HW02 - Logistic Regression

Kyle Barisone

```r
knitr::opts_chunk$set(warning=FALSE, message=FALSE)
depress <- read.delim("https://norcalbiostat.netlify.com/data/depress_081217.txt",
                  sep="\t", header=TRUE)


library(kableExtra)
library(ggplot2)
library(dplyr)
library(leaps)
library(pander)
library(ROCR)
library(caret)
```

##PMA 12.1 p = .20 = 1/5 so the odds are: 5 - 1 = 4 so the odds are 1:4.

I would rather be told the second option because it sounds as if for every 4 unsuccessful attempts, I am guaranteed at least 1 hit whereas if the probability of getting a hit is 20% i may go much longer than 4 unsuccessful attempts before i get a hit.

PMA6 12.7 & 12.8

#First we will fit the acute model

```r
acute.model <- glm(formula = acuteill ~ age + educat + income + cesd + drink,
                      family = 'binomial', data = depress)

backward_aic_model <- step(acute.model, direction = "backward", trace = 1)
```

```
## Start:  AIC=353.12
## acuteill ~ age + educat + income + cesd + drink
## 
##          Df Deviance    AIC
## - income  1   331.25 351.25
## - drink   1   331.53 351.53
## - educat  6   342.81 352.81
## <none>        331.12 353.12
## - cesd    1   333.62 353.62
## - age     1   340.63 360.63
## 
## Step:  AIC=351.25
## acuteill ~ age + educat + cesd + drink
## 
##          Df Deviance    AIC
## - drink   1   331.64 349.64
## - educat  6   342.81 350.81
## <none>        331.25 351.25
## - cesd    1   333.98 351.98
## - age     1   340.64 358.64
##
## Step:  AIC=349.64
## acuteill ~ age + educat + cesd
##
##          Df Deviance    AIC
## - educat  6   343.01 349.01
## <none>        331.64 349.64
## - cesd    1   334.32 350.32
## - age     1   341.64 357.64
##
## Step:  AIC=349.01
## acuteill ~ age + cesd
##
##        Df Deviance    AIC
## - cesd  1   344.50 348.50
## <none>      343.01 349.01
## - age   1   353.97 357.97
##
## Step:  AIC=348.5
## acuteill ~ age
##
##        Df Deviance    AIC
## <none>      344.50 348.50
## - age   1   357.13 359.13
```

```r
pander(summary(acute.model))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1762	1.315	-0.134	0.8934
age	-0.02503	0.008327	-3.006	0.002649
educatBS	-0.469	1.218	-0.3852	0.7001
educatHS Grad	0.2303	1.165	0.1976	0.8434
educatMS	1.404	1.299	1.081	0.2797
educatPhD	-1.05	1.585	-0.6623	0.5077
educatSome college	0.1622	1.205	0.1346	0.8929
educatSome HS	-0.356	1.194	-0.2981	0.7656
income	-0.003649	0.009908	-0.3683	0.7126
cesd	0.02412	0.01518	1.589	0.112
drink	0.2256	0.3526	0.64	0.5222

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	357.1 on 293 degrees of freedom
Residual deviance:	331.1 on 283 degrees of freedom

#Next we fit the chronic model

```
chronic.model <- glm(formula = chronill ~ age + educat + income + cesd + drink,
                      family = 'binomial', data = depress)

backward_aic_model <- step(chronic.model, direction = "backward", trace = 1)
```

```
## Start:  AIC=399.42
## chronill ~ age + educat + income + cesd + drink
## 
##          Df Deviance    AIC
## - educat  6   384.08 394.08
## - income  1   377.50 397.50
## - drink   1   378.52 398.52
## <none>        377.42 399.42
## - cesd    1   383.11 403.11
## - age     1   395.32 415.32
## 
## Step:  AIC=394.08
## chronill ~ age + income + cesd + drink
## 
##          Df Deviance    AIC
## - income  1   384.08 392.08
## - drink   1   385.82 393.82
## <none>        384.08 394.08
## - cesd    1   389.08 397.08
## - age     1   400.45 408.45
## 
## Step:  AIC=392.08
## chronill ~ age + cesd + drink
## 
##         Df Deviance    AIC
## - drink  1   385.85 391.85
## <none>       384.08 392.08
## - cesd   1   389.36 395.36
## - age    1   401.32 407.32
## 
## Step:  AIC=391.85
## chronill ~ age + cesd
## 
##        Df Deviance    AIC
## <none>      385.85 391.85
## - cesd  1   391.30 395.30
## - age   1   405.00 409.00
```

