Omikron

Künstliche Intelligenz im Unternehmen

Saubere Daten, starke KI: Ihr Praxisguide zur Datenvorbereitung

Für Entscheider, KI-Berater & Projektmanager

Richard Leibrandt

Vorwort

In den Gesprächen mit Vorständen spüren wir den immensen Druck, KI-Initiativen schnell zu implementieren. Gebremst werden die Initiativen von internen Datensilos, einer veralteten IT-Infrastruktur und einem berechtigten Misstrauen gegenüber den eigenen Daten. Wer Geschwindigkeit in KI-Projekten will, braucht gute Daten. Denn qualitativ hochwertige Daten sind der Raketentreibstoff für KI-Modelle. Darum geht es in diesem Handbuch.

Keiner der genannten Effekte tritt alleine auf. Vielmehr vermischen und verketten sich die Fehler in den Daten, was zu der generischen Aussage führt: mit unseren Daten geht das nicht. Es geht, aber es braucht die richtigen Werkzeuge und Expertise.

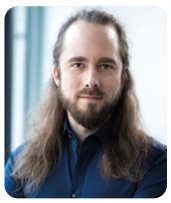
Unser Ziel mit diesem Handbuch ist es, Ihnen unsere Datenwelt näher zu bringen. Eine Welt, in der wir Woche für Woche Unternehmen dabei helfen, zur daten-getriebenen Organisation zu werden.

Ich wünsche Ihnen viel Spaß beim Lesen.

Albert Pusch
Geschäftsführer
Omikron

Inhalt

Über den Autor



Richard Leibrandt, Data Scientist

Richard Leibrandt entdeckte bereits während seines Studiums seine Faszination für Data Science und vertiefte sich anschließend gezielt in das Feld der Künstlichen Intelligenz. Über die letzten zehn Jahre hat er durch seine Expertise praxisorientierte Projekte unterstützt. Dabei hat er sein Wissen und seine Erfahrung nicht nur als Spezialist genutzt, sondern auch als Trainer, Berater und Redner weitergegeben. Sein Engagement erstreckt sich zudem auf die universitäre Forschung und parallel dazu in der Privatwirtschaft, in der er innovative Ansätze im maschinellen Lernen vorantreibt.


Wenn wir das nächste Mal eine KI mit Daten füttern, sollten wir vorher vielleicht mehr Wert auf die Datenqualität legen!
Wuuusch
CLOUD-SCIENCE.DE

Ist Ihre Organisation bereit für KI?

1

1. Effektive KI-Modelle nur mit guten Daten

In der heutigen, digitalisierten Welt steht Künstliche Intelligenz (KI) im Zentrum zahlreicher Innovationen. KI-Modelle sind die Motoren, die diese Innovationen antreiben. Doch kein Motor läuft ohne Schmieröl und Treibstoff. Was "Schmieröl" und "Treibstoff" für Motoren sind, sind Daten für KI-Modelle.

Kein Wunder also, dass The Economist schrieb "The world's most valuable resource is no longer oil, but data"[1]. So wie Treibstoff den Motor antreibt, so treiben Daten die Algorithmen und Modelle an, die die Grundlage für alle intelligenten Systeme bilden. Doch es ist nicht die schiere Menge an Daten, die zählt. Es ist die Qualität und Vielfalt dieser Daten, die maßgeblich beeinflussen, ob Modelle effektiv und zuverlässig trainiert werden können.[2]

Wie bei jedem Rohstoff ist es die Qualität, die den Unterschied macht. Hochwertige Daten führen zu effektiveren Modellen, die genauere Vorhersagen und Analysen ermöglichen. Ohne Daten sind die leistungsstärksten Methoden nichts als konzeptionelle, leere Hüllen. **Ohne Datenqualität sind die trainierten Modelle nichts als wilde Zufallszahlengeneratoren.** Die Datenqualität entscheidet also darüber, ob KI Ihre Organisation in den nächsten Monaten und Jahren voranbringen wird oder ob Wettbewerber an Ihnen vorbeiziehen.

1.1. Daten, Ihr Wachstums-Treibstoff der Zukunft

Datenqualität zu erreichen ist jedoch kein triviales Unterfangen. Damit die Daten eine zuverlässige Grundlage für KI-Modelle bilden, müssen sie zudem sorgfältig analysiert werden. Dies ist der unschöne Arbeitsbereich vieler Datenwissenschaftler.

50-80%

Data Scientists verbringen etwa 50 bis 80 Prozent ihrer Zeit mit dem Aufbereiten der Daten.

In einer Welt, in der Daten das 'Öl' der KI sind, ist Qualität die Währung, die den Unterschied ausmacht. In einer Welt, in der Daten der Treibstoff sind, ist die Qualität der Daten die Energiedichte[3], die unsere lernenden Maschinen antreibt. Je höher die Energiedichte, umso mehr Nutzen kann man aus einer Treibstoffmenge ziehen. Je höher die Datenqualität, umso mehr Nutzen kann man aus einer Datenmenge ziehen. **Daher benötigt es weniger Daten, um ein Modell gut zu trainieren, wenn die Datenqualität höher ist.** Qualitativ hochwertige Daten schaffen die besseren Modelle.

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[2] "It is not only the size of datasets that counts. The better the data, the better the model." laut https://www.economist.com/business/2023/08/13/ai-is-setting-off-a-great-scramble-for-data

[3] Energiedichte ist die Menge an Energie, die pro Volumeneinheit des Kraftstoffs freigesetzt wird. Ein Kraftstoff mit höherer Energiedichte kann mehr Leistung erzeugen.

1.2. Die Schmerzpunkte aller KI-Berater lindern

Ich bin selbst Datenwissenschaftler, sowohl in der Industrie als auch in der akademischen Forschung und war jahrelang als beratender Data Scientist tätig, sowohl im Big Data Umfeld als auch in der Verarbeitung kleiner Datenmengen. Die Welt von uns Datenwissenschaftlern ist oft abstrakt und komplex, aber sie wird noch komplexer, wenn wir uns mit schlechter Datenqualität auseinandersetzen müssen.

Die Folgen schlechter Datenqualität: Unsere Vorhersagen und Ergebnisse werden ungenau, unsere wertvolle Zeit wird in Datenbereinigung investiert und die Modellvalidierung wird zu einer undurchsichtigen Herausforderung, genauer:

- Ungenauigkeiten in Vorhersagen und Ergebnissen führen zu falschen Schlussfolgerungen.
- Überanpassung von Modellen aufgrund unrepräsentativer oder zu homogener Daten bedroht die Universalität und Anpassungsfähigkeit unserer Lösungen.
- Datenschutzbedenken schwingen im Hintergrund mit, wenn personenbezogene Daten fehlerhaft oder unsachgemäß verarbeitet werden.
- Unvollständige Datenmengen, obwohl nicht falsch, können uns wichtige Informationen vorenthalten, unser Verständnis einschränken, und uns dazu veranlassen, wichtige Aspekte zu übersehen, die für eine effektive KI-Modellierung entscheidend sind.
- Und die Validierung unserer Modelle wird zu einem Ratespiel, wenn die Qualität der Daten zweifelhaft ist.

Datenbereinigung ist eine undankbare Aufgabe. Das Ergebnis ist, dass ein beträchtlicher Teil unserer Zeit und Ressourcen für die Reinigung dieser Daten aufgewendet wird. Und ganz ehrlich gesagt: eigentlich wollen sich Data Scientists mit etwas anderem beschäftigen als Datensäuberungs-Pipelines aufzubauen.

Folgen einer schlechten Datenqualität

Fehlerhafte Daten führen zu falschen Ergebnissen und Schlussfolgerungen

Daten-Homogenität begrenzt KI-Flexibilität (Modellüberanpassung)

Datenschutzprobleme bei fehlerhaft verarbeiteten Personendaten

Unvollständige Datensätze verschleiern wichtige Aspekte

Unsichere Modellvalidierung anhand zweifelhafter Datenqualität

2

2. Wie lernt künstliche Intelligenz

Künstliche Intelligenz (KI) und insbesondere maschinelles Lernen (ML) lernen durch einen Prozess, der in zwei Hauptphasen unterteilt werden kann: Training und Inferenz.

1

Training: In dieser Phase lernt die KI ein Modell der Realität zu erstellen. Hierfür wird eine vordefinierte Struktur (Modell-Struktur) benötigt, die von Menschen entwickelt wurde. Diese Struktur ist wie ein leerer Rahmen, der darauf wartet, mit Wissen gefüllt zu werden. Um diesen Rahmen zu füllen, werden Daten benötigt, die Aspekte der Realität repräsentieren, die die KI verstehen soll. Diese Daten könnten alles Mögliche sein: Bilder, Texte, Zahlen, je nachdem, was das Modell lernen soll. Während des Trainingsprozesses passt die KI die Parameter ihres Modells an, um die in den Daten enthaltenen Muster und Beziehungen abzubilden. Das Ziel ist es, ein Modell zu entwickeln, das die Realität so genau wie möglich abbildet.

2

Inferenz: Nachdem das Modell trainiert wurde, kann es verwendet werden, um Schlussfolgerungen zu ziehen oder Vorhersagen zu machen. In dieser Phase erhält die KI neue Daten, die sie noch nicht gesehen hat, und wendet das gelernte Modell an, um Informationen aus diesen Daten abzuleiten oder Vorhersagen darüber zu treffen, was als nächstes passieren könnte. Dieser Schritt ist entscheidend, da er zeigt, ob das Modell gut trainiert wurde und ob es in der Lage ist, das Gelernte auf neue, unbekannte Situationen anzuwenden.

Zusammenfassend lernt KI durch das Erkennen von Mustern in Daten während des Trainings und die Anwendung dieser Erkenntnisse auf neue Daten während der Inferenz. Es ist ein iterativer Prozess, der oft viele Durchläufe durch Training und Inferenz erfordert, um ein präzises und zuverlässiges Modell zu entwickeln.

