ANGEWANDTE PROGRAMMIERUNG

Einführung Machine Learning

Vorlesung 06

Dennis Glüsenkamp

29. April 2022

Daten-getriebene Entscheidungsunterstützung

Arten von Machine Learning

Live-Übungen mit generierten Daten

Rückbezug: Data Analysis Lifecycle

Daten-getriebene Entscheidungsunterstützung

Daten und Entscheidungen

- Entscheidungsprozesse können durch die Nutzung von Daten unterstützt werden
- Organisatorische Entscheidungen werden somit nicht allein aufgrund von Intuition getroffen
- Heute verfügbare Daten(mengen) und analytische Tools erweitern die traditionellen Möglichkeiten
- *Data-driven decision-making (DDDM)* zeichnet sich daher durch folgende Aspekte aus:
  - Validierung von Handlungen und Maßnahmen über Datenanalytik
  - Transformierung von Ergebnissen in handlungsorientierte Erkenntnisse
  - Nutzung von Daten, Statistiken und Metriken um hinsichtlich Geschäftszielen zu steuern

Wege und Ansätze zu DDDM

- Fragestellung und geschäftliche Zielsetzung bestimmen den Ansatz des Lösungswegs
- Zielsetzungen der Analyse können sein:
  - Aggregationen von Daten für Erfassung des Gesamtzusammenhangs $\rightarrow$ *BI*
  - Ableitung von statistischen Erkenntnissen, Erkennung von Mustern und Zusammenhängen $\rightarrow$ *Data Science*
  - Erstellung von Algorithmen zur Klassifizierung, Verarbeitung und Prädiktion $\rightarrow$ *Machine Learning*
- Begrifflichkeiten können je nach Standpunkt, Schwerpunkt oder Fachrichtung auch anders definiert oder eingesetzt werden

Arten von Machine Learning

Notation und Bezeichnungen (1/2)

- Im Rahmen von Machine Learning arbeiten wir mit relationen Daten
- Verschiedene Bezeichnungen kommen dabei in der Beschreibung der Daten vor
- **Objekte:**
  - Entitäten in der beobachteten Situation
  - Beispiele: Personen, Produkte, Ereignisse
  - Synonyme: Beobachtungen, Zeilen
- **Attribute:**
  - Eigenschaften eines Objekts
  - Beispiele: Körpergröße, Farbe, Eintrittsdatum
  - Synonyme: Features, Spalten, Datenfelder, *Input*

Notation und Bezeichnungen (2/2)

- **Labels:**
  - Besondere Ausprägung eines *Attributs*, dass eine Zugehörigkeit oder ein Ergebnis widerspiegelt
  - Beispiele: Testergebnis, Preis, Zustand
  - Synonym: Target, *Output*
- Benennung hier mit folgender Notation
  - alle relationale Daten ohne Labels: X
  - alle Labels: y

Supervised Learning - überwachtes Lernen

- Objekte im Datensatz besitzen ein Label
- Ziel ist das Lernen einer Regel, die den Input auf den Output abbildet
- Regel sollte so gewählt sein, dass der Fehler der Abbildung minimiert ist
- Gelingt eine hinreichend gute Minimierung, dann *generalisiert* der Algorithmus das Problem
- Algorithmus kann nachfolgend den Output von neuen Beobachtungen (ohne Label) prognostizieren

	f_1	...	y
X_1	1	...	A
X_2	2	...	B
...	...	...	...
X_n	3	...	C

Unsupervised Learning - unüberwachtes Lernen

- Objekte im Datensatz besitzen kein Label
- Ziel ist das Erkennen von strukturellen Zusammenhängen zwischen den Objekten
- Im Regelfall kein absolutes Ergebnis, da von Parametrisierung abhängig
- Kann auch zur Generierung von Labeln eingesetzt werden $\rightarrow$ *Feature Learning*

	f_1	...	f_n
X_1	1	...	A
X_2	2	...	B
...	...	...	...
X_n	3	...	C

Weitere Kategorien

- **Reinforcement Learning:**
  - Eigene, bedeutende Kategorie
  - Optimierung findet in dynamischem Umfeld statt
  - Algorithmus erhält Feedback in Form von Belohnungen
  - Belohnung soll maximiert werden
- Semi-supervised Learning
- Anomaly Detection
- Association Rule Learning

