Izbira najboljšega modela

Podatki

Najprej sem izbral pod množico podatkov, ker pri velikem številu podatkov računalnik ni prenesel velikosti. Iz vseh podatkov z referenco sem vzel 1% podatkov, nato sem jih razdelil na učno in testno množico v razmerju 3:1, torej 75% učni.

Metrika

Naslednji korak je bil izbira metrik za primerjavo modelov. Za primerjavo sem izbral natančnost prečnega preverjanja, natančnost na testni množici in ploščina pod ROC krivuljo (AUC). Pri modelu za napovedovanje pozidanega območja sem primerjal tudi glede sensitivity in specifity.

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:

Najboljši model

Kot vidno zgoraj se je najbolje izkazal rf, kjer je pri najpomembnejših metrikah dobil najboljše rezultate. Optimiziral sem parametra mtry in ntree. ­Edina slaba lastnost rf se je pokazala pri časovni zahtevnosti in težavi pri veliko številu vrstic.

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 pod vzorčenja.

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