3

3. Die vier Arten von Datenproblemen

Datenqualitätsmängel können vielfältig sein und so Daten für das Training einer KI unbrauchbar machen. Hier sind einige häufige Datenqualitätsprobleme, die in Datensätzen auftreten können:

All die im folgenden genannte Datenqualitätsmängel führen zu fehlerhaften oder ungenauen Analysen und Modellen. Diese Mängel beeinflussen Algorithmen negativ und bergen das Potenzial für falsche Interpretationen. Wenn Datenprobleme identifiziert und behoben werden, verbessert sich die Datenqualität, was zu präziseren Analysen und verlässlicheren Modellvorhersagen führt.

Die 4 Arten von Datenqualitätsproblemen

Fehlerhafte Daten

- Inkonsistente Daten
- Ungenaue Daten
- Beliebige Daten
- Veraltete Daten

Fehlende Daten

- Lückenhafte Daten
- Fehlende Merkmale
- Fehlende gelabelte Daten
- Isolierte Datenmengen

Übermäßige Daten

- Irrelevante Daten
- Ausreißer
- Duplikate
- Daten, die nicht den Daten-schutzbestimmungen entsprechen

Unrepräsentative Daten

- Schiefe Daten, Mangel an Heterogenität
- Andere Arten unrepräsentativer Daten

Technologie und Expertise können viele dieser Probleme beheben, z.B. der Data Quality Server.

3.1. Fehlerhafte Daten

3.1.1. Inkonsistente Daten

Jede Organisation hat eine Vielzahl von IT-Systemen. Wenn Daten aus verschiedenen Quellen (z.B. IT-Systemen, Datenbanken) stammen oder von verschiedenen Menschen gesammelt wurden, können Inkonsistenzen auftreten. Dies kann unterschiedliche Schreibweisen, Formate oder Maßeinheiten umfassen.

Entstehung: Inkonsistenzen in den Daten können durch die Nutzung unterschiedlicher Datenquellen, menschliche Fehler, systembedingte Unterschiede oder die Verwendung unterschiedlicher Skalen und Einheiten entstehen.

Fallbeispiel

In einem Unternehmen werden Kundeninformationen in verschiedenen Systemen gespeichert: Salesforce.com CRM für das Kundenbeziehungsmanagement und Shopware als Online Shop – jedes System hat seine ganz eigenen Formate und Standards, die sich nicht einfach zusammenlegen lassen.

Beispielsituationen, die zu Inkonsistenzen führen:

- **Unterschiedliche Datenquellen:** Salesforce.com, SAP und Shopware haben jeweils eigene Datenformate und Standards.
- **Menschliche Fehler:** Manuelle Eingaben führen oft zu Variationen in der Schreibweise von Kundennamen, Adressen oder anderen Details.
- **Systembedingte Unterschiede:** Jedes System hat eigene Felder und Formate. Beispielsweise könnte Salesforce.com die Telefonnummern im internationalen Format speichern, während Shopware sie ohne Ländervorwahl speichert.
- **Unterschiedliche Skalen und Einheiten:** Kundensegmente könnten in Salesforce.com als “klein”, “mittel”, “groß” kategorisiert sein, während in Shopware andere Kriterien wie Umsatzgrößenklassen verwendet werden.

Nachteile: Inkonsistente Daten machen es schwierig, Daten zu kombinieren und zu analysieren. Sie führen zu Problemen bei der Datenintegration und -analyse.

Vorteile: Konsistente Datenmengen können auf zweierlei Weise einfach miteinander kombiniert werden:

1. Datenpunkte aus mehreren Datenmengen können zu einer Datenmenge vereinigt werden. Im Ergebnis haben wir mehr Trainingsdaten, die zu robusteren Modellen führen können. Daten aus unterschiedlichen Quellen können etwa so kombiniert werden, dass die Datenqualität an anderer Stelle verbessert wird: z.B. kann die Heterogenität erhöht werden und damit der Grad, zu dem die Trainingsdaten die echte Population repräsentieren.

2. Datenmengen von unterschiedlichen Attributen, welche dieselben Datenpunkte beschreiben, können leicht miteinander fusioniert werden. Diese Anreicherung mit Information hilft bei den Analysen, denn auf diese Weise können Strukturen erkannt und berücksichtigt werden, die nur in hochdimensionalen Räumen sichtbar werden. Dies ist insbesondere der Fall, wenn sich die multivariate Verteilung der Datenpunkte, nicht aus den jeweiligen Randverteilungen bestimmen lässt, weil die jeweiligen Zufallsvariablen nicht stochastisch unabhängig sind.

Lösungen: Zur Behebung solcher Probleme kann eine Standardisierung der Daten durchgeführt werden, bei der einheitliche Formate, Einheiten und Skalen festgelegt werden. Datenvalidierungsverfahren können auch zur Identifizierung und Behebung von Inkonsistenzen eingesetzt werden.

3.1.2. Ungenaue Daten

Dies beinhaltet Daten, die zwar in der "Nähe" der richtigen Daten sind, aber auch nicht exakt sind.

Ungenaue Daten können durch menschliche Fehler, oder durch technische Fehler in der Datenerfassung entstehen.

Fallbeispiel

In einem Unternehmen werden Verkaufsdaten aus verschiedenen Filialen zentral gesammelt mit dem Ziel, Einblicke in das Kundenverhalten zu gewinnen. Jede Filiale verwendet ein eigenes Kassensystem, um Verkäufe zu erfassen. Einige Kassensysteme liefern nur das Datum (den Tag) der Transaktion, nicht aber die genaue Uhrzeit, was einzelne Analysen im BI-System erschwert.

Weitere beispielhafte Entstehung ungenauer Daten:

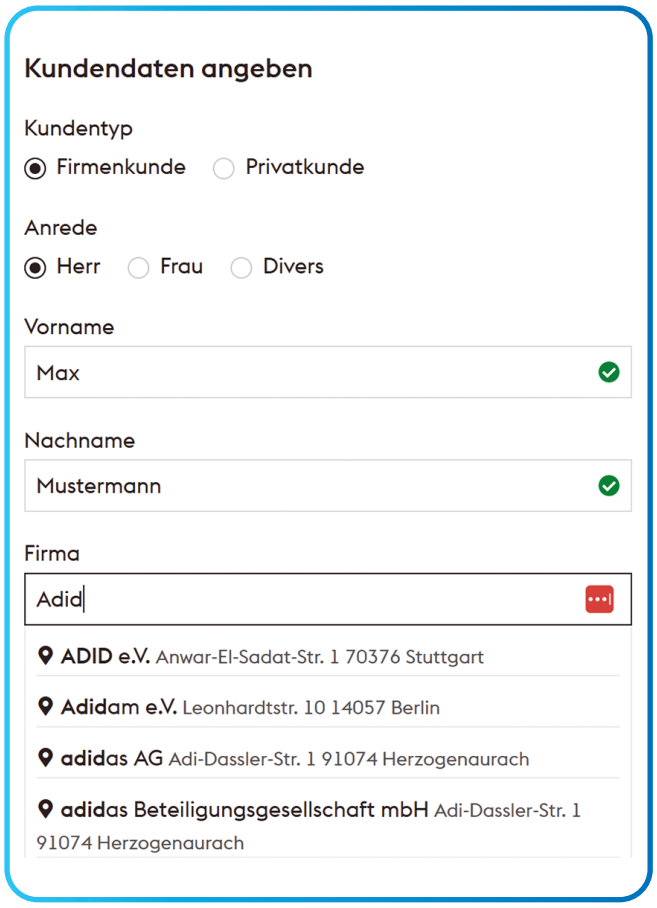
- **Manuelle Dateneingabe:** Mitarbeiter in einem Unternehmen erfassen Kundeninformationen manuell in einem CRM-System. Tippfehler oder Missverständnisse führen zu ungenauen Einträgen wie falschen Adressen, Namen oder Geburtsdaten.
- **Veraltete Messgeräte:** In einem Fertigungsbetrieb werden alte Messgeräte eingesetzt, die nur Messwerte in größeren Schritten anzeigen können. Dadurch können feinere Abweichungen in den Produktabmessungen nicht genau erfasst werden, was zu Ungenauigkeiten führt.
- **Datentransformationen:** Bei der Migration von Daten zwischen verschiedenen Systemen werden komplexe Daten, wie z.B. Zeitstempel, manchmal vereinfacht oder abgerundet, um den Anforderungen des neuen Systems zu entsprechen. Diese Transformationen können zu einem Verlust an Genauigkeit und Detail führen.

Nachteile: Ungenaue Daten können zu ungenauen Analysen führen. Wenn eine systemische Ungenauigkeit vorliegt, dann führen ungenaue Daten tendenziell zu einer stärkeren Modellverzerrung (Bias). Wenn randomisierte Ungenauigkeiten vorliegen, dann führen ungenaue Daten tendenziell zu erhöhter Modellvarianz.

Vorteile: Wenn solche Ungenauigkeiten identifiziert und behoben werden, führt dies zu präziseren und zuverlässigeren Daten, die bessere Analysen und Vorhersagen ermöglichen, da Modellverzerrung und Modellvarianz reduziert werden. Der große Vorteil hier ist, dass wir sowohl die Modellverzerrung als auch die Modellvarianz verbessern können. Dies ist mit anderen Ansätzen, wie der Einstellung von Regularisierungsparametern[4], oft nicht möglich – dort müssen wir oft einen Kompromiss eingehen.

Lösungen: Validierungsprozesse können helfen, Ungenauigkeiten in den Daten zu identifizieren, und Datenreinigungsprozesse helfen, diese zu korrigieren. Zudem kann die Verwendung genauerer Datenerfassungswerkzeuge hilfreich sein. Genauer: Zum Beispiel kann die Detektion von Concept Drift oder Data Drift helfen.

[4]Regularisierungsparameter sind mitverantwortlich für die Einstellung von Modellkomplexitäten. Richtig eingestellt, stellen sie eine Balance zwischen Modellverzerrung und Modellvarianz ein.



FLYERALARM

Wie hier bei unserem Kunden Flyeralarm werden bereits beim Eintippen des Firmennamens im Registrierungsprozess „Adidas“, passende Unternehmen mit ihrer Adresse vorgeschlagen. Das reduziert Datenfehler. Außerdem werden im Hintergrund Daten wie Branche, Unternehmensgröße, Umsatzsteuer-ID etc. angereichert. Dies verbessert die Datenlage vom Online-Shop bis hin zum CRM.

