HMD (2023) 60:144–161
https://doi.org/10.1365/s40702-022-00842-3

SPEKTRUM

Hierarchische Eignungsprüfung von externen (Open) Data Sets für unternehmensinterne Analytics- und Machine-Learning-Projekte

Matthias Kaiser · Dominic Stirnweiß · Lars Wederhake

Eingegangen: 26. Oktober 2021 / Angenommen: 1. Februar 2022 / Online publiziert: 7. März 2022

Zusammenfassung Unternehmen erkennen zunehmend die Bedeutung evidenzbasierter Entscheidungen. Insbesondere die zunehmende Nutzung unternehmensexterner und offener Datensätze (Open Data) fördert die Möglichkeiten evidenzbasierter Entscheidungen. Dabei basieren evidenzbasierte Entscheidungen mit diesen Datensätzen immer häufiger auf Analysen, welche mittels maschineller Lernverfahren bzw. Machine Learning (ML) vorbereitet oder durchgeführt werden. Weil der Inhalt und die Qualität und damit der Nutzen eines Datensatzes für solche Analyseverfahren im Vorfeld ungewiss ist, stellt die Auswahl und die Beschaffung von geeigneten Daten unabhängig vom ML-Verfahren eine Kernherausforderung dar. Dieser Beitrag stellt deshalb zum Zwecke der Effizienz ein hierarchisches Vorgehen vor. Mit diesem können schemabasierte Datensätze strukturiert und effektiv dahingehend überprüft werden, ob deren Qualität und inhaltliche Fit für einen bestimmten Anwendungsfall (z. B. eine wiederkehrende Entscheidungssituation) ausreichend ist. Im Beitrag beschreiben wir einen Anwendungsfall aus dem Bereich der datengestützten Energieverbrauchsprognose für Wohngebäude, bei dem der Aufwand für die Datensatzauswahl reduziert werden konnte.

Schlüsselwörter Datenqualität · Datenqualitätsbewertung · Data Analytics · Open Data · Machine Learning

Matthias Kaiser (✉) · Lars Wederhake
Project Group Business and Information Systems Engineering, Fraunhofer Institute for Applied Information Technology FIT, Augsburg, Deutschland
E-Mail: matthias.kaiser@fit.fraunhofer.de

Dominic Stirnweiß
Adesso SE, Adessoplatz 1, 44269 Dortmund, Deutschland

Lars Wederhake
FIM Research Center, University of Bayreuth, Bayreuth, Deutschland

credium GmbH, Katharinengassse 13, 86150 Augsburg, Deutschland

Hierarchical Suitability Check of External (Open) Data Sets for Internal Analytics and Machine Learning Projects

Abstract Companies are increasingly recognizing the importance of evidence-based decisions. In particular, the increasing use of company-external and open data sets (open data) additionally promotes the possibilities of evidence-based decisions. More often, evidence-based decisions on these data sets are based on analyses that are prepared or carried out using machine learning (ML). Because the content and quality and thus the usefulness of a dataset for such analysis are uncertain in advance, the selection and acquisition of suitable data is a core challenge independent of the specific ML procedure. This paper therefore presents a hierarchical and efficiency-oriented procedure to check schema-based data sets in a structured and effective way to determine whether their quality and content fit is sufficient for a specific use case (e.g. a recurring decision situation). In the paper, we describe a use case from the field of data-based energy consumption forecasting for residential buildings, where the effort for data set selection could be reduced.

Keywords Data quality · Data quality assessment · Data analytics · Open Data · Machine Learning

1 Motivation

Evidenzbasierte Entscheidungen – Entscheidungen auf Basis von Daten anstelle von Erfahrung und „Bauchgefühl“ – sind branchenübergreifend zu einem Geschäftsziel mit höchster Priorität für Unternehmen geworden (Dursun 2019). Die steigende Verfügbarkeit an unternehmensinternen, vor allem aber -externen Daten ist Voraussetzung für evidenzbasierte Entscheidungen (Donnelly et al. 2021). Insbesondere die Entwicklung zur verbreiteten Nutzung offener Datensätze (Open Data Sets) unterstützt evidenzbasierte Entscheidungen maßgeblich (Binzen 2021). Dabei basieren evidenzbasierte Entscheidungen immer häufiger auf Analysen, welche mittels maschineller Lernverfahren beziehungsweise Machine Learning (ML) auf diesen Datensätzen vorbereitet oder durchgeführt werden (Hofmann et al. 2020). Genauer werden Datensätze eingesetzt, um ML-Verfahren zu trainieren und deren Genauigkeit zu validieren.

Die Auswahl und Beschaffung von Datensätzen stellt jedoch unabhängig vom ML-Verfahren eine Kernherausforderung dar (Agrawal et al. 2019). Dies gelingt regelmäßig für unternehmensinterne, bekannte und ggf. umfangreich dokumentierte Datensätze. Für die immer umfänglicher verfügbaren unternehmensexternen Datensätze wird das Potenzial jedoch (noch) nicht voll ausgeschöpft (Döbel et al. 2018; Reinsel et al. 2018). Ein Grund ist, dass Inhalt und Qualität dieser unternehmensexternen Datensätze vor der Beschaffung meist unbekannt sind und somit der zu erwartende Nutzen unsicher ist (Agrawal et al. 2019). Die Beschaffung unternehmensexterner Datensätze ist darüber hinaus nicht selten mit Kapitalinvestitionen, wie beispielsweise Lizenzgebühren oder aber zumindest mit Personalaufwand im Rahmen des Beschaffungsprozesses (insbesondere für die Analyse), verbunden. Ge-

geben der Kosten wirft dies die Frage auf, wie die Unsicherheit bzgl. dem Nutzen bei der Beschaffung dieser Datensätzen reduziert werden kann.

Dieser Beitrag stellt ein hierarchisches Vorgehen vor, mit dem Datensätze strukturiert und effektiv dahingehend überprüft werden können, ob deren Qualität und inhaltliche Fit für einen bestimmten Anwendungsfall (z. B. eine wiederkehrende Entscheidungssituation) ausreichend ist, d. h. ob eine Eignung vorliegt. Das entwickelte Vorgehen basiert auf dem bewährtem Datenbewertungsframework von Wang und Strong (1996) sowie darauf aufbauenden Forschungsarbeiten (Batini et al. 2009; Otto et al. 2007; Görz und Kaiser 2012; Heinrich et al. 2007; Jarke et al. 1999). Das Framework wird in dieser Arbeit ML-spezifisch für unbekannte und unternehmensexterne Datensätze weiterentwickelt und in ein prozessuales Vorgehen eingebettet. Die Validierung des Vorgehens fand in einem angewandten und öffentlichen Forschungsprojekt statt, in dem ML-Verfahren zur Bewertung energetischer Sanierungsmaßnahmen bei privaten Ein- und Zweifamilienhäusern erforscht und entwickelt worden sind. Dieses Projekt dient diesem Beitrag als Fallbeispiel.

