Informatik Spektrum (2022) 45:13–19
https://doi.org/10.1007/s00287-021-01420-5

HAUPTBEITRAG

Data Preprocessing as a Service – Outsourcing der Datenvorverarbeitung für KI-Modelle mithilfe einer digitalen Plattform

Marko Kureljusic[1] · **Erik Karger**[1]

Angenommen: 5. Oktober 2021 / Online publiziert: 29. Oktober 2021

Zusammenfassung

Sowohl in der Praxis als auch in der Wissenschaft kam es in den vergangenen Jahren zu einem zunehmenden Interesse an datenintensiven Verfahren, wie der künstlichen Intelligenz. Die Mehrheit dieser Data-Science-Projekte fokussierte sich auf den Erklärungsgehalt und die Robustheit der Modelle. Vernachlässigt wurde hierbei häufig der Prozess der Datenvorverarbeitung, obwohl dieser ca. 80 % der Zeit eines Data-Science-Projekts beansprucht. Im Rahmen der Datenvorverarbeitung, welche auch als Data Preprocessing bezeichnet wird, werden Daten akquiriert, bereinigt, transformiert und reduziert. Das Ziel dieser Vorgehensweise ist die Generierung eines Datensatzes, welcher sich für Trainings- und Testzwecke der Data-Science-Modelle eignet. Somit ist das Data Preprocessing ein erforderlicher Prozessschritt, der für das maschinelle Erlernen von korrekten Mustern und Zusammenhängen notwendig ist. Häufig scheitern Data-Science-Projekte jedoch an der mangelhaften Datenvorverarbeitung. So werden beispielsweise fehlerhafte Daten nicht vorab identifiziert, wodurch möglicherweise falsche Zusammenhänge erlernt werden. Dies führt dazu, dass der Erklärungsgehalt der Data-Science-Modelle signifikant verringert wird. Eine Möglichkeit, dieses Problem zu lösen, ist das Outsourcing der Datenvorverarbeitung an spezialisierte Fachkräfte. Mithilfe einer Plattform kann ein sicherer und automatisierter Datenaustausch zwischen Kunden und Dienstleistern gewährleistet werden. Der vorliegende Beitrag thematisiert, wie die Plattform für das Data Preprocessing genutzt werden kann, um eine effizientere und schnellere Bereitstellung der Daten zu ermöglichen.

Einleitung

Mangelnde Datenqualität wird laut aktuellem Stand der Forschung als einer der wichtigsten Gründe für das Scheitern von Data-Science-Projekten angesehen [1]. In den bisherigen empirischen Studien wird eine Ex-post-Betrachtung vorgenommen, indem diese nachträglich die Ursachen für gescheiterte Projekte untersuchen. Es mangelt an empirischen Untersuchungen, die aufzeigen, dass eine angemessene Datenvorverarbeitung das Problem der mangelhaften Datenqualität eliminieren oder zumindest reduzieren kann. Jedoch wird dies in der qualitativen Forschung bereits als elementarer Bestandteil zur Gewährleistung der Datenqualität angesehen [2].

Das Aufbereiten von Rohdaten, welches im Folgenden auch als Data Preprocessing bezeichnet wird, ist in der Praxis eine der technisch anspruchsvollsten Aufgaben bei Data-Science-Projekten. Vor allem das Extrahieren von Merkmalen aus Daten führt zu einer erhöhten Komplexität und einem hohen zeitlichen Aufwand [3]. Um Algorithmen, wie Machine-Learning-basierte Analyseverfahren, auf den Daten anzuwenden, muss das Data Preprocessing erfolgreich abgeschlossen sein. Laut aktuellen Umfragen beansprucht das Data Preprocessing ca. 80 % der Zeit eines Data-Science-Projekts und ist maßgeblich für dessen Erfolg [4].

Data Science ist eine multidisziplinäre Wissenschaft, die Expertise in verschiedenen Bereichen erfordert. So sind insbesondere bei dem Data Preprocessing Kenntnisse der Programmierung, Statistik oder dem Datenmanagement notwendig [8]. Darüber hinaus setzt eine qualitativ hochwertige Aufbereitung von Rohdaten voraus, dass deren Herkunft und Inhalt verstanden werden. Neben technischer Expertise ist daher, je nach Branche, Unternehmen oder Problemstellung, auch ein bestimmtes domänenspezifisches Know-how der Data-Scientists erforderlich. Aufgrund der unterschiedlichen Rahmenbedingungen und Zielsetzungen von Data-Science-Projekten gibt es oftmals kein standardisiertes Vorgehen, auf das bei der Datenvorverarbeitung zurückgegriffen werden kann. Aus diesem Grund besteht die Not-

✉ Marko Kureljusic
marko.kureljusic@uni-due.de

[1] Campus Essen, NRW, Universität Duisburg-Essen, Universitätsstraße 12, 45141 Essen, Deutschland

wendigkeit, dass das Data Preprocessing durch einen Data-Scientist durchgeführt wird, der auf bestimmte Branchen spezialisiert ist und Inhalt, Struktur sowie Herkunft der Daten versteht. Allerdings ist der Markt für spezialisierte Fachkräfte sehr kompetitiv, wodurch es für Unternehmen herausfordernd ist, diese zu akquirieren und zu halten. So geht aus einer vom Stifterverband und McKinsey durchgeführten Studie hervor, dass bis zum Jahr 2023 ca. 455.000 zusätzliche Fachkräfte für komplexe Datenanalysen gesucht werden [5].