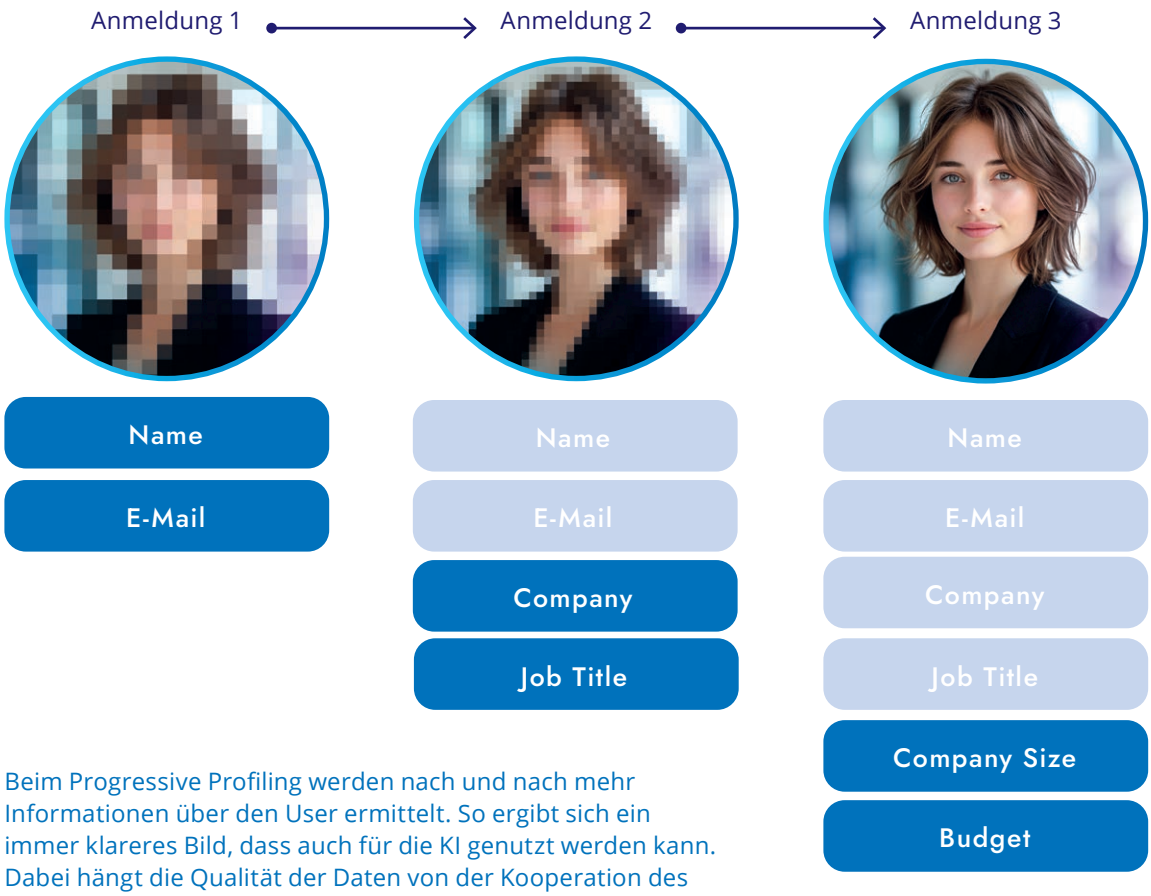
3.1.3. Beliebige Daten

Dies beinhaltet Daten, die falsch sind. Zum Beispiel könnte eine Person bei einer Umfrage ihre Altersinformation falsch eingeben.

Entstehung: Falsche Daten können durch menschliche Fehler, Fehlfunktionen von Sensoren oder anderen Datenerfassungswerkzeugen, oder durch betrügerische Absichten entstehen.

Fallbeispiel

In einem Unternehmen wurden Daten durch *Progressive Profiling* erhoben, um mehr über die Kunden zu erfahren. *Progressive Profiling* ist eine Technik im digitalen Marketing, bei der schrittweise und über die Zeit hinweg zusätzliche Informationen über einen Kunden gesammelt werden. Anstatt von Kunden zu verlangen, dass sie bei der ersten Interaktion eine Vielzahl von Daten preisgeben, werden Informationen in kleineren Portionen und über mehrere Interaktionen hinweg erfasst. Die Befragung erfolgt über eine Online-Plattform, auf der Kunden verschiedene Fragen zu sich beantworten.



Beim Progressive Profiling werden nach und nach mehr Informationen über den User ermittelt. So ergibt sich ein immer klareres Bild, dass auch für die KI genutzt werden kann. Dabei hängt die Qualität der Daten von der Kooperation des Users ab. Unabhängige externe Datenquellen könnten eine Alternative darstellen.

Gute Daten sind das Fundament im KI-Bereich – dies gilt nicht nur für die Generative AI. Auch traditionelle KI-Methoden und Data Science Modelle sind auf korrekte und präzise Daten angewiesen. Die Effektivität von Lagerprognosen, Kunden-Personalisierungen und viele andere Anwendungen hängen maßgeblich von der Datenqualität ab. Sind diese Daten fehlerhaft oder ungenau, verlieren solche Systeme schnell ihre Nutzbarkeit. In der Finanzbranche können unzureichende Daten zu fehlerhaften Risikoanalysen führen, im Marketing die Zielgruppenansprache verzerren und in der Fertigungsindustrie die Produktionsplanung beeinträchtigen. Kurz gesagt, die Qualität der Daten bestimmt die Effizienz und Genauigkeit der KI-gestützten Prozesse und ist damit entscheidend für den unternehmerischen Erfolg.

Carsten Kraus
Omikron Gründer &
KI-Vordenker

Beispielhafte Entstehung falscher Daten:

- **Menschliche Fehler:** Einige Kunden könnten versehentlich falsche Informationen eingeben, z.B. ein falsches Alter oder eine falsche E-Mail-Adresse.
- **Absichtliche Fehler:** In einigen Fällen könnten Teilnehmer absichtlich falsche Angaben machen, sei es aus Spaß, zur Verschleierung echter Daten oder um die Ergebnisse der Umfrage zu manipulieren.
- **Fehlfunktionen der Datenerfassungswerkzeuge:** Technische Probleme mit der Online-Plattform sind wohl eher selten, könnten aber ebenfalls dazu führen, dass einige Antworten falsch aufgezeichnet werden.

Nachteile: Sie können zu irreführenden Analysen und ungenauen Modellvorhersagen führen, die das Vertrauen in datenbasierte Entscheidungsprozesse untergraben. Tatsächlich führen falsche Daten zu sehr unterschiedlichen Problemen.

Vorteile: Durch das Erkennen und Beheben von Fehlern in den Daten verbessert sich die Datenintegrität, was zu verlässlicheren Analysen und Entscheidungen führt.

Lösungen: Datenvalidierung und -reinigung können dazu beitragen, falsche Daten zu identifizieren und zu korrigieren. Auch ein Feedback-Loop mit den Datenlieferanten kann hilfreich sein, um Fehlerquellen zu minimieren. Hier können Methoden aus der Anomaliedetektion, aber auch Methoden aus anderen Bereichen helfen.

3.1.4. Veraltete Daten

Daten, die nicht mehr aktuell sind oder deren Relevanz mit der Zeit abnimmt, können das Ergebnis der Datenanalyse beeinflussen. Zum Beispiel könnte ein Datensatz, der das Kundenkaufverhalten von vor zehn Jahren darstellt, nicht mehr relevant sein, um aktuelle Kaufmuster vorherzusagen.

Entstehung: Daten können veralten, wenn sie über einen längeren Zeitraum gespeichert werden und die Realität, genauer gesagt, die Entität, die modelliert werden soll, sich inzwischen verändert hat.

Fallbeispiel

Ein Händler kaufte die Assets eines insolventen Handelsunternehmens, u.a. die Kundendaten. Die Daten wurden einfach in die bestehende SAP Marketing Cloud hochgeladen. Da beide Händler eine betagte Zielgruppe bewerben, sind viele Personen bereits verstorben oder teilweise sogar doppelt im System (siehe Duplikate).

Beispielhafte Entstehung veralteter Daten:

- **Kundeninformationen:** Das System enthält nun Kundenkontaktdaten und Kaufhistorie, die über Jahre hinweg gesammelt wurden. Viele dieser Daten, besonders Kontaktdetails und Kaufpräferenzen, haben sich im Laufe der Zeit geändert, wurden aber nicht aktualisiert.
- **Produktinformationen im Online-Shop:** Der zugekaufte Online-Shop auf Shopware-Basis könnte veraltete Produktinformationen oder Preise enthalten, da diese nicht regelmäßig mit den aktuellen Daten aus dem ERP-System (MS Navision) abgeglichen werden.
- **Vertriebsdaten in MS Navision:** Verkaufszahlen oder Kundeninformationen in MS Navision könnten veraltet sein und nicht die aktuellen Marktbedingungen oder Kundenbedürfnisse widerspiegeln.

Nachteile: Veraltete Daten können dazu führen, dass Analysen und Modelle nicht mehr aktuell sind und nicht mehr relevante oder nützliche Vorhersagen liefern.

Vorteile: Die Aktualisierung von Daten sorgt dafür, dass Analysen und Modelle auf den neuesten und relevantesten Informationen basieren, was zu genaueren und nützlicheren Vorhersagen führt.

Lösungen: Es ist wichtig, regelmäßige Datenaktualisierungen durchzuführen und veraltete Daten zu entfernen oder zu archivieren. Ein effektives Datenmanagement kann dabei helfen, die Aktualität der Daten zu gewährleisten. Gerade um diese Datenmängel zu verhindern, reicht es nicht einmalige Datenqualitätsmaßnahmen durchzuführen. Stattdessen müssen kontinuierliche Prozesse etabliert werden, welche die Datenqualität grundsätzlich überprüfen und aufrechterhalten.

