D. Demmler, D. Krupka, H. Federrath. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 1055

Teilautomatisierte Datenqualitätsbewertung und Fehlerkorrektur zur Senkung der Einstiegshürde von Datenanalysemethoden

Philipp Schlunder[1]

Abstract: Zur Nutzung aktueller Methoden der Datenanalyse müssen Daten häufig zunächst aufbereitet werden. Hierzu werden Kenntnisse benötigt, die in viele Organisationen noch nicht vorliegen. Um diese Einstiegshürde zu verringern, wird ein Konzept zur interaktiven, teilautomatisierten Behebung von Datenqualitätsproblemen vorgestellt, dessen Anwendung weniger Vorkenntnisse bedarf, indem eine automatische Bewertung der Datenqualität und des Informationsgehalts bereitgestellt wird, mit der Option, erkannte Probleme durch verschiedene Ansätze direkt korrigieren zu lassen.

Keywords: Datenqualität; Maschinelles Lernen; KMU; Einstiegshürde

1 Motivation und Stand der Technik

Deutsche KMU[2] stellen knapp 99 % der Unternehmen dar, doch nur 6 % von ihnen nutzen bereits Künstliche Intelligenz (KI) [BMWi]. Knapp über 50 % der Unternehmen nennen fehlende Expertise und knapp 28 % fehlende finanzielle Ressourcen als Grund für den mangelnden Einsatz [BVMW20]. In den meisten Fällen müssen Daten vor einer Analyse aufbereitet werden. Dazu gehört die Bewertung der vorliegenden Datenqualität, sowie des Informationsgehaltes und damit die Eignung für eine anschließende Analyse. Bestehende Methoden zur Qualitätsbewertung [Af20; Ei19; Gu21], die teils interaktiv und detailliert Probleme aufzeigen (Pandas-Profiling [PP], Tensorflow Data Validator [Ma15]), setzen Programmierkenntnisse voraus, sind eingeschränkt auf eine vordefinierte Korrekturmaßnahme pro Fehler (Great Expectations [Go]) oder erlauben Fehlerkorrekturen über ein GUI vorzunehmen, bieten aber keine Handlungsempfehlung an (RapidMiner TurboPrep [RMTP18]). Im Vergleich der bewerteten Werkzeuge wird deutlich, dass aktuell zur Bewertung und Behebung von Datenqualitätsproblemen, sowie zur Einschätzung des Informationsgehaltes eines Datensatzes entweder Programmierkenntnisse oder Kenntnisse über Datenanforderungen zum Einsatz von KI benötigt werden.

[1] daibe UG, pschlunder@daibe.io

[2] Klein und mittelständische Unternehmen mit weniger als 250 Mitarbeitenden

doi:10.18420/inf2022_87

2 Konzept

Zielgruppe des vorgestellten Konzepts sind Personen einer Organisation, die über Domänenwissen verfügen, aber über geringe bis keine Programmier- und Datenanalysekenntnisse. Sie besitzen lediglich ein erstes Grundverständnis vom Zusammenhang zwischen Daten und abgebildeten Informationen (Data Literacy). Das Konzept soll diese Personen befähigen, eigenständig ein erstes Verständnis über die Qualität und Eignung von Daten für weitere Analysezwecke zu erlangen, damit sie eigenständig Fehler im betrachteten Datensatz informiert beheben können. Um dieses Bewusstsein herzustellen, werden Informationen zur erhobenen Datenqualität in verschiedenen Granularitätsstufen bereitgestellt, und unterschiedliche Behebungsmechanismen angeboten, die zusammen mit einer Erklärung, Nutzenden erlauben Korrekturen an geladenen Datensätzen vorzunehmen.

