11_27_logistic_Vincent

Ziwen Wang

11/27/2017

```
df<-read.csv("cleandata.csv", header = T)
#View(df)
head(df)
```

```
##   X targdol    datead6    datelp6 lpuryear slstyr slslyr sls2ago sls3ago
## 1 1       0 1980-01-01 1980-01-01       NA      0      0       0       0
## 2 2       0 1980-01-01 1980-01-01       NA      0      0       0       0
## 3 3       0 1980-01-01 1980-01-01       NA      0      0       0       0
## 4 4       0 1980-01-01 1980-01-01       NA      0      0       0       0
## 5 5       0 1980-01-01 1980-01-01       NA      0      0       0       0
## 6 6       0 1980-01-01 1980-01-01       NA      0      0       0       0
##   slshist ordtyr ordlyr ord2ago ord3ago ordhist falord sprord train
## 1       0      0      0       0       0       0      0      0     0
## 2       0      0      0       0       0       0      0      0     1
## 3       0      0      0       0       0       0      0      0     0
## 4       0      0      0       0       0       0      0      0     0
## 5       0      0      0       0       0       0      0      0     0
## 6       0      0      0       0       0       0      0      0     0
##   recentseason recency_year years_since_purchase consistencycategory
## 1       Spring         1980                   32                 Ref
## 2       Spring         1980                   32                 Ref
## 3       Spring         1980                   32                 Ref
## 4       Spring         1980                   32                 Ref
## 5       Spring         1980                   32                 Ref
## 6       Spring         1980                   32                 Ref
```

```
#change the targdol to 0:1
df$targdol[df$targdol!=0]=1
```

```
#buy rate
sum(df$targdol)/nrow(df)
```

```
## [1] 0.09426585
```

```
table(df$lpuryear, df$recency_year, useNA = "ifany")
```

```
##
##         1980  2002  2003  2004  2005  2006  2007  2008  2009  2010  2011
##   2003     0     0  1958    46     0     0     1     0     2     1    17
##   2004     0     0     0  3105     4     0     0     0     0     6    13
##   2005     0     0     0     0  5438     6     0     0     0    12    22
##   2006     0     0     0     0     0  6728    11     0     1    11    46
##   2007     0     0     0     0     0     0  9156     8   205    29    81
##   2008     0     0     0     0     0     0     0  3443  8349   281   135
##   2009     0     0     0     0     0     0     0     0  4208 11004   215
##   2010     0     0     0     0     0     0     0     0     0  4405 10507
##   2011     0     0     0     0     0     0     0     0     0     0  7256
##   2012     0     0     0     0     0     0     0     0     0     0     0
##   <NA>    17   687    12     6     0     0     0     0     0     0     3
```

```
##
##         2012
##   2003  2322
##   2004    24
##   2005    41
##   2006    36
##   2007    66
##   2008   119
##   2009   178
##   2010    19
##   2011 12477
##   2012  8812
##   <NA>     3
```

```
df$lpuryear[is.na(df$lpuryear) == TRUE] <- df$recency_year[is.na(df$lpuryear) == TRUE]
head(df)
```

```
##   X targdol    datead6    datelp6 lpuryear slstyr slslyr sls2ago sls3ago
## 1 1       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 2 2       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 3 3       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 4 4       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 5 5       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 6 6       0 1980-01-01 1980-01-01     1980      0      0       0       0
##   slshist ordtyr ordlyr ord2ago ord3ago ordhist falord sprord train
## 1       0      0      0       0       0       0      0      0     0
## 2       0      0      0       0       0       0      0      0     1
## 3       0      0      0       0       0       0      0      0     0
## 4       0      0      0       0       0       0      0      0     0
## 5       0      0      0       0       0       0      0      0     0
## 6       0      0      0       0       0       0      0      0     0
##   recentseason recency_year years_since_purchase consistencycategory
## 1       Spring         1980                   32                Ref
## 2       Spring         1980                   32                Ref
## 3       Spring         1980                   32                Ref
## 4       Spring         1980                   32                Ref
## 5       Spring         1980                   32                Ref
## 6       Spring         1980                   32                Ref
```

