Week 03 : Modeling the Expert

Description

```
### Format

# The variables in the dataset quality.csv are as follows:

# * __MemberID__ numbers the patients from 1 to 131, and is just an identifying number.

# * __InpatientDays__ is the number of inpatient visits, or number of days the person spent in the hosp

# * __ERVisits__ is the number of times the patient visited the emergency room.

# * __OfficeVisits__ is the number of times the patient visited any doctor's office.

# * __Narcotics__ is the number of prescriptions the patient had for narcotics.

# * __DaysSinceLastERVisit__ is the number of days between the patient's last emergency room visit and

# * __Pain__ is the number of visits for which the patient complained about pain.

# * __TotalVisits__ is the total number of times the patient visited any healthcare provider.

# * __ProviderCount__ is the number of providers that served the patient. Medical Claims is the number

# * __ClaimLines__ is the total number of medical claims.

# * __StartedOnCombination__ is whether or not the patient was started on a combination of drugs to tre

# * __AcuteDrugGapSmall__ is the fraction of acute drugs that were refilled quickly after the prescript

# * __PoorCare__ is the outcome or dependent variable, and is equal to 1 if the patient had poor care,
```

Video 04

```r
# Read in dataset

quality = read.csv("F:/SkyDrive/Studying/MIT_COURSES/15-071-TheAnalyticsEdge/lecture/dataset/quality.cs

# Look at structure
str(quality)
```

```
## 'data.frame':    131 obs. of  14 variables:
##  $ MemberID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ InpatientDays       : int  0 1 0 0 8 2 16 2 2 4 ...
##  $ ERVisits            : int  0 1 0 1 2 0 1 0 1 2 ...
##  $ OfficeVisits        : int  18 6 5 19 19 9 8 8 4 0 ...
##  $ Narcotics           : int  1 1 3 0 3 2 1 0 3 2 ...
##  $ DaysSinceLastERVisit: num  731 411 731 158 449 ...
##  $ Pain                : int  10 0 10 34 10 6 4 5 5 2 ...
##  $ TotalVisits         : int  18 8 5 20 29 11 25 10 7 6 ...
##  $ ProviderCount       : int  21 27 16 14 24 40 19 11 28 21 ...
##  $ MedicalClaims       : int  93 19 27 59 51 53 40 28 20 17 ...
##  $ ClaimLines          : int  222 115 148 242 204 156 261 87 98 66 ...
##  $ StartedOnCombination: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ AcuteDrugGapSmall   : int  0 1 5 0 0 4 0 0 0 0 ...
##  $ PoorCare            : int  0 0 0 0 0 1 0 0 1 0 ...
```

```
# Table outcome
table(quality$PoorCare)
```

```
##
##  0  1
## 98 33
```

```
# Baseline accuracy
98/131
```

```
## [1] 0.7480916
```

```
# Install and load caTools package
# install.packages("caTools")
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.1.2
```

```
# Randomly split data and make sure 75% of the data for training set and 25% for
# the tesing set and the ratio of good care is 75% both in the training set and
# testing set

set.seed(88)
?caTools
```

```
## starting httpd help server ... done
```

```
?sample.split
split = sample.split(quality$PoorCare, SplitRatio = 0.75)
# TRUE for training set and FALSE for testing set
split
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
##  [12] FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
##  [23]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
##  [34] FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
##  [56]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [78]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
##  [89]  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [100]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE
## [111] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
## [122]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
```

```r
# Create training and testing sets
qualityTrain = subset(quality, split == TRUE)
qualityTest = subset(quality, split == FALSE)

# Logistic Regression Model using generalized linear model
QualityLog = glm(PoorCare ~ OfficeVisits + Narcotics, data=qualityTrain, family=binomial)
# It accounts for the number of variables used compared to the number of
# observations. It provides a means for model selection. The preferred model is the one with the minimum
summary(QualityLog)
```

```
##
## Call:
## glm(formula = PoorCare ~ OfficeVisits + Narcotics, family = binomial,
##     data = qualityTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8512  -0.6082  -0.4866  -0.1397   2.1642
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.77042    0.54219  -5.110 3.23e-07 ***
## OfficeVisits  0.07846    0.02995   2.620  0.00879 **
## Narcotics     0.14708    0.05146   2.858  0.00426 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 111.888  on 98  degrees of freedom
## Residual deviance:  82.405  on 96  degrees of freedom
## AIC: 88.405
##
## Number of Fisher Scoring iterations: 5
```

```r
# Make predictions on training set
predictTrain = predict(QualityLog, type="response")

# Analyze predictions
summary(predictTrain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05894 0.10780 0.14800 0.25250 0.25840 0.99760
```

```r
tapply(predictTrain, qualityTrain$PoorCare, mean)
```

```
##         0         1
## 0.1738683 0.4853497
```

```r
# Video 5

# Confusion matrix for threshold of 0.5
table(qualityTrain$PoorCare, predictTrain > 0.5)
```

```
##
##     FALSE TRUE
##   0    71    3
##   1    15   10
```

```
# Sensitivity and specificity
10/25
```

```
## [1] 0.4
```

```
70/74
```

```
## [1] 0.9459459
```

```
# Confusion matrix for threshold of 0.7
table(qualityTrain$PoorCare, predictTrain > 0.7)
```

```
##
##     FALSE TRUE
##   0    73    1
##   1    17    8
```

```
# Sensitivity and specificity
8/25
```

```
## [1] 0.32
```

```
73/74
```

```
## [1] 0.9864865
```

```
# Confusion matrix for threshold of 0.2
table(qualityTrain$PoorCare, predictTrain > 0.2)
```

```
##
##     FALSE TRUE
##   0    54   20
##   1     8   17
```

```
# Sensitivity and specificity
16/25
```

```
## [1] 0.64
```

```
54/74
```

```
## [1] 0.7297297
```

```r
# Video 6

# Install and load ROCR package
# install.packages("ROCR")
library(ROCR)
```

