Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Unüberwachtes Lernen: Explorative Datenanalyse (EDA) Theorie

Violeta Vogel

Technik und Informatik, BFH

Kursinhalt

1. Datenverständnis
2. Explorative Datenanalyse
3. Feature Engineering
4. Maschine Learning:
    1. Clustering
    2. Dekomposition
    3. Klassifikation
    4. Validierung

Überwachtes Lernen vs. Unüberwachtes Lernen

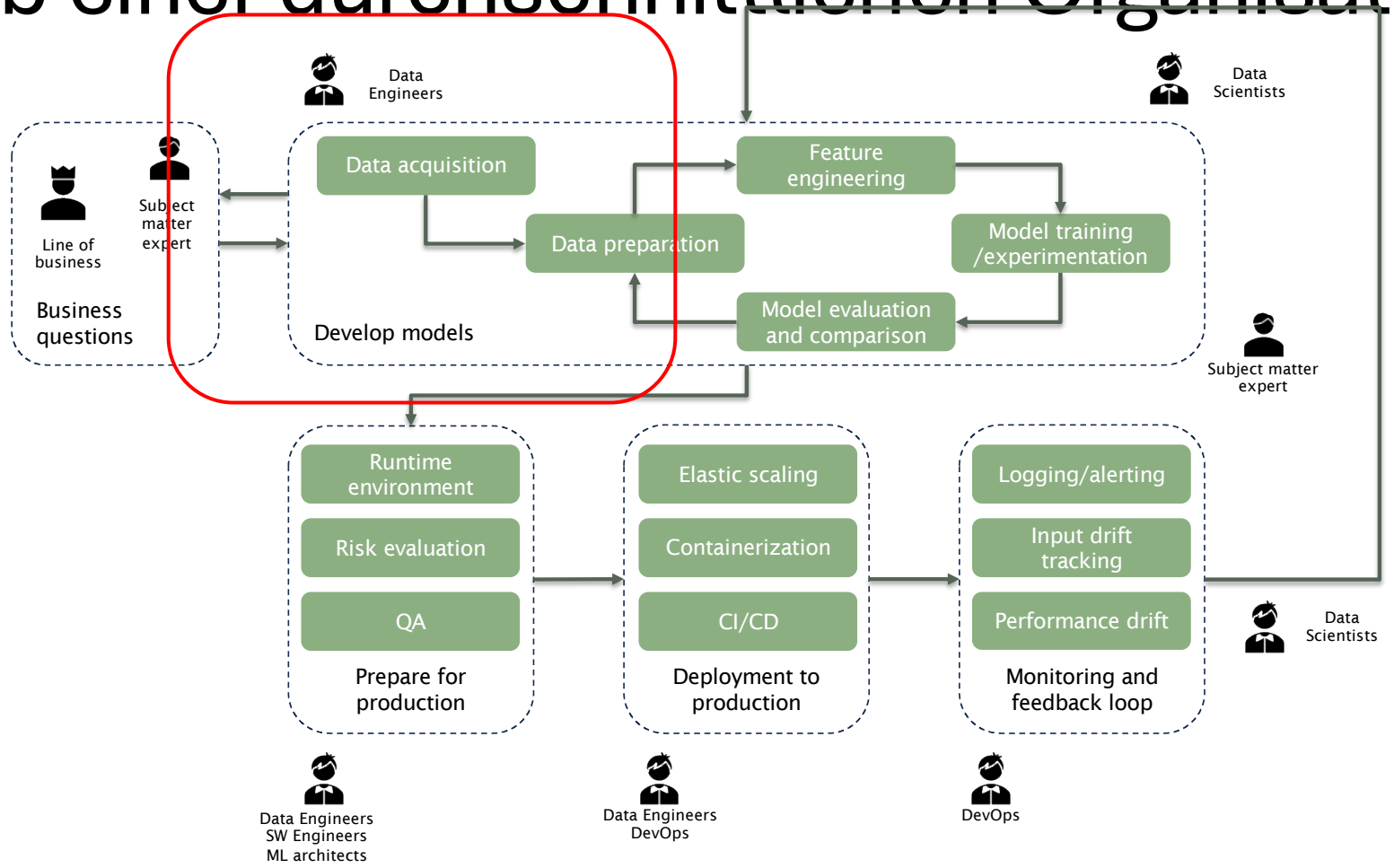
- Beim überwachten Lernen kennt das Modell die richtigen Antworten (Labels) und lernt, diese vorherzusagen.
- Beim unüberwachten Lernen gibt es keine Labels - das Modell sucht selbst Strukturen und Muster in den Daten.

