Kommentar von Dr. Theo Steininger, Erium

So löst die Astrophysik Big-Data-Probleme

Autor / Redakteur: Dr. Theo Steininger / Nico Litzel

Unstrukturierte Rohdaten zu analysieren und so aufzubereiten, dass sie eine wertvolle und verlässliche Basis für wichtige Unternehmensentscheidungen liefern, bedarf einer großen Vorleistung: Sind die Daten fehlerhaft, weisen Lücken auf oder sind in ihrem Umfang sehr eingeschränkt, kommen auch Datenwissenschaftler schnell an ihre Grenzen. Hinzu kommt, dass Unternehmen für einzelne Anwendungsfälle noch nicht genug Daten zur Verfügung haben. Dennoch können Unternehmen von zuverlässigen Prognosen profitieren. Inspiration hierzu liefert die Astrophysik mit Methoden zur Erforschung des Kosmos.

Der Autor: Dr. Theo Steininger ist CEO von Erium
(Bild: Nadine Rupp)

Wenn es beispielsweise um die Erforschung der 3D-Struktur der Milchstraße geht, haben Astrophysiker folgendes Problem: Es gibt nur eine einzige Perspektive, mit der man von der Erde oder einem Satelliten aus ins All blicken kann. Dieser zweidimensionale Blickwinkel erschwert das Erzeugen eines 3D-Abbildes der verschiedenen Bestandteile der Milchstraße. Zudem lässt sich die Auflösung der Teleskope aufgrund von physikalischen Grenzen nicht beliebig weit erhöhen.

Astrophysiker haben allerdings den Vorteil, dass es in der Milchstraße die verschiedensten Sterntypen gibt, deren Lichtspektrum – sozusagen der Farb-Fingerabdruck – man gut verstanden hat. Wenn nun das Licht dieser Sterne beispielsweise durch interstellaren Staub reist, wird der Fingerabdruck von diesem verändert. Aus diesen Veränderungen lassen sich Rückschlüsse auf die Staubart und -dichte zwischen der Erde und den Sternen ziehen.

Das funktioniert so: Mithilfe eines physikalischen Machine-Learning-Modells kann man für eine beliebige Kombination aus (hypothetischen) Dichteverläufen und Messdaten eine Wahrscheinlichkeit berechnen, dass diese zusammenpassen. Mit echten Beobachtungsdaten lässt sich damit also beantworten, wie plausibel beispielsweise verschiedene Annahmen über die Staubdichte in einem bestimmten Teil der Milchstraße sind. Es lässt sich aber auch jene (milchstraßenweite) Staubverteilung berechnen, die die Beobachtungsdaten insgesamt am wahrscheinlichsten verursacht hat.

Wenige Daten in Wissen verwandeln

Es geht also darum, die wenigen vorhanden Daten, die zur Verfügung stehen, über ein statistisches Modell mit dem Wissen von Experten, in diesem Fall Astrophysiker, zu verknüpfen, nämlich: „Wie hat das Sternenlicht ursprünglich ausgesehen?“ und „Wie wird Licht verändert, wenn es interstellaren Staub durchdringt?“. Auf diese Weise entsteht eine statistische Analysemethodik, die mit so wenigen Daten wie möglich auskommt. Das Spannende dabei: Diese Methode lässt sich auf jedes andere beliebige Szenario unabhängig von bestimmten Branchen anwenden. Das ist auch notwendig, denn gerade in der Praxis ist in vielen Unternehmen die tatsächlich verwertbare Datenmenge sehr gering.

Das liegt zum Teil an den hohen Kosten der Datenerzeugung oder daran, dass sich Prozesse undokumentiert über die Zeit verändern oder verändert werden. Wer diese Daten dann in einen Analysetopf wirft, arbeitet genauso präzise wie jemand, der Äpfel mit Birnen vergleicht. Da die wirtschaftlich zu erlangende Datenmenge sehr begrenzt ist, ist man vom Wissensstand her gesehen, auf einer ähnlichen Ebene wie in der Astrophysik.

