Multiple Logistic regression Churn data

Lecture/Practical 6

04/08/2021

Multiple Logistic Regression

- More than one predictor variable is used to classify the binary response variable.
- The logit of the multiple logistic regression model is

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- The logistic regression model is

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x + \ldots + \beta_p x_p}}$$

Multiple Logistic Regression: Churn data

- We examine whether a relationship exists between churn and the following set of predictors.

- International Plan, a flag variable
- Voice mail Plan, a flag variable
- CSC_Hi, a flag variable
- Account Length , continuous
- Day Minutes, Continuous
- Evening Minutes, Continuous
- Night Minutes, Continuous
- International Minutes, Continuous

Partitioning training and test data

- smp_size<-floor(0.75*nrow(churn))
- set.seed(124)
- trainingdata <- sample(seq_len(nrow(churn)),size=smp_size)
- training<-churn[trainingdata,]
- testing<-churn[-trainingdata,]
- table(training$Churn)
- table(testing$Churn)
- write.csv(training, "C:\\Users\\.....\\training.csv")

Proportion of churners in training and testing

- tab2<-table(training$Churn)
- prop.table(tab2)
- False True.
- 0.8543417 0.1456583
- tab3<-table(testing$Churn)
- prop.table(tab3)
- False True.
- 0.8573141 0.1426859

Check for Multicollinearity

Obtain VIFs

library(car)

- vif(lm(training$Account.Length ~ training$CSC_Hi +training$Intl.Calls +training$VMP.ind +training$Day.Mins + training$Eve.Mins +training$Intl.Mins, data=training))

```
  training$CSC_Hi training$Intl.Calls    training$VMP.ind
         1.000532            1.000615            1.001983
training$Day.Mins   training$Eve.Mins  training$Intl.Mins
         1.001733            1.001810            1.001566
```

R Zone

- churn$IntlP.ind<- ifelse(churn$Int.l..Plan == "yes",1,0)
- churn$VMP.ind<- ifelse (churn$VMail.Plan=="yes",1,0)
- lr4<-glm (training$Churn ~ training$Account.Length + training$CSC_Hi +training$Int.l..Plan + training$VMP.ind +training$Day.Mins + training$Eve.Mins +training$Night.Mins +training$Intl.Mins, data=training, family= "binomial", maxit = 500)
- summary(lr4)

Output

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5106  -0.4744  -0.3363  -0.2037   3.0908

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.0642639  0.6227305 -12.950  < 2e-16
training$Account.Length  0.0001064  0.0016690   0.064 0.949187
training$CSC_Hi          2.5273546  0.1827241  13.832  < 2e-16
training$Int.l..Planyes  1.9357416  0.1681649  11.511  < 2e-16
training$VMP.ind        -1.0269748  0.1740374  -5.901 3.62e-09
training$Day.Mins        0.0135529  0.0012821  10.571  < 2e-16
training$Eve.Mins        0.0070950  0.0013424   5.285 1.26e-07
training$Night.Mins      0.0044910  0.0013085   3.432 0.000598
training$Intl.Mins       0.0814177  0.0240564   3.384 0.000713

(Intercept)             ***
training$Account.Length
training$CSC_Hi         ***
training$Int.l..Planyes ***
training$VMP.ind        ***
training$Day.Mins       ***
training$Eve.Mins       ***
training$Night.Mins     ***
training$Intl.Mins      ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insignificant predictor?

- It can be seen from the output that, the variable *Account Length* is not significant as the p-value associated with Z-wald test is 0.949 which is more than the level of significance 0.05.
- Hence we remove the variable *Account length* from the model and run the regression.

Output (after removing Account Length)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5107  -0.4746  -0.3363  -0.2036   3.0893

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -8.052114   0.592699 -13.585  < 2e-16 ***
training$CSC_Hi         2.527373   0.182713  13.833  < 2e-16 ***
training$Int.l..Planyes 1.936194   0.168019  11.524  < 2e-16 ***
training$VMP.ind       -1.026664   0.173962  -5.902 3.60e-09 ***
training$Day.Mins       0.013552   0.001282  10.571  < 2e-16 ***
training$Eve.Mins       0.007094   0.001342   5.285 1.26e-07 ***
training$Night.Mins     0.004487   0.001307   3.433 0.000597 ***
training$Intl.Mins      0.081386   0.024051   3.384 0.000715 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Contd….

- All predictors included in the model are significant.
- Estimated logit is

$\hat{g}(x)$=-8.0521+2.5274CSC_Hi(=1) +1.9362Int.l..Planyes-1.0267VMP.ind(=yes)+0.0136Day.Mins+0.0071Eve.Mins
+0.0045Night.Mins+0.0814Intl.Mins

- A high usage customer belonging to the International plan but not the voice mail plan with many calls to customer service . This customer has 300 day, evening, and night minutes and 20 international minutes. Find the probability of churning?

$\hat{g}(x)$=-8.0521+2.5274(1) +1.9362(1)-1.0267VMP.ind(0)+0.0136(300)+0.0071(0) +0.0045(300)+0.0814(20)

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$