Determining Predictors of a Portuguese Bank Marketing Campaign using Binary Logistic Regression

Nkosi and Teddy

December 10, 2019

Table of Contents

[Abstract 1](#_Toc27237635)

[Data Characteristics 2](#_Toc27237636)

[Exploratory Analysis 2](#_Toc27237637)

[First Order logistic regression model 9](#_Toc27237638)

[Model Selection 13](#_Toc27237639)

[Stepwise Regression on the first order model 13](#_Toc27237640)

[Fit a model with interactions 17](#_Toc27237641)

[Stepwise Regression 19](#_Toc27237642)

[Model Diagnostics 26](#_Toc27237643)

[Residual Plots 27](#_Toc27237644)

[Conclusion 28](#_Toc27237645)

[ROC Curve 28](#_Toc27237646)

# Abstract

The data being surved came from Kaggle.com under a dataset called Bank Marketing. It is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on number of phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (‘yes’) or not (‘no’) subscribed. The dataset included all types of employees so for this study,it was narrowed to just 875 students and 1720 retired employees. A selection of 5 continous variables were used to try and predict if a customer would subscribe a term deposit.The variables are:

Response Variable: y (term deposit) which is coded as 1 for yes and 0 for no

Predictor Variables: Age, Campaign (number of contacts performed during this campaign and for this client), Previous (number of contacts performed before this campaign and for this client), CCI(Consumer Confidence Index), NE(Number of Employees at the bank)

The goal is to predict if the client will subscribe a term deposit (variable y). Also, we would like to find an optimal logistic regression model using these 5 predictor variables. The binary logistic regression model could then be utilized for helping the bank determine which sectors are doing well so that they could target their marketing campaigns to those industries.

# Reading in the data  
library(readxl)  
bm <- read\_excel("C:/Users/Nkosilenhle/Desktop/Project 2/BM.xlsx")

## New names:  
## \* `` -> ...1

attach(bm)  
  
ftable(y, campaign)

## campaign 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 27 30 42  
## y   
## 0 853 458 229 142 66 37 20 23 12 7 6 5 3 1 1 4 4 2 2 2 1 2 2 2 1 1  
## 1 369 191 82 38 10 9 5 2 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

ftable(y, previous)

## previous 0 1 2 3 4 5 6  
## y   
## 0 1524 255 78 18 10 1 0  
## 1 402 173 83 36 11 3 1

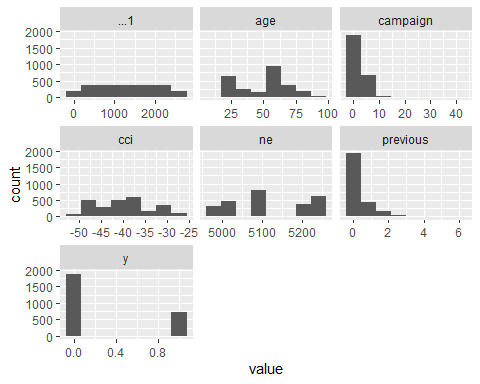
# Data Characteristics

The first frequency table shows that each customer was at least contacted once or more during this campaign. During the campaign, customers who were contacted the most see to have made a deposit during the campaign. The second frequency table indicates that each customer in the data table was contacted before the campaign started. Even before the marketing campaign started, a large majority of customers were already making deposits to the bank.

