DRUHÝ SEMINÁŘ

AI jako datový analytik

Co skutečně spočítala, co odvodila a co doplnila?

Minule: Co delegovat?
Dnes: Co se při delegování analýzy vlastně děje?

Mgr. Markéta Trnečková, Ph.D.

ZADÁNÍ

Předpověď kvality ovzduší

Cíl: odhadnout PM2.5 na jedné stanici za 24 hodin.

SLOUPCE V NAŠÍ TABULCE

Měření	čas, teplota, vlhkost, rychlost větru
Provoz a ovzduší	intenzita dopravy, PM2.5, PM10
Další sloupce	PM2.5 za 24 hodin, centrovaný 24hodinový průměr PM2.5

Centrovaný průměr v čase t: měření od $t - 12$ do $t + 11$ hodin.

Návrh pracovního postupu

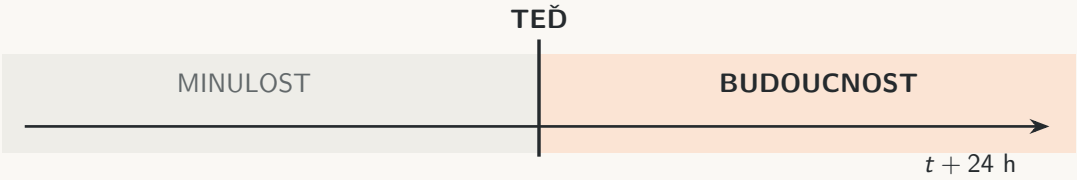
Navrhni co nejpřesnější model pro předpověď PM2.5 za 24 hodin. Použij naši tabulku.

AI ASISTENT

1. **Příprava vstupů** Použijeme všechny numerické sloupce jako vstupy.
2. **Rozdělení dat** Náhodně vyčleníme 80 % pro trénink a 20 % pro test.
3. **Modelování** Vstupy standardizujeme. Porovnáme Random Forest, XGBoost a neuronovou síť.
4. **Vyhodnocení** Vybereme model s nejnižší RMSE na testovacích datech.

Použili byste tento postup? Co byste si chtěli ověřit?

Co máme k dispozici v čase předpovědi?



Cíl mezi vstupy

PM2.5 za 24 hodin je hledaná odpověď.

Centrovaný průměr

Centrovaný průměr v čase t ještě neznáme.

Náhodný test

Trénink a test míchají časová období.

Odborné znalosti potřebujeme i ke kontrole přesvědčivého návrhu.

Didaktická ukázka sestavená podle zadání. Nejde o záznam odpovědi konkrétního modelu.

Od návrhu ke skutečné analýze

U PM2.5 jsme kontrolovali návrh.

Jsou vstupy dostupné?
Odpovídá test budoucímu použití?

Teď AI dostane skutečný soubor.

Co z něj opravdu spočítá?
Jak to dokážeme ověřit?

Potřebujeme dvě kontroly:

Dává postup smysl? Byl skutečně proveden?

Kterou z těchto otázek zodpoví samotný běh Pythonu?

Jedna tabulka, mnoho rozhodnutí

Výsledky dovednostního testu před programem a po něm.

CSV

Co potřebujeme vědět, než začneme porovnávat skupiny?

experiment.csv

Syntetická výuková data.

`id, date, group, age, height_cm,`
`weight_kg, temperature, score_before, score_after`

Význam sloupců najdete v popisu dat. Zatím nic neopravujte.

První zadání

Prozkoumej tato data a řekni mi, co je v nich důležité a co bych měl před další analýzou řešit.

Nahrajte `experiment.csv` a přiložte popis dat.
Použijte běžné nastavení svého systému.

Uložte odpověď i viditelné výstupy nástrojů.

Zapište systém, označení modelu, čas a dostupné nástroje.

ZÁZNAM VLASTNÍ ZKUŠENOSTI

Tvrzení potřebuje oporu

AI zjistila	Ještě neprověřila	Tvrdí, ale ověřím

U jednoho tvrzení dohledáte cestu:

soubor / použitý postup / výstup

„Zjistila“ zapisujte s konkrétní evidencí. Zatím neznáte celý klíč.

Viděli jste výpočet, nebo jen hotovou odpověď?

Kód v chatu ještě nedokládá spuštění.

Hledejte záznam běhu, načtený soubor a skutečný výstup.
Pokud běh není viditelný, označte stav jako neověřený.