Das Konzept sieht eine Web-App vor, die Personen vom Einlesen ihrer Daten über eine Bewertung und optionalen Fehlerkorrektur der Datenqualität bis hin zu einer Anschlussverwertung begleitet. Im Kern besteht der Prozessfluss des Konzepts aus folgenden Schritten:

1. Exemplarische Daten per Drag & Drop anbinden.
2. Darstellung eines Datenauszugs mit automatischer Erkennung vorliegender Datentypen und Identifikation einer als ID geeigneten Spalte.
3. Auswahl der zu analysierenden Teilmenge der Daten.
4. Bereitstellung einer Bewertung der Datenqualität, sowie des Informationsgehalts in allgemeiner und detaillierter Form.
5. Möglichkeit zur (teil-)automatischen Korrektur von Datenqualitätsproblemen.
6. Bereitstellung reparierter und entfernter Daten, sowie Qualitäts- und Änderungsberichte. Option, durchgeführte Korrekturlogik auf neuen Daten anzuwenden, indem je nach Wunsch eine Web-App, ein REST-API oder eine CLI bereitgestellt wird.

Die Bereitstellung als Web-App mit der Möglichkeit, Daten in gängigen Formaten via Drag & Drop zu laden, ermöglicht es auch nicht IT-affinen Personen, das Werkzeug einfach bedienen zu können. Bei jedem der Schritte ist Nutzenden freigestellt, ob sie Automatisierungen verwenden oder manuell eingreifen, indem Sie beispielsweise selbst eine ID-Spalte auswählen; entscheiden, ob und wie ein Datenqualitätsproblem behoben wird; oder der Überprüfungskontext angepasst wird, um regionsspezifische Einstellungen und durchzuführende Qualitätsüberprüfungen zu verändern.

Insbesondere das Laden heterogener, teilstrukturierter Dateien, sowie die Auswahl von Fehlerkorrekturen wird durch verschiedene Automatisierungen und Einstellungsmöglichkeiten unterstützt. Um zu gewährleisten, dass reale Daten eingebunden werden können, wird beispielsweise das Encoding einer Datei erkannt und bei XLSX-Dateien eine Bereichserkennung auf Datenblättern durchgeführt. Bei CSV-Dateien wird der eigentliche Zeilenbereich

herausgefiltert, welcher oft durch Meta-Informationen zu Beginn oder Ende der Dateien umgeben ist.

Datenqualitätsprobleme werden zusammenfassend beschrieben aufgeführt, mit der Option, detaillierte statistische Beschreibungen, sowie Visualisierungen der Problematik anzeigen zu lassen; einen Fehler zu ignorieren; eine automatische Korrektur zu verwenden, oder diese manuell anzupassen. Dazu werden verschiedene Korrekturmechanismen angeboten, wobei der empfohlene ausgewählt ist und gängige Einstellungen anpassbar sind. Jedoch sind die Parameterbeschreibungen in eine anwendergerechte Sprache überführt und mit dem Datensatz entsprechenden Parametern vorbefüllt. Je nach Fehlertyp wird eine Vorauswahl getroffen, ob ein Fehler behoben werden soll oder nicht. Somit ist es Nutzenden möglich, nach der Überprüfung eine Korrektur automatisch mit vorgeschlagenen Mechanismen vorzunehmen, oder diese anzupassen. Bereitgestellte Informationen umfassen dabei immer auch Auflistungen problematischer Einträge, während Korrekturempfehlungen umgangssprachliche Beschreibungen des Mechanismus bereitstellen.

Durch die Befähigung zur Qualitäts- und Informationsgehaltsbewertung können Nutzende so bestehende Datensätze anpassen und vergleichen, um eine Vorauswahl an Daten zu erhalten, die für ein etwaiges Analyseprojekt verwendet werden können. Diese Vorarbeiten können dabei helfen, anfängliche Aufwände zur Datensammlung und -Vorverarbeitung durch Expert*innen zu reduzieren, und somit Kosten zusenken. Zudem lernen Nutzende, welche potenziellen Fehler in Daten auftreten können und was für Korrekturmechanismen existieren, um so gegebenenfalls interne Datenerhebungsprozesse proaktiv anzupassen.

