Algorithms for Data Guided Business Intelligence

Home Work

Topic 2, Part 1

By Rohit Mandge

Unity ID: ***RNMANDGE***

R code for preprocessing the data:

```
ebayData = read.table('/home/rnmandge/R/eBayAuctions.csv', header=TRUE, sep=",")
table2 <- table(ebayData$currency, ebayData$Competitive.)
totalCount<-table2[,1]+table2[,2]
table2[,1] <- table2[,1]/totalCount
table2[,2] <- table2[,2]/totalCount
table2
t(table2)
pivotTable1<-t(table2)

myTable<-table(ebayData$Category, ebayData$Competitive.)
totalCount<-myTable[,1]+myTable[,2]
myTable[,1] <- myTable[,1]/(myTable[,1]+myTable[,2])
myTable<-table(ebayData$Category, ebayData$Competitive.)
myTable[,1] <- myTable[,1]/totalCount
myTable[,2] <- myTable[,2]/totalCount
pivotTable2 <- t(myTable)

myTable1<-table(ebayData$endDay, ebayData$Competitive.)
totalCount<-myTable1[,1]+myTable1[,2]
myTable1[,2] <- myTable1[,2]/totalCount
myTable1[,1] <- myTable1[,1]/totalCount
pivotTable4 <- t(myTable1)
pivotTable4

myTable2<-table(ebayData$Duration, ebayData$Competitive.)
totalCount<-myTable2[,1]+myTable2[,2]
myTable2[,2] <- myTable2[,2]/totalCount
myTable2[,1] <- myTable2[,1]/totalCount
pivotTable3 <- t(myTable2)

ebayData$currency[ebayData$currency=='EUR'] <- 'US'

ebayData$Duration[ebayData$Duration=='7'] <- '3'
ebayData$Duration[ebayData$Duration=='10'] <- '1'

ebayData$endDay[ebayData$endDay=='Sat'] <- 'Fri'
ebayData$endDay[ebayData$endDay=='Sun'] <- 'Wed'

ebayData$Category[ebayData$Category =='Computer'] <- 'Business/Industrial'
ebayData$Category[ebayData$Category =='Pottery/Glass'] <- 'Automotive'

ebayData$Category[ebayData$Category =='Clothing/Accessories'] <- 'Books'
```

```
ebayData$Category[ebayData$Category =='Collectibles'] <- 'Antique/Art/Craft'

ebayData$Category[ebayData$Category =='Photography'] <- 'Electronics'

ebayData$endDay.f <- factor(ebayData$endDay)
ebayData$currency.f <- factor(ebayData$currency)
ebayData$Category.f <- factor(ebayData$Category)
ebayData$Duration.f<- factor(ebayData$Duration)
contrasts(ebayData$Duration.f) <- contr.treatment(3)
contrasts(ebayData$currency.f) <- contr.treatment(2)
contrasts(ebayData$endDay.f) <- contr.treatment(5)
contrasts(ebayData$Category.f) <- contr.treatment(13)

contrasts(ebayData$Duration.f)
factoredData <- ebayData
factoredData$Duration.f <- NULL
factoredData$currency.f <- NULL
factoredData$Category.f <- NULL
factoredData$endDay.f <- NULL

set.seed(345)
indexes <- sample(1:nrow(factoredData), size=0.4*nrow(factoredData))
validationData = factoredData[indexes,]
trainData = factoredData[-indexes,]

fit.full <- glm(Competitive. ~ Category + currency + sellerRating+Duration +endDay +ClosePrice
+OpenPrice,data = trainData,family = binomial(link="logit"))
summary(fit.full)
```

Question 1

```
fit.single <- glm (Competitive. ~ Category == "Automotive",data = trainData,family =
binomial(link="logit"))
summary(fit.single)
```

Question 4

```
fit.reduced <- glm(Competitive. ~ (Category == "Automotive") + (Category == "Books") +
(Category == "EverythingElse") + (Category == "Health/Beauty") + (currency == "US") +
sellerRating + (Duration == 5) + (endDay == "Mon") + (endDay == "Thu") + ClosePrice +
OpenPrice, data = trainData, family = binomial(link="logit"))
summary(fit.reduced)
anova(fit.reduced, fit.full, test = "Chisq")
```

Question 5

Over dispersion = Residual deviance/ Residual Df
 = 1135.0 / 1172
 = 0.968430034
Thus, the constructed model is not over dispersed.

Q. 1.) y_n (category) = "Automotive"

$$= 1/1+e^{-(\beta_0+\beta_1 x)}$$

a) Prob(y_{yes} | $x_n = x$)

$$= \frac{1}{1+e^{-(1.749e+00) - 9.220e-01 \; * \text{category} = \text{Automotive})}}$$

b) odds: prob(y = yes)

$$= e^{1.749e+00 - 9.220e-01 \; *(\text{category} = \text{"Automotive"})}$$

$$\doteq e^{\beta_0 + \beta_1 x} = \text{odds}$$

c) logit $= \beta_0 + \beta_1 x$

$$= 1.~~749~~e+00 \; (-9.220e-01 \; * \text{category} = \text{"Automotive"})$$

Q. 2) Top 4 predictions with highest estimates

1) category Automotive → x_1
2) Category Books → x_2
3) category Electronics → x_3
4) category coin/stamps → x_4

where, $\beta_1 = -9.220e-01$

$\beta_2 = -9.184e-01$

$\beta_3 = 8.059e-01$

$\beta_4 = -7.933e-01$

$\beta_0 \rightarrow$ 1.749e+00

a) logit functions as a function of predictors

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

b) The odds as a function of the predictors.

$$= e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4}$$

c) Probability as a function of predictors.

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}}$$

3) x_k = category = "Automotive"

$$\frac{odds\,(x_k + 1, x_2, \cdots x_q)}{odds\,(x_k, x_2, \cdots x_q)} = \frac{e^{\beta_0 + \beta_k (x_k + 1) + \beta_2 x_2 + \beta_3 x_3 + \cdots \beta_q x_q}}{e^{\beta_0 + \beta_k (x_k) + \beta_2 x_2 + \cdots \beta_q x_q}}$$

$$= e^{\cancel{\beta_k x_k} + \beta_k - \cancel{\beta_k x_k}}$$

$$= e^{\beta_k} = \boxed{e^{-9.220e-01}}$$

4) The reduced model is equivalent to the full model. The value obtained after applying chi-square test is greater than 0.05. Hence, it is equivalent.