Churn Data Modeling Model Accuracy

Lecture 8 (Practical)

07/08/2021

Model Diagonostics;overall model significance

```
> #Model Diagonostics;overall model significance
> with(lr5,pchisq(null.deviance-deviance,df.null-df.residual, lower.tail = F))
[1] 5.069035e-103
```

- The p-value is small, hence the overall model is significant at 0.01.

```
> 1-pchisq(2056.9 - 1560.6, df=(2498-2491))
[1] 0
```

- The p-value associated with chi-square test for deviance also shows that the model is significant at 0.01

The analysis of deviance is given below.

```
> anova(lr5, test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: training$Churn

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                  2498     2056.9              
training$CSC_Hi       1  167.504      2497     1889.4 < 2.2e-16 ***
training$Int.l..Plan  1  127.410      2496     1762.0 < 2.2e-16 ***
training$VMP.ind      1   35.063      2495     1726.9 3.191e-09 ***
training$Day.Mins     1  117.573      2494     1609.4 < 2.2e-16 ***
training$Eve.Mins     1   25.484      2493     1583.9 4.460e-07 ***
training$Night.Mins   1   11.551      2492     1572.3 0.0006771 ***
training$Intl.Mins    1   11.689      2491     1560.6 0.0006289 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- All variables are significant at 0.01

Prediction using the fitted model

```
> p<-predict(lr5,type = 'response')
> # type = "response" gives the predicted probabilities.
> head(p)
      1345       2215       1002       1435       2309       2776 
0.21231506 0.12893098 0.03515824 0.01900706 0.09720877 0.05467187 
```

- The first five observations in the training data gives the probabilities as given above. (View Training data for more details of row numbers)
- As the first observation in the training data(row 1345) is predicted with low probability of churning (considering cut-off probability 0.5), the predicted class is non-churner. But view the data, row 1345 customer is a churner. So there is a mis-classification.
- Let us find the mis-classification error now.

Mis-classification error

```
pred<-ifelse(p>0.5, 1,0)
tab5<-table(predicted = pred, actual = training[,21])
tab5
          actual
 predicted False. True.
         0   2047   294
         1     93    65
```

- The sum table will give more details

```
sumtable5<-addmargins(tab5,FUN=sum)
sumtable5
          actual
predicted False. True.  sum
      0     2047   294 2341
      1       93    65  158
      sum   2140   359 2499
```

Mis-classification error

```
          actual
predicted False. True.  sum
      0      2047   294 2341
      1        93    65  158
      sum    2140   359 2499
```

- It says that out of total 2499 customers, the training data has 359 churners, but the model predicted only 158 as churners. Also the training data has 2140 non-churners, but the model predicted as 2341 non-churners.
- That is the model incorrectly classified some actual churners as non-churners.
- To check the accuracy of the model, we can find out sensitivity, specificity etc.

Sensitivity

```
#TAN : Total Actually Negative= TN + FP
#TAP : Total Actually Positive= FN + TP

TAP <- sum(tab5[,2])
TAN <- sum(tab5[,1])

TP <- tab5[2,2]
TN <- tab5[1,1]
FP <- tab5[2,1]
FN <- tab5[1,2]

TPR <- TP / TAP
TPR
```

```
[1] 0.1810585
```

- The TPR is sensitivity which is 0.18 only. Also we have already seen that the model classified some actual churners as non-churners.

Specificity

```
FPR <- FP / TAN
FPR

[1] 0.04345794

specificity <- TN/TAN
specificity

[1] 0.9565421
```

- The false negative rate is 0.04 and accordingly specificity is 0.95. Which means the model predicted actual non-churners as non-churners.

Mis-classification error

```
#The proportion of observations correctly classified is
sum(diag(tab5))/sum(tab5)

[1] 0.8451381
```

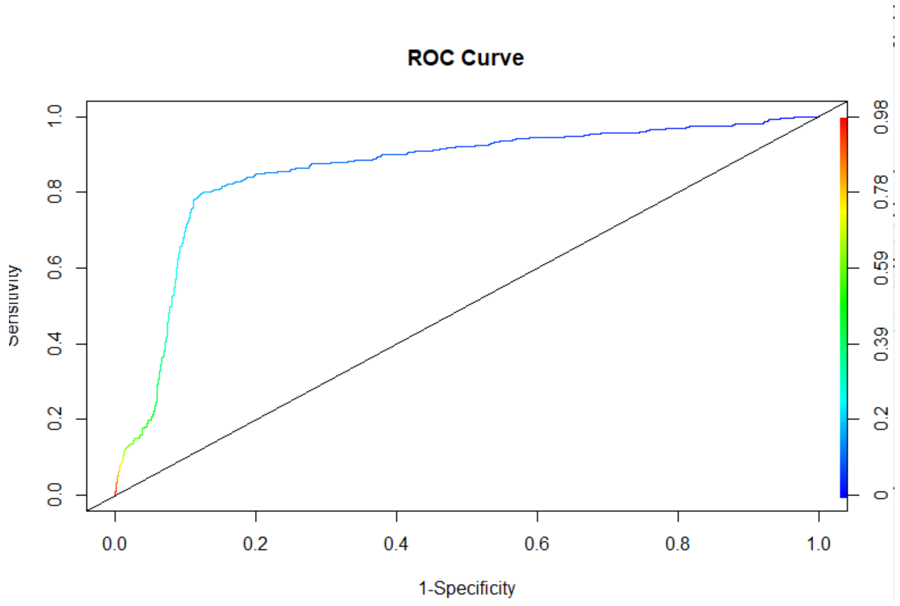
- The correct classification rate is 0.845 which is more than 0.8 and looks fine to implement the model in practice. Accordingly the mis-classification rate is
- ```
  #The proportion of observations misclassified is
  1-sum(diag(tab5))/sum(tab5)

  [1] 0.1548619
  ```
- 15% is the mis-classification error.
  ```

ROC Curve

```
library(ROCR)

ROCRpred <- prediction(p, training[,21])
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
ROCRperf
plot(ROCRperf, colorize=T, main= "ROC Curve", ylab = "Sensitivity", xlab = "1-Specificity")
abline(a=0,b=1)
```



- The curve is almost closer to left-hand border and top border, so the model accuracy is good.

Area under the curve

```
auc <- performance(ROCRpred,measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
[1] 0.8566045
```

- The area under the curve is 0.8566 which is more than 0.8 and can be considered as a good model and can be applied in practical situation.