```
#drop the index
df <- subset(df, select = -c(X) )
```

```
#relevel
#factor
df$consistencycategory <- relevel(df$consistencycategory, ref = "Ref")
is.factor(df$consistencycategory)
```

```
## [1] TRUE
```

```
contrasts(df$consistencycategory)
```

```
##     1 10 11 2 3 4 5 6 7 8 9
## Ref 0  0  0 0 0 0 0 0 0 0 0
## 1   1  0  0 0 0 0 0 0 0 0 0
## 10  0  1  0 0 0 0 0 0 0 0 0
## 11  0  0  1 0 0 0 0 0 0 0 0
## 2   0  0  0 1 0 0 0 0 0 0 0
```

```
## 3   0  0  0 0 1 0 0 0 0 0 0
## 4   0  0  0 0 0 1 0 0 0 0 0
## 5   0  0  0 0 0 0 1 0 0 0 0
## 6   0  0  0 0 0 0 0 1 0 0 0
## 7   0  0  0 0 0 0 0 0 1 0 0
## 8   0  0  0 0 0 0 0 0 0 1 0
## 9   0  0  0 0 0 0 0 0 0 0 1
```

```
#change season category to 0(fall) and 1(spring)
is.factor(df$recentseason)
```

```
## [1] TRUE
```

```
df$recentseason<-as.character(df$recentseason)
df$recentseason[df$recentseason=='Spring']=1
df$recentseason[df$recentseason=='Fall']=0
df$recentseason<-as.factor(df$recentseason)
contrasts(df$recentseason)
```

```
##   1
## 0 0
## 1 1
```

```
#yearsince _pur and fallord
summary(df$falord)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.429   2.000 106.000
```

```
summary(df$years_since_purchase)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.725   5.000  32.000
```

```
#createc the new predictor named the month of last purchase
#check the type of date and year
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
class(df$datelp6)
```

```
## [1] "factor"
```

```
df$datelp6<-as.Date(df$datelp6)
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d", tz = "GMT"): unknown timezone
## 'zone/tz/2017c.1.0/zoneinfo/America/Chicago'
```

```
df$lpmonth<-month(df$datelp6)
df$lpmonth<-as.integer(df$lpmonth)
```

```
#change date to integer to facilitate the following modeling
#df$datelp6<-as.integer(df$datelp6)
```

```
#df$datead6<-as.integer(df$datead6)
```

```
#split the train and test
train<-df[df$train==1,]
test<-df[df$train==0,]
```

```
#choose the predictors
#train<-subset(train, select = -c(datead6,datelp6,lpuryear,train) )

train<-subset(train, select = -train)
head(train)
```

```
##    targdol    datead6    datelp6 lpuryear slstyr slslyr sls2ago sls3ago
## 2        0 1980-01-01 1980-01-01     1980      0      0       0       0
## 8        0 1980-01-01 1980-01-01     1980      0      0       0       0
## 9        0 1980-01-01 1980-01-01     1980      0      0       0       0
## 10       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 11       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 12       0 2005-09-17 2012-05-03     2012      0      0       0       0
##    slshist ordtyr ordlyr ord2ago ord3ago ordhist falord sprord
## 2        0      0      0       0       0       0      0      0
## 8        0      0      0       0       0       0      0      0
## 9        0      0      0       0       0       0      0      0
## 10       0      0      0       0       0       0      0      0
## 11       0      0      0       0       0       0      0      0
## 12       0      1      0       0       0       6      5      1
##    recentseason recency_year years_since_purchase consistencycategory
## 2             1         1980                    32                 Ref
## 8             1         1980                    32                 Ref
## 9             1         1980                    32                 Ref
## 10            1         1980                    32                 Ref
## 11            1         1980                    32                 Ref
## 12            1         2012                     0                 Ref
##    lpmonth
## 2        1
## 8        1
## 9        1
## 10       1
## 11       1
## 12       5
```

```
#correct the data type
train$targdol<-as.factor(train$targdol)
train$recentseason<-as.factor(train$recentseason)
```

```
#choose the variables in test
#test<-subset(test, select = -c(datead6,datelp6,lpuryear,train) )
test<-subset(test, select = -train)
head(test)
```