Der vorliegende Beitrag stellt mit Data Preprocessing as a Service einen neuen, plattformbasierten Ansatz für das Outsourcing der Datenvorverarbeitung vor. Auf Basis einer Plattform entsteht ein neuer Dienstleistungsmarkt, der es ermöglicht, die Datenvorverarbeitung an spezialisierte Fachkräfte outzusourcen. Hierdurch soll die Datenqualität optimiert werden, um die Erfolgswahrscheinlichkeit von Data-Science-Projekten zu erhöhen.

Zunächst werden im zweiten Kapitel die theoretischen Grundlagen beschrieben, indem auf die Schritte des Data Preprocessings sowie auf digitale Geschäftsmodelle eingegangen wird. Daraufhin wird im dritten Kapitel das Geschäftsmodell Data Preprocessing as a Service vorgestellt. In Kapitel vier erfolgt eine kritische Diskussion dieses Ansatzes hinsichtlich der praktischen und theoretischen Notwendigkeit sowie der Nachteile und Risiken, die sich durch die Anwendung ergeben können. Der Artikel schließt mit einem Fazit in Kapitel fünf, welches die zentralen Aussagen zusammenfasst.

Theoretische Grundlagen

Im Folgenden wird das Data Preprocessing tiefergehend beleuchtet, indem auf sämtliche relevanten Schritte, von der Datensammlung bis hin zur optimalen Datenbereitstellung für Machine-Learning-Anwendungen, eingegangen wird. In diesem Zusammenhang wird auch die Relevanz dieser Schritte thematisiert. Da das Data Preprocessing in der Praxis aufgrund von zeitlichen, finanziellen oder personellen Gründen häufig vernachlässigt wird, stellt sich die Frage, welchen Mehrwert es für Data-Science-Projekte bieten kann. Im Anschluss an die Beantwortung dieser Frage werden sowohl die Grundlagen digitaler Plattformen vorgestellt als auch deren Notwendigkeit wissenschaftstheoretisch begründet.

Data Preprocessing

In der Regel werden für Data-Science-Projekte Daten aus unterschiedlichen Ursprungsquellen herangezogen, die sich sowohl von ihrer Struktur als auch ihrer potenziellen Informationsqualität differenzieren [6]. Grundsätzlich lassen sich strukturierte und unstrukturierte Daten sowie interne und externe Daten vorfinden, die gemeinsam für eine Problemstellung genutzt werden können. Das Ziel von Data Science besteht darin, komplexe Muster und Zusammenhänge in einem Datensatz zu identifizieren, um wichtige Erkenntnisse aus den Daten zu gewinnen [7]. Die identifizierten Muster bilden die Basis für Machine-Learning-Anwendungen, welche in der Lage sind, anspruchsvolle Regressions- und Klassifikationsaufgaben zu lösen. Da im Vorfeld eines Data-Science-Projekts meistens unklar ist, wie stark die Eingangsvariable die Zielvariable beeinflusst, wird häufig eine Vielzahl unterschiedlicher Daten verwendet. Mithilfe von iterativen Testverfahren können die wesentlichen Ergebnistreiber im Datensatz identifiziert werden [8]. Dies setzt jedoch voraus, dass die Daten maschinell verarbeitbar sind. In den meisten Fällen müssen die Rohdaten erst aufwendig angepasst werden, um anschließend Algorithmen erfolgreich darauf anzuwenden. Die einzelnen Schritte einer Datenvorverarbeitung werden in Anlehnung an einschlägige Praxishandbücher für Machine Learning, u. a. Géron [9] und Chollet [10], nachfolgend dargestellt.

Die Abb. 1 veranschaulicht, dass sämtliche Schritte der Datenvorverarbeitung bei falscher Anwendung Risiken nach sich ziehen. Um den Erfolg eines Data-Science-Projekts nicht zu gefährden, sollte das Data Preprocessing daher nicht vernachlässigt werden. Vor dem Hintergrund, dass es mittlerweile AutoML-Tools gibt, die grundsätzlich ein End-to-end-Learning ermöglichen, stellt sich jedoch die Frage, welchen Mehrwert eine manuelle Datenvorverarbeitung gegenüber AutoML-Tools bietet. Bekannte Beispiele für AutoML-Tools sind unter anderem Auto-sklearn [11], Hyperopt [12] und Auto-Keras [13]. Das Ziel sämtlicher AutoML-Tools besteht darin, die Hyperparameter von Machine-Learning-Algorithmen zu optimieren, um so das bestmögliche Ergebnis aus der zugrundeliegenden Datenmenge zu erzielen [14]. Zwar unterstützen AutoML-Tools einige Schritte der Datenvorverarbeitung, wie die Vektorisierung von kategorialen Variablen, jedoch ersetzen sie keine vollständige Datenvorverarbeitung durch Data-Scientists [15]. Vor allem ist im Rahmen der Merkmalsextraktion ein tiefgründiges Verständnis der Datenmengen und ihrer Zusammenhänge notwendig, um neue Eingangsvariablen zu kreieren, die das maschinelle Erlernen von komplexen Mustern erleichtern [16].

