XAI (Explainable Artificial Intelligence)

XAI je soubor metod a technik, které umožňují lidským uživatelům porozumět výsledkům a výstupům vytvořeným algoritmy strojového učení a důvěřovat jim. Zatímco klasické modely (např. rozhodovací stromy) jsou přirozeně srozumitelné, moderní hluboké sítě jsou tak komplexní, že jejich vnitřní logika je lidem skrytá.

Cílem XAI je dosáhnout rovnováhy mezi **vysokou přesností** modelu a jeho **interpretovatelností**.

1. Proč XAI potřebujeme?

S rostoucím nasazením AI v kritických oblastech vyvstávají zásadní otázky:

- **Etika a spravedlnost:** Nediskriminuje model určitou skupinu lidí (např. při žádosti o půjčku)?
- **Bezpečnost:** Proč autonomní vozidlo vyhodnotilo překážku jako stín?
- **Právo:** Podle nařízení EU (GDPR) mají občané „právo na vysvětlení“ u automatizovaných rozhodnutí, která se jich týkají.
- **Korekce:** Pokud víme, proč model udělal chybu, můžeme ho lépe opravit.

2. Metody vysvětlování

XAI využívá různé techniky k odhalení vnitřního fungování modelů:

A. Lokální vysvětlení (např. LIME)

Snaží se vysvětlit jedno konkrétní rozhodnutí.

- **Příklad:** „Tato konkrétní žádost o půjčku byla zamítnuta kvůli nízkému zůstatku na účtu a krátké době v zaměstnání.“

B. Globální vysvětlení

Snaží se popsat celkové chování modelu – které faktory jsou pro něj obecně nejdůležitější.

- **Příklad:** „Model pro předpověď počasí přikládá největší váhu atmosférickému tlaku.“

C. Feature Attribution (Atribuce příznaků)

Metody jako **SHAP** (SHapley Additive exPlanations) přiřazují každému vstupnímu parametru číselnou hodnotu podle toho, jak moc přispěl k výsledku.

D. Saliency Maps (Mapy pozornosti)

Používají se u obrazových dat (CNN). Zvýrazňují pixely v obrázku, na které se model „díval“, když identifikoval objekt. [Image showing a photo of a dog and its corresponding saliency map highlighting the ears and snout]

3. Vztah mezi složitostí a vysvětlitelností

Existuje nepřímá úměra:

- **Vysoká interpretovatelnost:** Lineární regrese, rozhodovací stromy (víme přesně, co se děje, ale přesnost může být nižší).
- **Vysoká přesnost:** Hluboké neuronové sítě, soubory modelů (skvělé výsledky, ale nikdo přesně neví, jak vznikly).

4. Praktické využití XAI

Oblast	Využití XAI
Medicína	Lékař potřebuje vědět, proč AI označila snímek jako rizikový, než zahájí léčbu.
Finance	Banka musí klientovi vysvětlit důvody zamítnutí hypotéky.
Právo	Analýza soudních rozhodnutí a identifikace možných systémových zaujatostí.
Průmysl	Pochopení příčin poruchy stroje, kterou AI předpověděla.

5. Budoucnost: Integrovaná vysvětlitelnost

Moderní trendy směřují k vytváření modelů, které jsou „vysvětlitelné už od návrhu“ (Interpretable by Design), místo aby se vysvětlení hledalo až dodatečně na hotovém modelu.

> **Zajímavost:** Výzkum ukazuje, že lidé mají tendenci AI důvěřovat více, pokud jim poskytne jakékoli vysvětlení, i kdyby bylo velmi jednoduché. To se nazývá „placebo efekt vysvětlení“ a je to jedna z psychologických pastí, na které si tvůrci XAI musí dávat pozor.

Zpět na AI rozcestník