Logistic Regression for Online Shopper's Intention Data (with Cross Validation)

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#Importing required libraries
library(gmodels)
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(boot)

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

options(warn = -1)

#Importing the Online shoppers intention data file
OSI_df =
read.csv(file="C:/Users/delld/Downloads/online_shoppers_intention.csv",header
= TRUE, stringsAsFactors = TRUE)

#Check fo missing data
TM<-sum(is.na(OSI_df))
cat("Total missing values:",TM)

## Total missing values: 0

#Descriptives
summary(OSI_df)

##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
```

```
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser           Region
##  May    :3364   Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mar    :1907   Median :2.000    Median : 2.000   Median :3.000
##  Dec    :1727   Mean   :2.124    Mean   : 2.357   Mean   :3.147
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000
##  (Other):1337
##   TrafficType                VisitorType     Weekend          Revenue
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422
##  Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

```
#Logistic Regression for the complete data
OSI_Logit = glm(Revenue~.,data = OSI_df,family = binomial)
summary(OSI_Logit)
```

```
##
## Call:
## glm(formula = Revenue ~ ., family = binomial, data = OSI_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.1072  -0.4663  -0.3328  -0.1648   3.3801
```

```
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.619e+00  2.012e-01  -8.046 8.56e-16 ***
## Administrative                 5.108e-03  1.100e-02   0.464 0.642439
## Administrative_Duration       -1.225e-04  1.943e-04  -0.630 0.528372
## Informational                  3.341e-02  2.703e-02   1.236 0.216373
## Informational_Duration         7.117e-05  2.217e-04   0.321 0.748197
## ProductRelated                 1.718e-03  1.153e-03   1.490 0.136228
## ProductRelated_Duration        6.075e-05  2.705e-05   2.245 0.024744 *
## BounceRates                   -3.788e+00  3.254e+00  -1.164 0.244333
## ExitRates                     -1.559e+01  2.399e+00  -6.498 8.17e-11 ***
## PageValues                     8.217e-02  2.415e-03  34.021  < 2e-16 ***
## SpecialDay                    -1.228e-01  2.362e-01  -0.520 0.603109
## MonthDec                      -5.944e-01  1.821e-01  -3.263 0.001101 **
## MonthFeb                      -1.750e+00  6.384e-01  -2.741 0.006131 **
## MonthJul                       8.142e-02  2.184e-01   0.373 0.709261
## MonthJune                     -3.094e-01  2.751e-01  -1.125 0.260729
## MonthMar                      -5.071e-01  1.802e-01  -2.815 0.004879 **
## MonthMay                      -5.520e-01  1.739e-01  -3.174 0.001502 **
## MonthNov                       5.467e-01  1.627e-01   3.360 0.000780 ***
## MonthOct                      -2.521e-04  2.018e-01  -0.001 0.999003
## MonthSep                       3.101e-03  2.123e-01   0.015 0.988347
## OperatingSystems              -7.930e-02  3.892e-02  -2.037 0.041602 *
## Browser                        4.301e-02  1.874e-02   2.295 0.021731 *
## Region                        -1.228e-02  1.310e-02  -0.937 0.348523
## TrafficType                    3.322e-03  8.302e-03   0.400 0.689095
## VisitorTypeOther              -5.010e-01  5.524e-01  -0.907 0.364480
## VisitorTypeReturning_Visitor -3.267e-01  8.576e-02  -3.810 0.000139 ***
## WeekendTRUE                    1.026e-01  7.102e-02   1.444 0.148694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10624.8  on 12329  degrees of freedom
## Residual deviance:  7153.1  on 12303  degrees of freedom
## AIC: 7207.1
##
## Number of Fisher Scoring iterations: 7

#Dividing Data into training (70%) and testing (30%) set
sample_size = floor(0.70*nrow(OSI_df))
train_ind = sample(seq_len(nrow(OSI_df)),size = sample_size)
train =OSI_df[train_ind,]
test=OSI_df[-train_ind,]

#Logistic Model for the training data
OSI_Logit1 = glm(Revenue~.,data = train,family = binomial)
summary(OSI_Logit1)
```

```
##
## Call:
## glm(formula = Revenue ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8194  -0.4494  -0.3225  -0.1604   3.4427
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.634e+00  2.453e-01  -6.660 2.73e-11 ***