Neben AutoML-Tools gibt es auch Machine-Learning-Dienstleistungen, die auf Basis von Cloud Computing angeboten werden. Charakteristisch hierfür ist, dass große Datensätze mit modernen Algorithmen analysiert werden können, ohne die eigene Hardware zu beanspruchen [17]. Diese werden auch als Machine Learning as a Service bezeichnet. Bekannte Beispiele hierfür sind IBM Watson, Azure, Google Cloud und AWS. Analog zu AutoML-Tools können im Rahmen von Machine Learning as a Service grundlegen-

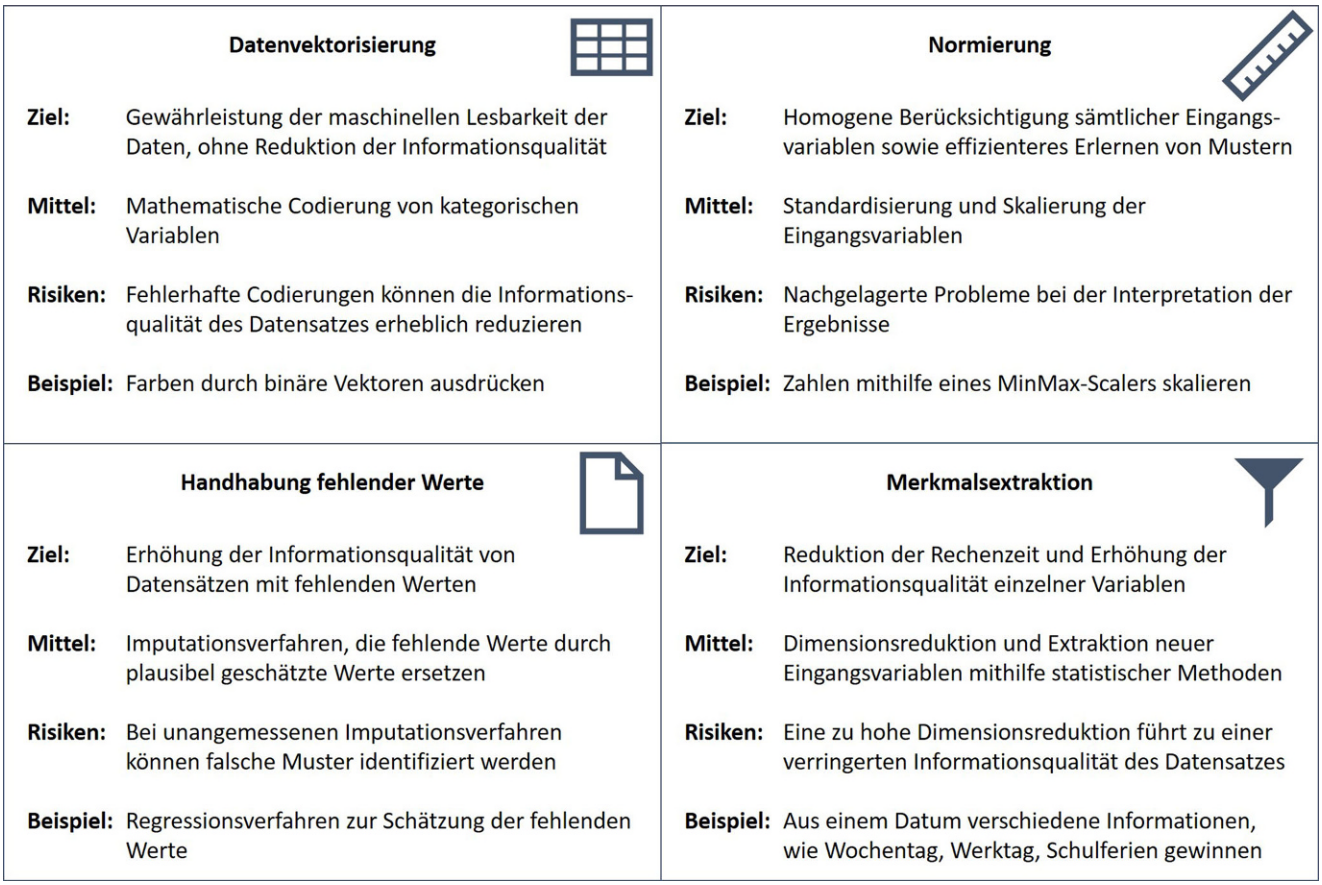

Abb. 1 Bestandteile des Data Preprocessings. (eigene Darstellung)

de Schritte der Datenvorverarbeitung automatisiert werden. Allerdings ersetzen diese ebenfalls keine vollwertigen Data-Scientists, da sie auf ähnlichen Algorithmen wie AutoML-Tools basieren und daher wichtige Schritte der Merkmalsextraktion nicht automatisiert werden können. Hierzu bedarf es einem tiefen Verständnis über die zugrundeliegenden Daten, welches nur durch eine starke künstliche Intelligenz automatisiert werden könnte. Allerdings ist dies laut aktuellem Stand der Technik noch nicht möglich und erst mittel- bis langfristig realisierbar [18].