2 Grundlagen der Eignungsprüfung und Stand der Forschung

Zur Realisierung von evidenzbasierten Entscheidungen müssen Unternehmen Daten und Datensätze verfügbar machen, verarbeiten und analysieren (Kessler und Gómez 2020). Dabei spielen wie eingangs dargestellt ML-Verfahren eine zunehmend wichtige Rolle (Polyzotis et al. 2018). Die Umsetzung solcher ML-Verfahren findet in der Regel in definierten Projekten statt. In wissenschaftlichen Kreisen wird dazu in der Regel auch heute – mehr als 20 Jahre nach Veröffentlichung – auf den Cross Industry Standard Process for Data Mining (CRISP-DM) verwiesen (Martinez-Plumed et al. 2021; Wirth und Hipp 2000). Dabei wird CRISP-DM als Vorgehensmodell aufgrund seiner Branchen- und Technologieunabhängigkeit wertgeschätzt. Für CRISP-DM gilt zudem, dass die Projektzieldefinition zu Beginn eines Projekts zunächst scharf ausgearbeitet werden soll. Dies charakterisiert die Projektart, die auch diesem Beitrag zugrunde liegt. Für primär explorative Projekte und Methoden verweisen wir auf (Martinez-Plumed et al. 2021). CRISP-DM ist weniger explizit, wenn es um die Festlegung der Datenbeschaffung geht. In der Praxis werden stattdessen häufig feingranulare Projektvorgehensmodelle angewendet – so z. B. der Team Data Science Process (TDSP) – eine agile und iterative Methodik zur Entwicklung von ML-Lösungen (Microsoft 2020). Das mehrphasige Vorgehen betont in der Phase *Data Acquisition & Understanding* die Wichtigkeit von externen Datensätzen und deren Datenqualität zum Trainieren von ML-Modellen. Dabei geht TDSP nicht weiter auf die Implikationen im Beschaffungsprozess ein. Ein detaillierteres Vorgehen, hinsichtlich der Datenverfügbarkeit und -beschaffung, gibt dafür der Domino Data Science Life Cycle (DDSL) vor (Domino Data Lab 2017). Dieser stellt heraus, dass der Nutzen externer Datensätze bei der Datenakquise häufig unbekannt ist. Dies könnte neben den Beschaffungskosten zu weiteren Kosten, wie z. B. für aufwändige Datenaufbereitung und/oder -nacherhebung, führen, wenn sich der Datensatz während eines Projekts als ungeeignet herausstellt (Domino Data Lab 2017). Die Notwendigkeit einer Eignungsprüfung von Datensätzen, im Rahmen der

Beschaffung von Datensätzen, wird in beiden Vorgehensmodellen hervorgehoben. Ein konkretes Vorgehen, in welchem Umfang Datensätze und welche Bestandteile in welcher Reihenfolge nach welchem Vorgehen untersucht werden müssen, stellen diese Ansätze jedoch auch nicht vor. Polyzotis et al. (2018) greifen diese Problematik auf und führen in ihrem Lebenszyklusmodell für ML-Projekte deshalb eine eigene Phase *Sanity Check* ein. Diese hat zum Ziel, Datensätze deskriptiv zu analysieren, um die Qualität eines Datensatzes und die Aussagekraft der Datenquelle vor der Nutzung abschätzen zu können. In der praktischen Anwendung muss der Nutzen von Datensätzen für konkrete Projekte schnell und einfach evaluierbar sein. Der *Sanity Check* beinhaltet zwar beispielhafte Evaluierungsmechanismen, aber ist damit für die Praxis nur bedingt direkt anwendbar.

In diesem Zusammenhang können Datenqualitätsmanagement (DQM) -Ansätze zum Einsatz kommen. DQM-Ansätze strukturieren die Planung, die Durchführung und die Überprüfung von von Datenqualitätssicherungsmaßnahmen in Unternehmen. Cichy und Rassa (2019) unterscheiden bei DQM-Ansätzen zwischen allgemein anwendbaren und spezialisierten Ansätzen, wie derjenige in diesem Beitrag. Unter den allgemeinen Ansätzen genießt Total Data Quality Management (TDQM) (Wang 1998; Lee et al. 1998), als einer der frühen Ansätze, laut Otto et al. (2007) und Weber et al. (2009) größte Popularität. Daneben wurden in jüngerer Zeit, unter der Berücksichtigung neuerer Erkenntnisse, die Observe-Orient-Decide-Act Methodology for Data Quality (OODA DQ) (Sundararaman und Venkatesan 2017) und die Task-based Data Quality Method (TBDQ) (Vaziri et al. 2016) vorgestellt. Umfangreich diskutiert werden in der wissenschaftlichen Literatur das Data Quality Assessment Framework (DQAF) (Sebastian-Coleman 2012) und die Comprehensive Methodology for Data Quality Management (CDQ) (Batini et al. 2006). Aber auch TDQM und das von English (1999) präsentierte Total Information Quality Management (TIQM) haben heute, mehr 20 Jahre nach ihrer Vorstellung, weiterhin Relevanz. Für eine umfassende Analyse und den komparativen Vergleich von verbreiteten allgemeinen DQM-Ansätzen verweisen wir auf Cichy und Rass (2019) sowie auf Batini und Scannapieco (2016). Den zuvor referenzierten DQM-Ansätzen ist gemein, dass ihnen eine Datenqualitätsanalyse entlang mehrerer Datenqualitätsdimensionen zugrunde liegt. So referenziert TDQM analog zu diesem Beitrag die von Wang und Strong (1996) beschriebenen Datenqualitätsdimensionen. Die Dataqualitätsanalyse stellt das verbindende Glied zwischen den Ansätzen dar. Der Anlass und der Kontext, zu denen die DQM-Ansätze angewendet werden, variieren. Die unterschiedlichen Eigenschaften der Ansätze können, je nach Anlass und Kontext, mehr beziehungsweise weniger geeignet sein. So referenzieren Cichy und Rass (2019) dreizehn spezialisierte Ansätze mit Fokus auf bestimmte Kontexte (Medizin, Finanzen, Produktion), Informationssystemen (Web, Data Warehouses, etc.) und Organisationstypen (z. B. Interorganisationale Strukturen). Jedoch findet sich auch unter dieser aktuellen Zusammenstellung kein auf Analytics- und ML-Projekte spezialisierter Ansatz, obwohl die speziellen Eigenschaften bzw. Anforderungen solcher Projekte einen Bedarf an einem zugeschnittenen DQM-Ansatz begründen können.

Zielorientierung und Datennutzung Analytics- und ML-Projekte zielen in der Regel auf einen spezifischen Informationsbedarf einer Nutzergruppe ab. Sie kön-

nen ihr Ergebnis dann als Informationsprodukt managen (Wang 1998). Projekte mit Zielorientierung liegen diesem Beitrag zugrunde. Nicht im Fokus dieses Beitrags stehen solche Analytics- und ML-Projekte, welche über die Datenexploration der Frage nachgehen: „für welche Einsatzzwecke können bestehende Daten zuträglich sein“. Im Sinne der Zielorientierung legen Vorgehensmodelle für Analytics- und ML-Projekte üblicherweise – so auch CRISP-DM – zu Beginn eines Projekts einen Schwerpunkt auf die Konkretisierung und Definition der Zielsetzung. Mit dieser Zielorientierung können spezialisiertere Ansätze eine Datenqualitätsprüfung kontextspezifisch und im Sinne des Einsatzzwecks („Fit-for-Use“) optimieren (Wang und Strong 1996; Lee et al. 2002). Allgemeine Ansätze müssen den Anspruch berücksichtigen, Datenqualität für unterschiedliche Datenabnehmer bereitzustellen. Die Datenqualität auf ein einzelnes Ziel wie in einem Projekt auszurichten, könnte dem entgegenstehen oder zumindest aus Gründen der Koordination der unterschiedlichen Datenabnehmer zu zusätzlichem Aufwand führen.

Zudem steht im Zentrum eines Analytics- und ML-Projekts typischerweise die Entwicklung eines bzw. mehrerer ML-Modelle auf Basis maschineller Lernverfahren, die auf Daten(-sätzen) trainiert wurden (Bhavsar und Ganatra 2012). Dieser Fokus zeigt zwei, von den allgemeinen Ansätzen potenziell abweichende, Eigenschaften auf: Zum einen bemisst sich der Projekterfolg indirekt über die Güte der Modellausgaben (Keim und Sattler 2021), wohingegen in operativen Datenverarbeitungssystemen und klassischen dispositiven Datenhaltungssystemen in der Regel die Güte der Daten direkt Analyse- und Bewertungsgegenstand ist. Zum anderen ist für Analytics- und ML-Projekte die Erstellung eines Trainings- und Validierungsdatensatzes sowie gegebenenfalls eines separaten Evaluierungsdatensatzes eine Kernherausforderung. Dabei sind Trainings- und Validierungsdaten potenziell exklusiv für diesen Anwendungszweck zusammenzustellen. Diese Daten können unternehmensextern bezogen oder ergänzt werden. Dieser Vorgang ist charakteristisch für Analytics- und ML-Projekte und höchst zeitaufwändig (Ratner et al. 2017). Daher könnte ein zugeschnittener Ansatz bei dieser Zusammenstellung zusätzliche Struktur und damit Unterstützung geben.

