Interpretierbarkeit

Target Leakage	Das Modell schummelt, weil es beim Training Infos bekommt, die es beim echten Einsatz nicht zur Verfügung hat**.** Das Modell wirkt im Training super, aber versagt dann im echten Einsatz.
Argumente für Interpretierbarkeit	Warum ist es wichtig, dass Modelle verständlich und nachvollziehbar sind. **1. Debugging** a) Data Leakage erkennen - das bedeutet, dass das Modell Informationen verwendet, die es beim echten Einsatz gar nicht haben darf - Interpretierbarkeit hilft, so etwas zu erkennen b) Umgang mit Ausnahmefällen - Manche Situationen im echten Leben sind ungewöhnlich oder selten - Bspw. Ein autonomes Auto erkennt Fahrradfahrer. Das Modell achtet nur auf zwei sichtbare Räder. Problem: Fahrräder mit Gepäcktaschen, die die Räder verdecken, werden nicht erkannt - Verständliche Modell helfen, solche Schwächen zu entdecken c) Unerwünschten Bias erkennen - Bspw. Ein Modell bevorzugt männliche Bewerber, weil es aus verzerrten Daten gelernt hat - Nur wenn wir das Modell verstehen, können wir Diskriminierung verhindern **2. Vertrauen** - Menschen müssen die Modellentscheidungen nachvollziehen können, bevor sie auf ihnen basierend entscheiden, da sie verantwortlich sind **3. Wissenschaftliche Erkenntnis** - Durch verständliche Modelle können Forscher und Experten neue Muster oder Zusammenhänge entdecken, die vorher nicht bekannt waren
Arten von Interpretierbarkeit	**Intrinsisch**(modell kann man anschauen) <-> **post-hoc(**modell kann man nicht anschauen) **Modellspezifisch**(funktioniert nur für ein Modell) <-> **modell-agnostisch**(funktioniert für jedes Modell) **Lokal**(erklärt eine Instanz) <-> **global**(erklärt das ganze Modell)
Intrinsisch interpretierbare Modelle (Modelle von Natur aus leicht verständlich sind)	**Lineare Modelle (Lineare Regression, Naive Bayes etc.): Vorteile:** - gut erforscht - gut für numerische Merkmale geeignet - die Gewichte zeigen direkt, wie stark jedes Merkmal die Vorhersage beeinflusst

	Nachteile: - Wechselwirkungen zwischen Merkmalen werden nicht erfasst - Kann keine komplexen Zusammenhänge erkennen **Logikbasierte Modelle (Entscheidungsbäume):** **Vorteile:** - Sehr einfach zu verstehen - Erkennt Interaktionen zwischen Merkmalen **Nachteile:** - Numerische Werte werden immer diskretisiert - Bäume können instabil sein und teilweise schwer zu interpretieren
Permutation Feature Importance	Methode, um zu bestimmten, wie wichtig ein einzelnes Merkmal für ein Modell ist. Idee: Wie stark steigt der Fehler des Modells, wenn man das Attribut «kaputtmacht», also die Werte zufällig mischt. → wenn ein Feature wichtig ist, dann wir das Modell viel schlechter Vorteile: - Wenn ein Feature z. B. nur zusammen mit einem anderen wichtig ist, wird das bei der Permutation trotzdem erkannt, weil die Beziehung dabei auch zerstört wird Nachteile: - Benötigt Zugang zu den Labels der Daten - Wenn man zwei stark zusammenhängende Features hat, dann scheint keines von beiden wichtig, weil das Modell die Info doppelt bekommt, das macht die Interpretation nicht immer logisch nachvollziehbar
Partial Dependence Plots	Verstehen, wie ein einzelnes Merkmal die Modellvorhersage beeinflusst, im Durchschnitt über alle Datenpunkte Beispiel: Wenn du wissen willst, wie sich „Alter = 50“ auf die Vorhersage auswirkt, dann: - nimm alle Datenpunkte, - setze deren Alter auf 50, - rechne für jeden die Vorhersage neu aus, - und bilde **den Durchschnitt dieser Vorhersagen** → das ist der PDP-Wert bei „50“. Vorteile: - intuitiv verständlich Nachteile: - funktioniert nur gut für 1 oder 2 Features gelichzeitig - ignoriert Korrelationen

Shapely-Werte (lokal)	Man stellt sich vor, dass die Merkmale einer Instanz nacheinander in zufälliger Reihenfolge ins Modell «hineingehen» - Jedes Feature trägt etwas zur **Gesamtvorhersage** bei - Der **Shapley-Wert** eines Merkmals ist: → Wie viel ändert sich die Vorhersage im Schnitt, wenn dieses Feature dem Team beitritt? Beispiel: - Wenn Alter „als letztes reinkommt“ – wie stark ändert sich dann die Vorhersage? - Man wiederholt das mit allen möglichen Reihenfolgen und mittelt Vorteile: - Mathematisch fundiert - Kann vollständige Erklärungen liefern Nachteile: - Rechenintensiv - Nicht sofort intuitiv
SHAP Summary Plot(global)	- 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)
Individual Conditional Expecatiton	- 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. Vorteile: - Zeigt mehr Details (Verteilung) als PDPs - Ziemlich intuitiv Nachteile: - Nur 1 Feature darstellbar - Bildet keine Interaktionen/Korrelationen ab