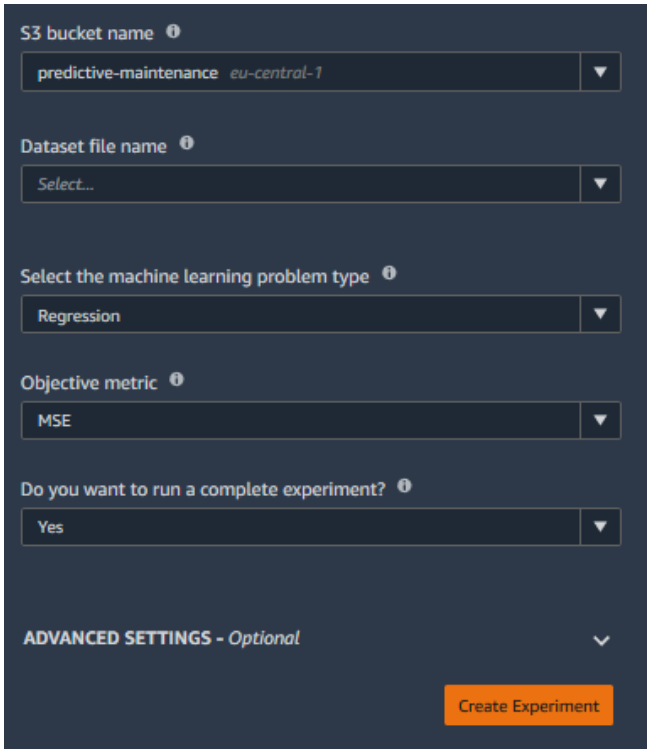


Abbildung 41: Detaillierte Konfiguration des Experiments[95].

Die berechneten Ergebnisse werden wieder in ein dafür anzugebendes S3-Bucket gespeichert, welches unter dem »S3 bucket name« gesetzt werden kann. Es ist möglich, unterschiedliche Buckets zu verwenden, wenn man Datenhaltung und berechnete Ergebnisse voneinander trennen möchte. Andernfalls kann hier auch der schon vorher verwendete Bucket gewählt werden. Durch die Wahl der Machine-Learning-Problemstellung legen die Benutzer*innen den Fokus des Experiments und die zu berechnenden Metriken fest. Da dieser Anwendungsfall eine möglichst genaue Vorhersage des Verschleißwertes VB anstrebt, was ein Regressionsproblem darstellt, wird eben diese Einstellung gewählt. Die verwendete MSE-Metrik ist in diesem Experiment allerdings nicht änderbar. Somit kann abschließend das Experiment durch Klick auf »Create Experiment« erstellt werden.

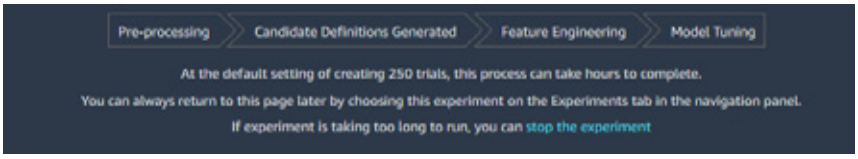


Abbildung 42: Übersicht der AutoML-Pipeline-Schritte und des aktuellen Bearbeitungsstands[96].

95 *Quelle: AWS SageMaker Autopilot.*

96 *Quelle: AWS SageMaker Autopilot.*

Die Oberfläche zeigt daraufhin die Reihenfolge der AutoML-Phasen an, die während des Experiments abgearbeitet werden. Dazu gehören unter anderem die Vorverarbeitung der Daten, das Erstellen möglicher Modelle, die sich für diese Problemstellung eignen, Feature Engineering sowie das Model Tuning. Nach Abschluss aller Phasen erhalten die Nutzer*innen eine Übersicht über alle abgearbeiteten Schritte während des Experiments.

Charts | Metrics | Parameters | Artifacts | AWS settings | Debugger

Name	Minimum	Maximum	Final value
ObjectiveMetric	0.00691	0.05794	0.00691
train:rmse	0.08304	0.24073	0.08304
train:mse	0.0069	0.05795	0.0069
validation:mse	0.00691	0.05794	0.00691
validation:rmse	0.08314	0.24071	0.08314

Abbildung 43: Zusammenfassung der Ergebnisse des Experiments[97].

Aus der Liste der »Trial Components« kann nach Abschluss der Berechnungen der beste »Tuning Job« ausgewählt und dessen Metriken eingesehen werden. Für diesen Anwendungsfall und dieses Experiment ergibt sich hierfür ein RMSE von 0.07066 und ein MSE von 0.00493, wobei niedrigere Werte ein besseres Modell und somit eine bessere Approximation des Verschleißwerts VE darstellen. Das so berechnete Modell kann im Anschluss daran als Dienst bereitgestellt werden, um es zum Beispiel zur Echtzeitvorhersage des Verschleißwerts anhand neuer Sensormesswerte zu verwenden.

Google Cloud Platform Die Google-Cloud-Plattform ermöglicht das automatisierte Training von Modellen durch die Anwendung von AutoML. Dabei stehen unterschiedliche Varianten zur Auswahl. In dem hier betrachteten Anwendungsfall werden tabellarische Daten verarbeitet, weswegen »AutoML Tables« verwendet wird. Nach Anlage eines neuen Projekts können Daten aus unterschiedlichen Datenquellen importiert werden. Dazu zählen unter anderem CSV-Dateien, die in der Cloud Storage gelagert sind oder sich auf dem Computer befinden. Der Import nimmt eine gewisse Zeit in Anspruch, wobei den Benutzer*innen keine Abschätzung oder Zeitanzeige geboten wird. Nach erfolgreichem Import werden die Nutzer*innen per E-Mail benachrichtigt und können zum nächsten Schritt übergehen.