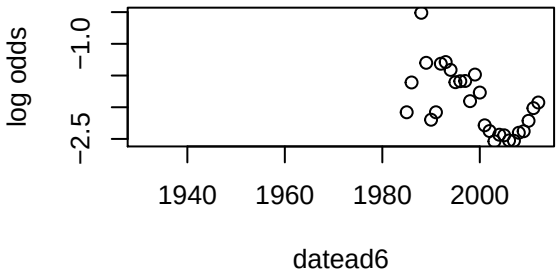
```
##   targdol    datead6    datelp6 lpuryear slstyr slslyr sls2ago sls3ago
## 1       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 3       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 4       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 5       0 1980-01-01 1980-01-01     1980      0      0       0       0
## 6       0 1980-01-01 1980-01-01     1980      0      0       0       0
```

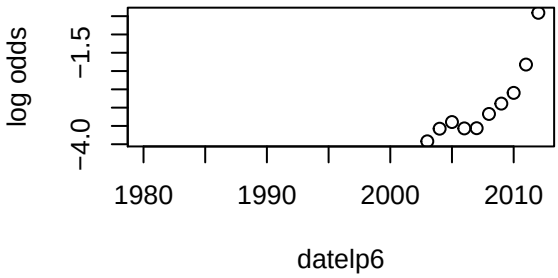
```
## 7       0 1980-01-01 1980-01-01     1980      0      0       0       0
##   slshist ordtyr ordlyr ord2ago ord3ago ordhist falord sprord recentseason
## 1       0      0      0       0       0       0      0      0            1
## 3       0      0      0       0       0       0      0      0            1
## 4       0      0      0       0       0       0      0      0            1
## 5       0      0      0       0       0       0      0      0            1
## 6       0      0      0       0       0       0      0      0            1
## 7       0      0      0       0       0       0      0      0            1
##   recency_year years_since_purchase consistencycategory lpmonth
## 1         1980                   32                 Ref       1
## 3         1980                   32                 Ref       1
## 4         1980                   32                 Ref       1
## 5         1980                   32                 Ref       1
## 6         1980                   32                 Ref       1
## 7         1980                   32                 Ref       1
```

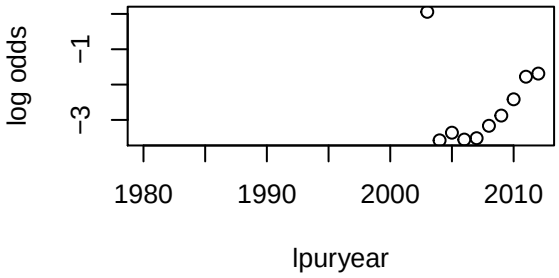
```
test$targdol<-as.factor(test$targdol)
test$recentseason<-as.factor(test$recentseason)
```

```
#frequency tables/logistic transform plot (see pg 131 of book)
frequency_table <- function(df,predictor, name) {
  tab <- as.data.frame.matrix(table(predictor, df$targdol))
  tab$buy_prop <- tab[,2]/rowSums(tab)
  tab$log_odds <- log(tab$buy_prop/(1 - tab$buy_prop))
  return (plot(x = as.numeric(rownames(tab)),
               y = tab$log_odds, xlab = name,
               ylab = 'log odds'))
}
```

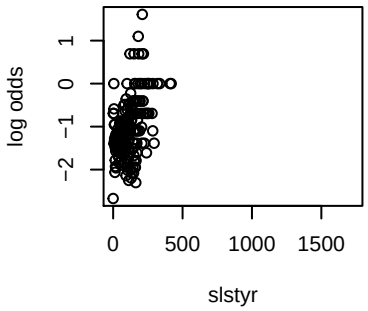
```
#plot
par(mfrow = c(2, 2))
frequency_table(train,year(train$datead6), 'datead6')
frequency_table(train,year(train$datelp6), 'datelp6')
frequency_table(train,train$lpuryear, 'lpuryear')
```

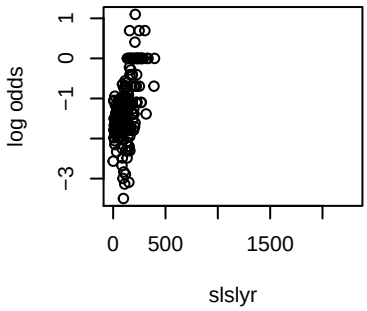


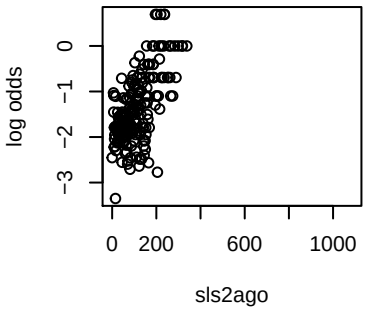


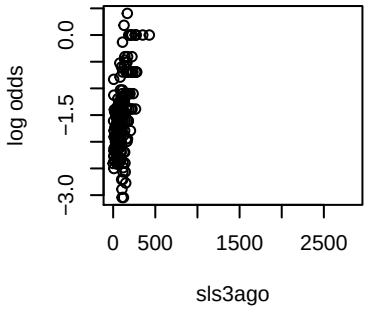


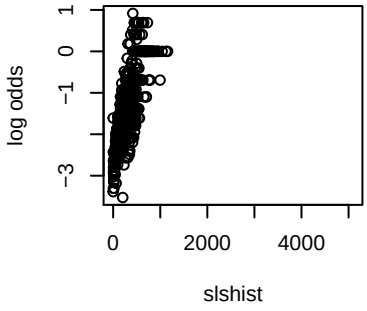

```r
par(mfrow = c(2, 3))
frequency_table(train,train$slstyr, 'slstyr')
frequency_table(train,train$slslyr, 'slslyr')
frequency_table(train,train$sls2ago, 'sls2ago')
frequency_table(train,train$sls3ago, 'sls3ago')
frequency_table(train,train$slshist, 'slshist')
```