Digitale Plattformen

Das digitale Zeitalter ist durch technologische Entwicklungen und Phänomene wie das Internet of Things, Big Data und smarte Produkte charakterisiert. Die ökonomische Nutzung der Potenziale, welche den digitalen Technologien und Daten innewohnt, begünstigt das Entstehen neuer Formen von Wertschöpfung [19]. Eine Ausprägung sind dabei digitale Plattformen, wie beispielsweise Airbnb, Uber oder Alibaba. Grundsätzlich stellen Plattformen keine neue Entwicklung dar, sondern existieren bereits seit Langem, wie etwa in Form von Einkaufszentren, die Verbraucher und Händler verbinden [20].

Die Informationstechnologie hat jedoch die Notwendigkeit, eine eigene physische Infrastruktur und eigene Vermögenswerte zu besitzen, stark reduziert. Auch werden durch IT eine einfache Skalierung digitaler Plattformen sowie die Anbindung vieler Plattformteilnehmer möglich. Dies verstärkt die Netzwerkeffekte und ermöglicht es den Plattformbetreibern, riesige Datenmengen zu erfassen, zu analysieren und auszutauschen [20]. Zusätzlich können IT-Dienstleistungen und IT-Infrastruktur zunehmend über das Internet bezogen werden. Dieser als Cloud Computing bezeichnete Fremdbezug von IT bietet Vorteile, wie eine flexible Nutzung und Abrechnung, beispielsweise nach der Anzahl der Nutzer oder der Nutzungsdauer. Die häufigsten über das Cloud Computing bezogenen Dienstleistungen sind Software (Software as a Service, SaaS), Hardware (Infrastructure as a Service, IaaS) oder Plattformen für die Entwicklung eigener Lösungen (Platform as a Service, Paas) [21].

Das stetig zunehmende Volumen an Daten, welches ein Resultat der zunehmenden Verbreitung digitaler Technologien ist, kann ebenso die Basis für Geschäftsmodelle darstellen. Chen et al. [22] betonen den Wert und die Möglichkeiten, die Daten für Unternehmen sowie die Generierung von Wettbewerbsvorteilen bieten können. Big Data sowie die Technologien und Fähigkeiten für deren Analyse kön-

nen für Unternehmen entscheidend für die Wertschöpfung und strategische Weiterentwicklung sein. Aus Daten lässt sich auf verschiedene Art und Weise ein Mehrwert generieren. Zum einen können Daten intern verwendet werden, um auf Basis von Analysen interne Prozesse effizienter zu gestalten. Zum anderen können durch eine fundierte Analyse von Daten angebotene Produkte und Dienstleistungen verbessert oder Kundenwünsche besser erfüllt werden.

Die Notwendigkeit von digitalen Plattformen lässt sich auch wissenschaftstheoretisch begründen. Grundsätzlich lassen sich durch die Auslagerung von Tätigkeiten, die für das Unternehmen mit hohen Opportunitätskosten verbunden sind, komparative Kostenvorteile realisieren [23]. Dies ist darauf zurückzuführen, dass externe Fachkräfte für spezialisierte Aufgaben weniger Zeit aufwenden müssen als interne, ungeschulte Mitarbeiter. Da die Zusammenbringung von externen Fachkräften und Unternehmen für beide Seiten ebenfalls hohe Opportunitätskosten verursacht, können digitale Plattformen diese Aufgabe übernehmen. Die effiziente Allokation von Ressourcen zwischen Anbietern und Dienstleistern ermöglicht eine Reduktion von Transaktionskosten für beide Parteien. Transaktionskosten, die durch digitale Plattformen verringert werden können, sind Anbahnungs- und Matching-Kosten sowie Informationskosten [24, 25]. Ein Grund hierfür ist, dass sich externe Fachkräfte und Unternehmen über eine digitale Plattform schneller miteinander vernetzen können.

Darüber hinaus kann eine digitale Plattform dazu beitragen, Informationsasymmetrien zwischen Nachfrager und Dienstleister zu reduzieren. Ein Mittel hierfür ist die Gewährleistung eines Qualitätsstandards hinsichtlich der erbrachten Dienstleistung, welches das Vertrauen auf der Nachfrageseite erhöht und opportunistisches Verhalten minimiert [26]. Ferner zeigen Tabarrok und Cowen [27], dass die Möglichkeit einer gegenseitigen Bewertung ebenfalls einen Abbau von Informationsasymmetrien begünstigt. Somit lässt sich die Notwendigkeit digitaler Plattformen, neben der Theorie des komparativen Kostenvorteils, auch durch die Prinzipal-Agenten-Theorie begründen. Der Prinzipal ist in diesem Fall der Nachfrager, der den Agenten über eine digitale Plattform beauftragt, eine Dienstleistung zu erbringen. Der Agent hat zwar einen Wissensvorsprung gegenüber dem Prinzipal, kann diesen aber nicht für opportunistische Zwecke nutzen, da die Plattform als dritte Partei die ordentliche Erfüllung des Dienstleistungsvertrags sicherstellt [28].

