Seminarska naloga ITAP - 1.del

Vito Rozman

12. junij 2022

Podatki Prvi korak je bil zmašanje število vrstic, ker pri velikem številu podatkov računalnik ni prenesel računske zahtevnosti. Iz vseh podatkov z referenco sem vzel 1% podatkov, nato sem jih razdelil na učno in testno množico v razmerju $3:1$, torej 75% učna in 25% testna.

Metrika Za primerjavo modelov sem izbral metriko natančnosti prečnega preverjanja, natančnost na testni množici in ploščina pod ROC krivuljo (AUC). Pri modelu za napovedovanje pozidanega območja sem primerjal tudi glede občutljivosti in specifičnosti.

Modeli Primerjal sem modele *glm*, *knn*, *svm* in *rf*. Med primerjavo sem optimiziral še parametre pri *knn* (število sosedov) in pri *svm* (kompleksnost). Dobil sem naslednje rezultate:

Model - gozd	GLM	KNN	SVM	RF
Natančnost - testna	0.9739	0.9717	0.9772	0.9770
Natačnost - cv	0.9737	0.9715	0.9756	0.9764
Ploščina pod ROC (AUC)	0.9955	0.9903	0.9957	0.9963

Model - pozidano	GLM	KNN	SVM	RF
Natančnost - testna	0.9242	0.9016	0.9362	0.9377
Natačnost - cv	0.9195	0.9204	0.9271	0.9355
Ploščina pod ROC (AUC)	0.9651	0.9235	0.9702	0.9793
Občutljivost	0.8863	0.9722	0.9218	0.9262
Specifičnost	0.9327	0.9213	0.9353	0.9483

Najboljši model: Kot vidno zgoraj se je v obeh primerih najbolje izkazal *rf* (naključni gozdovi), saj je pri najpomembnejših metrikah dobil najboljše rezultate. Po izbiri modela sem optimiziral parametra mtry in ntree. Edina slaba lastnost *rf* se je pokazala pri časovni zahtevnosti in težavi pri velikem številu vrstic. V primeru da bi optimiziral še časovno komponento, bi izbral *glm* ali pa *svm*.

Za napovedovanje pozidanega območja, sem moral podatke malo prilagoditi, ker bi se lahko pojavila težava pri prekomernem napovedovanju negativnih izidov, saj je razmerje pozidanega (pozitivni) proti nepozidanemu (negativni) precej manjša kot pri podatkih za gozdove. Uporabil sem metodo podvzorčenja.

Končni model za napoved gozd sem natreniral na 20% vseh podatkov z referenco (ker mi računalnik ni dopuščal več), za napoved pozidanega območje pa sem vzel 90% vseh podatkov z referenco na katerih sem uporabil podvzorčenje.