97 *Quelle: AWS SageMaker Autopilot.*

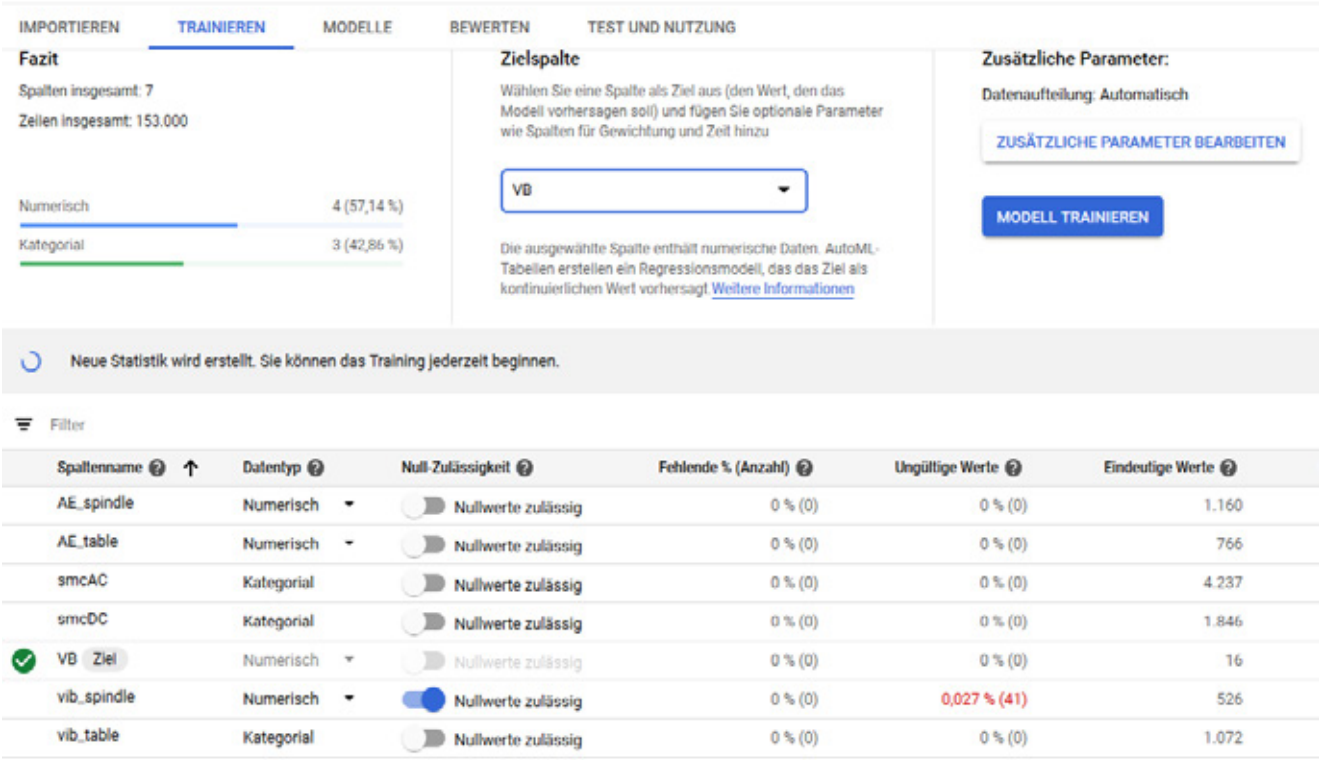


Abbildung 44: Konfiguration eines AutoML-Experiments[98].

In der Übersicht zum Modelltraining erhalten die Nutzer*innen eine Vielzahl an Informationen über die im Datensatz vorhandenen Spalten sowie deren Eigenschaften. Zusätzlich haben sie die Möglichkeit, eine Zielspalte auszuwählen, nach der das AutoML-Verfahren die Modelle entwickelt. Der zu erstellende Modelltyp wird durch die Plattform automatisiert festgelegt und richtet sich nach dem verwendeten Datentyp der Zielspalte. Für diesen Anwendungsfall wird der Verschleißwert VB anhand der restlichen Spalten vorhergesagt, wobei die Nutzer*innen die zu verwendenden Spalten gezielt an- und abwählen können, um so unterschiedliche Ansätze des Modelltrainings zu testen und miteinander vergleichen zu können. Nach Eingabe eines Rechenbudgets in Form der maximal verwendeten Anzahl an Knotenstunden kann mit dem Modelltraining begonnen werden.

Ähnlich wie beim Import von Datensätzen wird auch in der folgenden Übersichtsseite der Plattform keine Zeitangabe gemacht oder den Nutzer*innen zusätzliche Informationen über den Trainingsfortschritt bereitgestellt. Es ist somit nur schwer abschätzbar, wann das Modelltraining abgeschlossen ist, weswegen auch hier eine gesonderte E-Mail durch die Plattform versendet wird, die über den Abschluss der Berechnungen informiert.

98 *Quelle: GCP AutoML Tables.*

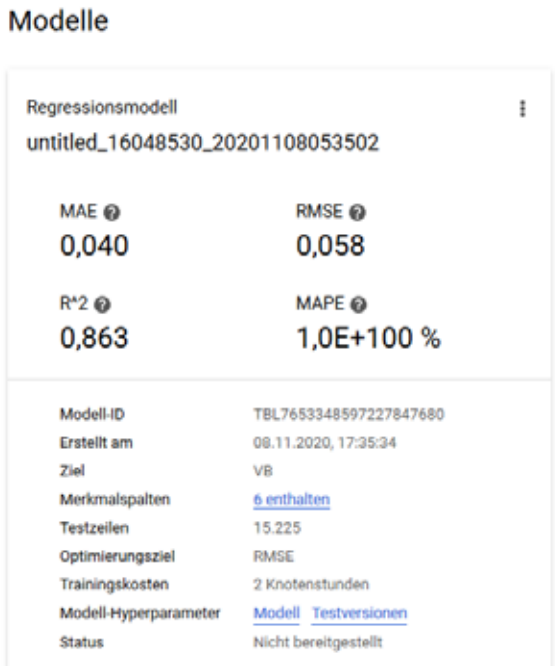


Abbildung 45: Übersicht der Leistungsdaten eines Modells anhand unterschiedlicher Metriken[45].

Nach Abschluss des Modelltrainings hat man die Möglichkeit, das berechnete Modell anhand unterschiedlicher Metriken wie MAE, RMSE, R^2 und MAPE zu bewerten. Es fällt auf, dass keine detaillierten Informationen über das Modell oder den verwendeten Algorithmus bereitgestellt werden, der dieses berechnet hat. Außerdem erhält man keine Informationen darüber, welche unterschiedlichen Verfahren getestet und zum Modelltraining verwendet wurden. Das trainierte Modell lässt sich lediglich anhand der vorliegenden Merkmalswichtigkeiten seiner Features überprüfen.

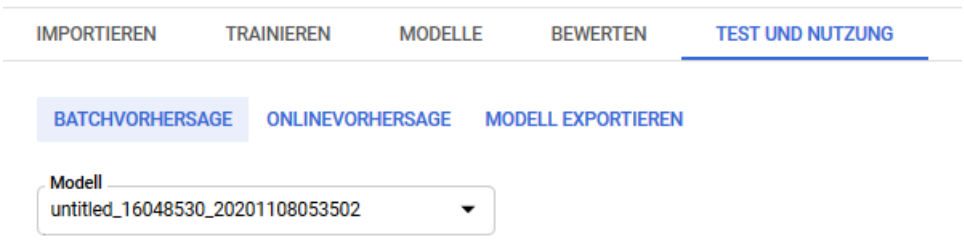


Abbildung 46: Übersichtsseite zur weiteren Verwendung des Modells nach erfolgtem Modelltraining[100].

