starter\_gbm.R

Administrator

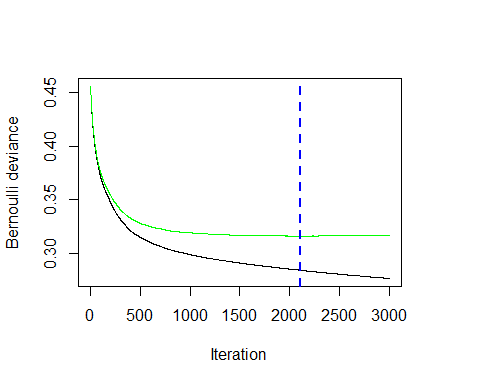
Sun May 17 12:09:21 2015

library(Metrics)  
library(data.table) ## load data in quickly with fread  
setwd("E:/Dropbox/kaggle/West Nile Virus Prediction")  
x <- fread("data/train3.csv")  
test <- fread("data/test3.csv")  
  
## prep the species column by moving the test-only UNSPECIFIED CULEX to CULEX ERRATICUS, and re-doing the levels  
## logistic regression will complain otherwise  
vSpecies<-c(as.character(x$Species),as.character(test$Species))  
vSpecies[vSpecies=="UNSPECIFIED CULEX"]<-"CULEX ERRATICUS"  
vSpecies[-which(vSpecies == "CULEX PIPIENS" |  
 vSpecies == "CULEX PIPIENS/RESTUANS" |  
 vSpecies == "CULEX RESTUANS")] = "CULEX OTHER"  
vSpecies<-factor(vSpecies,levels=unique(vSpecies))  
  
## data.table syntax for adding a column; could overwrite the existing column as well  
x[,Species2:=factor(vSpecies[1:nrow(x)],levels=unique(vSpecies))]

#test[,Species2:=factor(vSpecies[(nrow(x)+1):length(vSpecies)],levels=unique(vSpecies))]  
  
# we'll set aside 2011 data as test, and train on the remaining  
my.x = data.frame(x[,list(WnvPresent,Year,Week, Species2, Latitude, Longitude,Block,NumMosquitos,Tmax, Tmin, Tavg, Depart, DewPoint,WetBulb,Heat,Cool,PrecipTotal)])  
x1<-my.x[x$Year!=2011,]  
x2<-my.x[x$Year==2011,]  
  
## GAMboost modelling  
require(gbm)

## Loading required package: gbm  
## Loading required package: survival  
## Loading required package: lattice  
## Loading required package: splines  
## Loading required package: parallel  
## Loaded gbm 2.1.1

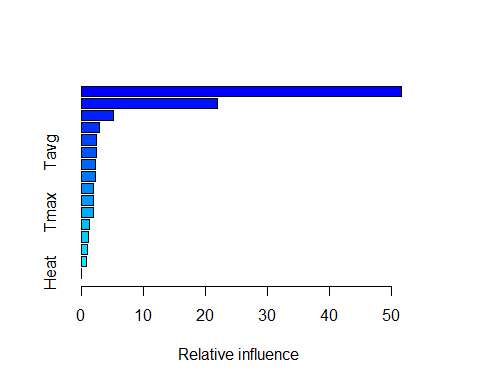
set.seed(2)  
fitCv = gbm(WnvPresent ~ Year+Week +Species2+NumMosquitos+Latitude+Longitude+Block+Tmax+ Tmin+ Tavg+ Depart+DewPoint+WetBulb+Heat+Cool+PrecipTotal, data = x1, n.trees = 3000, interaction.depth = 2,shrinkage = 0.01,cv.folds=5,distribution = "bernoulli")  
  
#distribution = "adaboost"  
best.iter <- gbm.perf(fitCv,method="cv")



best.iter

## [1] 2109

summary(fitCv,n.trees=best.iter)

