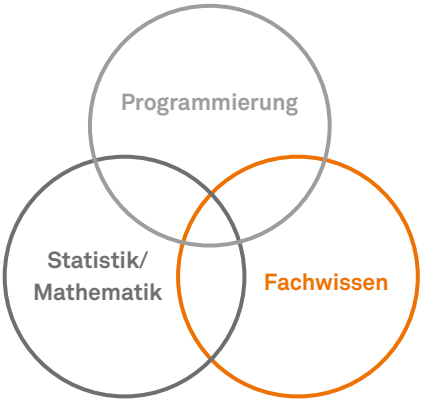
Markus Hausmann

Data Analytics und künstliche Intelligenz

Was ist Data Science?

Im Zuge der fortschreitenden Digitalisierung erfreut sich derzeit kaum ein anderes Thema größerer Beliebtheit als das der künstlichen Intelligenz. In diesem Kontext sind Begriffe wie „Maschinelles Lernen“, „Big Data“, „Data Science“, „Predictive Analytics“, „Advanced Analytics“ und „Künstliche Intelligenz“ (Artificial Intelligence - AI) omnipräsent.

Viele Unternehmen beschäftigen sich aktuell mit dem Nutzen und den Potenzialsteigerungen, die durch Einführung eines Data-Analytics-Systems auf Basis künstlicher Intelligenz entstehen können. Obwohl die Methodiken der künstlichen Intelligenz im wissenschaftlichen Bereich schon seit Langem erforscht werden, haben sie sich bei vielen Unternehmen erst in den vergangenen Jahren etabliert. Maßgeblich verantwortlich für den raschen Erfolg der künstlichen Intelligenz ist zum einen die immer schneller werdende Rechenleistung der Computer, auf Basis derer die Algorithmen der künstlichen Intelligenz laufen. Zum anderen sind die Speicherkapazitäten enorm angestiegen, sodass immer größer werdende Datenmengen kostengünstig archiviert werden können.



Die wesentliche Intention von Data Science besteht darin, eine (bank-)betriebswirtschaftliche Fragestellung in eine datengetriebene Fragestellung zu transformieren. Durch den Einsatz von Algorithmen wird nach Mustern in den Daten gesucht, die zur Lösung der Fragestellung verwendet werden.

Um Data Science betreiben zu können, sind drei Dinge nötig: Idealerweise Kenntnisse

DATA SCIENCE

Als Data Science (Wissenschaft der Daten) werden Methoden und Verfahren bezeichnet, Datenmengen zu strukturieren, zu analysieren, und daraus Entscheidungen abzuleiten. Ein Teilbereich von Data Science ist beispielsweise das maschinelle Lernen.

in einer der dafür relevanten Programmiersprachen (Python, R, Java, Scala ...). Zum anderen Kenntnisse in Mathematik/Statistik, um die Modelle nachvollziehen und erweitern zu können. Und drittens ist domänenspezifisches Fachwissen für die Ergebnisinterpretation und zur passenden visuellen Darstellung erforderlich.

Die Komponente (betriebswirtschaftliches) Fachwissen ist in jedem Finanzinstitut vorhanden und bildet die Basis jedes Data-Science-Anwendungsfalls. Darauf aufbauend können die beiden verbleibenden Komponenten durch darauf spezialisierte externe Dienstleister eingekauft werden. Das Zielbild eines Data-Science-Projektes ist somit die technische Umsetzung einer bankbetriebswirtschaftlichen Fragestellung in eine datengetriebene Fragestellung. Idealerweise erfolgt die Umsetzung direkt in einem Anwendungsprogramm. Somit sind zur späteren Bedienung der Software keine Programmierkenntnisse mehr notwendig.

Um tiefergehende Erkenntnisse aus den Datenmengen ziehen zu können, werden Methoden des maschinellen Lernens verwendet.