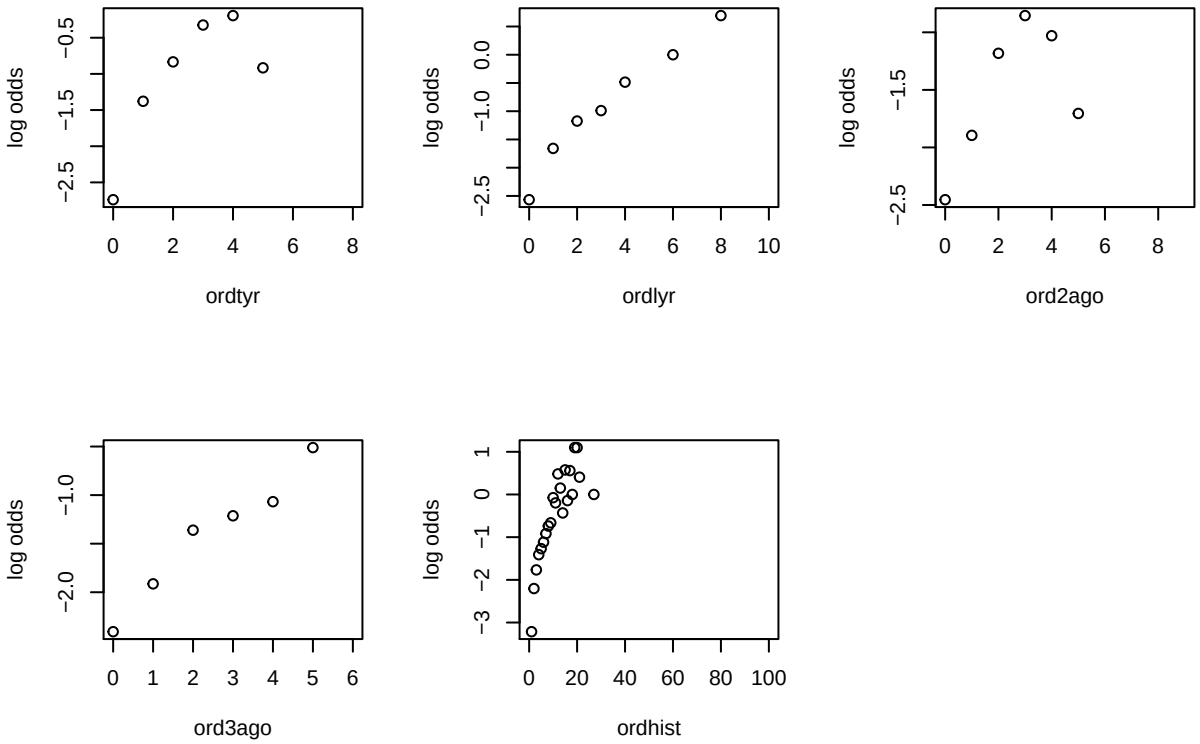




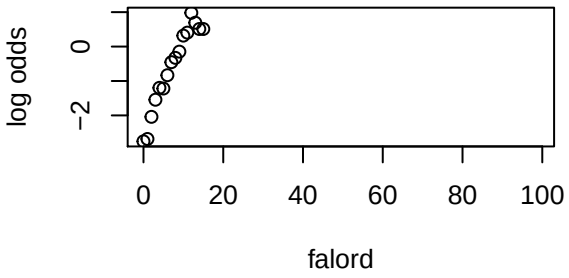


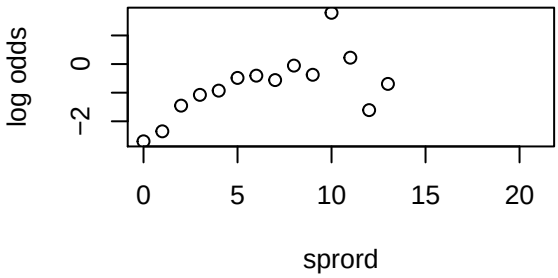


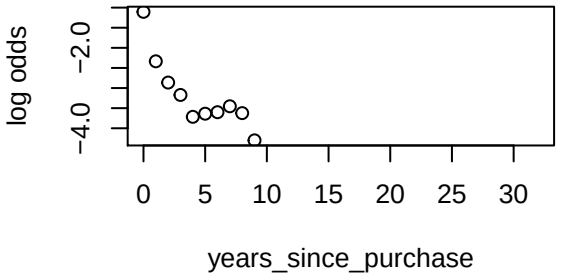
```r
par(mfrow = c(2, 3))

frequency_table(train,train$ordtyr, 'ordtyr')
frequency_table(train,train$ordlyr, 'ordlyr')
frequency_table(train,train$ord2ago, 'ord2ago')
frequency_table(train,train$ord3ago, 'ord3ago')
frequency_table(train,train$ordhist, 'ordhist')
```




```r
par(mfrow = c(2, 2))


frequency_table(train,train$falord, 'falord')
frequency_table(train,train$sprord, 'sprord')
frequency_table(train,train$years_since_purchase, 'years_since_purchase')
```








```
#change date to integer to facilitate the following modeling
train$datelp6<-as.integer(train$datelp6)
train$datead6<-as.integer(train$datead6)
test$datelp6<-as.integer(test$datelp6)
test$datead6<-as.integer(test$datead6)
```

```
#first_logistic_regression
model1 <- glm(targdol ~.,family=binomial(link='logit'),data=train)
summary(model1)
```

```
##
## Call:
## glm(formula = targdol ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9122  -0.3883  -0.2780  -0.1876   3.0042
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.784e+02  9.722e+01   2.863 0.004196 **
## datead6               -1.102e-04  2.916e-05  -3.778 0.000158 ***
## datelp6                1.024e-03  1.456e-04   7.032 2.04e-12 ***
## lpuryear              -4.776e-01  1.206e-02 -39.602  < 2e-16 ***
## slstyr                 9.225e-04  5.542e-04   1.664 0.096024 .
## slslyr                 4.097e-04  5.713e-04   0.717 0.473315
## sls2ago               -2.312e-04  7.371e-04  -0.314 0.753806
## sls3ago                6.088e-04  6.034e-04   1.009 0.312953
## slshist               -6.163e-04  2.577e-04  -2.392 0.016773 *
## ordtyr                 3.249e-01  5.094e-02   6.378 1.80e-10 ***
## ordlyr                 8.022e-02  4.754e-02   1.688 0.091497 .
```

```
## ord2ago               -1.361e-02  5.182e-02  -0.263 0.792883
## ord3ago               -7.532e-02  5.249e-02  -1.435 0.151285
## ordhist               -4.456e-02  2.257e-02  -1.974 0.048400 *
## falord                 2.696e-01  2.347e-02  11.487  < 2e-16 ***
## sprord                        NA         NA      NA       NA
## recentseason1          3.930e+00  1.324e-01  29.678  < 2e-16 ***
## recency_year           3.277e-01  4.855e-02   6.749 1.49e-11 ***
## years_since_purchase          NA         NA      NA       NA
## consistencycategory1   1.239e+00  1.417e-01   8.746  < 2e-16 ***
## consistencycategory10  4.958e-01  1.186e-01   4.181 2.90e-05 ***
## consistencycategory11  3.337e-02  1.355e-01   0.246 0.805464
## consistencycategory2   7.170e-01  1.346e-01   5.328 9.91e-08 ***
## consistencycategory3   7.337e-01  1.348e-01   5.441 5.30e-08 ***
## consistencycategory4   5.192e-01  1.614e-01   3.217 0.001295 **
## consistencycategory5   7.899e-01  1.482e-01   5.330 9.84e-08 ***
## consistencycategory6   6.827e-01  8.994e-02   7.590 3.19e-14 ***
## consistencycategory7   4.313e-01  1.051e-01   4.102 4.10e-05 ***
## consistencycategory8   3.209e-01  1.117e-01   2.872 0.004074 **
## consistencycategory9   2.037e-01  1.175e-01   1.733 0.083035 .