Datenstruktur und Record Linkage Die meisten Ansätze fokussieren insbesondere strukturierte Daten in relationalen Datenbanken (Batini et al. 2009). In diesem Beitrag liegt der Fokus auf externen schemabasierten Datensätzen, die zumeist in Dateiformaten vorliegen (z. B. csv, parquet) und deren Schema beim Speichern definiert wird. Zusätzlich erfordert die Analyse von externen Datensätzen die Berücksichtigung von Verknüpfungen zwischen zwei oder mehreren physisch getrennten Datensätzen. Durch Objektidentifikation können Eigenschaften eines Objekts der Realwelt zusammengeführt werden. Dieser Vorgang wird in der Literatur als Record Linkage bezeichnet (Fellegi und Sunter 1969). Da dies eine typische Aktivität bei der Prüfung von externen und gegebenenfalls offenen Datensätzen darstellt, sollte ein geeigneter Ansatz dies berücksichtigen können.

Umfang und Bezugsrahmen Schließlich beschreibt der organisationale, zeitliche und budgetäre Rahmen eines unternehmensinternen Analytics- und ML-Projektes andere Anforderungen an das DQM, als ein fortdauerndes unternehmensweites oder

unternehmensübergreifendes DQM, das vielen allgemeinen Ansätzen als angenommener Rahmen unterliegt (Otto 2012). Analytics- und ML-Projekte sind zeitlich begrenzt, häufig iterativ und werden in einzelnen meist cross-funktionalen Teams ausgeführt (Saltz und Grady 2017). Aufwandsintensive Managementprozesse, insbesondere zur Datenqualitätsverbesserung in externen Systemen, können ineffektiv sein. Ein für Analytics- und ML-Projekte zugeschnittener Ansatz sollte daher auf eine strukturierte und effektive Prüfung abzielen. Datensätze, die nicht wertstiftend sind, sollten entsprechend möglichst früh im Prüfprozess verworfen werden können.

Mit Bezug auf die oben genannten Eigenschaften und im Abgleich mit existierenden DQM-Ansätzen nach Batini und Scannapieco (2016) und Cichy und Rass (2019) besteht der Bedarf an einem Ansatz für die Eignungsprüfung von schemabasierten externen Datensätzen für unternehmensinterne Analytics- und ML-Projekte. Konkret soll damit geprüft werden können, ob identifizierte **Datenquellen,** die für einen vorliegenden Verwendungszweck notwendigen **Inhalte** in ausreichender **Qualität** bereitstellen können. Die Datenquelle und das Vorhandensein der richtigen Inhalte können anhand der Metadaten geprüft werden, während die Datenqualität an den Nutzdaten festgestellt werden kann.

- Die **Datenquelle** zeichnet sich durch den Herkunftsort, an dem Daten entstanden sind oder Informationen digital verfügbargemacht wurden, und den Lieferanten, der den Datenzugang verantwortet, aus. Eine Datenquelle ist passend, wenn Daten vor dem Hintergrund ihrer Herkunft und des Herausgebers interpretiert werden können.
- Ein Datensatz wird durch seine Attribute beschrieben, welche dessen thematischen **Inhalte** charakterisieren. Ein Attribut besteht aus einer Attributbezeichnung, z. B. „Augenfarbe“, und seiner Attributausprägung, z. B. „blau, grün, braun“. Der inhaltliche Fit eines Datensatzes zum Verwendungszweck eines bestimmten Projekts ist dann gegeben, wenn die Attribute des Datensatzes den Projektkontext adäquat beschreiben.
- **Datenqualität** ist nach ISO-Norm (International Organization for Standardization 2008) definiert als der Grad, zu dem die Eigenschaften von Daten die Anforderungen erfüllen, die zur Lösung einer konkreten Problemstellung notwendig sind.

Dieser Artikel fokussiert die fachliche Eignung von Datensätzen, welche durch nicht fachliche Kriterien (z. B. Lizenzkosten oder Datenschutzanforderungen) ergänzt werden können. Für Praktiker stellt das in diesem Beitrag präsentierte und in der wissenschaftlichen Forschung entwickelte hierarchische Vorgehen ein Werkzeug dar, das Datensätze strukturiert und effektiv auf ihren anwendungsspezifischen Nutzen prüft und Entscheidungen bei der Beschaffung von Datensätzen rationalisiert. Im Folgenden wird das Vorgehen zur Eignungsprüfung präsentiert. Dieses wird für den jeweils vorliegenden Anwendungsfall adäquat adaptiert (siehe Kap. 4).

3 Vorgehen zur Eignungsprüfung von Datensätzen

Das entwickelte hierarchische Vorgehen zur Eignungsprüfung von schemabasierten Datensätzen gliedert sich in drei Phasen (vgl. Abb. 1):

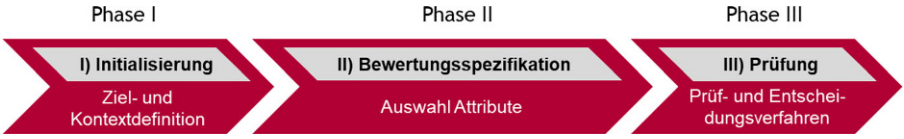

Abb. 1 Die drei Phasen zur kontextspezifischen Eignungsprüfung von Datensätzen

In **Phase I** (*Initialisierung*) wird zunächst der Kontext eines vorliegenden ML-Projekts definiert. Der Kontext eines Projekts beschreibt den thematischen Rahmen und die Projektziele.

In **Phase II** (*Bewertungsspezifizierung*) werden auf Basis des definierten Kontexts des Projekts die bewertungsrelevanten Attribute ausgewählt. Die in diesem Beitrag vorgestellten Bewertungsdimensionen werden für den Kontext geeignet spezifiziert.

In **Phase III** (*Prüfung*) werden schließlich das genaue Bewertungsvorgehen in einem Prüf- und Entscheidungsprozess definiert. Darin wird beschrieben in welcher Reihenfolge die Bewertungsdimensionen geprüft werden, sodass effizient entschieden werden kann, ob der Datensatz zur Erreichung des Gesamtprojektziels oder zumindest für Zwischen- bzw. Etappenziele geeignet ist.

Das Vorgehen ist sowohl für einzelne Datensätze als auch für miteinander verknüpfte Datensätze geeignet. Ein verknüpfter Datensatz ist ein aus mehreren Datensätzen – über eine eindeutige Objektreferenz – zusammengesetzter Datensatz (Record Linkage).

Nachfolgend erläutern wir den Prozess inklusive der konkreten Schritte, welche innerhalb der Phasen zum Einsatz kommen. Kap. 4 präsentiert die Anwendung des Vorgehens an einem Fallbeispiel (Eignungsprüfung im Bereich Energieeffizienz von Gebäuden).

Phase I: Initialisierung Die methodische Vielfalt von ML und die breite Einsatzmöglichkeit über alle Branchen hinweg, erfordert eine eindeutige Kontextdefinition eines Projekts oder einer Analyseanwendung. ML-Anwendungen können je nach Modell und Einsatzzweck sehr unterschiedliche Datenqualitätsanforderungen haben. Ebenfalls stellt das Einsatzgebiet unterschiedliche Anforderungen an die Inhalte eines Datensatzes. Beispielsweise gibt die Kundenklassifizierung in der Finanzdienstleistungsbranche andere inhaltliche Anforderungen an den Datensatz vor als die Prognose von Ausfallwahrscheinlichkeiten von Maschinen im produzierenden Gewerbe. Zur Eignungsprüfung wird also der sogenannte Kontext eines Projekts durch das Projektziel beschrieben. In einem Projekt kann außerdem die Definition von Etappenzielen nützlich sein, die zur Erreichung eines Gesamtprojektziels beitragen.