Der Vorteil des maschinellen Lernens besteht darin, dass Algorithmen den Datenberg weitgehend automatisiert nach Mustern durchsuchen. Im Vergleich zu einer manuellen Datenaufbereitung durch den Menschen sind Algorithmen emotionslos. Sie können Muster in sehr großen und komplexen Datenstrukturen erkennen, die dem Menschen aufgrund ihrer Größe und Komplexität oft verborgen bleiben. Und sie benötigen im Gegensatz zur klassischen statistischen Datenanalyse keine explizit definierten Modelle.

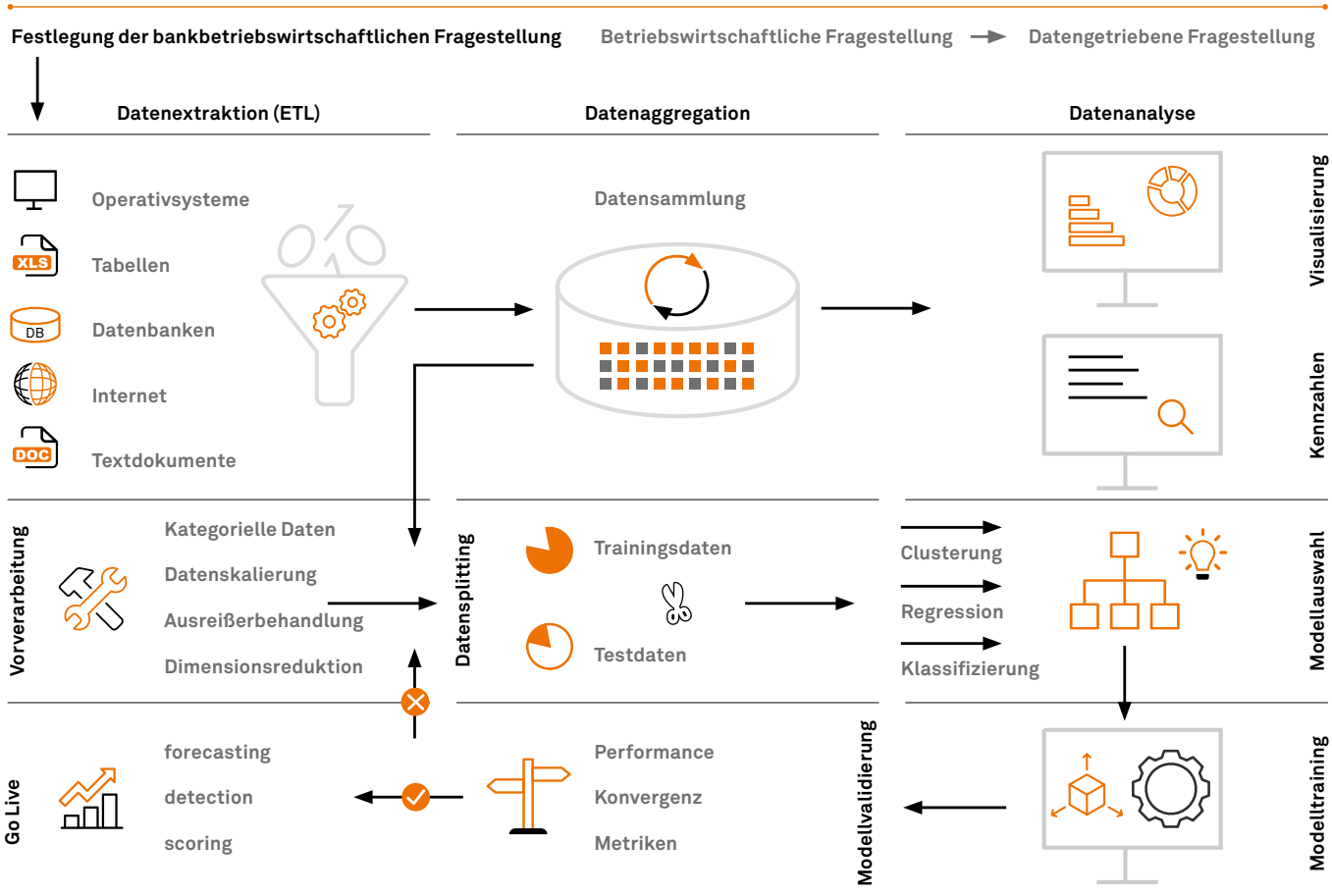
DATA SCIENCE ENTWICKELT SICH RASANT

Der Einsatz von künstlicher Intelligenz startete seine Erfolgsgeschichte in den Unternehmen der Digital Economy, wie Google, Amazon, Facebook etc. Auch die Branche »

