20160315_statsathon

Andrew Nguyen

2016-March-15

load libraries

```
library(ggplot2)#plotting
library(rpart)
library(tree)
library(randomForest)
```

```
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

```
#library(dismo)
#library(gbm)
#library(caret)
library(ipred)
library(rpart.plot)
```

data

```
training<-read.csv("AIS_train70.csv")
str(training)
```

```
## 'data.frame':    174671 obs. of  1232 variables:
##  $ INC_KEY  : int  13000003 13000007 13000015 13000021 13000024 13000028 13000029 13000037 13000046
##  $ died     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ AIS110099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110600: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110800: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110806: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110808: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS113000: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS115099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS115999: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS116002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS116004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120206: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120806: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121006: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122006: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122406: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130608: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140206: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140208: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140210: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140212: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140214: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140216: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140218: int  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```r
dim(training)
```

```
## [1] 174671   1232
```

```r
test<-read.csv("AIS_test30.csv")
str(test)
```

```
## 'data.frame':    74858 obs. of  1231 variables:
##  $ INC_KEY  : int  13000000 13000005 13000013 13000026 13000045 13000062 13000065 13000079 13000082
##  $ AIS110099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110600: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110800: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110806: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS110808: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS113000: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS115099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS115999: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS116002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS116004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120206: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120806: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS120899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121006: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS121899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122006: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122406: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS122899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130606: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130608: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS130899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131002: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131004: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131402: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131499: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131602: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131604: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131802: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131804: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS131899: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132099: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132299: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132404: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS132699: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140202: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140204: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140206: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140208: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140210: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140212: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140214: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140216: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140218: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AIS140299: int  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```r
dim(test)
```

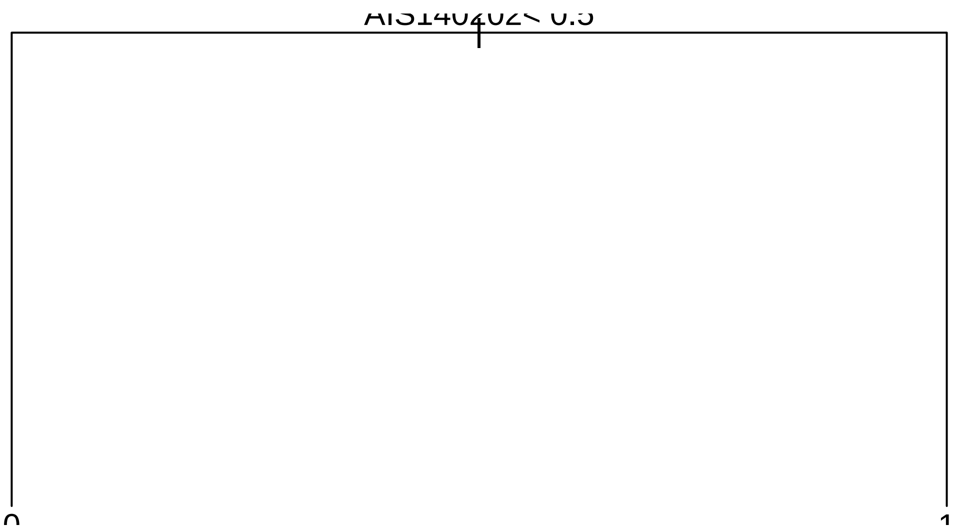
```
## [1] 74858  1231
```

trying boosted regression trees

```r
t<-head(training,10000)

#
#brt.mod3<-gbm.step(data=training,gbm.x= 3:1231,gbm.y= 2,family="bernoulli",tree.complexity=1,learning.

####
#pred<-predict(brt.mod3,vars[,-1],n.trees=brt.mod3$gbm.call$best.trees,type="response")

#d<-as.data.frame(cbind(vars[,1],pred))
#d$pred_point5<-ifelse(d$pred > 0.5,1,0)
#"good" predictability at .5 prob of finding species cut off
#sum(ifelse(d$V1==d$pred_point5,1,0))/nrow(d)
```

Trying rpart

```
###try rpart
form<-as.formula(died~.)
testing<-rpart(form,data=training[,-1],control=rpart.control(minsplit=1),method="class")

#quick and dirty plots
plot(testing)
text(testing)
```



```
printcp(testing) # look at complexity parameter and cross validation error
```

```
##
## Classification tree:
## rpart(formula = form, data = training[, -1], method = "class",
##     control = rpart.control(minsplit = 1))
##
## Variables actually used in tree construction:
## [1] AIS140202
##
## Root node error: 7154/174671 = 0.040957
##
## n= 174671
##
##         CP nsplit rel error xerror     xstd
## 1 0.011602      0    1.0000 1.0000 0.011578
## 2 0.010000      1    0.9884 0.9884 0.011514
```

```
#predict training set
pred<-predict(testing,training[,-1:-2],type="class")

#accuracy
sum(ifelse(training$died==pred,1,0))/length(pred)
```

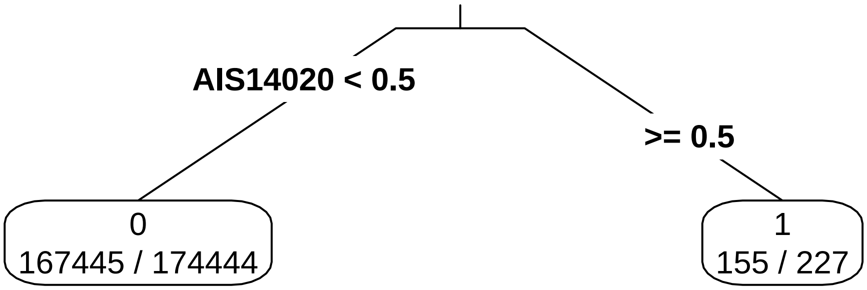
```
## [1] 0.9595182
```

```
#confusion matrix
table(training$died,pred)
```

```
##    pred
##          0      1
##   0 167445     72
##   1   6999    155
```

```
rpart.plot(testing,type=3,extra=2,main="Displays the classification rate at the node,\n expressed as nu
```

**Displays the classification rate at the node,
expressed as number of correct classifcations
and number of observations in the node**



```
#let's predict the test set
pred.test.set<-predict(testing,test[,-1],type="class")
```

Final answer

```
# call on pred.test.set
dat<-as.data.frame(cbind(test[,1],pred.test.set))
names(dat)<-c("ais_data_test30.INC_KEY","died")
write.csv(dat,"20160321_ANBE_model_predictions.csv",row.names=FALSE)
```