Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

2024 FS CAS PML
1 Feature Engineering
1.5 Selektion

Werner Dähler 2024

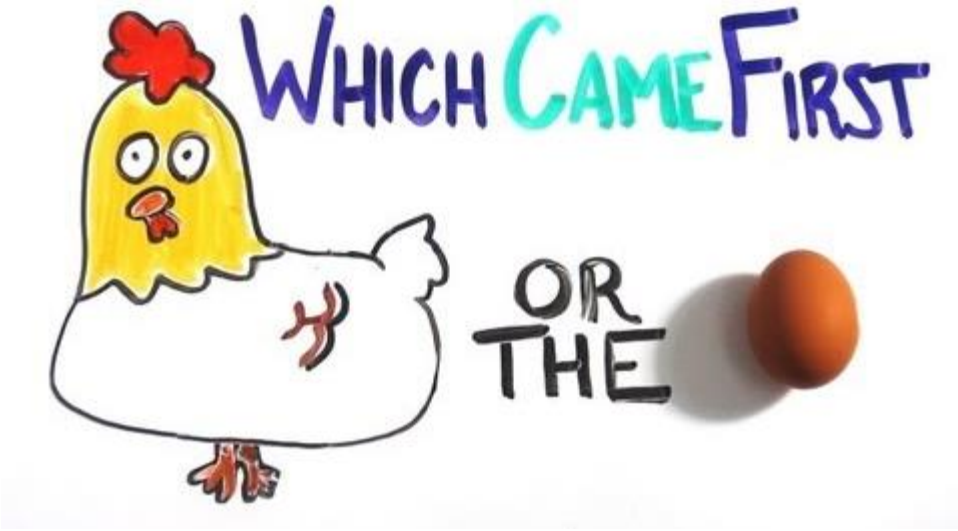
1 Feature Engineering - AGENDA

11. Einführung
12. Exploration
13. Transformation
14. Konstruktion
15. Selektion
    151. Automatische Selektion
    152. Selektion mit Expertenwissen
16. Implementation
17. Nachträge

1.5 Feature Engineering - Selektion

1.5.1 Automatische Selektion

- wie bereits erwähnt, setzt erfolgversprechendes Maschinelles Lernen vorgängig kompetentes Feature Engineering voraus
- andererseits basieren insbesondere FE Methoden wie die Beurteilung der Wichtigkeit einzelner Variablen auf ML Methoden
- wo also beginnen? (das Henne-Ei-Dilemma)



- daher werden einzelne Teile dieser Thematik vorerst nur rudimentär behandelt
- eine eingehendere Diskussion derselben erfolgt nach dem Kursteil zum Überwachten Lernen in 1.7 Feature Engineering - Nachträge

1.5 Feature Engineering - Selektion

1.5.1 Automatische Selektion

- unter diesem Thema werden drei Methodengruppen genannt
  1. Univariate Methoden
  2. Modellbasierte Selektion (Details folgen später)
  3. Iterative Selektion (Details folgen später)

1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

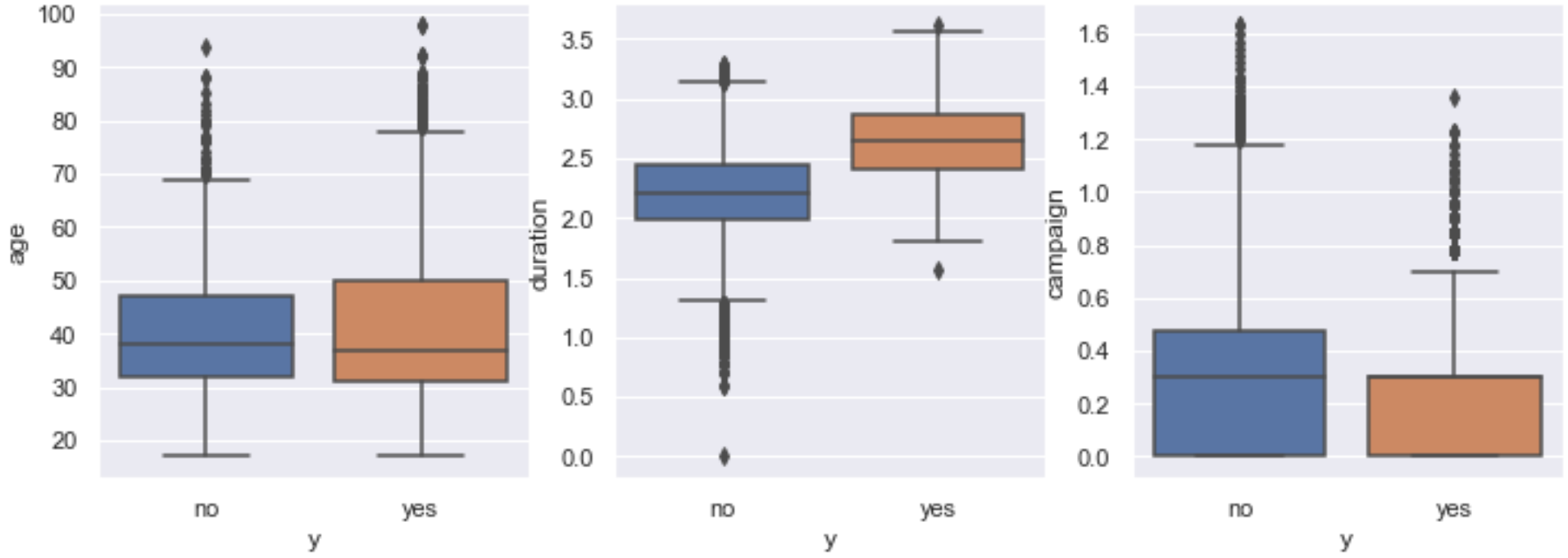
- die Methode untersucht die Wichtigkeit der Variablen (einzeln) in Bezug auf deren Differenzierbarkeit des Targets (Vorschau auf Supervised Learning)