Co vytváří jazykový model

Text se reprezentuje posloupností tokenů.

$$x_1, \quad x_2, \quad \ldots, \quad x_n$$

$$P_\theta(x_{n+1} \mid x_1, \ldots, x_n)$$

Model postupně generuje pokračování podle kontextu.
Token může být část slova, čísla nebo interpunkce.

Zjednodušený pohled na autoregresivní LLM. Samotný vzorec nepopisuje celou aplikaci.

Odkud se bere analytický postup?

Naučené parametry

Vzory jazyka, vztahy mezi pojmy, postupy a struktury kódu.

Aktuální kontext

Zadání, zpřístupněná data, historie a výsledky nástrojů.

Znalost obvyklé EDA
$\neq$ provedená EDA nad tímto souborem.

Příloha může vstoupit přes náhled, výběr textu nebo nástroj.

Samotné nahrání souboru neříká, kterou část model skutečně zpracoval.

Tři situace, tři možnosti ověření

1. **Dotaz → LLM → odpověď**
   Návrh postupu bez přiložených dat.

2. **Data v kontextu + dotaz → LLM → odpověď**
   Model může správně odvozovat i počítat, ale běh nástroje není doložen.

3. **LLM → výpočetní nástroj → výstup → LLM**
   Výsledek lze spojit s konkrétním kódem a daty.

Pro reprodukovatelné souhrny celého CSV chceme doložit skutečný výpočet.

Výpočet může být správně provedený

A přesto odpovídat na jinou otázku.

Co ověří nástroj?

Že zvolený kód nad vstupem vytvořil tento výstup.

Co posuzujeme my?

Zda jsme vybrali správná data, výpočet a interpretaci.

Příklad: průměr všech řádků versus průměr za účastníky.

Druhé kolo: doložená analýza

Proveď systematický průzkum datasetu. Vytvoř vhodné souhrny číselných i kategoriálních proměnných. Zkontroluj chybějící hodnoty, duplicity, rozsahy a neobvyklá pozorování.

Pokud máš výpočetní nástroj, skutečně ho použij.
Uveď soubor, kód, počty zahrnutých řádků a výstupy, z nichž závěry vycházejí. Bez běhu nevydávej odhady za vypočtené statistiky.

Data zatím neměň. Bez nástroje vygeneruj skript pro místní spuštění.

Změnu nástroje nebo modelu zapiš do pracovního listu.

Co se mezi koly změnilo?

Více nálezů? Lepší evidence? Nové otázky?

U tří závěrů porovnejte první a druhou odpověď.
Čím je nyní výsledek doložený?

Který výsledek dokážete zopakovat mimo chat?
Který krok by stále potřeboval kontrolu?

První kolo mohlo použít nástroj také.

Srovnáváme skutečné podmínky. Rozdíl automaticky nepřipisujeme lepšímu promptu.

EDA pokračuje další otázkou

Co se o datech potřebujeme dozvědět?



NEJPRVE BEZ AI

Na co se analýza ještě nezeptala?

Najděte jednu otázku,
která může změnit závěr studie.

Napište ji dříve, než znovu otevřete chat.

Pokud AI našla vše podstatné, vyberte neověřený předpoklad nebo chybějící informaci o sběru dat.

Teď prověřte svou otázku

„Prověř právě tuto hypotézu. . . “

Jaký výsledek by vaše podezření podpořil?
Co ho může oslabit?

Který výpočet nebo graf potřebujete?
Co se z CSV samotného nedozvíte?

Oddělte výsledek kontroly, jeho interpretaci a to, co zůstává neznámé.

Po společné diskusi: kolik je pozorování?

204

řádků v CSV

201

po odebrání 3 přesných kopií

200

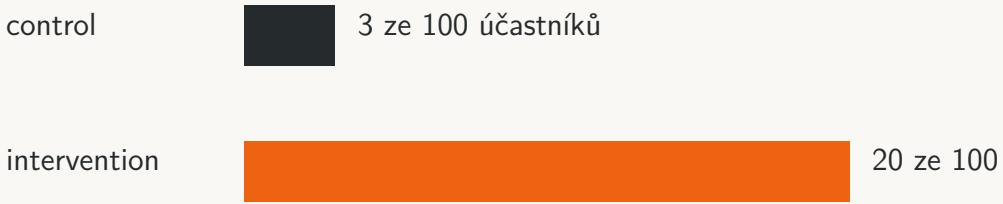
různých účastníků

P042 má dvě různá měření.
Opakované ID samo o sobě nedokazuje chybu.

Tři nadbytečné kopie znamenají šest řádků v duplicitních dvojicích.
Vždy uveďte, kterou z těchto veličin počítáte.

Chybějí výsledky všude stejně?