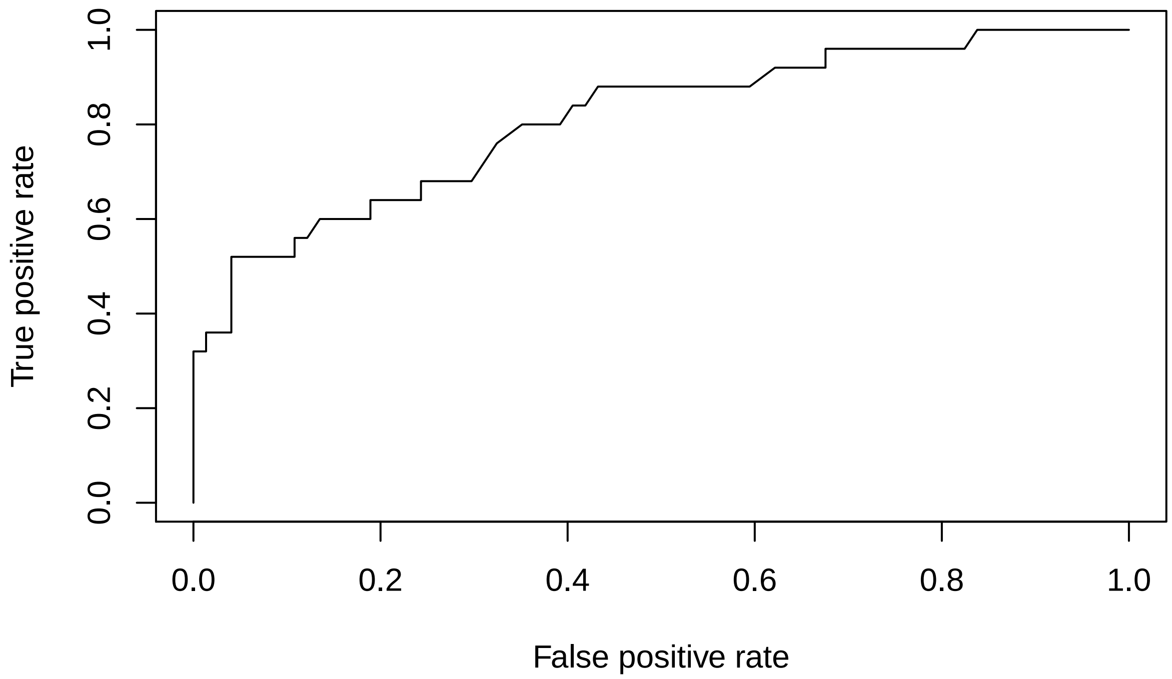
```
## Warning: package 'ROCR' was built under R version 3.1.2
```

```
## Loading required package: gplots
```

