

Princípy dátovej vedy: Domáca úloha

Termín odovzdania: 3.11.2024, 23:55

Pracovať budeme s dátami o ekonomike a zdraví v jednotlivých krajinách. Dáta sú vopred rozdelené na:

trénovací dataset: <http://www.iam.fmph.uniba.sk/ospm/Rosa/PDV/countries-train.csv>

testovací dataset: <http://www.iam.fmph.uniba.sk/ospm/Rosa/PDV/countries-test.csv>

Dôkladnejší popis dát je na konci tohto zadania.

Vykonajte nasledovné analýzy:

1. Spravte pre tieto dáta vhodné vizualizácie a vypočítajte vhodné deskriptívne štatistiky. Výstupy interpretujte! Teda čo zaujímavého ste sa z dát dozvedeli (napr. o charakteristikách jednotlivých premenných a o vzťahoch medzi nimi; alebo identifikujte "výnimočné" krajiny apod.). Uveďte aspoň tri zaujímavé pozorovania.
2. Aplikujte lineárnu regresiu na modelovanie dĺžky dožitia v krajinách podľa ich ostatných charakteristík. Zhodnoťte kvalitu tohto modelu (zatiaľ bez použitia testovacieho datasetu).
3. Použite validáciu alebo krosvalidáciu na vylepšenie tohto modelu. Napríklad môžete pomocou týchto metód nájsť lepší model s menším počtom premenných, alebo nájsť model s vhodne transformovanými premennými. Nemusíte nájsť najlepší možný model, ale nejaké zlepšenie v (kros)validačnej chybe by ste mali dosiahnuť.
4. Na testovacej vzorke zhodnoťte, ako dobre váš finálny model vie predikovať očakávanú dĺžku života v krajinách. Slovné zhodnoťte a zdôvodnite, či je podľa vás model dokonalý/dobrý/slabý apod.
5. Stručne napíšte, aké ďalšie dáta, ktoré sa reálne dajú získať, by ste do modelu skúsili zahrnúť (a prečo), aby ste zlepšili predikcie dĺžky prežitia po krajinách [toto už neprogramujte]. Napíšte tiež, aké iné zmysluplné otázky by sa dali s týmto datasetom skúmať (okrem modelovania dĺžky života v krajinách).

Svoje výstupy odovzdajte cez Google classroom ako okomentovaný Python notebook alebo okomentovaný Rkovský skript (odovzdávate teda jeden súbor).

Hint: Reálne datasety typicky nie sú také pekné a jednoduché, že ich stačí zobrať, pustiť na nich model a hneď sme hotoví. To platí aj pre tento dataset, t.j. sú v ňom nejaké jasné nedostatky, ktoré potom škodia lineárnej regresii. Už v bode 3 by ste mali pracovať s tak upraveným datasetom, aby v sebe nemal z pohľadu lineárnej regresie evidentné hlúposti.

Poznámky:

a) Nejde o to získať extrémne dobrý predikčný model, ale o prvé praktické aplikovanie "princípov dátovej vedy". Kvalita finálneho modelu nebude hodnotená - pokiaľ nedostatočná kvalita nebude dôsledkom chyby, alebo toho, že ste nespravili niečo potrebné. Pre porovnanie, môj základný model v bode 3 mal MSE_train okolo 11,9.

b) Komentáre výstupov sú kľúčové, nezabudnite v nich uviesť všetky vaše zistenia.

c) Úlohu treba vypracovať samostatne. Špeciálne nebudem vnímať a hodnotiť pozitívne, ak budete mať hromadne rovnaké nápady (napr. na analýzy v bode 1 a vylepšenia v bode 5). Tiež sa vyvarujte používaniu ChatGPT a podobne. Viac detailov:

http://www.iam.fmph.uniba.sk/ospm/Rosa/PDV/PDV_hodnotenie.pdf

d) Ak je niečo nejasné, spýtajte sa ma (naživo, mail, google classroom).

Dáta:

Pre každú z krajín v datasete (či už v trénovacej alebo testovacej vzorke) máme o nej informácie:

- Country: jej názov
- child_mort: miera detskej úmrtnosti (počet úmrtí na 1000 detí do 5 rokov)
- exports: export krajiny, udávaný ako percento HDP
- health: výdavky na zdravotnú starostlivosť, udávané ako percento HDP
- imports: import krajiny, udávaný ako percento HDP
- income: čistý príjem na obyvateľa
- inflation: ročná miera inflácie
- life_expec: očakávaná dĺžka života (koľko očakávame, že sa v priemere dožije novonarodené dieťa)
- total_fer: miera plodnosti (~ priemerný počet detí na ženu)
- gdpp: HDP na obyvateľa
- region: 1 = Ázia, 2 = Európa, 3 = Afrika, 4 = Amerika (severná alebo južná), 5 = Austrália a Oceánia