```
pander(summary(chronic.model))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.206	1.112	-1.983	0.04737
age	0.03117	0.00762	4.091	4.302e-05
educatBS	0.7286	1.002	0.7269	0.4673
educatHS Grad	0.968	0.9565	1.012	0.3115

	Estimate	Std. Error	z value	Pr(>|z|)
educatMS	1.643	1.124	1.462	0.1438
educatPhD	-0.3894	1.268	-0.307	0.7588
educatSome college	1.055	0.9994	1.056	0.2909
educatSome HS	0.6696	0.9701	0.6903	0.49
income	-0.002706	0.009257	-0.2924	0.77
cesd	0.03492	0.01498	2.331	0.01976
drink	-0.3288	0.3147	-1.045	0.2961

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	407.5 on 293 degrees of freedom
Residual deviance:	377.4 on 283 degrees of freedom

Depression (CESD) and age are both good predictors however depression seems to predict the chronic model much more accurately than the accute model. Meaning depression could have an effect on whether or not someone is chronically ill.

##3 a)

```
EQ <- read.delim("https://norcalbiostat.netlify.com/data/Earthq.txt",
                  sep="", header=TRUE)
EQ$NEWETHN <- ifelse(EQ$NEWETHN==".", NA, EQ$NEWETHN)
EQ$RSEX <- factor(EQ$RSEX, labels=c("male", "female"))
EQ$NEWETHN <-factor(EQ$NEWETHN, labels = c("White","Hispanic", "Native American", "Asian/P
acific Islander", "African American", "Other"))

summary(EQ$NEWETHN)
```

```
##                    White                 Hispanic          Native American
##                      247                      150                        5
## Asian/P\nacific Islander         African American                    Other
##                       31                       46                        6
##                     NA's
##                       21
```

b)

```
#removes other variables other than own and rent
EQ <- EQ[EQ$V449 %in% c("1", "5"),]

EQ$V449 <- factor(EQ$V449, labels=c("own", "rent"))
summary(EQ$V449)
```

```
##  own rent
##  236  265
```

```
EQ$W238 <- factor(EQ$W238, labels=c("yes", "no", "NA"))
inj.model <- glm(W238 ~ V449 + RSEX + NEWETHN + RAGE, data=EQ, family="binomial")
```

#creates table or.out <- data.frame(ODDS = exp(coef(inj.model)), Lower_control = exp(confint(inj.model))[,1], Upper_control = exp(confint(inj.model))[,2], p = format.pval(coef(summary(inj.model))[,4], digits=1, eps=.001))

rownames(or.out) <- c(“Intercept”, “Rent”, “Female”, “Hispanic”, “Native American”, “Asian/Pacific Islander”, “African American”, “Other”, “Age”) kable(or.out[-1,], digits=2) %>% kable_styling(full_width = FALSE, “striped”) %>% add_header_above(c(" “=2,”95% CI“=2,” "=1))

This code was giving me an error saying fitted probabilityies numerically 0 or 1 occured so it wouldnt knit. I couldnt figure out the problem

```
EQ$RAGE <- as.numeric(EQ$RAGE)
pander(summary(EQ$RAGE))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	12	20	24.29	34	70

```
#removed the outcome "don't know" since we are not interested in this outcome

EQ$V173 <- ifelse(EQ$V173=="8", NA, EQ$V173)
EQ$V173 <- factor(EQ$V173, labels=c("yes", "no"))
```

D)

```
#find best predictive variables
summary(regsubsets(V173 ~ MMI + V127 + W220 + W238 + RAGE + RSEX + V461 + NEWETHN + V455 + V449,
data=EQ))
```