99 *Quelle: GCP AutoML Tables.*

100 *Quelle: GCP AutoML Tables.*

Das erstellte Modell lässt sich unter dem Menüpunkt »Test und Nutzung« zur Batch- oder Onlinevorhersage verwenden, was für die Onlinevorhersage mit dem Bereitstellen des Modells einhergeht. Zusätzlich zur Bereitstellung haben die Nutzer*innen die Möglichkeit, ihr erstelltes Modell in ihr Google Cloud Storage zu exportieren. Leider ist das exportierte Modell und dessen Quellcode nicht einsehbar. Es liegt allerdings als TensorFlow-Paket vor, wodurch es zu einem späteren Zeitpunkt in einem Docker-Container ausgeführt werden kann.

IBM Watson Die IBM-Cloud-Plattform bietet Nutzer*innen über Watson Studio AutoML-Funktionalität an. Um darauf zugreifen zu können, müssen zuerst zwei Services zur Ressourcenliste hinzugefügt werden. Dies ist zum einen der »Watson Studio Service« und zum anderen eine »Machine Learning Ressource«. Hierbei ist darauf zu achten, dass sich beide Ressourcen in der gleichen Region befinden müssen, da diese sonst aus Datenschutzgründen nicht miteinander interagieren können.

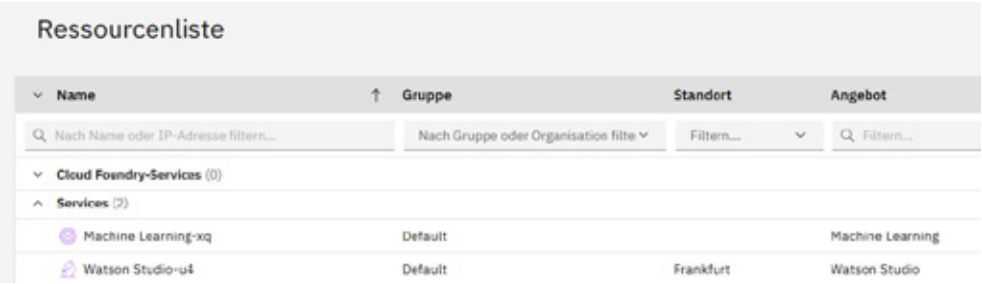


Abbildung 47: Übersicht einer Ressourcenliste[101].

Nach erfolgreicher Erstellung der Services kann die IBM-Watson-Studio-Oberfläche geöffnet werden, die als Ausgangspunkt für die Anwendung der AutoML-Komponente dient. Nutzer*innen können über die Oberfläche ein neues Projekt anlegen, das durch sogenannte »Assets«, wie der Datenquelle und einer AutoML-Komponente, erweitert werden muss, um den Anwendungsfall bearbeiten zu können.

101 Quelle: IBM Watson.

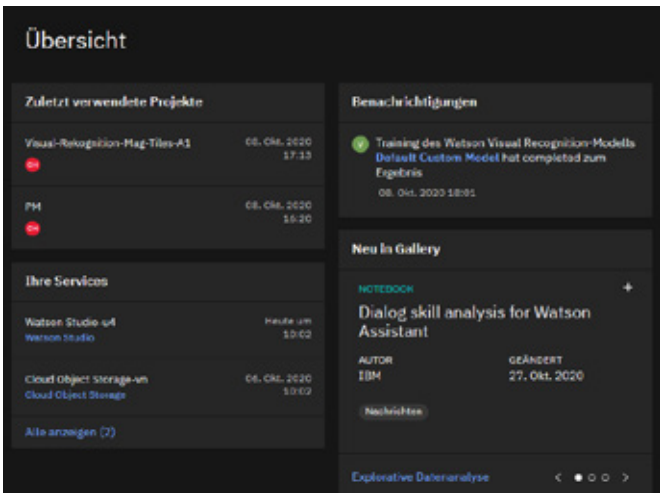


Abbildung 48: Inhalt des Watson Studio Dashboards[102].

Beim Hinzufügen von Datenquellen kann man den betreffenden Datensatz entweder per Drag-and-drop oder durch Auswahl in einem Datei-Explorer hinzufügen. Grundsätzlich kann eine beliebige Anzahl an Datenquellen zum Projekt hinzugefügt werden. Allerdings kann für die Durchführung des Modelltrainings selbst nur eine einzige Datenquelle ausgewählt werden, auch wenn mehrere dieser Datenquellen das gleiche Schema aufweisen. Sollte es notwendig sein, mehrere verteilte Datenquellen einzubinden, muss daraus im Vorfeld zuerst ein konsolidierter Datensatz erstellt werden.

Nach Auswahl der Datenquelle erfolgt eine Vorabanalyse des verwendeten Datenschemas durch die Plattform. Die abgeleiteten Namen werden den Benutzer*innenunter anderem zur Auswahl der Zielspalte für das AutoML-Experiment angeboten. Das Experiment selbst kann durch Festlegen weiterer Einstellungen noch genauer eingestellt werden. Dazu zählen unter anderem die Aufteilung der Daten in Trainings-, Test- und Validierungsdaten, die für das Modelltraining zu verwendenden Spalten, die Art der Vorhersage (Klassifikation, Regression), die zu verwendenden Metriken für den Modellvergleich sowie die Gesamtlaufzeit und Anzahl an Iterationen bei der Modelloptimierung.

102 Quelle: IBM Watson Studio.

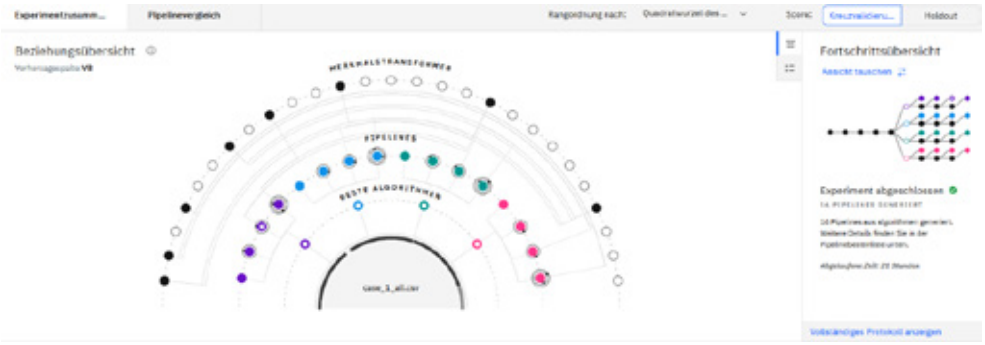


Abbildung 49: Übersicht des Modelltrainings und durchgeführter Schritte[103].