Auswahl von Modellen/Algorithmen

- Regression Analysis
- Entscheidungsbaum (Decision Tree)
- Random Forest
- Support Vector Machine
- Neuronales Netz (Neural Network)
- Clustering

Live-Übungen mit generierten Daten

Confusion-Matrix

- Gütebestimmung bei Klassifikationsproblemen
- Auf beliebige Anzahl von Klassen erweiterbar
- Kennzahlen für Performance direkt ableitbar:
  - Accuracy $a = \frac{TP+TN}{TP+TN+FP+FN}$
  - Recall/Sensitivity $r = \frac{TP}{TP+FN}$
  - Specificity $s = \frac{TN}{TN+FP}$
  - Precision $p = \frac{TP}{TP+FP}$
  - ...

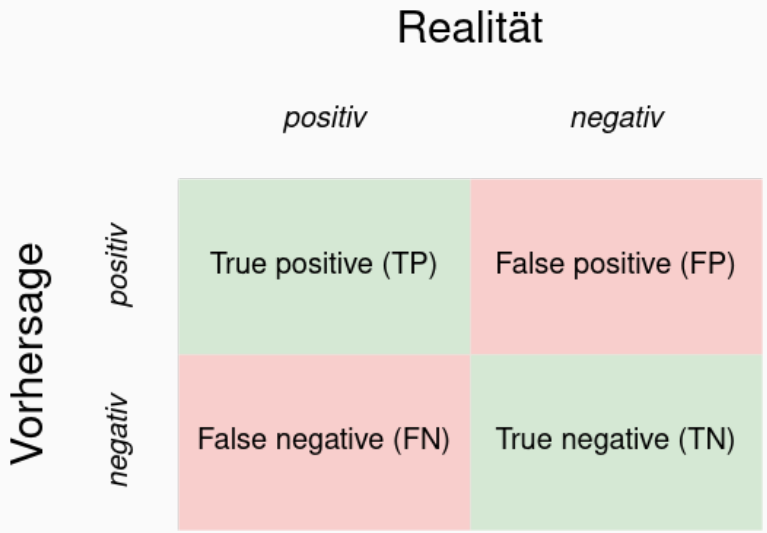

Figure 1: Confusion-Matrix

Rückbezug: Data Analysis Lifecycle

CRISP-DM Life Cycle

- **CR**oss-**I**ndustry **S**tandard **P**rocess for Data Mining [1]
- 1996 im Rahmen von EU-Förderprojekt entwickelt
- Offenes, freies Prozessmodell zur Durchführung von Data Mining Vorhaben
- Prozess kann flexibel und unabhängig von Branche, Toolset und Anwendung verwendet werden