```
## Subset selection object
## Call: regsubsets.formula(V173 ~ MMI + V127 + W220 + W238 + RAGE + RSEX +
##     V461 + NEWETHN + V455 + V449, data = EQ)
## 15 Variables  (and intercept)
##                                 Forced in Forced out
## MMI                                 FALSE      FALSE
## V127                                FALSE      FALSE
## W220                                FALSE      FALSE
## W238no                              FALSE      FALSE
## W238NA                              FALSE      FALSE
## RAGE                                FALSE      FALSE
## RSEXfemale                          FALSE      FALSE
## V461                                FALSE      FALSE
## NEWETHNHispanic                     FALSE      FALSE
## NEWETHNNative American              FALSE      FALSE
## NEWETHNAsian/P\nacific Islander     FALSE      FALSE
## NEWETHNAfrican American             FALSE      FALSE
## NEWETHNOther                        FALSE      FALSE
## V455                                FALSE      FALSE
## V449rent                            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          MMI V127 W220 W238no W238NA RAGE RSEXfemale V461 NEWETHNHispanic
## 1  ( 1 ) " " "*"  " "  " "    " "    " "  " "        " "  " "
## 2  ( 1 ) " " "*"  " "  "*"    " "    " "  " "        " "  " "
## 3  ( 1 ) "*" "*"  " "  "*"    " "    " "  " "        " "  " "
## 4  ( 1 ) "*" "*"  " "  "*"    " "    "*"  " "        " "  " "
## 5  ( 1 ) "*" "*"  " "  "*"    " "    " "  " "        "*"  " "
## 6  ( 1 ) "*" "*"  " "  "*"    "*"    "*"  " "        "*"  " "
## 7  ( 1 ) "*" "*"  " "  "*"    "*"    " "  " "        "*"  "*"
## 8  ( 1 ) "*" "*"  " "  "*"    "*"    " "  " "        "*"  "*"
##          NEWETHNNative American NEWETHNAsian/P\nacific Islander
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    "*"
## 8  ( 1 ) " "                    "*"
##          NEWETHNAfrican American NEWETHNOther V455 V449rent
## 1  ( 1 ) " "                     " "          " "  " "
## 2  ( 1 ) " "                     " "          " "  " "
## 3  ( 1 ) " "                     " "          " "  " "
## 4  ( 1 ) " "                     " "          " "  " "
## 5  ( 1 ) " "                     " "          " "  "*"
## 6  ( 1 ) " "                     " "          " "  " "
## 7  ( 1 ) " "                     " "          " "  " "
## 8  ( 1 ) " "                     " "          " "  "*"
```

```r
EQ$W238 <- factor(EQ$W238, labels=c("yes", "no", "NA"))
pander(confint(inj.model))
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.4175	2.598
V449rent	-0.7113	0.2446
RSEXfemale	-1.202	-0.3016
NEWETHNHispanic	-0.9875	0.03634
NEWETHNNative American	-1.587	3.62
NEWETHNAsian/P acific Islander	-0.3508	1.685
NEWETHNAfrican American	-1.545	0.03228
NEWETHNOther	-122.3	NA
RAGE18	-2.467	1.357
RAGE19	-0.9903	4.275
RAGE20	-1.392	3.944
RAGE21	-1.522	2.076
RAGE22	-2.128	1.106
RAGE23	-0.9007	2.926
RAGE24	-2.157	1.089
RAGE25	-1.466	1.799
RAGE26	-1.923	1.319
RAGE27	-1.422	1.615
RAGE28	-0.9754	2.027
RAGE29	-1.585	1.314
RAGE30	-1.488	1.333

	2.5 %	97.5 %
RAGE31	-1.704	1.251
RAGE32	-1.842	1.329
RAGE33	-2.462	1.156
RAGE34	-1.322	2.713
RAGE35	-1.842	1.206
RAGE36	-2.612	0.6386
RAGE37	-1.217	4.025
RAGE38	-2.274	1.343
RAGE39	-2.687	0.5233
RAGE40	-1.533	1.316
RAGE41	-1.311	3.931
RAGE42	-2.233	0.9035
RAGE43	-1.945	3.467
RAGE44	-3.426	0.1249
RAGE45	-1.779	3.497
RAGE46	-2.448	1.287
RAGE47	-2.27	0.6484
RAGE48	-2.241	3.065
RAGE49	-1.512	2.047
RAGE50	-1.656	2.547
RAGE51	-1.272	4.13
RAGE52	-2.382	3.14
RAGE53	-152.8	NA
RAGE54	-2.166	3.37
RAGE55	NA	396.1
RAGE56	-4.897	0.651
RAGE57	-3.112	1.55
RAGE58	-2.707	1.317
RAGE59	-274.5	NA
RAGE60	-3.527	0.62
RAGE61	-200.4	NA
RAGE62	-4.47	2.306
RAGE63	-277.2	NA
RAGE64	-789.2	NA
RAGE65	-209.8	NA
RAGE66	-274.5	NA
RAGE67	-2.011	2.076
RAGE68	-419.5	NA
RAGE69	-5.049	0.7258
RAGE70	-1.965	3.651
RAGE71	-4.47	2.306
RAGE72	NA	786.4
RAGE73	-3.737	3.03
RAGE74	-2.049	3.397
RAGE75	NA	410.7
RAGE76	NA	787.2
RAGE77	-1.802	3.699
RAGE78	-3.841	2.887
RAGE80	-2.225	3.244
RAGE82	NA	419.6
RAGE83	NA	419.4
RAGE84	-1.79	3.641