3 Technische Umsetzung

Es existiert bereits eine erste Umsetzung des Konzeptes in Python, wobei `streamlit`[3] zur Erstellung des Front-Ends eingesetzt wird, und das Back-End auf einer neu entwickelten Bibliothek zur Überprüfung und Korrektur von Datenqualitätsproblemen aufbaut. Im Folgenden wird der Aufbau der Bibliothek beschrieben. Zentrale Konzepte sind:

- **CheckContext**: Verwaltet Meta-Informationen zur Konfiguration einer Qualitätsprüfung. Dies beinhaltet eine Liste an durchzuführenden *QualityChecks*, samt Voreinstellungen, sowie regionale Einstellungen, wie die typische Darstellung von Zeitstempeln, der Zeitzone, Währungseinstellungen und der Zielsprache.
- **DataQualityHandler**: Lädt einen *CheckContext*, wendet die definierten *QualityChecks* auf bereitgestellte Daten an und verwaltet erkannte Probleme und ausgewählte Korrekturen. Vor der Anwendung der Checks auf Daten werden diese bedingt durch regionale Einstellungen vorkonfiguriert.
- **QualityChecks**: Mechanismen zur Überprüfung einzelner Datenqualitätsmetriken, die vordefinierte *QualityFixes* enthalten.

[3] Python-Bibliothek zur Erstellung von Web-Apps aus Analyse-Skripten, `https://streamlit.io/`

- **Warnings**: Bei Anwendung eines Checks auf Daten werden Qualitätsmetriken erhoben, *QualityFixes* angepasst und ein empfohlener Fix vorausgewählt. Dabei werden regionale Einstellungen aus der Vorkonfiguration durch den *DataQualityHandler* berücksichtigt. Diese datenbezogenen Ergebnisse eines *QualityChecks* werden als *Warning* abgespeichert.
- **QualityFixes**: Sind Korrekturmaßnahmen zur Behebung von Datenqualitätsproblemen, wobei ein Fix für mehrere Checks als Maßnahme benutzt werden kann. *QualityFixes* enthalten allgemeinverständliche Beschreibungen ihrer Methodik und Parametrisierung.
- **Reports**: *DataQualityHandler* sammeln generierte *Warnings* in einem *Report*, welcher alle durchgeführten *QualityChecks* und angewandten *QualityFixes* verwaltet. Diese können als textueller Bericht und Konfigurationsdatei ausgegeben werden.

Es existieren bereits Datenqualitätschecks und Korrekturmechanismen zur Handhabung von Problemen der Vollständigkeit, Konsistenz kategorischer Ausprägungen, unerwarteter Sprungstellen und doppelter Einträge. Jeder Datenqualitätscheck berechnet verschiedene Kennzahlen zur Bewertung, ob ein Problem vorliegt, und welche Methode zur Korrektur vorgeschlagen wird. Eine Kombination dieser Metriken wird zudem verwendet, um eine einzelne, prozentuale Kennzahl als erstes Maß für die Gesamtqualität anzubieten. Der Informationsgehalt wird aktuell über die Entropie, die Korrelation der Spalten untereinander, sowie der Güte der (erkannten) ID-Spalte bestimmt. Dabei wird angenommen, dass eine höhere Entropie und eine geringe Korrelation zwischen Spalten im Bezug auf ihre Anzahl ein höheres Potenzial zum Finden von Strukturen in den Daten bietet, sowie berücksichtigt, dass in den meisten Analyse-Szenarien eine konsistente ID benötigt wird.

