Vorworte

Vorwort von Sarah Detzler

Wieso arbeitest du als Data Scientist bei SAP? Diese Frage höre ich sehr oft. Als großer Fan von Büchern bin ich der festen Überzeugung, dass man sich mit einem guten Buch als Grundlage in fast jedes Thema sehr gut einarbeiten kann. Deswegen freut es mich umso mehr, dass die Frage, warum ich als Data Scientist bei SAP arbeite, in diesem Buch beantwortet wird.

Mir persönlich war es immer wichtig, mit Machine Learning und Data Science Business-Prozesse zu verbessern, effizienter zu gestalten und das Leben der Endnutzer*innen zu vereinfachen. Viele der Daten, die hierfür notwendig sind, stammen aus SAP-Anwendungen oder werden in SAP-Systemen gesammelt. Und als Data Scientist ist es eine gute Idee, nah an den Daten zu sein, auch um Datenduplizierung und damit zusammenhängende Risiken zu vermeiden. In der SAP-Welt bedeutet dies oft, mit den analytischen Engines in SAP HANA zu arbeiten.

Natürlich bin ich auch wie die meisten meines Berufstandes ein großer Fan der Open-Source-Welt. Würde man mir R oder Python verbieten, würde ich sofort meinen Job kündigen. In den meisten Fällen müssen die Daten für die Nutzung der Open-Source-Bibliotheken aber aus der eigentlichen Datenquelle abgezogen und in eine separate Data-Science-Umgebung gebracht werden. Wenn ich mit solch einer Umgebung arbeite, bekomme ich leider oft gegen Ende des Projektes Probleme, die Logik nachhaltig produktiv zu setzen und in die Business-Welt zu integrieren.

In diesem Buch wird hingegen beschrieben, wie es SAP HANA ermöglicht, die Algorithmen zu den Daten zu bringen, und eben nicht die Daten zu den Algorithmen bringen zu müssen. Unnötige Datenduplizierung wird also vermieden.

Konkret werden hierzu in SAP HANA native Bibliotheken angeboten, die das Machine Learning direkt dort ausführen, wo die Daten bereits liegen. Das Beste daran ist, dass diese Bibliotheken auch aus R und Python heraus aufrufbar sind. So kann ich in meinem gewohnten Jupyter Notebook bleiben und mich direkt mit SAP HANA verbinden und dort komplett ohne Datentransfer die Algorithmen in der Datenbank ausführen. Auf diese Weise habe ich das Beste aus beiden Welten: meine geliebte Open-Source-Welt (Python und R) und die Vorteile von SAP HANA. Zudem steht mir so alles zentral auf einem System zur Verfügung, was das Produktivsetzen erheblich erleichtert.

Keine Panik, liebe Leser*innen, wenn Sie noch keine Ahnung von SAP haben: In diesem Buch wird dies alles Schritt für Schritt erklärt. Es gibt viele anschauliche Beispiele

und (mein persönlicher Favorit) viel ausführbaren Code, den man direkt verwenden und so gleich loslegen kann.

Sarah Detzler
Competence Lead Data Science und Machine Learning bei SAP

Vorwort von Pavlos Panagiotidis

Im Jahr 2022 ist ein »intelligentes Unternehmen« ohne die Einbindung von Data Science in seine Prozesse kaum noch vorstellbar. Dennoch sehen wir in unserer täglichen Arbeit mit Unternehmen (und trotz der enormen Entwicklungen in diesem Bereich und der Verfügbarkeit von Daten und Open-Source-Technologien) immer noch eine Diskrepanz zwischen Data Science, Datenmanagement und Unternehmens-IT.

Das vorliegende Buch ist ein praktischer Leitfaden, der zeigt, wie Sie mit SAP HANA diese Kluft überbrücken und die Vorteile von Data Science für Ihr Unternehmen nutzen können. SAP hat vor zwölf Jahren SAP HANA mit dem Ziel eingeführt, die Verarbeitung von Transaktions- und Analysedaten in ein und demselben System bei sehr hoher Performanz zu ermöglichen.

