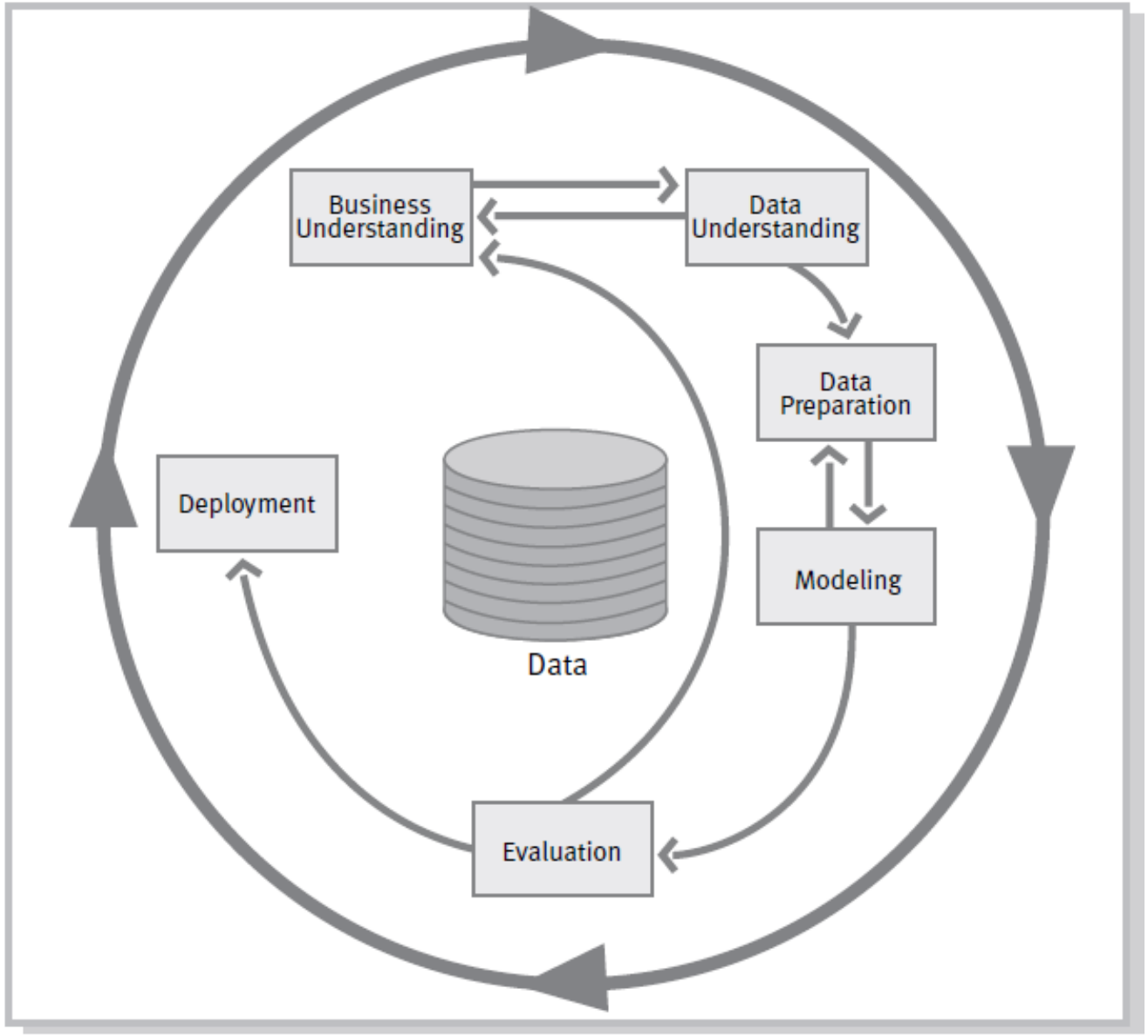
Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Maschinelles Lernen: Interpretierbarkeit

