

Buchrezension



Blum, Lothar B. (Hrsg.): Angewandte Data Science – Projekte, Methoden, Prozesse
Verlag: Springer Vieweg, erschienen 27. Juni 2023
280 + XXXI Seiten
Softcover: € 32,99 (ISBN 978-3-658-39624-4)
eBook: € 24,99 (ISBN 978-3-658-39625-1)

Wirtschaftsinformatik & Management
 2023 • 15 (6): 431–433
<https://doi.org/10.1365/s35764-023-00497-9>
 Angenommen: 16. Oktober 2023
 Online publiziert: 17. Januar 2024
 © The Author(s) 2023

Das Thema Data Science war vor wenigen Jahren noch ein großer Hype, der aber inzwischen von Artificial Intelligence (AI) oder in deutsch Künstliche Intelligenz (KI) fast abgelöst wurde. Insofern stellt sich zunächst die Frage, warum ein Buch über Data Science, wenn doch KI das aktuellere Thema ist. Diese vielleicht etwas provokativ anmutenden Frage hat einen ernsten Hintergrund. Der Data-Science-Hype ist vielleicht gar nicht abgebrochen, sondern nur von dem noch größeren KI-Hype überlagert worden. Wenn dem so ist, dann wäre ein Buch über Data Science sehr zeitgerecht, wenn es zumindest Stellung bezieht zum KI-Hype. Das tut das Buch „Angewandte Data Science“, herausgegeben von Lothar Blum, das jüngst im Springer Verlag erschienen ist (<https://link.springer.com/book/10.1007/978-3-658-39625-1>). Der Herausgeber hat zahlreiche Autoren aus sehr unterschiedlichen Bereichen dazu gewinnen können, einen Beitrag zu seinem Werk zu schreiben. Die einzelnen Beiträge befassen sich mit beidem, Data Science und KI. Der feine Unterschied mag aber darin gesehen werden, dass Data Science letztendlich versucht, aus Daten Wert zu schöpfen. Viele KI-Projekte haben derzeit weniger die Wertschöpfung aus Daten im Vordergrund als vielmehr die Grenzen der KI weiter und weiter nach vorn zu treiben. Insofern unterscheiden sich Data Science und KI. Aber, eine Diskussion von ChatGPT fehlt dabei im Buch genauso wenig wie ein Überblick über die Möglichkeiten von (bild-)generierenden Modellen. Das Buch geht noch darüber hinaus und zeigt in zahlreichen spannenden Projektbeschreibungen, wie KI und auch Statistik zum Einsatz kommen können, um Mehrwert zu schaffen.

Das Buch basiert auf einer vom Herausgeber Lothar Blum initiierten Veranstaltungsreihe mit dem Namen *Data Science Darmstadt*. Viele Autoren haben hier vorgetragen und den Vortrag im Nachgang verschriftlicht. Das ist lobenswert und spannend und man kann dem Herausgeber nur danken (siehe auch <https://angewandtedatascience.de>).

Es stellt sich die Frage, wer die Leserschaft von solch einem Werk sein soll. Das Buch ist sicher kein Lehrbuch und auch kein klassisches Textbuch. Es ist ein Sammelwerk und hat auf dem Markt sicher Konkurrenz mit ähnlicher Ausrichtung. Für wen also wäre das Buch eine spannende Lektüre? Mir fallen hier gleich mehrere Adressaten ein. Zunächst einmal Kolleginnen und Kollegen meiner Profession, sprich Hochschullehrerinnen und Hochschullehrer. Wir lehren unsere Studierenden Werkzeuge der Statistik, des maschinellen Lernens und führen sie, je nach Studienrichtung in Natural Language Processing (NLP) oder Computer Vision ein. Aber was die Studierenden dann im Berufsalltag erleben werden, das entzieht sich zum Teil unserer Kenntnis. Hier betrachte ich das Buch als eine spannende Quelle, eine Art Lagebericht, was jenseits von Hochschulen im Bereich Data Science so passiert. Aber auch für die Data-Scientists in Unternehmen ist das Buch eine Inspiration. Man erfährt, woran Kolleginnen und Kollegen arbeiten und

kann Vernetzungen aufbauen, eine der grundlegenden Ideen von *Data Science Darmstadt* und auf nationaler Ebene der German Data Science Society, deren Vorsitzender ich zur Zeit bin (www.gds-society.de). Insofern wird das Buch sicher dankenswerte Abnehmer im beruflichen Alltag finden. Aber aus meiner Sicht ist das Buch auch ganz hervorragend für Studierende geeignet. Diese erhalten einen lebendigen Einblick in die Arbeitswelt der Data-Scientists und bekommen damit neben der reinen Methodenausbildung den praktischen Aspekt vermittelt.

