

I-SUNS: Zadanie č.2

STROMY, STROJE, HLASOVANIA A REDUKCIA DIMENZIE

Vo vybranom programovacom jazyku implementujte program, ktorý bude predpovedať cenu auta. 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 **16.11.2022 o 14:00/15:00/16:00 (pred vašim 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**, outliers, 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 tréningovú 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 skóre¹ - pri najlepšom modeli aspoň 0.5):
 - 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 tréningovej 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**

¹Nezamieňajte 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 vyfarbíte 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 zafarbíte 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 zhľukovaní použite viac príznakov než pri zobrazovaní v grafe, nepoužívajte výstupný parameter - cenu). **1b**
 - Natrénujte Váš najlepší model pre jednotlivé kategórie vzniknuté zhľukovaním a porovnajte medzi sebou aj s pôvodným výsledkom. **1b**
- Natrénujte umelú neurónovú sieť - pozor na zmenu typu problému, nejedná sa o klasifikáciu. Je potrebné prispôbiť sieť aj analýzu výsledkov. **1b**

Upresnenie stĺpcov

- *ID* - Identifikátor záznamu. Odstráňte ho skôr než budete odstraňovať duplikáty.
- *Price* (spojitá číselná hodnota) - Cena auta v \$ - výstupná hodnota pre Vaše modely.
- *Levy* (textová hodnota - treba upraviť aby bola spojitá číselná hodnota) - Odvody.
- *Manufacturer* (kategorická hodnota) - Výrobca auta.
- *Model* (kategorická hodnota) - Model auta.
- *Prod. year* (celočíselná hodnota) - Rok výroby auta.
- *Category* (kategorická hodnota) - Typ vozidla.
- *Leather interior* (kategorická hodnota) - Určuje, či má vozidlo kožený interiér.
- *Fuel type* (kategorická hodnota) - Typ paliva.
- *Engine volume* (treba upraviť aby bola spojitá číselná hodnota) - Objem motora.
- *Mileage* (textová hodnota - treba upraviť aby bola spojitá číselná hodnota) - Najazdené kilometre.
- *Cylinders* (celočíselná hodnota) - Počet valcov motora.
- *Gear box type* (kategorická hodnota) - Typ prevodovky.
- *Drive wheels* (kategorická hodnota) - Pohon auta (predný, zadný, 4x4).
- *Doors* (kategorická hodnota) - Počet dverí.
- *Left wheel* (kategorická hodnota) - Určuje, či má vozidlo volant naľavo.
- *Color* (kategorická hodnota) - Farba auta.
- *Airbags* (celočíselná hodnota) - Počet airbagov.
- *Turbo engine* (kategorická hodnota) - Určuje, či je v aute turbo motor.