Die hohe Akzeptanz dieser Technologie der SAP-Kunden hat eine In-Memory-Revolution ausgelöst und zu der neuen Kategorie von *translytischen Datenplattformen* geführt. Noel Yuhanna, Principal Analyst bei Forrester, beschreibt diese Kategorie wie folgt: »*Translytical* ist eine vereinheitlichte Datenbank, die Transaktionen, Analysen, operative Erkenntnisse und weitere zusätzliche Anforderungen in Echtzeit unterstützt, ohne die transaktionale Integrität, die Performanz oder die Skalierbarkeit zu beeinträchtigen«.

Da Data Science sowohl einen analytischen als auch einen transaktionalen Charakter hat, ist SAP HANA wie gemacht dafür, Data Science in die unternehmerischen Prozesse zu bringen. Die Datenbank hat einen analytischen Charakter, da sie oft mit Daten arbeitet, die bereits für Business-Intelligence-Anforderungen (BI) zusammengetragen wurden. Ebenso hat sie einen transaktionalen Charakter, denn während BI in erster Linie mit Datenaggregationen und Filtern arbeitet, muss Data Science auf der Ebene des einzelnen Datensatzes arbeiten. Sie muss neue Datensätze generieren und Entscheidungen auf der Ebene der einzelnen Transaktion treffen (z. B. Abschätzen, ob ein der Versicherung gemeldeter Schadensfall betrügerisch sein könnte).

SAP HANA bietet also durch die translytischen Fähigkeiten eine mächtige Umgebung für Data-Science-Implementierungen und vereinfacht als vielseitig verwendbares System gleichzeitig das Datenmanagement und die Unternehmens-IT.

Mit den vielen Beispielen in diesem Buch geben Andreas Forster und Stojan Maleschlijski einen praktischen Leitfaden, der für verschiedene Personengruppen geeignet ist (z. B. Data Scientists, Business Analysten, Anwendungsentwicklerinnen und die SAP-IT von Unternehmen), um sie auf dem Weg zum intelligenten Unternehmen zu unterstützen.

Ich wünsche Ihnen viel Spaß bei der Lektüre dieses Buchs und dem Ausprobieren der Beispiele. Hoffentlich ist es für Sie ein wertvoller Ratgeber, den Sie für Ihre Data-Science-Projekte mit SAP HANA neben Ihren Computer legen werden.

Pavlos Panagiotidis
Global Vice President Data Science und
Head of AI Business Process Improvement bei SAP

Einleitung

Warum noch ein Buch über Data Science? Und was ist Data Science überhaupt? Und wie kann ich daraus einen Mehrwert extrahieren? Was auch immer Ihre Fragen sind, wir freuen uns, dass Sie Interesse an dem Thema haben, und hoffen, dass dieses Buch Sie als praktischer Ratgeber bei Ihren Data-Science-Aktivitäten begleitet und unterstützt. Mit diesem Buch möchten wir pragmatisch anhand relevanter Use Cases, an denen wir als Autoren gearbeitet haben, das Thema für Sie greifbar und nutzbar machen.

Der Fokus liegt auf der konkreten Nutzung, das Buch beinhaltet keine wissenschaftlichen oder theoretischen Erklärungen von verschiedenen Algorithmen. Die Hauptbereiche wie Clustern, Klassifizieren, Regression, Zeitreihenvorhersagen etc. werden eingeführt, jedoch nicht in wissenschaftlicher Tiefe behandelt. Vielmehr haben wir versucht, ein praxisnahes Handbuch für strukturierte Daten mit Tipps und Tricks zu erstellen, das immer wieder zurate gezogen werden kann.

Mit dem Buch erhalten Sie außerdem Zugriff auf den von uns genutzten Code als Download. So gelingt Ihnen ein schneller und praktischer Einstieg in das Thema und Sie können unsere Beispiele als Kick-off für Ihre eigenen Projekte nutzen. Das Download-Material können Sie unter *www.sap-press.de/5539* im Bereich **Material** herunterladen.