## Administrative                6.530e-03  1.325e-02   0.493 0.622012
## Administrative_Duration      -7.526e-05  2.310e-04  -0.326 0.744597
## Informational                 6.646e-02  3.271e-02   2.032 0.042184 *
## Informational_Duration       -8.911e-05  2.807e-04  -0.317 0.750921
## ProductRelated                5.744e-04  1.404e-03   0.409 0.682404
## ProductRelated_Duration       6.797e-05  3.207e-05   2.119 0.034051 *
## BounceRates                  -7.098e+00  4.184e+00  -1.696 0.089824 .
## ExitRates                    -1.418e+01  2.908e+00  -4.874 1.09e-06 ***
## PageValues                    9.088e-02  3.127e-03  29.059  < 2e-16 ***
## SpecialDay                    1.476e-01  2.718e-01   0.543 0.587105
## MonthDec                     -6.399e-01  2.236e-01  -2.862 0.004212 **
## MonthFeb                     -1.749e+00  7.750e-01  -2.257 0.024006 *
## MonthJul                      8.208e-02  2.647e-01   0.310 0.756507
## MonthJune                    -3.959e-01  3.435e-01  -1.152 0.249121
## MonthMar                     -5.900e-01  2.200e-01  -2.682 0.007317 **
## MonthMay                     -6.553e-01  2.125e-01  -3.084 0.002045 **
## MonthNov                      4.935e-01  1.986e-01   2.485 0.012951 *
## MonthOct                     -1.554e-01  2.508e-01  -0.619 0.535633
## MonthSep                     -1.417e-01  2.652e-01  -0.534 0.593093
## OperatingSystems             -4.894e-02  4.766e-02  -1.027 0.304470
## Browser                       4.642e-02  2.283e-02   2.033 0.042049 *
## Region                       -1.463e-02  1.604e-02  -0.912 0.361808
## TrafficType                  -1.053e-02  1.051e-02  -1.002 0.316584
## VisitorTypeOther             -3.446e-01  7.000e-01  -0.492 0.622535
## VisitorTypeReturning_Visitor -3.459e-01  1.048e-01  -3.302 0.000961 ***
## WeekendTRUE                   8.664e-02  8.700e-02   0.996 0.319317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7336.1  on 8630  degrees of freedom
## Residual deviance: 4786.3  on 8604  degrees of freedom
## AIC: 4840.3
##
## Number of Fisher Scoring iterations: 7
```

```r
#Predict the probabilities and classes for the test data
OSI_Logit1.prob = predict.glm(OSI_Logit1,newdata = test,type = "response")
OSI_Logit1.prob[1:10]
```

```
##            5            8           11           13           17
18
## 0.0108920212 0.0003656176 0.0088082736 0.0153762272 0.0003110017
0.0223087271
##           19           25           29           30
## 0.0055178758 0.0003454820 0.0138295103 0.5870577667
```

```r
OSI_Logit1.pred <- rep("FALSE",nrow(test))
nrow(test)
```

```
## [1] 3699
```

```r
OSI_Logit1.pred[OSI_Logit1.prob>0.5]<-"TRUE"

#Display the Cross Table for pred vs actual responses
CrossTable(OSI_Logit1.pred,test$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3699
##
##
##                  | test$Revenue
## OSI_Logit1.pred |     FALSE |      TRUE | Row Total |
## ----------------|-----------|-----------|-----------|
##           FALSE |      3007 |       371 |      3378 |
##                 |    11.298 |    58.125 |           |
##                 |     0.890 |     0.110 |     0.913 |
##                 |     0.971 |     0.616 |           |
##                 |     0.813 |     0.100 |           |
## ----------------|-----------|-----------|-----------|
##            TRUE |        90 |       231 |       321 |
##                 |   118.897 |   611.667 |           |
##                 |     0.280 |     0.720 |     0.087 |
##                 |     0.029 |     0.384 |           |
##                 |     0.024 |     0.062 |           |
## ----------------|-----------|-----------|-----------|
```

```
##    Column Total |      3097 |       602 |      3699 |
##                 |     0.837 |     0.163 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
#Display the Confusion Matrix
test$Revenue=as.factor(test$Revenue)
Predicted=as.factor(OSI_Logit1.pred)
confusionMatrix(Predicted,test$Revenue, positive = "TRUE")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3007  371
##      TRUE     90  231
##
##                Accuracy : 0.8754
##                  95% CI : (0.8643, 0.8859)
##     No Information Rate : 0.8373
##     P-Value [Acc > NIR] : 4.803e-11
##
##                   Kappa : 0.4368
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.38372
##             Specificity : 0.97094
##          Pos Pred Value : 0.71963
##          Neg Pred Value : 0.89017
##              Prevalence : 0.16275
##          Detection Rate : 0.06245
##    Detection Prevalence : 0.08678
##       Balanced Accuracy : 0.67733
##
##        'Positive' Class : TRUE
##
```

```
#Cross Validation using train function
train_control <- trainControl(method = "cv", number = 10)
train$Revenue=as.factor(train$Revenue)
OSI_Logit_cv <- train(Revenue~.,
               data = train,
               trControl = train_control,
               method = "glm",
               family=binomial())
summary(OSI_Logit_cv)
```

```
##
## Call:
```

```
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8194  -0.4494  -0.3225  -0.1604   3.4427
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.634e+00  2.453e-01  -6.660 2.73e-11 ***