	2.5 %	97.5 %
RAGE86	-4.47	2.306
RAGE89	-789.2	NA
RAGE91	-788.3	NA
RAGE97	NA	787.4

##1) Perform a binary logistic regression analysis using the Parental HIV data to model the probability of having been absent from school without a reason (variable HOOKEY). Find the variables that best predict whether an adolescent had been absent without a reason or not. Use a hefty dose of common sense here, not all variables are reasonable to use (e.g. using the # of times a student skips school to predict whether or not they will predict school)

```
hiv <- read.delim("https://norcalbiostat.netlify.com/data/PARHIV_081217.txt",
                  sep="\t", stringsAsFactors = FALSE, header=TRUE)

hiv$HOOKEY <- ifelse(hiv$HOOKEY=="1", 0, 1)
hiv$GENDER <- ifelse(hiv$GENDER=="Male", "0", "1")

summary(regsubsets(HOOKEY ~ AGE + GENDER + LIVWITH + SIBLINGS + JOBMO + EDUMO + HOWREL + ATTSERV + AGEAL
```

```
## Subset selection object
## Call: regsubsets.formula(HOOKEY ~ AGE + GENDER + LIVWITH + SIBLINGS +
##     JOBMO + EDUMO + HOWREL + ATTSERV + AGEALC + FINSIT + AGESMOKE +
##     LIKESCH, data = hiv)
## 12 Variables  (and intercept)
##          Forced in Forced out
## AGE          FALSE      FALSE
## GENDER1      FALSE      FALSE
## LIVWITH      FALSE      FALSE
## SIBLINGS     FALSE      FALSE
## JOBMO        FALSE      FALSE
## EDUMO        FALSE      FALSE
## HOWREL       FALSE      FALSE
## ATTSERV      FALSE      FALSE
## AGEALC       FALSE      FALSE
## FINSIT       FALSE      FALSE
## AGESMOKE     FALSE      FALSE
## LIKESCH      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          AGE GENDER1 LIVWITH SIBLINGS JOBMO EDUMO HOWREL ATTSERV AGEALC FINSIT
## 1  ( 1 ) "*" " "     " "     " "      " "   " "   " "    " "     " "    " "
## 2  ( 1 ) "*" " "     " "     " "      " "   " "   " "    " "     " "    " "
## 3  ( 1 ) "*" " "     " "     " "      "*"   " "   " "    " "     " "    " "
## 4  ( 1 ) "*" " "     " "     " "      "*"   " "   " "    " "     "*"    " "
## 5  ( 1 ) "*" " "     " "     "*"      "*"   " "   " "    " "     "*"    " "
## 6  ( 1 ) "*" " "     " "     "*"      "*"   "*"   " "    " "     "*"    " "
## 7  ( 1 ) "*" " "     "*"     "*"      "*"   "*"   " "    " "     "*"    " "
## 8  ( 1 ) "*" " "     "*"     "*"      "*"   "*"   " "    " "     "*"    "*"
##          AGESMOKE LIKESCH
## 1  ( 1 ) " "      " "
## 2  ( 1 ) "*"      " "
## 3  ( 1 ) "*"      " "
## 4  ( 1 ) "*"      " "
## 5  ( 1 ) "*"      " "
## 6  ( 1 ) "*"      " "
## 7  ( 1 ) "*"      " "
## 8  ( 1 ) "*"      " "
```

```
hookey_model=glm(HOOKEY~AGE + GENDER + LIVWITH + SIBLINGS + JOBMO + EDUMO + HOWREL + ATTSERV + AGEALC +
```

The variables that best predict hookey from the model i made seems to be age, the job of the mother, and the age they started smoking as well as drinking.

