Domaći zadatak

Miljana Milošević BI2/2020

Lena Marković BI7/2020

Ime baze: Car Price Prediction Challenge
Link:
https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge/data

1. **Definisati u 2-3 rečenice problem koji će se u projektu rešavati.**

   Predviđanje cene automobila na osnovu širokog spektra atributa i funkcija. Cilj nam je da razvijemo model učenja koji tačno procenjuje cenu različitih modela automobila koristeći dataset koji sadrži detalje o automobilu kao što su dimenzije, specifikacije motora, tip goriva i tako dalje . U pitanju je problem regresije.

2. **Koliko ima uzoraka u bazi?**
   Ima 19 237 uzoraka.
3. **Jednom rečenicom objasniti šta predstavlja jedan uzorak u konkretnoj bazi.**
   Jedan uzorak u ovoj bazi predstavlja podatke o jednom automobilu.
4. **Koliko ima obeležja u bazi?**

   Ima 18 obeležja.
5. **Navesti sva obeležja (jasnim imenom na srpskom ili opisno, nebitan je naziv u samoj bazi).**

   - Identifikacioni brojevi automobila;
   - Cena automobila;
   - Porez na cenu ;
   - Ime modela automobila;
   - Proizvođač automobila;
   - Godina proizvodnje;
   - Tip karoserije;
   - Kožni enterijer;
   - Tip goriva (benzin, dizel, struja, gas...);

- Zapremina motora;
- Broj cilindara;
- Kilometraža;
- Tip menjača;
- Pogon;
- Broj vrata;
- Volan (da li je sa leve ili sa desne strane automobila);
- Boja automobila;
- Airbag (koliko ih ima);

6. **Koliko ima numeričkih obeležja?**

   Cena, porez (kada se " –" pretvori u "0") , godina proizvodnje, kilometraža(kada se skloni "km"), broj airbagova.
   Takođe, ID je numeričko obeležje, ali to svakako izbacujemo.

7. **Ako ima kategoričkih obeležja, navesti koje od njih ima najmanji broj kategorija i koje su, i navesti ono koje ima najveći broj kategorija i koliko ih je.**
   Najmanje ima volan, koji može da bude sa leve ili desne strane automobila, i kožni enterijer (da ili ne). Najveće kategoričko obeležje je model automobila kojih ima 1590.

8. **Ako se rešava regresioni problem: navesti opseg, sr.vr. i medijanu obeležja koje će se predvidjati.**
   Min: 1
   Max: 26307500
   Mean: 18556
   50%: 13172

9. **Da li postoje obeležja u bazi koja smatraš da treba izbaciti iz baze? Koja su to i zašto smatraš da ih treba izbaciti?**

   Identifikacioni broj vozila možemo izbaciti jer su svi različiti pa ne možemo ništa predvideti na osnovu toga. Takođe, možemo izbaciti i broj vrata, jer ne utiče mnogo na predikciju automobila zato što ima samo tri opcije.
   Isto važi i za boju automobila jer ni ona ne utiče na cenu.

10. **Da li u bazi ima nedostajućih vrednosti? Ako ima, navesti za svako od obeležja koliko vrednosti mu procentualno nedostaje?**

    Obeležje poreza ima 5819 '-' obeležja (30%) koja su prakticno NaN-ovi al nisu uočljiva na prvi pogled.

11. **Da li ima nevalidnih vrednosti u bazi? Ako ima, navesti za svako od obeležja koje su vrednosti nevalidne i zašto se smatraju nevalidnim.**

    Vrednost kilometraže je predstavljena sa "km", broj vrata pokazuje datum umesto ispravne automobilske oznake za vrata, neke vrednosti poreza imaju "-", neki automobili u koloni za zapreminu automobila imaju u svom nazivu turbo pa bi bilo dobro da su svi podaci istog tipa, odnosno da izbacimo "turbo" iz naziva.

12. **Ako ima nedostajućih i/ili nevalidnih vrednosti u bazi, za svako od obeležja navesti kako ce problem biti rešen.**

    - Kilometražu bi bilo dobro prebaciti u broj, tj. treba izbaciti "km". (replace('km',''),astype('int64'))
    - Broj vrata isto nije validna vrednost, jer automatski prikazuje datum umesto 4.5 ili 2.3. (drop)
    - Porez ima NaN vrednosti, pa bi trebalo da se prebaci u 0. (replace('-','0'), astype('float64'))
    - U zapremini motora može turbo da se izbaci iz naziva i da se sve vrednosti pretvore u float, zato što uglavnom svi stariji automobili nemaju turbo, a svi novi imaju - tako da ako imamo godinu proizvodnje nije presudno naglasiti da li je turbo ili ne. (replace('Turbo',''), astype('float64'))
    - Godinu proizvodnje mozemo pretvoriti u novu kolonu Age koja predstavlja 'starost' automobila. (data['Age']=dtime.year - data['Prod. year], drop prod year)

13. **Kada je završeno izbacivanje, dopuna, i drugo, navesti koliko je u sređenoj bazi ostalo uzoraka, a koliko obeležja.**

    (18924, 15)

14. Da li neka od obeležja sadrže autlajere? Navesti koja obeležja ih sadrže.

Sadrže ih sva numerička obeležja (price, levy, engine volume, mileage, cylinders, age) osim airbagova.

*Napomena: kod cene i zapremine motora postoji jedan outlier koji baš odudara od ostalih. Da li on treba da se gleda kao greška i samim tim da se izbaci iz baze?

15. Da li postoje parovi obeležja korelisani više od 0.7? Navesti takve parove obeležja.

Cilindar i zapremina motora imaju korelaciju iznad 0.7, tačnije 0.78.

16. Ako se rešava regresioni problem: utvrditi koliko je odstupanje raspodele varijable koja se predviđa od normalne raspodele dobijene korišćenjem uzoračke sr.vr. i standardne devijacije (asimetričnost i spljoštenost)?

Koef. asimetrije: 135.38

Koef. spljoštenosti: 18519.2