Die Durchführung des Experiments kann im Watson Studio Dashboard genau verfolgt werden. Somit sind immer der aktuelle Status und Gesamtfortschritt ersichtlich, wodurch ungefähr abgeschätzt werden kann, wann das Experiment abgeschlossen ist. Die erzeugten Modelle können im Anschluss entweder als Notebook oder Watson-ML-Modell gespeichert werden, wobei das Notebook den Python-Programmcode enthält, der dem Experiment zugrunde liegt. Das Watson-ML-Modell kann hingegen zur Bereitstellung als Dienst und zur damit verbundenen Verarbeitung neuer Daten weiterverwendet werden.

Microsoft Azure Die Microsoft-Azure-Plattform bietet Einsteiger*innen und IT-affinen Benutzer*innen zum Einstieg in den Bereich Machine Learning eine eigene AutoML-Komponente. Bevor auf diese zugegriffen werden kann, muss zuerst eine neue Machine-Learning-Ressource angelegt werden.

Zur Bearbeitung dieses Anwendungsfalls müssen allerdings keine besonderen Einstellungen vorgenommen werden und es kann mit den Standardeinstellungen fortgefahren werden. Nach dem Erstellen der Ressource folgt deren Bereitstellung. Im Anschluss daran kann die erstellte Ressource mit dem »Azure Machine Learning Studio« geöffnet und verwendet werden.

103 Quelle: IBM Watson Studio.

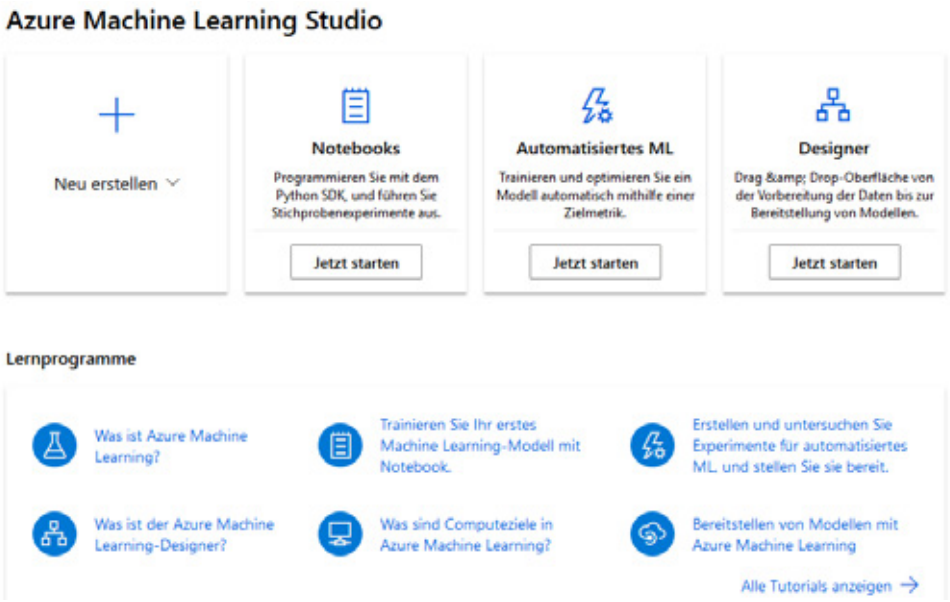


Abbildung 50: Übersichtsseite des Azure Machine Learning Studios[104].

Die Plattform bietet dem Benutzer*innen die Möglichkeit, aus verschiedenen Machine-Learning-Ansätzen zu wählen. Dabei werden neben dem klassischen Programmieransatz, wie zum Beispiel Notebooks, auch automatisierte Ansätze wie AutoML und ein graphischer Designer angeboten. Anhand der Anwendungsfallbeschreibung wird für dieses Experiment der AutoML-Ansatz zur Bearbeitung ausgewählt. Danach haben Benutzer*innen die Möglichkeit, spezifische Einstellungen durchzuführen. Dies umfasst in einem ersten Schritt die Wahl des zu verwendenden Datensatzes sowie zugehörige Parameter. Im nächsten Schritt erfolgt die Konfiguration des durchzuführenden Experiments, was unter anderem die Wahl der Zielspalte im Datensatz und das zu verwendende Rechencluster beinhaltet. Hier kann man sich dann zwischen verschiedenen Varianten entscheiden, die je nach Anwendungsfall unterschiedliche Rechenkapazität bereitstellen.

Im dritten und letzten Schritt können Nutzer*innen noch die Art der Aufgabe festlegen, wobei der Anwendungsfall die Vorhersage des Verschleißwerts VB zum Ziel hat, was eine Regressionsaufgabe darstellt. Ein wichtiger Punkt in diesem Schritt ist die Auswahl der Metrik, anhand derer der Modellvergleich erfolgt und wonach Modelloptimierungen durchgeführt werden. Für Regressionsaufgaben haben Nutzer*innen die Wahl zwischen der Spearman-Korrelation, RMSE, R^2 oder MAE.

104 *Quelle: Microsoft Azure Machine Learning Studio.*

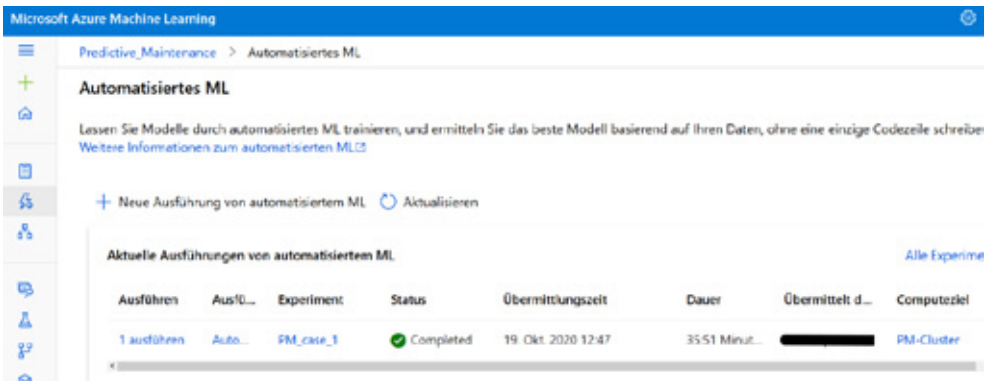


Abbildung 51: Übersichtsseite des laufenden Experiments[105].

Nach dem Erstellen des Experiments erfolgt die automatisierte Ausführung der AutoML-Aufgabe, die über das Azure Machine Learning Dashboard im Unterpunkt »Automatisiertes ML« kontinuierlich verfolgt werden kann.