##2) Use the default value for the predict() function to create a vector of predictions for each student.

```
model.pred.prob <- predict(hookey_model, type='response')
set.seed(12345)
plot.mpp <- data.frame(pred.prob = model.pred.prob,
                       pred.class = rbinom(n=length(model.pred.prob), size=1, p=model.pred.prob),
                       truth = hookey_model$y)

head(plot.mpp)
```

```
##    pred.prob pred.class truth
## 4  0.4091943          1     1
## 7  0.5155421          0     0
## 11 0.9037584          1     1
## 12 0.6003828          0     1
## 13 0.9448776          1     0
## 15 0.9130180          1     1
```

```
plot.mpp <- plot.mpp %>%
mutate(pred.class = factor(pred.class, labels=c("No",
"Hookey")), truth = factor(truth,
labels=c("No", "Hookey")))
table(plot.mpp$pred.class, plot.mpp$truth)
```

```
##
##          No Hookey
##   No      8      7
##   Hookey 12     45
```

##3) Create a confusion matrix for these predictions and interpret: accuracy, balanced accuracy, sensitivity, specificity, PPV, NPV.

```
confusionMatrix(plot.mpp$pred.class, plot.mpp$truth, positive="Hookey")
```

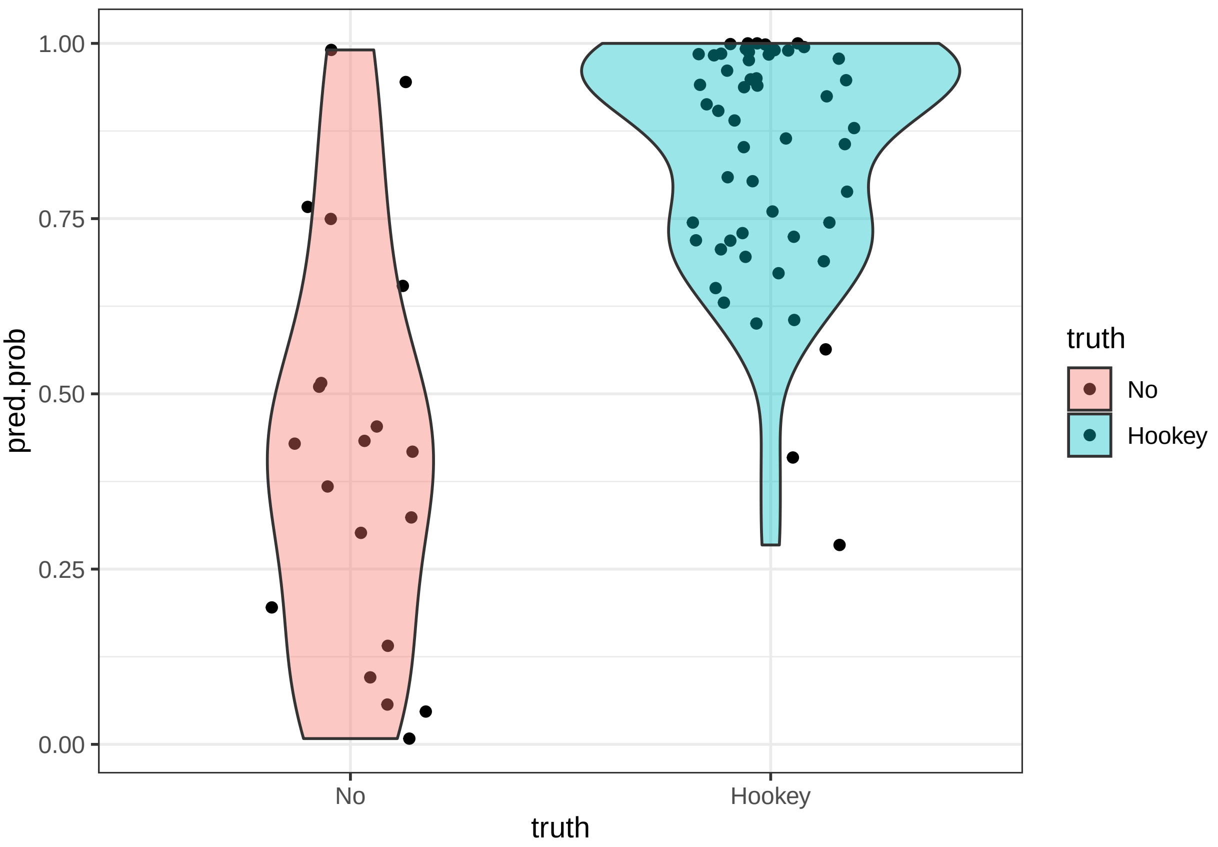
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Hookey
##     No      8      7
##     Hookey 12     45
```