3.2. Fehlende Daten

3.2.1. Lückenhafte Daten

Dies sind Datenmengen, die fehlende Werte aufweisen.

Entstehung: Lücken können durch verschiedene Faktoren verursacht werden, darunter technische Fehler während der Datenerhebung, menschliche Fehler oder Situationen, in denen bestimmte Daten einfach nicht erfasst werden können oder dürfen.

Fallbeispiel

In einem Konzern werden bestimmte Gewohnheiten und Vorlieben der Kunden systematisch erfragt. In den asiatischen Ländern konnten die Daten erhoben werden, während die DSGVO dies in EU-Ländern verhinderte. Selbst in einzelnen Ländern kam es zu lückenhaften Daten, so wurden in den USA die Kundenverhalten in einzelnen Staaten erhoben, aber in Kalifornien wurde dies durch die CCPA verhindert.

Beispielhafte Entstehung lückenhafter Daten:

- **Unterschiedliche Datenschutzbestimmungen:** In Dubai wurden detaillierte personenbezogene Daten der Kunden erfasst, wie etwa die Freundlichkeit am Telefon, die aufgrund strengerer Datenschutzbestimmungen und fehlendem berechtigtem Interesse in Deutschland nicht erhoben werden durften. Dies führt zu lückenhaften Datensätzen, wenn man die Kundendaten aus beiden Ländern vergleicht.
- **Fehlende Integration zwischen Systemen (MS Navision und Online-Shop):** Es gibt keine vollständige Integration zwischen dem ERP-System (MS Navision) und dem Online-Shop. Bestimmte Transaktionsdaten, die im Online-Shop erfasst werden, wie spezifische Kaufpräferenzen in einzelnen Kategorien, werden nicht automatisch in das ERP-System übertragen.
- **Unvollständige Kundendatenerfassung (Salesforce.com CRM):** In Salesforce.com werden Kundenkontakte und -interaktionen manuell eingegeben. Aufgrund menschlicher Fehler oder Zeitmangels bei den Mitarbeitern sind diese Datensätze oft unvollständig, was zu Lücken in der Kundenhistorie führt.
- **Fehlende Merkmale im eCommerce:** Ein Beispiel-Händler erhält Modeartikel von unterschiedlichen Herstellern. Während alle Hersteller ein Minimum an Beschreibungstexten („Ein rubinrotes Abendkleid…“) an das bestehende PIM-System liefern, fehlen häufig die verfügbaren Farbvarianten des Produkts. Präferenzen der Käufer können nicht in der Marketing-Kommunikation genutzt werden. Das Daten-Team muss die Merkmale aus den Beschreibungstexten (z.B. die Farbe rot) herausfiltern, um sie für Analysen zu nutzen, aber auch um sie in der Such-Navigation als Filter für den Nutzer anzubieten.

Nachteile: Viele Methoden können nicht mit Lücken umgehen; in solchen Fällen müssen Daten weggelassen oder imputiert werden.

Vorteile: Vollständige Datenmengen eröffnen uns die Möglichkeit, alle möglichen herkömmlichen Machine Learning Methoden zu verwenden.

Lösungen: Methoden zur Behandlung lückenhafter Daten umfassen das Weglassen von Attributen oder von Datenobjekten – letzteres ist natürlich nur eine Option während des Trainings. Alternativ können wir Imputationsverfahren zur Schätzung fehlender Werte verwenden. Oder wir wenden Methoden an, die lückenhafte Daten aus anderen Merkmalen ableiten können.

3.2.2. Fehlende Merkmale

Manche Datenmengen enthalten zwar gewisse Informationen über den betrachteten Sachverhalt, aber nicht Informationen über alle Merkmale, die nötig sind, damit die Datenmenge repräsentativ ist. Oft meint man mit fehlenden Markmalen solche Daten, die für einen trainierten Algorithmus den Eingang, nicht den Ausgang repräsentieren; in anderen Worten, Label sind hier oft nicht gemeint, auch wenn diese streng genommen ebenfalls Merkmale sind.

Entstehung: Typische Gründe für fehlende Merkmale sind, dass keine entsprechende Datenerhebung stattfand, und dass Datenmengen isoliert voneinander sind.

Fallbeispiel

Ein Unternehmen aus der Telekommunikationsbranche möchte entscheiden, wo Glasfaserleitungen in einzelnen Gewerbegebieten verlegt werden sollen. Besonders Gewerbegebiete mit einer hohen Dichte an IT-Firmen sollen bevorzugt werden, allerdings fehlt die Brancheninformation bei den Unternehmensdaten aus den Gewerbegebieten.

Beispielhafte Entstehung fehlender Merkmale:

- **Unvollständige Datenerhebung:** Bei der Erfassung der Unternehmensdaten wurden keine spezifischen Informationen über die Branchenzugehörigkeit der Unternehmen erhoben.
- **Isolierte Datenquellen:** Die Daten zu den Gewerbegebieten sind möglicherweise auf verschiedene Systeme verteilt und enthalten nicht alle relevanten Informationen. Zum Beispiel könnten Brancheninformationen in einem anderen System gespeichert sein, das nicht mit dem Hauptdatensystem verbunden ist.

Nachteile: Durch die mangelnde Information können anspruchsvolle Modelle und einfache Statistiken nicht erstellt und manche Erkenntnisse können nicht gewonnen werden.

Vorteile: Wenn Merkmale vollständig vorhanden sind, eröffnen sich völlig neue Herangehensweisen Erkenntnisse zu gewinnen. Dies kann oft einen großen Unterschied machen für den generierten Nutzen.

Lösungen: Die wichtigsten Lösungen sind Daten ordentlich zu erheben und Datensilos zusammenführen. Letzteres ist dabei meist preiswerter und sauberer. Fehlertolerante Matching-Verfahren (z.B. Omikrons FACT®, WorldMatch®) erhöhen die Datenqualität zusätzlich, weil mehr gefunden wird und die Zuordnung verlässlicher geschieht.

3.2.3. Fehlende Daten-Labels

Manchmal fehlen Daten Labels, die für das Training benötigt werden oder es fehlen Datenpunkte, die gelabelt sind. Welche dieser beiden Fälle vorliegt, ist oft eine praktische Frage der Problembeschreibung. Es ist auch möglich, fehlende Labels als das Fehlen von Merkmalen aufzufassen.

Entstehung: Daten-Labels können fehlen oder nur begrenzt verfügbar sein aufgrund von Kosten, Zeit oder, wenn menschliche Expertise fehlt, die notwendig ist, um korrekte Labels zu erstellen.[5]

Fallbeispiel

Ein Unternehmen arbeitet im ERP mit Materialstammdaten und sogenannten ECLASS-Klassifizierungen, ein Datenstandard für die Klassifizierung von Produkten. Während ein Großteil der Bauelemente aus Eisen eine ECLASS-Klassifizierung aufweist, sind bei den Maschinenteilen nur wenige Datensätze mit einem ECLASS-Label versehen.

Beispielhafte Entstehung fehlender Daten-Labels:

- **Unvollständige Erfassung von Kundeninteraktionen:**
  Ein Dienstleistungsunternehmen erfasst Kundeninteraktionen über verschiedene Kanäle, aber nicht alle Interaktionen werden mit den entsprechenden Ergebnissen (z.B. Verkauf abgeschlossen, Supportanfrage gelöst) gelabelt. Dies erschwert die Analyse der Kundenzufriedenheit und die Optimierung der Serviceprozesse.
- **Fehlende Kundenbewertungen in Kundenstammdaten:**
  Ein Einzelhandelsunternehmen sammelt Kundendaten, aber viele Kundenprofile enthalten keine Informationen über frühere Käufe oder Kundenbewertungen. Diese fehlenden Label begrenzen die Möglichkeit, personalisierte Marketingstrategien zu entwickeln und die Kundenbindung zu verbessern.
- **Fehlende Attribute in Produktfeeds:**
  Für einen Produktfeed, der an Preissuchmaschinen und Marktplätze gesendet wird, fehlen häufig spezifische Attribute wie Farbe oder Größe, die für die Darstellung im Online-Marketing entscheidend sind. Dadurch sind die Produkte weniger auffindbar und können von potenziellen Kunden übersehen werden.

Nachteile: Das Fehlen von gelabelten Daten kann die Anwendung von überwachten Lernalgorithmen einschränken und dazu führen, dass Modelle ungenau oder unzuverlässig sind. Es kann auch den Trainingsprozess verlangsamen, da das Modell möglicherweise mehr ungelabelte Daten benötigt, um nützliche Erkenntnisse zu gewinnen – dafür müssen allerdings besondere semi-überwachte Methoden verwendet werden, die oft nicht verfügbar sind.

[5]"As a consequence, model builders are working hard to improve the quality of the inputs they already have. Many ai labs employ armies of data annotators to perform tasks such as labelling images and rating answers." laut https://www.economist.com/business/2023/08/13/ai-is-setting-off-a-great-scramble-for-data

Eine wesentliche Herausforderung für die Datenqualität sind Silos im Unternehmen. Datensilos entstehen, wenn die einzelnen Bereiche eines Unternehmens über eigene Datensammlungen verfügen.[6]

Hörtipp: Wir empfehlen den Podcast „My Data Is Better Than Yours“ von Jonas Rashedi mit spannenden Gästen aus der Welt der Daten.

Jonas Rashedi
Podcaster und Buchautor

[6] Jonas Rashedi, “Datengetriebenes Marketing: Wie Unternehmen Daten zur Skalierung ihres Geschäfts nutzen können”, Springer Gabler, 2020, ISBN: 9783658308414, S. 16.

Vorteile: Das Vorhandensein von gelabelten Daten ermöglicht die Anwendung von überwachten Lernalgorithmen und führt zu genaueren und verlässlicheren Modellvorhersagen. Dies kann auch den Trainingsprozess beschleunigen, da das Modell mit weniger Daten effektive Vorhersagen treffen kann.

