**Zadatak 4:**

Cilj ovog zadatka bilo je odrediti koeficijente theta koji zajedno sa odlikama dovode do predikcije cena automobila. Inicijalno, koeficijenti su postavljeni na nula I oni dobijaju svoju konacnu vrednost (koja ucestvuje u predikciji) tehnikom gradijentnog spusta. Koeficijenti theta smesteni su unutar matrice koja ima za jedan vise polja od broja odlike koji ucestvuju u predickiji (koeficijentu theta0 pridruzena je vrednost 1 za x0). Kako bih postigao sto optimalnije resenje, u obzir su uzimane karakteristike automobila koji na osnovu datog skupa podataka najvise uticu na vrednost cene. Ovime je smanjena preterana prilagodjenost modela trening podacima. Za potrebe toga koriscenja je heatmapa koja ukazuje na to koje odlike su vise povezane sa odlikom za koju vrsimo predikciju.

Chart

Description automatically generated with medium confidence  
Za vrednost brzine ucenja odabrano je 0.01. Za manje vrednosti ucenje postaje sporije, to jest potrebno je obaviti veci broj koraka. Za vece vrednosti brzine ucenja, moze doci do preskakanja lokalnog minimuma ili divergencije. Sto se tice cost funkcije, primecuje se njena konvergencija vec nakon 200 iteracija.

Shape

Description automatically generated

Ono sto je dosta uticalo na rad modela jeste I tehnika normalizacije odlika. Najbolje se pokazao z-score metod odnosno standardizacija.  
Kao metrika za procenu kvaliteta modela koriscen je RMSE (Root Mean Squared Error).  
Ocekivano, sto vise koriscenih podataka za trening uglavnom dovodi do manje greske. U nastavku je tabelarno prikazan procentualni odnos trening-test seta kao I dobijeni RMSE (napomena: s’ obzirom da se vrednosti koriscene za trening odnosno test set generisu proizvoljno, ovi rezultati ne bi bili identicni za svako pokretanje programa).

|  |  |
| --- | --- |
| **Procentualni odnos trening-test seta** | **RMSE** |
| 60-40 | 15613.23 |
| 70-30 | 8729.727 |
| 80-20 | 7819.933 |
| 90-10 | 7708.407 |
| 99-1 | 5282.135 |

**Zadatak 5:**

Osnovni izazov prilikom realizacije KNN algoritma, jeste odabir broja suseda koji ucestvuju u predikciji (K). Program inicijalno za K uzima kvadratni koren broja podataka, ali pruza mogucnost I da korisnik sam odabere vrednost parametra K. S’ obzirom da je vrednost cene numericka, potrebno ju je podeliti u klase prema zadatku 3.E. Kao I u prethodnom zadatku I ovde su takodje koriscenje odlike koje su najvise povezane na osnovu matrice korelacija, kako bi se izbegla preterana prilagodjenost modela. Takodje, koriscena je ista tehnika normalizacije kao u prethodnom zadatku. Prilikom predikcije modela, moze se koristiti euklidska ili Menhetn razdaljina.   
Kao metrika koriscena je tacnost (accuracy). Sto veci broj podataka ode u trening set, ocekivano je da ce tacnost biti veca. Tacnost takodje zavisi I od parametra K. U nastavku je prikazan odnos tacnosti I vrednosti parametra K za procentualni odnos trening-test podataka od 80-20.

Chart, bar chart

Description automatically generated

Ovo ponasanje u teoriji nije ocekivano, jer sa povecanjem vrednosti K ocekivano bi bilo da preciznost raste, medjutim sa grafika mozemo videti da preciznost osciluje oko 0.7.

Chart, bar chart

Description automatically generated

I ovde se takodje primecuje odstupanje od teorijski ocekivanog ponasanja.   
Sto se tice preciznosti, primecujemo da je preciznost priblizno slicna bilo da koristimo Euklidsku ili Menhetn razdaljinu.

Naravno, ova dva modela imaju I svoje nedostatke. Na primer, modeli ne uzimaju u obzir da neko moze greskom da unese cenu automobila koja je drasticno veca od one koja bi trebalo da bude. Takodje, modeli ne uzimaju u obzir oldtajmere. Ocekivano je da stariji auto povlaci sa sobom nesto nizu cenu, medjutim kod oldtajmera to nije slucaj. Dakle, potrebno je nesto detaljnije predprocesiranje podataka u smislu odbacivanja pojedinih. Takodje, kako bi se smanjio stepen overfitting-a, kod linearne regresije moguce je koristiti I tehniku grebene ili LASSO regresije. Takodje, jos jedan nacin sprecavanje overfitting-a bi bilo koriscenje unakrsne validacije.