Untitled

Vinoth Aryan Nagabosshanam

March 21, 2017

Logistic Regression

here in this problem we are use logistic regression to panitents use the attorney or not based on the date set given so to find we are apply logistic regression model

```
cla<-read.csv("B:\\data science courses\\Datasets_BA 2\\Claimants.csv")
head(cla)
```

```
##   CASENUM ATTORNEY CLMSEX CLMINSUR SEATBELT CLMAGE   LOSS
## 1       5        0      0        1        0     50 34.940
## 2       3        1      1        0        0     18  0.891
## 3      66        1      0        1        0      5  0.330
## 4      70        0      0        1        1     31  0.037
## 5      96        1      0        1        0     30  0.038
## 6      97        0      1        1        0     35  0.309
```

```
str(cla)
```

```
## 'data.frame':    1340 obs. of  7 variables:
##  $ CASENUM : int  5 3 66 70 96 97 10 36 51 55 ...
##  $ ATTORNEY: int  0 1 1 0 1 0 0 0 1 1 ...
##  $ CLMSEX  : int  0 1 0 0 0 1 0 1 1 0 ...
##  $ CLMINSUR: int  1 0 1 1 1 1 1 1 1 1 ...
##  $ SEATBELT: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ CLMAGE  : int  50 18 5 31 30 35 9 34 60 NA ...
##  $ LOSS    : num  34.94 0.891 0.33 0.037 0.038 ...
```

```
# convert int to factor for few variable
cla$CLMSEX<-as.factor(cla$CLMSEX)
cla$CLMINSUR<-as.factor(cla$CLMINSUR)
cla$SEATBELT<-as.factor(cla$SEATBELT)
str(cla)
```

```
## 'data.frame':    1340 obs. of  7 variables:
##  $ CASENUM : int  5 3 66 70 96 97 10 36 51 55 ...
##  $ ATTORNEY: int  0 1 1 0 1 0 0 0 1 1 ...
##  $ CLMSEX  : Factor w/ 2 levels "0","1": 1 2 1 1 1 2 1 2 2 1 ...
##  $ CLMINSUR: Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ SEATBELT: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ CLMAGE  : int  50 18 5 31 30 35 9 34 60 NA ...
##  $ LOSS    : num  34.94 0.891 0.33 0.037 0.038 ...
```

```
summary(cla)
```

```
##     CASENUM          ATTORNEY        CLMSEX     CLMINSUR     SEATBELT
##  Min.   :    0   Min.   :0.0000   0   :586   0   : 120   0   :1270
##  1st Qu.: 4177   1st Qu.:0.0000   1   :742   1   :1179   1   :  22
##  Median : 8756   Median :0.0000   NA's: 12   NA's:  41   NA's:  48
##  Mean   :11202   Mean   :0.4888
```

```
##   3rd Qu.:15702   3rd Qu.:1.0000
##   Max.   :34153   Max.   :1.0000
##
##      CLMAGE           LOSS
##   Min.   : 0.00   Min.   :  0.000
##   1st Qu.: 9.00   1st Qu.:  0.400
##   Median :30.00   Median :  1.069
##   Mean   :28.41   Mean   :  3.806
##   3rd Qu.:43.00   3rd Qu.:  3.781
##   Max.   :95.00   Max.   :173.604
##   NA's   :189
```

fiting a Logistic Regression model

```r
model_lgm<-glm(cla$ATTORNEY~cla$CLMSEX+cla$CLMINSUR+cla$SEATBELT+cla$CLMAGE+cla$LOSS,family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(model_lgm)
```

```
##
## Call:
## glm(formula = cla$ATTORNEY ~ cla$CLMSEX + cla$CLMINSUR + cla$SEATBELT +
##     cla$CLMAGE + cla$LOSS, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.74474  -1.01055  -0.02547   0.95764   2.78320
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.199978   0.246769  -0.810  0.41772
## cla$CLMSEX1    0.432996   0.135706   3.191  0.00142 **
## cla$CLMINSUR1  0.602173   0.231030   2.606  0.00915 **
## cla$SEATBELT1 -0.781079   0.566125  -1.380  0.16768
## cla$CLMAGE     0.006487   0.003324   1.952  0.05097 .
## cla$LOSS      -0.385044   0.034845 -11.050  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1516.1  on 1095  degrees of freedom
## Residual deviance: 1287.8  on 1090  degrees of freedom
##   (244 observations deleted due to missingness)
## AIC: 1299.8
##
## Number of Fisher Scoring iterations: 6
```

