Feature Selection

-

Wie man aus Sensordaten effiziente Vorhersagemodelle entwickelt

13.02.2025

Dr. David Geisel

CDO

CV

Berufliche Laufbahn

- 2019-2021: PhD Physik Uni Marburg
- 2022-2023: Ailio GmbH - Data Scientist
- seit 2024: Ailio GmbH - CDO

Themen und Fokus

- Data Science
- Machine & Deep Learning
- Microsoft Azure AI Engineer Associate
- Databricks Machine Learning Professional

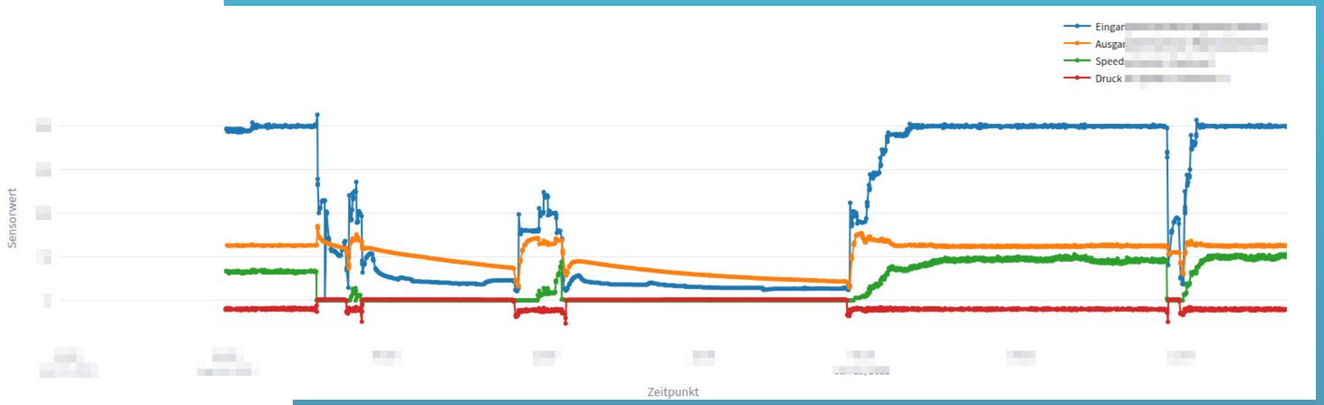
Warum Feature Engineering und Feature Selection?

Feature Engineering

- Rohdaten (oft) nicht für ML-Modelle verwendbar
- Rauschen, fehlerhafte Daten und unsynchronisierte Werte erschweren die Mustererkennung
- Features = bereinigte, normierte und skalierte Informationen aus den Rohdaten

Feature Selection

- Auswahl der für das Modell hilfreichen Features
- Verbessert
  - Modell-Performance und Generalisierbarkeit
  - Interpretierbarkeit
  - Aufwand und Komplexität



	feature_1	feature_2	feature_3	feature_4	feature_5	target
0	-0.440043	-1.523969	0.799146	0.925127	1.410907	102.727401
1	-1.829682	0.497788	-0.362348	-1.003439	-0.893595	-334.791078
2	-0.060331	-0.744331	2.344146	1.356951	-1.544654	-231.906975
3	-0.643925	-0.113965	-0.239783	-0.137655	-0.320272	87.164866
4	-2.832744	0.821210	1.037193	-1.741216	-0.791959	-54.956348