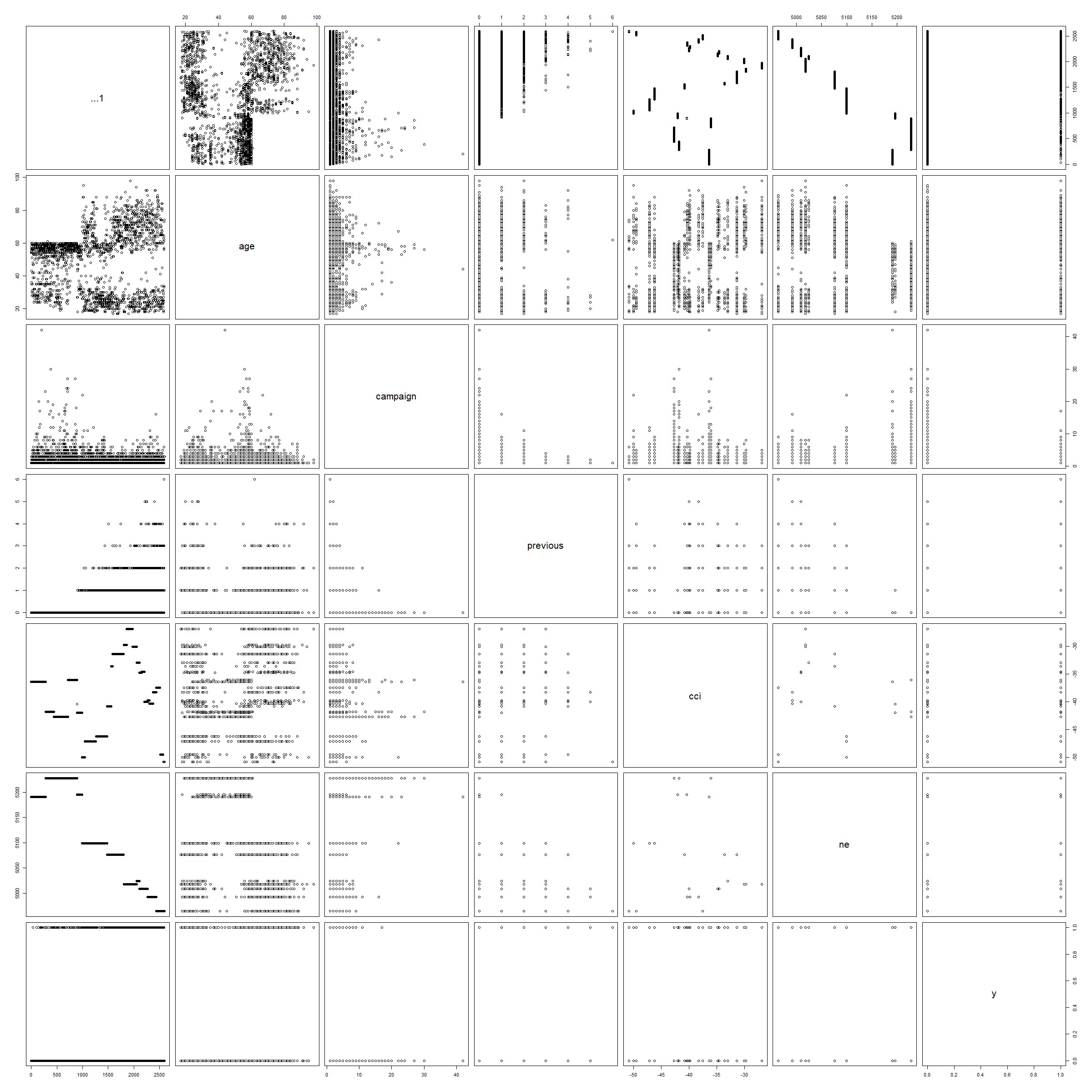
# Exploratory Analysis

Check distributions of the predictor variables

library (tidyr)  
library (ggplot2)  
ggplot(gather(bm), aes(value)) +   
 geom\_histogram(bins = 8) +   
 facet\_wrap(~key, scales = 'free\_x')



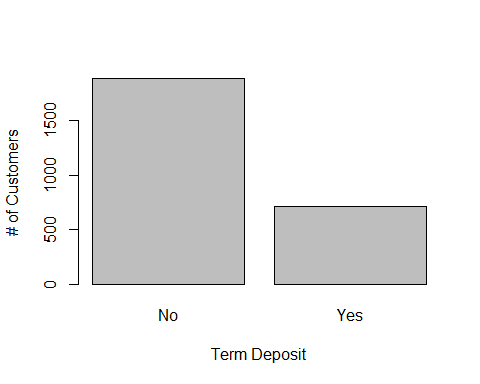
plot(bm)



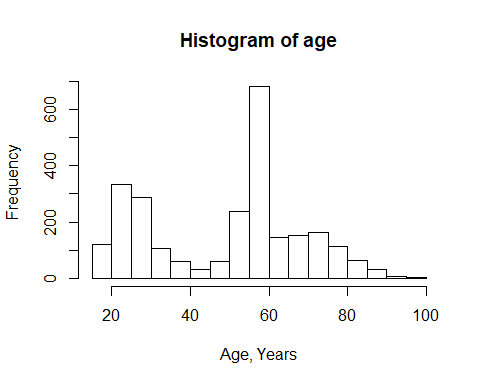
None of our continous predictor variables require a transformation.