## lpmonth                4.372e-01  1.615e-02  27.066  < 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: 31907  on 50417  degrees of freedom
## Residual deviance: 22688  on 50389  degrees of freedom
## AIC: 22746
##
## Number of Fisher Scoring iterations: 6
```

```
anova(model1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: targdol
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 50417      31907
## datead6               1     73.4     50416      31833 < 2.2e-16 ***
## datelp6               1   4588.0     50415      27245 < 2.2e-16 ***
## lpuryear              1   2714.7     50414      24530 < 2.2e-16 ***
## slstyr                1     38.6     50413      24492 5.207e-10 ***
## slslyr                1     38.5     50412      24453 5.400e-10 ***
## sls2ago               1     35.4     50411      24418 2.654e-09 ***
## sls3ago               1     18.4     50410      24399 1.746e-05 ***
## slshist               1     23.3     50409      24376 1.391e-06 ***
## ordtyr                1    174.2     50408      24202 < 2.2e-16 ***
## ordlyr                1    189.3     50407      24013 < 2.2e-16 ***
## ord2ago               1    104.1     50406      23908 < 2.2e-16 ***
```

```
## ord3ago               1     64.3     50405      23844 1.080e-15 ***
## ordhist               1     88.4     50404      23756 < 2.2e-16 ***
## falord                1     48.5     50403      23707 3.245e-12 ***
## sprord                0      0.0     50403      23707
## recentseason          1    124.2     50402      23583 < 2.2e-16 ***
## recency_year          1      8.2     50401      23575  0.004241 **
## years_since_purchase  0      0.0     50401      23575
## consistencycategory  11    131.1     50390      23444 < 2.2e-16 ***
## lpmonth               1    755.4     50389      22688 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#second_logistic_regression
model2 <- glm(targdol ~ .-sprord-years_since_purchase,family=binomial(link='logit'),data=train)
summary(model2)
```

```
##
## Call:
## glm(formula = targdol ~ . - sprord - years_since_purchase, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9122  -0.3883  -0.2780  -0.1876   3.0042
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.784e+02  9.722e+01   2.863 0.004196 **