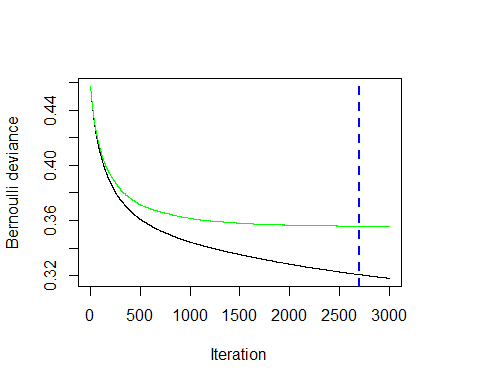


## var rel.inf  
## NumMosquitos NumMosquitos 51.5383507  
## Week Week 21.8755968  
## Longitude Longitude 5.1103265  
## DewPoint DewPoint 2.8495506  
## Year Year 2.3360538  
## Tavg Tavg 2.2886759  
## Block Block 2.2489705  
## Latitude Latitude 2.2144810  
## Tmin Tmin 1.9003675  
## Species2 Species2 1.8543285  
## Tmax Tmax 1.8033913  
## WetBulb WetBulb 1.2807802  
## PrecipTotal PrecipTotal 1.0350333  
## Depart Depart 0.8881890  
## Cool Cool 0.7759043  
## Heat Heat 0.0000000

p2<-predict(fitCv, newdata = x2,n.trees = best.iter, type = "response")  
auc(x2$WnvPresent,p2)

## [1] 0.9127568

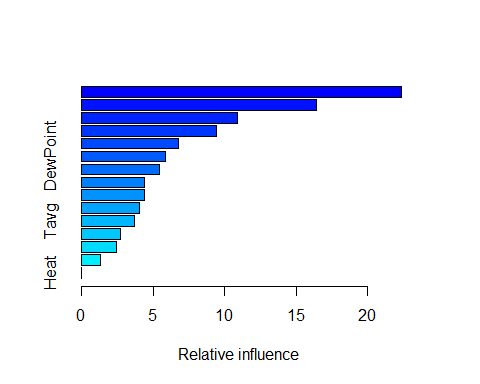
set.seed(2)  
fitCv = gbm(WnvPresent ~ Year+Week +Species2+Latitude+Longitude+Block+Tmax+ Tmin+ Tavg+ Depart+DewPoint+WetBulb+Heat+Cool+PrecipTotal, data = x1, n.trees = 3000, interaction.depth = 2,shrinkage = 0.01,cv.folds=5,distribution = "bernoulli") #distribution = "adaboost"  
best.iter <- gbm.perf(fitCv,method="cv")



best.iter

## [1] 2696

summary(fitCv,n.trees=best.iter)



## var rel.inf  
## Week Week 22.352107  
## Longitude Longitude 16.414952  
## Latitude Latitude 10.922866  
## Year Year 9.444314  
## Tmin Tmin 6.776944  
## DewPoint DewPoint 5.821737  
## Species2 Species2 5.415356  
## WetBulb WetBulb 4.407530  
## Block Block 4.402174  
## Tmax Tmax 4.036639  
## Tavg Tavg 3.652658  
## Cool Cool 2.673376  
## PrecipTotal PrecipTotal 2.398743  
## Depart Depart 1.280604  
## Heat Heat 0.000000

p2<-predict(fitCv, newdata = x2,n.trees = best.iter, type = "response")  
auc(x2$WnvPresent,p2)

## [1] 0.8104933

if(1==0)  
{  
# dWeek +Species2+Latitude+Longitude+Block+NumMosquitos  
## now fit a new model to all the data, so that our final submission includes information learned from 2011 as well  
fitSubmit <- gbm(WnvPresent ~ dYear+dWeek +Species2+Latitude+Longitude+Block+NumMosquitos, data = my.x, n.trees = 12000, interaction.depth = 2,cv.folds=5,distribution = "bernoulli")  
best.iter <- gbm.perf(fitSubmit,method="cv")  
best.iter  
summary(fitSubmit,n.trees=best.iter)  
pSubmit<-predict(fitSubmit, newdata = test, type = "response")  
  
submissionFile<-cbind(test$Id,pSubmit)  
colnames(submissionFile)<-c("Id","WnvPresent")  
options("scipen"=100, "digits"=8)  
write.csv(submissionFile,"submission/starter\_gbm.csv",row.names=FALSE,quote=FALSE)  
}