Optimale Positionierung von Anbauteilen bei BMW

Ein Anwendungsbeispiel für ein solches statistisches Machine-Learning-Modell liefert der bayerische Automobilhersteller aus München. Jede einzelne Komponente der verschiedenen Fahrzeugmodelle muss perfekt positioniert sein – das gilt natürlich auch für die Fahrzeugtüren. Optimale Spaltmaße sind dabei sowohl für die Funktionalität als auch das äußere Erscheinungsbild Pflicht.

Die Montage der Türen bedeutete jedoch eine große Herausforderung, da zum Zeitpunkt des Einbaus weder Tür noch Karosserie lackiert sind und Ausstattung, Fensterscheiben und notwendige Dichtungen fehlen. Der Einfluss dieser Faktoren auf die Türposition – beispielsweise durch Verformung und zusätzliches Gewicht – muss dementsprechend antizipiert und kompensiert werden. Ein äußerst schwieriges Unterfangen, bei dem sich die richtige Antwort auch noch stetig verändert, weshalb die Türen in der Praxis nach Abschluss der Montage noch einmal in aufwendiger und zeitintensiver Nacharbeit von Hand angepasst wurden. Die verantwortlichen Teams wollten jedoch genau und in Echtzeit berechnen können, mit welcher Einbauposition sich die besten Spalt- und Versatzqualitäten erreichen lassen. Aufgrund der stetigen Prozessveränderungen fehlte es allerdings auch hier an Big Data <https://www.bigdata-insider.de/was-ist-big-data-a-562440/> , übliche Deep-Learning-Ansätze kamen nicht in Frage.

Nun kennt BMW seine Autos allerdings mindestens genauso gut wie Astrophysiker die Sterne der Milchstraße. Und so formulierte BMW ein kausales Prozessmodell: Als Informationsbasis dienten der Ablauf des Produktionsprozesses sowie Messdaten der tatsächlichen Spaltmaße, CAD-Koordinaten und Stammdaten wie die Fugenpläne der jeweiligen Baureihen, um die Frage nach relevanten Einflussfaktoren und Parametern zu beantworten. Zusätzlich wurde das Wissen der Prozessexperten eingebunden, beispielsweise Informationen über das typische Drehverhalten durch zusätzliches Gewicht oder Verformungen durch die Dichtung. Die Analogie: So wie Sternenlicht durch den interstellaren Staub fliegt und von diesem verändert wird, fahren die Fahrzeugkarosserien mit ihren Spalt- und Versatzmaßen durch die Fertigung, welche diese ebenfalls verändert.

BMW konnte auf diese Weise ermitteln, an welchen Stellen die Stammdaten und die Spezifikation des Prozesses noch nicht in Einklang mit den Daten standen. So wurden keine Daten gesammelt, die später aufgrund einer fehlerhaften Dokumentation doch nicht verwendet werden konnten. Im nächsten Schritt trainierte der Automobilhersteller die Modellparameter anhand der bestehenden Daten. Das Modell konnte auf diese Weise vorhersagen, wie genau eine Fahrzeugtür im Laufe der Fertigung verdreht oder verformt wird und welche finale Position sich daraus ergibt. Dabei ließ sich auch kontinuierlich die Wiederholgenauigkeit des Prozesses überwachen. BMW konnte zudem berechnen, welche Einstellungsparameter notwendig sind, um die Türen optimal einzupassen. Mit jedem neuen Fahrzeug werden nun die Modellparameter nachtrainiert und die optimalen Einstellungsparameter neu ermittelt.