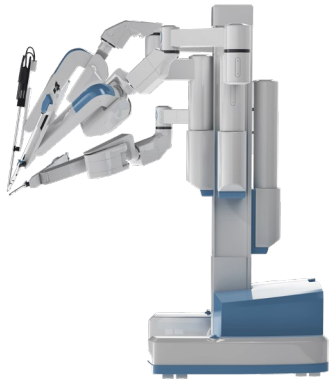


– Hans Friedrich Witschel, Andreas Martin

Interpretierbarkeit: Gedankenexperiment

- Stellt euch vor: ihr habt einen Tumor, der operiert werden muss
- Ihr habt 2 Optionen:

85% Überlebenschance



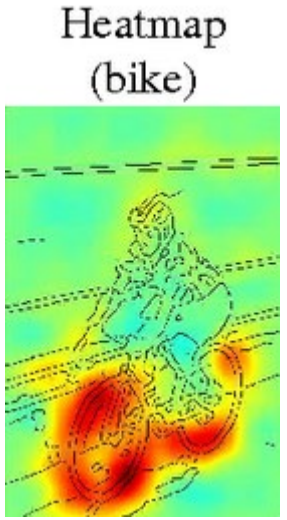
98% Überlebenschance

- Was wählt ihr?

Argumente für Interpretierbarkeit

1. Debugging, z.B.
   - Data Leakage erkennen (Verwendung von Informationen, die zum Zeitpunkt der Vorhersage unbekannt sind)
   - Umgang mit «Ausnahmefällen»

> Machine learning models take on real-world tasks that require **safety measures** and testing. Imagine a self-driving car automatically detects cyclists based on a deep learning system. You want to be 100% sure that the abstraction the system has learned is error-free, because running over cyclists is quite bad. An explanation might reveal that the most important learned feature is to recognize the two wheels of a bicycle, and this explanation helps you think about edge cases like bicycles with side bags that partially cover the wheels.



   - Unerwünschten Bias erkennen (Bsp.: Modell bevorzugt männliche Bewerber)
2. Vertrauen: Menschen sind meist verantwortlich für Entscheidungen
3. (Wissenschaftliche) Erkenntnis, die aus Mustern gezogen werden kann

Arten von Interpretierbarkeit

Intrinsisch (Tree, LR, …) ↔ post-hoc

modellspezifisch ↔ modell-agnostisch

lokal ↔ global

Intrinsisch interpretierbare Modelle

Lineare Modelle
(Lineare / logistische Regression, Naive Bayes, GLM, GAM)

Vorhersage = gewichtete Summe (von Log.)

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- Gut untersucht
- GAMs können nicht-lineare Effekte erfassen
- „geschmeidige“ Handhabung numerischer Features

- Wechselwirkungen zwischen Merkmalen werden nicht erfasst
- Erläuterung der Gewichte: geht immer davon aus, dass andere Merkmale unverändert sind
- Erweiterungen (z. B. GAMs) sind weniger interpretierbar

Logikbasierte Modelle
Entscheidungsbäume / -tabellen, Regellerner)

Vorhersage = Menge logischer (Wenn-dann-)Regeln

- Wirklich einfache Interpretation
- Erfasst die Interaktionen zwischen den Merkmalen

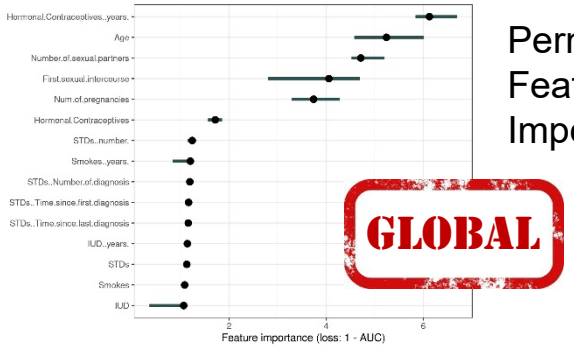
- diskretisiert immer numerische Merkmale
- Instabil
- Grosse Bäume sind weniger interpretierbar

