

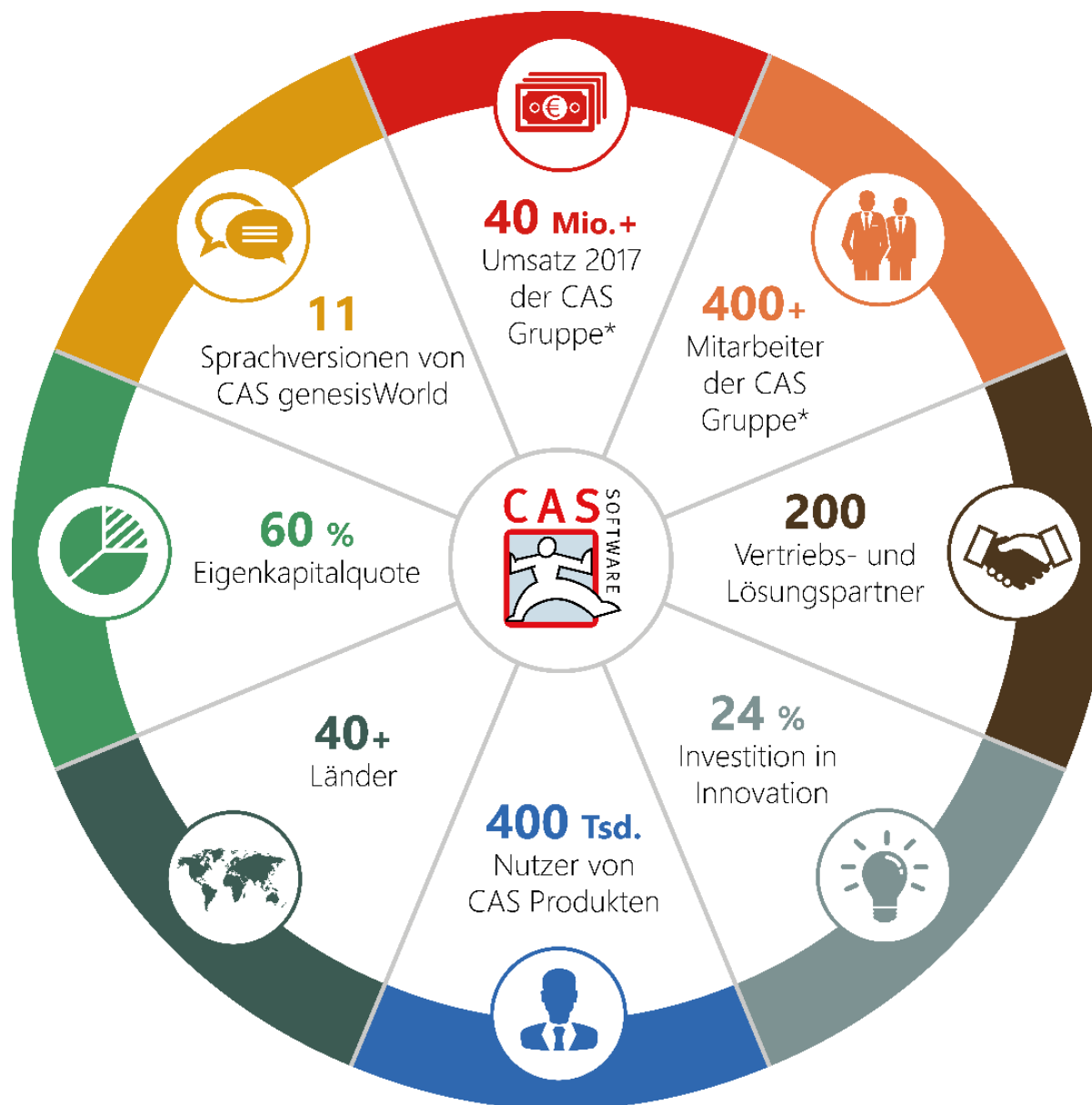
Business Analytics Day Predictive Sales in CRM



Business Analytics Day | 07.03.2019
Artur Felic, Pol Zeimet | CAS Software AG

CAS Software AG

in Zahlen



*CAS Software AG und Beteiligungen anteilig

CAS Software AG – eine Netzwerkorganisation

Jedem Kunden seine CAS

CAS Mittelstand: Marketing & Vertrieb der CRM-Produkte

CAS Drive: CRM-Lösungen für Autohäuser

CAS Education: Flexible Lösungen für Bildungseinrichtungen

CAS Communities: CRM-Lösungen für Netzwerke und Vereine

CAS SmartCustomization: Produktkonfiguratoren,
Vertriebs- & Produktplanungslösungen für Großunternehmen

CAS Aviation: Produktkonfiguratoren für Luftfahrtindustrie

CAS Engineering: Produktkonfiguratoren für den Maschinenbau

CAS Ecosystems: Organisationsübergreifender
Kollaborations-Support

CAS Future Labs



CAS Aviation



CAS Merlin



CAS Mittelstand



CAS Education



CAS Ecosystems



CAS Communities



CAS UserXperience

CAS FUTURE LABS

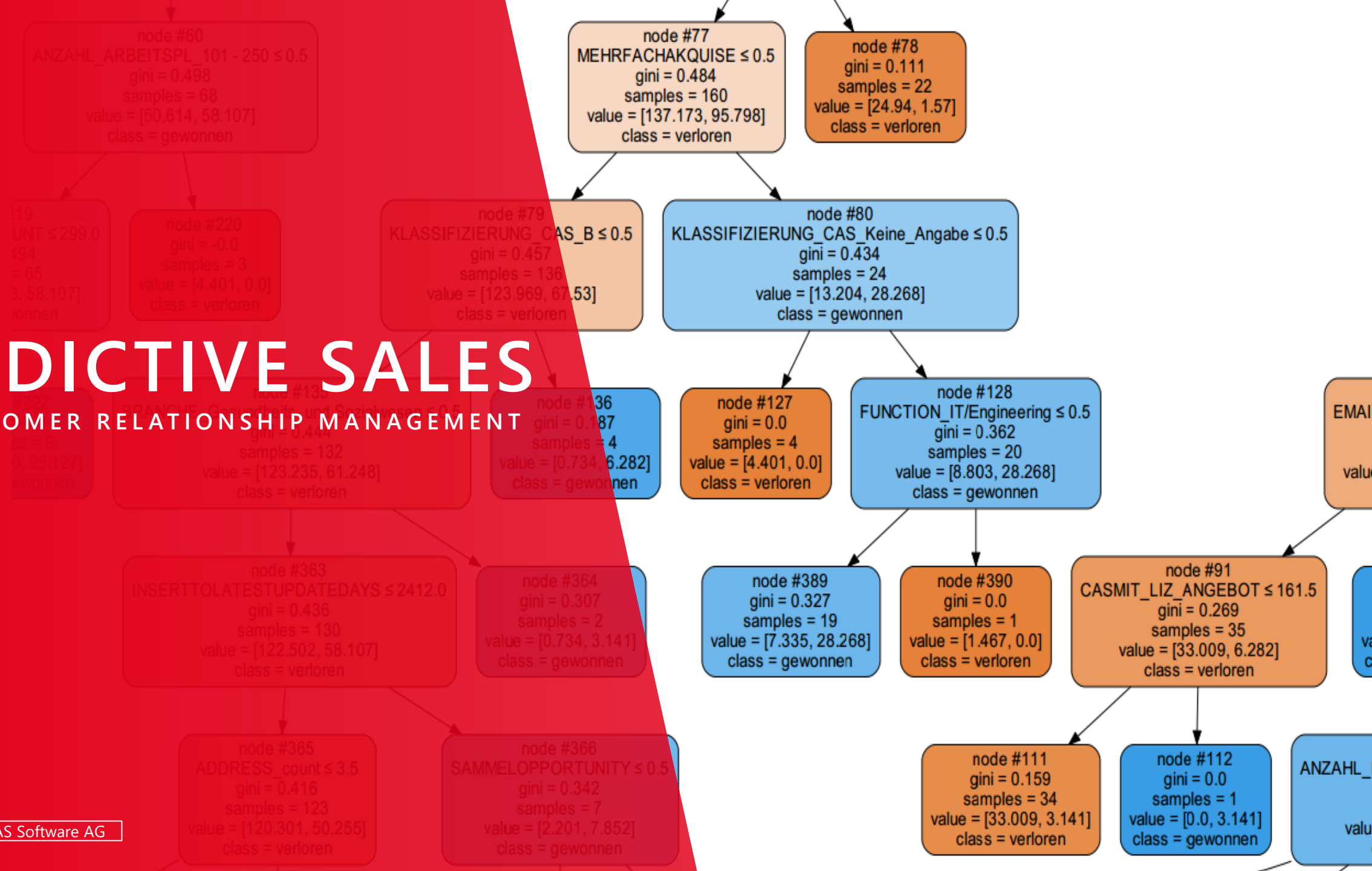
Visions. Grow. Innovation.