Lösungen: Zur Generierung von gelabelten Daten können Techniken wie Data Augmentation, semi-supervised learning oder active learning verwendet werden. Zudem kann die Zusammenarbeit mit Fachleuten zur Generierung von Labels hilfreich sein. In einigen Fällen können auch automatisierte Labeling-Tools oder Crowdsourcing-Plattformen zur Generierung von gelabelten Daten beitragen. Hier kann der Human-in-the-Loop Ansatz (siehe 4.2) einen entscheidenden Beitrag liefern.

3.2.4. Isolierte Datenmengen

Wenn Daten in unterschiedlichen Tabellen, Datenbanken, in Silos eines Unternehmens liegen oder in anderer Art und Weise voneinander isoliert sind, kann das Problem isolierter Datenmengen vorliegen. Hier können einzelne Datenpunkte nicht einander zugeordnet werden, weil die Verbindungs- oder Zuordnungsinformation fehlt. Wenn einem diese Zuordnung fehlt und man nur eine Datenmenge besitzt, dann fehlen effektiv relevante Merkmale (siehe 3.2.2) und die Daten sind nicht repräsentativ (siehe 3.4.1).

Entstehung: Datenisolation entsteht z.B. wenn unterschiedliche Abteilungen oder Projektgruppen unterschiedliche Merkmale pflegen und keine Absprache stattfindet.

Fallbeispiel

Ein Unternehmen wurde von einem Private Equity Investor zum Kauf weiterer Unternehmen gedrängt. Jedes neue Unternehmen bringt eigene Datensilos zu den bereits bestehenden IT-Systemen hinzu. Erst durch den Aufbau einer MDM-Plattform können Silos aufgelöst werden und die Daten dem BI-Team verfügbar gemacht werden, die nun mit Power-BI strategische Empfehlungen in die Fachbereiche geben.

Beispielhafte Entstehung isolierter Datenmengen:

- **Unterschiedliche Datenquellen:** Jedes dieser Systeme sammelt und speichert Daten unabhängig voneinander. Beispielsweise erfasst MS Navision Finanz- und Bestandsdaten, während SAP Marketing Cloud und Hubspot Informationen über Kundeninteraktionen und Marketingaktivitäten sammeln.
- **Keine einheitliche Kunden-ID:** Es gibt keine standardisierte Kunden-ID, die über alle Systeme und Unternehmenstöchter hinweg verwendet wird. Dies erschwert die Zusammenführung und Analyse der Daten aus verschiedenen Quellen.

Nachteile: Beispielsweise, wenn eine Datenbank mit Kauftransaktionen vorliegt und ein Newsletter-Tool das Verhalten der Nutzer enthält, wie Öffnungs- und Click-Through-Raten, und kein Zusammenhang zwischen den Datenquellen hergestellt werden kann, dann kann kein reguläres Model über beide Datenmengen hinweg trainiert werden.

Vorteile: Werden Datenmengen verbunden, erhält man eine Gesamtdatenlage, welche repräsentativ für den betrachteten Sachverhalt wird. Auf diese Weise können Modelle erstellt werden, welche zuvor nicht erstellt werden konnten. Über anspruchsvolle Modelle, aber auch mit Hilfe simpler Statistiken, können nun Erkenntnisse abgeleitet werden, die zuvor unzugänglich waren.

Lösungen: In unseren Projekten bei großen Organisationen beginnt die Arbeit meist mit einer Detektivarbeit. Welche Silos existieren eigentlich? Anschließend ist Technologie und Expertise gefragt, um eine verlässliche Verbindung zwischen den Datenmengen herzustellen. Diese Verbindung wird mit Hilfe einer ID hergestellt, die in mehreren Datenmengen vorhanden sein muss. Wir stellen diese ID über die sogenannte Omikron-ID her, die wir mit Hilfe unserer Software bestimmen.

3.3. Übermäßige Daten

3.3.1. Irrelevante Daten

Dies sind Daten, die für die Analyse oder Vorhersage nicht relevant sind. Zum Beispiel könnten in einem Datensatz, der dazu verwendet wird, Kundenkaufverhalten vorherzusagen, Informationen über das Wetter enthalten sein, die für die spezifische Analyse irrelevant sein könnten.

Entstehung: Irrelevante Daten können auftreten, wenn zu viele Daten gesammelt werden oder wenn der Datensammelprozess nicht gut auf die spezifischen Ziele der Datenanalyse abgestimmt ist.

Fallbeispiel

In einem Unternehmen wurde eine strategische Geschäftseinheit aufgelöst. Bei einer Analyse der Kundendaten sind allerdings ein Großteil der Daten von Kunden, die ausschließlich Produkte dieser Geschäftseinheit gekauft haben.

Beispielhafte Entstehung übermäßiger Daten:

- **Datenüberfluss:** Ihr Konzern sammelt umfangreiche Daten aus verschiedenen Quellen in einem Data Lake. Der Aufbau des Data Lakes erfolgte ohne klare Zielsetzung und ist zu durcheinander, um daraus zielgerichtet eine KI zu trainieren.
- **Kundenkaufverhalten:** Sie möchten das Kundenkaufverhalten analysieren, um zukünftige Verkaufsstrategien zu optimieren. Der Datensatz enthält jedoch auch Informationen wie Wetterdaten oder allgemeine Nachrichten, die für diese spezifische Analyse irrelevant sein könnten.

Nachteile: Sie können dazu führen, dass Ressourcen verschwendet werden und dass Analysen und Modelle durch unnötige Daten überladen werden.

Vorteile: Durch die Identifizierung und Entfernung irrelevanter Daten wird der Datensatz effizienter und fokussierter, was zu effizienteren Analysen führt und besserer Produktivität der Analysten führt.

Lösungen: Die Verwendung geeigneter Datenextraktionswerkzeuge und die Anwendung von Feature-Auswahlverfahren können dazu beitragen, irrelevante Daten zu identifizieren und zu entfernen. Dazu können Verfahren gehören, die direkt in die finalen Machine Learning Modelle integriert werden. Diese Verfahren sind sehr angepasst an den Anwendungsfall. Das ist sowohl ein Vorteil als auch ein Nachteil. Wenn wir stattdessen Verfahren verwenden, die unabhängig von dem finalen Machine Learning sind, können wir die entsprechenden Daten einerseits für unterschiedliche Anwendungsfälle nutzen und andererseits sind wir frei in unserer Kreativität die Machine Learning Methoden zu verwenden, die wir als beste ansehen.

3.3.2. Ausreißer

Diese sind Datenpunkte, die sich stark von anderen Punkten in der Datenmenge unterscheiden. Manchmal sind solche Ausreißer nicht fälschlicherweise in den Daten vorhanden, sondern sie sind tatsächlich ein extremes Beispiel schiefer Daten.

Entstehung: Ausreißer können durch Fehler, ungewöhnliche Ereignisse oder andere extreme Variationen in den Daten entstehen.

Beispielhafte Entstehung von Ausreißern:

1. **Ungewöhnlich hohe Verkaufstransaktion im ERP-System (MS Navision):**
   Ein Datensatz in MS Navision zeigt eine Verkaufstransaktion, die zehnmal höher ist als der Durchschnitt. Dies könnte auf einen Eingabefehler zurückzuführen sein oder eine tatsächlich stattgefundene Großbestellung repräsentieren. Ohne weitere Überprüfung könnte dieser Ausreißer die Analyse von Verkaufstrends und Prognosen verzerren.

2. **Extreme Klickrate in einer E-Mail-Kampagne (SAP Marketing Cloud):**
   In der SAP Marketing Cloud wird eine E-Mail-Kampagne aufgezeichnet, die eine außergewöhnlich hohe Klickrate hat, die weit über dem Durchschnitt anderer Kampagnen liegt. Sie stellen fest, dass es sich um einen technischen Fehler im Tracking handelt.

3. **Anomalie in Kundenfeedback (Salesforce.com CRM):**
   Im Salesforce.com könnte ein Kundendatensatz außergewöhnlich negatives Feedback zeigen, das stark von der allgemeinen Kundenzufriedenheit abweicht. Dies könnte entweder einen Einzelfall eines sehr unzufriedenen Kunden oder einen Fehler bei der Erfassung des Feedbacks darstellen.

4. **Unrealistische Lagerbestände im Online-Shop (Shopware):**
   Im Online-Shop könnte ein Artikel mit einer extrem hohen Lagermenge geführt werden, was realistisch nicht möglich ist. Dies könnte auf einen Synchronisationsfehler mit dem ERP-System oder einen Tippfehler bei der manuellen Bestandspflege hinweisen.

Nachteile: Ausreißer können das Ergebnis von Analysen verzerren (also einen Bias in das Modell einführen), wenn sie nicht richtig behandelt werden.

Vorteile: Durch die korrekte Erkennung und Behandlung von Ausreißern wird die Datenqualität so verbessert, dass – aufgrund von weniger Modellverzerrung - präzisere Analysen und verlässlichere Modellvorhersagen gemacht werden können und das ohne die Modellvarianz zu verschlechtern.

Lösungen: Es gibt verschiedene Techniken zur Erkennung und Behandlung von Ausreißern, einschließlich statistischer Methoden und maschinellen Lernverfahren.

3.3.3. Duplikate

Datenmengen aus Organisationen enthalten oft doppelte Datenpunkte. Duplikate sind ein häufiger Grund für schiefe Daten. Duplikate sind ein Spezialfall fehlerhafter Daten – eigentlich hätte nur ein Datenpunkt in den Daten sein sollen. Es ist beispielsweise nicht selten, dass für die größten Kunden einer Organisation auch die meisten Duplikate bestehen.

Duplikate verstärken das fehlerhafte Trainieren der Modelle. Weichen die Duplikate leicht voneinander ab, dann wird es umso wichtiger, dass professionelle Tools eingesetzt werden, um diese zu matchen.

Entstehung: Duplikate können durch Fehler in der Datenerfassung, mangelnde Kontrolle bei der Dateneingabe (Data Quality Gate) oder durch die Zusammenführung von Daten aus verschiedenen Quellen entstehen (z.B. Zukäufe, Schnittstellen zu anderen Systemen).

Fallbeispiel

Das Datenteam eines Technologieunternehmens verlässt sich auf die integrierte Dublettenprüfung von MS Dynamics. Die Vertriebsmitarbeiter schlagen Alarm, weil sie mit den Daten unzufrieden sind. Erst während der Analyse einer Stichprobe stellt sich eine 20% Dubletten-Quote heraus.

