KI IM STAMMDATENMANAGEMENT

VIELE ANWENDUNGSFELDER, GUTE DATEN UNERLÄSSLICH

KÜNSTLICHE INTELLIGENZ HAT AUCH IM STAMMDATEN-MANAGEMENT EINZUG GEHALTEN. IHR WEITERER ERFOLG WIRD DAVON ABHÄNGEN, OB ES GELINGT, DIE DATENQUALITÄT ZU SICHERN UND DIFFERENZIERTE, AUSSAGEKRÄFTIGE DATEN ZU GENERIEREN.

Monika Pürsing, Geschäftsführerin, zetVisions GmbH, www.zetVisions.de

Für die Suche nach „artificial intelligence" liefert die Suchmaschine Google - selbst ein Paradebeispiel für die Anwendung künstlicher Intelligenz - über 150 Millionen Ergebnisse in 0,56 Sekunden. Für „machine learning", ein Teilgebiet der künstlichen Intelligenz (KI), liefert der Algorithmus 116 Millionen Ergebnisse. Beeindruckende Zahlen, die zeigen, wie sehr künstliche Intelligenz und maschinelles Lernen diskutiert wurden und werden. Für KI gibt es eine ganze Reihe von Anwendungsgebieten, wie etwa Expertensysteme (Beispiel Watson), Gesichts- und Spracherkennung, Predictive Analytics und Robotik. KI lässt sich natürlich auch für das Management von Stammdaten wirkungsvoll einsetzen.

Stammdaten sind die Voraussetzung, um Daten überhaupt nutzen zu können. Ohne Stammdaten fehlt es an Definitionen und Kontext, ohne Stammdaten fehlt die Möglichkeit, Daten zu verstehen, verknüpfen zu können, zu interpretieren und richtig zu verwenden. Der Haken ist – wie so oft – die Qualität der Daten. Man kann das auf eine ganz einfache Formel bringen: Gute Daten verbessern die künstliche Intelligenz. Das Gegenteil trifft leider auch zu. „Schlechte Datenqualität ist Feind Nummer eins für den weit verbreiteten, profitablen Einsatz des maschinellen Lernens", schrieb Thomas C. Redman, der „Data-Doc", vor zwei Jahren im Harvard Business Review. Während die bissige Beobachtung ‚garbage-in, garbage-out' die Analytik und Entscheidungsfindung seit Generationen geplagt habe, enthalte sie für das maschinelle Lernen eine besondere Warnung. Die Qualitätsanforderungen an das maschinelle Lernen seien hoch, und schlechte Daten könnten ihm zweimal den Kopf verdrehen – erstens die historischen Daten, die zum Training des Vorhersagemodells verwendet werden, und zweitens die neuen Daten, die von diesem Modell für zukünftige Entscheidungen verwendet werden.

Daten als immaterieller Unternehmenswert

Es ist das alte Lied von der Datenqualität. „The Machine Learning Race Is Really a Data Race", lautete Ende 2018 die Überschrift eines Beitrags im Sloan Management Review. Daten werden zu einem Unterscheidungsmerkmal, weil viele Unternehmen nicht über die benötigten Daten verfügen. Die wertvollen, nützlichen Daten, die sie in die Lage versetzen, beispielsweise im Finanzbereich nicht nur materielle Vermögenswerte, sondern vor allem immaterielle Vermögenswerte zu messen. Dass Daten zu diesen immateriellen Unternehmenswerten gehören, diese Sichtweise ist noch nicht sehr weit verbreitet. Christine Legner und Martin Fadler vom Competence Center Corporate Data Quality in St. Gallen bemängeln: „Trotz der zunehmenden Relevanz von Daten im Kontext der Digitalisierung wird bisher in nur wenigen Unternehmen dem Management der Daten die gleiche Aufmerksamkeit zuteil, wie anderen Unternehmenswerten." In ihrer Studie „Managing Data as an Asset with the Help of Artificial Intelligence" (2019) kommen Legner und Fadler zu der Einsicht, in traditionellen Unternehmen seien Daten eine wichtige, aber vor allem unterstützende Ressource in Geschäfts- und Entscheidungsprozessen; in einer zunehmend digitalisierten Welt würden sie zu einem Wert an sich, weil sie die unabdingbare Voraussetzung für digitale Geschäftsmodelle und Strategien seien.