PREDICTIVE SALES

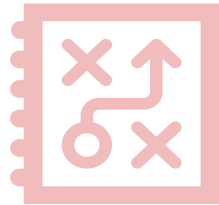
IM CUSTOMER RELATIONSHIP MANAGEMENT





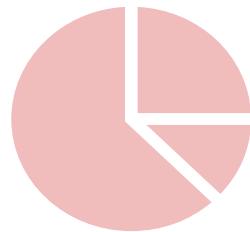
Das Projekt

- Was ist CRM?
- Predictive Sales
- Maschinelles Lernen



Die Vorgehensweise

- Analyse
- Vorhersage
- Datenaufbereitung



Das Ergebnis

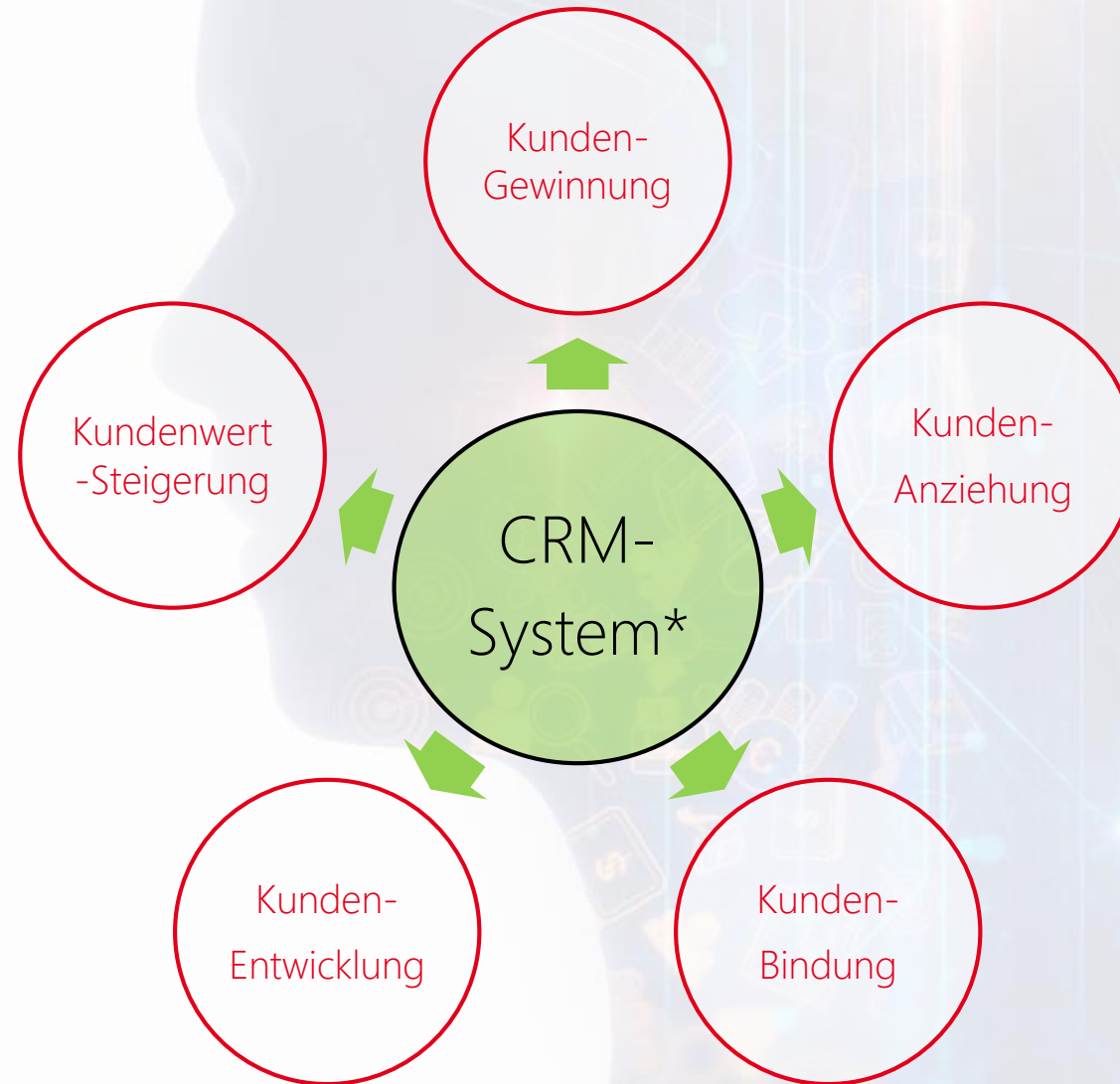
- Modell-Performance
- Aussagekraft und Mehrwert des Modells
- Verbesserungsmöglichkeiten und Ausblick

DAS PROJEKT



Das Projekt

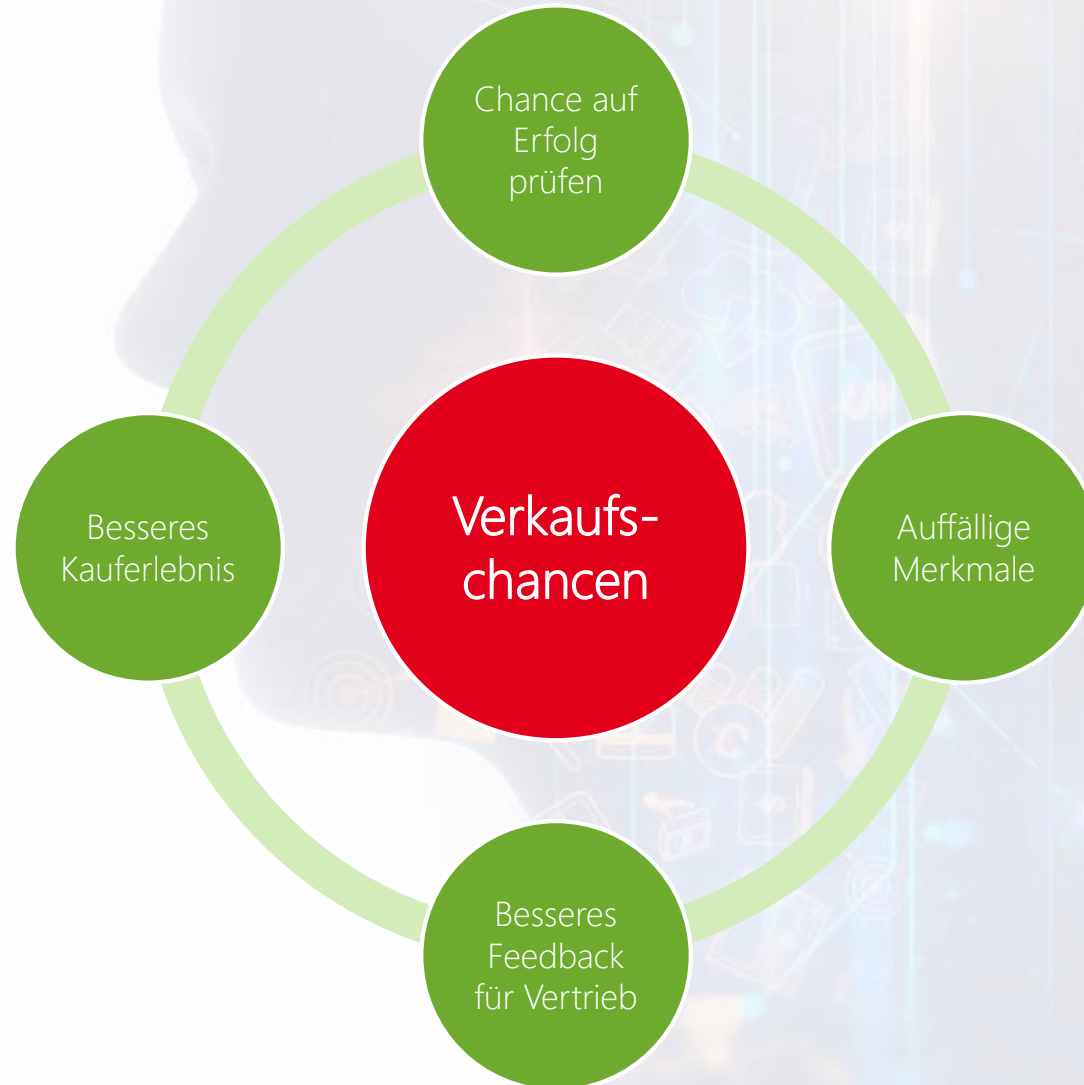
Was ist ein CRM?