Methode	**Überwachtes Lernen**	**Unüberwachtes Lernen**
Labels vorhanden	Ja	Nein
Ziel	Vorhersage	Muster finden
Beispiele	Spamfilter, Preisprognosen	Kundensegmente, Anomalien
Schwierigkeit	Benötigt viele gelabelte Daten	Interpretation der Ergebnisse oft schwieriger

Überwachtes Lernen vs. Unüberwachtes Lernen Aufgabe, 15 min.

- Welche Methoden verwenden folgende Modelle:
  - GPT – Modelle
  - Claude – Modelle
  - Llama – Modelle
  - Mistral – Modelle
  - Gemini – Modelle

Das realistische Bild eines ML-Lebenszyklus innerhalb einer durchschnittlichen Organisation



Datenverständnis

Hauptaufgaben

1. Sammeln Sie erste Daten
2. Daten beschreiben
3. Daten erkunden
4. Datenqualität prüfen

Datenverständnis

Aufgabe 1: Sammeln von Anfangsdaten

- Beschaffen Sie sich die in den Projektressourcen aufgeführten Daten (oder den Zugang zu den Daten).
- Diese erste Datenerfassung umfasst das Laden von Daten, sofern dies für das Datenverständnis erforderlich ist.
  - Wenn Sie beispielsweise ein bestimmtes Tool zur Datenanalyse verwenden, ist es sinnvoll, Ihre Daten in dieses Tool zu laden. Dies kann erste Schritte zur Datenaufbereitung nach sich ziehen.
  - Hinweis: Wenn Sie mehrere Datenquellen verwenden, stellt die Integration ein zusätzliches Problem dar, entweder hier oder in der späteren Datenaufbereitungsphase.

Datenverständnis

Ergebnis von Aufgabe 1: Sammeln von Anfangsdaten

- **Erster Datenerfassungsbericht (Datenkatalog)**
  - Listen Sie die erfassten Datensätze zusammen mit ihren Speicherorten, den Methoden, mit denen sie erfasst wurden, und allen aufgetretenen Problemen auf.
  - Dokumentieren Sie aufgetretene Probleme und deren Lösungen. Dies erleichtert die zukünftige Wiederholung dieses Projekts oder die Durchführung ähnlicher Projekte.

Datenverständnis

Aufgabe 2: Daten beschreiben

- **Daten beschreiben**
  - Untersuchen Sie die „groben“ oder „oberflächlichen“ Eigenschaften der erfassten Daten und berichten Sie über die Ergebnisse.

Datenverständnis

Ergebnis von Aufgabe 2: Daten beschreiben

- **Datenbeschreibungsbericht (Datenkatalog)**
  - Beschreiben Sie die erhobenen Daten, einschließlich:
    - das Format der Daten,
    - die Datenmenge (zum Beispiel die Anzahl der Datensätze und Felder in jeder Tabelle),
    - die Identitäten der Felder,
    - und alle anderen Oberflächenmerkmale, die entdeckt wurden.
  - Prüfen Sie, ob die erhobenen Daten die relevanten Anforderungen erfüllen.

Datenverständnis

Aufgabe 3: Daten erkunden

- **Daten erkunden**
  - Diese Aufgabe befasst sich mit Fragestellungen des Data Science mithilfe von Abfrage-, Visualisierungs- und Berichtstechniken. Dazu gehören:
    - Verteilung der Schlüsselattribute (zum Beispiel des Zielattributs einer Vorhersageaufgabe)
    - Beziehungen zwischen Paaren oder kleinen Anzahlen von Attributen,
    - Ergebnisse einfacher Aggregationen,
    - Eigenschaften signifikanter Teilpopulationen,
    - und einfache statistische Analysen

Datenverständnis

Ergebnisse von Aufgabe 3: Daten erkunden

- **Datenexplorationsbericht**
  - Beschreiben Sie die Ergebnisse dieser Aufgabe, einschließlich erster Erkenntnisse oder anfänglicher Hypothesen und deren Auswirkungen auf den weiteren Projektverlauf.
  - Fügen Sie gegebenenfalls Grafiken und Diagramme hinzu, um Datenmerkmale aufzuzeigen, die eine weitere Untersuchung interessanter Datenteilmengen nahelegen.

