Prasanta Panja

Target Marketing of Simmons Catalogue

Prasanta Panja

Table of Contents

[Logistic Regression Problem Statement 4](#_Toc490166862)

[Solution 4](#_Toc490166863)

[Observations: 4](#_Toc490166864)

[Interpretation of Results 5](#_Toc490166865)

[Information Value of the predictor variables 5](#_Toc490166866)

[Linear Regression Model Accuracy 6](#_Toc490166867)

[Accuracy of the model 7](#_Toc490166868)

[ROC Curve 8](#_Toc490166869)

[Area under Curve 8](#_Toc490166870)

[C-Statistic 9](#_Toc490166871)

[KS Statistics 9](#_Toc490166872)

[Conclusion 9](#_Toc490166873)

# Logistic Regression Problem Statement

Simmons’ catalogs are expensive and Simmons would like to send them to only those customers who have the highest probability of making a $200 purchase using the discount coupon included in the catalog. Simmons’ management thinks that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon to make a $200 purchase.

Simmons conducted a study by sending out 100 catalogs, 50 to customers who have a Simmons credit card and 50 to customers who do not have the card. At the end of the test period, Simmons noted for each of the 100 customers: 1) the amount the customer spent last year at Simmons, 2) whether the customer had a Simmons credit card, and 3) whether the customer made a $200 purchase. The data file that contains the information is in Logit-Simmons.xls Develop a logistic regression model, obtain the output and interpret the results

# Solution

## Hypothesis:

H0: β1 = β2 = 0 (Spending and Card will not able to explain Purchase)

Ha: β1 ≠ 0, β2 ≠ 0 (Spending and Card will be able to explain Purchase)

## Observations:

There are 100 observations with the following variables:

1. Customer - Customer ID
2. Spending = Amount customer spent last year at Simmons
3. Card = whether the customer had Simmons Credit Card
4. Purchase = If the customer has made a $200 purchase
5. SpendCat= Status of the Simmons catalog was sent based on the purchase value

**We have to predict those customers who have the highest probability of making a $200 purchase using the discount coupon included in the catalog. We will use Logistic Regression to solve the problem.**

R-code:

#Logistic Regression for Simmons customer prediction

library(caret)

library(ggplot2)

library(Information)

library(caTools)

library(stringr)

library(car)

library(ROCR)

library(MASS)

library(gmodels)

library(dummies)

library(Hmisc)

# Import Data

setwd("E:/Personal/PGPBABI/Assignments/PM/GA")

library(readr)

Simmons\_Cust\_data <- read.csv("Logit-Simmons.csv")

View(Simmons\_Cust\_data)

str(Simmons\_Cust\_data)

summary(Simmons\_Cust\_data)

#training data

set.seed(220)

split <- sample.split(Simmons\_Cust\_data, SplitRatio = 0.8)

#Information value to determine strong predictor variables

IV\_test <- create\_infotables(Simmons\_Cust\_data,y="Purchase",ncore = 2)

IV\_test$Tables

IV\_test$Summary

# Split data set into Training and Test Data set

train<- subset(simmons, split == TRUE)

dim(train)

test<- subset(simmons, split == FALSE)

dim(test)

# logistic Regression model for Predicting customer who is likely to spend 200$ worth of items of more

model1 <- glm(Purchase~Spending+Card, data=train, family=binomial)

summary(model1)

# Stepwise selection of variables

best\_model = step(model1,direction = "both",data=train,family=binomial)

summary(best\_model)

vif(model1)

predTrain <- predict(model1, type = "response") # in-sample accuracy

table(train$Purchase, predTrain >= 0.5)