*Nach Tala Mirzaei,
Lakshmi Iyer und
Sean Kelly

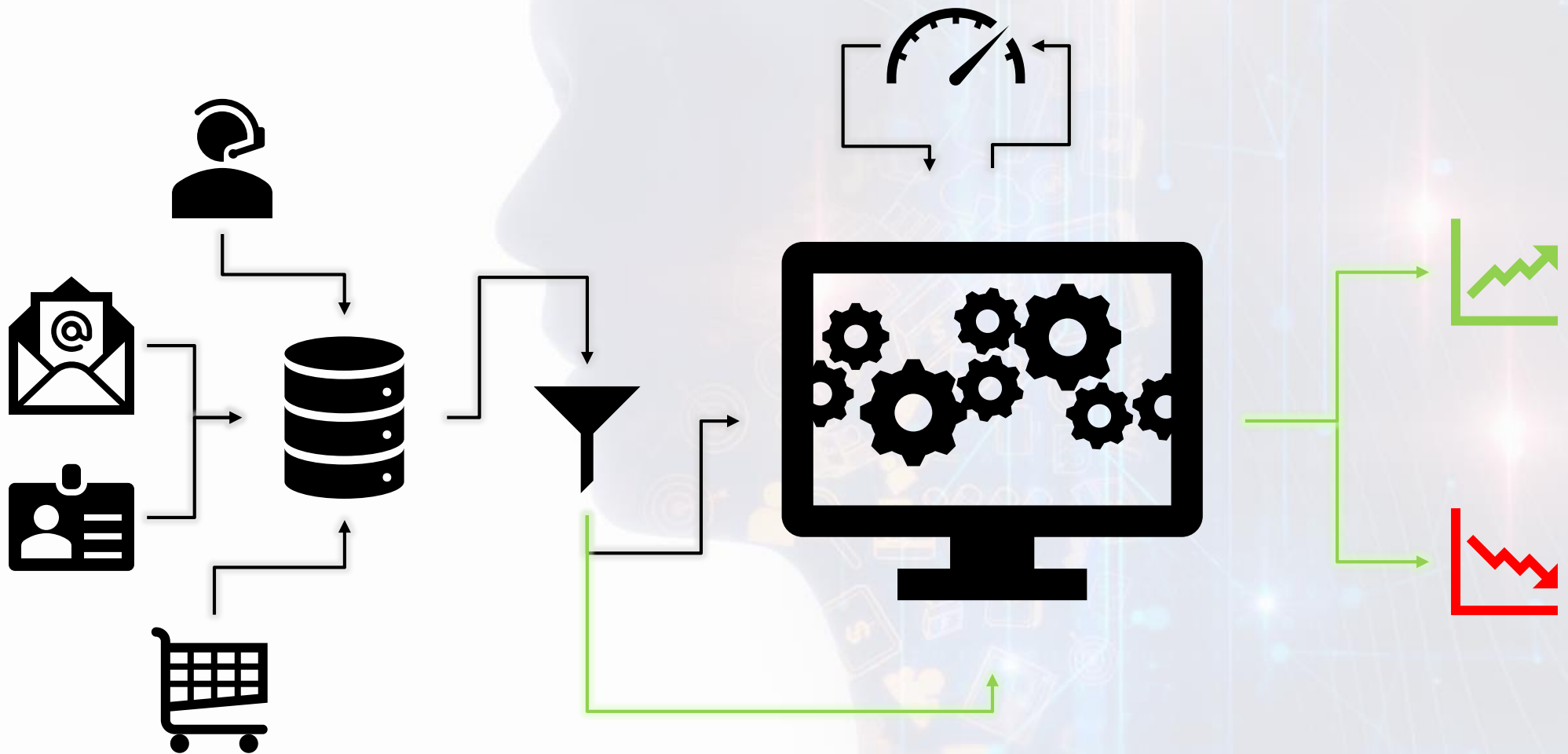
Das Projekt

Predictive Sales

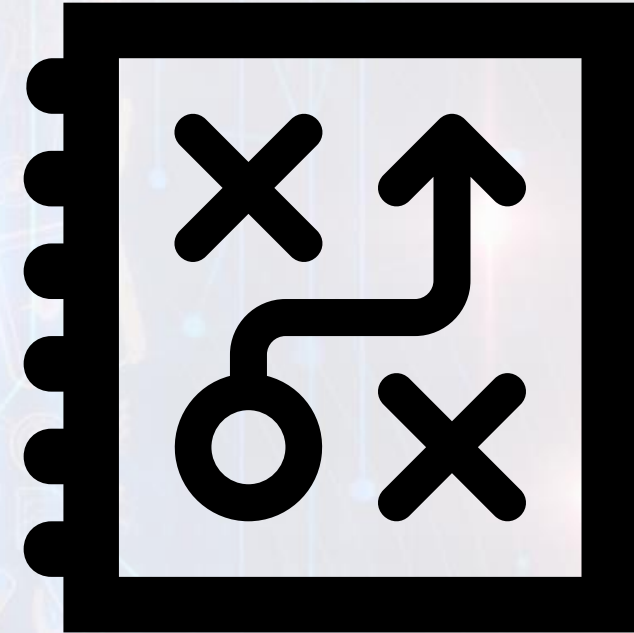


Das Projekt

Maschinelles Lernen



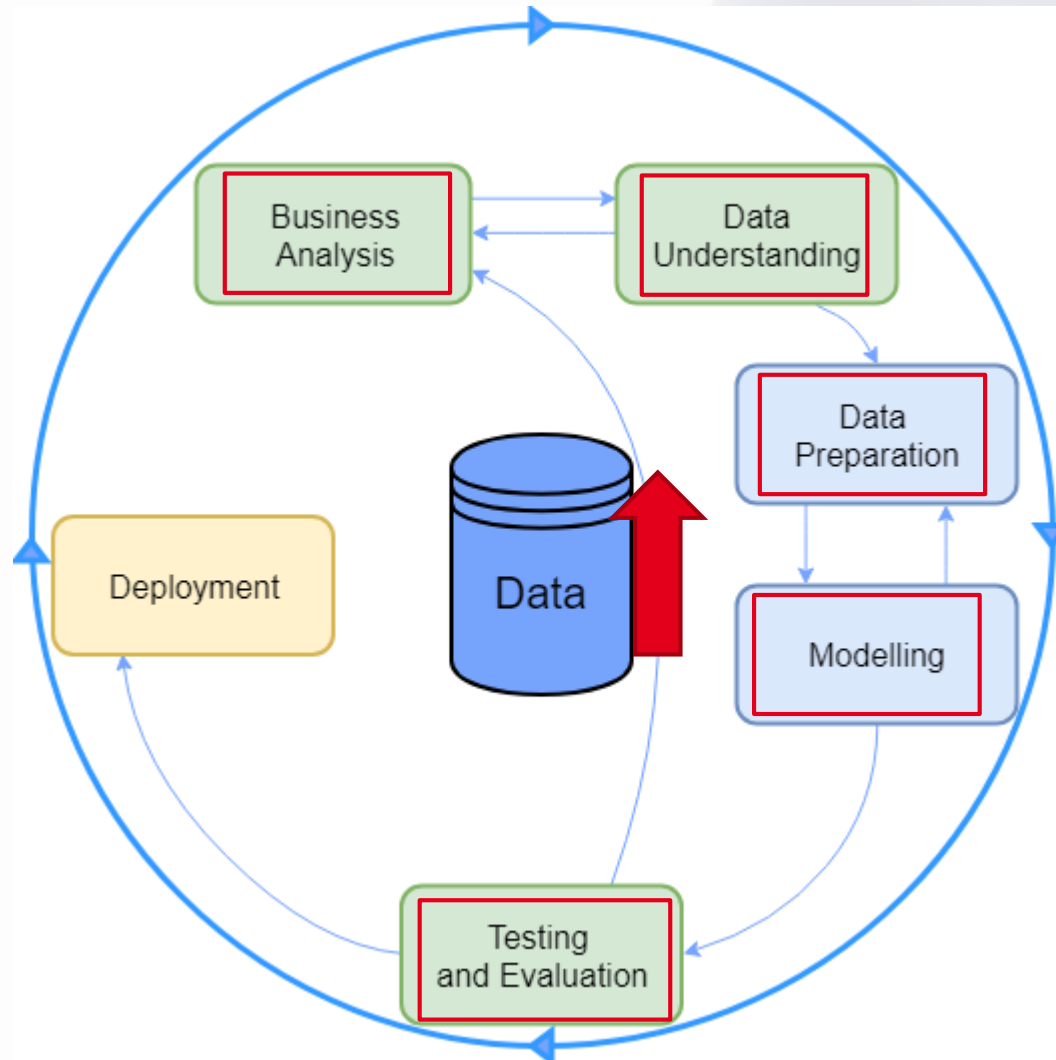
DIE VORGEHENSWEISE



Vorgehensweise

CRISP-DM

Cross Industry Standard Process for Data-Mining



• Prozessanalyse

• Datenanalyse

• Datenaufbereitung

• Modellierung

• Evaluierung

Vorgehensweise

Prozess- und Datenanalyse

Wie Funktioniert der Prozess?