We can plot the individual variables below:

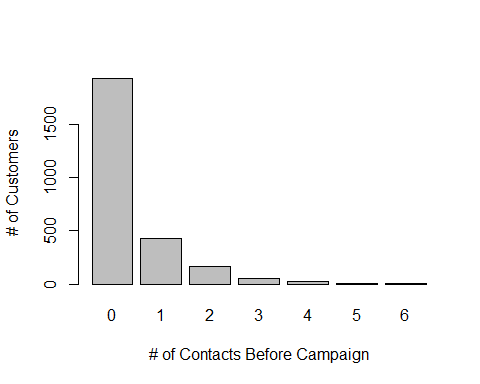
barplot(table (ifelse (y==1,"Yes","No")), xlab="Term Deposit", ylab="# of Customers")



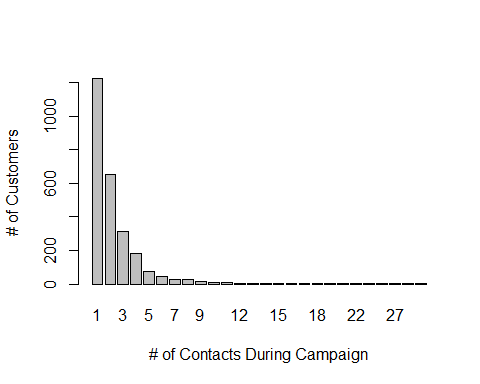
hist (age, xlab="Age, Years")



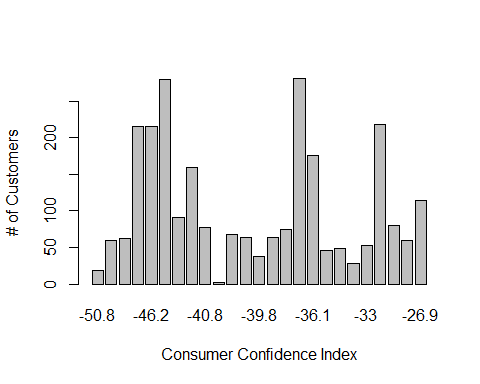
barplot(table (previous), xlab="# of Contacts Before Campaign", ylab= "# of Customers")



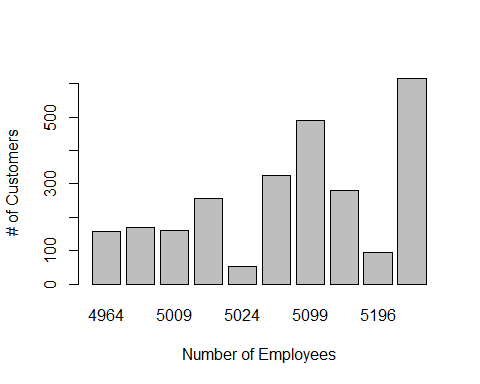
barplot(table (campaign), xlab="# of Contacts During Campaign", ylab= "# of Customers")



barplot(table (cci), xlab="Consumer Confidence Index", ylab= "# of Customers")



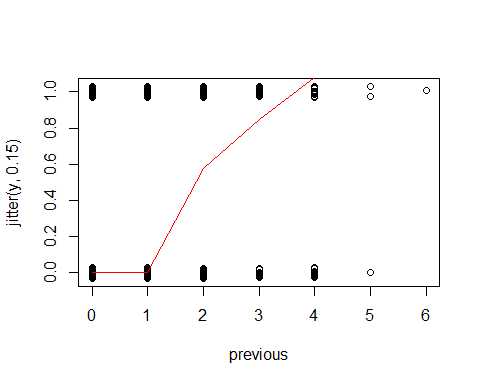
barplot(table (ne), xlab="Number of Employees", ylab= "# of Customers")



summary(previous)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.3938 1.0000 6.0000

plot(previous, jitter(y,0.15))  
lines(lowess(y~previous),col="red")



# First Order logistic regression model

A first-order binary logistic regression was fit using all 5 predictor variables.

# Fitting the regression model  
y.logit <- glm(y ~ age + campaign + previous + cci + ne, family=binomial)  
summary(y.logit)

##   
## Call:  
## glm(formula = y ~ age + campaign + previous + cci + ne, family = binomial)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6138 -0.7803 -0.4064 0.9802 2.7156   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 54.4368411 3.5138921 15.492 < 2e-16 \*\*\*  
## age 0.0030542 0.0023182 1.317 0.18768   
## campaign -0.0744243 0.0305305 -2.438 0.01478 \*   
## previous 0.1617345 0.0595722 2.715 0.00663 \*\*   
## cci 0.0028116 0.0074582 0.377 0.70619   
## ne -0.0108825 0.0006974 -15.603 < 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: 3043.6 on 2594 degrees of freedom  
## Residual deviance: 2556.1 on 2589 degrees of freedom  
## AIC: 2568.1  
##   
## Number of Fisher Scoring iterations: 5

exp (y.logit$coefficients)