# Interpretation of Results

## Information Value of the predictor variables

> IV\_test$Tables

$Customer

Customer N Percent WOE IV

1 [1,9] 9 0.09 -0.8472979 0.05648652

2 [10,19] 10 0.10 -0.9808293 0.13822230

3 [20,29] 10 0.10 0.0000000 0.13822230

4 [30,39] 10 0.10 -0.4418328 0.15663199

5 [40,49] 10 0.10 -1.7917595 0.38060193

6 [50,59] 10 0.10 0.8109302 0.44817944

7 [60,69] 10 0.10 0.4054651 0.46507382

8 [70,79] 10 0.10 1.2527630 0.62166920

9 [80,89] 10 0.10 0.8109302 0.68924671

10 [90,100] 11 0.11 -0.1541507 0.69181589

$Spending

Spending N Percent WOE IV

1 [1.06,1.39] 9 0.09 -0.8472979 0.05648652

2 [1.4,1.88] 9 0.09 -1.6739764 0.23783397

3 [1.91,2.12] 10 0.10 -0.4418328 0.25624367

4 [2.13,2.31] 10 0.10 0.0000000 0.25624367

5 [2.32,2.63] 10 0.10 -0.9808293 0.33797944

6 [2.68,3.25] 11 0.11 -0.1541507 0.34054862

7 [3.32,3.57] 10 0.10 0.8109302 0.40812614

8 [3.92,4.96] 10 0.10 0.8109302 0.47570365

9 [5,6.07] 9 0.09 1.0986123 0.58556488

10 [6.18,7.08] 12 0.12 0.4054651 0.60583814

$Card

Card N Percent WOE IV

1 [0,0] 50 0.5 -0.5389965 0.1347491

2 [1,1] 50 0.5 0.4855078 0.2561261

$SpendCat

SpendCat N Percent WOE IV

1 [0,0] 54 0.54 -0.6443570 0.2040464

2 [1,1] 46 0.46 0.6678294 0.4155257

> IV\_test$Summary

Variable IV

1 Customer 0.6918159

2 Spending 0.6058381

4 SpendCat 0.4155257

3 Card 0.2561261

From the table shown above, excluding customer ID, we see that SpendCat is strong predictor and Card is medium predictor. Spending is suspicious or too good to be true. As the attribute " SpendCat" mentioned in the data set is the status of Simmons Catalog sent based on the purchase value, we’ll not consider this as a dependent variable. So, we’ll consider Purchase as the target and Card, Spending as independent variables.

## Logistic Regression Model Accuracy

Multi-collinearity check:

> vif(model1)

Spending Card

1.011865 1.011865

**As VIF value is below 2, we see that the two variables are not collinear.**

> summary(model1)

Call:

glm(formula = Purchase ~ Spending + Card, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7786 -0.9825 -0.6808 1.0685 1.7903

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.9617 0.6161 -3.184 0.00145 \*\*

Spending 0.3513 0.1492 2.355 0.01852 \*

Card 1.0632 0.4881 2.178 0.02938 \*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 109.097 on 79 degrees of freedom

Residual deviance: 98.251 on 77 degrees of freedom

AIC: 104.25

Number of Fisher Scoring iterations: 4

**From the above data it can be seen that the AIC is 104.25 . We evaluate the best model by step wise selection of variables.**

> best\_model <- step(model1,direction = "both", data=train,family=binomial)

Start: AIC=104.25

Purchase ~ Spending + Card

Df Deviance AIC

<none> 98.251 104.25

- Card 1 103.186 107.19

- Spending 1 104.230 108.23

> summary(best\_model)

Call:

glm(formula = Purchase ~ Spending + Card, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7786 -0.9825 -0.6808 1.0685 1.7903

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.9617 0.6161 -3.184 0.00145 \*\*

Spending 0.3513 0.1492 2.355 0.01852 \*

Card 1.0632 0.4881 2.178 0.02938 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 109.097 on 79 degrees of freedom

Residual deviance: 98.251 on 77 degrees of freedom

AIC: 104.25

Number of Fisher Scoring iterations: 4

**From the above model –** *model1***, based on p-value we reject the NULL hypothesis hence we conclude that both Spending and Card is statistically significant. So, making a $200 purchase probability is dependent on both Spending and Card. We continue our analysis with** *model1.*

### 

### Accuracy of the model

> predTrain <- predict(model1, type = "response")

> table(train$Purchase, predTrain >= 0.5)

FALSE TRUE

0 37 9

1 16 18

**So, Accuracy percentage for Training data set is ((37+18)/80)\*100 = 68.75%.**

> predTest <- predict(model1, newdata = test, type = "response")

> table(test$Purchase, predTest >= 0.5)