- für die folgende Diskussion werden einige Vorarbeiten vorausgesetzt (vgl. [ipynb]):
  - neu laden
  - NAs entfernen
  - Features - Target - Split (mehr dazu später)
  - nur numerische Features

1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

- von drei ausgewählten Variablen ("age", "duration", "campaign") werden vorerst gruppierte Boxplots erstellt
  - Gruppierung: Wert des Targets (y)
  - "duration" und "campaingn" werden ausserdem im 10er Logarithmus dargestellt



1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

- "age" und "campaign" zeigen keine bemerkenswerten Unterschiede zwischen den zwei Gruppen, "duration" dagegen wohl (die beiden Boxen überschneiden sich mehrheitlich)

- eine Rückbesinnung auf CAS DA (für die, welche betroffen sind)
  - ANOVA (Analysis of Variance) untersucht den "Einfluss" einer oder mehrerer kategorialen Variablen(!) auf eine metrische Variable
  - der F-Wert (Statistik) ist grösser, je mehr die metrische Variable differenziert

- anova steht (unter anderem) als folgende Funktion in Python zur Verfügung: scipy.stats.f_oneway

- im beigelegten Code Beispiel ([ipynb]) wird der F-Wert für die oben genannten Variablen ermittelt

1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

- das Ergebnis:

```
F-Statistik
age      : 18.265609177886155
duration : 2950.21220692035
campaign : 161.66280010318036
```

  - grösster F-Wert für "duration"
  - kleinster für "age"

  was mit den obigen Boxplots korrespondiert

- im Folgenden eine Vorschau auf Modellieren mit `sklearn`

1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

- die hier verwendete Funktion SelectKBest sucht die n am besten trennenden Variablen und hinterlegt sie in einer Bool-Maske

```
from sklearn.feature_selection import SelectKBest
select = SelectKBest(k=5)
select.fit(X, y)
mask = select.get_support()
print(pd.DataFrame({
    'index': X.columns,
    'mask': pd.Series(mask)}))
```

	index	mask
0	age	False
1	duration	True
2	campaign	False
3	pdays	True
4	previous	False
5	emp.var.rate	True
6	cons.price.idx	False
7	cons.conf.idx	False
8	euribor3m	True
9	nr.employed	True

- danach kann mask verwendet werden, um die Variablen auf die oben identifizierten einzuschränken, z.B. so:

```
X_red = X.loc[:, mask]
```

1.5 Feature Engineering - Selektion

1.5.1.1 Automatische Selektion - Univariate Methoden

- diese Methode untersucht also den Einfluss einzelner Variablen auf das Target (oder umgekehrt)
- mögliche Interaktionen (Korrelationen, lineare Abhängigkeiten) werden dabei **nicht** berücksichtigt
- dieser Makel kann durch ML basierte Methoden behoben werden
- daher werden die folgenden zwei hier kurz vorgestellt und dann **nach** dem Kursteil Supervised Learning noch vertiefter diskutiert
  - Modellbasierte Selektion
  - Iterative Selektion

1.5 Feature Engineering - Selektion

1.5.1.2 Automatische Selektion - Modellbasierte Selektion (Ausblick)

- einige der Trainingsmethoden, welche im Kursteil Überwachtes Lernen vorgestellt werden, beurteilen zwecks Modellbildung intern die Wichtigkeit der einzelnen beteiligten Variablen
- diese Informationen können mit geeigneten Methoden auch aus den trainierten Modellen extrahiert werden, um im Nachhinein als Feature Importance zu beurteilen
- Methoden, welche hierzu zur Verfügung stehen
  - Klassifikation
    - Decision Tree
    - Random Forest
  - Regression
    - Lasso Regression
- mehr dazu dann unter dem Thema 1.7.2.1 Modellbasierte Selektion im Anschluss an Überwachtes Lernen

1.5 Feature Engineering - Selektion

1.5.1.3 Automatische Selektion - Iterative Selektion (Ausblick)

- im Gegensatz zu den Modellbasierten Methoden wird hier der Einfluss der einzelnen Variablen auf die Performance des Modells beurteilt
- dabei wird schrittweise jeweils die Variable mit dem geringsten Einfluss entfernt, bis die geforderte Anzahl Variablen erreicht ist
- diese Methode kann auf (praktisch) alle Trainingsmethoden angewendet werden

- mehr dazu unter dem Thema 1.7.2.2 Iterative Methoden im Anschluss an den Kursteil zum Überwachten Lernen

1.5 Feature Engineering - Selektion

1.5.2 Selektion mit Expertenwissen

- für Feature Engineering (ins. F. Selektion) kann auch Expertenwissen eine wichtige Rolle spielen
- während beim Einsatz einiger Machine Learning Algorithmen auch Information zur Wichtigkeit der einzelnen Variablen anfällt, sollte das Potential von Expertenwissen nicht unterschätzt werden
- oft können Business Experten dazu beitragen, sinnvolle Variablen zu identifizieren, welche klar informativer sind als die ursprüngliche Datenrepräsentation