## (Intercept) age campaign previous cci   
## 4.381469e+23 1.003059e+00 9.282777e-01 1.175548e+00 1.002816e+00   
## ne   
## 9.891765e-01

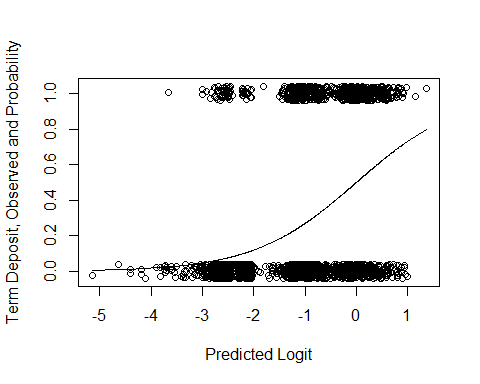
exp (confint (y.logit))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 4.875181e+20 4.712487e+26  
## age 9.985174e-01 1.007635e+00  
## campaign 8.720615e-01 9.826160e-01  
## previous 1.046633e+00 1.322228e+00  
## cci 9.882636e-01 1.017594e+00  
## ne 9.878063e-01 9.905121e-01

Campaign (number of contacts performed during this campaign and for this client), Previous (number of contacts performed before this campaign and for this client) and NE(Number of Employees) are significant predictors in the initial model. However, age of the customer and CCI(Consumer Confidence Index) are not significant predictor variables in the first order model. The probability of a customer having subscribed a term deposit is higher with customers who have been contacted during the marketing campaign, customers who were contacted before the campaign started and the number of employees at the bank at that time. Looking at it, the odds of subscribing a term deposit is 0.93 times higher for each extra phone call during the campaign, 1.1 times higher for each extra phone call before the campaign and 0.99 times higher with each additional employee at the bank. The 95% confidence interval for campaign is 0.87 to 0.98% higher odds of a term deposit per additional phone call during the campaign.The 95% confidence interval for previous is 1.05 to 1.32% higher odds of a term deposit per additional phone call before the campaign. The 95% confidence interval for number of employees at the bank is 0.9787 to 0.9905% higher odds of a term deposit per additional employee added to the bank.

# plot the model with the data.  
predpr = predict (y.logit, type='response')  
predlogit = predict (y.logit)  
plot (jitter (y, 0.2) ~ predlogit, xlab="Predicted Logit",  
 ylab="Term Deposit, Observed and Probability")  
pred.ord = order (predlogit)  
lines (predlogit[pred.ord], predpr[pred.ord])



# ROC curve - install package ROCR  
par (mfrow=c(1,1))  
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

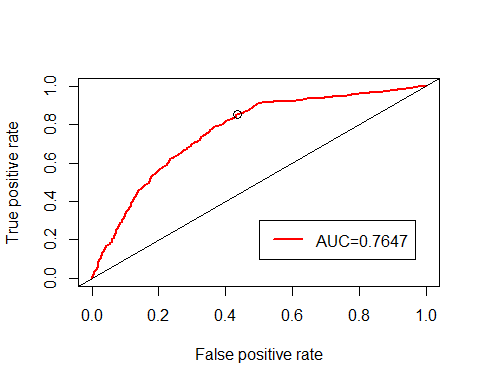
pred1 <- prediction(y.logit$fitted.values, bm$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]  
auc1

## [1] 0.7646567

plot(perf1, lwd=2, col=2)  
abline(0,1)  
legend(0.5, 0.3, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)  
  
# Extract the X and Y values from the ROC plot, as well as the probability cutoffs  
roc.x = slot (perf1, "x.values") [[1]]  
roc.y = slot (perf1, "y.values") [[1]]  
cutoffs = slot (perf1, "alpha.values") [[1]]  
  
auc.table = cbind.data.frame(cutoff=pred1@cutoffs, tp=pred1@tp, fp=pred1@fp, tn=pred1@tn, fn=pred1@fn)  
names (auc.table) = c("Cutoff", "TP", "FP", "TN", "FN")  
auc.table$sensitivity = auc.table$TP / (auc.table$TP + auc.table$FN)  
auc.table$specificity = auc.table$TN / (auc.table$TN + auc.table$FP)  
auc.table$FalsePosRate = 1 - auc.table$specificity  
auc.table$sens\_spec = auc.table$sensitivity + auc.table$specificity  
  
# Find the row(s) in the AUC table where sensitivity + specificity is maximized  
  
auc.best = auc.table [auc.table$sens\_spec == max (auc.table$sens\_spec),]  
auc.best