Methoden für Feature Selection

Filter-Methoden

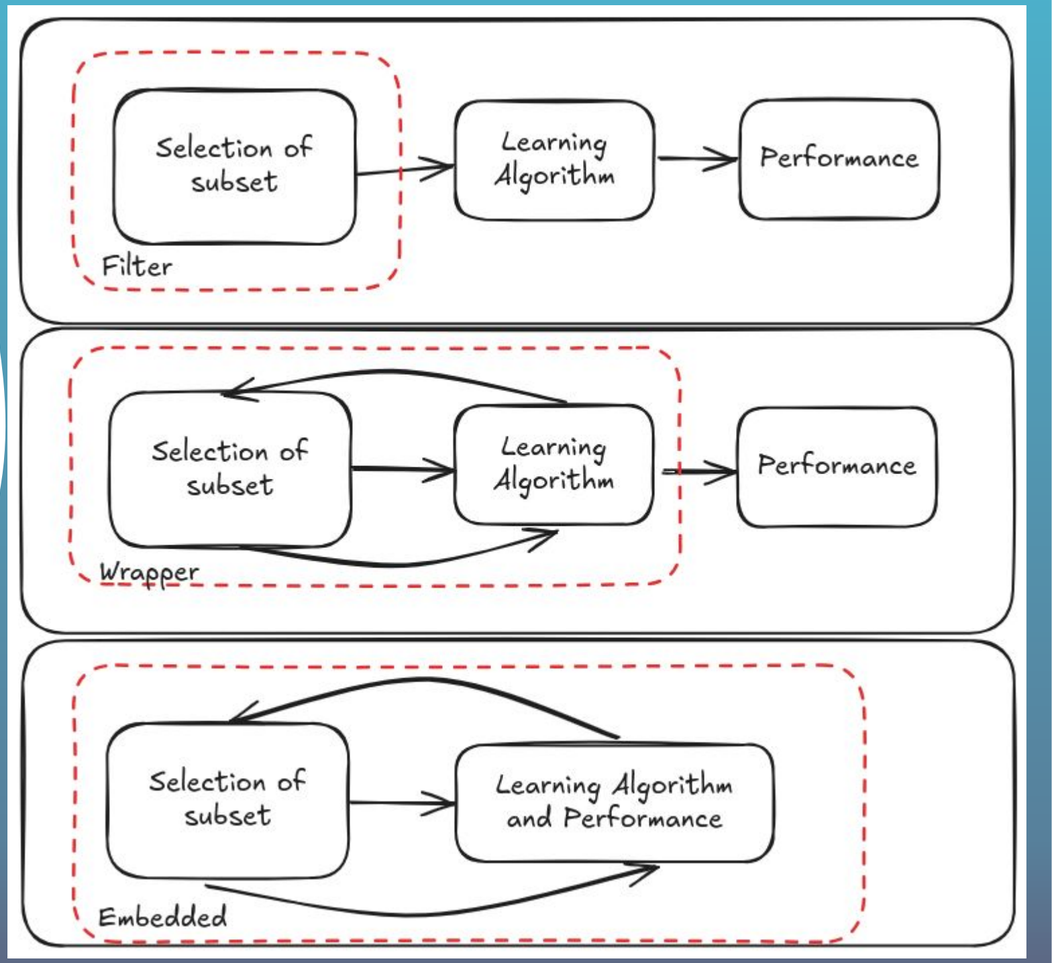
- basierend auf Feature-Statistiken
- *vor dem Modelltraining*
- z.B. Chi-Quadrat Test, Pearson Korrelation, Mutual Information
- einfach und schnell, unabhängig vom Modell
- schwieriger, Wechselwirkungen zu berücksichtigen

Wrapper-Methoden

- *modellabhängige* Feature-Auswahl
- Modelltraining mit unterschiedlichen Features -> beste Auswahl wird gewählt
- z.B. forward oder backward selection
- direkte Optimierung der Modell-Performance
- oft nur “one feature at a time” -> Kombinationen können verpasst werden

Eingebettete Methoden

- nutzt *feature importance* Ausgabe der Modelle (z.B. Random Forest Gini impurity)
- sehr effiziente Methode
- hohe Modellabhängigkeit