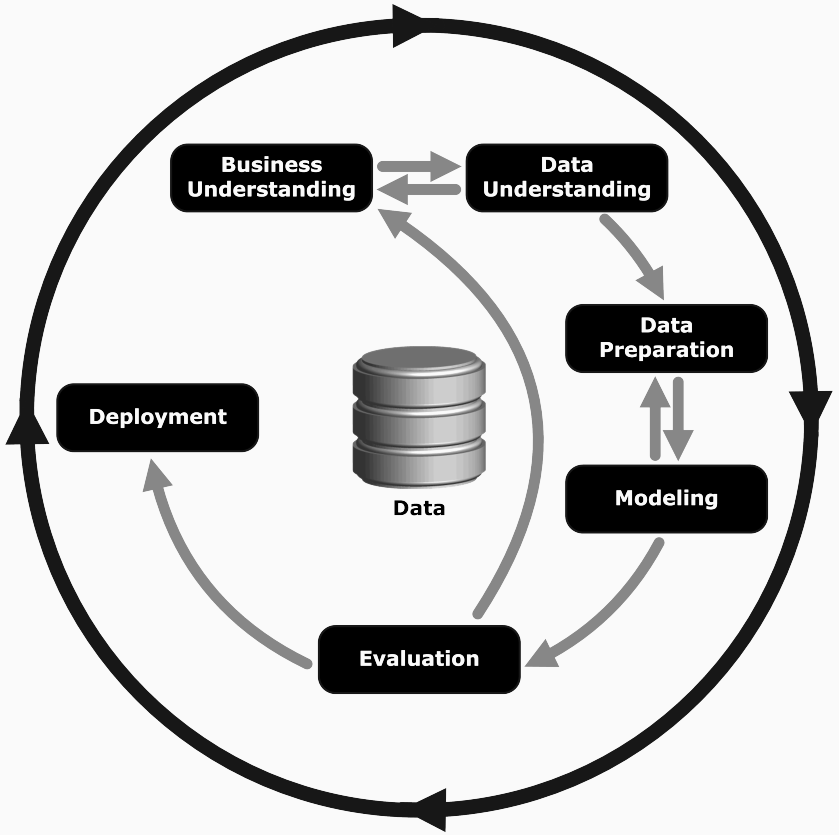

Figure 2: CRISP-DM Prozessmodelldiagramm (Quelle: Kenneth Jensen, CC BY-SA 3.0)

Phasen von CRISP-DM (1/6)

1. **Business Understanding:**
   - Formulierung von konkreten Fragestellungen und Zielen
   - Abgleich von Aufgaben und Erwartungen
   - Vereinbarung eines Vorgehens/einer Planung
   - Identifikation von wichtigen Einflussfaktoren
   - Verständnis des Geschäftsmodells
   - Definition von Erfolgskriterien

Phasen von CRISP-DM (2/6)

2. Data Understanding:

- Betrachtung des Datenbestands
- Auswertung der Datenverfügbar-keit, -reliabilität, -qualität
- (Statistische) Auffälligkeiten in den Daten
- Abstimmung zum Datenschutz

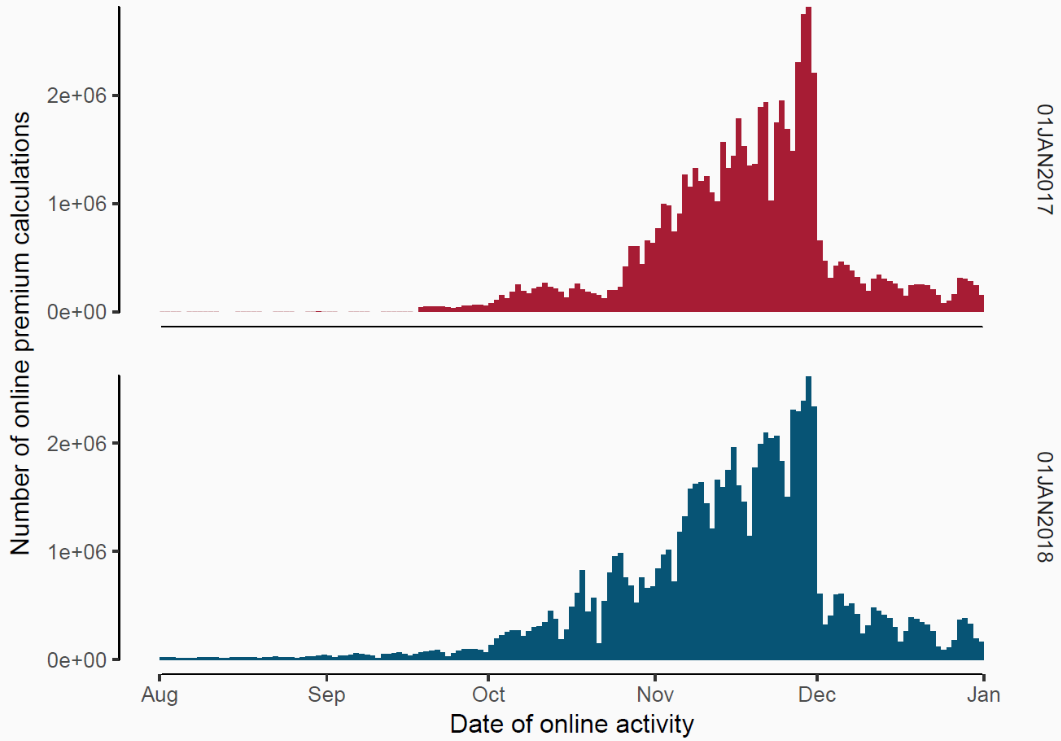

Figure 3: Anzahl von Preisanfragen für Kfz-Versicherungen über Aggregator- bzw. Vergleichswebsites bei einem deutschen Versicherer [2]