Abbildung 52: Übersicht aller trainierten Modelle und Vergleich anhand der voreingestellten Metrik[106].

Nach Beendigung der automatisierten Bearbeitung erhalten die Nutzer*innen unter dem Punkt »Models« eine Übersicht aller trainierten Modelle und können diese anhand der vorher gewählten Metrik miteinander vergleichen, wobei in diesem Experiment RMSE gewählt wurde. In der Übersicht erhält man für die gewählte Metrik eine geordnete Auflistung der Modelle und wie diese abgeschnitten haben, wobei sich in diesem Anwendungsfall das StackEnsemble-Modell mit einem RMSE von 0.10692 als das beste herausgestellt hat. Zusätzlich können die Benutzer*innen neben dem reinen Modellvergleich die jeweiligen Modelle über die Auswahl in der Liste bereitstellen oder herunterladen.

105 Quelle: Microsoft Azure Machine Learning Studio.

106 Quelle: Microsoft Azure Machine Learning Studio.

Zusammenfassung der Ergebnisse

	AWS	Google	IBM	Microsoft
Verfügbarkeit und automatische Wahl geeigneter Schritte und Algorithmen	4 (Keine Infos über Algorithmen und geeignete Schritte auf der Oberfläche)	4 (Keine Infos über Algorithmen und geeignete Schritte, da die Oberfläche nichts anzeigt)	5 (Viele verschiedene Algorithmen getestet und optimiert. Zudem eine Übersicht aller getesteten Verfahren)	5 (Viele verschiedene Algorithmen getestet und optimiert. Zudem eine Übersicht aller getesteten Verfahren)
Benutzerfreundlichkeit während des Trainings und der Evaluation	5 (Übersichtlich und klar strukturiert)	3 (Übersichtlich und klar strukturiert, aber keine Infos während des Trainings und danach über gewähltes Modell)	5 (Übersichtlich und klar strukturiert.)	5 (Übersichtlich und klar strukturiert)
Leichte Wiederverwendbarkeit/Integrierbarkeit der trainierten Modelle	5 (Sehr hohe Wiederverwendbarkeit. Python-Skripte sind einsehbar und können wiederverwendet werden.)	3 (Leicht wiederverwendbar, allerdings ist das erstellte Modell nicht auf anderen Plattformen ausführbar und man ist auf die Google-Plattform beschränkt.)	5 (Sehr hohe Wiederverwendbarkeit. Python-Skripte sind einsehbar und können wiederverwendet werden.)	3 (Modelle sind wiederverwendbar, allerdings sind diese in einem besonderen Format, das nicht ohne Weiteres eingesehen werden kann.)
Erforderliche Programmiererkenntnisse	5 (Keine Kenntnisse notwendig)	5 (Keine Kenntnisse notwendig)	5 (Keine Kenntnisse notwendig)	5 (Keine Kenntnisse notwendig)

Tabelle 16: Bewertung der Anbieter hinsichtlich des Anwendungsfalls zur vorausschauenden Wartung.

Dienst	MSE	RMSE	MAE	R^2
Amazon	0.004	0.065	-	-
Google	-	0.058	0.040	0.863
IBM	-	0.069	0.051	0.808
Microsoft	-	0.107	-	-

Tabelle 17: Übersicht der Ergebnisse nach Plattform und Metrik.

6 EMPFEHLUNGEN UND FAZIT

In der vorliegenden Studie wurden die führenden vier MLaaS-Plattformen von AWS, Google, IBM und Microsoft basierend auf technischen und geschäftlichen Allgemeinkriterien sowie auf der Grundlage von vier realitätsnahen Anwendungsfällen des Machine Learnings verglichen. Gerade bei der manuellen Durchführung der vier Use Cases wurde deutlich, wie rasant sich die angebotene Produktpalette bzw. deren Funktionsumfang ändert. Die Anbieter investieren viele Ressourcen in die Weiterentwicklung ihrer Angebote, um Mitbewerbern im heiß umkämpften MLaaS-Markt Anteile abzujagen. Für die Kundinnen und Kunden hat dies sowohl Vor- als auch Nachteile: Einerseits bieten sich vielfältige Möglichkeiten, eigene Visionen umzusetzen. Selten scheitern Projekte an fehlenden Funktionen aufseiten der Cloud-Anbieter. Viele Use Cases, die im Alltag in Unternehmen auftreten, sollten auf den betrachteten Plattformen sinnvoll durchführbar sein. Andererseits erschwert die Vielzahl der Dienste, die schiere Anzahl an Fachbegriffen und das oftmals unüberschaubare Netz von Teilbereichen der Plattformen gerade für KI-Neueinsteiger*innen den Einstieg. Neu hinzukommende Funktionalitäten sind darüber hinaus oft noch nicht ausgereift. Dies führt dazu, dass man als Endkund*in eine gehörige Portion Tüftlercharakter und Beständigkeit mitbringen sollten. Bis man sich einen Überblick über die Plattformen verschafft hat und die korrekten Dienste ausgewählt sind, kann einige Zeit vergangen sein. Inwieweit einmal eingerichtete Modelle und Schnittstellen auf Dauer Bestand haben, ohne dass sie regelmäßig aufgrund plattformbedingter Änderungen aktualisiert werden müssen, darf und muss kritisch hinterfragt werden.

Im Rahmen dieser Studie eine eindeutige, allgemeingültige Empfehlung hinsichtlich der zu verwendenden MLaaS-Plattform zu geben, ist nicht möglich. Zu vielfältig und vielschichtig sind die Einsatzmöglichkeiten und die damit verbundenen Stärken und Schwächen der einzelnen Anbieter. Bezogen auf die vier betrachteten Use Cases jedoch können wir Empfehlungen aussprechen: Im Falle tabellarischer Daten (5.4.1) sehen wir IBM Watson und Microsoft Azure im Vorteil, auch wenn Google Cloud oder AWS teils geringfügig bessere Ergebnisse lieferten. Textdaten betreffend (5.4.2) sehen wir Microsoft Azure vorne. Hier waren die Benutzerführung und der Funktionsumfang ausschlaggebend, wenngleich Microsoft auch hier nicht die beste Performance erzielt. Für die Bildklassifizierung (5.4.3) bieten Microsoft und Google ausgereifte und intuitiv benutzbare Lösungen, die auch einen Export der erstellten Modelle in diverse Technologien ermöglichen. Bei der Analyse von Zeitreihendaten (5.4.4) hat sich IBM Watson

knapp vor AWS und Microsoft Azure als am besten geeignet herausgestellt. Dies spiegelt sich unter anderem in der intuitiven Anwendbarkeit und den erzielten Ergebnissen wider, was Nutzer*innen mit geringen Vorkenntnissen einen leichten Einstieg ermöglicht.