Phase II: Bewertungsspezifikation In Phase II werden auf Basis der Zieldefinition die konkreten Inhalte abgeleitet und spezifiziert, welche in einem Datensatz enthalten sein müssen und nach welchen Bewertungsdimensionen der inhaltliche Fit und die Qualität der Daten überprüft werden müssen.

Attribute beschreiben maßgeblich den Datensatz. Dabei können die Attribute in Basisattribute (BA) und Zusatzattribute (ZA) eingeteilt werden. BA beschreiben die-

jenigen Attribute, die auf jeden Fall in einem Datensatz vorhanden sein müssen, um das Gesamtprojektziel erreichen zu können. Attribute, die zusätzlich zur Problemlösung beitragen, werden als ZA aufgefasst. Wichtig ist, dass sowohl BA als auch ZA von der zu lösenden Problemstellung abgeleitet werden. Denn je nach Projektkontext sind andere Inhalte und damit Attribute essenziell relevant oder haben ergänzenden Charakter. Letzteres liegt vor, wenn z. B. eine weitere Verbesserung bezüglich des Gesamtprojektziels oder eines Etappenziels damit erreicht werden kann.

Der inhaltliche Fit eines Datensatzes kann folglich anhand des Vorhandenseins von BA und ZA festgestellt werden. Somit kann auch durch das Fehlen oder fehlerhafte Vorliegen der BA die Nichteignung eines Datensatzes festgestellt werden. Um das Vorhandensein von Attributen zu überprüfen und deren Qualität effizient zu bewerten, sind die richtigen Bewertungsdimensionen notwendig. Das von Wang und Strong (1996) entwickelte allgemeine Datenqualitätsframework umfasst mehrere Bewertungsdimensionen. Alle acht konkret definierbaren Bewertungsdimensionen sind berücksichtigt und für die Eignungsprüfung von Datensätzen ausgewählt worden (vgl. Tab. 1). Die Bewertungsdimensionen werden dabei in drei Hierarchieebenen eingeteilt (vgl. Abb. 2).

- Der **Datenquellenebene** sind die Dimensionen Reputation und Objektivität zugeordnet, da diese die Qualität der Datenquelle auf Basis ihrer Herkunft bzw. ihres Lieferanten bewerten.
- Auf der **Datensatzebene** darunter bewerten die Dimensionen Verständlichkeit, Relevanz und Vollständigkeit, ob der Datensatz die notwendigen Inhalte enthalten kann.
- Auf **Attributebene** sind die Bewertungsdimensionen Interpretierbarkeit, Fehlerfreiheit und Aktualität, die die Eignung eines Attributs auf Basis der Qualität ihrer Datenwerte bewertet, anzuwenden.

Für eine konkrete Quantifizierung der Bewertungsdimensionen existieren in der Literatur eine ganze Reihe von Ansätzen (Helfert 2002; Hinrichs 2002; Lee et al. 2002). Eine Kernherausforderung ist, dass die Messung der Datenqualität einiger Bewertungsdimensionen auf subjektiven Qualitätseinschätzungen basiert und Datennutzer je nach konkretem Anwendungszweck häufig unterschiedliche Zielsetzungen verfolgen (Heinrich und Klier 2011).

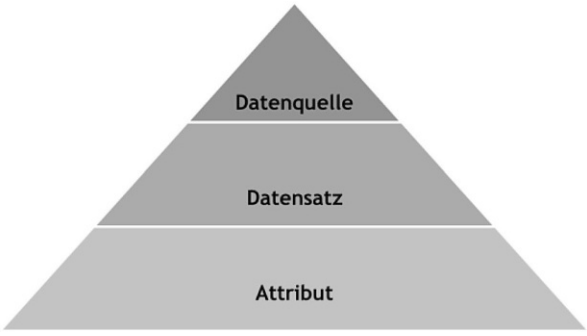

Abb. 2 Hierarchieebenen der Bewertungsdimensionen zur Eignungsprüfung von Datensätzen

Tab. 1 Bewertungsdimensionen auf Basis von Wang und Strong (1996) für die vorgestellte Eignungsprüfung

Bewertungs-dimension	Beschreibung
Datenquellenebene	
Reputation	Ein Datensatz besitzt eine hohe Reputation, wenn die Datenquelle bzw. die herausgebende Instanz einen hohen sozialen Status und eine große gesellschaftliche Akzeptanz erfährt. Bspw. ist ein Datensatz, der von einer renommierten Universität unter Nennung der Nutzungsbedingungen veröffentlicht wird, reputierlicher als ein Datensatz, der ohne Angaben der Nutzungsbedingungen von unbekannten Autoren veröffentlicht ist
Objektivität	Ein Datensatz gilt als objektiv, wenn die enthaltenen Daten sachlich und wertfrei veröffentlicht sind. Datensätze, die bspw. von einem Unternehmen für unternehmerische Zwecke veröffentlicht wurden, tragen das Risiko dem Anspruch der Objektivität weniger bzw. nicht gerecht zu werden. Im Vergleich dazu können Datensätze, die von Forschungsinstituten für öffentlichen Erkenntnisgewinn bereitgestellt werden, objektiver bewertet werden
Datensatzebene	
Verständlichkeit	Ein Datensatz ist verständlich, wenn die Bezeichner und Beschreibungen identifiziert und gedeutet werden können sowie wenn dadurch die Objekteigenschaften aus der Realwelt, welche durch die Attribute repräsentiert werden, verstanden werden. Zum Beispiel ist ein Datensatz verständlich, wenn kodierte Attributbezeichnungen über zusätzlich vorliegende Legenden dekodiert werden können. Ist eine Dekodierung und damit die Deutung der Attributbezeichnungen nicht möglich, gilt der Datensatz als nicht verständlich
Relevanz	Ein Datensatz gilt als relevant, sobald alle Basisattribute (BA) im Datensatz vorhanden sind, sodass die definierten (Etappen-) Ziele erreicht werden können. Zum Beispiel muss ein Datensatz für Prognosezwecke die prognostizierende Zielvariable zwingend enthalten, um relevant zu sein
Vollständigkeit	Die Vollständigkeit misst den Grad an identifizierten Zusatzattributen in einem Datensatz. Sind zur Gesamtzielerreichung alle notwendigen ZA im Datensatz enthalten, gilt er als vollständig, während ein teilweise vollständiger Datensatz nur ZA zur Erreichung von Etappenzielen enthält. Ist mit den vorhandenen ZA kein definiertes Ziel erreichbar, gilt der Datensatz als unvollständig
Attributebene	
Interpretierbarkeit	Ein Attribut gilt als interpretierbar, wenn die einzelnen Datenwerte in ihrer vorliegenden Form korrekt gedeutet werden können. Kodierungen – text- oder zahlenbasierte Kategorisierung von Rohdaten – müssen einheitlich sein und dekodiert werden können. Zu Größen- oder Mengenparametern muss eine Einheit zuordenbar sein
Aktualität	Ein Datensatz ist aktuell, wenn die (kontextspezifischen) Anforderungen an den zeitlichen Bezug von Attributen erfüllt sind. Zum Beispiel könnte die Anforderung sein, dass die Daten von einem oder mehreren Attributen einen Zeitstempel tragen müssen oder die Datenerhebung aus einem bestimmten Zeitraum stammt
Fehlerfreiheit	Die Fehlerfreiheit misst für jedes Attribut die Anzahl an Datenwerte, die innerhalb des (von Domänenexperten gesetzten) Definitionsbereiches liegen. Für einige Anwendungsfälle ist darüber hinaus der Grad der Anzahl von fehlenden Datenwerten gegenüber aller Datenwerten eines Attributs zu messen. Insb. Außerhalb der Definitionsbereiche liegen fehlende Datenwerte z. B. gekennzeichnet durch „N/V“, „N/A“, „“, „NULL“, „–1“ „999999“, etc

