1. domača naloga -Izbrane teme iz analize podatkov

Metod Jazbec, 27151006

Šolsko leto 2017/2018

Podatke sem naprej uvozil in naredil osnovni 'preprocessing' (odstranitev čudnih vnosov itd.). Potem ko sem si definiral funkcijo *F_ beta*, sem najprej poskusil naučiti model z uporabo *kNN* metode. Mera natančnosti (torej F_2) se je gibala okoli 0.65. Opazil sem tudi nenavadno obnašanje, saj sem ob različnih zagonih modela za optimalno vrednost parametra k dobival različne vrednosti (varirale so vse od 3 pa do približno 120).
Za naslednjo metodo sem vzel linearni model (torej logistično regresijo, ker gre za problem klasifikacije). Poleg modela, ki je vseboval le linearne člene, sem poizkusil še s členi višjih stopenj, vendar se je izkazalo, da najboljši rezultat da osnovna verzija. F_2 je tu znašal okoli 0.68. Model bi lahko morda še izboljšal, če bi poleg členov višjih stopenj dodajal še interaktivne člene (produkte) ali pa bi določene spremenljivke celo odstranil iz modela (ti. backward elimination pristop).
Na koncu sem uporabil še *random forest* model, vendar nisem izboljšal natančnosti linearnega modela. Tudi tu sem, podobno kot pri *kNN*, dobival različne vrednosti parametra *mtry* ob različnih zagonih. Kot končni model za napoved na testni množici sem tako uporabil linearni model. Poročilu prilagam csv datoteko z napovedmi.
Natančnost bi verjetno lahko izboljšal, če bi podatke na začetku normaliziral ali standariziral. Seveda bi lahko poskusil uporabiti tudi kakšen drug model (SVM, odločitvena drevesa). Vsa koda je dostopna tukaj.

Menim, da bi morala biti vrednost *F_ beta* na oddanih podatkih podobna kot tista, ki sem jo dobil pri testiranju modela logistične regresije (torej okoli 0.68). Pri treniranju modelov sem namreč uporabil prečno preverjanje, ki učno množico razbije na podmnožice in nato vrne povprečje napak izračunanih na podmnožicah, ki niso bile vključene v treniranje modela.