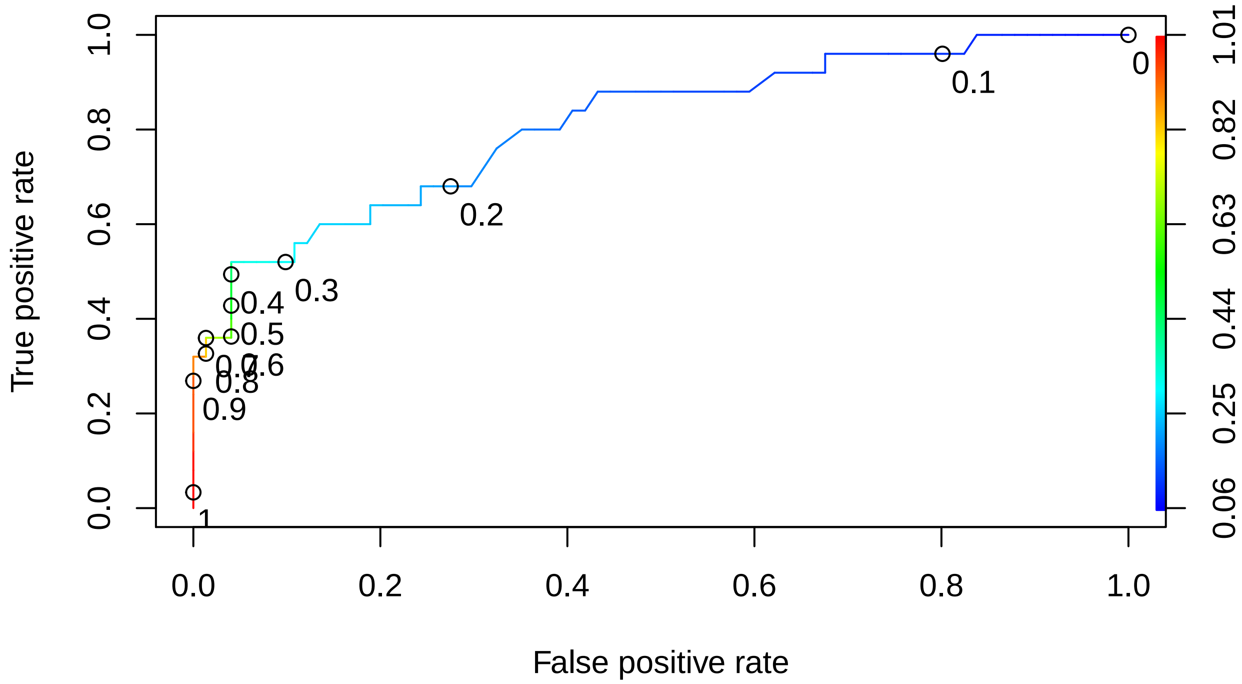
```
## Warning: package 'gplots' was built under R version 3.1.2
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
# Prediction function
ROCRpred = prediction(predictTrain, qualityTrain$PoorCare)

# Performance function
ROCRperf = performance(ROCRpred, "tpr", "fpr")

# Plot ROC curve
plot(ROCRperf)
```



```
# Add colors
plot(ROCRperf, colorize=TRUE)

# Add threshold labels
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))
```


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1.01
0.82
0.63
0.44
0.25
0.06
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1