Praxisbeispiel: Sensordaten

Beispieldatensatz

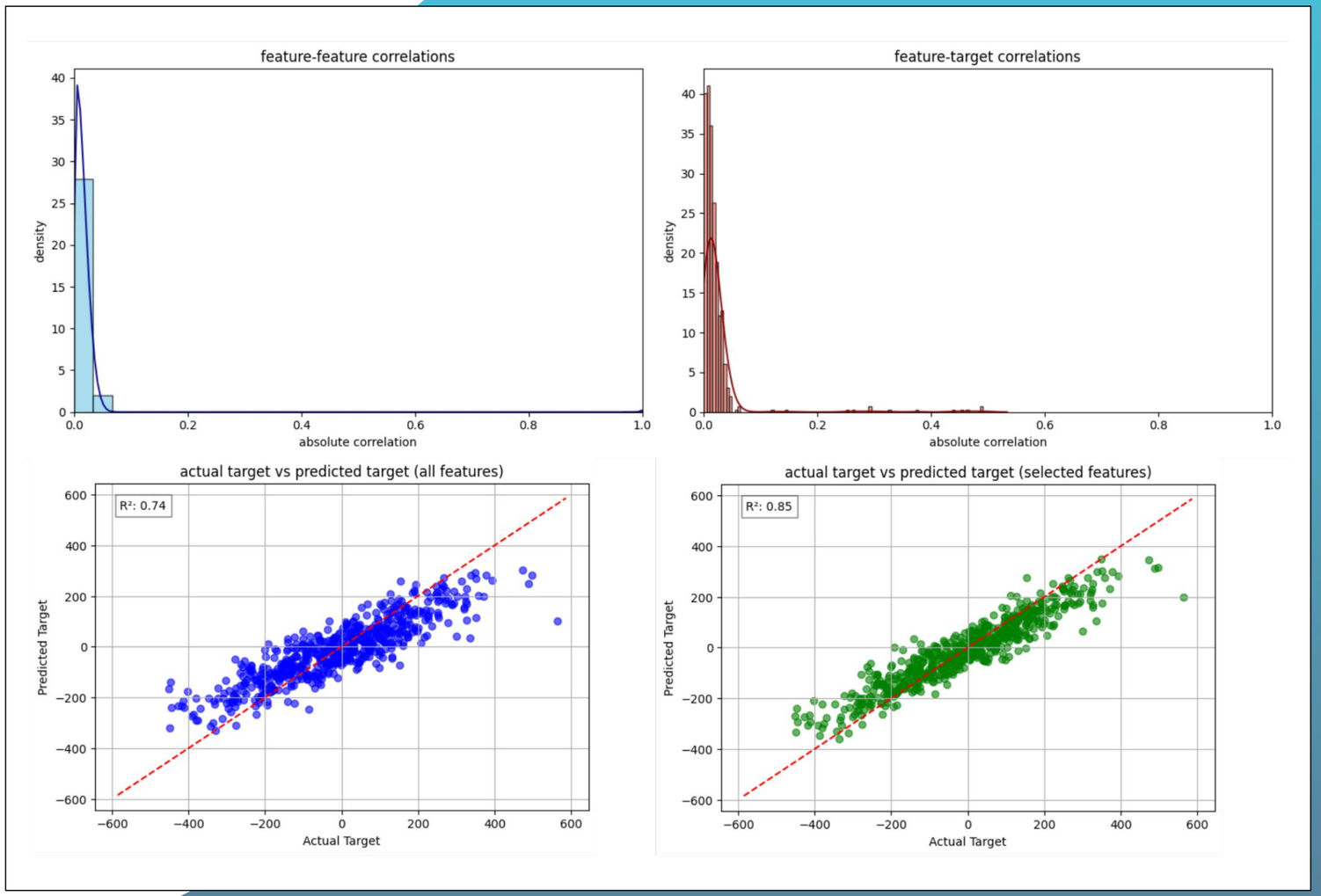
- 600 verschiedene Sensoren als Rohdaten
- Ziel: Vorhersage der Produkteigenschaft (Target)
- nur wenige Features mit Target korreliert
- Features sind auch untereinander korreliert

Feature Selection mit mRMR

- mRMR = Maximum Relevance - Minimum Redundancy
- suche Feature, die eine hohe Relevanz für die Zielvariable aufweisen und untereinander geringe Redundanz besitzen
- Ziel: finde die **beste Kombination** von Features, nicht einfach **die besten Features**

Ergebnis

- deutlich besseres Modell mit den 15 von mRMR ausgewählten Features
- auch ein Random Forest profitiert von der Vorauswahl



Merke:
Typischerweise noch Feature Engineering, z.B. unter Verwendung von TSFresh
-> noch mehr Feature

Praxis-Erfahrungen

Domänwissen nutzen!

- Data Scientists neigen zum Over-Engineering
- wichtig, Prozessexperten einzubeziehen
- Prozess- und Datenverständnis spart Entwicklungsstunden

“Data Science zum Anfassen”

- Modelle und Daten visualisieren
- Interaktion der Experten fördern
- wir nutzen interaktive Apps, um Feedback zu bekommen