## Cutoff TP FP TN FN sensitivity specificity FalsePosRate  
## 1036 0.2297128 607 818 1068 102 0.8561354 0.5662778 0.4337222  
## sens\_spec  
## 1036 1.422413

# Plot the maximum point(s) on the ROC plot  
  
points (auc.best$FalsePosRate, auc.best$sensitivity, cex=1.3)



# Model Selection

## Stepwise Regression on the first order model

Following the plots, we apply step-wise regression to the first order model

m2 = step (y.logit, direction='both')

## Start: AIC=2568.13  
## y ~ age + campaign + previous + cci + ne  
##   
## Df Deviance AIC  
## - cci 1 2556.3 2566.3  
## - age 1 2557.9 2567.9  
## <none> 2556.1 2568.1  
## - campaign 1 2562.9 2572.9  
## - previous 1 2563.6 2573.6  
## - ne 1 2842.7 2852.7  
##   
## Step: AIC=2566.27  
## y ~ age + campaign + previous + ne  
##   
## Df Deviance AIC  
## - age 1 2558.2 2566.2  
## <none> 2556.3 2566.3  
## + cci 1 2556.1 2568.1  
## - campaign 1 2563.3 2571.3  
## - previous 1 2563.7 2571.7  
## - ne 1 2854.8 2862.8  
##   
## Step: AIC=2566.24  
## y ~ campaign + previous + ne  
##   
## Df Deviance AIC  
## <none> 2558.2 2566.2  
## + age 1 2556.3 2566.3  
## + cci 1 2557.9 2567.9  
## - campaign 1 2565.2 2571.2  
## - previous 1 2565.2 2571.2  
## - ne 1 2863.0 2869.0

summary (m2)

##   
## Call:  
## glm(formula = y ~ campaign + previous + ne, family = binomial)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6339 -0.7912 -0.3965 0.9789 2.7296   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 55.1817026 3.4826198 15.845 < 2e-16 \*\*\*  
## campaign -0.0748247 0.0304639 -2.456 0.01404 \*   
## previous 0.1559798 0.0594312 2.625 0.00868 \*\*   
## ne -0.0110196 0.0006854 -16.078 < 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: 3043.6 on 2594 degrees of freedom  
## Residual deviance: 2558.2 on 2591 degrees of freedom  
## AIC: 2566.2  
##   
## Number of Fisher Scoring iterations: 5

beta2 = coefficients(m2)  
exp (beta2)

## (Intercept) campaign previous ne   
## 9.228029e+23 9.279062e-01 1.168803e+00 9.890409e-01

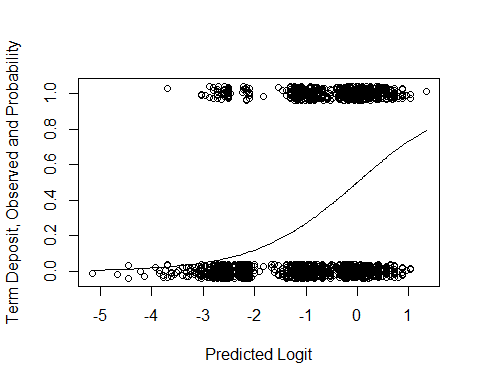
exp (confint (m2))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 1.088435e+21 9.300676e+26  
## campaign 8.718492e-01 9.821312e-01  
## previous 1.040923e+00 1.314284e+00  
## ne 9.876952e-01 9.903537e-01

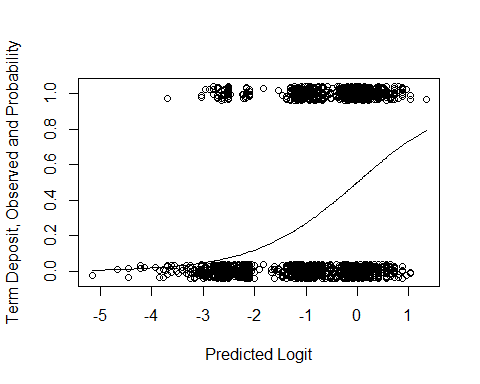
Age and Consumer COnfidence Index variables were removed from the model. The model with only 3 predictor variables has a lower AIC value which could mean that it is a better model. The significance of campaign, previous and number of employees seem closely related to the first order model.The odds of a term deposit is 0.87 to 0.98% higher per additional phone call during the campaign.The odds of a term deposit is 1.04 to 1.32% higher per additional phone call before the campaign. The odds of a term deposit is 0.9877 to 0.9035% higher per additional employee added to the bank.