Beispielhafte Entstehung von Duplikaten:

1. **Mehrfache Kundenregistrierungen (Salesforce.com CRM):**
   Unterschiedliche Besteller eines Unternehmens bestellen über unterschiedliche Accounts und wurden so mehrfach im CRM-System registriert. Die tatsächliche Größe des Accounts wird nicht deutlich.

2. **Bestellung im Shopware Online Shop:**
   Ein B2C-Kunde nutzt für mehrere Bestellvorgänge leicht abweichende Namen oder E-Mail-Adressen, weil er den Login vergessen hat.

3. **Wiederholte Transaktionseinträge (MS Navision):**

Aufgrund von Fehlern in der Datenerfassung könnten identische Verkaufstransaktionen mehrfach in MS Navision erfasst werden.

4. **Datenzusammenführung aus verschiedenen Quellen (SAP Marketing Cloud und Hubspot):**
   Beim Zusammenführen von Kundendaten aus verschiedenen Marketing-Tools könnten identische Kundeninformationen ohne entsprechende Abgleichung zu Duplikaten führen.

Nachteile: Duplikate sorgen für schiefe Daten. Es werden gewisse Gruppen oder Datenpunkte auf eine Art und Weise betont, die nicht den wahren Sachverhalt wiedergibt. Die Nachteile beschränken sich aber nicht nur auf Machine Learning Modelle, sondern auch auf Entscheidungen (wie z.B. an wen ein Katalog versandt werden soll), die direkt aus den Daten abgeleitet werden.

Vorteile: siehe 3.1.2 Ungenaue Daten, 3.1.3 Beliebige Daten, 3.4.1 Schiefe Daten, Mangel an Heterogenität.

Lösungen: Datenreinigungsprozesse und Duplikaterkennungstools können zur Identifizierung und Entfernung von Duplikaten verwendet werden.

3.3.4. Daten, die nicht den Datenschutzbestimmungen entsprechen

Es ist möglich, dass Daten gegen Datenschutzbestimmungen verstoßen.

Fallbeispiel

In einem Unternehmen speichert ein Call-Center Team aus Unwissenheit regelmäßig Daten über bestimmte Verhaltensweisen der Kunden, die auf charakterliche Eigenschaften schließen lassen („sehr patzig am Telefon“).

Entstehung: Solche Daten können entstehen, wenn Datenschutzrichtlinien nicht ausreichend befolgt oder verstanden werden, oder wenn Daten ohne die erforderlichen Zustimmungen erfasst oder verwendet werden.

Beispielhafte Entstehung datenschutzrechtlicher Probleme:

- **Unzureichende Befolgung von Datenschutzrichtlinien:** Möglicherweise werden Datenschutzbestimmungen bei der Erfassung, Speicherung oder Verwendung von Kundendaten nicht ausreichend beachtet, z.B. indem zu viele persönliche Informationen gesammelt werden, ohne eine klare Notwendigkeit dafür.
- **Fehlendes Verständnis der Datenschutzbestimmungen:** Mitarbeiter verstehen möglicherweise nicht vollständig die Datenschutzgesetze und -regelungen, was zu Fehlern im Umgang mit sensiblen Daten führt.

- **Daten ohne erforderliche Zustimmungen:** Kundendaten werden möglicherweise ohne explizite Zustimmung der Kunden oder in einer Weise verwendet, die nicht mit ihrer Einwilligung übereinstimmt.

Nachteile: Der Umgang mit datenschutzrechtlich problematischen Daten kann zu rechtlichen Konsequenzen führen und das Vertrauen der Nutzer oder Kunden in die Organisation untergraben.

Vorteile: Durch die Einhaltung von Datenschutzbestimmungen wird das Vertrauen der Nutzer und Kunden in die Organisation gestärkt und das Risiko rechtlicher Konsequenzen minimiert.

Lösungen: Es ist wichtig, dass Organisationen Datenschutzrichtlinien implementieren und befolgen und dass sie geeignete Datenschutzwerkzeuge und -praktiken einsetzen.

3.4. Unrepräsentative Daten

3.4.1. Schiefe Daten, Mangel an Heterogenität

Ein weiteres Problem, das auftreten kann, ist der Mangel an Heterogenität oder Diversität in den Daten. Wenn die Daten nicht repräsentativ für die gesamte Zielpopulation sind, also die Entität, die modelliert wird, kann auch kein Modell entwickelt werden, welches nützlich ist. Zum Bespiel können bestimmte Gruppen über- oder unterrepräsentiert sind. Solche Daten werden auch als schief bezeichnet. In extremen Fällen fehlen Gruppen vollständig oder sind nur als vereinzelte Ausreißer (siehe 3.3.2) enthalten.

Entstehung: Unrepräsentative Daten können entstehen, wenn die Datenerhebung nicht ausreichend divers ist oder bestimmte Gruppen über-, unterrepräsentiert oder gar nicht repräsentiert sind. Die Gründe dafür sind vielzählig, Beispiele sind Duplikate (siehe 3.3.3), fehlerhafte Datenaufnahme, Fehler in vorhergehender Verarbeitung, Fehler beim Zusammenlegen von Daten aus unterschiedlichen Quellen (siehe 3.2.4).

Fallbeispiel

Bei einem Unternehmen wird ein Machine-Learning-Modell entwickelt, um Kundenverhalten vorherzusagen und personalisierte Marketingstrategien zu entwerfen. Die Trainingsdaten für das Modell stammen hauptsächlich aus Online-Transaktionen des Unternehmens. Daten aus den 200 Filialen fehlen und sind damit unterrepräsentiert.

Beispielhafte Entstehung schiefer Daten:

- **Unzureichende Datendiversität:** Die meisten Daten stammen von Online-Käufen, was möglicherweise nicht die gesamte Kundenbasis repräsentiert. Ältere Kunden, die seltener online einkaufen, wurden systematisch unterrepräsentiert.

- **Überrepräsentation bestimmter Gruppen:** Wenn beispielsweise die Mehrheit der Online-Kunden in urbanen Gebieten lebt, könnten die ländlichen Kunden in den Daten unterrepräsentiert sein.
- **Fehler beim Zusammenlegen von Daten aus unterschiedlichen Quellen:** Wenn Online-Daten mit Daten aus anderen Quellen, wie dem CRM-System, zusammengeführt werden, könnten Inkonsistenzen oder Fehler in der Datenintegration zu einer Verzerrung führen.

Nachteile: Dies verfälscht resultierende Modelle und Analysen. Dies kann die Universalität und Anpassungsfähigkeit von Machine-Learning-Modellen oder datengesteuerten Lösungen bedrohen, da sie möglicherweise nicht gut auf verschiedene oder neue Datengruppen reagieren. Neue Datenpunkte, die aus Gruppen stammen, die in den Trainingsdaten überhaupt nicht auftauchten, können nur schwer beurteilt werden.

Vorteile: Durch das Gewährleisten von Repräsentativität und Diversität in den Daten wird die Genauigkeit und Fairness der Modelle verbessert, was zu besseren Vorhersagen und Analysen führt. Durch das Erkennen und Entfernen von Duplikaten werden Modelle für mehr Datengruppen geeignet, da mit den Daten auch die Modelle repräsentativer für die Gesamtdatenmenge werden. In vielen Anwendungsfällen ist aber genau dies der Mehrwert. Wenn unser Modell für jeden neuen Datenpunkt prädiktiert, dass dieser zur Gruppe A gehört, und (fast) nie, dass ein neuer Datenpunkt zu Gruppe B gehört, hätten wir von vorneherein keinen Machine Learning Algorithmus, ja, keine algorithmische Entscheidung entwickeln müssen.

Lösungen: Um Repräsentativität und Diversität in den Daten zu gewährleisten, ist es wichtig, die Datenquellen und die Datenerhebungsmethoden zu diversifizieren und auf eine ausgewogene Repräsentation der Zielgruppen zu achten.

4

4. Schlüsselbegriffe in der KI-Datenwelt

4.1. Einmal-Projekte vs. Kontinuierliche Projekte

Einmalige Projekte und kontinuierliche Projekte repräsentieren zwei kontrastierende Ansätze zur Verbesserung der Datenqualität für die KI-Entwicklung. Diese beiden Projekttypen ergänzen sich und haben ihre Stärken und Schwächen. Einmalige Datenreinigungsprojekte können gezielt auf ein bestimmtes KI-Projekt angepasst werden. Diese Spezialisierung erlaubt es schnell Ergebnisse zu liefern, löst Datenqualitätsprobleme aber nicht nachhaltig. Im Kontrast dazu versprechen kontinuierliche Projekte eine fortlaufende Anpassung an sich verändernde Datenlandschaften. Sie bauen auf der Idee auf, dass Datenqualität kein statisches Ziel ist, sondern ein dynamisches Bestreben. In einem Umfeld, in dem KI-Modelle ständig optimiert werden müssen, stellt sich die entscheidende Frage: **Wie balancieren wir zwischen dem kurzfristigen Bedarf und langfristiger Vision?**

In jedem Fall besteht der erste Schritt zur Verbesserung der Datenqualität darin, ein Bewusstsein für die Bedeutung guter Daten zu schaffen. Wenn Erfolge vorgezeigt werden können, fällt es leichter eine Datenkultur in Ihrem Unternehmen zu etablieren; dies können einmalige Projekte leisten. Um diese sichtbaren Erfolge in einen echten Mehrwert für das jeweilige Unternehmen zu wandeln, sind kontinuierliche Prozesse unabdingbar. In beiden Fällen können unsere Software-Technologien dazu beitragen, diese Datenqualität zu schaffen.

Ein Beispiel, bei dem der Unterschied deutlich wird, ist die Bildung von Golden Records, bei denen Kundendaten in einem vordefinierten Datenmodell gespeichert werden. Golden Records sind jene Daten, die als korrekt angesehen werden und deren Qualitätsmängel behoben sind. In einem einmaligen Projekt kann eine Datenbasis aufgebaut werden.

