Great Lakes PGPBABI -Hyderabad

Predictive Modelling – Assignment

Logistic Regression

Group -9

By

A Ramakrishna Reddy

Jayati Nevatia

Karnam Uday

Rajesh Kumar Pandey

**V Ramayya Nadipalli**

**Problem statement:** Develop a Logistic Regression Model on Simmons Catalogue dataset: Logit – Simmons.xls. Need to understand customer purchase behavior.

**Approach:** The team followed the basic steps needed for Logistic Regression model creation. There are 100 observations and 5 variables. The variables descriptions are provided below:

* Customer = Customer ID
* Spending = Amount customer spent last year at Simmons (in $1000)
* Card = Whether customer has Simmons Credit card (value 1) or not (value 0)
* Purchase = Whether customer made purchase worth $200 (value 1) or not (value 0)
* SpendCat = High spending group (value 1) vs low spending group (value 0)

Based on the data provided for this case study, team looked into below customer behavior:

* Purchase to Card, Spending and spend category relationship
* Purchase to Card and Spending relationship
* Purchase to Spending and spend category relationship.
* Purchase to Card and spend category relationship.
* Purchase to Card relationship.
* Purchase to Spending relationship.
* Purchase to spend category relationship.

The R Markdown file generated from R studio is attached for review.

Group\_Assignment\_Logistic\_Regression\_Group9\_Submission.R

Wed Apr 12 00:34:57 2017

######################Download data in R ###########################  
library(readr)

## Warning: package 'readr' was built under R version 3.3.3

Logit\_Simmons\_data <- read\_csv("C:/GLIM/PM/Group Assignment/Logit-Simmons.csv")

## Parsed with column specification:  
## cols(  
## Customer = col\_integer(),  
## Spending = col\_double(),  
## Card = col\_integer(),  
## Purchase = col\_integer(),  
## SpendCat = col\_integer()  
## )

######################Data Analysis#################################  
Sim\_data <- data.frame(Logit\_Simmons\_data)  
head(Sim\_data)

## Customer Spending Card Purchase SpendCat  
## 1 1 2.291 1 0 0  
## 2 2 3.215 1 0 1  
## 3 3 2.135 1 0 0  
## 4 4 3.924 0 0 1  
## 5 5 2.528 1 0 0  
## 6 6 2.473 0 1 0

str(Sim\_data)

## 'data.frame': 100 obs. of 5 variables:  
## $ Customer: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Spending: num 2.29 3.21 2.13 3.92 2.53 ...  
## $ Card : int 1 1 1 0 1 0 0 0 1 0 ...  
## $ Purchase: int 0 0 0 0 0 1 0 0 1 0 ...  
## $ SpendCat: int 0 1 0 1 0 0 0 1 0 1 ...

## check missing values  
sapply(Sim\_data,function(x) sum(is.na(x)))

## Customer Spending Card Purchase SpendCat   
## 0 0 0 0 0

# found none  
  
## more data Insights  
library(MASS)

## Warning: package 'MASS' was built under R version 3.3.2

sapply(Sim\_data,sd)

## Customer Spending Card Purchase SpendCat   
## 29.0114920 1.7412979 0.5025189 0.4923660 0.5009083

##################### Build the Logistic Model #####################  
## We are not considering Customer variable as it is unnecessary  
## Taking all the variables  
logit\_S=glm(Purchase~Spending+Card+SpendCat, data=Sim\_data, family=binomial)   
## Taking only Spending and Card  
logit\_S1=glm(Purchase~Spending+ Card, data=Sim\_data, family=binomial)   
## Taking only Spending and Spend category  
logit\_S2=glm(Purchase~Spending+SpendCat, data=Sim\_data, family=binomial)   
## Taking only Card and spend category  
logit\_S3=glm(Purchase~Card+SpendCat, data=Sim\_data, family=binomial)   
## Taking only Card   
logit\_S4=glm(Purchase~Card, data=Sim\_data, family=binomial)   
## Taking only spending  
logit\_S5=glm(Purchase~Spending, data=Sim\_data, family=binomial)   
## Taking only spend category  
logit\_S6=glm(Purchase~SpendCat, data=Sim\_data, family=binomial)   
#  
## Check the model parameters   
summary(logit\_S)