predpr = predict (m2, type='response')  
predlogit = predict (m2)  
plot (jitter (y, 0.2) ~ predlogit, xlab="Predicted Logit",  
 ylab="Term Deposit, Observed and Probability")  
pred.ord = order (predlogit)  
lines (predlogit[pred.ord], predpr[pred.ord])



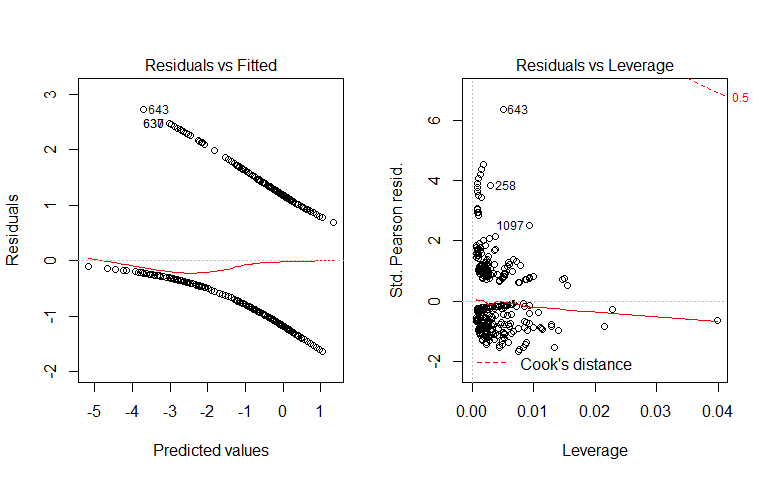
We can also show a general plot of response vs predicted:

predpr = predict (m2, type='response')  
predlogit = predict (m2)  
plot (jitter (y, 0.2) ~ predlogit, xlab="Predicted Logit",  
 ylab="Term Deposit, Observed and Probability")  
pred.ord = order (predlogit)  
lines (predlogit[pred.ord], predpr[pred.ord])



C

par (mfrow=c(1,2))  
plot (m2, which=c(1,5))



Residual plots of m2 above seem to indicate that the model is relatively reasonable. With the first plot of Residuals vs Fitted, the fitted lowess line is close to 0 and relatively flat. With the second plot of Residuals vs Leverage, there is not a clear pattern in the data or many outliers in the data. The Cook’s distance contours appear a little on the plot which indicates that there could be a few large values for the influence measure.

# Fit a model with interactions

We fit a full model that consisted of all two-way interaction effects.

bm$campaign.c = campaign - mean(campaign)  
bm$previous.c = previous - mean(previous)  
bm$ne.c = ne - mean(ne)  
bm$age.c = age - mean(age)  
bm$cci.c = cci - mean(cci)  
m3 = glm (y ~ (age.c + campaign.c + previous.c + cci.c + ne.c)^2, family=binomial, data=bm)  
summary (m3)

##   
## Call:  
## glm(formula = y ~ (age.c + campaign.c + previous.c + cci.c +   
## ne.c)^2, family = binomial, data = bm)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8585 -0.7754 -0.4086 0.9049 2.5086   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.374e+00 7.349e-02 -18.692 < 2e-16 \*\*\*  
## age.c 4.745e-03 3.528e-03 1.345 0.17866   
## campaign.c -8.627e-02 2.969e-02 -2.906 0.00366 \*\*   
## previous.c 1.018e-02 1.421e-01 0.072 0.94290   
## cci.c -2.853e-02 1.172e-02 -2.434 0.01492 \*   
## ne.c -1.197e-02 8.049e-04 -14.866 < 2e-16 \*\*\*  
## age.c:campaign.c -1.553e-04 1.613e-03 -0.096 0.92331   
## age.c:previous.c 2.119e-03 2.576e-03 0.822 0.41080   
## age.c:cci.c -6.489e-04 3.358e-04 -1.932 0.05335 .   
## age.c:ne.c 2.651e-05 4.378e-05 0.605 0.54489   
## campaign.c:previous.c -4.408e-02 5.440e-02 -0.810 0.41783   
## campaign.c:cci.c 6.313e-04 5.315e-03 0.119 0.90544   
## campaign.c:ne.c 5.129e-04 3.379e-04 1.518 0.12898   
## previous.c:cci.c 2.622e-02 9.400e-03 2.789 0.00528 \*\*   
## previous.c:ne.c -1.261e-03 1.279e-03 -0.986 0.32407   
## cci.c:ne.c -4.915e-04 1.514e-04 -3.247 0.00117 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3043.6 on 2594 degrees of freedom  
## Residual deviance: 2520.6 on 2579 degrees of freedom  
## AIC: 2552.6  
##   
## Number of Fisher Scoring iterations: 5

exp (m3$coefficients)