FALSE TRUE

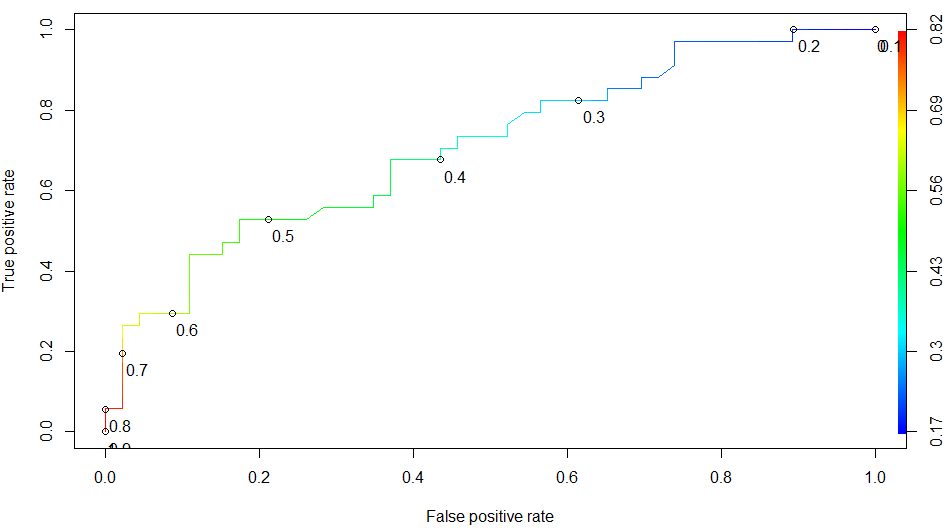
0 10 4

1 1 5

**So, Accuracy percentage of test data is ((10+5)/20)\*100 = 75%.**

### ROC Curve

> plot(ROCRperf, colorize = TRUE, print.cutoffs.at = seq(0,1,0.1), text.adj = c(-0.2, 1.7))



**ROC curve above shows the tradeoff between sensitivity and specificity.The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.Here we are not getting the good curve .**

### Area under Curve

> auc.tmp <- performance(ROCRpred,"auc");

> auc <- as.numeric(auc.tmp@y.values)

> auc

[1] 0.7081202

**Area under curve is fair.**

### C-Statistic

> library(Hmisc)

> train$predicted\_prob = predict(model1, type = "response")

> rcorr.cens(train$predicted\_prob,train$Purchase)



> test$predicted\_prob = predict(model1, newdata = test,type = "response")

> rcorr.cens(test$predicted\_prob,test$Purchase)



**The C statistic for the test model is 0.827 which can be considered as good.**

### KS Statistics

> model\_score <- prediction(train$predicted\_prob,train$Purchase)

> model\_perf <- performance(model\_score, "tpr", "fpr")

> ks\_table <- attr(model\_perf, "y.values")[[1]] - (attr(model\_perf, "x.values")[[1]])

> ks = max(ks\_table)

> which(ks\_table == ks)

[1] 26

> ks

[1] 0. 3554987

> model\_score\_test <- prediction(test$predicted\_prob,test$Purchase)

> model\_perf\_test <- performance(model\_score\_test, "tpr", "fpr")

> ks\_table\_test <- attr(model\_perf\_test, "y.values")[[1]] - (attr(model\_perf\_test, "x.values")[[1]])

> ks1=max(ks\_table\_test)

> which(ks\_table\_test == ks1)

[1] 7

> ks1

[1] 0. 6904762

**KS value seems to be good.**

## 

## Conclusion

1. Accuracy of the Model is 75% in case of test dataset.
2. Explanation of the model using odds ratio:

**Simmons estimated logistic regression equation:**

**y-hat** = **e-1.9617+.3513\*Spending+1.0632\*Card**

**1 +e-1.9617+.3513\*Spending+1.0632\*Card**

**Using the estimated logistic regression equation:**

**For customers that spend $2000 annually and do not have a Simmons credit card:**

**y-hat** = **e-1.9617+.3513\*(2)+1.0632\*(0)** = **0.2839** = **.2211**

**1 +e-1.9617+.3513\*(2)+1.0632\*(0) 1+ 0.2839**

**For customers that spend $2000 annually and do have a Simmons credit card:**

**y-hat** = **e-1.9617+.3513\*(2)+1.0632\*(1)** = **0.8221** = **.4512**

**1 +e-1.9617+.3513\*(2)+1.0632\*(1) 1+ 0.8221**

**Comparing Odds:**

Suppose we want to compare the odds of making a $200 purchase for customers who spend $2000 annually and have a Simmons credit card to the odds of making a $200 purchase for customers who spend $2000 annually and do not have a Simmons credit card.