##   
## Call:  
## glm(formula = Purchase ~ Spending + Card + SpendCat, family = binomial,   
## data = Sim\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5845 -0.9583 -0.5785 0.8610 1.9409   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.78531 0.61453 -2.905 0.00367 \*\*  
## Spending 0.04019 0.21540 0.187 0.85199   
## Card 1.14784 0.45446 2.526 0.01155 \*   
## SpendCat 1.30000 0.76018 1.710 0.08724 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 134.60 on 99 degrees of freedom  
## Residual deviance: 117.99 on 96 degrees of freedom  
## AIC: 125.99  
##   
## Number of Fisher Scoring iterations: 4

# All the parameters looks statistically significant.  
# A unit increase for customers with Card increases the log odds by 1.15   
# while same 1 unit increase in spending category increases the log odds by 1.30  
  
## Run the anova() function on the model to analyze the table of deviance   
anova(logit\_S, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Purchase  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 99 134.60   
## Spending 1 7.2182 98 127.38 0.007217 \*\*  
## Card 1 6.4103 97 120.97 0.011346 \*   
## SpendCat 1 2.9828 96 117.99 0.084156 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# There is a drop in deviance of around 1 unit by adding card to spending and 4 units by  
# adding spending category. The model seems to be doing good compared to null model (NULL deviance)  
# Check the McFadden R2 index equivalent to R2 for linear regression to access model fit.  
library(pscl)

## Warning: package 'pscl' was built under R version 3.3.2

## Loading required package: lattice

## 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(logit\_S)

## llh llhNull G2 McFadden r2ML r2CU   
## -58.9955701 -67.3011667 16.6111932 0.1234094 0.1530486 0.2068988

#llh The log-likelihood from the fitted model  
#llhNull The log-likelihood from the intercept-only restricted model  
#G2 Minus two times the difference in the log-likelihoods  
#McFadden McFadden's pseudo r-squared  
#r2ML Maximum likelihood pseudo r-squared  
#r2CU Cragg and Uhler's pseudo r-squared  
#  
# CIs using profiled log-likelihood :: confint function to obtain confidence  
# intervals of coefficient estimates  
confint(logit\_S)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) -3.0549241 -0.6254534  
## Spending -0.3834657 0.4683343  
## Card 0.2774595 2.0704407  
## SpendCat -0.1747875 2.8296337

# CI using standard errors  
confint.default(logit\_S)

## 2.5 % 97.5 %  
## (Intercept) -2.9897697 -0.5808427  
## Spending -0.3819913 0.4623708  
## Card 0.2571094 2.0385737  
## SpendCat -0.1899221 2.7899250

#Exponentiate the coefficients to interpret the results as odds-ratios  
# odds ratios  
exp(coef(logit\_S))

## (Intercept) Spending Card SpendCat   
## 0.1677457 1.0410083 3.1513834 3.6693019

# Odds ratios and 95% CI  
exp(cbind(OR = coef(logit\_S), confint(logit\_S)))

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 0.1677457 0.0471263 0.5350188  
## Spending 1.0410083 0.6814954 1.5973313  
## Card 3.1513834 1.3197727 7.9283167  
## SpendCat 3.6693019 0.8396354 16.9392542

##################Calculate the predicted probability and accuracy ##########  
# We aded a new variable PredPurchase to original dataset to see how our model is predicting.  
# Note that Type = response tells R to calculate predicted probability  
Sim\_data$PredPurchase <- predict(logit\_S, type='response')   
head(Sim\_data$PredPurchase)

## [1] 0.3669346 0.6882063 0.3654795 0.4188246 0.3691500 0.1563136

Sim\_data$PredPurchase <- ifelse(Sim\_data$PredPurchase > 0.5,1,0)  
misClassificationError <- mean(Sim\_data$PredPurchase != Sim\_data$Purchase)  
print(paste('Accuracy of Model is : ', 1 - misClassificationError))

## [1] "Accuracy of Model is : 0.75"

################### Test the Model accuracy ################################  
## Check the measures to see how well our model fits using Likelihood ratio test.  
library(lmtest)

## Warning: package 'lmtest' was built under R version 3.3.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.3.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

lrtest(logit\_S)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Spending + Card + SpendCat  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 4 -58.996   
## 2 1 -67.301 -3 16.611 0.0008495 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S1)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Spending + Card  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 3 -60.487   
## 2 1 -67.301 -2 13.628 0.001098 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S2)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Spending + SpendCat  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 3 -62.370   
## 2 1 -67.301 -2 9.8618 0.00722 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S3)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Card + SpendCat  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 3 -59.013   
## 2 1 -67.301 -2 16.576 0.0002515 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S4)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Card  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 2 -64.265   
## 2 1 -67.301 -1 6.0723 0.01373 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S5)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ Spending  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 2 -63.692   
## 2 1 -67.301 -1 7.2182 0.007217 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lrtest(logit\_S6)