```
##
##                Accuracy : 0.7361
##                  95% CI : (0.619, 0.833)
##     No Information Rate : 0.7222
##     P-Value [Acc > NIR] : 0.4554
##
##                   Kappa : 0.2875
##
##  Mcnemar's Test P-Value : 0.3588
##
##             Sensitivity : 0.8654
##             Specificity : 0.4000
##          Pos Pred Value : 0.7895
##          Neg Pred Value : 0.5333
##              Prevalence : 0.7222
##          Detection Rate : 0.6250
##    Detection Prevalence : 0.7917
##       Balanced Accuracy : 0.6327
##
##        'Positive' Class : Hookey
##
```

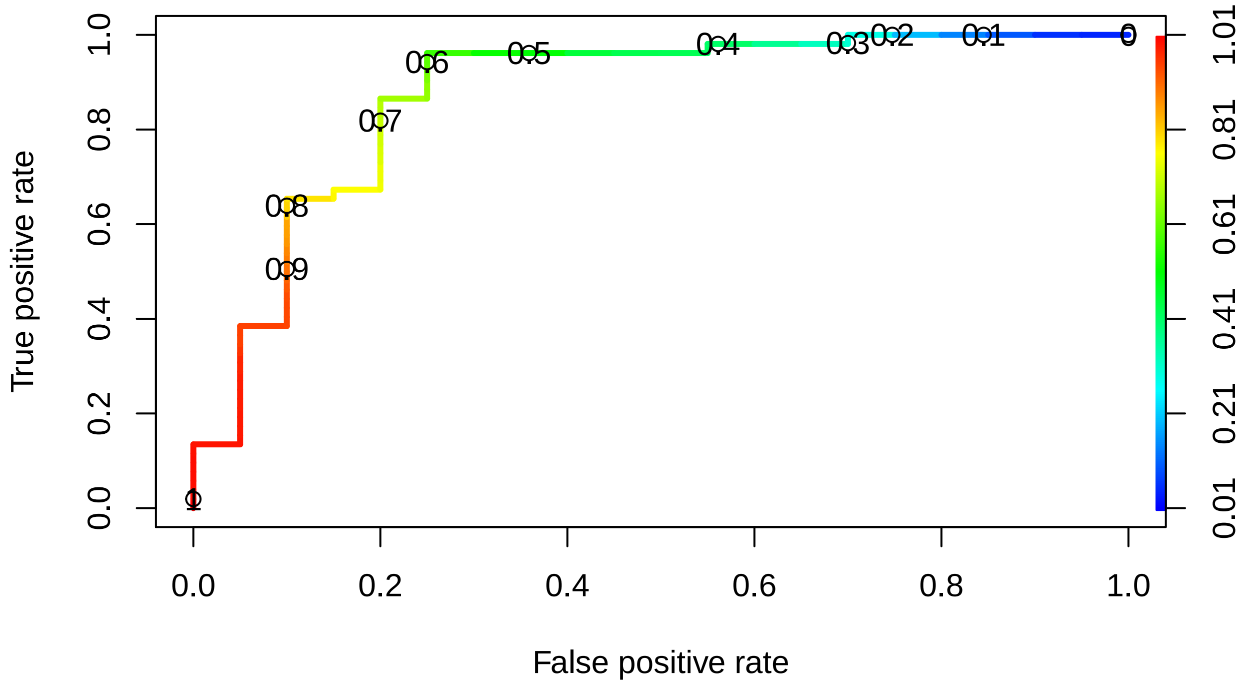
Accuracy: 73.61% of the time we are able to predict whether someone plays hookey. Balanced Accuracy: 63.27% Sensitivity (True Positive Rate): 86.54% were correctly predicted to play hookey. Specificity (True Negative Rate): 40% were correctly predicted to not skip school. PPV (Positive Predictive Value): 78.95% of individuals who were predicted to play hookey were predicted correctly. NPV (Negative Predicted Value): 53.33% of individuals who were predicted to not skip school were predicted correctly.

##4) Describe the distribution of predicted probabilities by true group membership. Use a violin + jitter plot as shown in the notes. What do you notice?