MASCHINELLES LERNEN

Unter maschinellem Lernen (Machine Learning) werden, vereinfacht gesprochen, Algorithmen zur Mustererkennung in Daten bezeichnet. Diese Algorithmen werden auf Basis bestehender Datensätze trainiert. Anschließend werden die Algorithmen zur Entscheidungsfindung auf bestehende oder neue Datensätze angewendet.

DATA-SCIENCE-PROZESSPIPELINE



Automotive beschäftigt sich intensiv mit künstlicher Intelligenz, beispielsweise beim autonomen Fahren. Doch auch die Finanzbranche hat die Vorteile durch den Einsatz von Systemen des maschinellen Lernens erkannt und beschäftigt sich intensiv mit möglichen Einsatzgebieten. Beispielsweise können Algorithmen zunehmend zum Erkennen von Betrugsversuchen (fraud detection) oder zur Identifikation von Vertriebspotenzialen verwendet werden. Auch die Aufsicht setzt sich derzeit intensiv mit dem Thema künstliche Intelligenz auseinander. So schrieb das BaFin Journal (Juni 2018)[1]:

„Aus Perspektive des Marktes zeigt die Studie, dass Big Data und künstliche Intelligenz sowohl bestehenden als auch potenziell neuen Marktteilnehmern erhebliche Wettbewerbschancen bieten."

Dazu betont BaFin-Präsident Felix Hufeld: „Der Innovationswettlauf um Finanzdaten hat längst begonnen." Und er fügt hinzu: „Die Ergebnisse zeigen deutlich, wie wichtig es ist, dass wir uns aufsichtlich und regulatorisch mit diesen Themen befassen."

Somit steht die Notwendigkeit, sich mit den Thematiken auseinanderzusetzen außer Frage. Diejenigen Marktteilnehmer, die diese Aufgabe erfolgreich meistern, werden langfristig einen Vorteil gegenüber ihren Mitbewerbern haben.

ANWENDUNGSGEBIETE

Schon einige ausgewählte Anwendungsfälle aus der Praxis zeigen die Möglichkeiten, die künstliche Intelligenz bietet.

Zum Beispiel werden in der Betrugserkennung (fraud detection) Algorithmen auf Basis des maschinellen Lernens in die Lage versetzt, Betrugsmuster zu erkennen, um der Bank eine frühzeitige Handlungsreaktion zu ermöglichen.

Auch in der Vertriebssteuerung bieten sich mehrere Anwendungsmöglichkeiten. Beispielsweise wird in der Kundensegmentierung maschinelles Lernen dazu verwendet, um Kunden in homogene Segmente einzuteilen. Im Gegensatz zur klassischen Kundensegmentierung müssen die Segmentgrenzen nicht vorgegeben werden, da dies automatisiert durch den Algorithmus erfolgt.
In der Kundenpotenzialanalyse sollen Predictive-Analytics-Modelle auf Basis des bereits bestehenden Datenbestands erkennen, welche Kriterien einen potenziellen Kunden auszeichnen. Über diese Kriterien werden durch den Algorithmus Neukunden identifiziert, bei denen Affinitäten zum Kauf bestimmter Produkte (zum Beispiel Kreditkarten, Fonds etc) bestehen.

Auch in der Datenvisualisierung und Bestandsdatenanalyse bieten sich Einsatzmöglichkeiten, denn viele bankbetriebswirtschaftliche Fragestellungen basieren auf Daten. Doch gerade im Zusammenführen der dafür benötigten Daten aus verschiedenen Teilsystemen ist auch technisches Know-how zu Datenbankabfragen erforderlich, um die Daten effizient zusammenzuführen, auszuwerten und zu visualisieren.