Zielgruppe des Buchs

Das Buch richtet sich an die folgenden Zielgruppen mit Interesse an SAP HANA, SAP HANA Cloud oder SAP Data Warehouse Cloud:

- Business-Analysts, die sich praktisch in das Thema Data Science einarbeiten möchten
- Data Scientists und Data Engineers, die lernen möchten, wie die oben genannten SAP-Umgebungen für Data-Science-Aufgaben angewandt werden können
- SAP-IT-Spezialistinnen und Entwickler, die die Data-Science-Welt von SAP anhand von praktischen Beispielen kennenlernen möchten
- Berater-Profis, die Ihren Erfahrungsschatz mit Data-Science-Kompetenz erweitern möchten
- SAP-Partner, die ihren Kunden Data-Science-basierte Anwendungen anbieten möchten
- alle Interessierten, wie Studierende oder generell Neugierige, die einen praktischen Einstieg in die Data-Science-Welt suchen

Sie, als Leser*in, sollten technisch interessiert sein. Falls Sie noch keine Erfahrungen mit Data Science gesammelt haben sollten, liefert das Buch Ihnen einen schnellen Einstieg. Falls Sie bereits Data Scientist sind, erklärt Ihnen das Buch die Nutzung innerhalb der SAP Data Warehouse Cloud, SAP HANA Cloud und SAP-HANA-On-Premise-Umgebung.

Das Buch kann sowohl als Nachschlagewerk benutzt werden als auch als durchgängige Lektüre über das Thema Data Science mit SAP. Es werden Beispiele für verschiedenste Data-Science-Anwendungsfälle diskutiert und durch Python-Code, der die Analyse-Engines von SAP nutzt, gelöst.

Inhalt und Aufbau des Buchs

Das Buch besteht aus drei Teilen. Im ersten Teil werden die Grundlagen für das Verständnis von Data Science und deren Nutzung gelegt. Im zweiten Teil werden die verschiedensten Data-Science-Methoden praktisch angewendet. Im letzten und dritten Teil des Buchs wird die Operationalisierung der Methoden erläutert und mit verschiedenen Tipps und Tricks abgerundet.

Kapitel 1, »Einführung«, liefert Ihnen eine kurze Einführung in die relevantesten Data-Science-Themen und grenzt diese voneinander ab. Anschließend erhalten Sie einen Überblick der Nutzung von Data-Science-Fähigkeiten in verschiedenen SAP-Anwendungen.

In **Kapitel 2**, »SAP HANA als Data-Science-Umgebung«, werden die verschiedenen SAP-HANA-Umgebungen (On-Premise, SAP HANA Cloud, SAP Data Warehouse Cloud) vorgestellt. Sie erhalten für diese Systeme Anleitungen, wie Sie sie für die Data-Science-Nutzung einrichten und vorbereiten. Ebenso wird erläutert, wie Sie eine lokale Python-Umgebung aufsetzen, aus welcher heraus Sie mit den SAP-Umgebungen arbeiten werden.

Erste praktische Erfahrungen mit den Systemen sammeln Sie in **Kapitel 3**, »Erste Schritte«. Nach einer kurzen Einführung in die Nutzung von Python verbinden Sie sich aus Python heraus mit Ihrem SAP-HANA-System und laden erste Daten hoch. Zusätzlich werden die SQL- und R-Schnittstellen für SAP HANA vorgestellt.

Im zweiten und umfangreichsten Teil des Buchs nutzen Sie die aufgesetzte SAP-HANA-Umgebung für verschiedene Data-Science-Aufgaben. In **Kapitel 4**, »Explorative Datenanalyse und Datenvorbereitung«, analysieren Sie Tabellen und Daten, mit denen Sie arbeiten, bevor Sie die Daten weiter für die Nutzung von Data-Science-Methoden aufbereiten.

