Block 1: Zusammenfassung

Grundlagen & Empirisches Fitting

January 11, 2026

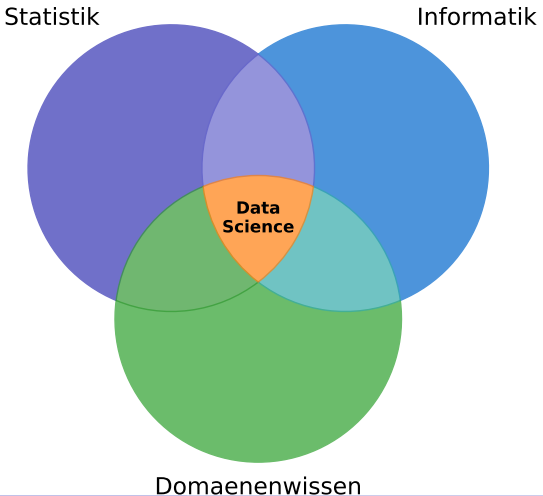
1. Data Science Definition

Data Science = Schnittmenge von:

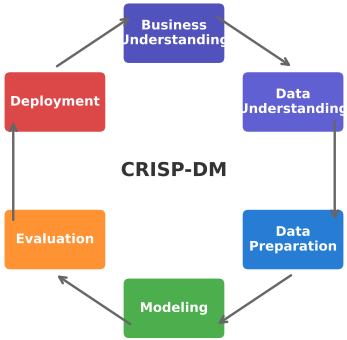
- Statistik
- Informatik
- Domaenenwissen

Kernidee: Modellbasierte Abstraktion der Realität

Data Science als Schnittmenge



2. CRISP-DM Prozess



6 Phasen: Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

3. R-Umgebung und Tidyverse

R-Projekt Best Practices

- Immer `.Rproj` nutzen
- Relative Pfade verwenden
- Strukturierte Ordner

Tidyverse Kernpakete

- `dplyr`: filter, mutate, summarize
- `ggplot2`: Visualisierung
- `tidyr`: Datenbereinigung

Pipeline-Syntax: `daten %>% filter() %>% group_by() %>% summarize()`

Tidyverse macht R-Code lesbar und reproduzierbar.

4. Explorative Datenanalyse (EDA)

Vor jeder Modellierung:

- Dimensionen und Struktur prüfen: `dim()`, `str()`
- Verteilungen visualisieren: Histogramme, Boxplots
- Zusammenhaenge erkunden: Scatterplots, Korrelationen
- Datenqualität: Fehlende Werte, Ausreißer

Ziel: Daten verstehen, Hypothesen generieren, Probleme identifizieren

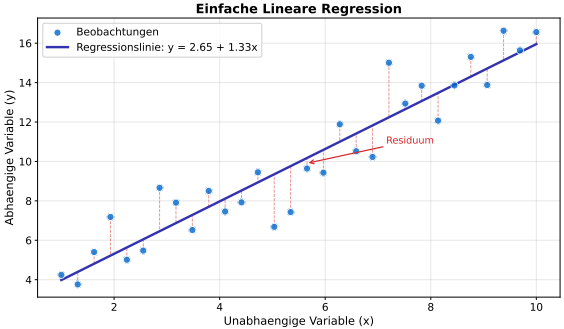
---

Nie blind modellieren – EDA ist Pflicht!

5. Einfache Lineare Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0: Achsenabschnitt
- β_1: Steigung (Effekt von X auf Y)
- ε: Fehlerterm



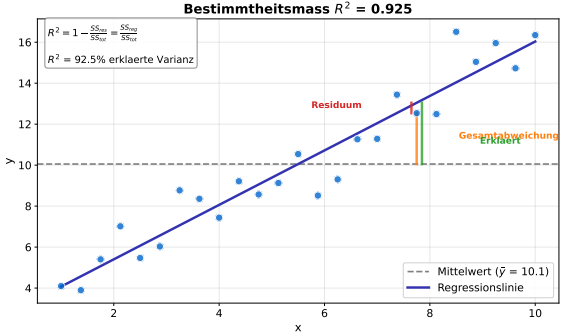
β_1 = Um wie viel aendert sich Y, wenn X um 1 Einheit steigt?

6. Bestimmtheitsmass R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

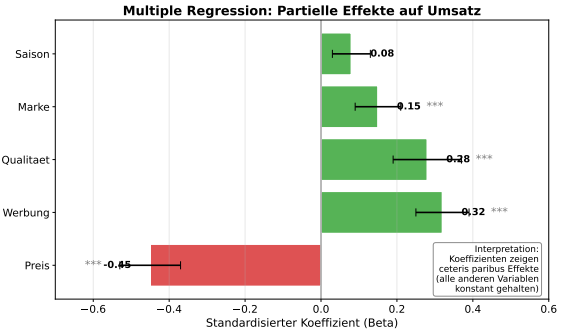
- Anteil erklärter Varianz
- Wert zwischen 0 und 1
- Hoeher = bessere Anpassung

Warnung: R^2 allein sagt nichts über Kausalität oder Generalisierung!



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \varepsilon$$

Partielle Effekte: Jeder β_i zeigt den Effekt von X_i, wenn alle anderen Variablen *konstant gehalten* werden (ceteris paribus).



Interaktionsterme: Effekt einer Variable hängt von einer anderen ab.

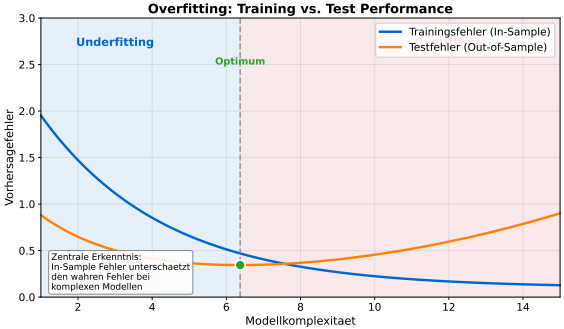
8. Overfitting

Problem: Modell lernt Rauschen statt Signal

- Gut auf Trainingsdaten
- Schlecht auf neuen Daten

Lösung:

- Train/Test Split
- Kreuzvalidierung
- Regularisierung



Ziel ist Generalisierung, nicht Memorierung!

Block 1: Checkliste

Konzepte verstanden?

- ☐ Data Science Venn-Diagramm
- ☐ CRISP-DM 6 Phasen
- ☐ EDA vor Modellierung
- ☐ Regressionskoeffizienten
- ☐ R^2 Interpretation
- ☐ Overfitting

Skills anwendbar?

- ☐ R-Projekt erstellen
- ☐ Tidyverse Pipeline schreiben
- ☐ ggplot2 Visualisierungen
- ☐ lm() ausführen und interpretieren
- ☐ Train/Test Split durchführen

Naechster Block: Inferenz, Unsicherheit & Entscheidungslogik