Pozdrav svima,

U prilogu datoteka koja se koristi za provjeru znanja iz drugog modula, Uvod u Podatkovnu znanost.

U vašem datsertu imate podatke o osobama (starost, spol, primanja, staž) I cijanu varijablu - iznos odobrenog kredita.

Cilj je istražiti koje varijable utječu na iznos odobrenog kredita.

Vaši zadaci - EDA proces

1. Učitati dataset I osnovna analiza (head, describe, info). Opisati što vidite u komentarima (tipa starost se kreće u tom i tom rasponu, sumnivo je to i to, itd) - 2 boda
2. Upoznati se s kategoričkim varijablama - koliko podataka imamo npr u koloni spo - 2 boda
3. Čišćenje podataka
    - Srediti duplikate - 2 boda
    - Srediti missing values - 4 boda
    - Srediti outliere - 4 boda
    - Srediti krive upise (na koliko načina imate napisan spol) - 4 boda
4. Otkriti veze medju podacima (korelacija, grafički prikazi, grupiranja itd) - 6 bodova
5. Odrediti koje varijable ostaju u datasetu, a koje mičete (4 boda) - objasniti zašto ih mičete. Koje brojke su vas navele da nešto maknete?  Ili možda vaš domain knowledge?

Cijeni se kod s komentarima  u kojima objašnjavate što radite. Uzmite u obzir da ćete se možda morati vraćati I repetitivno ponalvjati neke stvari dok niste zadovoljni. Ovo je znanost kao i vještina, slijedite gut feeling. Ako koristite gut feeling, objasnite što radite.
Dataset nije savršen, možda se ne slaže s vašim domain knowledgeom- to je ono što imate. Ako dobijete "glupi graf", ostavite ga, ali pojasnite kako iz njega nema korisnih informacija.

Rad se šalje u obliku Jupyter notebooka na moj mail, rok je subota, 26.07. do 06:59 h (duga je noć, pazite kako vozite ako budete išli ujutro za Lipik).

Vaša obrana: usmeno branite svoj projekt - ja vas pitam SAMO ono što ste napravili. Ako ne znate objasniti svoj kod, žao mi je, vidimo se opet. Za prolaz morate imati minimum 50 % bodova.

Sretno i zabavite se!

Igor