Data Preprocessing as a Service – Vorstellung einer plattformbasierten Lösung

Das vorherige Kapitel hat die theoretische Fundierung digitaler Plattformen erörtert sowie die Notwendigkeit und Relevanz der Datenvorverarbeitung im Kontext von Data-Science-Projekten thematisiert. Dabei wurde insbesondere auf die Probleme und Herausforderungen eingegangen, die sich im Rahmen der Datenvorverarbeitung ergeben können. Im Folgenden wird mit dem Begriff Data Preprocessing as a Service ein digitaler, plattformbasierter Lösungsansatz vorgestellt. Der Zweck dieser Plattform besteht darin, aktuelle praktische Probleme der Datenqualität zu reduzieren, indem der Prozess der Datenvorverarbeitung effizient, anonym und sicher an einen Data-Scientist ausgelagert wird. Dieses Outsourcing ermöglicht auch kleinen und mittleren Unternehmen mit geringeren finanziellen Ressourcen den Zugang zu dem notwendigen Know-how für das Data Preprocessing. Mit der Plattform entsteht so ein neuer Markt, der durch die effiziente Ressourcenallokation ein Wertschöpfungspotenzial sowohl für Datennutzer als auch für Data-Scientists bietet. Nachfolgend wird die Data-Preprocessing-Plattform visuell dargestellt und beschrieben. (Abb. 2).

Rohdaten stellen für Datennutzer zunächst die Grundlage dar, aus der mithilfe von Machine-Learning-Anwendungen Erkenntnisse gewonnen werden können. Hierfür ist jedoch der Schritt der Datenvorverarbeitung notwendig, welcher durch die Plattform an erfahrene Data-Scientists ausgelagert werden kann. Die Datennutzer stellen die Nachfrageseite der Plattform dar, wohingegen die Data-Scientists das Aufbereiten von Rohdaten als Dienstleistung anbieten. Hierbei kann es sich sowohl um einzelne Personen als auch Unternehmen handeln, die auf die Vorverarbeitung von Daten spezialisiert sind. Die Data-Preprocessing-Plattform ist multifunktional konzipiert und bietet neben einer Matching-Funktion zwischen Datennutzer und Data-Scientist auch die Möglichkeit eines sicheren Datenaustausches sowie eine Datenqualitätskontrolle.

Im Rahmen von Data Preprocessing as a Service laden die Datennutzer zunächst ihre unverarbeiteten Rohdaten auf die Plattform hoch. Ein häufiges Problem von digitalen Plattformen ist das Bedenken der Datennutzer, etwa hinsichtlich des Datenschutzes. Eine Möglichkeit dieses Problem zu beheben, ist die Pseudonymisierung von Daten, die nach der DSGVO wie folgt definiert ist: „Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können ...“ (Art 4 Nr. 5 DSGVO). Die Pseudonymisierung personenbezogener Daten erfolgt häufig mit Verschlüsselungsverfahren, welche ein Teilgebiet der Kryptografie sind. Mithilfe eines Schlüssels werden sensible Daten in Chiffretext transformiert und ermöglichen, nach Abschluss der Datenverarbeitung, eine spätere Rückführung der Pseudonymisierung [29]. Dieses Verfahren eignet sich auch für die Data-Preprocessing-Plattform. Vor dem Hintergrund der hoch spezialisierten Aufgabe, sollte sowohl die