Während Heinrich und Klier (2011) konkrete Metriken zur Quantifizierung der Bewertungsdimensionen Vollständigkeit, Fehlerfreiheit und Aktualität vorstellt, existieren für die stärker qualitativ zu bemessenden und kontextabhängigen Bewertungsdimensionen wie Reputation, Objektivität, Verständlichkeit, Relevanz und Interpretierbarkeit, vermehrt Leitplanken und Methoden der Konsensbildung zur Qualitätsbemessung (Helfert 2002; Lee et al. 2002). Da der Fokus des vorliegenden Beitrags auf der Eignungsprüfung von Datensätzen für einen bestimmten Anwendungsfall liegt, steht die Quantifizierung einzelner Bewertungsdimensionen nicht im Vordergrund. Um Praktikern ein Gefühl zu vermitteln, nach welchen Kriterien die ausgewählten Bewertungsdimensionen bemessen werden können, werden in Tab. 1 die Bewertungsdimensionen beispielhaft beschrieben.

Die Eignungsprüfung anhand dieser Bewertungsdimensionen setzt voraus, dass auf einen Datensatz zugegriffen wird oder dieser Verfügbar gemacht werden kann. Das mag in der Praxis mit Administrations- und Orchestrierungsaufwand verbunden sein. Jedoch sollte eine rationale Entscheidung – ganz im Sinne der evidenzbasierten Entscheidung – auf vorliegenden Daten basieren.

Phase III: Prüfung In Phase III des Vorgehens erfolgt die Entwicklung eines Prüf- und Entscheidungsprozesses, der die Anwendung der Bewertungsdimensionen vorgibt (vgl. Abb. 3). Wichtig ist, dass vor der Prüfung Datensätze über gemeinsame Objekte referenziert werden. Der Prüfprozess besteht aus fünf Schritten, welche sich aus Effizienzgründen an den Hierarchieebenen orientieren.

Im ersten Schritt **(1)** findet die **Bewertung der Datenquelle** statt. Ein Datensatz besitzt *Reputation*, wenn die Datenquelle in der einschlägigen Community des thematischen Kontexts bekannt ist und die Datenschutzkonformität über den gesamten Datenerhebungs- und Aufbereitungsprozess transparent dargelegt ist. Die *Objektivität* eines Datensatzes hängt maßgeblich davon ab, ob der Herausgeber den Datensatz aus eigennützigem Interesse veröffentlicht. Öffentlichen Einrichtungen, Universitäten oder Forschungsinstituten verfolgen in der Regel weniger subjektive Interessen als privatwirtschaftliche Unternehmen, weshalb die Art des Herausgebers ein wichtiger Indikator für die Objektivität von Daten sein kann. Falls ein Datensatz aus

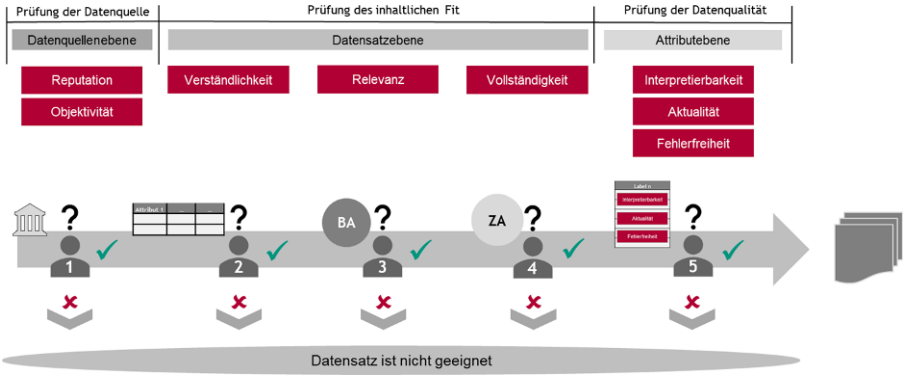

Abb. 3 Hierarchischer Prüf- und Entscheidungsprozess zur kontextspezifischen Eignungsprüfung von Datensätzen

einer nicht reputierlichen Quelle stammt oder wenig objektive Daten enthält, müssen alle darauf aufbauenden Analysen und Ergebnisse entsprechend der Reputation und Objektivität interpretiert werden.

Im Sinne der Effizienz wird auf **Datensatzebene** zuerst geprüft, ob die Attribute *verständlich* sind, um damit Klarheit über den Inhalt eines Datensatzes zu schaffen, bevor BA und ZA ausgewählt werden können **(2)**. Dazu wird untersucht, in welcher Form die Attributbezeichnungen vorliegen und ob diese direkt und intuitiv verstanden werden, maschinell verarbeitbar sind oder dekodiert vorliegen. Können die Attribute eines Datensatzes nicht verstanden oder interpretiert werden, können die Inhalte nicht mit Sicherheit identifiziert werden und der Datensatz ist ungeeignet. Anschließend wird der Datensatz nach BA durchsucht **(3)**. Nur wenn diese vollständig im Datensatz vorhanden sind, gilt der Datensatz als *relevant*. Fehlt mindestens ein BA besteht der Datensatz die Relevanzprüfung nicht und ist ungeeignet. Sind die BA vorhanden, wird bei der *Vollständigkeitsprüfung* evaluiert, welche ZA der Datensatz enthält **(4)**. Abhängig von den vorhandenen ZA können Gesamt- oder Etappenziele erreicht werden. Das alleinige Vorhandensein der BA und ZA reicht in der Regel noch nicht aus, um abschließend über die Eignung des Datensatzes zu urteilen.

Auf **Attributebene** müssen die Datenwerte eines Attributs *interpretierbar, aktuell* und ausreichend *fehlerfrei* vorliegen **(5)**. Bspw. könnten die Anforderungen sein, dass die fehlenden Dateneinträge innerhalb eines Attributs einen gewissen relativen oder absoluten Anteil nicht überschreiten dürfen oder, dass Dateneinträge einen bestimmten Zeitstempel tragen müssen. Außerdem könnten funktionale Abhängigkeiten zu prüfen sein, falls mehrere Attribute gleichzeitig Datenwerte aufweisen müssen.

4 Anwendungsfall: Eignungsprüfung im Bereich Energieeffizienz von Gebäuden

Im Folgenden wird obiges Vorgehen im Kontext eines realen ML-Projekts angewendet. Dabei werden die Erzeugung einer konkreten Instanz des Schemas und die einzelnen Schritte am konkreten Beispiel veranschaulicht. Bei dem herangezogenen Beispielprojekt handelt es sich um ein Konsortial-Forschungsprojekt aus dem Bereich der Energieeffizienz von Gebäuden. Das Ziel des Forschungsprojekts ist es, Energieverbräuche von Ein- und Zweifamilienhäusern vor und nach der Durchführung energetischer Gebäudesanierungsmaßnahmen unter Berücksichtigung des Verhaltens der Bewohner zu prognostizieren. Zur Energieverbrauchsprognose soll ein ML-Ansatz verwendet werden, der im Stande ist, Zusammenhänge zwischen dem Energieverbrauch, Gebäudeparametern, Bewohnerinformationen und Informationen zu Sanierungsmaßnahmen zu erkennen. Diese Definition des Gesamtprojektziels ist bereits Teil der Initialisierung (Phase I) und beschreibt den Kontext des Projekts. Zur Verringerung der Komplexität und der Entwicklungsrisiken wurden im Projekt zwei Etappenziele definiert. *Etappenziel 1* beschränkt sich auf die Energieverbrauchsprognose, auf Basis von Gebäudeparameter und Bewohnerinformationen, ohne jedoch

Kategorie	Attribut	Gesamtziel	Etappenziel 1	Etappenziel 2
Energieverbrauchsdaten				
BA	Energieverbrauch			
Gebäudeeigenschaften				
BA	Gebäudetyp (EH/ZH)			
ZA	Wohnfläche			
ZA	Baujahr			
ZA	Dachform			
ZA	…			
Bewohnerdaten				
ZA	Anzahl der Bewohner			
ZA	Alter der Bewohner			
ZA	Anzahl minderjähriger Kinder			
ZA	…			
Sanierungsdaten				
ZA	Sanierungsmaßnahme (ja/nein)			
ZA	Art der Sanierung			
ZA	Zeitpunkt der Sanierung			
ZA	…			

Abb. 4 Auszug der Basisattribute (BA) und Zusatzattribute (ZA) des Beispielprojekts

Sanierungsmaßnahmen zu bewerten. Weiterhin stellt *Etappenziel 2* die Energieverbrauchsprognose auf Basis von Gebäudeparameter dar.