Nutzen bei: vielen numerischen Features

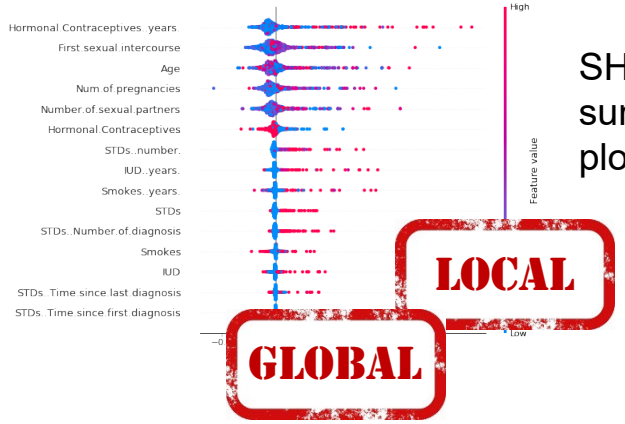
Nutzen bei: multivariaten Mustern

Interpretierbarkeit: Modell-agnostische post-hoc Ansätze

- Wie wichtig sind einzelne Features?

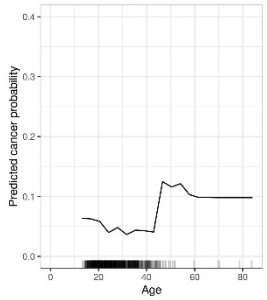


Permutation Feature Importance

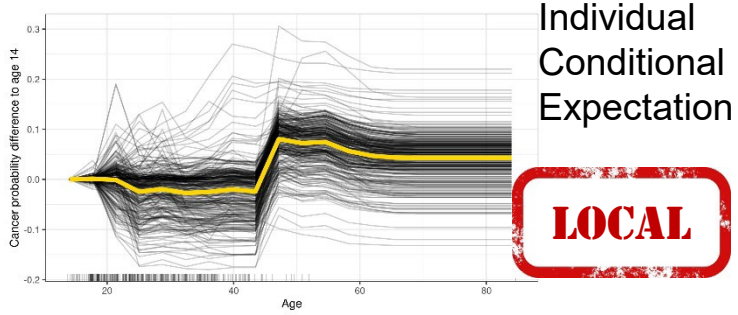


SHAP summary plot

- Wie beeinflussen die Attributwerte die Vorhersage?



Partial Dependence Plot



Individual Conditional Expectation

Ideen hinter den Ansätzen: Permutation Feature Importance

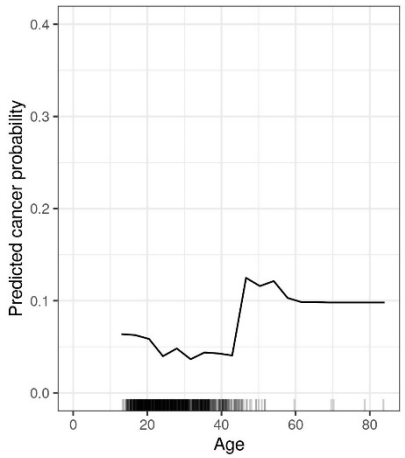
Idee:
Wichtigkeit eines Attributs = Anstieg des Fehlers, wenn die Information des Attributs durch Vertauschung der Attributwerte zwischen Instanzen zerstört wird

- Permutation zerstört auch die Interaktion eines Features mit anderen
  → Interaktionen, d.h. die Wichtigkeit von Features in Abhängigkeit von anderen wird berücksichtigt

- Benötigt Zugang zu den Labels der Daten
- Manchmal instabile (wegen randomisierter Permutationen)
- Permutationen können unrealistische Instanzen erzeugen (z.B. 2m grosse Person, die 30kg wiegt) → nicht unbedingt schlimm…
- Hinzufügen von Features, die mit vorhandenen stark korrelieren, verringern deren Wichtigkeit teils drastisch (nicht immer intuitiv…)