Datenverständnis

Aufgabe 4: Datenqualität überprüfen

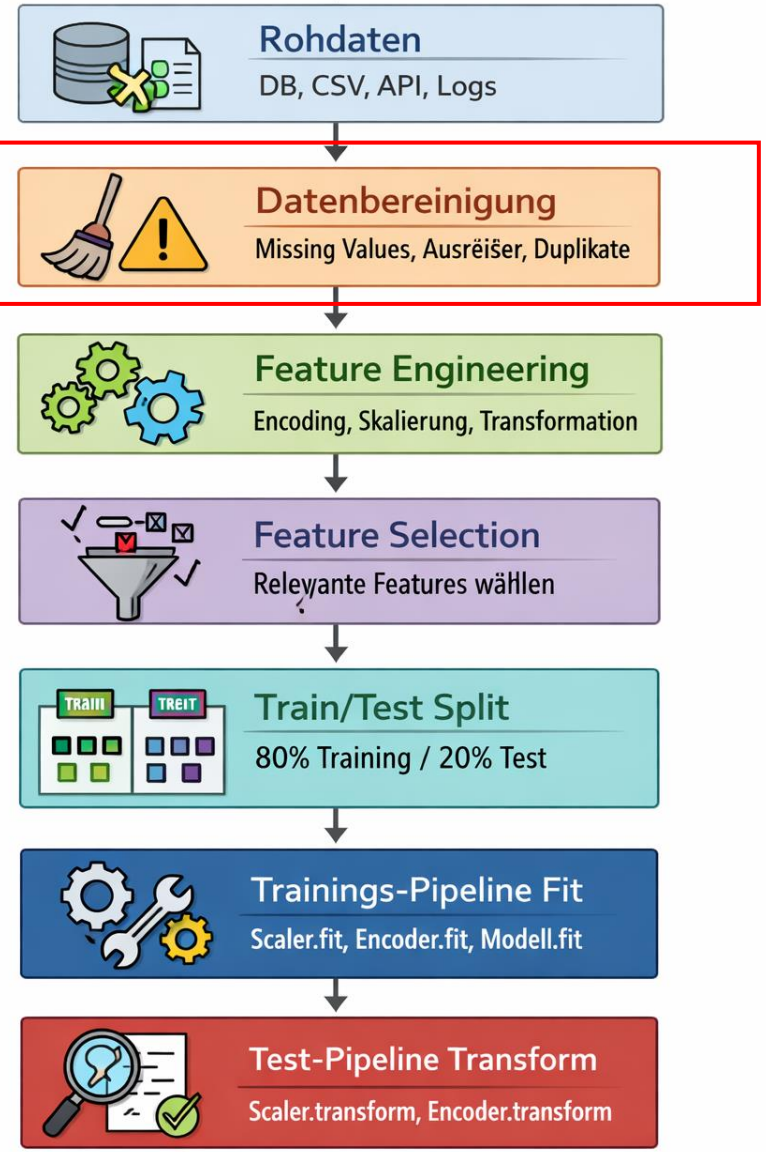
- Prüfen Sie die Qualität der Daten und beantworten Sie dabei Fragen wie:
  - Sind die Daten vollständig (decken sie alle erforderlichen Fälle ab)?
  - Ist es korrekt, oder enthält es Fehler, und falls es Fehler gibt, wie häufig sind diese?
  - Gibt es fehlende Werte in den Daten?
  - Falls ja, wie werden sie dargestellt, wo kommen sie vor und wie verbreitet sind sie?

Datenverständnis

Ergebnis von Aufgabe 4: Datenqualität überprüfen

- Datenqualitätsbericht
  - Listen Sie die Ergebnisse der Datenqualitätsprüfung auf;
  - Falls Qualitätsprobleme bestehen, listen Sie mögliche Lösungen auf.
    - Die Lösung von Datenqualitätsproblemen hängt im Allgemeinen stark von Daten- und Geschäftskenntnissen ab.

ML Datenpipeline



Datentypen

- pandas, die bevorzugte Library zum Datenhandling, bietet Datentypen an, welche zumeist auf NumPy basieren, wie
  - float
  - Int
  - Bool
  - string
- für die Tätigkeiten in Feature Engineering und Machine Learning ist primär die folgende Unterscheidung relevant:
  - numerical (num: int und float)
  - not numerical (string, im Folgenden kategorial)
- bool wird intern als int dargestellt
  - 0 = False
  - 1 (und alle anderen) = True

Skalierung

Bezeichnung DA	Bezeichnung ML	Bedeutung	Dtype	Ops
nominal	kategorial (ev. nominal)	ungeordnete Gruppen	num/str	=,≠
ordinal	kategorial (ev. ordinal)	geordnete Gruppen	num/str	<,≤,≥,>
intervall	metrisch	Zahlen ohne def. Nullpunkt	num	+,-
ratio	metrisch	Zahlen mit def. Nullpunkt	num	*,/

- Beispiele
- - nominal: Geschlecht, Religion, Augenfarbe
- - ordinal: Grössenklasse, Einkommensklasse, Schulnoten (!)
- - intervall: Temperatur in Grad Celsius
- - ratio: Temperatur in Grad Kelvin

Skalierung

- die Unterscheidung zwischen intervall und ratio bei Metrischen Daten spielt im ML keine Rolle
- dagegen kann bei numerischen Variablen aufgrund der Verteilung zusätzlich folgende Differenzierung angebracht sein
  - stetig (meist Datentyp float)
  - diskret: (meist Datentyp int), verschiedene Interpretationsmöglichkeiten
    - Zählwerte
    - numerische Codierungen von Kategorialen Merkmalen
- (dies wäre im Zweifelsfall mit dem Fach abzuklären)

