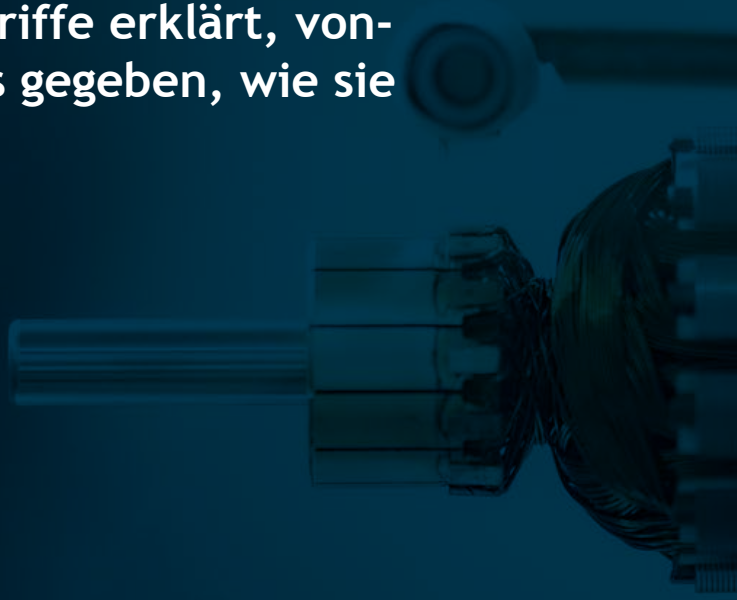
DATA SCIENCE & MACHINE LEARNING

WORUM GEHT ES DABEI EIGENTLICH?

Stand 08.2020

BIG DATA, DATA SCIENCE, MACHINE LEARNING

Was verbirgt sich hinter diesen Begriffen? In diesem Artikel werden die Begriffe erklärt, voneinander abgegrenzt und Tipps gegeben, wie sie eingesetzt werden können.



“Daten sind das neue Öl”

Dieser Spruch wird häufig zitiert, wenn es um Zukunftsthemen in der IT geht. Konkret geht es dabei für Unternehmen meist um die Frage, wie sie aus der Vielzahl der zur Verfügung stehenden Daten die richtigen Schlüsse ziehen können. Diese Frage lässt sich nochmals unterteilen in „Wie bekomme ich eine gewünschte Information aus den Daten“ und „Welche Informationen sind sonst noch in den Daten versteckt, von denen ich möglicherweise noch nichts weiß, die mir aber trotzdem enorme Wettbewerbsvorteile bringen würden, wenn sie mir denn bekannt wären?“. Die zur Verfügung stehenden Daten wiederum lassen sich unterteilen in Unternehmens-interne Daten sowie öffentlich zugängliche Daten, mit denen beispielsweise Lücken in den Unternehmens-internen Daten aufgefüllt werden können oder aus denen wertvolle Zusatzinformationen gezogen werden können.

Ein Beispiel für eine Fragestellung, bei denen Methoden aus den Bereichen Data Science, Big Data und Machine Learning zum Einsatz kommen können: Ein Unternehmen möchte die Gründe für Kündigungen von Kunden seiner Webseite erfahren. Gesucht sind dann Aussagen dieser Art: „Wenn ein Kunde am Wochenende lange im Teilbereich <X> der Webseite unterwegs war, dann hat er mit einer Wahrscheinlichkeit von <y>% am Montag seine Mitgliedschaft gekündigt“. Datenbasis für eine solche Aussage kann die Logdatei der Webanwendung sein. Aus dieser lassen sich zwar die benötigten technischen Abhängigkeiten erkennen, meist ist dies aber aufgrund der Größe der Logdateien nur mit speziellen Tools aus dem Bereich „Big Data“ möglich. Noch interessanter, gleichzeitig aber auch noch komplexer wird die Aufgabe, wenn man weitere Datenquellen hinzuziehen will, um die Kündigungswahrscheinlichkeit noch besser bestimmen zu können. Beispiel: Welche Informationen sind für den geografischen Bereich verfügbar, aus dem der kündigungswillige Kunde stammt? Ist dies eine Region mit besonders hohen oder niedrigen Haushaltsnettoeinkommen? Gibt es dort aktuell besondere Marketingmaßnahmen von Konkurrenzanbietern? Liegen solche Daten vor und können Sie mit hinreichender Genauigkeit zueinander in Beziehung gesetzt werden, so kann das Unternehmen aussagekräftige Reports über vergangene Auswertungszeiträume erzeugen.

Liegen solche umfassenden Datenauswertungen für vergangene Zeiträume vor, dann kann man mit Data Science-Methoden noch einen Schritt weiter gehen und Vorhersagen über das zukünftige Verhalten von Kunden gewinnen und versuchen, in Realzeit die passenden Maßnahmen daraus abzuleiten. In unserem Beispiel könnte das z.B. eine Maßnahme der folgenden Art sein: „Wenn ein Kunde am Wochenende schon länger als <x> Minuten in einem bestimmten Teilbereich unserer Webseite unterwegs ist, und wir wissen, dass er aus einer Region mit

niedrigem Haushaltsnettoeinkommen stammt, in der unsere Konkurrenz in der letzten Woche umfassende Rabattaktionen gestartet hat, so schicken wir ihm umgehend per Mail ein Freiguthaben zu, bevor er auf den Kündigungsbutton drückt“. Ein Vorhersagesystem, dass solche Vorhersagen in Realzeit auf der Basis von historischen Daten trifft, ist das Ergebnis der Arbeit von Data Scientisten.