## datead6               -1.102e-04  2.916e-05  -3.778 0.000158 ***
## datelp6                1.024e-03  1.456e-04   7.032 2.04e-12 ***
## lpuryear              -4.776e-01  1.206e-02 -39.602  < 2e-16 ***
## slstyr                 9.225e-04  5.542e-04   1.664 0.096024 .
## slslyr                 4.097e-04  5.713e-04   0.717 0.473315
## sls2ago               -2.312e-04  7.371e-04  -0.314 0.753806
## sls3ago                6.088e-04  6.034e-04   1.009 0.312953
## slshist               -6.163e-04  2.577e-04  -2.392 0.016773 *
## ordtyr                 3.249e-01  5.094e-02   6.378 1.80e-10 ***
## ordlyr                 8.022e-02  4.754e-02   1.688 0.091497 .
## ord2ago               -1.361e-02  5.182e-02  -0.263 0.792883
## ord3ago               -7.532e-02  5.249e-02  -1.435 0.151285
## ordhist               -4.456e-02  2.257e-02  -1.974 0.048400 *
## falord                 2.696e-01  2.347e-02  11.487  < 2e-16 ***
## recentseason1          3.930e+00  1.324e-01  29.678  < 2e-16 ***
## recency_year           3.277e-01  4.855e-02   6.749 1.49e-11 ***
## consistencycategory1   1.239e+00  1.417e-01   8.746  < 2e-16 ***
## consistencycategory10  4.958e-01  1.186e-01   4.181 2.90e-05 ***
## consistencycategory11  3.337e-02  1.355e-01   0.246 0.805464
## consistencycategory2   7.170e-01  1.346e-01   5.328 9.91e-08 ***
## consistencycategory3   7.337e-01  1.348e-01   5.441 5.30e-08 ***
## consistencycategory4   5.192e-01  1.614e-01   3.217 0.001295 **
## consistencycategory5   7.899e-01  1.482e-01   5.330 9.84e-08 ***
## consistencycategory6   6.827e-01  8.994e-02   7.590 3.19e-14 ***
## consistencycategory7   4.313e-01  1.051e-01   4.102 4.10e-05 ***
## consistencycategory8   3.209e-01  1.117e-01   2.872 0.004074 **
## consistencycategory9   2.037e-01  1.175e-01   1.733 0.083035 .
```

```
## lpmonth                4.372e-01  1.615e-02  27.066  < 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: 31907  on 50417  degrees of freedom
## Residual deviance: 22688  on 50389  degrees of freedom
## AIC: 22746
##
## Number of Fisher Scoring iterations: 6
```

```
anova(model2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: targdol
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 50417      31907
## datead6               1     73.4     50416      31833 < 2.2e-16 ***
## datelp6               1   4588.0     50415      27245 < 2.2e-16 ***
## lpuryear              1   2714.7     50414      24530 < 2.2e-16 ***
## slstyr                1     38.6     50413      24492 5.207e-10 ***
## slslyr                1     38.5     50412      24453 5.400e-10 ***
## sls2ago               1     35.4     50411      24418 2.654e-09 ***
## sls3ago               1     18.4     50410      24399 1.746e-05 ***
## slshist               1     23.3     50409      24376 1.391e-06 ***
## ordtyr                1    174.2     50408      24202 < 2.2e-16 ***
## ordlyr                1    189.3     50407      24013 < 2.2e-16 ***
## ord2ago               1    104.1     50406      23908 < 2.2e-16 ***
## ord3ago               1     64.3     50405      23844 1.080e-15 ***
## ordhist               1     88.4     50404      23756 < 2.2e-16 ***
## falord                1     48.5     50403      23707 3.245e-12 ***
## recentseason          1    124.2     50402      23583 < 2.2e-16 ***
## recency_year          1      8.2     50401      23575  0.004241 **
## consistencycategory 11    131.1     50390      23444 < 2.2e-16 ***
## lpmonth               1    755.4     50389      22688 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#prediction
#first prediction