Chybějící `score_after` podle skupiny

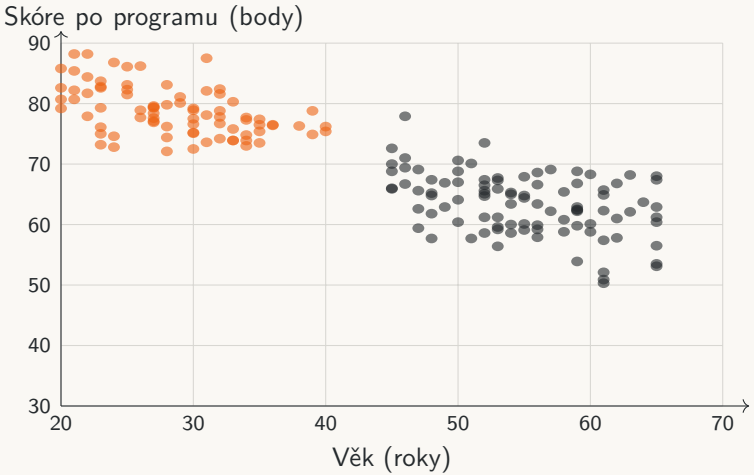


Víme, kde hodnoty chybějí. Z toho ještě neplyne, proč chybějí.

Audit CSV: po přesné deduplikaci a výběru prvního záznamu každého ID.

Původní CSV má 24 NA ve score_after, tato tabulka 23. Hodnota 999 není NA.

Vyšší skóre. Srovnatelné skupiny?



intervention
20–40 let

control
45–65 let

Věk se mezi skupinami nepřekrývá.

Citlivostní pohled: jeden záznam na ID, podezřelé buňky nahrazeny NA.

Graf zahrnuje 95 control a 80 intervention s dostupným věkem a skóre. Bez imputace.

Co přesně porovnáváme?

Skóre po programu

control: **63,38**
intervention: **78,76**

Průměr individuální změny

control: **4,98**
intervention: **5,55**

Ani jeden rozdíl sám nedokazuje účinek programu.

Skupiny se liší věkem. Chybějící výsledky mohou ovlivnit oba souhrny.

Stejný citlivostní pohled jako v grafu. Skóre i párové změny: n = 96 / 80.

Není to automaticky „vyčištěná pravda“. Bez ověření zdroje neopravujeme jednotlivé hodnoty.

„Data jsou přibližně normální.“

Na čem tento závěr stojí?

Která proměnná? V jaké skupině? Po jakých úpravách?
Máme histogram, Q–Q graf nebo výsledek testu?

Potřebuje zamýšlená metoda normalitu této proměnné,
nebo například reziduí modelu?

Nevýznamný test normality není důkaz normality.

Jde o výukovou formulaci k posouzení, ne o výstup z vašeho běhu.

Pravděpodobné pokračování a pravdivost

$$P_\theta(\text{další token} \mid \text{kontext}) \neq P(\text{tvrzení je pravdivé})$$

Plynulá odpověď může obsahovat nepodložené tvrzení.
Výstup nástroje poskytne evidenci, kterou je ještě třeba správně vyložit.

Požadavek „spočítej to“ pomůže jen tehdy, když systém výpočet skutečně provede.

AI navrhuje. Nástroj počítá. Člověk posuzuje.

Co dnes delegujete s větší jistotou?

A jakou evidenci budete vyžadovat?

Odborník často musí poznat, kterou otázku je potřeba položit.

Zdroje a původ výukových dat

Brown a kol. (2020): Language Models are Few-Shot Learners.

Schick a kol. (2023): Toolformer.

Ji a kol.: Survey of Hallucination in Natural Language Generation.

NIST/SEMATECH: What is EDA?

Vaswani a kol. (2017): Attention Is All You Need.

Dataset je syntetický. Graf a souhrny vznikly spuštěním kontrolního skriptu nad přiloženým CSV. Nejde o výsledky reálné studie ani skutečný výstup studentského AI systému.

Detaily, předpoklady a odkazy na dokumentaci jsou v handoutu.

Další slide je volitelný a není součástí základního 90minutového plánu.

Pod kapotou: proč záleží na kontextu

Tokeny mají číselné reprezentace.

Attention kombinuje informace z dostupných tokenů podle jejich vztahů. Poziční informace pomáhá zachytit pořadí.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

Q: dotazy, K: klíče, V: hodnoty. U autoregresivního modelu maskování brání použití budoucích tokenů.
Tento mechanismus sám neotevírá CSV ani nespouští Python.

To zajišťuje propojení modelu s nástroji v aplikaci.