Die einzelnen Kapitel sind in sich selbstständig und eine Kreuzreferenzierung wurde nicht verfolgt. Das ist nicht schädlich und vielleicht sogar natürlicher, weil die Anwendungsbeispiele schon sehr weit auseinanderliegen. Ich will exemplarisch auf einige Kapitel eingehen, die aus meiner Sicht interessante Aspekte angesprochen haben.

Zunächst beziehe ich mich auf Kapitel 3, verfasst von Felix Bode und Florian Staffel. In dem Kapitel geht es um den Einsatz von KI in der Kriminologie. Die Autoren berichten von einem größeren Projekt in NRW und geben gut nachvollziehbar an, wie einzelne Datenquellen zu einem System kombiniert wurden, um polizeiliche Prognosemodelle zu erstellen. Der technische Fachbegriff hierfür ist *Predictive Policing*. Bei solch einem Projekt ist natürlich Transparenz ein entscheidender Faktor, im Fachjargon als *Explainable AI* oder *Explainable ML* bezeichnet, was auch in anderen Kapiteln im Vordergrund steht. In dem Kapitel wird dem Rechnung getragen durch die Anwendung von Entscheidungsbäumen, die aber bei umfangreichen und komplexen Datenkonstellationen dann auch ihre Grenzen erreichen. Insofern befindet man sich in der Pareto-Front, die auch im Kapitel 4 von Christophe Krech angesprochen wird. Je komplexer die Modelle, desto besser die Prognosegüte, aber desto weniger transparent und erklärbar sind die Modelle. Ein Dilemma, dem man sich stellen muss, insbesondere in entscheidungsrelevanten Anwendungen wie im Bereich *Predictive Policing*. Ein Aspekt, der in diesem Zusammenhang vielleicht etwas zu kurz kommt, ist das Thema Fairness. Wann ist ein Algorithmus fair? Das schon angesprochene Kapitel 4 versucht dieser Sache zumindest dahin gehend nachzukommen, als dass hier gängige Verfahren wie etwa die SHAP Werte vorgestellt werden.

Einer ausgesprochen spannenden Frage widmet sich Kapitel 2. Es wird hinterfragt, ob generative Modelle kreativ sind oder ob Kreativität dem Menschen überlassen ist. Der Autor Vladimir Alexeev zeigt in zahlreichen Beispielen auf, dass generative Modelle durchaus als kreativ betrachtet werden können. Zugleich spiegelt das Kapitel die ungeheure aktuelle Dynamik in dem Bereich wider. ChatGTP kommt mit immer neuen Versionen und in 5 Jahren werden wir über die vom Autor gegebenen Beispiele vielleicht nur noch schmunzeln. Insofern ist das Kapitel ein wertvolles Zeitzeugnis, was es zu archivieren gilt.

Herausstellen möchte ich noch drei weitere Kapitel. Zum einen den Beitrag von Bernhard Rummel, der Verweildauern, sprich Zeiten bis zum Eintritt eines Ereignisses betrachtet. Die Analyse erfordert entsprechendes statistisches Repertoire und es ist erfreulich zu lesen, dass dieses Thema Eingang

in das Buch gefunden hat. Zu nennen ist auch der Beitrag des Herausgebers, Lothar Blum, der einen Abriss über Sankey-Diagramme gibt, den meisten Lesern wohl eher als Darstellung von Wählerwanderungen zwischen den Parteien bekannt. Und schließlich sei noch auf das Kapitel von Ulrich Walter verwiesen, der die aus der Beschreibung von Big Data bekannten 3 „Vs“ (volume, velocity, variety) auf 7 „Vs“ im Bereich Data Science erweitert.

Auch die weiteren Kapitel des Buches sind spannend und sehr lesbar und die selektive Auswahl in dieser Rezension ist nicht dahin gehend zu interpretieren, dass die anderen Kapitel weniger spannend sind. Es ist aber sicher eine sehr persönliche Auswahl, das waren die Sachen, die ich am spannendsten fand. Andere Leser werden dem vehement widersprechen. Und das ist gut so. Insofern ist die detaillierte Vorstellung der Autoren und deren Themen, mit der das Buch startet, sicher hilfreich, wenngleich ungewöhnlich. Statt in medias res zu steigen, beginnt das Buch eher wie ein langes Vorwort. Das ist aber letztendlich hilfreich, denn so kann man sich durchaus auch die Themen heraussuchen, die einen (besonders) interessieren.

Subsumierend habe ich das Buch mit großem Interesse gelesen. Es gibt einen Einblick in die Arbeitswelt von Data-Scientists und macht sich gut im Bücherschrank, nach dem Lesen, versteht sich. Insofern gratuliere ich dem Herausgeber und natürlich allen Beitragenden zu dem Werk.

Hinweis des Verlags

Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.

Funding. Open access funding provided by Projekt DEAL.

Open Access. Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.