Dieser Anwendungsfall ist definitiv nichts Neues, dennoch nutzen viele Finanzinstitute moderne Auswertungsprogramme zur Visualisierung von komplexen Daten in Dashboards erst in geringem Maße. Hier liefert der Aufbau einer Datenpipeline von

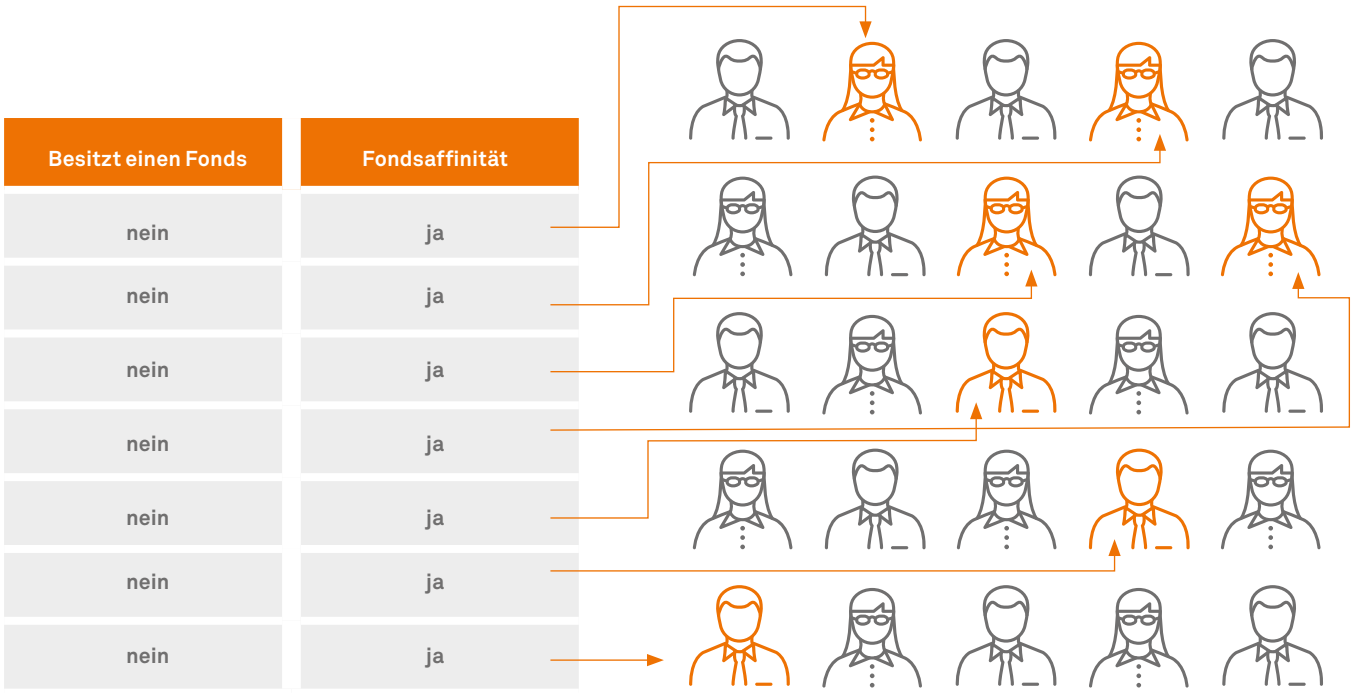


Abbildung 3: Beispiel Provisionsgeschäft: Identifikation von potenziellen Fondskunden durch Algorithmen

1 https://www.bafin.de/SharedDocs/Downloads/DE/BaFinJournal/2018/bj_1806.html

» Neben der eigenen Erstellung in einer Programmierumgebung wie Python oder R stehen auch Standardtools zur Datenanalyse zur Verfügung. «

der Datensammlung bis hin zur Abbildung der komplexen Zusammenhänge in einem Dashboard erhebliche Mehrwerte in der Steuerung.