Ideen hinter den Ansätzen: Partial Dependence Plots

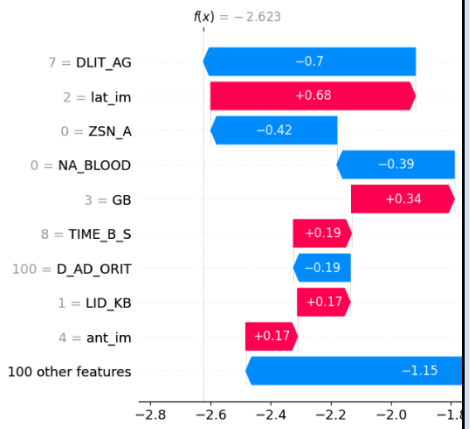
Idee:
Wert der PDP-Funktion für einen Attributwert = Durchschnittlicher Vorhersagewert des Modells, wenn alle Instanzen diesen Wert hätten



- intuitive Interpretation
- Kann als kausale Beziehung zwischen Feature-Wert und Vorhersage interpretiert werden (die laut dem Modell existieren!!!)

- Nur für 1 bis max. 2 Features
- Man sollte die Verteilung der Werte (an der x-Achse) in Betracht ziehen
- Ignoriert Korrelationen und Interaktionen von Features
- Bildung des Durchschnitts kann stark divergierende Werte verdecken (siehe ICEs)

Ideen hinter den Ansätzen: Shapley-Werte (lokal)



Idee:
Die Attributwerte einer Instanz betreten in zufälliger Reihenfolge einen Raum. Alle Attributwerte tragen zur Vorhersage bei.

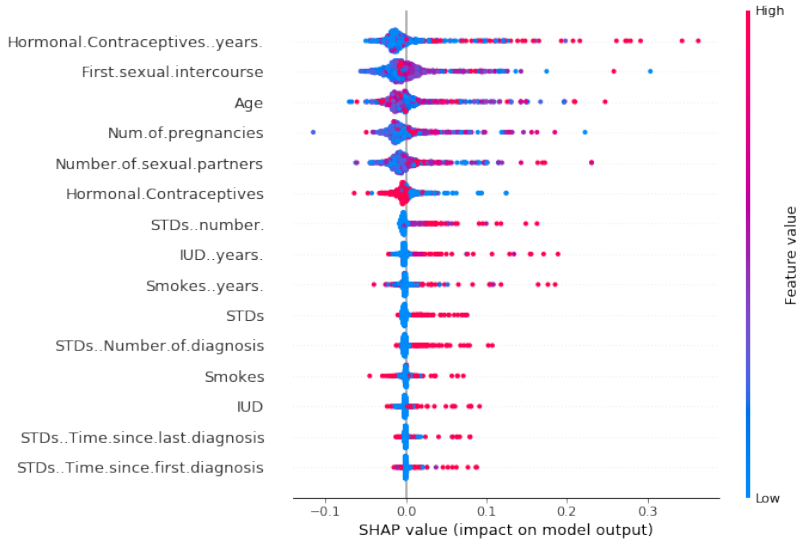
Shapley-Wert eines Attributwerts = durchschnittliche Änderung der Vorhersage der "Koalition", wenn das Feature den Raum betritt.

- Solide Theorie, kann vollständige Erklärungen liefern (im Gegensatz z.B. zu LIME)

- Rechenintensiv
- Nicht so einfach zu interpretieren
- Die Interpretation muss alle Merkmale berücksichtigen (und Menschen mögen kurze Erklärungen!!)

Genauere Erklärung:
https://www.youtube.com/watch?v=VB9uV-x0gtg
(bis 02:55 oder weiter…)