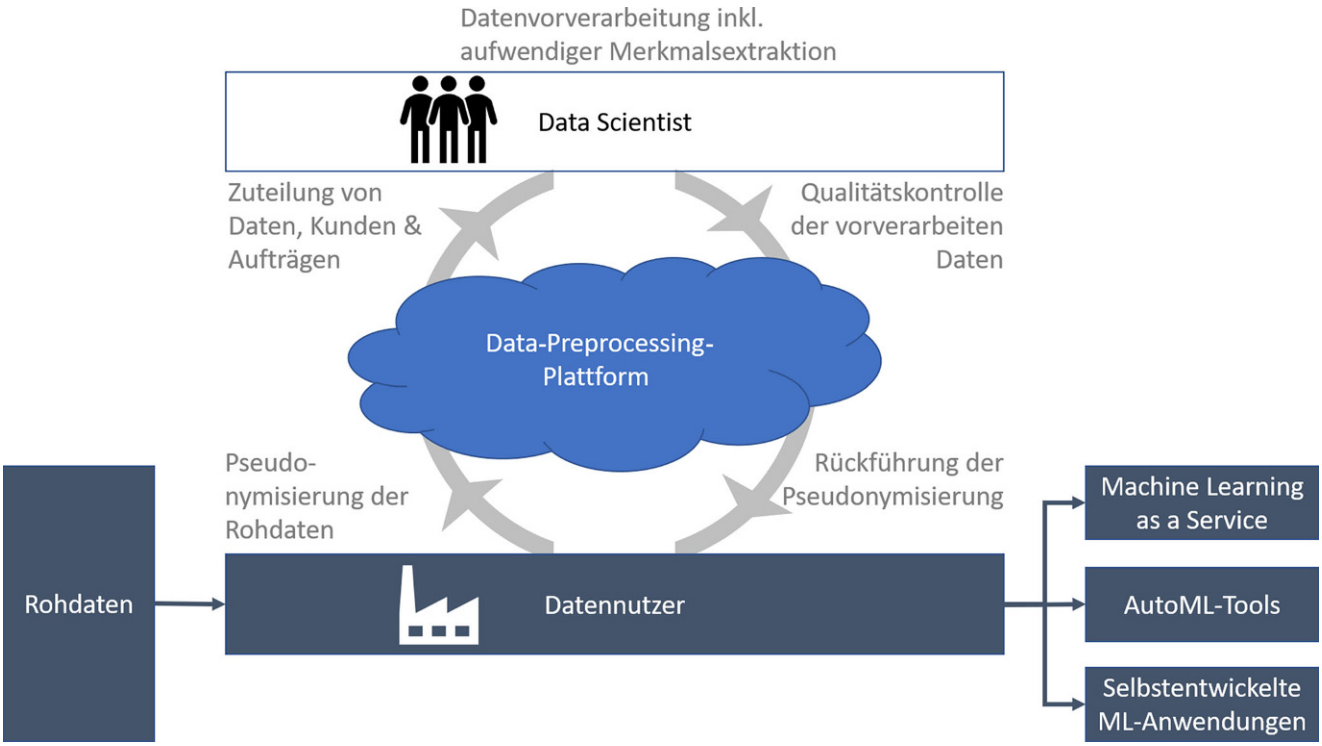

Abb. 2 Data Preprocessing as a Service. (eigene Darstellung)

Pseudonymisierung als auch die Rückführung der Pseudonymisierung von der Plattform selbst erfolgen.

Der Datennutzer kann der Plattform zudem mitteilen, für welchen Zweck die Daten genutzt werden sollen. Daraufhin werden von der Plattform geeignete Data-Scientists ausgewählt bzw. dem Datennutzer vorgeschlagen. Der Datennutzer kann hierbei auswählen, ob er die Entscheidung vollständig an die Plattform auslagern möchte oder stattdessen selbst unter den vorgeschlagenen Anbietern auswählen möchte. Die Eignung verschiedener Data-Scientists für einen bestimmten Data-Preprocessing-Auftrag wird anhand verschiedener Auswahlkriterien ermittelt. Hierzu zählen u. a. die Anzahl erfolgreich abgeschlossener Projekte, Fachexpertisen, Referenzen und Auszeichnungen. Nachdem der Data Scientist ausgewählt wurde, entsteht zwischen ihm und dem Datennutzer ein Vertragsverhältnis, welches über die Plattform abgeschlossen wird. Der Plattformbetreiber könnte für die Durchführung dieses Auswahlprozesses und der Gewährleistung der Datensicherheit in Abhängigkeit von der Größe der Datenmenge eine Gebühr erheben. Daraufhin werden dem Data-Scientist die hochgeladenen und pseudonymisierten Daten bereitgestellt, sodass das Data Preprocessing durchgeführt werden kann. Nach erfolgreichem Abschluss dieses Schrittes werden die vorverarbeiteten Daten durch die Plattform hinsichtlich ihrer Datenqualität überprüft. Hierdurch soll gewährleistet werden, dass der Data-Scientist die Datenvorverarbeitung ordnungsgemäß erfüllt hat. Anschließend werden dem Datennutzer die vorverarbeiteten Daten zum Download bereitgestellt. Dieser kann die angepassten Daten nun verwenden, um Machine-Learning-Algorithmen darauf anzuwenden. Eine gegenseitige Bewertung von Nachfrager und Dienstleister fördert den Abbau von Informationsasymmetrien und kann so zur Optimierung zukünftiger Auswahlverfahren beitragen.

Kritische Diskussion

Im Folgenden wird die Data-Preprocessing-Plattform hinsichtlich ihrer Notwendigkeit kritisch diskutiert, indem der Ansatz von bereits bestehenden Plattformdiensten abgegrenzt wird. Darauf aufbauend wird der Ansatz vor dem Hintergrund der Prinzipal-Agenten-Theorie und der Theorie des komparativen Kostenvorteils wissenschaftstheoretisch gewürdigt. Abschließend werden Nachteile und Risiken aufgezeigt, die sich durch die Anwendung der Plattform ergeben können.

Bereits heute lassen sich auf Crowdsourcing-Plattformen Data-Science-Dienstleistungen beziehen. Bekannte Beispiele sind unter anderem Kaggle, Explorium und AICrowd. Sämtliche genannten Plattformen bieten die Möglichkeit, Unternehmen mit Data-Scientists zu vernetzen. Allerdings setzt dies zunächst voraus, dass personenbezogenen Daten pseudonymisiert werden, da ansonsten datenschutzrechtliche Bestimmungen gemäß Art. 32

DSGVO verletzt werden. An dieser Stelle differenziert sich die vorgeschlagene Data-Preprocessing-Plattform von herkömmlichen digitalen Plattformen, da die Pseudonymisierung als zusätzliche Dienstleistung durchgeführt wird, um dem Unternehmen unnötige Opportunitätskosten zu ersparen. Eine weitere Abgrenzung gegenüber herkömmlichen Plattformen ist die Qualitätskontrolle der vorverarbeiteten Daten. Komplexe Tätigkeiten wie das Data Preprocessing haben bei einer crowdbasierten Lösung den Nachteil, dass die Qualität der Datenverarbeitung bei fehlender Qualifikation des Data-Scientisten beeinträchtigt werden kann [30]. Die Data-Preprocessing-Plattform setzt an diesem Problem an, indem die Datennutzer bei der Qualitätsbeurteilung der erbrachten Dienstleistung unterstützt werden.