Machine Learning ist nicht gleich Machine Learning

Gängige Lösungen stoßen aufgrund von komplexen Szenarien, schlechter Datenqualität <https://www.bigdata-insider.de/was-ist-data-quality-a-649900/> oder fehlender Kollaborationsmöglichkeiten schnell an ihre Grenzen. Data Scientists <https://www.bigdata-insider.de/was-ist-ein-data-scientist-a-600907/> sollten daher bei der Wahl auf folgende Kriterien setzen:

Smarte Daten: Data Scientists müssen mit ihrer Software Fachwissen in Form von statistischen Modellen abbilden können, damit Analysen eben auch mit den kleinsten Datensätzen auskommen. Falls notwendig, schließt eine geeignete Software Lücken im Datensatz automatisch mit konsistenten Schätzwerten oder generiert Beispieldaten, wenn keine Daten vorhanden sind. So lassen sich trotzdem theoretische Szenarien durchspielen und Machbarkeitsstudien durchführen. Data Scientists müssen außerdem stets darauf verlassen können, wie es um die Wahrscheinlichkeit ihrer Ergebnisse bestellt ist. Nur so können sie schon vorab beurteilen, welche Daten überhaupt erhoben werden sollten. Ein weiterer Pluspunkt ist die Modellerstellung mit physikalischen Parametern, die sie individuell bestimmen können und deren Analysen bereits die ersten Erkenntnisse bringen. Wenn ein Prozess zu kompliziert ist, sollte es möglich sein, diesen für eine höhere Übersichtlichkeit in mehrere Teilprozesse aufzugliedern. Ist es dann noch möglich, dem Modell eine kausale Struktur zu geben, gehören auch Scheinkorrelationen der Vergangenheit an.

- **Transparente Analysen**: Eine geeignete Software-Lösung setzt auf Whiteboxing. Auf diese Weise behalten Data Scientists jeder Zeit den Überblick über alle Inputs und Outputs sowie jede Zwischenstufe und jeden Teilprozess im Modell. Diese sollten sich dazu gesondert trainieren und mit zusätzlichen Daten anreichern lassen. Eine passende Lösung ist zudem in der Lage, die Genauigkeit der Ergebnisse zu ermitteln bzw. wie sich die Unsicherheit des Inputs auf die Genauigkeit der Prognosen auswirkt. Idealerweise berechnet die Lösung dabei nicht nur Standardabweichungsintervalle, sondern gleich die gesamte Wahrscheinlichkeitsverteilung der Ergebnisse, denn: In der Praxis ist nicht immer alles normalverteilt. Viele Machine-Learning-Lösungen scheitern an dieser Stelle. Ebenfalls wichtig: Die Lösung sollte Maßnahmen aufweisen, um Overfitting zu erkennen und beispielsweise durch richtig gesetzte Prior-Wahrscheinlichkeiten zu vermeiden.
- **Effiziente Teamarbeit**: Wenn ein Projekt dringend abgeschlossen werden muss und der Zeitdruck groß ist, ist es mit gängigen Machine-Learning-

> Lösungen häufig nicht praktikabel, spontan einen weiteren Kollegen dazuzuholen oder bereits bestehende Analyseergebnisse mit einzubauen. Der Einarbeitungsaufwand ist schlichtweg zu groß. Dementsprechend sollten Data Scientists bei der Wahl ihrer Lösung darauf achten, dass sie über umfassende Kollaborations-, Reporting- und Darstellungsfunktionen sowie Schnittstellen zu etablierten Werkzeugen wie beispielsweise Microsoft Power BI verfügt.

So können nicht nur mehrere Kollegen gleichzeitig und ohne Abhängigkeiten an einem Modell arbeiten, alle Analysen, Zwischenergebnisse und Prognosen lassen sich noch dazu anschaulich darstellen und visualisieren. Ein großes Plus, wenn es darum geht, die eigene Arbeit für andere Abteilungen oder den Vorstand verständlich aufzubereiten. Eine geeignete Lösung speichert dabei auch bereits vorhandenes und Expertenwissen und ist dank standardisierter Formate zu jeder Zeit und an jedem Ort für das gesamte Team zugänglich.

Fazit

Machine Learning und Big Data sind nicht immer die ultimative Kombination, wenn es darum geht, Prozesse zu optimieren und wertvolle Erkenntnisse für das eigene Unternehmen zu gewinnen. Es braucht stattdessen eine geeignete Machine-Learning-Lösung, smarte Daten und manchmal auch den Blick nach oben zu den Sternen.