DATABOOSTER

Deep Dive in ihre Daten

» Innovationen sind für das Überleben von Unternehmen notwendig. Die von der Innosuisse geförderten NTN Innovationbooster verfolgen das Ziel, durch geeignete Kombination von Wirtschaft und Wissenschaft radikale Innovationen umzusetzen. Der Databooster bietet die Möglichkeit, mit einem Deep Dive die Machbarkeit einer Idee anhand eigener Daten zu prüfen.

Autoren: Philipp Schmid und Patricia Deflorin

Viele haben in der eigenen Firma bereits angefangen, Daten zu sammeln. Häufig liegen viele Excel Tabellen mit Tausenden von Einträgen auf deren Server und sie haben Mühe, daraus neue Erkenntnisse zu ziehen. Dies ist die ideale Voraussetzung für einen Deep Dive in diese Daten. Der NTN Innovationbooster *Databooster* unterstützt darin, die richtigen Experten zu finden, und finanziert eine vertiefte Analyse der Daten.

DATA-SCIENCE-PYRAMIDE – ES GIBT KEINE ABKÜRZUNGEN

Die Basis für eine erfolgreiche Datenauswertung beginnt bei den Rohdaten selbst, die aus einer Vielzahl von Quellen stammen können, in verschiedenen Formaten vorliegen und in riesigen Mengen vorhanden sind (siehe Abbildung 1). Eine Schicht höher, beim Data Engineering, wird der Kontext und die Struktur hergestellt, welche erforderlich sind, damit Daten zu Informationen werden. Master Data Management und Governance bauen auf dieser Prämisse auf, um die Qualität sicherzustellen, bevor die Daten in die letzten beiden Phasen übergehen. Reporting und Business Intelligence stellen den Beginn der Erkenntnisgewinnung dar, dabei werden Daten visualisiert und Informationen aggregiert. Data Science schliesslich stellt den Höhepunkt der Datenumsetzung dar und baut oft auf neuronalen Netzwerken auf, welche mit leistungsstarken statistischen Ansätzen kombiniert werden.

Bild: Sense Corp

Abbildung 1: Data-Science-Pyramide: das Fundament bilden die Daten, aber jede Schicht ist wichtig.

Die Data-Science-Pyramide ist ein Indikator für das Wertpotenzial; wenn ein Unternehmen nicht bereits eine solide Datengrundlage geschaffen hat, ist es in den meisten Fällen nicht ratsam, die Ebenen zu überspringen. In den letzten Jahren wurde oft versucht das Fundament, die Rohdaten, direkt mit künstlicher Intelligenz zu prozessieren. Dieser Ansatz ist meist erfolglos – es gibt leider keine Abkürzung durch die Datenpyramide. Ein datenzentriertes Modell ist in der Regel nur so gut wie die Daten, mit denen es trainiert wird, was wiederum das direkte Ergebnis aller Schichten der Pyramide darstellt. Es ist ratsam sich mit jeder der einzelnen Schichten vertieft auseinanderzusetzen.

Wie sieht nun aber ein Databooster Deep Dive im Detail aus?

PHASE #1: DATA-WRANGLING – DIE DATEN ZÄHMEN

Etwa 80 Prozent der Arbeit in Datenprojekten geht auf das Konto der Datenaufbereitung, auch Data-Wrangling genannt. Wie im wilden Westen muss der IT-Cowboy versuchen die verstreuten Daten einzusammeln, zu ordnen und für die weitere Verwendung vorzubereiten. Meist liegen die Daten in vielen Excel-Tabellen verstreut. Es gibt fehlende und falsche Einträge, oft fehlt die physikalische Einheit, die Skala zur Normierung oder eine eindeutige Bezeichnung. Ärgerlich sind auch verschiedene Zeitstempel, welche zueinander einen Versatz haben, weil sie nicht vom gleichen Taktgeber stammen. Damit die Daten effizient und automatisch weiterverwendet werden können, braucht es einen aufwendigen Prozess der Bereinigung, Anreicherung, Validierung und Umwandlung in ein Format, welches die spätere Prozesspipeline vereinfacht. In dieser Phase entstehen noch keine Erkenntnisse, sie bildet aber die notwendige Basis für die weiteren Schritte. Es ist ratsam die Daten in eine geeignete Datenbankstruktur zu integrieren. Speziell für die Arbeit mit reinen Zeitreihendaten optimiert, bieten Zeitreihendatenbanken, etwa der Marktführer InfluxDB (influxdata.com), effiziente und skalierbare Lösungen. Bei komplexeren Strukturen helfen offene hierarchische Architekturen wie HDF5 (hdfgroup.org/solutions/hdf5/).

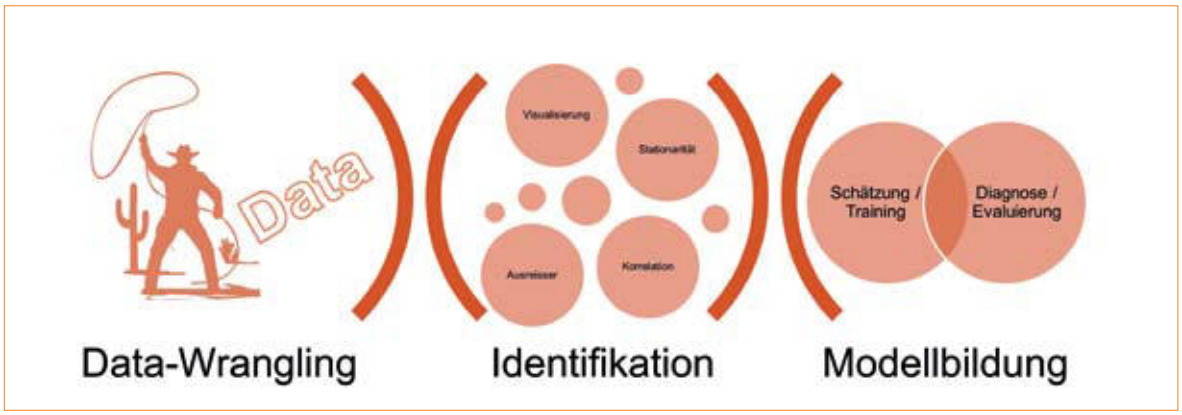

Abbildung 2: Schritte im Deep Dive Prozess.