## (Intercept) age.c campaign.c   
## 0.2531460 1.0047561 0.9173423   
## previous.c cci.c ne.c   
## 1.0102287 0.9718687 0.9881053   
## age.c:campaign.c age.c:previous.c age.c:cci.c   
## 0.9998447 1.0021214 0.9993514   
## age.c:ne.c campaign.c:previous.c campaign.c:cci.c   
## 1.0000265 0.9568811 1.0006315   
## campaign.c:ne.c previous.c:cci.c previous.c:ne.c   
## 1.0005131 1.0265659 0.9987394   
## cci.c:ne.c   
## 0.9995086

exp (confint (m3))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.2186663 0.2917368  
## age.c 0.9979015 1.0118077  
## campaign.c 0.8632509 0.9698363  
## previous.c 0.7593338 1.3270805  
## cci.c 0.9496239 0.9943113  
## ne.c 0.9865279 0.9896472  
## age.c:campaign.c 0.9967214 1.0030612  
## age.c:previous.c 0.9970861 1.0072217  
## age.c:cci.c 0.9986928 1.0000092  
## age.c:ne.c 0.9999415 1.0001132  
## campaign.c:previous.c 0.8587352 1.0635426  
## campaign.c:cci.c 0.9903614 1.0111602  
## campaign.c:ne.c 0.9998332 1.0011689  
## previous.c:cci.c 1.0079450 1.0458694  
## previous.c:ne.c 0.9961822 1.0012004  
## cci.c:ne.c 0.9992098 0.9998035

Centered campaign (number of phone calls done during the campaign), consumer confidence index, and number of employees are all statistically significant. The 2 interaction effects between centered previous (number of phone calls done before the campaign) and centered comsumer confidence index as well as centered consumer confidence index and centered number of employees at the bank are also significant in the model with interaction effects.

# Stepwise Regression

Following this, we ran a step-wise regression on the full interaction model.

# Stepwise with AIC criterion  
m.both = step (m3, directon='both')

## Start: AIC=2552.56  
## y ~ (age.c + campaign.c + previous.c + cci.c + ne.c)^2  
##   
## Df Deviance AIC  
## - age.c:campaign.c 1 2520.6 2550.6  
## - campaign.c:cci.c 1 2520.6 2550.6  
## - age.c:ne.c 1 2520.9 2550.9  
## - campaign.c:previous.c 1 2521.2 2551.2  
## - age.c:previous.c 1 2521.2 2551.2  
## - previous.c:ne.c 1 2521.6 2551.6  
## <none> 2520.6 2552.6  
## - campaign.c:ne.c 1 2522.8 2552.8  
## - age.c:cci.c 1 2524.3 2554.3  
## - previous.c:cci.c 1 2528.5 2558.5  
## - cci.c:ne.c 1 2531.3 2561.3  
##   
## Step: AIC=2550.57  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:previous.c +   
## age.c:cci.c + age.c:ne.c + campaign.c:previous.c + campaign.c:cci.c +   
## campaign.c:ne.c + previous.c:cci.c + previous.c:ne.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## - campaign.c:cci.c 1 2520.6 2548.6  
## - age.c:ne.c 1 2520.9 2548.9  
## - campaign.c:previous.c 1 2521.2 2549.2  
## - age.c:previous.c 1 2521.2 2549.2  
## - previous.c:ne.c 1 2521.6 2549.6  
## <none> 2520.6 2550.6  
## - campaign.c:ne.c 1 2522.8 2550.8  
## - age.c:cci.c 1 2524.3 2552.3  
## - previous.c:cci.c 1 2528.5 2556.5  
## - cci.c:ne.c 1 2531.3 2559.3  
##   
## Step: AIC=2548.58  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:previous.c +   
## age.c:cci.c + age.c:ne.c + campaign.c:previous.c + campaign.c:ne.c +   
## previous.c:cci.c + previous.c:ne.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## - age.c:ne.c 1 2520.9 2546.9  
## - campaign.c:previous.c 1 2521.2 2547.2  
## - age.c:previous.c 1 2521.3 2547.3  
## - previous.c:ne.c 1 2521.6 2547.6  
## <none> 2520.6 2548.6  
## - campaign.c:ne.c 1 2522.8 2548.8  
## - age.c:cci.c 1 2524.3 2550.3  
## - previous.c:cci.c 1 2528.5 2554.5  
## - cci.c:ne.c 1 2531.4 2557.4  
##   
## Step: AIC=2546.95  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:previous.c +   
## age.c:cci.c + campaign.c:previous.c + campaign.c:ne.c + previous.c:cci.c +   
## previous.c:ne.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## - age.c:previous.c 1 2521.4 2545.4  
## - campaign.c:previous.c 1 2521.6 2545.6  
## - previous.c:ne.c 1 2522.0 2546.0  
## <none> 2520.9 2546.9  
## - campaign.c:ne.c 1 2523.4 2547.4  
## - age.c:cci.c 1 2525.5 2549.5  
## - previous.c:cci.c 1 2528.9 2552.9  
## - cci.c:ne.c 1 2531.5 2555.5  
##   
## Step: AIC=2545.38  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:cci.c +   
## campaign.c:previous.c + campaign.c:ne.c + previous.c:cci.c +   
## previous.c:ne.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## - campaign.c:previous.c 1 2522.0 2544.0  
## - previous.c:ne.c 1 2522.5 2544.5  
## <none> 2521.4 2545.4  
## - campaign.c:ne.c 1 2523.8 2545.8  
## - age.c:cci.c 1 2525.8 2547.8  
## - previous.c:cci.c 1 2529.7 2551.7  
## - cci.c:ne.c 1 2531.9 2553.9  
##   
## Step: AIC=2543.97  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:cci.c +   
## campaign.c:ne.c + previous.c:cci.c + previous.c:ne.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## - previous.c:ne.c 1 2523.1 2543.1  
## <none> 2522.0 2544.0  
## - age.c:cci.c 1 2526.4 2546.4  
## - campaign.c:ne.c 1 2527.1 2547.1  
## - previous.c:cci.c 1 2530.4 2550.4  
## - cci.c:ne.c 1 2532.4 2552.4  
##   
## Step: AIC=2543.12  
## y ~ age.c + campaign.c + previous.c + cci.c + ne.c + age.c:cci.c +   
## campaign.c:ne.c + previous.c:cci.c + cci.c:ne.c  
##   
## Df Deviance AIC  
## <none> 2523.1 2543.1  
## - age.c:cci.c 1 2527.5 2545.5  
## - campaign.c:ne.c 1 2528.5 2546.5  
## - previous.c:cci.c 1 2530.5 2548.5  
## - cci.c:ne.c 1 2533.1 2551.1

