Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

2024 FS CAS PML
1 Feature Engineering
1.6 Implementation

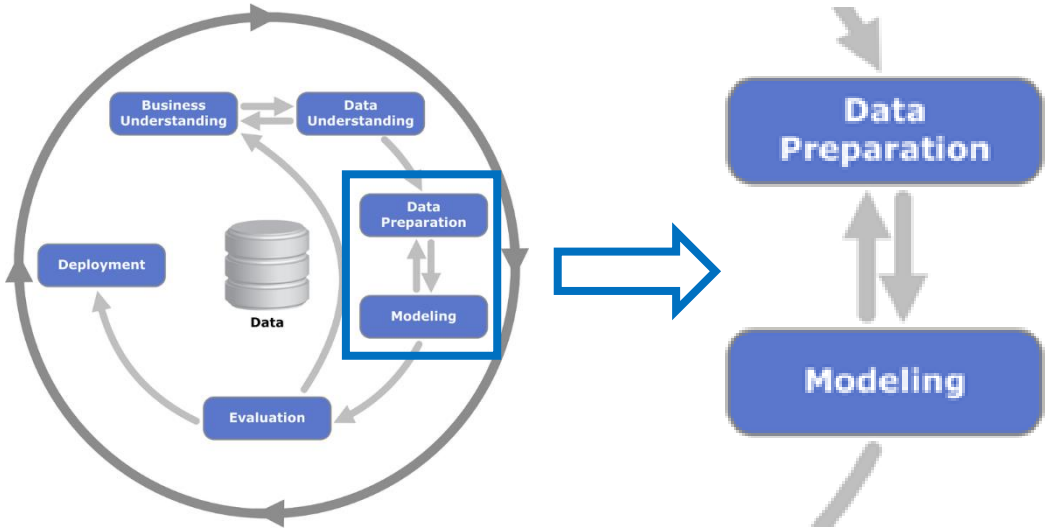
Werner Dähler 2024

1 Feature Engineering - AGENDA

11. Einführung
12. Exploration
13. Transformation
14. Konstruktion
15. Selektion
16. Implementation
17. Nachträge

1.6 Feature Engineering - Implementation

- die bisherigen Tätigkeiten können aus der Sicht als Data Understanding resp. Vorbereiten von Preprocessing verstanden werden
- aus praktischen Gründen kann es angebracht sein, das effektive Preprocessing in einen separaten Task auszulagern
- Begründung: der Doppelpfeil zwischen Data Preparation und Modelling weist darauf hin, dass im Rahmen der Modellierung, resp. Modelloptimierung neue Erkenntnisse eine Modifikation des Preprocessing nötig machen kann



- Best Practice in diesem Zusammenhang:
  - sammeln der Empfehlungen aus bisherigen Tätigkeiten
  - erstellen eines Drehbuchs für alle notwendigen Transformationen und in der richtigen Reihenfolge (Konfliktpotential!)
  - codieren eines einzigen Scripts (.py, .ipynb) zum Erstellen des für Machine Learning vorbereiteten Dataset

1.6 Feature Engineering - Implementation

- genanntes Vorgehen ist auch aus Sicht einer finalen Produktivsetzung (Deployment) eines trainierten Modells angebracht: Feature Engineering muss auf neuen Daten identisch durchgeführt werden

- beispielhaft zusammengestellt die Modifikationen, ausgehend von den Rohdaten des Bankkundendataset
- die einzelnen Schritte sind in den folgenden Blöcken organisiert:
  1. Modifikationen auf Data Frame
  2. Modifikationen kategoriale Variablen
  3. Modifikationen numerische Variablen
  4. Abschlussarbeiten inkl. Speichern des modifizierten Data Frames als CSV-File unter neuem Namen

1.6 Feature Engineering - Implementation

- das "Drehbuch" (Vorschlag)

Data Frame		
E1	Entfernen von Beobachtungen nach Bedingung	age > 100
E2	Entfernen von Duplikaten	
E3	Entfernen fragwürdiger Variablen	default, poutcome (ev. duration)
E4	Einsetzen von Werten für NAs	alle numerische: Median
		alle nicht numerische: Modalwert
Kategoriale Variablen		
E5	Reduzieren der Kardinalität	education: illiterate -> basic.4y
	Nummerisieren - Faktorisieren	hier keine
E6	Nummerisieren - Ordinal Encodieren	education, day_of_week, month
E7	Nummerisieren - Binär Encodieren	housing: no -> 0, sonst 1
		contact: cellular -> 1, sonst 0 (rename)
E8	Nummerisieren - Nominal Encodieren	alle jetzt noch nicht numerischen ausser y

1.6 Feature Engineering - Implementation

- das "Drehbuch" (Fortsetzung)

Numerische Variablen		
E9	logarithmieren	duration, campaign
E10	binär umcodieren	pdays: 999 -> 0, sonst 1
		previous: 0 -> 1 sonst 0
Andere Tätigkeiten		
	Konstruktion	unterbleibt hier
E11	Bereinigen der Variablennamen	
	Standardisieren	unterbleibt hier
E12	Speichern	als bank_data_prep.csv mit sep=',' (default)

- konkrete Implementierung: FE_6_Implementation.ipynb

1.6 Feature Engineering - Implementation

Workshop 03

Gruppen zu 3 bis 4, Zeit: 60'

- in Workshop 2 haben Sie das Melbourne Housing Dataset mit Sicht auf Supervised Learning untersucht und dabei erste Empfehlungen für Feature Transformation erarbeitet

- eine Zusammenstellung konsolidierter Empfehlungen finden Sie in `WS 03 Empfehlungen.xlsx`

- empfohlenes Vorgehen:
  - erstellen Sie eine Kopie des Notebooks zu Kap. 1.6 (`1.6 Feature Engineering - Implementation.ipynb`)
  - modifizieren Sie dies mit Sicht auf die neue Fragestellung
  - erstellen Sie vom Ergebnis ein neues Dataset unter einem anderen Namen, z.B. `melb_data_prep.csv`