Izvestaj o cetvrtom i petom zadatku

Stefan Teslic, 2021/3069

June 17, 2022

1 Proces realizacije Linearne Regresije

Doneta je odluka da se iskoristi *Ridge Regression* umesto standardne linearne regresije. Razlog za to je donet jer cene znaju da budu prilicno velike i treba nam nesto da "smiri" azuriranje tezina. Parametri obucavanja, odnosno, α, λ, i broj iteracija obucavanja su odradjeni pomocu *Grid Search* sa parametrima:

- $\alpha = [0.1, 0.5, 0.01, 0.05, 0.001]$
- $\lambda = [0.1, 0.5, 0.01, 0.05, 0.001]$
- Iterations $= [50, 100, 1000, 5000, 10000]$

Odluka o parametrima zavisi od implementacije *predict* metode. Ukoliko se koristi numpy, odabrana vrednost je $(\lambda, \alpha, \textit{iterations}) = (0.001, 0.5, 10000)$, u slucaju da se koristi cist Pandas, onda vrednosti koje se dobijaju su $(\lambda, \alpha, \textit{iterations}) = (0.5, 0.5, 10000)$.

Ociceni su takodje *outlier-i* nastali greskom unosa, konkretno za kubikazu i kilometrazu. Maksimalna kubikaza je ispod 10000, a ponekad se desava da je daleko iznad i takodje kilometraza je stavljena da bude ispod 500.000km jer se cesto zna desiti da su pogresno unete iako je jasno da postoji situacija kada neka kola imaju toliku kilometrazu.

Nakon svih ovih promena, raspodela podataka je sledeca (pogledati Figure 1) Napomena: Slika je

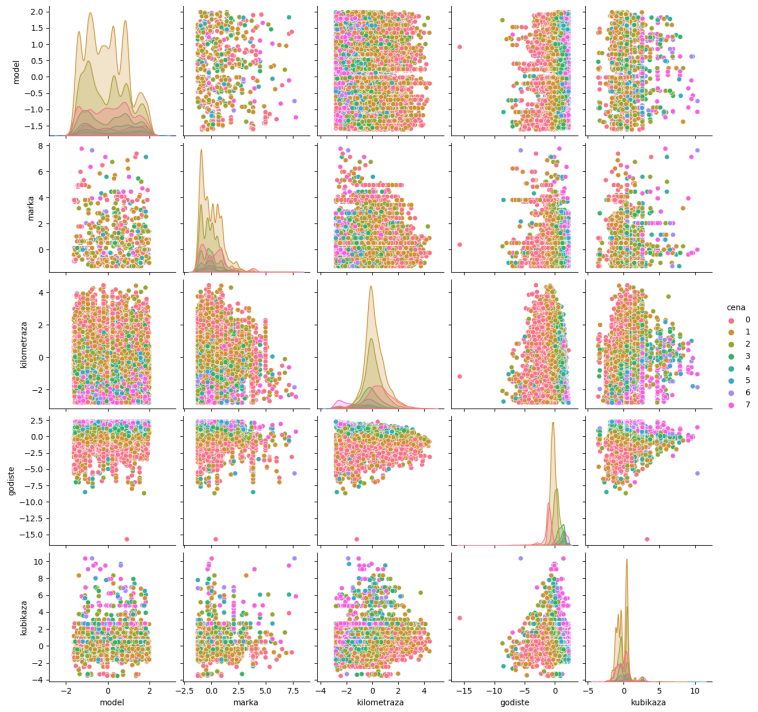

Figure 1: Raspodela podataka

pozajmljena od KNN algoritma, zato mogu da se vide razlicite klase podataka, ali sustinski su podaci ovako rasporedjeni.

Skaliranje koje se pokazalo kao najefektivnije je Min-Max skaliranje, Z-score skaliranje daje primetno losije performance.

Koriscena metrika performansi je R^2, odnosno koeficijent determinacije. Odabrano je ovo jer moze da nam kaze kako se nas model bori sa varijansom i zato sto se testira nad jednim skupom podataka (R^2 ne moze da se koristi za razlicite skupove podataka). Postignute performanse sa numpy implementacijom predikcije i R^2 metrikom su 0.63%

Predikcija nije na mnogo zavidnom nivou. Potencijalno resenje je formirati veci skup podataka i uraditi vecu normalizaciju podataka. U Figure 4 se vidi veliki disbalans podataka.