Anhand der Zielsetzung werden nachfolgend die Inhalte in Form von BA und ZA abgeleitet, die in einem Datensatz enthalten sein müssen, um für das Projekt geeignet zu sein (Phase II). Der Energieverbrauch eines Gebäudes ist im Projekt das zentrale Prognoseziel, weshalb der Energieverbrauch eines Gebäudes als BA definiert wird. Als weiteres BA wird der Gebäudetyp definiert, da aufgrund der Projektzielsetzung eindeutig die Zuordnung des Energieverbrauchs zu einem Ein- oder Zweifamilienhaus gegeben sein muss. Alle weiteren Attribute zu Gebäudeeigenschaften, Bewohnerdaten und Sanierungsmaßnahmen sind ZA. Abb. 4 zeigt einen Auszug der im Beispielprojekt definierten BA und ZA sowie deren Notwendigkeit zur Erreichung der Projektziele.

Um auf dieser Basis die Eignung von Datensätzen im Beispielprojekt feststellen zu können, werden der generische Prüfprozess (vgl. Abb. 3) und die definierten Bewertungsdimensionen (vgl. Tab. 1) angewendet (Phase III). Dazu wird im Folgenden

zunächst die Datenbeschaffung näher erläutert und anschließend als projektspezifischer Prüfprozess präsentiert.

Datenbeschaffung und Eignungsprüfung im Beispielprojekt Die Identifikation von potenziellen Datenquellen erfolgte systematisch und orientierte sich an Informations- und Datenlebenszyklusmodellen. Als Basis dazu diente das Lebenszyklusmodell der Informationswirtschaft von Krcmar (2015) und das POSMAD-Lebenszyklusmodell von MCGillvry (2008), welches in den ersten beiden Phasen insbesondere Planungsmaßnahmen, wie z. B. definieren von Zielen und Standards, und den Datenbeschaffungsprozess adressiert. Im Projekt wurden zunächst potenzielle Nutzer von Wohngebäudedaten identifiziert und in Interviews über Datenquellen und -lieferanten befragt. Open Data Sets wurden direkt eingeladen, während Datenlieferanten systematisch auf Fachmessen, über Unternehmensnetzwerke oder durch Direktansprache kontaktiert wurden. Ebenfalls wurden bei der Recherche regional hochauflösende Statistiken zu Wohngebäudeinformationen von Instituten, wie beispielsweise Zensus oder dem BBSR (Bundesamt für Bauwesen und Raumordnung), berücksichtigt. Insgesamt konnten durch diesen Prozess 37 potenziell geeignete Datenquellen bzw. -lieferanten ausgemacht werden, welche Datensätze über Wohngebäude besitzen. Unter anderem zählten dazu Energieversorger, Stadtwerke, Forschungsinstitute und Online-Vergleichsportale. Das strukturierte Vorgehen ermöglichte es bereits im Beschaffungsprozess der Datensätze anhand von Gesprächen und Eigenrecherche Informationen zu sammeln, um die jeweiligen Datenquellen in ihrer Reputation, Objektivität und Zugänglichkeit zu evaluieren. Bei 23 der 37 identifizierten Datenquellen und -lieferenten wurde die Reputation und Objektivität als unzureichend eingeschätzt.

Projektspezifisches Prüf- und Entscheidungsverfahren Anhand des generischen und hierarchischen Vorgehens (vgl. Abb. 3) wurde ein projektspezifischer Prüf- und Entscheidungsprozess (vgl. Abb. 5) abgeleitet. Dieser ermöglicht eine effiziente Überprüfung des inhaltlichen Fit und der Datenqualität der Datensätze. Dazu wurden die einzelnen Entscheidungsschritte und die Entscheidungslogik zur Zielerreichung projektspezifisch instanziiert. Wichtig ist, dass in der Praxis die inhaltliche Prüfung durch oder in Zusammenarbeit mit Domänenexperten durchgeführt wird, während für die Datenqualitätsprüfung Datenanalysten eingesetzt werden. In seltenen Fällen sind entsprechende Kenntnisse nicht intern vorhanden. Dies gilt es zu prüfen, so dass frühzeitig das Team durch Externe ergänzt werden kann. Die zur Prüfung notwendigen Anforderungen wurden in Gesprächen mit den jeweiligen Domänenexperten evaluiert und präzisiert. Anschließend wird auf Besonderheiten in den einzelnen Prüfschritten und auf zentrale Entscheidungen hinsichtlich der Verwendung der jeweiligen Datensätze, im Folgenden Datensatz A, B, C, D und E genannt, näher eingegangen. Die bildhaften Symbole in Abb. 5 stellen die für das Projekt relevanten Attribute (vgl. Abb. 4) und die möglichen Ziele des Projekts dar. Die grünen Haken und roten „X“ veranschaulichen den Prozessverlauf bei jeweils positiver oder negativer Eignungsprüfung der jeweiligen Bewertungsdimensionen (Kästen).

Durch die Überprüfung der Dimensionen Verständlichkeit, Relevanz und Vollständigkeit wurde in den ersten drei Prüfschritten über die inhaltliche Eignung