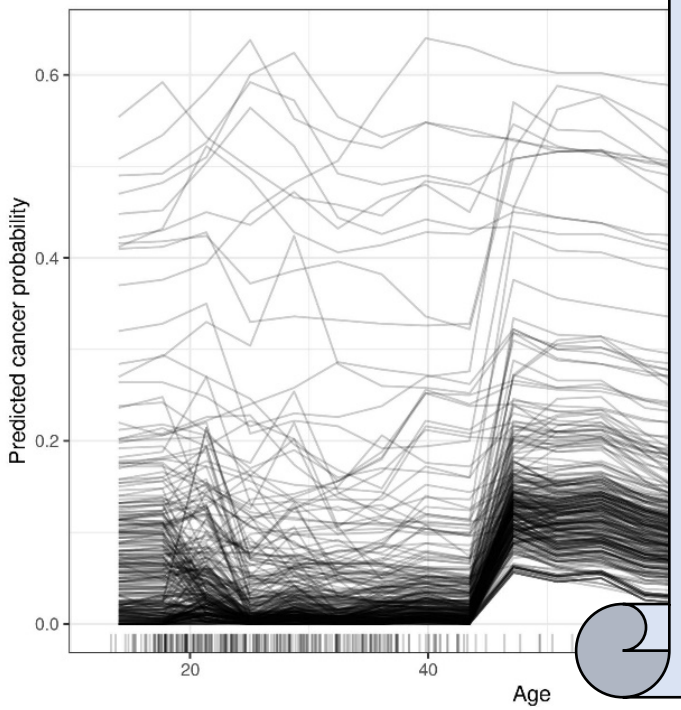
Ideen hinter den Ansätzen: SHAP Summary Plot (global)



Idee:
Jeder Punkt ist ein Shapley-Wert für eine Instanz und ein Attribut. Attribute sind nach Wichtigkeit absteigend sortiert, Attributwerte farblich kodiert (rot = grosse Werte). Werte auf der x-Achse entsprechen der Vorhersage(wahrscheinlichkeit)

Ideen hinter den Ansätzen (3)

– Individual Conditional Expectation



Idee:
Zeigt eine Linie pro Instanz, aus der man jeweils sieht, wie sich die Vorhersage für diese Instanz ändert, wenn der Attributwert (x-Achse) sich ändert.

- Zeigt mehr Details (Verteilung!) als PDPs
- Ziemlich intuitiv

- Nur 1 Feature darstellbar
- Bildet keine Interaktionen/Korrelationen ab

Interpretierbarkeit: Aufgabe 1. Teil

- **Hinweis**: Installiert zunächst das Add-on «Explain» in Orange und schaut euch die dort verfügbaren Widgets an…! Beachtet deren Input-Parameter!

1. **Gobale Interpretation**: Wir arbeiten wieder mit den Daten der Firma «FixIt» aus dem Kapitel «Klassifikation»
   a. Trainiert ein Gradient Boosting-Modell und prüft, ob es eine gute Performance hat bzw. konfiguriert es so, dass das der Fall ist! **Verwendet dabei zunächst alle Attribute**.
   b. Welche Features sind generell am wichtigsten?
   c. Wie lässt sich generell das Verhalten des Modells beschreiben? Wie genau hängen die Vorhersagen von den Werten der wichtigsten Features ab?
   d. Kann man die «Target Leakage» erkennen?
   e. Entfernt nun das Attribut «Days Open» und wiederholt die Schritte a-c! Erstellt einen «Steckbrief» des Modells, in dem ihr die Fragen b und c beantwortet!

Interpretierbarkeit: Aufgabe 2. und 3. Teil

2. **Vergleich** mit intrinsisch interpretierbaren Modellen: Vergleicht eure Interpretationen mit der Interpretation eines Entscheidungsbaums (von damals) und einer logistischen Regression.
   a. Gibt es Unterschiede? Welche?
   b. Was lässt sich mit den post-hoc-Methoden erreichen, was nicht? Ist das schlimm*?

3. **Lokale Interpretierbarkeit**:
   a. Wählt ein Ticket aus, welches zu einer SLA Violation geführt hat, sowie eines, bei dem dies nicht der Fall war. (Tipp: um Instanzen auszuwählen, kann man das «Data Table»-Widget nutzen!)
   b. Kreiert eine Erklärung für diese beiden Tickets und beschreibt sie in Worten!

** Was, wenn ihr z.B. ein Modell gelernt habt, das die Straffälligkeit von Personen vorhersagt, das zur Festsetzung von Strafen verwendet wird?*