Unabhängig von konkreten Zahlen oder Anwendungsfällen macht diese Studie auch die systembedingten Stärken und Schwächen von MLaaS-Lösungen deutlich. Zu den Stärken zählen wir die weite Fassung der Zielgruppe; von den Einsteiger*innen, die nur über grundlegende Kenntnisse im Bereich des Machine Learnings verfügen und stark auf hoch automatisierte Systeme oder Hilfestellung angewiesen ist, bis zu den Profis, die eigenen Code schreiben und nur die bloße Rechenpower aus der Cloud nutzen, kann jeder die Plattformen auf seine Weise gewinnbringend nutzen. Ein weiteres Argument für die Cloud-Dienstleister ist die nahezu permanente Verfügbarkeit von leistungsstarker Hardware. Somit werden rechenintensive Experimente ermöglicht, die ansonsten aus Kostengründen nicht durchführbar gewesen wären. Als Nachteil sehen wir die bereits angesprochene mangelhafte Überschaubarkeit der vielen Teildienste, deren Beziehung und Kommunikation untereinander sowie das zugrundliegende Preismodell. Zwar gibt es Budgetgrenzen und Preislisten, doch diese helfen nur begrenzt dabei, die entstehenden Kosten zuverlässig vorab zu verstehen. Hier wäre eine zukünftige Vereinfachung bzw. Vereinheitlichung der Teildienste, Preismodelle und Funktionen wünschenswert.

Für die Zukunft rechnen wir mit konsolidierten KI-Plattformen, die es Einsteiger*innen noch einfacher machen, Ergebnisse mittels maschinellen Lernens zu erzielen und auch zu interpretieren. Dies kann zum einen durch breiter ausgebaute und fortschrittliche AutoML-Module gelingen, die sich nicht nur auf einzelne Teilbereiche einer Plattform beziehen, sondern plattformweit einsetzbar sind. Zum anderen können erweiterte Ansätze der erklärbaren KI (Explainable AI) dabei helfen. Diese sind auf den meisten Plattformen derzeit in Ansätzen vorhanden, doch eine
umfassende Integration steht noch aus. Wir hoffen, dass solche Verfahren im Zuge fortschreitender Forschung im Bereich der Explainable AI entsprechend Einzug in MLaaS-Plattformen halten. Denn nur, wenn Ergebnisse maschineller Lernalgorithmen und damit die Nutzung von MLaaS-Lösungen transparent und erklärbar sind, wird es eine breite Akzeptanz, auch von Fachfremden, geben.

LITERATUR

[1] 115th Congress. 2018. S. 2383 - CLOUD Act. https://www.congress.gov/115/bills/s2383/BILLS-115s2383is.pdf.

[2] BITKOM. 2010. Leitfaden Cloud Computing – Was Entscheider wissen müssen.

[3] Bundesamt für Sicherheit in der Informationstechnik. 2012. Sicherheitsempfehlung für Cloud Computing Anbieter. Mindestanforderungen in der Informationssicherheit.

[4] Bundesministerium für Wirtschaft und Energie. Einsatz von Künstlicher Intelligenz in der Deutschen Wirtschaft. Stand der KI-Nutzung im Jahr 2019.

[5] CYBERARK. 2019. Global Advanced Threat Landscape 2019 Report – Special Edition EMEA.

[6] Europäisches Parlament. 2016. Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG.

[7] Ferrucci, D. A. 2012. Introduction to "This is Watson". IBM Journal of Research and Development, 56 (3.4), 1:1–1:15.

[8] Fraunhofer-Institut für Sichere Informationstechnologie. 2015. Vertraulichkeitsschutz durch Verschlüsselung. Strategien und Lösungen für Unternehmen.

[9] Gartner Research. 2020. Gartner Magic Quadrant for Cloud AI Developer Services. https://www.gartner.com/en/documents/3981253.

[10] Goldmedia GmbH Strategy Consulting, if(is) - Institut für Internet-Sicherheit. 2018. Kompass IT-Verschlüsselung. Entscheidungshilfen für kleine und mittlere Unternehmen.

[11] Hogan Lovells. 2019. Demystifying the U.S. CLOUD Act. Assessing the law's compatibility with international norms and the GDPR.

[12] Huang, Y., Qiu, C., Guo, Y., Wang, X., and Yuan, K. 2018 - 2018. Surface Defect Saliency of Magnetic Tile. In 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE). IEEE, 612–617. DOI=10.1109/COASE.2018.8560423.

[13] IDG Research Services. 2019. Studie Cloud Security 2019.

[14] KPMG. 2020. Cloud-Monitor 2020. Gut eingerichtet in der Cloud.

[15] Kriminologisches Forschungsinstitut Niedersachsen e.V. Cyberangriffe gegen Unternehmen. Ergebnisse einer repräsentativen Unternehmensbefragung 2018/2019.

[16] Michael Chui, James Manyika, Mehdi Miremadi, Nicolaus Henke, Rita Chung, Pieter Nel, Sankalp Malhotra. 2018. Notes From the AI Frontier. Insights from Hundreds of Use Cases. McKinsey Global Institute.

[17] Papp, H. and Hanussek, M. 2020. A Methodology for Deriving Evaluation Criteria for Software Solutions.

[18] Schabus, D., Skowron, M., and Trapp, M. 2017. One Million Posts: A Data Set of German Online Discussions. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17. Association for Computing Machinery, New York, NY, USA, 1241–1244. DOI=10.1145/3077136.3080711.

[19] SOPHOS. 2020. Sophos 2020 Threat Report.

KI-FORTSCHRITTSZENTRUM

Das KI-Fortschrittszentrum »Lernende Systeme« unterstützt Firmen dabei, die wirtschaftlichen Chancen der Künstlichen Intelligenz und insbesondere des Maschinellen Lernens für sich zu nutzen. In anwendungsnahen Forschungsprojekten und in direkter Kooperation mit Industrieunternehmen arbeiten die Stuttgarter Fraunhofer-Institute für Arbeitswirtschaft und Organisation IAO sowie für Produktionstechnik und Automatisierung IPA daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg.

Europas größte Forschungskooperation auf dem Gebiet der KI

Das KI-Forschungszentrum ist Forschungspartner des Cyber Valley, einem Konsortium aus den renommierten Universitäten Tübingen und Stuttgart, dem Max-Planck-Institut für intelligente Systeme und einigen führenden Industrieunternehmen. In gemeinsamen Forschungslabors werden Grundlagenforschung und anwendungsorientierte Entwicklung zu aktuellen wie auch zukünftigen Bedarfen behandelt und vorangetrieben.

