I-SUNS: Zadanie č.2

Stromy, stroje, hlasovania a redukcia dimenzie

Vo vybranom programovacom jazyku implementujte program, ktorý bude predpovedať cenu letenky. V tomto zadaní budete pracovať s dátami z AIS. Výstupný stĺpec pre toto zadanie je *Price*.

Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 10 bodov je **15.11.2023 o 9:00/11:00 (pred vaším cvičením)**. Každý týždeň omeškania je penalizovaný stratou dvoch bodov.

- Načítajte dáta a pripravte ich na spracovanie modelmi ML - odstráňte stĺpce s identifikátormi, odstráňte null hodnoty, **duplikáty**, outliery, správne spracujte textové hodnoty (pozor: použite vhodné kódovanie, myslite pri tom na to, že spracovanie stĺpca, ktorý má príliš veľa unikátnych hodnôt môže pridať do vášho datasetu príliš veľa nových stĺpcov. Modely s ktorými budete robiť na tomto zadaní pracujú ťažšie s veľkým počtom stĺpcov). **1b**

- Rozdeľte dáta na trénovaciu a testovaciu množinu (validačnú Vám v tomto zadaní netreba), následne na vstupnú a výstupnú množinu, potom normalizujte (výstup nenormalizujete).

- Trénujte nasledujúce modely (pre každý model dosiahnite kladné R2 skóre[1] - pri najlepšom modeli aspoň 0.7):

  - rozhodovací strom **0.5b**:
    - jeden strom z výsledkov aj zobrazte do dokumentácie - ak budú Vaše stromy príliš veľké, pre vizualizáciu natrénujte menší strom, aj keby mal mať horšie výsledky, je potrebné ho vedieť analyzovať; **0.5b**
  - Vami vybraný stromový súborový (*ensemble*) model **0.5b**:
    - vizualizujte dôležitosť vstupných parametrov - ak ich bude príliš veľa, zredukujte ich na podmnožinu najdôležitejších; **0.5b**
  - model SVM. **1b**

  Modely vyhodnoťte na trénovacej a testovacej množine pomocou MSE (príp. RMSE), R2 a výsledky vizualizujte tak, aby ste mohli aj analyzovať reziduály (pozor: treba vizualizovať reziduály, tj. **nie** očakávanú hodnotu vs. predpovedanú hodnotu). Navzájom modely porovnajte. **1b**

[1] Nezamieňajte si túto metriku s úspešnosťou, R2 skóre môže byť záporné, neuvádzajte ho v %!

- Sledujte, čo s dátami spraví redukcia dimenzie (na tomto zadaní pomocou 3D point grafov - scatter plot):
  - Vyberte si 3 príznaky (pred normalizáciou), ktoré budú na osiach. Snažte sa nájsť také príznaky, pri ktorých budete vedieť graf analyzovať. Dáta vyfarbite podľa výstupného parametra (ceny). Pokúste sa z grafu vyčítať nejakú závislosť. **1b**
  - Minimalizujte množinu (po normalizácii, bez výstupného parametra) na 3 dimenzie (pomocou PCA, UMAP ...), tie vyneste na osi, dáta opäť zafarbite podľa výstupného parametra (ceny). **1b**

  Grafy navzájom porovnajte.

- Vyberte podmnožinu príznakov , vyberte si najúspešnejší model z prvej časti zadania a opäť ho natrénujte pre zmenšenú množinu:
  - podľa korelačnej matice; **1b**
  - podľa dôležitosti príznakov z ensemble modelu; **1b**
  - podľa variancie pomocou PCA (zvoľte si hodnotu variancie, nie počet príznakov - t.j. nie 3). **1b**

  Výsledky porovnajte medzi sebou aj s pôvodným trénovaním pomocou MSE (príp. RMSE), R2 a reziduálov.

Nepovinné úlohy

Body za nepovinné úlohy sú udelené len v prípade, že sú vypracované správne:

- EDA. **1b**
- Zhlukujte vaše dáta (minimálne 3 kategórie):
  - Výsledky vizualizujte na 3D grafe (pozor: pri zhlukovaní použite viac príznakov než pri zobrazovaní v grafe, nepoužívajte výstupný parameter - cenu). **1b**
  - Natrénujte Váš najlepší model pre jednotlivé kategórie vzniknuté zhlukovaním a porovnajte medzi sebou aj s pôvodným výsledkom. **1b**
- Natrénujte umelú neurónovú sieť - pozor na zmenu typu problému, nejedná sa o klasifikáciu. Je potrebné prispôsobiť sieť aj analýzu výsledkov. **1b**

Upresnenie stĺpcov

- *ID* - Identifikátor záznamu. Odstráňte ho skôr než budete odstraňovať duplikáty.
- *Airline* Názov leteckej spoločnosti, ktorá prevádzkuje let.
- *Flight* Identifikátor letu.
- *Source city* Mesto, z ktorého lietadlo odlieta.
- *Departure time* Časový rámec odletu, ktorý popisuje, kedy lietadlo odlieta.
- *Stops* Počet zastávok letu.
- *Arrival time* Časový rámec príletu.
- *Destination city* Mesto, kde lietadlo pristane.
- *Class* Trieda cestovania, čiže typ kabíny, ktorú si cestujúci zvolil.
- *Duration* Dĺžka letu v numerickom formáte
- *Days left* Počet dní do odletu v čase, keď bola letenka rezervovaná.
- *Price* Cena letenky.