library(pscl)
```

```
## Warning: package 'pscl' was built under R version 3.4.2
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
```

```
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
#pR2(model)
test_predict <- predict(model2, newdata = test, type = 'response')
table(test$targdol, round(test_predict, 1), dnn = c('true', 'predicted'))
```

```
##     predicted
## true     0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9     1
##    0 26552 15834  2311   852   353   211   109    60    41    57     8
##    1   520  1611   546   372   142   118    98    80   174   743   322
```

```
test_predict[test_predict < .1] <- 0
test_predict[test_predict >= .1] <- 1
tab <- table(test$targdol, test_predict, dnn = c('true', 'predicted'))
tab
```

```
##     predicted
## true     0     1
##    0 38267  8121
##    1  1463  3263
```

```
cat('CCR:', sum(diag(tab))/sum(tab))
```

```
## CCR: 0.8124976
```

```
cat('\nTrue targdol of predicted buyers', sum(as.integer(test$targdol[test_predict == 1])))
```

```
##
## True targdol of predicted buyers 14647
```

```
#prediction
#optCutOff
#decide a optimal cutoff
library(InformationValue)
optCutOff <- optimalCutoff(test$targdol,test_predict)[1]
optCutOff
```

```
## [1] 0.01
```

```
#prediction
#second prediction - cutoff<- 0.01
library(pscl)
#pR2(model)
test_predict <- predict(model2, newdata = test, type = 'response')
table(test$targdol, round(test_predict, 1), dnn = c('true', 'predicted'))
```

```
##     predicted
## true     0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9     1
##    0 26552 15834  2311   852   353   211   109    60    41    57     8
##    1   520  1611   546   372   142   118    98    80   174   743   322
```

```
test_predict[test_predict < .4] <- 0
test_predict[test_predict >= .4] <- 1
tab <- table(test$targdol, test_predict, dnn = c('true', 'predicted'))
tab
```

```
##     predicted
## true     0     1
##    0 45755   633
```

```
##    1  3122  1604
```

```
cat('CCR:', sum(diag(tab))/sum(tab))
```

```
## CCR: 0.9265368
```

```
cat('\nTrue targdol of predicted buyers', sum(as.integer(test$targdol[test_predict == 1])))
```

```
##
## True targdol of predicted buyers 3841
```

```
#fitted.results2 = predict(mod_fit, newdata=test_balanced)
fitted.results <- predict(model2,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.4,1,0)
misClasificError <- mean(fitted.results != test$targdol)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.926536760965684"
```