4 Zusammenfassung und Ausblick

Zur Reduktion der Einstiegshürde in die Nutzung von Datenanalyse und KI wurde ein Konzept zur Überprüfung und Korrektur der Qualität von Datensätzen vorgestellt, welches es Personen ohne Programmier- und Datenanalyse-Kenntnissen ermöglichen soll, vorhandene Daten zu beurteilen und Qualitätsprobleme zu beheben. Dazu wurde das Konzept als prototypische Web-App basierend auf einer neuen Bibliothek zur Datenqualitätshandhabung umgesetzt, um teilautomatisierte Korrekturmechanismen bereitstellen zu können. Nutzende erhalten so die Möglichkeit, eigenständig und informiert Fehler zu erkennen, zu beheben und korrigierte Daten, sowie neu gewonnene Erkenntnisse als Start für Analysen oder Prozessanpassungen in ihrer Organisation zu verwenden.

In Zukunft wird die Bibliothek als open-source Lösung bereitgestellt und die Integration bestehender Datenvalidierungsframeworks geprüft.

Literatur

[Af20] Afzal, S.; C, R.; Kesarwani, M.; Mehta, S.; Patel, H.: Data Readiness Report./, Okt. 2020, URL: http://arxiv.org/abs/2010.07213.

[BMWi] Bundesministerium für Wirtschaft und Energie: Einsatz von Künstlicher Intelligenz in der Deutschen Wirtschaft, URL: https://www.bmwk.de/Redaktion/DE/Publikationen/Wirtschaft/einsatz-von-ki-deutsche-wirtschaft.pdf?__blob=publicationFile&v=8, Stand: 25. 07. 2022.

[BVMW20] Der Mittelstand, BVMW e. V.: Digitalisierung im Mittelstand: Aufschwung, aber kein Durchbruch, 2020, URL: https://www.bvmw.de/themen/digitalisierung/news/12285/digitalisierung-im-mittelstand-aufschwung-aber-kein-durchbruch/, Stand: 25. 07. 2022.

[Ei19] Eickelmann, M.; Wiegand, M.; Deuse, J.; Bernerstätter, R.: Bewertungsmodell zur Analyse der Datenreife. Zeitschrift für wirtschaftlichen Fabrikbetrieb 114/, S. 29–33, Feb. 2019, ISSN: 2511-0896, URL: https://www.degruyter.com/document/doi/10.3139/104.112037/html.

[Go] Gong, A.; Campbell, J.; Superconductive; Great Expectations: Great Expectations, URL: https://github.com/great-expectations/great_expectations.

[Gu21] Gupta, N.; Patel, H.; Afzal, S.; Panwar, N.; Mittal, R. S.; Guttula, S.; Jain, A.; Nagalapatti, L.; Mehta, S.; Hans, S.; Lohia, P.; Aggarwal, A.; Saha, D.: Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets./, Aug. 2021, URL: http://arxiv.org/abs/2108.05935.

[Ma15] Martín Abadi; Ashish Agarwal; Paul Barham; Eugene Brevdo; Zhifeng Chen; Craig Citro; Greg S. Corrado; Andy Davis; Jeffrey Dean; Matthieu Devin; Sanjay Ghemawat; Ian Goodfellow; Andrew Harp; Geoffrey Irving; Michael Isard; Jia, Y.; Rafal Jozefowicz; Lukasz Kaiser; Manjunath Kudlur; Josh Levenberg; Dandelion Mané; Rajat Monga; Sherry Moore; Derek Murray; Chris Olah; Mike Schuster; Jonathon Shlens; Benoit Steiner; Ilya Sutskever; Kunal Talwar; Paul Tucker; Vincent Vanhoucke; Vijay Vasudevan; Fernanda Viégas; Oriol Vinyals; Pete Warden; Martin Wattenberg; Martin Wicke; Yuan Yu; Xiaoqiang Zheng: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Software available from tensorflow.org, 2015, URL: https://www.tensorflow.org/.

[PP] Pandas Profiling, URL: https://github.com/ydataai/pandas-profiling, Stand: 25. 07. 2022.

[RMTP18] RapidMiner TurboPrep, 2018, URL: https://docs.rapidminer.com/9.3/studio/turbo-prep/, Stand: 24. 07. 2022.