Um sicherzustellen, dass die Qualität gleichbleibend gut bleibt, müssen neue Datenbestände durch ein Qualitätssicherungssystem fließen (Data Quality Server). So erhält das Datenteam kontinuierlich hochwertige Daten für ihre KI-Modelle.

4.2. Mensch und Maschine Interaktionen

Seit dem Aufkommen der Computerwissenschaften in den 1940er Jahren wurden Maschinen entwickelt, um menschliche Aufgaben zu automatisieren und unsere Fähigkeiten zu erweitern. In der heutigen wettbewerbsintensiven Geschäftswelt sind Unternehmen ständig auf der Suche nach Wegen, ihre Prozesse zu optimieren und ihre Ergebnisse zu verbessern.

Künstliche Intelligenz bietet beeindruckende Möglichkeiten zur Automatisierung und Effizienzsteigerung. Doch trotz der beeindruckenden Fortschritte in der Automatisierung und Künstlichen Intelligenz bleibt eine Tatsache bestehen: Maschinen denken nicht wirklich. In den letzten Jahren sind auch die ethischen Bedenken rund um KI-Anwendung gestiegen. Wie stellen wir sicher, dass unsere Technologien ethische Grundsätze einhalten, insbesondere in Bereichen, in denen die Entscheidungen der Maschinen weitreichende Folgen haben können?

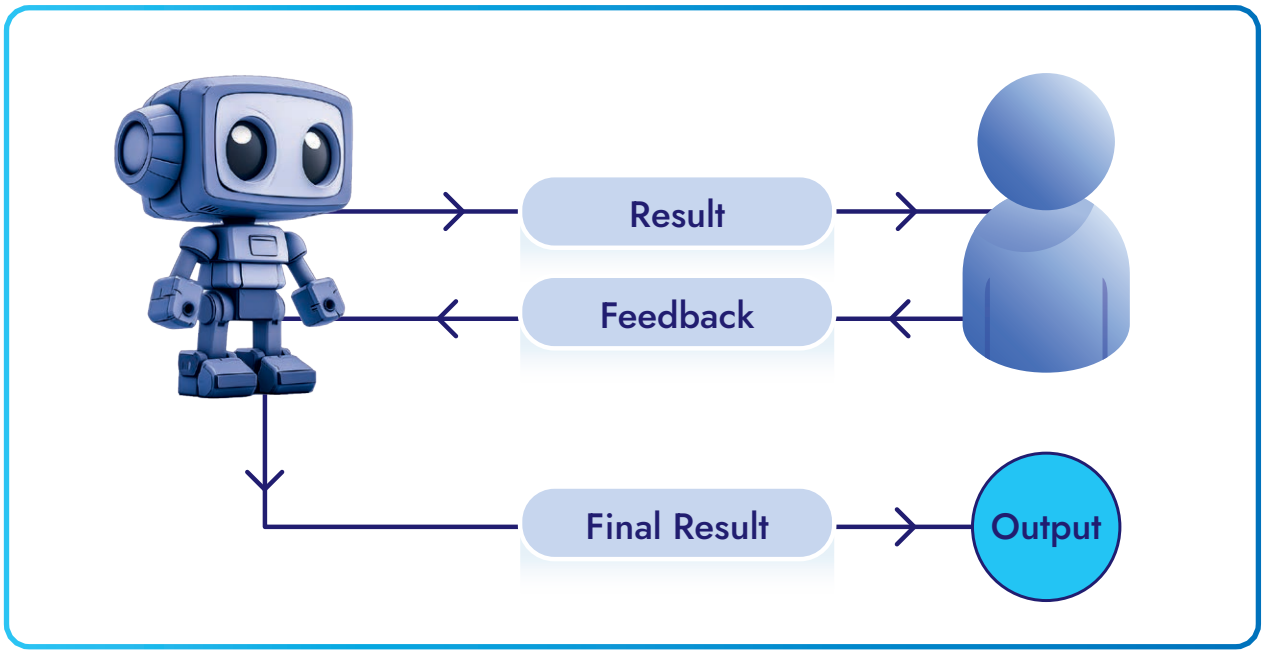
4.2.1. Human in the Loop und Machine in the Loop

Die Ansätze „Human-in-the-Loop“ (HITL) und „Machine-in-the-Loop“ (MITL) erkennen an, dass menschliche Intuition, Urteilskraft und Expertise unersetzlich sind, insbesondere in komplexen oder unvorhersehbaren Situationen. Diese Konzepte bringen den Menschen zurück in die Entscheidungsprozesse und verbinden das Beste aus beiden Welten: Es sind Ansätze, die sowohl Technologie als auch menschliches Talent wertschätzen. Durch die Integration menschlicher Urteilsbildung in automatisierte Systeme wird nicht nur die Genauigkeit und Zuverlässigkeit dieser Systeme erhöht, sondern auch ein ethischer Rahmen geschaffen, der die Werte und Prinzipien unserer Gesellschaft widerspiegelt. Es ist ein Bekenntnis zur Verantwortung, das Menschlichkeit in die Ära der Maschinen bringt.

Beide Ansätze beschreiben einen menschlichen Agenten und eine Maschine als zwei Komponenten. In beiden Ansätzen werden erst in einem Kreislauf Daten zwischen Agenten und Maschine ausgetauscht. Wer die „letzte“ Entscheidung trifft, unterscheidet sich jedoch: im HITL-Ansatz ist dies die Maschine, beim MITL-Ansatz ist dies der Mensch. Letzteres stellt damit den Menschen wieder ins Zentrum der Entscheidungsprozesse. Auch die Art der Information, die zwischen den beiden – Agent und Maschine – ausgetauscht wird, unterscheidet sich im Allgemeinen zwischen den beiden Ansätzen..

4.2.2. Human in the Loop

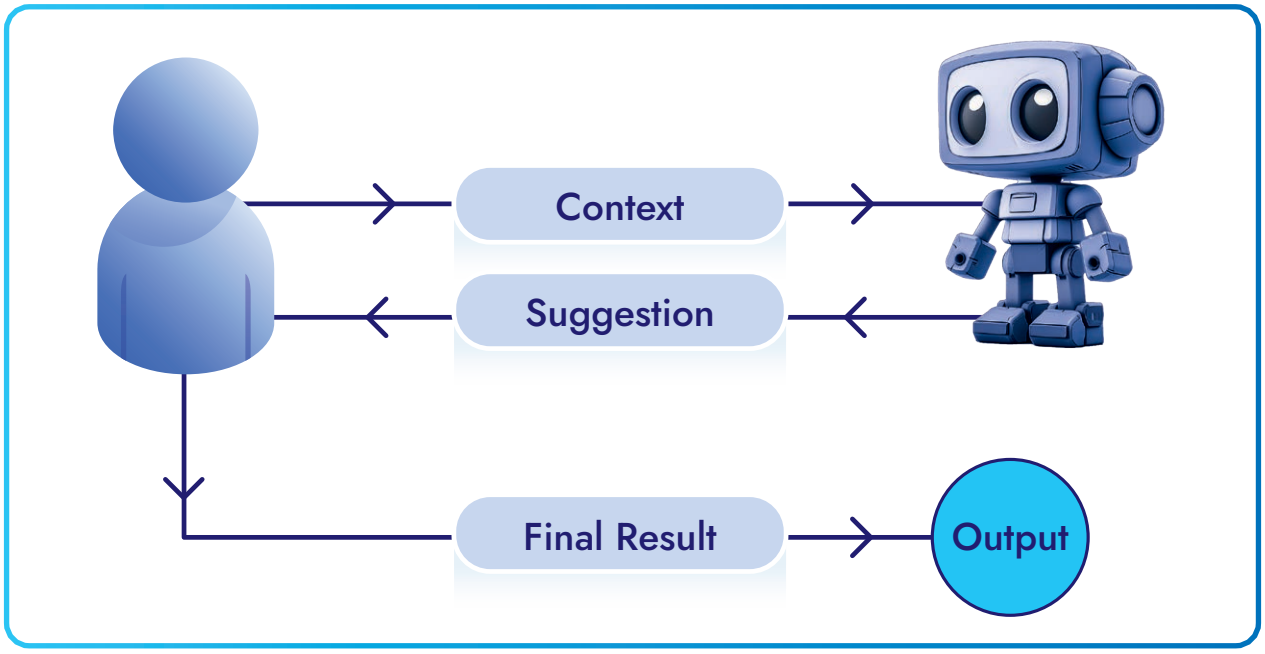
Während eines solchen Prozesses spielt die Maschine Resultate an den Agenten zurück. Dieser liefert Feedback, zum Beispiel Labels oder extrahierte Features oder andere Daten zurück an die Maschine, welche diese Daten in ihre Modelle integriert und entsprechende Methoden anwendet und anpasst. Anschließend werden weitere Resultate an den Agenten geliefert. Solch ein Prozess endet, wenn die Maschine kein weiteres Feedback mehr vom Agenten einholt und ein finales Ergebnis abliefert. Hier wird der Prozess von der Maschine gesteuert, insbesondere welche Anfragen an den Menschen gestellt werden und wann der Prozess beendet werden soll.



Human-in-the-Loop (HITL): Der Prozess illustriert, wie eine Maschine zunächst Ergebnisse erzeugt, die von einem Menschen bewertet werden. Der Mensch liefert Feedback wie Labels zurück an die Maschine. Diese Informationen werden von der Maschine genutzt, um ihre Algorithmen zu verbessern und angepasste Resultate zu produzieren. Der Zyklus wiederholt sich, bis die Maschine ein finales Ergebnis liefert, bei dem kein weiteres menschliches Feedback erforderlich ist. Dieser Ansatz ermöglicht es, menschliche Intuition und Expertise in automatisierte Entscheidungsprozesse zu integrieren und stellt sicher, dass die Maschine Handlungen gemäß ethischen Standards und gesellschaftlichen Werten ausführt.

