Kulturni centar "Kralj Fahd" Sarajevo

ISPIT

Python - napredni

Grupa: A

ID: C32G2

Ime prezime: _____

Ostvareni bodovi: _____

Negativni bodovi: _____

Ukupni bodovi: _____ **/100**

Popravni ispit: ☐ **DA** ☐ **NE**

Ponavljanje kursa: _____ _____

POTPIS PROFESORA

8. februar 2024.

Uvod:

Kompanija Geely Auto, kineski proizvođač automobila, želi proširiti svoje poslovanje na američko tržište uspostavljanjem proizvodnog pogona u Americi i lokalnom proizvodnjom automobila kako bi se natjecala s američkim i europskim konkurentima. Tom su prilikom angažovali konsulantsku automobilsku kompaniju (u kojoj ste vi zaposleni) kako bi razumjeli faktore koji utječu na cijenu automobila. Konkretno, žele razumjeti faktore koji imaju utjecaj na cijenu automobila na američkom tržištu, s obzirom na to da se oni mogu značajno razlikovati od kineskog tržišta.

Potrebno je kreirati model za predviđanje cijene automobila koristeći dostupne nezavisne varijable. Taj model će se koristiti kako bi se razumjelo kako cijene variraju ovisno o nezavisnim varijablama. Na temelju toga, uprava će moći prilagoditi dizajn automobila, poslovnu strategiju i druge faktore kako bi zadovoljila određene cjenovne razine. Osim toga, model će omogućiti upravi da bolje razumije dinamiku formiranja cijena na novom tržištu.

Vrijeme trajanja projekta je 3 mjeseca – jedan mjesec = 30 min. Sretno!

Dataset koji je potrebno učitati je pod nazivom: *Grupa_A.csv*

Odgovore pisati u komentaru ili *markdown* formatu u .ipynb *file*-u.

1. *Exploratory data analysis*/razumjevanje podataka i statistička analiza:

    a. Nakon učitavanja dataseta, potrebno je napisati kojih je dimenzija dostupni dataset (koliko redova, a koliko kolona). Također, potrebno je ispitati da li (i koliko) ima *NaN* vrijednosti kao i duplih redova.
    b. Koju kolonu možemo odmah odbaciti (nakon što smo pregledali dataset)? Napisati razlog zbog čega, te izvršiti trajno odbacivanje te kolone.
    c. Za kolonu *'price'*, napisati osnovne statističke karakteristike kao što su: *mean*, *std*, *min* i *max* vrijednost. Napisati šta svaka od tih vrijednosti predstavlja.
    d. Napraviti novu kolonu nad vašim datasetom pod imenom: '*car_volume'*, koja će biti produkt kolona: *'carlength'*, ' *carwidth'* i *'carheight'* . A nakon toga odbaciti kolone: *'carlength'*, ' *carwidth'* i *'carheight'.* Također možete odbaciti i kolonu pod nazivom: *'symboling'*.
    e. Ispisati red/ove koji u koloni '*CarName*' sadrži/e riječ *'macan'*.
    f. Napraviti novu kolonu nad vašim datasetom pod imenom: *'CarBrand*', tako što ćete na osnovu kolone *'CarName'* uzeti samo prvi element nakon splitanja vrijednosti unutar te kolone (split na osnovu praznog mjesta). A nakon toga odbaciti kolonu *'CarName'*. Te napisati koliko ima jedinstvenih vrijednosti unutar kolone *'CarBrand'*
    g. Izdvojiti samo one redove koji za kolonu brend automobila – imaju vrijednost: „bmw“ i naći: *min*, *max* i srednju vrijednost cijene (*'price'*) za redove pod tim brendom.

___ /35

Svaki tačno urađen podzadatak nosi 5 bodova.

2. *Exploratory data analysis*/vizualizacija podataka:

   a. Kreirati *histplot* za kolonu *'price'* i napisati komentar za dobijeni dijagram (parametar *bins* postaviti na vrijednost 8).
   b. Kreirati *boxplot* za kolonu *'price'* tako da na y-osi bude pomenuta kolona. Napisati komentar za dobijeni dijagram.
   c. Izdvojiti jedinstvene vrijednosti iz kolone *'CarBrand'* te obratiti pažnju na: *mazda, nissan, porsche, toyota i volkswagen/vw.* Šta uočavate? Uraditi odgovarajuću operaciju tako da se dobiju vrijednosti sa jedinstvenim imenima za prethodno pomenute brendove.
      Hint: upotrijebiti metodu replace.
   d. Izvršiti odbacivanje onog reda (brenda automobila) koji sadrži samo jednu vrijednost u koloni *'CarBrand'*.
   e. Kreirati *count* dijagram za broj vozila po kompaniji (brendu).
      Hint: koristiti sljedeće metode (za ljepši prikaz) – ove dvije linije je potrebno dodati nakon što pozovete funkciju koja će vam kreirati dijagram za broj vozila (za svaki brend. Napisati prigodan komentar nakon dijagrama.

      plt.xticks(rotation=90)
      plt.show()

   f. Kreirati *boxplot* dijagram za odnos *'CarBrand'* i *'price'*. Te na osnovu dijagrama komentarisati koji brend ima najnižu cijenu, najveću cijenu, te koji brendovi imaju (potencijalne) *outliere.*

   g. Kreirati *barplot* dijagram za *DataFrame* koji je dobijen grupisanjem po koloni *'CarBrand'*, te urađena srednja vrijednost po koloni *'price'*. Na x-osi treba da bude CarBrand, a a y-osi srednja vrijednost cijene „price“ (poredane po opadajućem redoslijedu). Te na osnovu dobijenog dijagrama komentarisati koji brend automobila imaju najveću srednju cijenu, a koji najmanju.

___ /35

Svaki tačno urađen podzadatak nosi 5 bodova.

3. Kreiranje *machine learning* modela:

   a. Za kolone: '*door_number'*, '*cylindernumber*', *'enginelocation'* i *'fueltype'* - uraditi mapiranje tih kolona, te objasniti razlog korištenja određene mape.
   b. Odbaciti sve one kolone koje nisu numeričkog tipa.
   c. Napisati koje kolone su korelaciji sa kolonom *'price'*, a koje nisu u nikakvoj (ili skoro nikakvoj) korelaciji.
   d. Naći predikciju cijene (*price*) za vrijednost *'enginsize'* = 150, kao i MSE, R^2.
   e. Izabrati *features* za *multiple (linear) regression*.
   f. Uporediti vrijednosti MSE i R^2 iz prethodnog koraka (korak d.) sa MSE i R^2 u slučaju *multiple (linear) regression.*

___ /30

Svaki tačno urađen podzadatak nosi 5 bodova.

Tokom rada spašavati izmjene.

Vaši dokumenti trebaju biti na mreži (This PC > Kursisti) u adekvatnom folderu da bi položili.