starter\_GAM1.R

Administrator

Mon May 11 22:13:21 2015

library(Metrics)  
library(data.table) ## load data in quickly with fread  
setwd("E:/Dropbox/kaggle/West Nile Virus Prediction")  
x <- fread("data/train.csv")  
test <- fread("data/test\_GAM.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))]

## also add some fields for components of the date using simple substrings  
x[,dMonth:=as.factor(paste(substr(x$Date,6,7)))]

x[,dYear:=as.numeric(paste(substr(x$Date,1,4)))]

x$Date = as.Date(x$Date, format="%Y-%m-%d")  
xsDate = as.Date(paste0(x$dYear, "0101"), format="%Y%m%d")  
x$dWeek = as.numeric(paste(floor((x$Date - xsDate + 1)/7)))  
  
test[,dMonth:=as.factor(paste(substr(test$Date,6,7)))]

test$Date = as.Date(test$Date, format="%Y-%m-%d")  
tsDate = as.Date(paste0(test$dYear, "0101"), format="%Y%m%d")  
test$dWeek = as.numeric(paste(floor((test$Date - tsDate + 1)/7)))  
## train set  
x$TrapNumber <- as.integer(substr(as.character(x$Trap), 2, 4))  
  
## test set  
test$TrapNumber <- as.integer(substr(as.character(test$Trap), 2, 4))  
  
  
  
# we'll set aside 2011 data as test, and train on the remaining  
my.x = data.frame(x[,list(WnvPresent, dYear,dMonth,dWeek, Species2, Latitude, Longitude,Block,TrapNumber,AddressAccuracy,NumMosquitos)])  
x1<-my.x[x$dYear!=2011,]  
x2<-my.x[x$dYear==2011,]  
  
## GAM modelling  
  
require(mgcv)

## Loading required package: mgcv  
## Loading required package: nlme  
## This is mgcv 1.8-6. For overview type 'help("mgcv-package")'.

fitCv = gam(WnvPresent ~ s(dWeek,k=3) +Species2 +s(Block,k=5)+s(NumMosquitos,k=5)+s(TrapNumber,k=3), data = x1, family = binomial)  
p2<-predict(fitCv, newdata = x2, type = "response")  
## check for a reasonable AUC of the model against unseen data (2011)  
auc(x2$WnvPresent,p2)

## [1] 0.9049935

##s(dWeek,Species2,bs="fs")+s(NumMosquitos,Species2,bs="fs")+ s(Latitude, Longitude)+s(TrapNumber)  
## now fit a new model to all the data, so that our final submission includes information learned from 2011 as well  
fitSubmit <- gam(WnvPresent ~ s(dWeek,k=3) +Species2 +s(Block,k=5)+s(NumMosquitos,k=5)+s(TrapNumber,k=3), data = my.x, family = binomial)  
pSubmit<-predict(fitSubmit, newdata = test, type = "response")  
## look at the predicted distribution (AUC doesn't care about probabilities; just ordering. It's still a good diagnostic)  
summary(pSubmit)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000000 0.000000 0.000000 0.004267 0.002735 0.360400

submissionFile<-cbind(test$Id,pSubmit)  
colnames(submissionFile)<-c("Id","WnvPresent")  
options("scipen"=100, "digits"=8)  
write.csv(submissionFile,"submission/submitGAM\_5\_11.csv",row.names=FALSE,quote=FALSE)