Korrelationen und Kausalitäten: Falsche Schlüsse und ihre Folgen

„Wer Korrelationen wie Kausalitäten behandelt, darf sich über negative Konsequenzen nicht wundern. Zugegeben, die Grenzen zwischen beiden sind fließend. Um so sorgsamer sollte damit umgegangen werden. Andernfalls kann es teuer werden“, ist Björn Heinen von Inform Datalab überzeugt. In einem Kommentar zeigt er auf, wie die beiden Wechselwirkungen, wenn sie sauber getrennt und klug eingesetzt werden, analytisch für Prozessoptimierungen und Handlungsempfehlungen genutzt werden können.

Bild: Inform

Björn Heinen ist Senior Data Scientist im Inform Datalab des Aachener Optimierungsspezialisten Inform

„Wie schnell man Korrelation und Kausalität verwechseln und daraus irreführende Fehlschlüsse ziehen kann, zeigt ein plakatives Beispiel aus den USA. Dort grassierten 2020 zwei Landkarten mit Diagrammen, die einen Zusammenhang zwischen der Zahl der registrierten Corona-Fälle und der Zahl der installierten 5G-Masten suggerieren. Die Korrelation zwischen beiden Werten ist offensichtlich. Daraus jedoch die Kausalität ableiten zu wollen, 5G-Masten beförderten Corona-Ausbrüche, wäre ein logischer Kurzschluss.

Der Grund für die Ähnlichkeit ist die schlichte Tatsache, dass dort, wo die größte Bevölkerungsdichte in den USA herrscht, nicht nur die Infektionsraten am höchsten waren, sondern eben auch die Anzahl der Handy-Nutzer. Kurz gesagt: Wo es viele 5G-Masten gibt, waren auch viele Menschen infiziert. Die Korrelationen sähen ähnlich aus, wenn man statt der 5G-Übersicht ein Diagramm der Verkehrsunfallzahlen oder der Kriminalitätsstatistik in den USA nähme. Auch daraus ließe sich keine direkte kausale Beziehung zwischen Verkehrsdichte, respektive Verbrechensrate, und der Häufigkeit von Corona-Fällen ableiten.

Wertschöpfende Unterscheidung von Korrelation und Kausalität

Für Unternehmen ist es wichtig, solche Verwechslungen zwischen Korrelation und Kausalität zu vermeiden. Nur so können sie beide sinnvoll für sich nutzen. Eine Korrelation ist zunächst einmal nichts weiter, als eine auffällige Übereinstimmung statistischer Größen. Kausalität dagegen liegt erst dann vor, wenn

Über das Inform Datalab

Der Aachener Optimierungsspezialist Inform gründete im November 2019 das Inform Datalab zur Entwicklung innovativer Lösungen für das Management und zur Analyse von Daten. Seitdem werden dort alle Datenprojekte gebündelt. Ziel ist es, Unternehmen bei ihrer digitalen Transformation zu unterstützen. Dazu entwickelt ein spezialisiertes Team aus Data Scientists, Softwareentwicklern und Consultants maßgeschneiderte Lösungen für jede Branche und entwirft eine globale Datenstrategie, die Unternehmen auf dem Weg zur „data-driven Company" begleitet.

Das Portfolio des Inform Datalab umfasst das gesamte Data Management zur gezielten Aufbereitung der Unternehmensdaten sowie sinnvolle Datenanalyse-Methoden. Zudem entwickelt das Lab auch Data-Science-Anwendungen auf Basis von Machine Learning und Künstlicher Intelligenz.

eine direkte Beziehung von Ursache und Wirkung besteht. Bei der unternehmerischen Nutzbarmachung von Korrelationen und Kausalitäten gilt es daher, im Kontext von Analytics und Data Science die Frage zu klären: Wollen wir bestimmte Zusammenhänge in Daten verstehen, oder wollen wir sie nur zu einem konkreten Zweck nutzen?

Viele Fragen im industriellen Kontext lassen sich durch Data Science und Machine Learning beantworten, ohne dass dabei explizit kausale Zusammenhänge identifiziert werden. Möchte etwa der Einkauf vorhersagen, wie lange der Lieferant wirklich braucht, bis er die Ware liefert oder möchte der Produktionsleiter im Vorfeld wissen, wie lange die Endmontage dauert, so ist oft nicht ausschlaggebend, woran das im Detail liegt, also welche Kausalketten dafür existieren. Zur Beantwortung dieser Fragen sind präzise Zahlen, ein niedrigerer Lagerbestand und höhere Termintreue meist viel wichtiger.

