

Inhalt

Vorwort	IX
1 Einleitung	1
Verwendete Bibliotheken.	1
Installation mit pip	4
Installation mit conda	5
2 Der Vorgang des maschinellen Lernens: Überblick	7
3 Klassifikation Schritt für Schritt: der Titanic-Datensatz	9
Vorschlag für das Projektlayout.	9
Importe	9
Eine Frage stellen.	10
Begriffe und Bezeichnungen für die Daten	10
Daten sammeln	12
Daten säubern	12
Merkmale gewinnen	18
Stichproben von Daten nehmen	20
Daten auffüllen	20
Daten normalisieren	21
Refaktorisieren	21
Vergleichsmodell	22
Verschiedene Algorithmenfamilien	23
Stacking	24
Ein Modell erstellen.	25
Das Modell auswerten.	25
Das Modell optimieren	26
Wahrheitsmatrix	27
Grenzwertoptimierungskurve (ROC-Kurve)	28
Trainingskurve.	29
Das Modell einsetzen.	30

4	Fehlende Daten	31
	Fehlende Daten untersuchen	31
	Fehlende Daten entfernen	35
	Daten auffüllen	35
	Indikatorspalten hinzufügen	36
5	Daten säubern	37
	Spaltennamen	37
	Fehlende Werte ersetzen	38
6	Erkunden	39
	Datenmenge	39
	Zusammenfassende Statistiken	39
	Histogramm	40
	Streudiagramm	41
	Kombidiagramm	42
	Paarmatrix	44
	Kasten- und Violinendiagramme	45
	Vergleich zweier Ordinalwerte	47
	Korrelation	48
	RadViz	52
	Parallele Koordinaten	53
7	Daten vorverarbeiten	57
	Standardisieren	57
	Den Wertebereich skalieren	58
	Dummy-Variablen	59
	Markierungen codieren	60
	Häufigkeitscodierung	61
	Kategorien aus Text gewinnen	61
	Weitere kategoriale Codierungen	62
	Datumsmerkmale konstruieren	64
	Ein Merkmal col_na hinzufügen	65
	Manuelle Merkmalskonstruktion	65
8	Merkmalsauswahl	67
	Kollineare Spalten	67
	Lasso-Regression	70
	Rekursiver Ausschluss von Merkmalen	71
	Wechselseitige Aussagekraft	72
	Hauptkomponentenverfahren	74
	Merkmalsgewichtung	74

9	Unausgeglichene Klassen	75
	Eine andere Metrik anwenden	75
	Baumalgorithmen und Ensembles	75
	Modelle mit Strafpunkten	75
	Minderheiten erweitern	76
	Minderheitsdaten erzeugen	77
	Mehrheiten verkleinern	77
	Erweitern und danach verkleinern	78
10	Klassifikation	79
	Logistische Regression	80
	Naiver Bayes-Klassifikator	83
	Supportvektormaschine	85
	K-nächste Nachbarn	88
	Entscheidungsbaum	89
	Random-Forest	95
	XGBoost	99
	Gradientenverstärkung mit LightGBM	107
	TPOT	111
11	Modellauswahl	115
	Validierungskurve	115
	Lernkurve	116
12	Metriken und Beurteilung der Klassifikation	119
	Wahrheitsmatrix	119
	Metriken	122
	Vertrauenswahrscheinlichkeit	122
	Trefferquote	123
	Genauigkeit	123
	F1 (F-Maß)	123
	Klassifikationstafel	124
	ROC-Kurve (Grenzwertoptimierungskurve)	124
	Kurve der Genauigkeit über der Trefferquote	125
	Kumulatives Gain-Diagramm	126
	Lift-Kurve	128
	Ausgeglichenheit der Klassen	129
	Klassenvorhersagefehler	130
	Ansprechschwelle	131
13	Interpretation von Modellen	133
	Regressionskoeffizienten	133
	Merkmalsgewichtung	133

LIME	133
Interpretation von Bäumen	135
Partielle Abhängigkeitsdiagramme	136
Stellvertretermodelle	139
Shapley	139
14 Regression	145
Vergleichsmodell.	147
Lineare Regression	147
Supportvektormaschinen (SVM)	150
K-nächste Nachbarn	152
Entscheidungsbaum	153
Random-Forest	158
XGBoost-Regression	161
Regression mit LightGBM	167
15 Metriken und Bewertung der Regression	171
Metriken	171
Residuendiagramm	173
Varianzheterogenität.	174
Normalverteilte Residuen	174
Diagramm des Vorhersagefehlers	176
16 Interpretation von Regressionsmodellen	177
Shapley	177
17 Dimensionsreduktion	183
Hauptkomponentenverfahren (PCA)	183
UMAP	197
t-SNE.	202
PHATE	205
18 Clustern	209
K-Means-Algorithmus.	209
Agglomeratives (hierarchisches) Clustern.	214
Cluster verstehen	216
19 Pipelines.	221
Klassifikationspipeline	221
Regressionspipeline.	223
Pipeline für das Hauptkomponentenverfahren.	224
Index.	225