Berufsbild Data Scientist

Data Scientist ist ein vergleichsweise neuer Beruf, der bereits im Jahr 2012 von Harvard Business Review als “sexiest job in the 21st century” bezeichnet wurde. Ein Data Scientist benötigt Kenntnisse aus der Informatik und der Mathematik (Statistik), kombiniert mit Anwendungswissen. Diese Kenntnisse kommen zum Tragen, um die richtigen Fragen und dazu passende verfügbare Daten zu finden (Anwendungskenntnisse), die Daten zu analysieren (mathematische Kenntnisse) und die daraus gewonnenen Erkenntnisse in ein produktives System umzusetzen (Informatikkenntnisse).

Im Projektalltag kommen dem Data Scientist unterschiedliche Aufgaben zu. Darunter fallen

- Datensuche (Welche Daten stehen zur Verfügung beziehungsweise welche lassen sich zusätzlich besorgen?),
- Datenbereinigung (Aufbereitung der Daten für die anschließende Analyse),
- Offline-Datenanalyse (Wie lassen sich aus den vorliegenden Daten die gewünschten Informationen extrahieren?) und
- Überführen der Ergebnisse in ein produktives System zur Online-Analyse (s. dazu obiges Beispiel zur Realtime-Vorhersage von Kundenverhalten)
- Methodische Grundlage dieser Ausgaben sind Algorithmen aus dem Bereich Machine Learning.

Machine Learning

Innerhalb von Machine Learning (ML)-Projekten geht es darum, dass eine Maschine (das kann in der Evaluationsphase der Laptop eines Entwicklers sein, im späteren produktiven Einsatz dann aber auch eine Cloud-Plattform mit der benötigten Rechenleistung) durch den Einsatz statistischer Algorithmen auf der Basis von bekannten Daten ein Wissen über „in den Daten verborgene Muster“ aufbaut, um dann später auf der Basis dieses angelernten Wissens Vorhersagen für neue bisher unbekannte Datensätze treffen zu können. Am oben beschriebenen Beispiel „Kündigungen durch gezielte Kundenbindungsmaßnahmen verhindern“ würde also ein Algorithmus durch bekannte Daten aus Logdateien und anderen Quellen angelernt, damit dieser dann später auf der Basis des Wissens aus der Anlernphase in Realzeit Vorhersagen bezüglich des zu erwartenden Verhaltens von Kunden treffen kann.

Die zur Verfügung stehenden ML-Algorithmen lassen sich einteilen in

- Unüberwachtes Lernen: Hier soll der Algorithmus selbständig nach bisher unbekannten Mustern suchen.

- Überwachtes Lernen: Der Algorithmus wird über Muster (Datensätze, bei denen man schon weiß, was das gesuchte Ergebnis ist) trainiert. Diese Datensätze legt man demAlgorithmus als Muster vor, so dass der Algorithmus daraus lernen kann. Anschließend muss der Algorithmus das zuvor Gelernte dann verwenden, um die Muster auch in bisher unbekannten Datensätzen zu finden.

Bereits an diesen Definitionen erkennt man, dass die Menge und die Qualität der Daten, die für das Training der entsprechenden Algorithmen erforderlich sind, neben der benötigten Hardware-Rechenleistung, die z.B. über Cloud-basierte Lösungen zusammengestellt werden kann, zentrale Punkte für den Erfolg von ML-Projekten sind.

Unterstützung erhalten ML-Projekte durch eine Reihe von – meist frei verfügbaren - Tools: Die Programmiersprache Python ist eine der in Machine Learning-Projekten am häufigsten eingesetzten Sprachen. Dementsprechend hoch ist auch die Zahl der Bibliotheken, mit denen die oben beschriebenen Schritte bei der Analyse von Daten sowie dem Anlernen und dem anschließenden produktiven Einsatz von Machine Learning-Algorithmen in der Praxis durchgeführt werden. Beispiele für solche Tools, in denen u.a. diverse statistische Methoden bereits implementiert sind, sind Scikit Learn und TensorFlow. Diese Tools nehmen den ML-Projektmitarbeitern zunächst eine Menge Implementierungsarbeit ab. Umso mehr Zeit kann und muss dann in den Teil der Projekte gesteckt werden, der am Ende über den Erfolg der Projekte entscheidet, nämlich die Aufbereitung der benötigten Daten, das Anlernen der Algorithmen mit diesen Daten sowie am Ende ein produktiver Einsatz der angelernten Modelle in einer mit ausreichend Ressourcen ausgestatteten IT-Umgebung.

ÜBER DEN AUTOR

Rudolf Jansen ist Diplom-Informatiker aus Aachen und arbeitet als freiberuflicher Softwareentwickler und Autor. Er schreibt über alle IT-Themen und unterstützt seine Kunden bei Projekteinsätzen als Business Analyst und Requirements Engineer.