Poročilo o nalogi 4

Tim Kmecl, UOZP 2022

Obdelava podatkov

Podatke iz csv datoteke sem prebral v pandas dataframe. Ciljna spremenljivka y je razlika med časom prihoda in odhoda (trajanje vožnje), za nadaljnjo obdelavo pa ohrani še linijo in datum/čas odhoda. Doda sledeče 0/1 atribute: ali je dan v tednu eden od prvih štirih dni (torej delovnik brez petka), petek, sobota ali nedelja. Ti atributi bodo torej imeli vrednost 1, če je vožnja na ustrezen dan v tednu, in 0 sicer.

Iz spleta sem pridobil še podatke o šolskih počitnicah tega leta. Dodaten atribut počitnice bo enak 1, če je vožnja v času počitnic, sicer 0. Podobno sem naredil še za praznike, ki so bili takrat med tednom (ne pa tudi v soboto ali nedeljo, saj sem predpostavil, da bo vpliv praznika dosti večji med tednom).

Čas odhoda sem pretvoril v dva številska atributa, in sicer minuto v dnevu in mesec, pri čemer sem štetje mesecev zamaknil tako, da se začne z junijem (indeks 0). To sem storil, ker se mi je zdelo smiselno, da so zimski meseci bolj blizu skupaj (pri običajnem štetju mesecev namreč december in januar nista en zraven drugega), zaradi kasnejše neuporabe poletnih mesecev pri učenju pa podoben preskok poleti ne bi smel delati težav.

Zadnji od atributov je linija (kombinacija št. linije in smeri, ne zgolj št. linije).

Napovedovanje

Za napovedovanje sem uporabil Gradient Boosting Trees (vgrajen v modul scikit-learn - specifično Histogram-based gradient boosting regressor tree, ki na velikih naborih podatkov dela hitreje od navadnih GBT), ki ga je na predavanjih posebej priporočil profesor.

Vse atribute razen časa v dnevu in številke meseca model obravnava kot kategorični atribut. S poskušanjem na lastni delitvi danih podatkov na testno in učno množico sem največje število dreves nastavil na 500 (pri večjem je bil rezultat na moji testni množici že slabši, verjetno je do neke mere prihajalo do overfittinga), minimalno število primerkov v listu pa na 5. Ostale parametre sem pustil na privetih vrednostih.

Pri učenju nisem uporabil poletnih mesecev učne množice (maj, junij, julij, avgust), saj so večji del tega časa šolske počitnice, prav tako pa je takrat lepše vreme, kar bi lahko slabo vplivalo na napovedovanje decembra.

S tem pristopom sem na lestvici dosegel 158.