Inhaltsverzeichnis

Vorwort 5

1 Einführung 11

2 Rollen und Teamstruktur 15
2.1 Rollen 15
2.1.1 Datenanalyst / Data Scientist 16
2.1.2 Data Engineer 18
2.1.3 Business Analyst / Fachbereich 19
2.1.4 Software-Entwickler / Systemadministrator 19
2.2 Teamaufbau 21
2.2.1 Teamstruktur 21
2.2.2 Hiring 22

3 Vorgehen 25
3.1 Arten von Projekten 25
3.2 Projektvorbereitung 27
3.3 Vorgehensmodell 35
3.3.1 Business-/Use-Case-Verständnis 36
3.3.2 Datenverständnis 39
3.3.3 Datenaufbereitung 44
3.3.4 Modellierung 46
3.3.5 Evaluation 48
3.3.6 Operationalisierung 48

4 Daten 57
4.1 Volume (Menge) 57
4.2 Velocity (Geschwindigkeit) 58
4.3 Variety (Vielfältigkeit) 59
4.4 Veracity (Glaubwürdigkeit) 62
4.5 Value (Wert) 63

5 Handwerkszeug 65
5.1 Methoden 65
5.1.1 Begrifflichkeiten 65
5.1.2 Deskriptive Analysen 69
5.1.2.1 Häufigkeitsverteilungen und Histogramme 69
5.1.2.2 Lage- und Streuungsmaße 69
5.1.2.3 Quartile, Whiskers und Boxplots 72
5.1.2.4 Streuungsdiagramme und -matrizen 74
5.1.2.5 Kovarianz und Korrelation 74
5.1.3 Datenvorverarbeitung 76
5.1.3.1 Aggregation und Pivot-Tabellen 76
5.1.3.2 Transformation von Zeichenketten 77
5.1.3.3 Normalisierung 79

5.1.3.4 Imputation – Auffüllen fehlender Daten ... 79
5.1.3.5 Selektion und Reduktion von Attributen ... 80
5.1.4 Zeitreihen ... 83
5.1.5 Supervised Learning ... 88
5.1.5.1 Regressionen ... 89
5.1.5.2 Logistische Regression ... 93
5.1.5.3 Support Vector Machine ... 95
5.1.5.4 k-Nearest Neighbor ... 97
5.1.5.5 Naive Bayes ... 98
5.1.5.6 Entscheidungsbäume ... 99
5.1.5.7 Random Forest ... 102
5.1.6 Unsupervised Learning ... 103
5.1.6.1 k-Means / k-Medoids ... 103
5.1.6.2 One-Class Support Vector Machine ... 104
5.1.6.3 Ausreißererkennung ... 105
5.1.7 Exkurs Deep Learning ... 108
5.1.8 Methodenüberblick ... 113
5.1.9 Evaluation und Optimierung von Modellen ... 114
5.1.9.1 Qualitätskennzahlen ... 114
5.1.9.2 Validierung der Qualität ... 119
5.1.9.3 Optimierungsmöglichkeiten ... 122
5.2 Technologien und Tools ... 125
5.2.1 Speichertechnologien ... 126
5.2.1.1 Relationale Datenbanksysteme ... 126
5.2.1.2 NoSQL-Datenbanksysteme ... 130
5.2.1.3 Hadoop-Ökosystem ... 133
5.2.2 ETL ... 135
5.2.3 Analytics ... 136
5.2.3.1 Visuelle / Workflow-basierte Tools ... 137
5.2.3.2 Programmiersprachen ... 138
5.2.3.3 Notebooks ... 140
5.2.4 Visualisierung ... 143

6 Use Cases ... 145
6.1 Prozesse ... 145
6.1.1 Beschreibung ... 145
6.1.2 Herangehensweise ... 147
6.1.3 Deskriptive Analysen ... 149
6.1.4 Process Mining ... 154
6.2 Berichte ... 157
6.2.1 Beschreibung ... 157
6.2.2 Herangehensweise ... 158
6.2.3 Vorbereitung ... 159
6.2.4 Modellentwicklung ... 164
6.3 Wartung ... 167
6.3.1 Beschreibung ... 167
6.3.2 Herangehensweise ... 170
6.3.3 Modellentwicklung ... 171

6.4 Transporte ... 176
6.4.1 Beschreibung ... 176
6.4.2 Vorbereitung ... 177
6.4.3 Visuelle Analysen ... 179

Abkürzungen ... 187

Quellenverzeichnis ... 189

Stichwortverzeichnis ... 195