Domača naloga 1 pri predmetu ITAP

Anej Rozman

Naloga 1

a)

Dobil sem dve celi števili in dve števili, ki sta praktično 0. Skratka model se je popolnoma prilegal podatkom, ker so bili simulirani z linearno funkcijo.
Koeficienti: $[1, 5, -4.00027234 \cdot 10^{-15},\ -7.21644966 \cdot 10^{-16}]$

b)

Pri tem delu naloge sem samo dodal stolpec enic matriki X, saj glede na to da obravnavamo podatke, ki so popolnoma linearno kolerirani med sabo, je to način, da se bo napovedni model (skoraj) popolnoma prilegal točnemu.

d)

Za preverjanje točnosti modela sem uporabil metodo zankanja oz. bootstrap. Glede na prvih par podatkov, nimamo opravka z linearno funkcijo, tako da je napaka nekoliko večja, kot v primeru b.

Naloga 2

a)

Izrišem nekaj standardnih statistik podatkov in jih standardiziram, da so vrednosti med seboj bolj primerljive.

b)

Točnosto modela ocenim s pomočjo prečnega preverjanja in izračunam 'accuracy score'. Nastavim 'random seed' na 42 za ponovljivost rezultatov. Model je stabilen, saj pri 10-kratnem prečnem preverjanju, 'accuracy score' ne preveč varira.

c)

Dodam spremenljivke, ki jih predlaga domenski ekspert in ponovno ocenim model s pomočjo prečnega preverjanja. Tokrat se 'accuracy score'poveča iz 0.959 na 0.987. Glede na koreliranost novih spremenljivk s starimi, sem poskušal z odstranjevanjem starih spremenljivk, ampak izboljšanje je bilo zanemarjljivo oz. ga ni bilo.

d)

Temu delu naloge sem se odločil pristopiti na dva načina.

1. Pogledal sem absolutne vrednosti koeficientov po tem ko sem model večkrat pognal in obdržal koeficiente, ki so bili večji kot 1 (seveda konstante nisem upošteval kot koeficient). Tako sem obdržal le 4 spremenljivke. Natančnost modela pa se praktično ni spremenila, saj je že izhodiščna bila visoka.

2. Treniral sem modele kjer sem izvzel i-to spremenljivko in nato izračunal 'accuracy score' s s prečnim preverjanjem. Nato sem izrisal graf spremembe natančnosti modela glede na izvzeto spremenljivko. Ni presenetljivo, da so na natančnost najbolj vpivale (skoraj vse) enake spremenljivke kot pri prvem pristopu. Zanimivo je, da je izvzetje 2. spremenljivke najbolj izboljšalo oceno modela, saj je imel 'accuracy score'nad 0.99.

Na podlagi zgornjih pristopov lahko sklepamo, da so spremenljivke z indeksi $0, 1, 4, 9$ najbolj pomembne za model, saj izvezetje ostalih ne povzroči vpada v oceni modela oz. koeficient pred njimi so najvišji.

Naloga 3

Nalgo sem začel s tem da sem odstranil spremenljivko 'zaporedna številka diamanta v bazi', saj očitno ni vsebinsko povezana z njegovo ceno (Če bi premešali podatke bi se vpliv popolnoma spremenil) Standardiziral sem napovedne spremenljivke. Nato sem si vizualiziral porazdelitev cene diamantov in izpisal Pearsonov korelacijski koeficient med napovednimi in odvisno spremenljivko, da sem dobil idejo kateri podatki so korelirani in kateri ne. Nato sem izraisal grafe, ki prikazujejo odvisnoti napovednih in odvisne spremenljivke. Nisem ravno upošteval grafov kategoričnih spremenljivk, saj so nekoliko nesmiselni. Poskusil sem ustvariti raznovrstne napovedne spremenljivke, npr. opazil sem, da je odvisnost med stKaratov in ceno kvadratična in sem zato uvedel spremenljivko stKaratov^2, in še marsikaj drugega. Nič ni dobro vplivalo na končno natančnost modelov, zato sem odločil, da ne bom dodal drugih napovednih spremenljivk (Kot primer sem pustil volumen diamantov). Odstranil sem 'outlierje' iz podatkov, saj ti slabo vplivajo na splošnost modela. Kategorične spremenljivke sem spremenil v 'dummy variable'. Končno sem s prečnim preverjanjem ocenil točnost knn regresijskega modela. Optimalen k je 3 in RMSE pri tem ka je približno 630, kar sicer ni zelo dober rezultat (praktično gledano), ampak moji opisani poskusi niso kaj dosti doprinesli k izboljšanju točnosti modela. Model je pa imel minimum variance napak pri $k = 5$ in glede na to da se RMSE ne razlikuje veliko, bi bilo bolj smiselno vzeti kot optimalen model s parametrom $k = 5$. Vseeno sem izbral $k = 3$ in dobil, da je R^2 približno 0.986, kar pomeni, da pojasnimo približno 98% variabilnosti (Precej dober rezultat).