## Likelihood ratio test  
##   
## Model 1: Purchase ~ SpendCat  
## Model 2: Purchase ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 2 -62.396   
## 2 1 -67.301 -1 9.8113 0.001734 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Compare the Log likelihood  
Compare\_models\_logLik <- rbind(logLik(logit\_S), logLik(logit\_S1), logLik(logit\_S2),logLik(logit\_S3),  
 logLik(logit\_S4),logLik(logit\_S5),logLik(logit\_S6))  
Compare\_models\_logLik

## [,1]  
## [1,] -58.99557  
## [2,] -60.48695  
## [3,] -62.37025  
## [4,] -59.01299  
## [5,] -64.26501  
## [6,] -63.69207  
## [7,] -62.39551

# Compare the Chi-Square  
Compare\_models\_Chisq <- rbind(with(logit\_S, null.deviance - deviance),  
 with(logit\_S1, null.deviance - deviance),  
 with(logit\_S2, null.deviance - deviance),  
 with(logit\_S3, null.deviance - deviance),  
 with(logit\_S4, null.deviance - deviance),  
 with(logit\_S5, null.deviance - deviance),  
 with(logit\_S6, null.deviance - deviance)  
 )  
Compare\_models\_Chisq

## [,1]  
## [1,] 16.611193  
## [2,] 13.628437  
## [3,] 9.861829  
## [4,] 16.576352  
## [5,] 6.072305  
## [6,] 7.218184  
## [7,] 9.811320

# We will go with Model1, that is logit\_S which has high Chi- sq value of 16.61.4 degress of freedom and P-value of 0.00085 and fits better than other 6 models.  
################### Visualize the Model performance #########################  
# Let us plot the ROC curve and calculate the AUC for performance measurements  
# Note : The Roc is curve generated by plotting the true positive rate(TPR) against the false   
# positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve.  
# A model with good predictive ability should have AUC close to 1 (ideal)  
library(ROCR)

## Warning: package 'ROCR' was built under R version 3.3.2

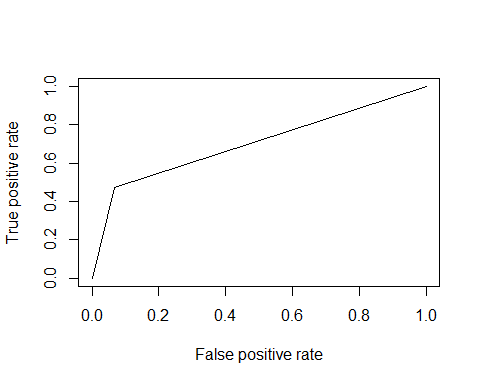
## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.3.2

##   
## Attaching package: 'gplots'

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

perf <- prediction(Sim\_data$PredPurchase, Sim\_data$Purchase)  
PerfMeas <- performance(perf, measure = "tpr", x.measure = "fpr")  
plot(PerfMeas)



#Area under the curve  
AUC <- performance(perf, measure = "auc")   
AUC <- AUC@y.values[[1]]  
print(paste('Area Under the Curve of Model is : ', AUC))

## [1] "Area Under the Curve of Model is : 0.704166666666667"

#### The model has an accuracy of 75 % and area under the curve is around 70%.  
#########################Further Insights from case study ##################  
# The odds of making a $200 purchase for customers who spend $2000 annually  
# and have a Simmons credit card   
Simon21=data.frame(Spending=2, Card=1, SpendCat=1)   
PredProb21=predict(logit\_S, Simon21, type='response')   
PredProb21

## 1   
## 0.6776332

Odds21=PredProb21/(1-PredProb21)   
Odds21

## 1   
## 2.102057

#The odds of making a $200 purchase for customers who spend $2000 annually  
# but do not have a Simmons credit card.  
Simon20=data.frame(Spending=2, Card=0, SpendCat=1)   
PredProb20=predict(logit\_S, Simon20, type='response')   
PredProb20

## 1   
## 0.4001296

Odds20=PredProb20/(1-PredProb20)   
Odds20

## 1   
## 0.6670266

## Comparison   
Odds=Odds21/Odds20   
Odds

## 1   
## 3.151383

# Clearly, Simmons credit card holders are top spenders annually.