In **Kapitel 5**, »Automated Predictive Library«, nutzen Sie mit der Automated Predictive Library ein automatisiertes Framework in SAP HANA, um verschiedene Data-

Science-Methoden anzuwenden und Vorhersagen zu erstellen. Anhand gängiger Anforderungen aus dem Geschäftsleben werden Klassifizierung, Regression und Zeitreihen vorgestellt und implementiert.

Individuelle Algorithmen der Predictive Analysis Library werden in **Kapitel 6**, »Predictive Analysis Library«, anwendet. Diese ermöglichen eine zusätzliche Feinjustierung der Data-Science-Methoden. Wiederum anhand gängiger Geschäftsanforderungen werden neben der Klassifizierung, der Regression und Zeitreihen auch weitere Themen wie Clustering, Ausreißeranalyse oder die Survival Analysis praktisch vorgestellt.

Über spezialisierte Analyse-Engines werden Sie in **Kapitel 7**, »Spezialisierte Analyse-Engines«, Analysen auf Basis von Geodaten, Graphen und Text erstellen.

Der dritte und abschließende Teil des Buchs fokussiert auf die fortwährende Nutzung der erstellten Data-Science-Modelle. In **Kapitel 8**, »Deployment-Optionen«, wird das Deployment über die SAP Business Technology Platform erläutert (SAP Data Intelligence, Cloud Foundry und Kyma).

Verschiedene praktische Hinweise, die Ihnen das Leben erleichtern können, erhalten Sie in **Kapitel 9**, »Tipps und Tricks«, z. B. zu den Bereichen Datenvorbereitung, Charting oder Logging.

Abschließend haben wir in **Anhang A**, »Checkliste«, noch einige kurze Empfehlungen zum Vorgehen in einem Projekt zusammengestellt, die auf leicht zu übersehende Aspekte aufmerksam machen, die aber dennoch sehr wichtig für erfolgreiche Data-Science-Projekte sind.

Um Sie auf wichtige Informationen hinzuweisen und Ihnen so die Arbeit mit diesem Buch zu erleichtern, verwenden wir im Text Kästen mit den folgenden Symbolen:

In Kästen, die mit diesem Symbol gekennzeichnet sind, finden Sie Informationen zu *weiterführenden Themen* oder *wichtigen Inhalten*, die Sie sich merken sollten. [«]

Mit diesem Symbol sind *Tipps und Hinweise* aus der Berufspraxis markiert, die praktische Empfehlungen geben, die Ihnen die Arbeit erleichtern können. [+]

Dieses Symbol weist Sie auf *Besonderheiten* hin, die Sie beachten sollten. Es *warnt* Sie außerdem vor häufig gemachten Fehlern oder Problemen, die auftreten können. [!]

Danksagung

Ohne die großartige Hilfe vieler Kollegen und Kunden wäre das Erstellen dieses Buchs nicht möglich gewesen. Wir möchten uns für die wunderbare Unterstützung aus den verschiedensten Abteilungen bedanken, vor allem bei: Remi Astier, Dmitry Buslov, Alain Charroux, Marc Daniau, Sarah Detzler, Cristiano Dias, Markus Fath,

John Gibson, Frank Gottfried, Thorsten Hapke, Ian Henry, Tomasz Janasz, Mathias Kemeter, Mike Khatib, Dimitrios Kostakis, Matthias Kretschmer, Bertrand Lamy, Yann Le Biannic, Dimitrios Lyras, Claude Philippe Medard, Christoph Morgen, Dirk Obert, Deniz Osoy, Rafael Pacheco, Pavlos Panagiotidis, Mike Paola, Frank Riesner, Witalij Rudnicki, Klaus-Peter Sauer, Yannick Schaper, Nidhi Sawhney, Gonzalo Hernán Sendra, Raymond Yao und Brook Zhao.

Ganz besonders möchten wir uns auch bei Frau Schweitzer und Frau Hohmann vom Rheinwerk Verlag bedanken, für die wunderbare Betreuung, Unterstützung und Geduld mit uns über das gesamte Projekt hinweg!

Wir wünschen Ihnen viel Spaß bei der Lektüre.

Andreas Forster und **Stojan Maleschlijski**