Nachdem die Data-Preprocessing-Plattform von bereits bestehenden digitalen Plattformen abgegrenzt wurde, soll im Folgenden evaluiert werden, ob die wissenschaftstheoretischen Ansätze zur Begründung einer digitalen Plattform auch im hier beschriebenen Fall gültig sind. In Bezug auf die Theorie der komparativen Kostenvorteile ist festzustellen, dass Datennutzer durch das Outsourcing der Datenvorverarbeitung finanzielle und zeitliche Ressourcen einsparen können, da ein Data-Scientist diese Tätigkeit effizienter erledigen kann. Darüber hinaus werden Transaktionskosten (Such- und Informationskosten) durch die Data-Preprocessing-Plattform reduziert. Der Datennutzer bekommt aufgrund der effizienten Allokation von Angebot und Nachfrage in vergleichsweise kurzer Zeit einen vorverarbeiteten Datensatz, den er für seine Data-Science-Projekte nutzen kann. Der Data-Scientist kann über die Plattform seine Dienstleistungen direkt anbieten und reduziert dadurch seine Akquisekosten. Somit realisieren beide Parteien komparative Kostenvorteile. Darüber hinaus werden mithilfe der Plattform Informationsasymmetrien zwischen Datennutzer und Data-Scientist abgebaut, da sowohl die Qualität des Auswahlverfahrens als auch die Qualität der Datenvorverarbeitung sichergestellt wird. Im Unterschied zu einem herkömmlichen Dienstleistungsvertrag können Data-Scientists nicht opportunistisch handeln, da ihre Tätigkeit einer externen Qualitätskontrolle durch die Plattform unterzogen wird. Somit lässt sich die Notwendigkeit einer Data-Preprocessing-Plattform auch mithilfe der Prinzipal-Agenten-Theorie begründen.

Neben den aufgeführten Vorteilen existieren jedoch auch Nachteile und Risiken, die sich für Plattformbetreiber oder Datennutzer ergeben können. Bezogen auf die Plattformbetreiber ist zunächst festzustellen, dass eine internationale Expansion aufgrund von unterschiedlichen länderspezifischen Datenschutzvorschriften erschwert ist. Die Berücksichtigung regionaler Gesetze ist für die Plattformbetreiber zwingend erforderlich, um die Rahmenbedingungen für das Geschäftsmodell juristisch zu legitimieren.

Hinsichtlich der Datennutzer ist kritisch anzumerken, dass das Outsourcing von IT-Dienstleistungen auch Risiken mit sich bringt, die durch empirische Studien belegt wurden [31]. Hierzu zählen vor allem die erhöhte Abhängigkeit von den Anbietern sowie der Verlust von kritischen Fähigkeiten und Kompetenzen aus Sicht des Datennutzers, da das technische Know-how der Datenvorverarbeitung außerhalb des Unternehmens liegt.

Fazit

Die Digitalisierung ist eine anhaltende Entwicklung, deren Geschwindigkeit in den kommenden Jahren noch zunehmen wird. Ein direktes Resultat dieser Entwicklung ist ein stetig steigendes Volumen an Daten, welches durch unterschiedliche Datenerzeuger und in verschiedenen Formaten anfällt. Für Unternehmen, die über große Mengen an Daten verfügen, ist es oftmals schwierig, einen Mehrwert und sinnvolle Erkenntnisse aus den Daten zu generieren. Data-Science-Projekte sind häufig zeit- und kostenintensiv und erfordern ein hohes Maß an Expertise sowie Know-how. Ein für die Datenanalyse erforderlicher Schritt ist zunächst das Data Preprocessing, also das Aufbereiten von Rohdaten für die maschinelle Verarbeitung. Diese Datenvorverarbeitung stellt bei Data-Science-Projekten jedoch einen erheblichen Zeitaufwand dar. Ebenfalls kann die Datenvorverarbeitung problem- und branchenspezifische Herausforderungen mit sich bringen, wodurch eine Durchführung durch erfahrene Data-Scientists erforderlich ist.

Data Preprocessing as a Service stellt eine Möglichkeit dar, den zeitintensiven und mitunter komplexen Vorgang des Data Preprocessings an spezialisierte Dienstleister auszulagern. Die Plattform fungiert als ein Bindeglied zwischen Datennutzer und Data-Scientist. Neben einer Matching-Funktion bietet die Plattform zudem einen sicheren Datenaustausch durch eine Pseudonymisierung sowie die Gewährleistung einer Mindestqualität hinsichtlich der Datenvorverarbeitung an. Auf diese Art und Weise kann das Data Preprocessing outgesourct werden, wodurch Unternehmen die Analyse ihrer zunehmenden Datenmengen ermöglicht wird.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