** to get the Odds Ratio we need to use the expo **

```r
exp(coef(model_lgm))
```

```
##   (Intercept)   cla$CLMSEX1 cla$CLMINSUR1 cla$SEATBELT1    cla$CLMAGE
```

```
##     0.8187490     1.5418701     1.8260829     0.4579119     1.0065085
##      cla$LOSS
##     0.6804208
```

** to get Confusion matrix table

```
prob <- predict(model_lgm,type=c("response"),cla)
#prob
confusion<-table(prob>0.5,cla$ATTORNEY)
confusion
```

```
##
##           0   1
##   FALSE 380 125
##   TRUE  198 393
```

** To get Accuracy of the model **

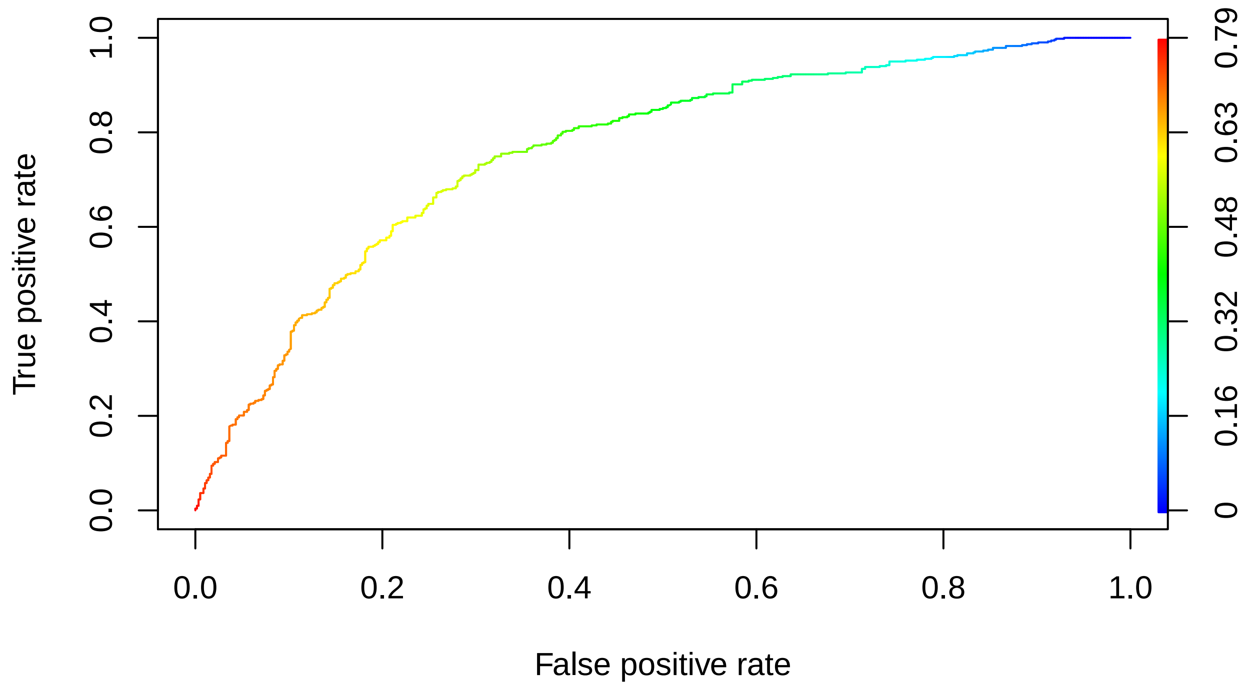
```
Accuracy<-sum(diag(confusion)/sum(confusion))
Accuracy
```

```
## [1] 0.705292
```

```
#install.packages("ROCR")
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.3.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.3.3

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
rocrpred<-prediction(prob,cla$ATTORNEY)
rocrperf<-performance(rocrpred,'tpr','fpr')
plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))
```


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0
0.16
0.32
0.48
0.63
0.79