Phasen von CRISP-DM (3/6)

3. Data Preparation:

- Datenbereinigung und Transformationen
- Datenverknüpfung und -aggregation
- Feature Engineering
- Feature Selection

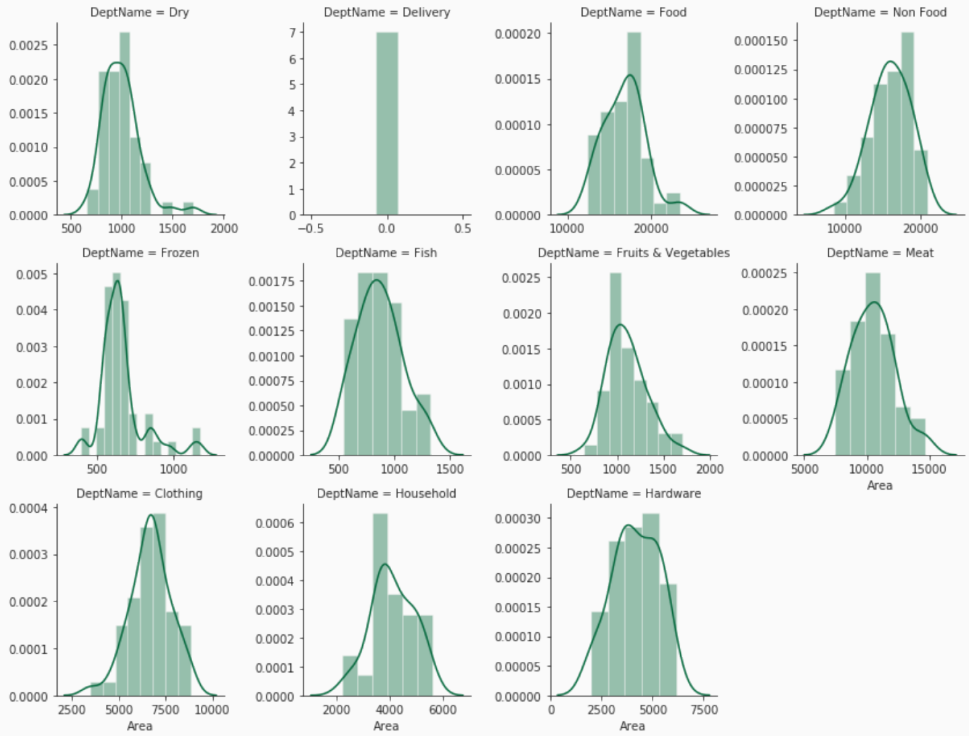

Figure 4: Verteilung von Verkaufsflächen von verschiedenen Märkten eines fiktiven Handelskonzerns, getrennt nach Fachabteilungen [3]

Phasen von CRISP-DM (4/6)

4. Modeling:

- Definition der Annahmen und Rahmenbedingungen der Modellierung
- Auswahl von geeigneten Algorithmen
- Test Design
- Training des Modells
- Tiefgreifende, zielgerichtete Datenexploration

Phasen von CRISP-DM (5/6)

5. Evaluation:

- Vergleich der verschiedenen Modelle anhand von Gütekriterien
- Betrachtung der Interpretierbarkeit des Modells
- Kritische Analyse des Modellierungsprozesses
- Abgleich mit (wirtschaftlichen) Erfolgskriterien
- Definition von Folgeaktivitäten

Phasen von CRISP-DM (6/6)

6. Deployment:

- Kommunikation der Ergebnisse
- Integration des Modells in die Systemlandschaft und Entscheidungsprozesse
- Wartung und Pflege des Modells
- Dokumentation der Erkenntnisse und Funktionsweise

Referenzen

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9, 13.

[2] Gluesenkamp, D. (2018). Prediction of customer churn with premium online calculation data in insurance business. DeMontfort University, Leicester, United Kingdom.

[3] Gluesenkamp, D. (2019). Wrangling and cleansing business data. Retrieved from https://dgluesen.github.io/wrangling-sales-workload/