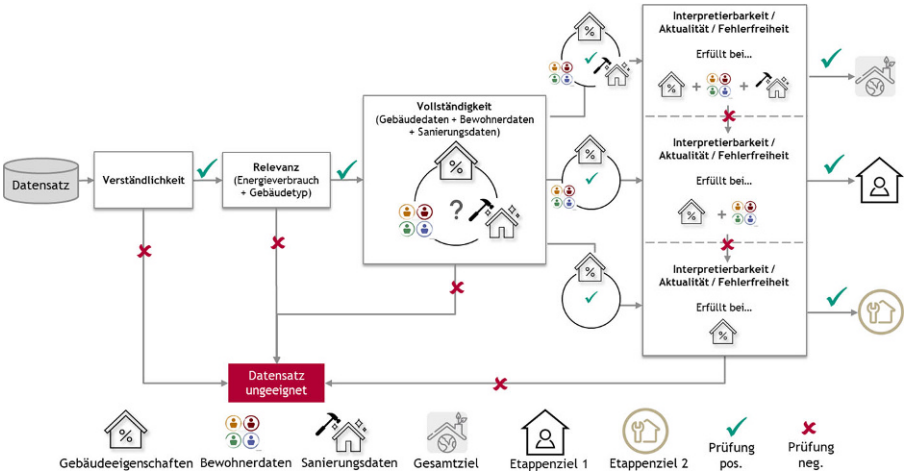

Abb. 5 Projektspezifischer Prüf- und Entscheidungsprozess

der Datensätze entschieden. Während Datensatz A intuitiv verständliche Attributbezeichnungen, wie zum Beispiel „Fassadenfläche“ oder „vermietbare Wohnfläche“, beinhaltete, bedurfte es bei den Attributausprägungen in den Datensätzen B, C und E einer initialen Dekodierung der kodierten Attributbezeichnungen anhand beigefügter Dekodierungstabellen. Datensatz D wurde aufgrund fehlender Informationen zur Dekodierung der Attributausprägungen verworfen. Ein Datensatz galt für das Projekt als relevant, wenn er Energieverbrauchsdaten von Gebäuden und gleichzeitig Angaben zum Gebäudetyp enthält. Die Datensätze C und E enthielten jeweils Energieverbrauchsdaten und Angaben zum Gebäudetyp. Beide Datensätze gaben außerdem an, ob es sich bei einem Wohngebäude um ein freistehendes oder angrenzendes Ein- bzw. Zweifamilienhäuser handelte. Den Datensätzen A und B fehlten Informationen zum Energieverbrauch. Sie waren damit für das Projekt nicht weiter relevant und wurden nicht weiter berücksichtigt. Zur Entscheidung, für welches Ziel die verbliebenen Datensätze C und E weiterhin geeignet sind, wurden beide Datensätze auf weitere Attribute (ZA) überprüft. Dabei konnte Datensatz C Gebäude-, Nutzer- und Sanierungsattribute aufweisen, während in Datensatz E Bewohnerdaten fehlten und nur Attribute zu Gebäudeparametern und Sanierungsdaten enthalten waren.

Auf Basis der inhaltlichen Bewertungsdimensionen wurde effizient identifiziert, dass die Datensätze A, B und D nicht zur Zielerreichung im Projekt beitragen und lediglich die Datensätze C und E das inhaltliche Potenzial dazu besitzen. Um deren finale Eignung festzustellen, wurden die Qualitätsdimensionen Interpretierbarkeit, Aktualität und Fehlerfreiheit auf Datenwertebene überprüft. In beiden Datensätzen C und E waren die Attribute zum Energieverbrauch und die physischen Gebäudeparameter in herkömmlichen Einheiten angegeben. Kodiert vorliegende Datenwerte konnten über vorhandene Dekodierungshinweise in beiden Datensätzen ebenfalls problemlos interpretiert werden. Da sich die Datenbasis der Wohngebäude über einen längeren Zeitraum erstreckt, Sanierungsmaßnahmen aber in der Regel zu einer Veränderung von Gebäudeparametern und Energieverbrauchswerten führen und

Bewohner aufgrund Eigentümer- oder Mieterwechsel über einen Zeitraum unterschiedlich sein können, war die Aktualität der Datenwerte im Projekt ein wichtiges Kriterium. Energieverbrauchswerte, Sanierungsdaten und Bewohnerdaten müssen aus dem gleichen Zeitraum stammen, um für das Prognosemodell geeignet zu sein. Datensatz C erfüllte diese Anforderungen nicht. Die enthaltenen Datenwerte stammen aus den Jahren 2006–2008, sodass in diesem Zeitraum die Bewohnerdaten im direkten Zusammenhang mit den angegebenen Energieverbrauchswerten standen. Die Datenwerte bilden in diesem Zeitraum jedoch keine Energieverbrauchswerte ab, die vor und nach einer Sanierungsmaßnahme aufgenommen wurden. Damit war dieser Datensatz nicht mehr geeignet Sanierungsmaßnahmen energetisch zu bewerten (Gesamtziel) eignete sich jedoch weiterhin zur Erreichung des Etappenziels. Für Datensatz E war die Aktualitätsdimension nicht relevant, da er inhaltlich ohnehin nur zur Energieverbrauchsprognose auf Basis von Gebäudedaten geeignet war und Bewohner- sowie Sanierungsdaten unberücksichtigt blieben. An die *Fehlerfreiheit* der Datenwerte legte das Projekt die Anforderung, dass mindestens 10.000 Datenwerte zu Gebäuden, zu deren Bewohner und zu Sanierungen enthalten sind. Die Schätzung beruhte auf Erfahrung von Domänenexperten und Datenanalysten gemeinsam. Je nach Umfang des Datensatzes ergeben sich so die Grade der Fehlerfreiheit – z. B. 40 % für Datensatz E.

Zusammengefasst war das Projektgesamtziel mit den identifizierten Datenquellen nicht erreichbar. Jedoch konnten jeweils ein Etappenziel nach Eignungsprüfung der Datensätze C und E erreicht werden (vgl. Abb. 6). Aus diesem Grund wurde Datensatz E im Rahmen des Projekts mit einer zusätzlichen Primärdatenerhebung versehen, um das Gesamtprojektziel zu verwirklichen.

Dass das Etappenziel 2 auf Basis des vorgestellten Vorgehens mit Datensatz E erreicht werden konnte, zeigten unteranderem Wenninger und Wiethe (2021). In ihrer Analyse verschiedener maschineller Lernverfahren zeigten sie, dass datengetriebene Prognoseverfahren gegenüber klassischen Verfahren zur Energiebilanzierung auf Basis physikalischer Gesetzmäßigkeiten über alle Gebäudealtersklassen exaktere Verbrauchsprognosen lieferten. Im Rahmen der Datennacherhebung zur Erreichung des Projektgesamtziels konnten Niemierko et al. (2019) feststellen, dass das Nutzerverhalten auf die Effektivität von energetischen Sanierungsmaßnahmen einwirkt. Die Ergebnisse unterstreichen, dass das präsentierte Vorgehen geeignet ist, Datensätze hinsichtlich der Eignung für spezifische Anwendungsfälle zu überprüfen.

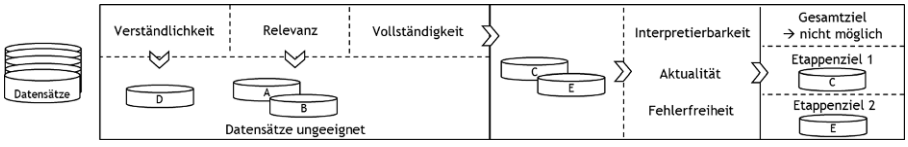

Abb. 6 Zusammenfassung der Eignungsprüfung im Beispielprojekt

5 Zusammenfassung

Das wachsende Angebot an unternehmensinternen und vor allem -externen Datensätzen (insb. Open Data Sets) verbessert die Grundlage evidenzbasierter Entscheidungen. Die Kernherausforderung in der Nutzbarmachung des Datenangebots liegt in der Datenvielfalt und der Notwendigkeit Datensätze systematisch auf ihre Eignung für eine spezifische Anwendung strukturiert und effektiv zu bewerten (Unsicherheit des Nutzens). Das vorgestellte systematische und hierarchische Vorgehen zur projektspezifischen Eignungsprüfung hilft Praktikern geeignete Datensätze für eine spezifische ML-Anwendung zu evaluieren und auszuwählen. Das Vorgehen zur Eignungsprüfung ergänzt dabei die in der Praxis verbreiteten Vorgehensmodelle für ML- bzw. Analytics-Projekte sinnvoll in der Datenbeschaffung. Es reduziert so u. a. die aufwändige Aufbereitung potenziell ungeeigneter Datensätze. Weiterhin besteht das Potenzial durch eine hierarchische und systematische Eignungsprüfung schnell und effizient Lücken in der Datenverfügbarkeit aufzudecken. Das vorgestellte Vorgehen gibt einen Rahmen, der Spielraum für projektspezifische Anpassungen gewährt und somit hinreichend allgemein ist. Zudem betont es die Involvierung von Domänenexperten in der frühen Phase der Eignungsprüfung, um den inhaltlichen Fit projektspezifisch zu bewerten. Die Anwendbarkeit des Vorgehens wurde im Rahmen eines mehrjährigen öffentlichen Forschungsprojekts validiert.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Agrawal P, Arya R, Bindal A, Bhatia S, Gagneja A, Godlewski J, Low Y, Muss T, Paliwal MM, Raman S, Shah V, Shen B, Sugden L, Zhao K, Wu M-C (2019) Data platform for machine learning, S 1803–1816. https://doi.org/10.1145/3299869.3314050