Wie werden welche Daten gepflegt?



Welche Daten enthalten nützliche Informationen?



Welche Daten können ignoriert werden, welche verwendet?

Vorgehensweise

Die Vorhersage

Vorhersage auf „STATUS“-Feld der Verkaufschance.

- 13 Ausprägungen

Multiklassifizierungs-Problem

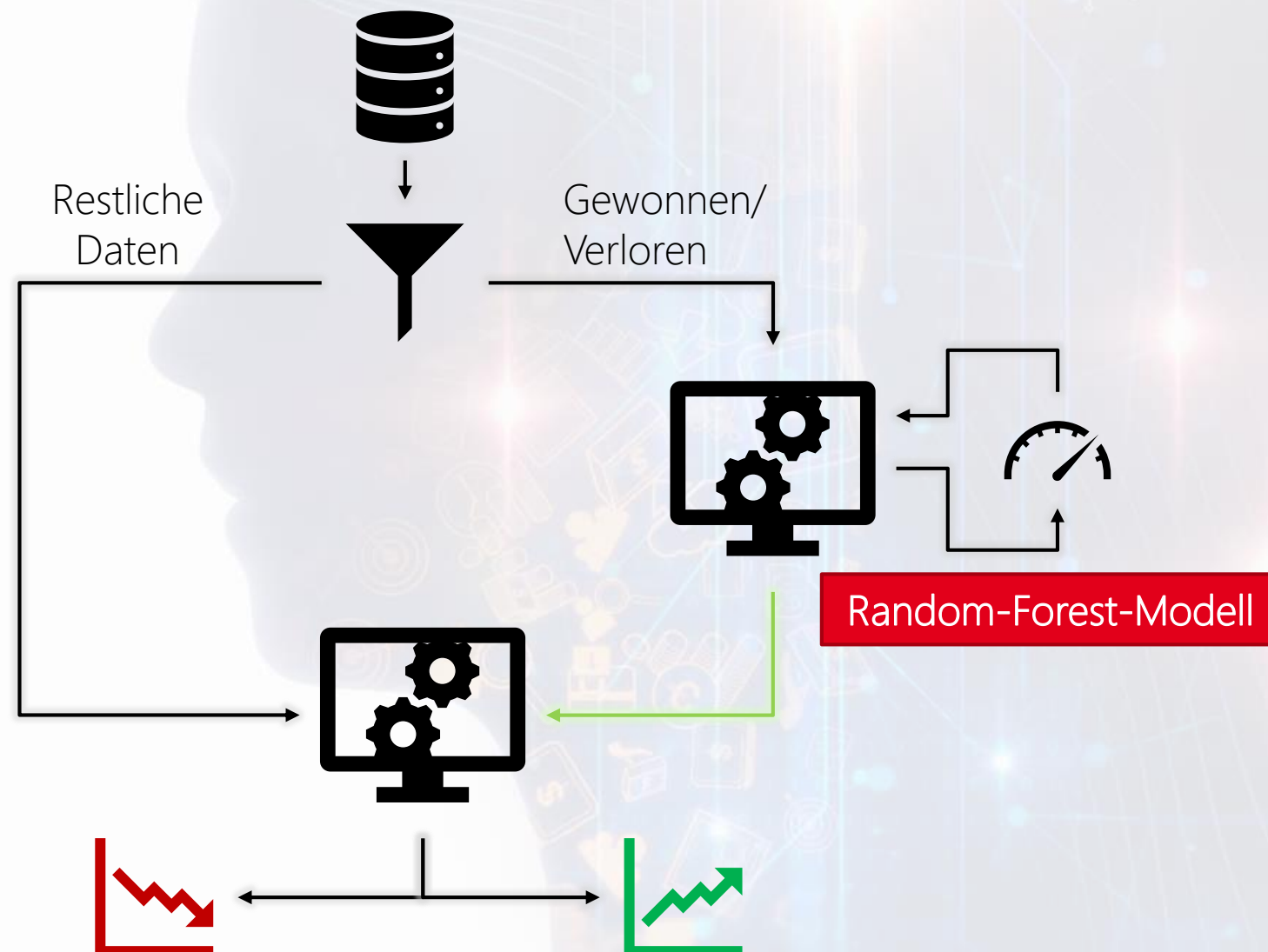
- Weniger Auswahl an Modellen
- Komplexere Modelle
- Schlechtere Performance

Vorschlag

- Beschränkung auf 2: Ausprägungen; gewonnen/verloren
 - Positiv aufgenommen da übersichtlicher
 - Performantere Modelle

Vorgehensweise

Die Vorhersage



Vorgehensweise

Datenaufbereitung - Herausforderungen

Datenzugriff

- Infrastruktur
- Datenschutz und DSGVO

Zustand der Daten

- Viele Objekte mit vielen Features
 - Sparse Data und Missing Data
 - Viele Freitextfelder, wenig Kohärenz
- => Über Zeit gewachsenes System

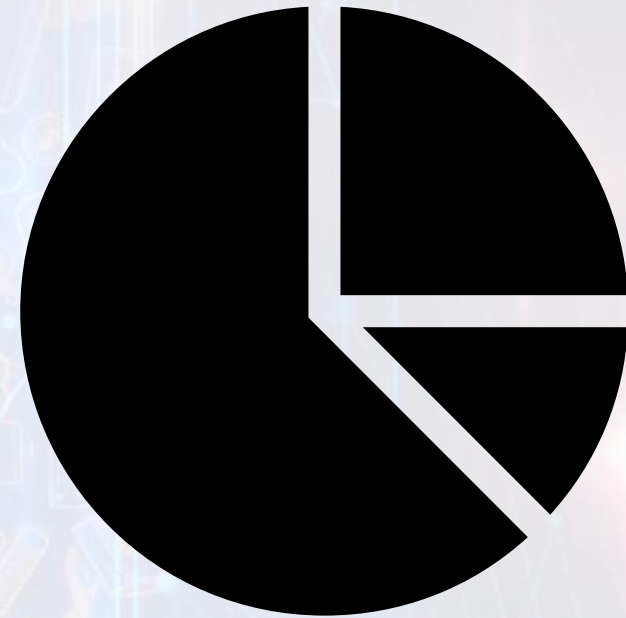
Vorgehensweise

Datenaufbereitung

Am Ende:

- ~75.000 Datenpunkte
- Davon ~10.000 für Modelltraining
- 131 Features

DAS ERGEBNIS



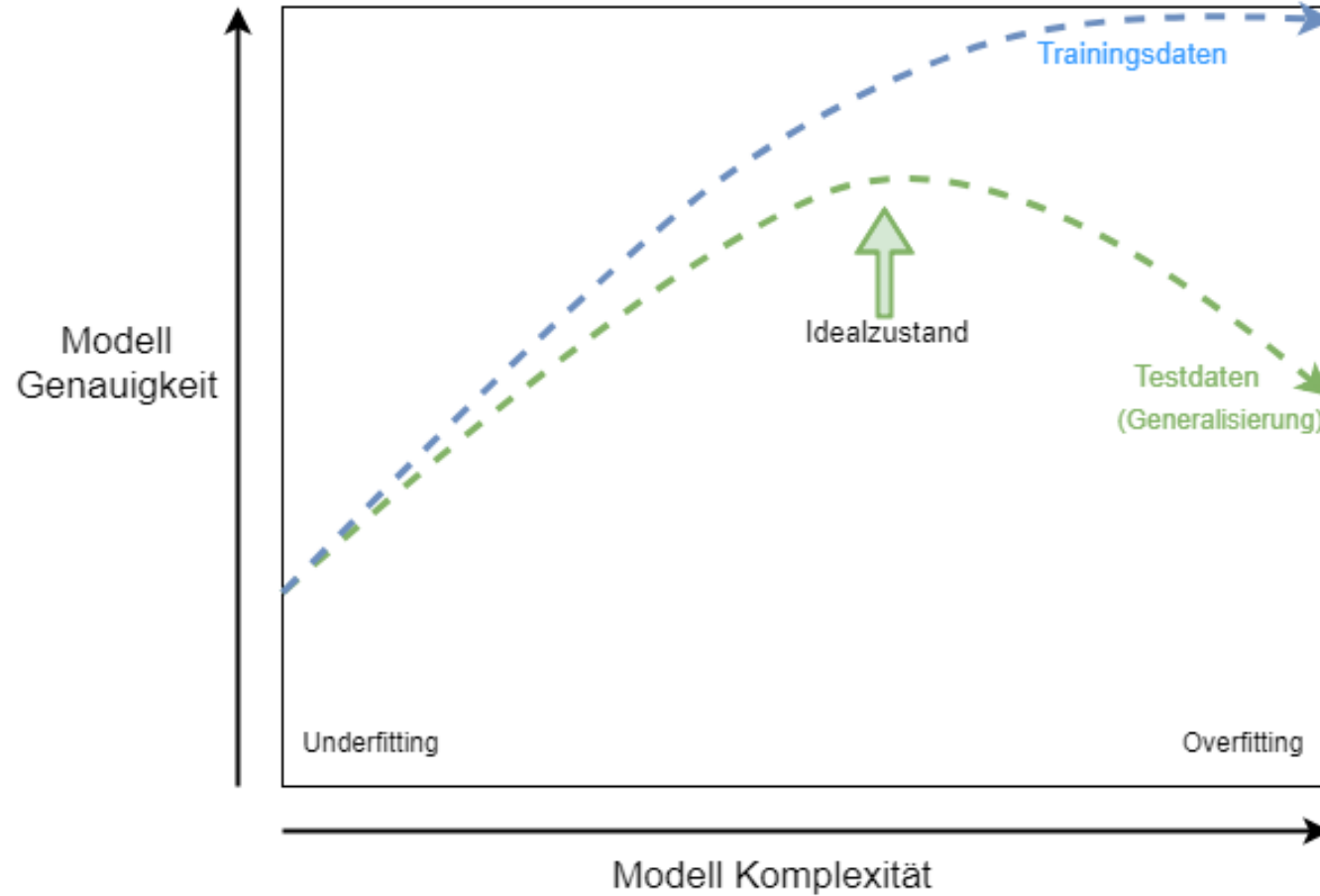
Das Ergebnis

Performance des Modells – Mit Testdaten

Klasse	Precision	Recall	F1-Score	Support
Verloren	0,93	0,94	0,94	1853
Gewonnen	0,88	0,84	0,86	866
Mittel/Total	0,91	0,91	0,91	2719
Micro F1-Score			0,911	
Macro F1-Score			0,897	
Mittlere Genauigkeit auf Trainingsdaten			0,943	
Mittlere Genauigkeit auf Testdaten			0,903	

Das Ergebnis

Performance des Modells – Mit Testdaten



Das Ergebnis

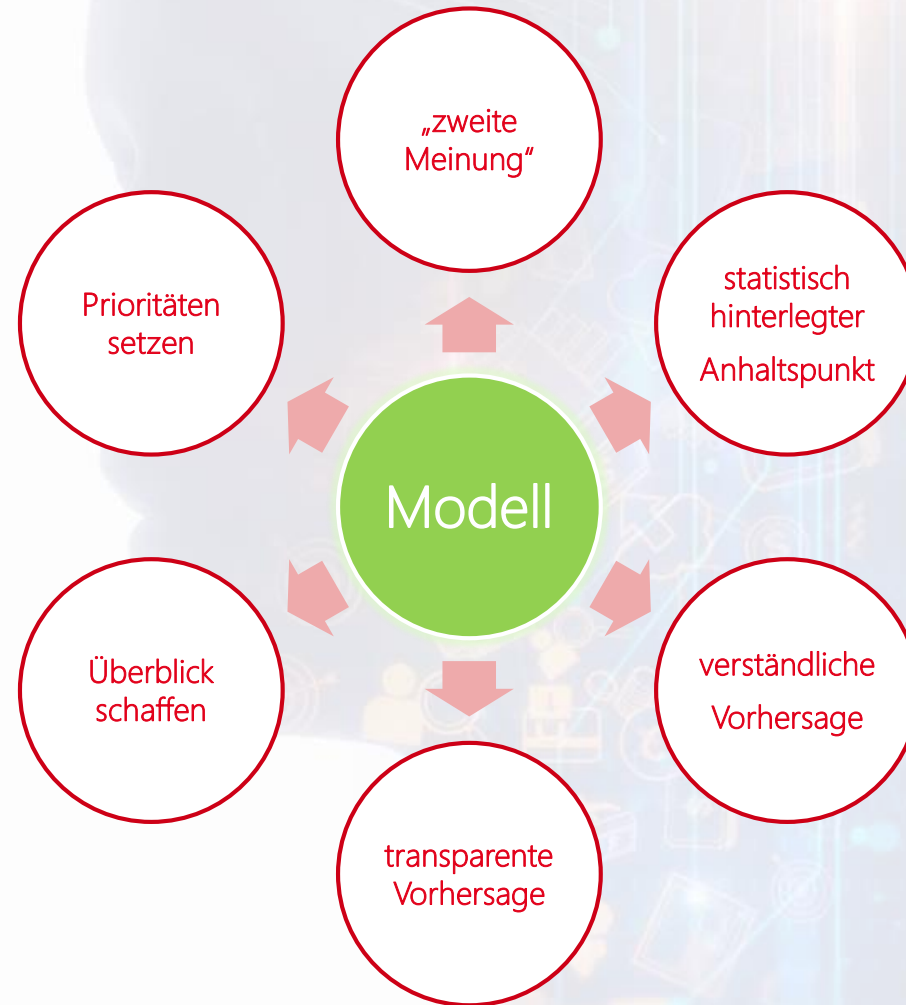
Aussagekraft der Vorhersage

reeller Wert	verloren	1750	103
	gewonnen	138	728
		verloren	gewonnen
		Vorhersage	

Keine_Angabe	28.00	41.00
Kontaktverbo	16.00	32.00
umverteilen	5.00	2.00
verschoben	19.00	81.00
Partner_kein_Inte	16.00	85.00
klären	14.00	116.00
nur_Infopost	13.00	321.00
storniert	82.00	427.00
CAS_kein_Interesse	65.00	606.00
offen	589.00	469.00
neu	168.00	977.00
verteilt	402.00	2005.00
zusammengeführt	839.00	5252.00
nicht_weiterverfolgt	2967.00	16340.00
übernommen	5226.00	29324.00
verloren	332.00	7081.00
gewonnen	3072.00	391.00
	gewonnen	verloren
	Prediction	

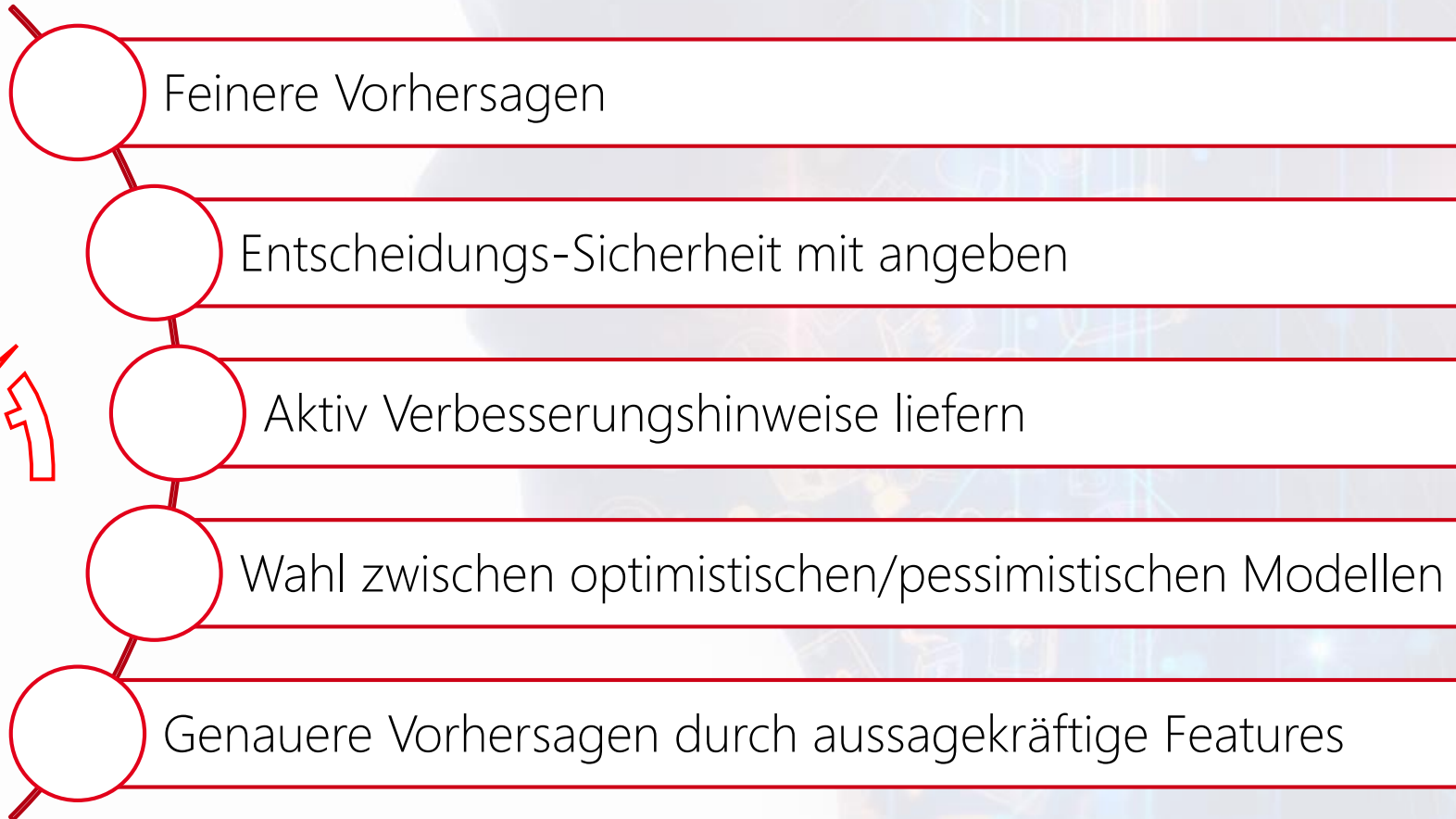
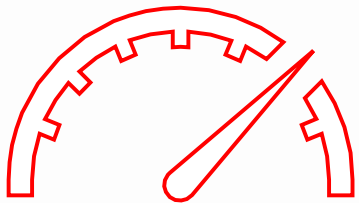
Das Ergebnis

