Logistic regression for a dichotomous predictor

Practical 5

28/07/2021

Logistic regression for a dichotomous predictor

- In the churn data, we are interested in predicting whether a customer would leave the cell phone company's service (churn), based on a set of predictor variables.
- Assume that only predictor variable available is *Voice Mail Plan,* a flag variable indicating membership in the plan.

Logistic regression for a dichotomous predictor

- cnts1<-table(churn$Churn, churn$VMail.Plan,
  dnn=c("Churn","Voice mail plan"))
- sumtable1<-addmargins(cnts1,FUN=sum)
- sumtable1

	Voice mail plan		
Churn	no	yes	sum
False.	2008	842	2850
True.	403	80	483
sum	2411	922	3333

Create dummy for Voice mail plan (VMP)#if required

- churn$VMP.ind<- ifelse (churn$VMail.Plan=="yes",1,0)

Run logistic regression

- lr<- glm (Churn ~ VMP.ind, data=churn, family ="binomial")
- summary(lr)

Output of Logistic Regression

- Call:
- glm(formula = Churn ~ VMP.ind, family = "binomial", data = churn)

- Deviance Residuals:
- Min 1Q Median 3Q Max
- -0.6048 -0.6048 -0.6048 -0.4261 2.2111

- Coefficients:
- Estimate Std. Error z value Pr(>|z|)
- (Intercept) -1.60596 0.05458 -29.422 < 2e-16 ***
- VMP.ind -0.74780 0.12910 -5.792 6.94e-09 ***
- ---
- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- (Dispersion parameter for binomial family taken to be 1)

- Null deviance: 2758.3 on 3332 degrees of freedom
- Residual deviance: 2720.3 on 3331 degrees of freedom
- AIC: 2724.3

- Number of Fisher Scoring iterations: 5

Interpreting Logistic regression for a dichotomous predictor

- Estimated logit

$$\hat{g}(x) = -1.60596 - 0.747795\,x$$

- For a customer belonging to the plan (x=1), the estimated probability of churning:

$$\hat{g}(1) = -2.3538 \qquad \hat{\pi}(1) = \frac{e^{\hat{g}(1)}}{1 + e^{\hat{g}(1)}} = 0.0868$$

Contd….

- For a customer not belonging to the voice mail plan (x=0), the estimated probability of churning:

$$\hat{g}(0) = -1.60596 \qquad \hat{\pi}(0) = \frac{e^{\hat{g}(0)}}{1+e^{\hat{g}(0)}} = 0.16715$$

- Indicating that not belonging to the voice mail plan may be slightly indicative of churning.
- Wald test Z_{wald}= -5.79
- P-value =P(|z|>5.79)=0.000 (approx).
- There is a strong evidence that voice mail plan is useful for predicting churn.

Odds Ratio and Confidence Interval
Churn data

- OR<-round(exp(coef(lr)),3)
- OR

- OR=0.47

- exp(confint(lr))
- exp(confint(lr,level=0.99))
- exp(confint(lr,level=0.90))
- 100(1-α)% C.I. for the Odds Ratio (OR)

$$\exp\left(b_1 \pm z.\, SE\left(b_1\right)\right) = (0.37, 0.61)$$

- We are 95 % confident that the OR for churning among voice mail plan members and non-members lies between 0.37 and 0.61.

Logistic regression for a polychotomous predictor

Interpreting Logistic regression for a polychotomous predictor

- For the churn dataset, suppose we categorize *Customer Service Calls* (CSC) as follows.
- *Zero or One CSC: CSC=Low*
- *Two or Three CSC: CSC= Medium*
- *Four or More CSC: CSC= High.*

R Zone

- churn$CSC<-factor(churn$CustServ.Calls)
- levels(churn$CSC)
- [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9"
- levels(churn$CSC)[0:2]<-"Low"
- levels(churn$CSC)[2:3]<-"Medium"
- levels(churn$CSC)[3:9]<-"High"
- churn$CSC_Med<-ifelse(churn$CSC =="Medium",1,0)
- churn$CSC_Hi<-ifelse(churn$CSC =="High",1,0)

R-zone

- table(churn$Churn,churn$CSC)
- Low Med High
- 0 1664 1057 129
- 1 214 131 138
- lr2<-glm (Churn ~ CSC_Med + CSC_Hi,data=churn, family ="binomial")
- summary(lr2)

Output

- glm(formula = Churn ~ CSC_Med + CSC_Hi, family = "binomial",

  data = churn)
- Deviance Residuals:
- Min 1Q Median 3Q Max
- -1.2062 -0.4919 -0.4919 -0.4834 2.0999

- Coefficients:
- Estimate Std. Error z value Pr(>|z|)
- (Intercept) -2.05100 0.07262 -28.243 <2e-16 ***
- CSC_Med -0.03699 0.11770 -0.314 0.753
- CSC_Hi 2.11844 0.14238 14.879 <2e-16 ***
- ---
- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- Null deviance: 2758.3 on 3332 degrees of freedom
- Residual deviance: 2526.7 on 3330 degrees of freedom

- The estimated logit is

$$\hat{g}(x) = -2.051 - 0.0369891\,(CSC\_Med) + 2.11844\,(CSC\_Hi)$$

Contd…

- For a customer with Low CSC, the probability of churning:

$$\hat{g}(0,0) = -2.051 \qquad \hat{\pi}(0,0) = \frac{e^{\hat{g}(0,0)}}{1+e^{\hat{g}(0,0)}} = 0.114$$

- For a customer with Medium CSC, the probability of churning:

$$\hat{g}(1,0) = -2.088 \qquad \hat{\pi}(1,0) = \frac{e^{\hat{g}(1,0)}}{1+e^{\hat{g}(1,0)}} = 0.110$$

- For a customer with High CSC, the probability of churning:

$$\hat{g}(0,1) = 0.06744 \qquad \hat{\pi}(0,1) = \frac{e^{\hat{g}(0,1)}}{1+e^{\hat{g}(0,1)}} = 0.5169$$

Contd…

- Clearly customers with high levels of customer service calls have a much higher estimated probability of churn. Company needs to focus customers who make four or more customer service calls.
- Wald test $Z_{wald\ (CSC\_Med)} = -0.31426$

  with a P-value =0.753.
- There is no evidence that the CSC_Med versus CSC_Low distinction is useful for predicting the churn. In the presence of other variables this variable is not adding any new information.

Contd…

- Wald test $Z_{wald\ (CSC\_High)} = 14.88$

  with a P-value $=P(|z|> 14.88)=0.000$.
- There is strong evidence that the CSC_High versus CSC_Low distinction is useful for predicting the churn.
- Same kind of analysis can be done when the predictor is a continuous variable. For eg. the predictor *Day Minutes* in Churn data set.