Batini C, Scannapieco M (Hrsg) (2016) Data and information quality. Springer, Cham

Batini C, Cabitza F, Cappiello C, Francalanci C, Di Milano P (2006) A comprehensive data quality methodology for web and structured data. In: 2006 1st International Conference on Digital Information Management. IEEE, S 448–456

Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Comput Surv 41:1–52. https://doi.org/10.1145/1541880.1541883

Bhavsar H, Ganatra A (2012) A comparative study of training algorithms for supervised machine learning. Int J Soft Comput Eng 2:2231–2307

Binzen M (2021) Open Data gewinnbringend einsetzen – Grundlagen und Hintergründe. HMD 58:359–376. https://doi.org/10.1365/s40702-021-00714-2
Cichy C, Rass S (2019) An overview of data quality frameworks. IEEE Access 7:24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751
Döbel I, Leis M, Vogelsang MM, Neustroev D, Petzka H, Riemer A, Rüping S, Voss A, Wegele M, Welz J (2018) Maschinelles Lernen; Eine Analyse zu Kompetenzen, Forschung und Anwendung. Fraunhofer-Gesellschaft. https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf. Zugegriffen: 01.10.2021
Domino Data Lab (2017) The practical guide to managing data science at scale; lessons from the field on managing data science projects and portfolios. https://www.dominodatalab.com/static/gfx/uploads/domino-managing-ds.pdf. Zugegriffen: 01.10.2021
Donnelly J, John A, Mirlach J, Osberghaus K, Rother S, Schmidt C, Voucko-Glockner H, Wenninger S (2021) Enabling the smart factory—A digital platform concept for standardized data integration https://doi.org/10.15488/11275
Dursun D (2019) Big Data für das Management. Control Manag Rev 63:46–52. https://doi.org/10.1007/s12176-018-0104-0
English LP (1999) Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Wiley,
Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183. https://doi.org/10.2307/2286061
Görz Q, Kaiser M (2012) An indicator function for insufficient data quality—A contribution to data accuracy. In: van der Aalst W, Mylopoulos J, Rosemann M, Shaw MJ, Szyperski C, Rahman H, Mesquita A, Ramos I, Pernici B (Hrsg) Knowledge and technologies in innovative information systems. Springer, Berlin, Heidelberg, S 169–184
Heinrich B, Klier M (2011) Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. In: Hildebrand K, Gebauer M, Hinrichs H, Mielke M (Hrsg) Daten- und Informationsqualität. Vieweg+Teubner, Wiesbaden, S 49–67
Heinrich B, Kaiser M, Mathias K (2007) How to measure data quality?—A metric based approach. In: Twenty Eighth International Conference on Information Systems Montreal, 2007
Helfert M (2002) Planung und Messung der Datenqualität in Data-Warehouse-Systemen. Universität St. Gallen
Hinrichs H (2002) Datenqualitätsmanagement in data warehouse-systemen. Universität Oldenburg
Hofmann P, Jöhnk J, Protschky D, Urbach N (2020) Developing purposeful AI use cases—A structured method and its application in project management. In: Gronau N, Heine M, Poustcchi K, Krasnova H (Hrsg) WI2020 Zentrale Tracks. GITO, S 33–49
International Organization for Standardization (2008) ISO/IEC 25012:2008; Software engineering—Software product Quality Requirements and Evaluation (SQuaRE)—Data quality model. https://www.iso.org/standard/35736.html. Zugegriffen: 16. Dez. 2020
Jarke M, Jeusfeld MA, Quix C, Vassiliadis P (1999) Architecture and quality in data warehouses: an extended repository approach. Inf Syst 24:229–253. https://doi.org/10.1016/S0306-4379(99)00017-4
Keim D, Sattler K-U (2021) Von Daten zu Künstlicher Intelligenz – Datenmanagement als Basis für erfolgreiche KI-Anwendungen. https://digitaleweltmagazin.de/2021/03/15/von-daten-zu-kuenstlicher-intelligenz-datenmanagement-als-basis-fuer-erfolgreiche-ki-anwendungen/. Zugegriffen: 22. Jan. 2022
Kessler R, Gómez JM (2020) Implikationen von Machine Learning auf das Datenmanagement in Unternehmen. HMD 57:89–105. https://doi.org/10.1365/s40702-020-00585-z
Krcmar H (2015) Informationsmanagement. Springer Gabler, Berlin, Heidelberg
Lee YW, Strong DM, Wang RY, Pipino LL (1998) Manage your information as a product. Sloan Manag Rev 39(4):95–105
Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inf Manag 40:133–146. https://doi.org/10.1016/S0378-7206(02)00043-5
Martinez-Plumed F, Contreras-Ochando L, Ferri C, Hernandez-Orallo J, Kull M, Lachiche N, Ramirez-Quintana MJ, Flach P (2021) CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Trans Knowl Data Eng 33:3048–3061. https://doi.org/10.1109/TKDE.2019.2962680
McGilvray D (2008) Executing data quality projects; ten steps to quality data and trusted information. Morgan Kaufmann/Elsevier, Burlington
Microsoft (2020) Was ist der Team Data Science-Prozess (TDSP)? https://docs.microsoft.com/de-de/azure/architecture/data-science-process/overview. Zugegriffen: 28. Juli 2021

Niemierko R, Töppel J, Tränkler T (2019) A D-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data. Appl Energy 233/234:691–708. https://doi.org/10.1016/j.apenergy.2018.10.025

Otto B (2012) Enterprise-wide data quality management in multinational corporations. University of St. Gallen,

Otto B, Wende K, Schmidt A, Osl P (2007) Towards a framework for corporate data quality management. In: Toleman M, Cater-Steel A, Roberts D (Hrsg) Proceedings of 18th Australasian Conference on Information Systems. The University of Southern Queensland, Toowoomba, S 916–926

Polyzotis N, Roy S, Whang SE, Zinkevich M (2018) Data lifecycle challenges in production machine learning. SIGMOD Rec 47:17–28. https://doi.org/10.1145/3299887.3299891

Ratner A, de Sa C, Wu S, Selsam D, Ré C (2017) Data programming: creating large training sets, quickly. Adv Neural Inf Process Syst 29:3567–3575

Reinsel D, Gantz J, Rydning J (2018) The digitization of the world; from edge to core. IDC. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Zugegriffen: 01.10.2021

Saltz JS, Grady NW (2017) The ambiguity of data science team roles and the need for a data science workforce framework. In: 2017 IEEE International Conference on Big Data (Big Data). IEEE, S 2355–2361

Sebastian-Coleman L (2012) Measuring data quality for ongoing improvement: a data quality assessment framework. Morgan Kaufmann/Elsevier, Burlington

Sundararaman A, Venkatesan SK (2017) Data quality improvement through OODA methodology. In: Proceedings of the 22nd MIT, S 1–14

Vaziri R, Mohsenzadeh M, Habibi J (2016) TBDQ: a pragmatic task-based method to data quality assessment and improvement. PLoS ONE 11:e154508. https://doi.org/10.1371/journal.pone.0154508

Wang RY (1998) A product perspective on total data quality management. Commun ACM 41:58–65. https://doi.org/10.1145/269012.269022

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12:5–33. https://doi.org/10.1080/07421222.1996.11518099

Weber K, Otto B, Österle H (2009) One size does not fit all—A contingency approach to data governance. J Data Inf Qual 1:1–27. https://doi.org/10.1145/1515693.1515696

Wenninger S, Wiethe C (2021) Benchmarking energy quantification methods to predict heating energy performance of residential buildings in Germany. Bus Inf Syst Eng 63:223–242. https://doi.org/10.1007/s12599-021-00691-2

Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, S 29–39