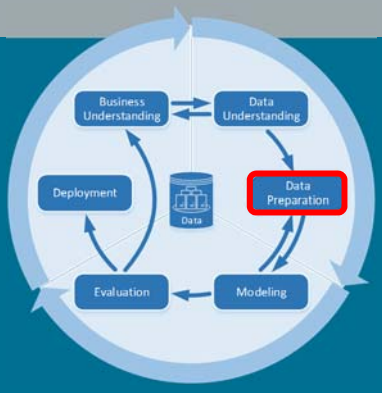
Fraunhofer
IESE

FRAUNHOFER-INSTITUT FÜR EXPERIMENTELLES SOFTWARE ENGINEERING IESE



Basierend auf: Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22

HANDS-ON „DATA PREPARATION“

Die Veränderungen im Zuge der Digitalisierung führen zu einem exponentiellen Wachstum der verfügbaren Daten und ermöglichen Unternehmen durch deren Nutzung einen entscheidenden Wettbewerbsvorteil. Die Zunahme der verfügbaren Daten geht einher mit der verstärkten Nutzung cyber-physischer Systeme (CPS) in produzierenden Unternehmen.

Aus prozesstechnischer Sicht wird die Datenanalyse nach dem branchenüber-greifenden Prozessmodell CRISP-DM (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) in sechs Phasen eingeteilt. Wenn das aktuelle und zukünftige Geschäftsmodell verstanden ist, geht es darum, die dafür notwendigen Daten zu identifizieren und zu verstehen. Die Daten müssen zunächst semantisch verstanden werden (hier ist oft Domänenfachwissen erforderlich) und potenzielle Defizite müssen identifiziert werden – das ist die Grundlage für die Integration und Aufbereitung der Daten. Für die Integration stehen vielfältige Technologien zur Auswahl – wobei es nicht „eine für alles“ gibt. Analyseziele und -verfahren (*Modeling*) beeinflussen die Wahl der Technologie ebenso wie die bestehende Infrastruktur. Die Analyse stellt konkrete Anforderungen an die vom Algorithmus zu verwendenden Daten. Die Aufgabe der **Datenaufbereitungsphase** (*Data Preparation*) ist es, aus den vorhandenen Daten durch entsprechende Datenaufbereitungsschritte (Transformationen, Bereinigungen, Filterungen, Ersetzen fehlender Werte etc.) die vom Analyseverfahren benötigten Daten zu extrahieren. Der explorative Charakter von Datenanalysen und der starke Einfluss, den Datenaufbereitungsverfahren auf die Analyseergebnisse haben können, führen zu wiederholt zu durchlaufenden Datenaufbereitungsschritten, in denen die verwendeten Parametrisierungen evaluiert und ggf. angepasst werden müssen. Zudem wird aufgrund des explorativen Charakters oftmals erst im Laufe einer Analyse ersichtlich, welche Daten tatsächlich wichtig bzw. unwichtig sind, wie Daten aufzubereiten sind und welche Daten zu einem besseren Analyseergebnis beitragen. So verbringen Analyseexperten sehr viel Zeit allein mit der Datenaufbereitung.

Die Entwicklung und Anwendung komplexer maschineller Lernverfahren hängt von einer effektiven und zielgerichteten Datenaufbereitung und der Sicherstellung einer guten Datenqualität ab. Viele Analysemethoden benötigen z.B. zwingend vollständige Daten. Deshalb muss ein Analyst neben der Datenqualität z.B. auch die Vollständigkeit der Daten überprüfen und fehlende Werte ggf. ergänzen (*Missing Data Imputation*). Weitere Datenaufbereitungsschritte wie z.B. Normalisieren, Balancieren, Feature/Instanz-Auswahl, Noise-Filtering, Sampling oder Diskretisieren können notwendig sein. Die Auswahl, Kombination und parametrische Anpassung dieser Verfahren hängt vom geplanten Analyseverfahren ab und ist oft ein zeitintensiver, iterativer Prozess.

In diesem Hands-on-Seminar zeigen wir Ihnen an konkreten Beispielen, wie Sie die Datenvorbereitung angehen und worauf Sie achten müssen. Dazu verwenden wir Fragestellungen aus dem Bereich der Prozessverbesserung. In praktischen Übungen (z.B. Jupyter Notebooks) vertiefen Sie Ihre Lernerfahrungen.

Sprache: Englisch oder Deutsch

Zielgruppen: Unternehmen, die Daten aus Prozessen verwenden, um besser zu werden, oder die schon Erfahrung mit Datenqualitätsproblemen gemacht haben.

Inhalte

Motivation, Technologien zur Datenintegration, Datenvorbereitung, Datenqualitätsanalyse, -sicherstellung, Behebung von Defiziten, Zusammenfassung und Ausblick.

Fraunhofer-Institut für Experimentelles Software Engineering IESE

Fraunhofer-Platz 1

67663 Kaiserslautern

Kontakt

Dr. Andreas Jedlitschka

Tel. +49 631 6800-2260

andreas.jedlitschka@iese.fraunhofer.de

www.iese.fraunhofer.de

2019-04