Für andere Anwendungsszenarien hingegen ist die Identifizierung von kausalen Zusammenhängen essenziell. So steigen beispielsweise die Kosten durch Ausschuss und Nacharbeit oft dann signifikant, wenn es einen Ressourcen-Engpass in der Montage gibt. Hier gibt es zwar keine direkte kausale Beziehung, aber eine gemeinsame kausale Ursache: eine sehr hohe Auftragslage. In dieser Situation wird die Montage zum Engpass, da Parallelisierung und Vorarbeiten hier schlechter möglich sind als an anderen Stellen. Bei Qualitätsproblemen muss daher immer zuerst die Frage nach den Ursachen gestellt werden. Und diese können ganz profan sein: Teile werden nicht im Lager, sondern auf dem Hof gelagert. Ein Lieferant liefert Material mit schwankender Güte. Eine Maschine steht nah an einem häufig geöffneten Tor, wodurch sich das verarbeitete Material verzieht. Oder es wurde wegen voller Auftragsbücher die 60 Jahre alte Fräse genutzt, anstatt der ausgebuchten, modernen Geräte. In all diesen Fällen geht es darum, die sogenannte „Root Cause" zu finden, also die auslösende kausale Ursache.

Methodische Ansätze zur Kausalitätsfindung

Die richtige Handhabung von Maschinen ist generell ein entscheidender Faktor bei der Kostenkalkulation von Fertigungsprozessen – und unterliegt eben auch seinen Kausalitäten. Fällt bei einer Galvanikanlage beispielsweise übermäßig viel Ausschuss an, dann versuchen Unternehmen in der Regel so gut wie möglich dafür Sorge zu tragen, dass alle Maschinenparameter im Soll-Bereich bleiben. Entsteht dennoch Ausschuss, stoßen sie mit diesem Ansatz schnell an ihre Grenzen. Eine genauere Betrachtung der Historie aller Teile, die die Galvanikanlage durchlaufen haben, kann hier Abhilfe schaffen: Setzen die Anwender diese Historie und deren Qualitätsinformationen in Relation zu den vorhanden Maschinendaten, besteht eine gute Chance, die echte Ursache für den Ausschuss zu finden. So können nicht nur die Temperaturen oder chemischen Konzentrationen in einzelnen Bädern Fehlerquellen sein. Es wäre auch möglich, dass einfach einer der Rüttelmotoren nicht richtig gelaufen ist und sich so während des Bads kleine Luftbläschen auf den Werkstücken bildeten. Damit wäre die tatsächliche Ursache für viele Qualitätsprobleme gefunden, die mit den Vorgabewerten des Herstellers gar nichts zu tun hat.

Die jeweilige analytische Vorgehensweise muss sich an der Zielsetzung des konkreten Projekts orientieren. Sollen Zusammenhänge lediglich beobachtet, und für die Vorhersage genutzt werden, reichen gut gewählte Trainings- und Testdatensätze. Der Testdatensatz hat dann hauptsächlich die Funktion, sicherzustellen, dass der Algorithmus keine zufälligen Zusammenhänge im Trainingsdatensatz beobachtet hat, die es ausschließlich dort gibt. Sollen aber kausale Zusammenhänge identifiziert werden, ist ein hypothesengestützter Ansatz meist unumgänglich. Echter Fortschritt bei der Prozess- und Kostenoptimierung wird nur durch methodisch saubere Analysen und die treffsichere Differenzierung von Korrelationen und Kausalitäten erreicht.

Wie sich das dann in der Praxis niederschlägt, verdeutlichen zwei Beispiele: Zum einen konnte ein Unternehmen aus dem Maschinen- und Anlagenbau durch die Einführung von Machine Learning seine Wiederbeschaffungszeiten um 42 % verbessern, was zu wesentlich zuverlässigeren Prognosen in der gesamten Wertschöpfungskette führte. Zum anderen wurde bei einem Kunden aus dem Bereich Fashion-Retail der Data-Management-Prozess von Unternehmens- und Online-Marketing-Kennzahlen von über 48 h auf weniger als 2 h reduziert."

www.inform-datalab.de