Menschzentrierte KI

Alle Aktivitäten des Zentrums verfolgen das Ziel, eine menschzentrierte KI zu entwickeln, der die Menschen vertrauen und die sie akzeptieren. Nur wenn Menschen mit neuen Technologien intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann deren Potenzial optimal ausgeschöpft werden. Daher konzentrieren sich die Forschungsaktivitäten unter anderem auf die Themen Erklärbarkeit, Datenschutz, Sicherheit und Robustheit von KI-Technologien.

Studienreihe »Lernende Systeme«

Die Studienreihe »Lernende Systeme« gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Nähere Informationen und die aktuellen Versionen der Studien finden Sie unter: https://www.ki-fortschrittszentrum.de/de/themen/studien.html

FRAUNHOFER-GESELLSCHAFT

Forschen für die Praxis ist die zentrale Aufgabe der Fraunhofer-Gesellschaft. Die 1949 gegründete Forschungsorganisation betreibt anwendungsorientierte Forschung zum Nutzen der Wirtschaft und zum Vorteil der Gesellschaft. Vertragspartner und Auftraggeber sind Industrie- und Dienstleistungsunternehmen sowie die öffentliche Hand.

Die Fraunhofer-Gesellschaft betreibt in Deutschland derzeit 74 Institute und Forschungseinrichtungen. Mehr als 28 000 Mitarbeiterinnen und Mitarbeiter, überwiegend mit natur- oder ingenieurwissenschaftlicher Ausbildung, erarbeiten das jährliche Forschungsvolumen von mehr als 2,8 Milliarden Euro. Davon entfallen mehr als 2,3 Milliarden Euro auf den Leistungsbereich Vertragsforschung. Rund 70 Prozent dieses Leistungsbereichs erwirtschaftet die Fraunhofer-Gesellschaft mit Aufträgen aus der Industrie und mit öffentlich finanzierten Forschungsprojekten. Rund 30 Prozent werden von Bund und Ländern als Grundfinanzierung beigesteuert, damit die Institute Problemlösungen entwickeln können, die erst in fünf oder zehn Jahren für Wirtschaft und Gesellschaft aktuell werden.

Internationale Kooperationen mit exzellenten Forschungspartnern und innovativen Unternehmen weltweit sorgen für einen direkten Zugang zu den wichtigsten gegenwärtigen und zukünftigen Wissenschafts- und Wirtschaftsräumen.

Mit ihrer klaren Ausrichtung auf die angewandte Forschung und ihrer Fokussierung auf zukunftsrelevante Schlüsseltechnologien spielt die Fraunhofer-Gesellschaft eine zentrale Rolle im Innovationsprozess Deutschlands und Europas. Die Wirkung der angewandten Forschung geht über den direkten Nutzen für die Kund*innen hinaus: Mit ihrer Forschungs- und Entwicklungsarbeit tragen die Fraunhofer-Institute zur Wettbewerbsfähigkeit der Region, Deutschlands und Europas bei. Sie fördern Innovationen, stärken die technologische Leistungsfähigkeit, verbessern die Akzeptanz moderner Technik und sorgen für die Aus- und Weiterbildung des dringend benötigten wissenschaftlich-technischen Nachwuchses.

Ihren Mitarbeiterinnen und Mitarbeitern bietet die Fraunhofer-Gesellschaft die Möglichkeit zur fachlichen und persönlichen Entwicklung für anspruchsvolle Positionen in ihren Instituten, an Hochschulen, in Wirtschaft und Gesellschaft. Studierenden eröffnen sich aufgrund der praxisnahen Ausbildung und Erfahrung an Fraunhofer-Instituten hervorragende Einstiegs- und Entwicklungschancen in Unternehmen.

Namensgeber der als gemeinnützig anerkannten Fraunhofer-Gesellschaft ist der Münchner Gelehrte Joseph von Fraunhofer (1787–1826). Er war als Forscher, Erfinder und Unternehmer gleichermaßen erfolgreich.

Fraunhofer IAO

Mensch und Technik in der digitalen Arbeitswelt, Wirtschaft und Gesellschaft

Digitale Technologien verändern unsere Arbeitswelt und haben tiefgreifende Auswirkungen auf Wirtschaft und Gesellschaft. Lang etablierte Methoden und Prozesse werden in kurzer Zeit modernisiert und revolutioniert. Das Fraunhofer IAO kooperiert eng mit dem Partnerinstitut IAT der Universität Stuttgart und entwickelt gemeinsam mit Unternehmen, Institutionen und Einrichtungen der öffentlichen Hand wirksame Strategien, Geschäftsmodelle und Lösungen für die digitale Transformation.

Die digitale Transformation und neue IT-Technologien eröffnen für Unternehmen viele Chancen: innovative Produktangebote für neue Zielgruppen, bessere und kostengünstigere Prozesse, eine »intelligentere« Kundenkommunikation und höhere Automatisierung. Dafür kommen innovative, vernetzte IT-Lösungen auf Basis von Big Data, Künstlicher Intelligenz, Cloud und Internetplattformen zum Einsatz.

Die richtige Strategie und IT sind eine wesentliche Grundlage für den Erfolg und die Wettbewerbsfähigkeit von Unternehmen. Voraussetzung für erfolgreiche Anwendungen ist ein klarer Nutzen für das Unternehmen, seine Kund*innen und seine Partner.

Unsere Leistungen basieren auf fundierter Technologie- und Marktkenntnis sowie branchenübergreifenden Erfahrungen. Durch den Einsatz unserer praxiserprobten Methoden und erfahrenen Mitarbeitenden sichern wir den Projekterfolg. Unser Fraunhofer-Netzwerk ermöglicht uns den Zugriff auf ein umfassendes Kompetenzspektrum.

Das Fraunhofer IAO und das IAT der Universität Stuttgart beschäftigen gemeinsam mehr als 650 Mitarbeitende und verfügen über rund 15 000 Quadratmeter Büroflächen, Demonstrationszentren sowie Entwicklungs- und Testlabors.

Der Forschungsbereich »Digital Business« unterstützt Sie dabei, die Chancen der Digitalisierung durch zukunftsfähige Strategien und neue IT-Lösungen zu nutzen. Bedeutende Anwendungsbranchen des Forschungsbereichs sind Versicherungen, Fertigungsunternehmen, Mobilitätsdienste und die Energiewirtschaft.