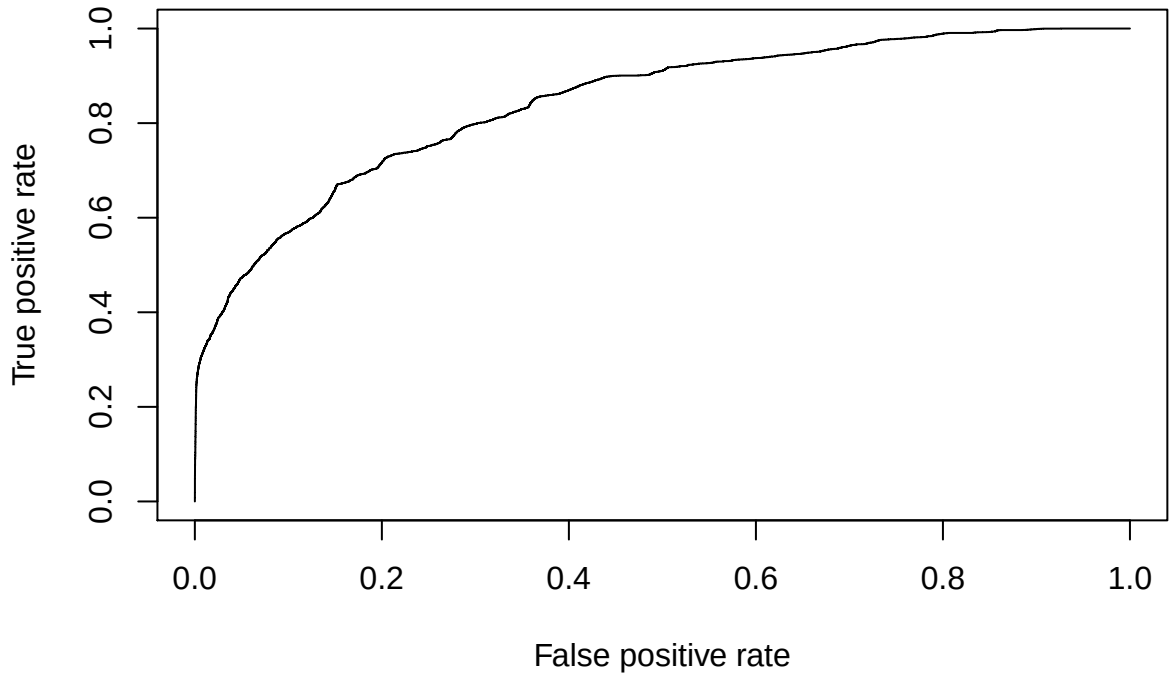
```
#ROC
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

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

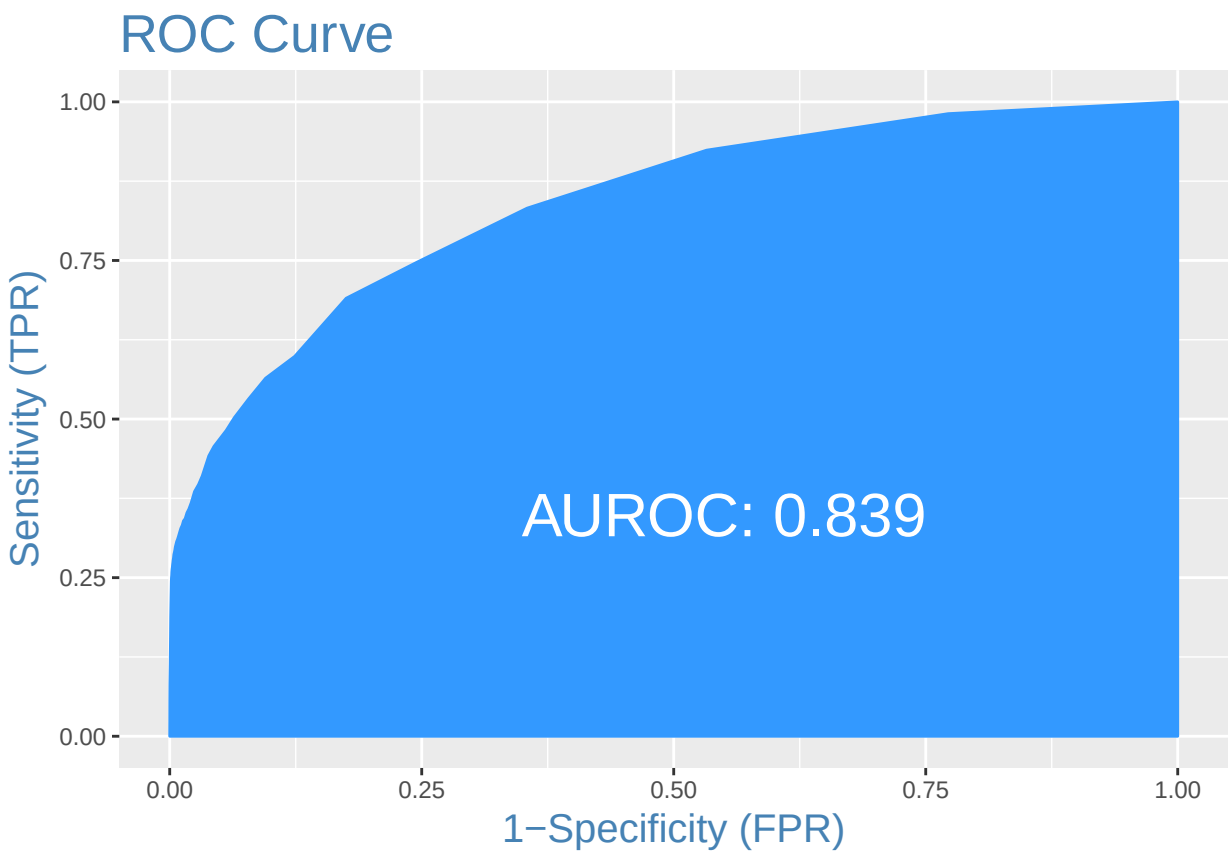
```
p <- predict(model2, newdata=test, type="response")

pr <- prediction(p, test$targdol)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```



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

```
## [1] 0.8426425
```

```
plotROC(test$targdol, p)
```




```
#
pr <- prediction(p, test$targdol)
test$targdol<-as.factor(test$targdol)
#p<-as.factor(p)
sensitivity(test$targdol, p, threshold = 0.4)
```

```
## [1] 0.3393991
```

```
specificity(test$targdol, p, threshold = 0.4)
```

```
## [1] 0.9863542
```

```
confusionMatrix(test$targdol, p, threshold = 0.4)
```

```
##       0    1
## 0 45755 3122
## 1   633 1604
```

```
# The columns are actuals, while rows are predicteds.

# Use your model to make predictions, in this example newdata = training set, but replace with your tes
```

```r
# use caret and compute a confusion matrix
confusionMatrix(test$targdol, p, threshold = 0.4)
```

```
##       0    1
## 0 45755 3122
## 1   633 1604
```