```
ggplot(plot.mpp, aes(x=truth, y=pred.prob, fill=truth)) +
      geom_jitter(width=.2) + geom_violin(alpha=.4) + theme_bw()
```



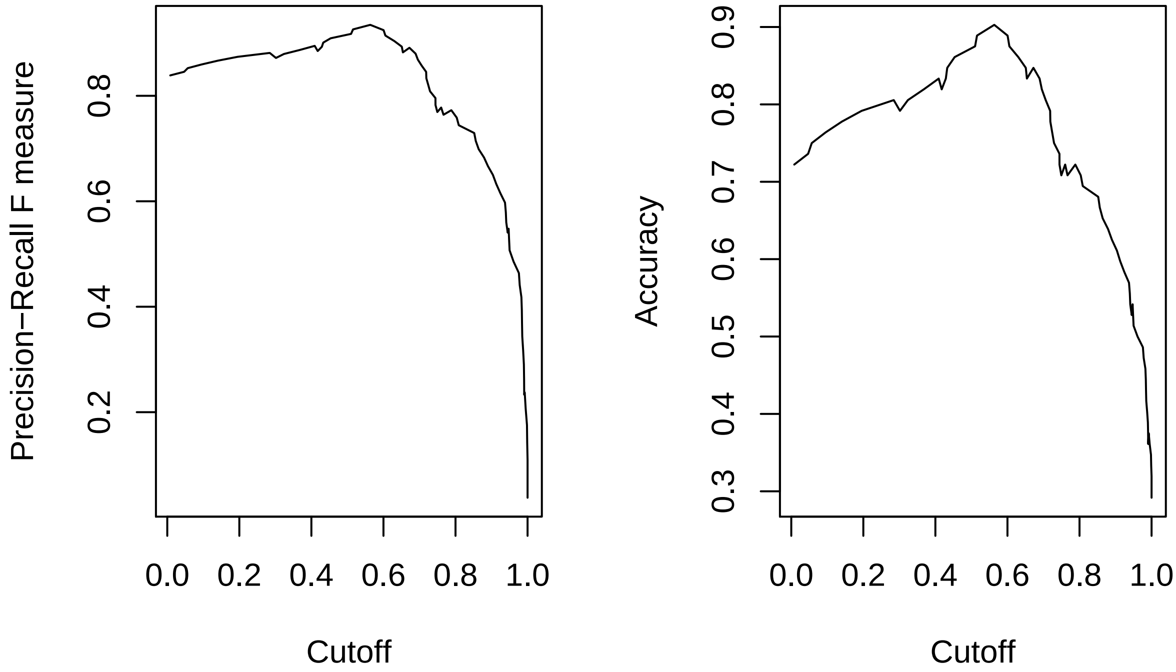
```r
pr <- prediction(model.pred.prob, hookey_model$y)
perf <- performance(pr, measure="tpr", x.measure="fpr")
plot(perf, colorize=TRUE, lwd=3, print.cutoffs.at=c(seq(0,1,by=0.1)))
```



##5) Find the best cutoff point to discriminate between adolescents who were absent without a reason and those who were not by using an ROC curve and maximizing accuracy.

```
perf.f1 <- performance(pr,measure="f")
perf.acc <- performance(pr,measure="acc")

par(mfrow=c(1,2))
plot(perf.f1)
plot(perf.acc)
```



```
(max.f1 <- max(perf.acc@y.values[[1]], na.rm=TRUE))
```

```
## [1] 0.9027778
```

```
(row.with.max <- which(perf.acc@y.values[[1]]==max.f1))
```

```
## [1] 56
```

```
(cutoff.value <- perf.acc@x.values[[1]][row.with.max])
```

```
##      135
## 0.563502
```