Mehrwert des Modells



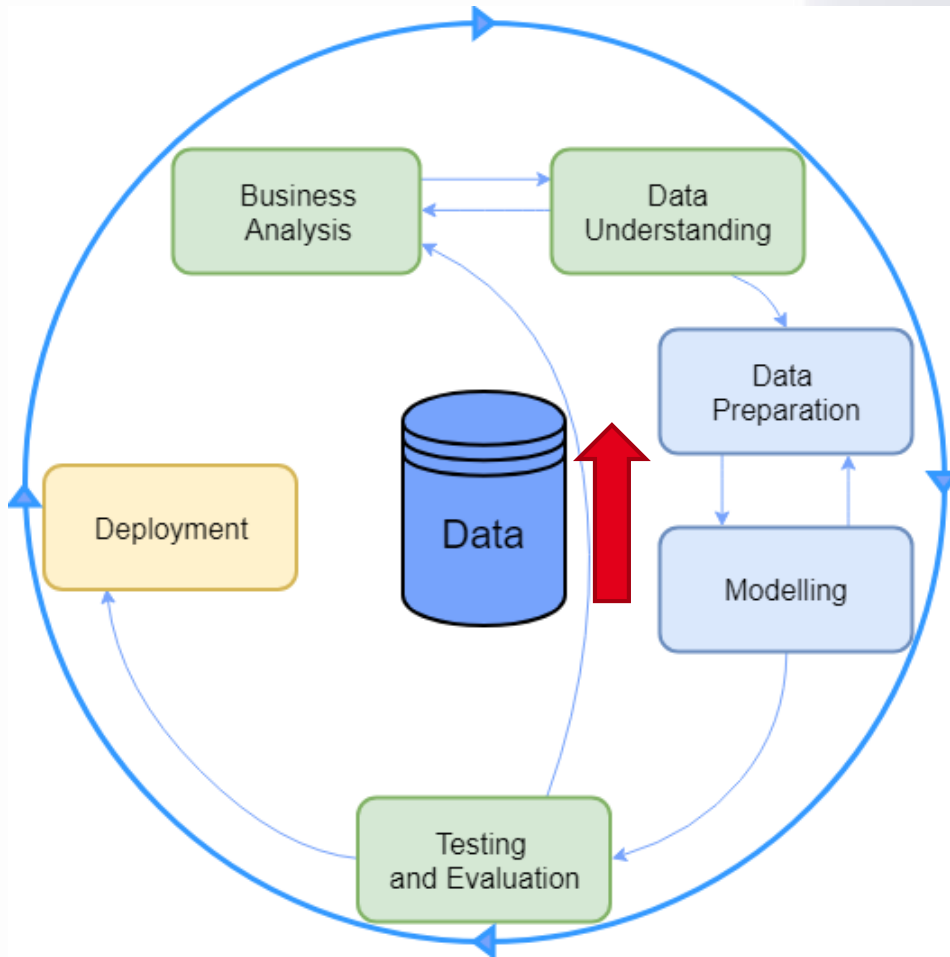
Das Ergebnis

Verbesserungspotential



Ergebnis

Ausblick - Wie geht es weiter?



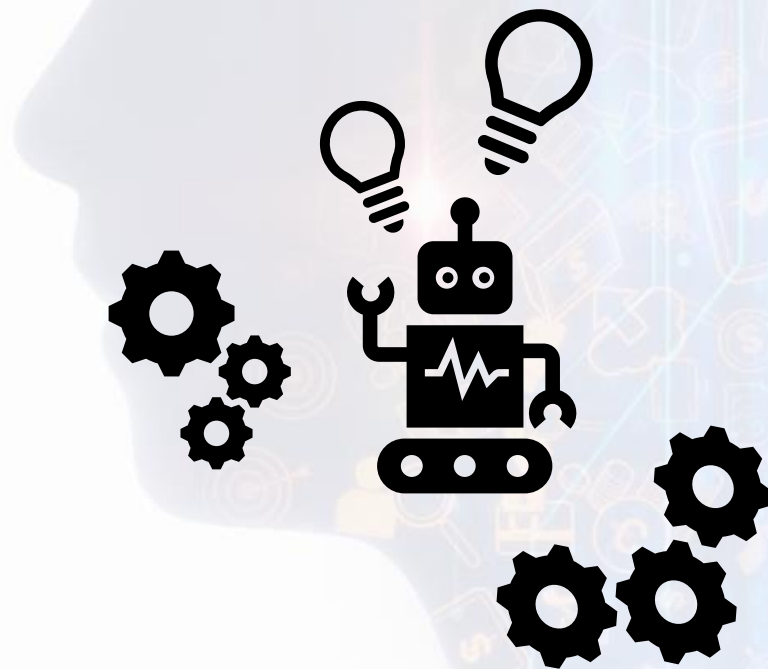
- Erste Iteration weitgehend abgeschlossen
- Nächste Entwicklungsiteration
- Implementierung ausgewählter Verbesserungsmöglichkeiten
- Steigerung der Vorhersage-Genauigkeit

Weiterentwicklung und Nutzung des Modelles oft gewünschte Fortsetzung

- Technologie die existierende Probleme angehen kann
- Hilfreiches, sinnvolles Features
- Attraktiveres Produkt falls in CRM-System eingebaut

Vielen Dank für ihre Aufmerksamkeit!

Fragen?