**Estimate of odds1** = .4512 = **.8221**

1-.4512

**Estimate of odds0** = .2211 = **.2839**

1-.2211

**Estimate of odds ratio** = .8221 = **2.9**

.2839

**Thus we can conclude that the estimated odds in favor of using the coupon for customers who spent $2000 last year and have a Simmons credit card are around 3 times greater than the estimated odds in favor of using the coupon for customers who spent $2000 last year and do not have a Simmons credit card .**

**So, those customers who has more spending and a Simmons card has the highest probability of making a $200 purchase using the discount coupon included in the catalog.**

**Final R-Code**

#Logistic Regression for Simmons customer prediction

library(caret)

library(ggplot2)

library(Information)

library(caTools)

library(stringr)

library(car)

library(ROCR)

library(MASS)

library(gmodels)

library(dummies)

library(Hmisc)

# Import Data

setwd("E:/Personal/PGPBABI/Assignments/PM/GA")

library(readr)

Simmons\_Cust\_data <- read.csv("Logit-Simmons.csv")

View(Simmons\_Cust\_data)

str(Simmons\_Cust\_data)

describe(Simmons\_Cust\_data)

#training data

set.seed(220)

split <- sample.split(Simmons\_Cust\_data, SplitRatio = 0.8)

#Information value to determine strong predictor variables

IV\_test <- create\_infotables(Simmons\_Cust\_data[,-1],y="Purchase",ncore = 2)

IV\_test$Tables

IV\_test$Summary

# Split data set into Training and Test Data set

train<- subset(simmons, split == TRUE)

dim(train)

test<- subset(simmons, split == FALSE)

dim(test)

# logistic Regression model for Predicting customer who is likely to spend 200$ worth of items of more

model1 <- glm(Purchase~Spending+Card, data=train, family=binomial)

summary(model1)

Vif(model1)

# Stepwise selection of variables

best\_model <- step(model1,direction = "both", data=train,family=binomial)

summary(best\_model)

predTrain <- predict(model1, type = "response") # in-sample accuracy

table(train$Purchase, predTrain >= 0.5)

predTest <- predict(model1, newdata = test, type = "response") # out-sample accuracy

table(test$Purchase, predTest >= 0.5)

# Model evaluation

#1 ROC Curve

ROCRpred <- prediction(predTrain, train$Purchase)

ROCRperf <- performance(ROCRpred, "tpr", "fpr")

plot(ROCRperf, colorize = TRUE, print.cutoffs.at = seq(0,1,0.1), text.adj = c(-0.2, 1.7))

auc.tmp <- performance(ROCRpred,"auc");

auc <- as.numeric(auc.tmp@y.values)

auc

#2 C-statistic

train$predicted\_prob = predict(model1, type = "response")

rcorr.cens(train$predicted\_prob,train$Purchase)

test$predicted\_prob = predict(model1, newdata = test,type = "response")

rcorr.cens(test$predicted\_prob,test$Purchase)

#KS-statistic

model\_score <- prediction(train$predicted\_prob,train$Purchase)

model\_perf <- performance(model\_score, "tpr", "fpr")

ks\_table <- attr(model\_perf, "y.values")[[1]] - (attr(model\_perf, "x.values")[[1]])

ks = max(ks\_table)

which(ks\_table == ks)

ks

model\_score\_test <- prediction(test$predicted\_prob,test$Purchase)

model\_perf\_test <- performance(model\_score\_test, "tpr", "fpr")

ks\_table\_test <- attr(model\_perf\_test, "y.values")[[1]] - (attr(model\_perf\_test, "x.values")[[1]])

ks1=max(ks\_table\_test)

which(ks\_table\_test == ks1)

ks1