1. Reggio G, Astesiano E (2020) Big-data/Analytics projects failure: a literature review. In: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications. IEEE, S 246–255
2. Ebubeogu F, Lee S (2019) Systematic literature review of preprocessing techniques for imbalanced data. IET Softw 13:479–496
3. Bradford L (2018) 8 real challenges data scientists face. https://www.forbes.com/sites/laurencebradford/2018/09/06/8-real-challenges-data-scientists-face/?sh=7a29592d6d99. Zugegriffen: 25. Nov. 2020
4. Konstantinou N, Paton NW (2020) Feedback driven improvement of data preparation pipelines. Inf Syst 92:101480
5. Kirchherr J, Klier J, Lehmann-Brauns C et al (2018) Future Skills: Welche Kompetenzen in Deutschland fehlen. https://www.stifterverband.org/pressemitteilungen/2018_09_17_future_skills. Zugegriffen: 30. Nov. 2020
6. Saleh H (2018) Machine learning fundamentals. Packt, Birmingham
7. Erl T, Buhler P, Khattak W (2016) Big data fundamentals: concepts, drivers & techniques. Prentice Hall, Boston
8. Saltz J, Shamshurin I, Crowston K (2017) Comparing data science project management methodologies via a controlled experiment. In: Proceedings of the 50th Hawaii International Conference on System Sciences Waikoloa Beach
9. Géron A (2018) Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme, 1. Aufl. O'Reilly, Heidelberg
10. Chollet F (2018) Deep learning with Python. Safari Tech Books Online. Manning, New York
11. Feurer M, Klein A, Eggensperger K et al (2019) Auto-sklearn: efficient and robust automated machine learning. In: Hutter F, Kotthoff L, Vanschoren J (Hrsg) Automated machine learning. Springer, Cham, S 113–134
12. Bergstra J, Komer B, Eliasmith C et al (2015) Hyperopt: a Python library for model selection and hyperparameter optimization. Comput Sci Discov 8:14008
13. Jin H, Song Q, Hu X (2019) Auto-Keras: an efficient neural architecture search system. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, S 1946–1956
14. He X, Zhao K, Chu X (2019) AutoML: a survey of the state-of-the-art
15. Lee D, Macke S, Xin D et al (2019) A human-in-the-loop perspective on autoML: milestones and the road ahead. IEEE Data Eng Bull 42:59–70
16. Elshawi R, Maher M, Sakr S (2019) Automated machine learning: state-of-the-Art and open challenges
17. Ribeiro M, Grolinger K, Capretz MA (2015) MLaaS: machine learning as a service. In: 2015 IEEE 14th International Conference on Machine Learning and Applications. IEEE, S 896–902
18. Buxmann P, Schmidt H (2018) Grundlagen der Künstlichen Intelligenz und des Maschinellen Lernens. In: Buxmann P, Schmidt H (Hrsg) Künstliche Intelligenz: Mit Algorithmen Zum Wirtschaftlichen Erfolg. Springer Gabler, Wiesbaden, S 3–19
19. Strobel G, Paukstadt U, Becker J et al (2019) Von smarten Produkten zu smarten Dienstleistungen und deren Auswirkung auf die Wertschöpfung. HMD 56:494–513
20. van Astyne MW, Parker GG, Choudary SP (2016) Pipelines, platforms, and the new rules of strategy. Harv Bus Rev 94:54–62
21. Alpar P, Alt R, Bensberg F et al (2016) Anwendungsorientierte Wirtschaftsinformatik: Strategische Planung, Entwicklung und Nutzung von Informationssystemen, 8. Aufl. Springer Vieweg, Wiesbaden
22. Chen H-M, Schutz R, Kazman R et al (2017) How Lufthansa capitalized on big data for business model renovation. MISQE 16:Article 4
23. Ang S, Straub DW (1998) Production and transaction economies and IS outsourcing: a study of the U. S. banking industry. MISQ 22:535. https://doi.org/10.2307/249554
24. BMWi (2019) Die volkswirtschaftliche Bedeutung von digitalen B2B-Plattformen im Verarbeitenden Gewerbe. https://www.de.digital/DIGITAL/Redaktion/DE/Digital-Gipfel/Download/2019/digitale-b2b-plattformen-im-verarbeitenden-gewerbe.pdf?__blob=publicationFile&v=3. Zugegriffen: 2. Febr. 2021
25. Benner MJ, Tushman ML (2015) Reflections on the 2013 decade award—"exploitation, exploration, and process management: the productivity dilemma revisited" ten years later. AMR 40:497–514. https://doi.org/10.5465/amr.2015.0042
26. Haucap J (2020) Plattformökonomie: neue Wettbewerbsregeln – Renaissance der Missbrauchsaufsicht. Wirtschaftsdienst 100:20–29. https://doi.org/10.1007/s10273-020-2611-9
27. Tabarrok A, Cowen T (2015) The End of Asymmetric Information. Cato Unbound J Debate. http://www.cato-unbound.org/2015/04/06/alex-tabarroktyler-cowen/end-asymmetric-information
28. Xinjian D, Junhai M (2011) Based on the theory of principal-agent model of enterprise outsourcing services platform's game complexity study. In: Zhang J (Hrsg) Applied informatics and communication: International conference, ICAIC 2011 Xi'an, August 20–21, 2011. proceedings, Bd. 228. Springer, Berlin, S 606–613
29. Schwartmann R, Weiß S (2017) Whitepaper zur Pseudonymisierung der Fokusgruppe Datenschutz der Plattform Sicherheit, Schutz und Vertrauen für Gesellschaft und Wirtschaft im Rahmen des Digital-Gipfels 2017, S 1–47
30. Hoßfeld T, Hirth M, Tran-Gia P (2012) Crowdsourcing. Informatik Spektrum 35:204–208. https://doi.org/10.1007/s00287-012-0610-y
31. Gonzalez R, Gasco J, Llopis J (2005) Information systems outsourcing risks: a study of large firms. Ind Manag Data Syst 105:45–62