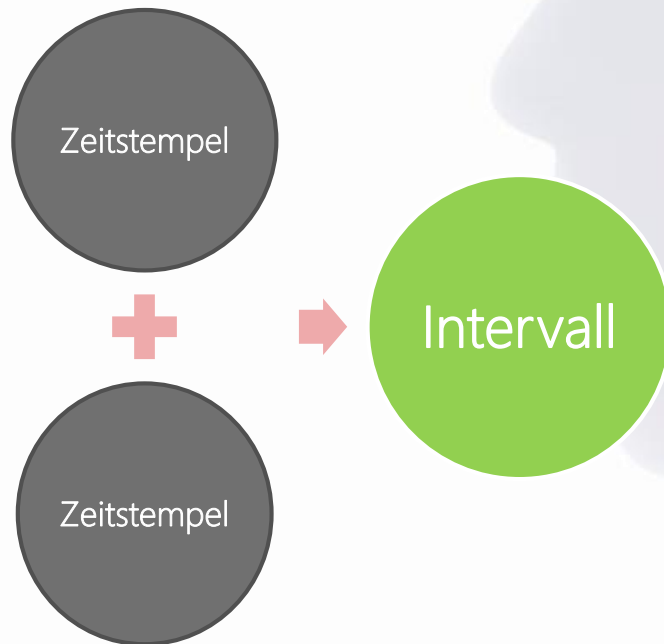
ANHANG



Vorgehensweise

Datenaufbereitung - Beispiele

Feature Engineering Am Beispiel Zeitstempel



Reguläre Ausdrücke und Binning

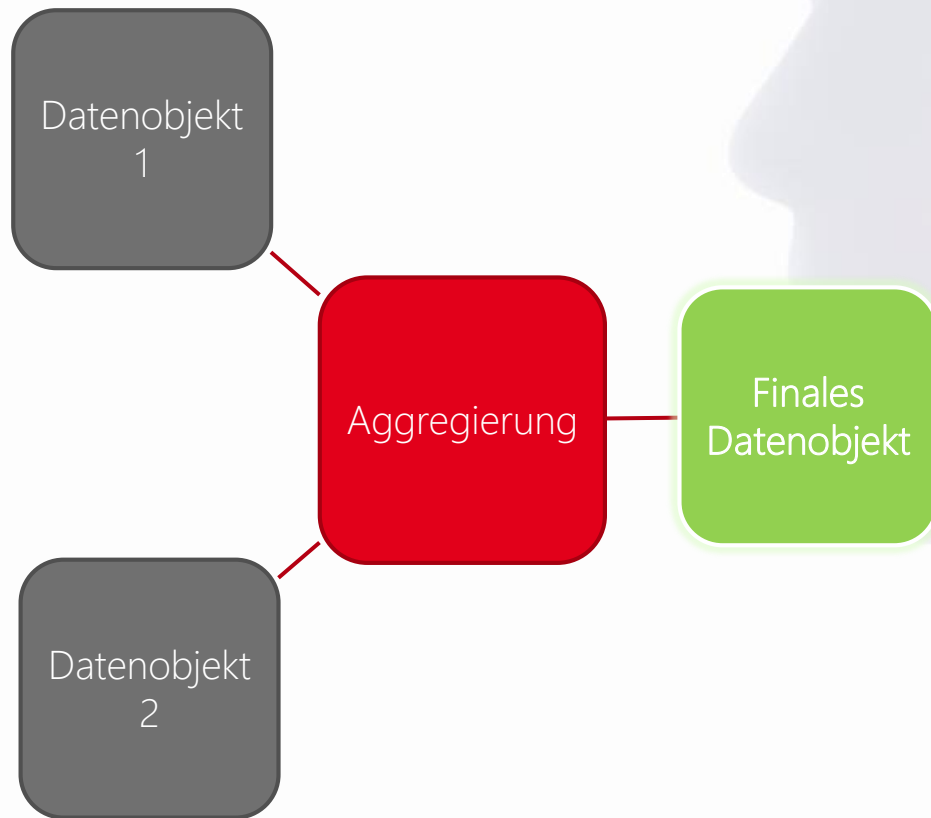


Kategorische Features
Und Bitvektoren

Vorgehensweise

Datenaufbereitung - Beispiele

Aggregation



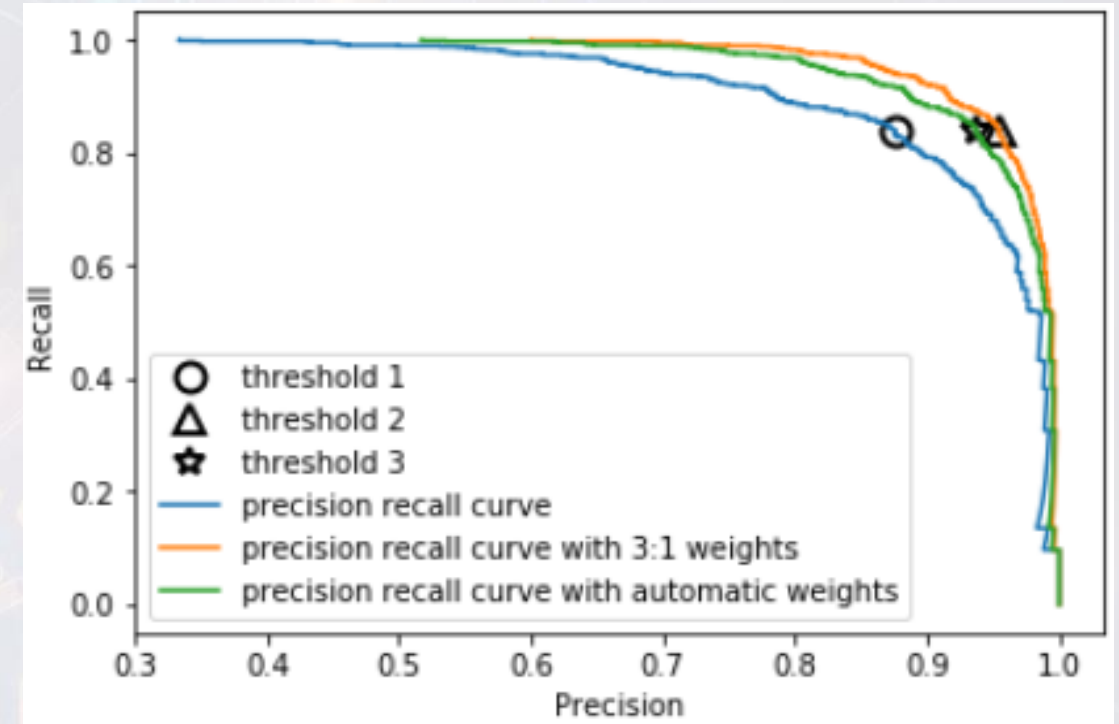
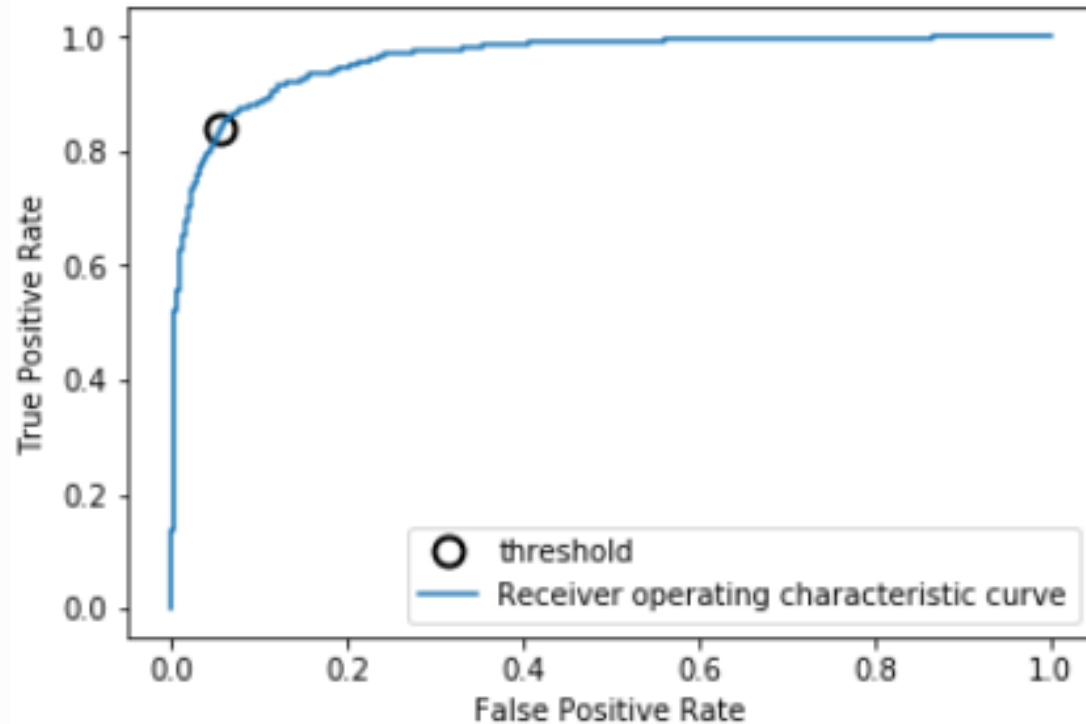
Dummy Variablen

Datenpunkt	Feature X
1	A
2	B
3	C
4	D

Dummy Variablen			
Datenpunkt	X_A	X_B	X_C
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Das Ergebnis

Performance des Modells – Mit Testdaten



ungewichtet	3:1 Gewichtung für gewonnen:verloren	Automatische Gewichtung
0.93318	0.97311	0.96395
AUC:	0.96242	

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\text{TNR} = \frac{\text{TN}}{N} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR}$$

precision or positive predictive value (PPV)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

negative predictive value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

miss rate or false negative rate (FNR)

$$\text{FNR} = \frac{\text{FN}}{P} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}$$

fall-out or false positive rate (FPR)

$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

false discovery rate (FDR)

