starter\_GAM22.R

Administrator

Sun May 24 16:29:07 2015

library(Metrics)  
library(data.table) ## load data in quickly with fread  
#setwd("E:/Dropbox/kaggle/West Nile Virus Prediction")  
#setwd("C:/Users/tshao/Dropbox/kaggle/West Nile Virus Prediction")  
x <- fread("E:/Dropbox/kaggle/West Nile Virus Prediction/data/train6.csv")  
test <- fread("E:/Dropbox/kaggle/West Nile Virus Prediction/data/test5.csv")  
  
  
  
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))  
  
  
x[,Species:=factor(vSpecies[1:nrow(x)],levels=unique(vSpecies))]

# we'll set aside 2011 data as test, and train on the remaining  
my.x = data.frame(x[,list(WnvPresent,WnvPresent1,Year,Week, Species, Latitude, Longitude,Block,NumMosquitos, DewPoint)])  
x1<-my.x[x$Year!=2011,]  
x2<-my.x[x$Year==2011,]  
  
  
#,Tmax, Tavg, DewPoint,WetBulb,Heat,Cool,PrecipTotal  
  
## GAM modelling  
require(gam)

## Loading required package: gam  
## Loading required package: splines  
## Loading required package: foreach  
## Loaded gam 1.12

fitCv = gam(WnvPresent ~ s(Week) + Species + lo(Latitude, Longitude)+s(Block)+s(Week):Species+s(NumMosquitos)+s(DewPoint), data = x1)  
p2<-predict(fitCv, newdata = x2, type = "response")  
## check for a reasonable AUC of the model against unseen data (2011)  
auc(x2$WnvPresent,p2)

## [1] 0.8695531

## now fit a new model to all the data, so that our final submission includes information learned from 2011 as well  
  
fitSubmit1 <- gam(WnvPresent ~ s(Year)+s(Week) + Species + lo(Latitude, Longitude)+s(Block)+s(Week):Species+s(NumMosquitos), data = my.x)  
  
pSubmit<-predict(fitSubmit1, newdata = test, type = "response")  
pSubmit=exp(pSubmit)/3  
summary(pSubmit)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3041 0.3476 0.3650 0.3694 0.3853 0.7159

submissionFile<-cbind(test$Id,pSubmit)  
colnames(submissionFile)<-c("Id","WnvPresent")  
options("scipen"=100, "digits"=8)  
write.csv(submissionFile,"E:/Dropbox/kaggle/West Nile Virus Prediction/submission/GAM\_train6\_test5\_speciescombined.csv",row.names=FALSE,quote=FALSE)