## Administrative                6.530e-03  1.325e-02   0.493 0.622012
## Administrative_Duration      -7.526e-05  2.310e-04  -0.326 0.744597
## Informational                 6.646e-02  3.271e-02   2.032 0.042184 *
## Informational_Duration       -8.911e-05  2.807e-04  -0.317 0.750921
## ProductRelated                5.744e-04  1.404e-03   0.409 0.682404
## ProductRelated_Duration       6.797e-05  3.207e-05   2.119 0.034051 *
## BounceRates                  -7.098e+00  4.184e+00  -1.696 0.089824 .
## ExitRates                    -1.418e+01  2.908e+00  -4.874 1.09e-06 ***
## PageValues                    9.088e-02  3.127e-03  29.059  < 2e-16 ***
## SpecialDay                    1.476e-01  2.718e-01   0.543 0.587105
## MonthDec                     -6.399e-01  2.236e-01  -2.862 0.004212 **
## MonthFeb                     -1.749e+00  7.750e-01  -2.257 0.024006 *
## MonthJul                      8.208e-02  2.647e-01   0.310 0.756507
## MonthJune                    -3.959e-01  3.435e-01  -1.152 0.249121
## MonthMar                     -5.900e-01  2.200e-01  -2.682 0.007317 **
## MonthMay                     -6.553e-01  2.125e-01  -3.084 0.002045 **
## MonthNov                      4.935e-01  1.986e-01   2.485 0.012951 *
## MonthOct                     -1.554e-01  2.508e-01  -0.619 0.535633
## MonthSep                     -1.417e-01  2.652e-01  -0.534 0.593093
## OperatingSystems             -4.894e-02  4.766e-02  -1.027 0.304470
## Browser                       4.642e-02  2.283e-02   2.033 0.042049 *
## Region                       -1.463e-02  1.604e-02  -0.912 0.361808
## TrafficType                  -1.053e-02  1.051e-02  -1.002 0.316584
## VisitorTypeOther             -3.446e-01  7.000e-01  -0.492 0.622535
## VisitorTypeReturning_Visitor -3.459e-01  1.048e-01  -3.302 0.000961 ***
## WeekendTRUE                   8.664e-02  8.700e-02   0.996 0.319317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7336.1  on 8630  degrees of freedom
## Residual deviance: 4786.3  on 8604  degrees of freedom
## AIC: 4840.3
##
## Number of Fisher Scoring iterations: 7

OSI_Logit_cv_pred = predict(OSI_Logit_cv,test)
CrossTable(OSI_Logit_cv_pred,test$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3699
##
##
##                    | test$Revenue
## OSI_Logit_cv_pred |     FALSE |      TRUE | Row Total |
## ------------------|-----------|-----------|-----------|
##             FALSE |      3007 |       371 |      3378 |
##                   |    11.298 |    58.125 |           |
##                   |     0.890 |     0.110 |     0.913 |
##                   |     0.971 |     0.616 |           |
##                   |     0.813 |     0.100 |           |
## ------------------|-----------|-----------|-----------|
##              TRUE |        90 |       231 |       321 |
##                   |   118.897 |   611.667 |           |
##                   |     0.280 |     0.720 |     0.087 |
##                   |     0.029 |     0.384 |           |
##                   |     0.024 |     0.062 |           |
## ------------------|-----------|-----------|-----------|
##      Column Total |      3097 |       602 |      3699 |
##                   |     0.837 |     0.163 |           |
## ------------------|-----------|-----------|-----------|
##
##
```

```
confusionMatrix(OSI_Logit_cv_pred,test$Revenue, positive = "TRUE")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3007  371
##      TRUE     90  231
##
##                Accuracy : 0.8754
##                  95% CI : (0.8643, 0.8859)
##     No Information Rate : 0.8373
##     P-Value [Acc > NIR] : 4.803e-11
##
```

```
##                   Kappa : 0.4368
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.38372
##             Specificity : 0.97094
##          Pos Pred Value : 0.71963
##          Neg Pred Value : 0.89017
##              Prevalence : 0.16275
##          Detection Rate : 0.06245
##    Detection Prevalence : 0.08678
##       Balanced Accuracy : 0.67733
##
##        'Positive' Class : TRUE
##
```

```
# CV using boot
# K-fold CV K=5 (accuracy); Cost as defined in the cost function
# Cost function for a binary classifier suggested by boot package
cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)
cat("Accuracy with cost function:",1-cv.glm(train,OSI_Logit,K=5,cost =
cost)$delta[1])
```

```
## Accuracy with cost function: 0.8063956
```

```
# K-fold CV K=5 (accuracy); Cost is default;average squared error function
cat("Accuracy with default cost function:",1-
cv.glm(train,OSI_Logit,K=5)$delta[1])
```

```
## Accuracy with default cost function: 0.8350696
```

```
#Displaying error in each of the cross-validation iteration
cv.err = rep(0,10)
for (i in 1:10){
  OSI_Logit = glm(Revenue~.,data = OSI_df,family = binomial)
  cv.err[i] = cv.glm(train,OSI_Logit,K=10)$delta[1]
}
cat("Error:\n",cv.err)
```

```
## Error:
##  0.1650221 0.1649781 0.1648348 0.1648761 0.1647916 0.1648799 0.1648519
0.1650141 0.1647913 0.1649275
```

```
cat("Accuracy:\n",1-cv.err)
```

```
## Accuracy:
##  0.8349779 0.8350219 0.8351652 0.8351239 0.8352084 0.8351201 0.8351481
0.8349859 0.8352087 0.8350725
```

```
cat("Average Accuracy:\n",mean(1-cv.err))
```

```
## Average Accuracy:
##  0.8351033
```