2 Proces realizacije KNN

Implementirane su Euklidska i Manhattan metrika. Odabir parametra k moze da se odradi na 2 nacina. Prvi je tipican:

- Ukoliko je rezultujuce k neparno $k = \sqrt{n}$
- Ukoliko je rezultujuce k parno $k = \sqrt{n} \rightarrow k_1 = k + 1$

Drugi nacin je radjen pomocu *Grid Search* da se ustanovi najveca preciznost. Dobijeni su grafici iz Figure 2 i Figure 3

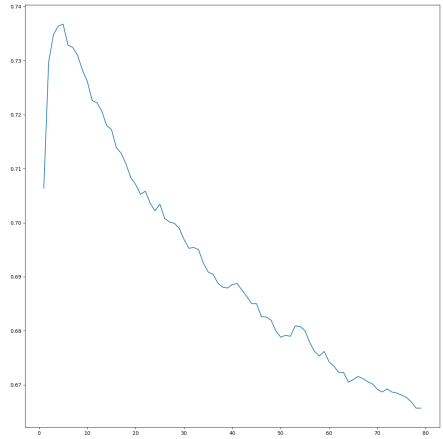

Figure 2: Euklidska metrika.

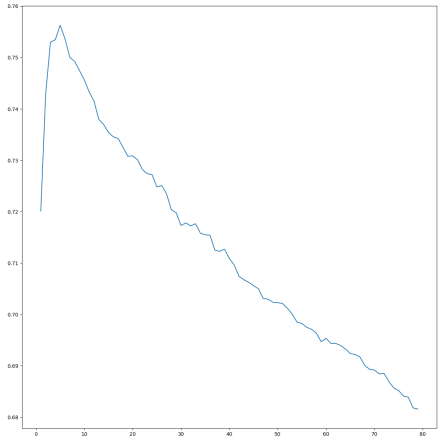

Figure 3: Manhattan metrika.

Testiranje je radjeno sa podelom na test i trening skup sa odnosom $20\% - 80\%$

Najbolje k, na osnovu *Grid Search* je $k = 9$, daje preciznost od oko 0.74%. "Skolsko" $k = 157$ daje preciznost od oko 0.67%.

Formatiranje podataka je odradjeno na isti nacin kao i za linearnu regresiju - stavise, marginalno razlicit upit se koristi za dohvatanje podataka. Skaliranje podataka koje je odradjeno u slucaju KNN je *Z-score*, daje bolje performanse u odnosu na *Min-Max*.

Grupe podataka, odnosno razlicite klase podataka su sledece:

1. manje od 2000
2. izmedju 2 000 i 4 999
3. izmedju 5 000 i 9 999
4. izmedju 10 000 i 14 999
5. izmedju 15 000 i 19 999
6. izmedju 20 000 i 24 999

7. izmedju 25 000 i 29 999
8. 30 000 ili vise

Raspodela automobila po ovim cenama se moze videti iz Figure 4.

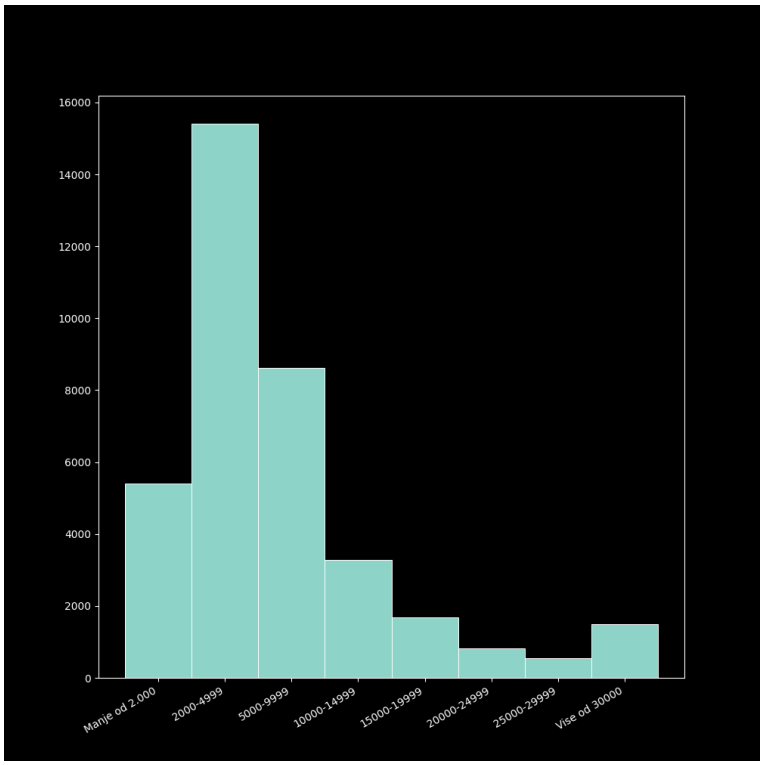

Figure 4: Raspodela automobila po ceni

Potencijalno poboljsanje moze da se postigne slicno kao i kod linearne regresije.