I-SUNS: Zadanie č.2

Stromy, stroje, hlasovania a redukcia dimenzie

Vo vybranom programovacom jazyku implementujte program, ktorý bude predpovedat cenu domu. V tomto zadaní budete pracovať s dátami z AIS. Pre dataset budú dostupné dva súbory - testovacie a trénovacie csv a takisto txt súbor, v ktorom je popis stĺpcov. Výstupný stĺpec pre toto zadanie je *SalePrice*.

Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 15 bodov je **22.11.2022 o 08:00/10:00/13:00 (pred vaším cvičením)**. Každý týždeň omeškania je penalizovaný stratou dvoch bodov.

- Načítajte dáta z oboch množín (trénovacej - *train.csv* a testovacej - *test.csv*) a pripravte ich na spracovanie modelmi ML - odstráňte null hodnoty, duplikáty, urobte korelačnú maticu (min. pozrite najviac korelované stĺpce), spracujte textové hodnoty, normalizujte[1]. **1b**

- Trénujte nasledujúce modely (pre každý model trénujte s *GridSearch* nad hyperparametrami a cross validáciou **1b**):

  – rozhodovací strom (jeden strom z výsledkov aj zobrazte do dokumentácie); **1b**

  – SVM (vyskúšajte aspoň dva typy kernelu a niekoľko hodnôt C a parametra gamma, výsledky jednotlivých konfigurácií *GridSearch* vizualizujte); **2b**

  – Vami vybraný stromový súborový (*ensemble*) model (vizualizujte dôležitosť vstupných parametrov). **2b**

  Najlepší model z *GridSearch* (z každej metódy) vyhodnoťte na testovacej množine pomocou MSE, R2 a výsledky vizualizujte tak, aby ste mohli aj analyzovať reziduály. **1b** Navzájom modely porovnajte.

- Sledujte, čo s dátami spraví redukcia dimenzie (na tomto zadaní pomocou 3D point grafov):

  – Vyberte 3 príznaky z originálnej databázy, ktoré budú na osiach. Snažte sa nájsť také príznaky, pri ktorých budete vedieť graf analyzovať. Dáta vyfarbite podľa výstupného parametra. **1b**

[1] Po diskusii na cvičeniach pribudli aj verzie súborov {*filename*}_*dummy.csv*, kde už je nad kategorickými hodnotami spravený one hot encoding.

  - Minimalizujte množinu (bez výstupného parametra) na 3 dimenzie (pomocou PCA, UMAP ...), tie vyneste na osi, dáta opäť zafarbite podľa výstupného parametra. **2b**

  Grafy navzájom porovnajte.

- Podľa poznatkov z predošlého trénovania, korelačnej matice aj 3D grafov, vyberte podmnožinu príznakov, ktoré zredukujete na X dimenzií. **1b** Pre X vyberte aspoň 5 hodnôt. Vyberte si najúspešnejší model z prvej časti zadania a opäť ho natrénujte pre zmenšené množiny (rôzne hodnoty X). Vyneste do grafu závislosť R2 skóre a času trénovania od veľkosti množiny. **2b** Snažte sa dopracovať k čo najúspešnejšiemu modelu. Výsledky navzájom porovnajte. **1b**

Nepovinné úlohy

- EDA. **1-2b**
- Zhlukujte vaše dáta. Výsledky vizualizujte na 3D grafe. **1-2b**
- Podľa výsledkov zhlukovania rozdeľte problém na viacero častí a pre ne trénujte modely samostatne. **1-3b**
- Rozšírte modely aj o štvrtú kategóriu - umelú neurónovú sieť. **1-2b**
- Vráťte sa k zadaniu 1 a analýze stĺpcov a aplikujte poznatky z redukcie dimenzie. Na zredukovanej databáze opäť trénujte sieť a výsledky porovnajte. **1b**