PHASE #2: IDENTIFIKATION – WAS VERBIRGT SICH IN DEN DATEN?

In dieser Phase soll ein geeigneter Ansatz für die Modellbildung identifiziert werden. Dazu müssen die Daten zuerst visualisiert werden. Anders als bei Bilddaten fällt eine Beurteilung von rohen Zeitreihen dem Menschen ohne Hilfswerkzeuge deutlich schwieriger. Dank der Nutzung professioneller Open-Source-Anwendungen, wie zum Beispiel Grafana (grafana.com), gelingen hochdynamische Visualisierungen in einer intuitiven Benutzeroberfläche. Auf diese Weise wird eine interaktive Arbeitsebene zwischen Fach- und Datenspezialisten geschaffen.

Diese ist nicht zu unterschätzen: Anhand von statistischen Analysen lassen sich Schlüsse über das Vorliegen von Trends, Saisonalitäten und Ausreissern ziehen. Dem Datenspezialisten gelingt es, durch die Wahl geeigneter Darstellungsformen, versteckte Informationen zu Datenverteilungen und in den Daten enthaltene Frequenzen sichtbar zu machen. Nur die Fachspezialisten der Firma haben das notwendige Wissen und Erfahrung, um die visualisierten Daten interpretieren zu können.

In der Schnittmenge dieser Kompetenzen entstehen erfolgsversprechende Konzepte, die suggerieren, mit welchen mathematischen Verfahren die Daten verarbeitet werden müssen. Oft zeigen sich in der Praxis bereits in dieser Phase die ersten wichtigen Aha-Momente, wenn die Fachexperten ihre Intuitionen visualisiert am Bildschirm analysieren können.

PHASE #3: MODELLBILDUNG – ERKENNTNIS SCHAFFEN

Die Modellbildung besteht aus zwei wichtigen Teilphasen: Training und Evaluierung. In kurzen Iterationszyklen versuchen die Datenspezialisten ein möglichst gutes mathematisches Modell zu entwickeln. Aufwind erleben dabei Verfahren aus dem Bereich der künstlichen Intelligenz. Zuerst muss der bestehende Datensatz aufgeteilt werden. Der Trainingsdatensatz wird für das Lernen und Optimieren des Modells, beispielsweise einem neuronalen Netzwerk, verwendet. Der Validierungsdatensatz dient als Referenz für die Evaluierung der Performance des Modelles. Die Modellparameter und -koeffizienten werden mit Hilfe unterschiedlicher Techniken geschätzt und immer weiter verbessert. Die Erkenntnisse der Identifikationsphase sind dabei enorm wichtig und definieren die besten Vorverarbeitungsschritte. Sobald erste Resultate vorliegen, können weitere interessante Analysen durchgeführt werden. In vielen Fällen wird mit einem hochwertigen Datensatz, dem Golden Sample, gestartet: sehr viele Sensoren kombiniert mit hoher Abtastrate. Für die Firma ist es nun sehr spannend zu verstehen, welche Sensoren wirklich relevant sind, welche miteinander korrelieren und wie das Verhältnis zwischen Performance des Modelles und Abtastfrequenz zusammenhängt. Durch automatisierte Tests und starke Rechenleistung können solche Analysen teilautomatisiert berechnet werden. Dadurch entsteht bereits in kurzer Zeit ein sehr guter Überblick über die Daten, den Prozess und das mögliche Potenzial.

WIE WEITER?

Am Ende einer Machbarkeitsanalyse erhält die Firma Visualisierungen, einen Erkenntnisbericht und eine Roadmap für die nächsten Schritte. Der Databooster Deep Dive ist ein schlankes, aber sehr mächtiges Instrument, um die ersten wichtigen Schritte auf dem Weg zur optimierten Prozesspipeline in Angriff zu nehmen. Es entstehen viele Erkenntnisse in kürzester Zeit, ideal um Unklarheiten zu beseitigen oder Risiken zu minimieren. Die Fachspezialisten gewinnen vertiefte Einblicke in ihre Maschinen oder Prozesse und die Datenspezialisten bauen in kurzer Zeit ein zwingend notwendiges Fachwissen auf. Beste Voraussetzungen, um die Innovationsidee in einem geförderten Innosuisse Projekt weiterzuverfolgen.

Wer Daten hat und neue Innovationsgelegenheiten sucht, um das eigene Unternehmen oder die eigene Organisation einen grossen Schritt weiterzubringen, und wer Partner sucht, die in diesen Innovationsvorhaben unterstützen, sollte auf den Databooster zugehen: databooster.ch. «

Philipp Schmid
Head Research and Business Development Industry 4.0 & Machine Learning am CSEM (Schweizer Forschungs- und Entwicklungszentrum mit Schwerpunkt Mikrotechnologie, Digitalisierung und Energie).

Prof. Dr. Patricia Deflorin
Dozentin für Innovationsmanagement, Forschungsleiterin Schweizerisches Institut für Entrepreneurship, Fachhochschule Graubünden.