Die gute Nachricht sei, so Legner und Fadler, dass durch substanzielle Fortschritte künstliche Intelligenz und maschinelles Lernen – was das Lernen aus Daten und die Automatisierung sich wiederholender Aufgaben betreffe – Unternehmen bei ihren Datenmanagement-Aktivitäten unterstützen könnten. Ihre Studie zeige, dass maschinelles Lernen in allen Phasen des Datenlebenszyklus angewendet werden könne, um Folgendes zu erreichen:

- Datenbestände auf effiziente, benutzerfreundliche Weise zu erstellen und anzureichern;
- Aufrechterhaltung qualitativ hochwertiger Daten durch Unterstützung aktiver und reaktiver Datenpflege sowie zur Datenvereinheitlichung;

- Management des Datenlebenszyklus, insbesondere bei sensiblen Daten und bei der Ausmusterung von Daten;

- Steigerung der Nutzung von Daten durch Verbesserung der Datenentdeckung durch Nutzer, insbesondere durch Data Scientists.

Datenlebenszyklusphasen

Für jede dieser Datenlebenszyklusphasen haben Legner und Fadler Anwendungsszenarien für maschinelles Lernen identifiziert.

Die Phase der Datenerstellung und -erfassung komme es zu Schreibfehlern, falschen oder ungültigen Dateneinträgen, leeren Feldern und manuellem Aufwand. Hier unterstütze maschinelles Lernen die Datenerstellung, zum Beispiel durch automatisches Ausfüllen von Werten in Formularen und automatisches Extrahieren von Daten, sowie die Datenanreicherung.

Die Problemfelder in der Phase der Datenvereinheitlichung und -pflege lägen etwa in der Datenintegration über mehrere Systeme hinweg (was zu Inkonsistenzen führe), in der Korrektur von Datenfehlern und in der Definition von Geschäftsregeln. Maschinelles Lernen unterstütze zum einen die Datenpflege aktiv durch Geschäftsregeln und reaktiv durch Datenkorrektur, zum anderen die Datenvereinheitlichung durch Abgleich und Eliminierung von Datendubletten.

In der dritten Phase stehen der Datenschutz und die Ausmusterung von Daten im Zentrum. Als problematisch erweise sich dabei die mangelnde Transparenz, wo Informationen gespeichert werden, die sich auf eine identifizierbare Person beziehen (personally identifiable information, PII), und damit verbunden die Einhaltung von Datenschutzbestimmungen. Künstliche Intelligenz und maschinelles Lernen unterstützten den Datenschutz – beispielweise durch die Identifizierung sensibler Daten und die Aufdeckung betrügerischen Verhaltens – und das „Data Retirement", wenn Daten ihr „Lebensende" erreicht haben.

Die Phase der Datenentdeckung und -nutzung sei gekennzeichnet durch Probleme beim Auffinden und bei der Bereinigung relevanter Daten sowie bei der Identifizierung von Datenbeziehungen. Hier könnten künstliche Intelligenz und maschinelles Lernen die Datenermittlung beispielsweise durch Empfehlungen und die Verknüpfung von Datensätzen unterstützen.

Fazit

Legner und Fadler kommen unter dem Strich zu dem Fazit, maschinelles Lernen habe das Potenzial, die Datenmanagementpraktiken erheblich zu verbessern und die Datenqualität zu steigern. Ein gutes Beispiel dafür liefere Bosch. Dort sei es gelungen, den aufwändigen Prozess der manuellen Zuweisung von Zolltarifnummern zu einem Produkt – im Außenhandel muss jedes Unternehmen seine Produkte als Voraussetzung für Export-/Importprozesse entsprechend klassifizieren – durch eine Lösung zu ersetzen, die mit Hilfe überwachter Machine Learning-Algorithmen eine automatisierte Zuweisung von Warencodes mit hoher Genauigkeit (90 Prozent) ermöglicht.

Monika Pürsing