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}$$

false omission rate (FOR)

$$\text{FOR} = \frac{\text{FN}}{\text{FN} + \text{TN}} = 1 - \text{NPV}$$

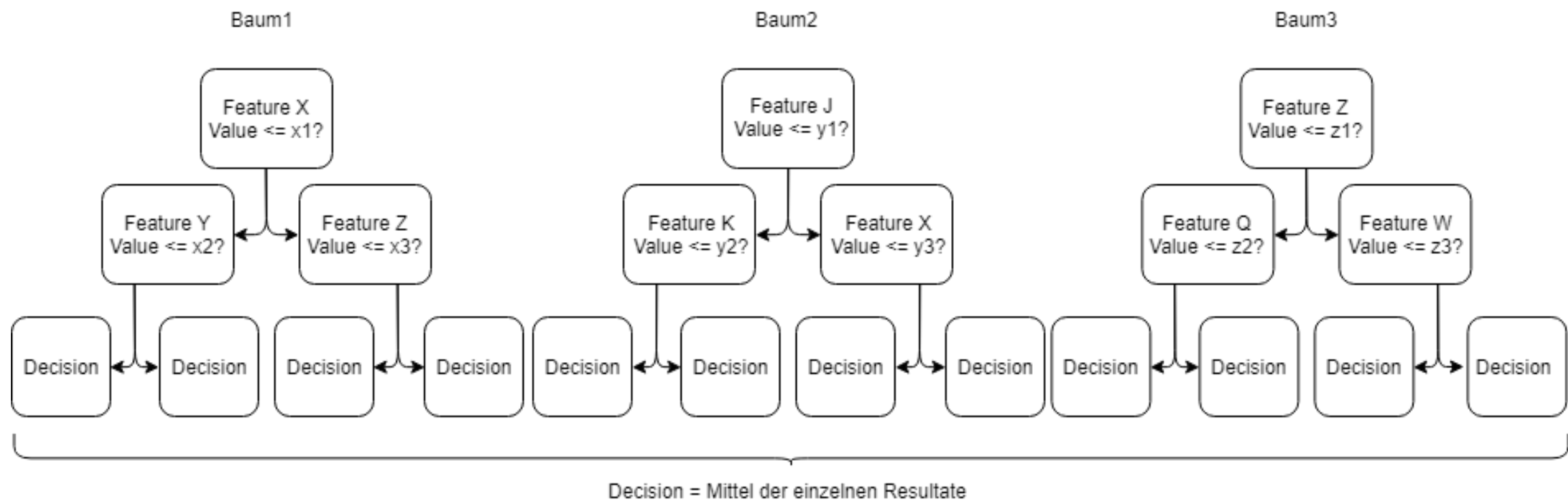
accuracy (ACC)

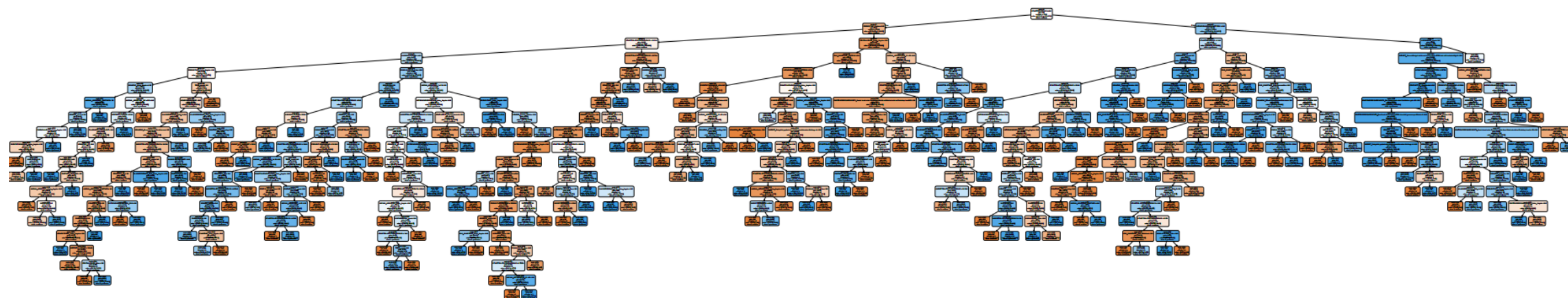
$$\text{ACC} = \frac{\text{TP} + \text{TN}}{P + N} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

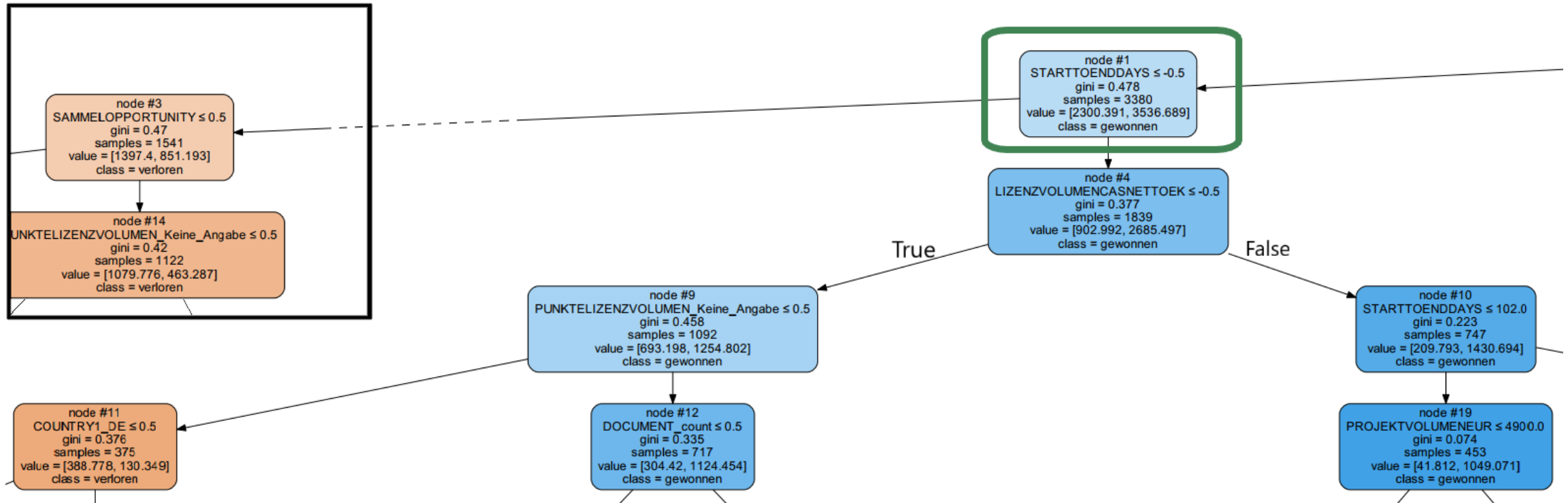
F1 score

is the **harmonic mean** of **precision** and **sensitivity**

$$F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$







node #0
BETREUENDERPARTNER_keine_Angabe ≤ 0.5
gini = 0.5
samples = 5143
value = [4079.234, 4076.93]
class = verloren

node #2
PROJEKTART_Partner ≤ 0.5
gini = 0.357
samples = 1763
value = [1778.842, 540.24]
class = verloren

node #5
MAILGIVEN ≤ 2.5
gini = 0.486
samples = 826
value = [714.471, 510.401]
class = verloren

node #7
SOURCE_kampagne ≤ 0.5
gini = 0.431
samples = 739
value = [704.201, 323.516]
class = verloren

node #87
OPPTOTALAMOUNT ≤ 457.55
gini = 0.437
samples = 719
value = [678.527, 323.516]
class = verloren

node #89
STARTTOLATESTUPDATEDAYS ≤ 632.0
gini = 0.377
samples = 633
value = [633.048, 213.583]
class = verloren

node #91
PRODUKT_PIA ≤ 0.5
gini = 0.483
samples = 320
value = [283.148, 194.738]
class = verloren

node #93
WERBUNG_True ≤ 0.5
gini = 0.437
samples = 286
value = [273.612, 130.349]
class = verloren

node #96
SOURCE_kunden_initiative ≤ 0.5
gini = 0.475
samples = 211
value = [192.922, 122.496]
class = verloren

node #97
FUNCTION_IT/Engineering ≤ 0.5
gini = 0.425
samples = 173
value = [170.182, 75.382]
class = verloren

node #105
SOURCE_messe ≤ 0.5
gini = 0.411
samples = 171
value = [170.182, 69.101]
class = verloren

node #121
SAMMELOPPORTUNITY ≤ 0.5
gini = 0.44
samples = 145
value = [139.373, 67.53]
class = verloren

node #124
gini = 0.0
samples = 20
value = [25.674, 0.0]
class = verloren