Die Fokusthemen des Forschungsbereichs »Digital Business« sind:

- Digitalisierungsstrategien und -lösungen,
  u. a. Innovationsnetzwerke, Zukunftsszenarien, Benchmarks, New Business Radar, Blockchain-Anwendungen, Data-Science-Seminare

- Datenbasierte Automatisierungslösungen und Künstliche Intelligenz,
  u. a. Automatisierung, digitale Assistenten, maschinelle Lernverfahren, wissensbasierte Arbeitsplätze, Chatbots, Big-Data-Lösungen
- Cloudbasierte Plattformen und Geschäftsmodelle,
  u. a. webbasierte Services, Smart Products und Services, serviceorientierte Architekturen, Sicherheit und Datenschutz

- Integrierte IT-Systeme für das Internet of Things (IoT),
  u. a. IoT-Lösungen, Echtzeit-Datenplattformen, Analysesysteme, Steuerungslösungen, autonome Systeme, Smart City

- Unternehmenssoftware und Prozessoptimierung,
  u. a. Enterprise Content Management, Geschäftsprozess- und Kundenbeziehungsmanagement, Stammdatenmanagement und Enterprise Resource Planning, Cloud-Lösungen für unterschiedliche Unternehmensbereiche

- Intelligente, vernetzte Energiesysteme,
  u. a. Elektromobilität und Lademanagement, Buchung und Dispositionsmanagement von Elektrofahrzeugen im Flottenbetrieb, Predictive Analytics, Simulation und Optimierung, Lastmanagement, Beschaffungsoptimierung, servicebasierte Plattformen

Unsere Leistungen basieren auf fundierter Technologie- und Marktkenntnis sowie branchenübergreifenden Erfahrungen. Durch den Einsatz unserer praxiserprobten Methoden und erfahrenen Mitarbeitenden sichern wir den Projekterfolg. Unser Fraunhofer-Netzwerk ermöglicht uns den Zugriff auf ein umfassendes Kompetenzspektrum.

Das Fraunhofer IAO und das IAT der Universität Stuttgart beschäftigen gemeinsam mehr als 650 Mitarbeitende und verfügen über rund 15 000 Quadratmeter Büroflächen, Demonstrationszentren sowie Entwicklungs- und Testlabors.

Fraunhofer IPA

Das Fraunhofer IPA – eines der größten Institute der Fraunhofer- Gesellschaft – wurde 1959 gegründet und beschäftigt annähernd 1000 Mitarbeiterinnen und Mitarbeiter. Unsere Zukunfts- und Leitthemen sind Biointelligente Wertschöpfung, Digitale Transformation im Rahmen von Industrie 4.0, Energiespeicher, Frugale Produktionssysteme, Künstliche Intelligenz in der Automatisierung, Leichtbau und Ressourceneffizienz.

Organisatorische und technologische Aufgabenstellungen aus der Produktion machen die Forschungs- und Entwicklungsschwerpunkte des Fraunhofer IPA aus. Methoden, Komponenten und Geräte bis hin zu kompletten Maschinen und Anlagen werden von uns entwickelt, erprobt und exemplarisch eingesetzt. Die 15 Fachabteilungen des Fraunhofer IPA decken den gesamten Bereich der Produktionstechnik ab. Sie werden koordiniert durch sechs Geschäftsfelder und arbeiten interdisziplinär mit Industrieunternehmen der Branchen Automotive, Maschinen- und Anlagenbau, Elektronik und Mikrosystemtechnik, Energie, Medizin- und Biotechnik sowie Prozessindustrie zusammen.

Künstliche Intelligenz in der Produktion

Das Fraunhofer IPA hat KI vor einigen Jahren zu einem seiner Leitthemen gemacht und sein Beratungs-, Förder- und Umsetzungsangebot strategisch intensiv ausgebaut. Zu diesen Angeboten gehören Machbarkeitsuntersuchungen, Quick Checks und Workshops vor Ort und auch die Entwicklung komplexer technischer Produktionsmodule, die auf ML basieren. Die Projektformen reichen von kompakten kurzen Zusammenarbeiten zu einer konkreten Fragestellung bis hin zu langfristigen strategischen Umsetzungen. Zudem engagiert sich das Institut in Vorträgen, Schulungen und ist mit zahlreichen Veröffentlichungen auch in der Wissenschaft präsent.

Digitale Werkzeuge in der Produktion

Der Forschungs- und Entwicklungsschwerpunkt des Kompetenzzentrums digITools liegt auf IT-Architekturen, Daten- und Anwendungsdiensten und Umsetzungsmethoden für die digitale Produktion. Wir unterstützen Unternehmen bei der Entwicklung und Integration von digitalen Werkzeugen in die Produktion.

Zu unseren Leistungen gehören neben der Beratung und Entwicklung von Lösungen rund um Computer- und Kommunikations-Infrastrukturen, wie dem 5G-Transferzentrum und der Virtual Fort Knox (VFK) Edge-Cloud-Plattform, auch Digitalisierungs- und Integrationslösungen wie der Manufacturing Service Bus (MSB) zur Anlagen- und Datenintegration. Von der Maschine auf dem Hallenboden über die Schnittstellen zu Produktionsdiensten bis hin zum digitalen Abbild der Produktion besitzen wir das Know-how, Werkzeuge und Technologien für eine vernetzte intelligente Produktion mittels cyberphysischer Produktionssysteme zu entwickeln. Unsere datengetriebenen Technologien und funktionalen IT-Lösungen für produzierende Unternehmen sind

unter anderem die Bausteine für Leuchtturmprojekte, wie unser vom Bundesministerium für Wirtschaft und Energie gefördertes offenes, verteiltes, echtzeitfähiges und sicheres Betriebssystem für die Fabrik, FabOS. Mit dem Mittelstand-4.0-Kompetenzzentrum und der Industrie-4.0-Seminarreihe haben Unternehmen die Möglichkeit, gemeinsam mit den digITools-Experten die neuesten Anwendungen rund um die digitale Produktion kennenzulernen und gemeinsam umzusetzen. Im Future Work Lab werden diese Entwicklungen und Ergebnisse greifbar für und mit Unternehmen dargestellt. Insbesondere kleine und mittlere Unternehmen erhalten so Unterstützung, die Potenziale von Industrie 4.0 für sich zu erschließen.

Kontaktadresse
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstraße 12, 70565 Stuttgart

Marc Hanussek
Telefon +49 711 970-5120
marc.hanussek@iao.fraunhofer.de

Arthur Grigorjan
Telefon +49 711 970-1662
arthur.grigorjan@ipa.fraunhofer.de

Herausgeber
Wilhelm Bauer, Thomas Bauernhansl, Marco Huber, Werner Kraus
Matthias Peissner, Thomas Renner

Titelbild
© Foto Pokki, lvnl – stock.adobe.com / Fraunhofer IAO

Satz und Gestaltung
Franz Schneider, Fraunhofer IAO

URN-Nummer
urn:nbn:de:0011-n-6221499

Online verfügbar als Fraunhofer-ePrint
http://publica.fraunhofer.de/dokumente/N-622149.html

Baden-Württemberg

MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S·TEC