EDA: Exploratyve Datenanalyse

- EDA ist
  - ein systematischer Ansatz zur Untersuchung von Datensätzen,
  - um Hauptmerkmale, Muster, Zusammenhänge und Ausreisser, ohne vorher festgelegte Hypothesen zu identifizieren

EDA: Ziele und Methoden

- Mustererkennung: Daten auf verborgene Strukturen und Korrelationen untersuchen
- Datenbereinigung: Anomalien, fehlende Werte und Ausreisser identifizieren
- Visualisierung: Nutzung von Grafiken, um Datensätze schnell zu erfassen
- Hypothesengenerierung: Daten erkunden, um Fragen zu entwickeln, die später getestet werden können.

EDA: Explorative Datenanalyse

- Anomalien: Datenpunkte, die stark vom Normalverhalten abweichen
- Mögliche Anomalien
  - Ausreisser: einzelne Datenpunkte, die signifikant von Rest abweichen. Z.B.: Eine extrem hohe Kreditkartenbuchung
  - Kontextbezogene Anomalien: Daten, die nur in einem bestimmten Kontext ungewöhnlich sind. Z. B.: Körpergrösse, Anstieg vom Energieverbrauch während der Ferien
  - Kollektive Anomalien: Eine Gruppe von Datenpunkten, die gemeinsam abweichen, auch wenn sie einzeln normal wirken. Z. B.: Mehrere IP-Adressen greifen gleichzeitig auf einen Server zu (DDoS-Angriff)

EDA: Explorative Datenanalyse

- Anomalien: Datenpunkte, die stark vom Normalverhalten abweichen
- Mögliche Anomalien
  - Anomalien in den Datenbanken:
    - Änderungsanomalie (Update): Daten werden nicht überall aktualisiert
    - Löschanomalie (Delete): Das Löschen der Daten führt zum Datenverlust
    - Einfügeanomalie: (Insert): Ein Datensatz kann nicht eingefügt wertden, weil notwendige Teilinformationen (Primärschlüssel) fehlen
  - Anomalien der IT-Systeme:
    - Unerwarteter Anstieg des Netzverkehrs
    - Häufige Verbindungsabbrücke
- Ziel der Bereinigung: Verbesserte Datenqualität, Validierung von Hypothesen und Vorbereitung für maschinelles Lernen

EDA: Explorative Datenanalyse

- mögliche Anomalien nach Variablenart
  - nicht numerische Daten:
    - Fehlende Werte
    - Duplikate
    - Kategoriale Variablen:
      - hohe Kardinalität: Viele eindeutige Werte (z.B.: Zeitstempel, KundenID)
      - nicht balancierte Daten
  - numerische Variablen
    - schiefe Verteilungen
    - Ausreisser
    - Korrelationen
    - diskrete Werte mit geringer Kardinalität: Wenige eindeutige Werte (z. B.: Geschlecht)

Klassierung

- Die Klassierung in der deskriptiven Statistik ordnet viele, unterschiedliche Rohdaten in wenige, überschaubare Klassen (Intervalle) ein.
- Dies erleichtert die Übersicht, Analyse und grafische Darstellung (z. B. Histogramme),
- führt jedoch zu Informationsverlust und geringerer Messgenauigkeit.
- Typische Darstellungen sind Häufigkeitstabellen mit Klassenbreiten und Klassenhäufigkeiten.

Klassierung

- **Zweck:** Reduzierung der Datenkomplexität, um Muster und Trends in großen Datensätzen erkennbar zu machen
- **Vorgehen:** Festlegung von Klassengrenzen
  - (z. B. [0;2),[2;4) [0;2),[2;4)),
  - wobei die Untergrenze meist geschlossen (inklusive) und die Obergrenze offen (exklusive) ist
- **Darstellung:** Ergebnisse werden häufig in Histogrammen dargestellt, bei denen die Fläche der Balken die Häufigkeiten repräsentiert.
- **Nachteile:** Durch die Gruppierung geht die exakte Messgenauigkeit verloren, da Einzelwerte nicht mehr erkennbar sind

Datenverständnis: Beispiel Klassierung

Rohe Daten, sortiert

1500	1500	1500	1500	1500
1600	1600	1600	1900	1900
1900	2000	2000	2000	2000
2100	2200	2300	2400	2500
2500	2700	2800	2900	2900
3000	3200	3300	3500	3500

Klassierte Daten

class	Interval	number	middle of the class	summ	% of total
1	[1500, 2000)	11	1750	18000	28
2	[2000, 2500)	9	2250	17000	26
3	[2500, 3000)	6	2750	16700	26
4	[3000, 3500)	5	3250	13500	21

Einsatz EDA für Uce Case Generierung