SO KANN ES GEHEN – ABLAUFBESCHREIBUNG EINES DATA-SCIENCE-PROJEKTES

Der elementare Baustein beim Aufbau eines Data-Science-Projektes ist die Definition der zugrunde liegenden bankbetriebswirtschaftlichen Fragestellung. Von dieser Fragestellung ausgehend wird ermittelt, welche Daten für ihre Beantwortung relevant sind.

Somit hängt die darauffolgende Datenextraktion unmittelbar von der vorangehenden Fragestellung ab. Wenn nicht nur auf interne Daten eines bestandsführenden Systems oder des Datawarehouses, sondern auch auf externe Datenquellen aus dem Internet zugegriffen werden soll, müssen entsprechende Schnittstellen zur Datenabfrage implementiert werden. Diese gesammelten Daten werden dann zusammengefasst (Datenaggregation) und strukturiert in einer Datenbank oder in einem bestimmten Datenformat (.csv, xml etc.) abgelegt. Diese Daten bilden die Basis für alle weiteren Prozessschritte. Im nächsten Schritt werden die gesammelten Daten analysiert (Datenanalyse). Dabei geht es im Wesentlichen um die Beschaffenheit der Daten und deren Eigenschaften. Neben der Berechnung der klassischen statistischen Kennzahlen (Maximalwerte, Mittelwert, Standardabweichung, Korrelation etc.) werden die Daten in Diagrammen grafisch aufbereitet. Gegebenenfalls können dabei schon erste erkennbare Abhängigkeiten visuell dargestellt werden.

Neben der Nutzung in einer Programmierumgebung wie Python oder R stehen auch Standardtools zur Datenanalyse zur Verfügung. Je nach Zielbild kann dies jedoch an manchen Stellen erforderlich sein.

Impliziert die Fragestellung ein Entscheidungsverfahren oder eine Voraussage, so können Modelle des maschinellen Lernens angewendet werden. Dabei werden die Datensätze zuerst vorverarbeitet. Beispielsweise können die Einflussfaktoren (Daten) ohne wesentlichen Informationsverlust reduziert werden, um die Modellkomplexität zu verringern. Anschließend werden die Datenbestände in Trainings- und Testdaten zerlegt.

Für den nächsten Schritt – die Auswahl eines passenden Modells – ist eine gewisse Erfahrung erforderlich. Je nach Fragestellung liegt zumeist entweder ein Clusterungs- (unüberwachtes Lernen) oder Klassifizierungsproblem/Regression (überwachtes Lernen) vor.

Dieses Modell wird mit den Trainingsdaten trainiert. Diese Phase wird als maschinelles Lernen bezeichnet. Nun ist das Modell in der Lage, mit den Testdaten validiert zu werden. Dabei sollten die klassischen Metriken – Korrektklassifizierungsrate, Genauigkeit und Sensitivität – betrachtet werden. Bei unzureichenden Ergebnissen sollte der Prozess ab dem Schritt der Vorverarbeitung wiederholt werden, um anschließend eine bessere Modellperformance zu erreichen. Liefert das Modell hingegen zufriedenstellende Ergebnisse, kann das Modell produktiv verwendet werden.

Ein Praxisbeispiel ist die gezielte Identifizierung von Potenzialkunden. So generieren viele Kreditinstitute Erträge durch Provisionserlöse im Fondsgeschäft. Mithilfe des maschinellen Lernens suchen Algorithmen nach Mustern in den Bestandsdaten. Dabei wird zunächst analysiert, welche Merkmalsausprägungen (zum Beispiel Einkommen, Alter etc.) Fondsbesitzer vorweisen. Anschließend wird in den Bestands- und Neukundendaten nach diesen Mustern gesucht. Diejenigen Kunden, die noch keine Fonds besitzen, aber die typischen Kennzeichen aufweisen, identifiziert der Algorithmus als Potenzialkunden.

Auf Basis dieser Berechnungen werden somit die Kunden identifiziert, bei denen entsprechende Vertriebsmaßnahmen vielversprechend sind. ■

Ansprechpartner:

Markus Hausmann
Business Consultant
markus.hausmann@msg-gillardon.de