4.2.3. Machine in the Loop

Während eines solchen Prozesses sendet die Maschine Vorschläge an den Agenten. Dieser liefert weiteren Kontext, z.B. eine Art „nächste Frage“, die mit weiteren Vorschlägen der Maschine beantwortet wird. Solch ein Prozess endet, wenn der Agent keine weitere Hilfe benötigt und ein finales Ergebnis abliefert. Hier wird der Prozess vom Agenten gesteuert, insbesondere welche Anfragen an die Maschine gestellt werden und wann der Prozess beendet werden soll.



Machine-in-the-Loop (MITL): Dieses Diagramm veranschaulicht den Prozess, bei dem eine Maschine Vorschläge an einen Menschen sendet. Der Mensch liefert dann weiteren Kontext – die nächste Frage – woraufhin die Maschine mit weiteren Vorschlägen antwortet. Der Prozess endet, wenn der Mensch entscheidet, dass keine weitere Unterstützung durch die Maschine notwendig ist und das finale Ergebnis selbst liefert.

4.2.4. Anwendungsbereiche für HITL und MITL

Beide Ansätze können genutzt werden, um Menschen aktiv an der Verbesserung der Datenqualität oder Modellentwicklung zu beteiligen, sowohl in einmaligen Projekten als auch in kontinuierlichen Prozessen. Sie sind insbesondere dann wichtig, wenn menschliche Kompetenz unabdingbar ist.

In einem Machine Learning Kontext beschreiben beide Ansätze, wie Menschen in die Trainings- oder Anwendungsprozesse eines Modells eingreifen können, z.B. können Trainingsdaten kuratiert, Parameter von Algorithmen eingestellt, oder die Simulationen von Modellen überprüft werden. Durch solch gezielten Einsatz von HITL kann die Genauigkeit und Robustheit von Machine Learning-Modellen signifikant verbessert werden, da sie von der Kombination aus maschinellem Lernen und menschlichem Expertenwissen profitieren.

Zusätzlich ist es möglich diese Ansätze bereits in der Datenqualitätssicherung einzusetzen. Beide Ansätze können z.B. dazu dienen, Daten zu korrigieren, zu labeln oder zu verfeinern.

Es ist sogar möglich alle drei – Qualitätssicherung, Training und Modellanwendung – als Gesamtheit in HITL oder MITL einzubetten.

HITL und MITL in kontinuierlichen Prozessen

Machine Learning ist nicht nur ein einmaliger Prozess, sondern erfordert oft mehrere Iterationen, um ein optimales Modell zu entwickeln. Der „Human-in-the-Loop“-Ansatz erkennt die dynamische Natur von iterativen Prozessen an, indem es das Modell optimiert und den Menschen als einen entscheidenden Faktor in diese Iterationen integriert. Statt eines Modells, das einmal trainiert und dann in der Produktion eingesetzt wird, ermöglicht HITL eine kontinuierliche Rückmeldung und Anpassung. Dies stellt sicher, dass das Modell nicht nur auf historischen Daten basiert, sondern aktuelle, von Menschen überprüfte Erkenntnisse berücksichtigt. Es ist ein Paradigma, das die adaptive und iterative Natur des Machine Learnings in den Vordergrund stellt.

5. Die Zeit zum Handeln ist jetzt

Die Qualität der Daten ist ein entscheidender Faktor für den Erfolg von KI-Projekten. Effektive und zuverlässige KI-Modelle erfordern nicht nur technische Expertise, sondern ein tiefgreifendes Verständnis für qualitativ hochwertige Daten. Um dieses Ziel zu erreichen, sind folgende Handlungsempfehlungen wichtig:

1. **Bewusstsein für Datenqualität schaffen:**
   Unternehmen müssen die Wichtigkeit von hochwertigen Daten erkennen. Das bedeutet, Aufklärungsarbeit ist der erste Schritt zur Sicherstellung effektiver KI-Modelle. Wer KI nutzen will, muss in Datenqualität investieren.

2. **Implementierung von Qualitätsprozessen:**
   Es ist essenziell, Prozesse zur Datenstandardisierung, -bereinigung und -validierung dauerhaft zu etablieren. Automatisierte Datenbereinigungsprozesse helfen dabei, Inkonsistenzen, Ungenauigkeiten und irrelevante Daten zu identifizieren und zu korrigieren, was letztlich die Grundlage für präzise Analysen und zuverlässige Modellvorhersagen bildet. Zudem erhöht es die Produktivität Ihrer Data Scientists.

3. **Kontinuierliche Überwachung und Aktualisierung:**
   Datenqualität ist kein einmal erreichter Zustand, sondern ein kontinuierlicher Prozess. Regelmäßige Überprüfungen und Aktualisierungen der Daten sind unerlässlich, um ihre Relevanz und Genauigkeit dauerhaft zu gewährleisten.

4. **Integration von Human in the Loop (HITL) und Machine in the Loop (MITL):**
   Die Einbindung menschlicher Expertise in den KI-Prozess ist entscheidend, um die Grenzen der Automatisierung zu überwinden. Durch die Kombination von menschlicher Intuition und maschinellem Lernen können Datenqualität und Modellgenauigkeit signifikant verbessert werden.

5. **Schulung und Weiterbildung:**
   Investitionen in die Schulung und Fortbildung von Mitarbeitern im Bereich Datenmanagement und KI sind unerlässlich. Ein gut ausgebildetes Team, das die Prinzipien der Datenqualität versteht und anwendet, ist ein unschätzbarer Vorteil in jedem KI-Projekt.

Gerne helfen wir Ihnen bei der Umsetzung dieser Schritte und stellen so gemeinsam sicher, dass ihre KI-Projekte auf einem soliden Fundament hochwertiger Daten aufbauen. Dies führt nicht nur zu besseren Ergebnissen, sondern auch zu einer Steigerung des Vertrauens in KI-basierte Entscheidungen, sowohl innerhalb des Unternehmens als auch bei den Endnutzern.

6. Über uns

Bei Omikron sind wir auf einer Mission: Wir pushen die datengetriebene Organisation. In einer digitalen Welt, die von KI und Software geformt wird, beeinflussen Daten alle Prozesse und Entscheidungen. Unsere Mission ist es, Unternehmen bei der Überwindung von drei zentralen Herausforderungen zu unterstützen:

1. **Historisch gewachsene Infrastruktur:**
   Viele Unternehmen kämpfen mit einer Vielzahl von IT-Systemen, die wertvolle Informationen über Kunden verstecken.

2. **Datenverteilung:**
   Typischerweise sind Kundendaten über 7 bis 20 IT-Systeme verstreut.

3. **Vertrauensmangel in Datenqualität:**
   Mitarbeiter zweifeln oft an der Zuverlässigkeit ihrer Unternehmensdaten.

Wir schaffen Abhilfe, indem wir Daten zusammenführen, Datenabläufe automatisieren und saubere, vertrauenswürdige Daten für Fachbereiche wie KI-Teams, Vertrieb, eCommerce und Marketing bereitstellen. Unsere Kunden profitieren von zentralisierten, bereinigten Daten, auf die sie sich für KI-Modelle, Personalisierung, Auswertungen und Berichte verlassen können.

Unsere Erfolge spiegeln sich in Projekten mit Kunden wie Mediamarkt, 1&1, Poco, Flyeralarm und globalen Konzernen wie TeamViewer und Siemens wider.

Wir glauben fest daran, dass Daten Wachstum schaffen, Datensilos aufgebrochen werden müssen und saubere Daten der Schlüssel zur Nutzung von KI sind. Unsere Überzeugung: Die Zukunft gehört denjenigen, die KI mit den richtigen Daten trainieren und nutzen.

Über 200 Marktführer setzen auf Omikron

LG · POCO · SIEMENS · 1&1 · TeamViewer · BERNER · KAISER+KRAFT · SCHÜCO · HELM · FLYERALARM · RADEBERGER GRUPPE · Beuth · Hailo · BECHTLE

Prüfen Sie Ihre Datenqualität kostenlos unter
www.omikron.com/datencheck
Nennen Sie das Code-Wort:
#ebook24

Omikron

The Data Strategy Platform

Jedes Unternehmen möchte daten-getrieben sein. In einer digitalen Welt, die von KI und Software verändert wird, sind es die Daten, die alle Prozesse und Entscheidungen beeinflussen. Mit der Data Strategy Platform von Omikron schaffen Sie Sicherheit, wo andere raten. Nutzen Sie die Macht der Daten, um Ihre Visionen und Ziele zu erreichen.

Kundenwissen aus allen Silos zusammenbringen

★★★★★

Automatisieren aller Datenprozesse

★★★★★

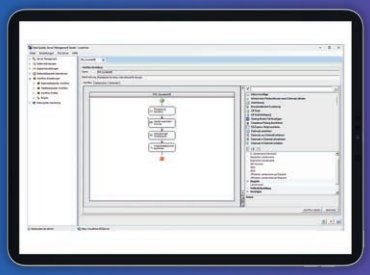
Korrekte Daten für alle Fachbereiche & Data-Teams

★★★★★

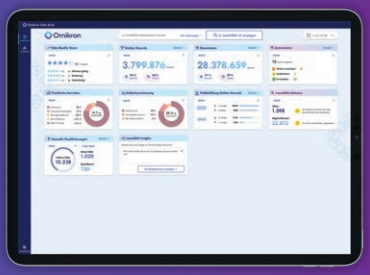
Alle Werkzeuge aus einer Hand für Ihre daten-getriebene Organisation

Daten matchen aus Silos

Datenprozesse automatisieren

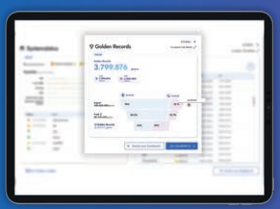


Kundendaten beobachten

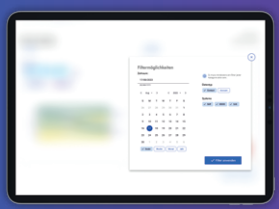


Tiefe BI-Analysen ermöglichen

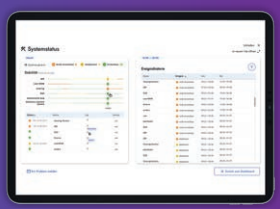
Daten anreichern



Personalisierung erleichtern



Dubletten verhindern



www.omikron.com