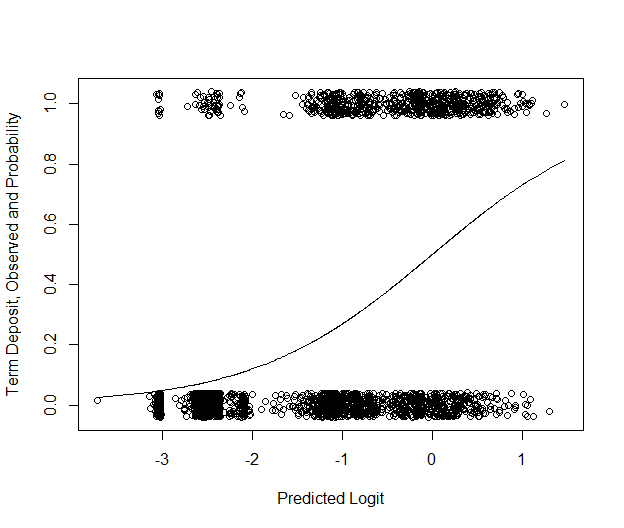
The stepwise regression with the AIC criterion retained 4 interaction effects. These were age centered and centered consumer confidence index, centered campaign(number of phone calls done during the campaign) and centered number of employees, centered previous(number of phone calls done before the campaign) and centered consumer confidence index as well as centered consumer confidence index and centered number of employees.

summary(m.both)

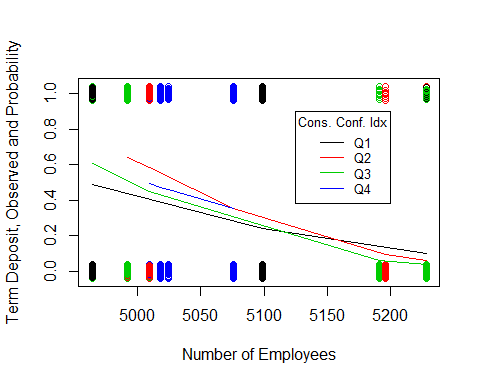
##   
## Call:  
## glm(formula = y ~ age.c + campaign.c + previous.c + cci.c + ne.c +   
## age.c:cci.c + campaign.c:ne.c + previous.c:cci.c + cci.c:ne.c,   
## family = binomial, data = bm)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7566 -0.7793 -0.4020 0.9421 2.4948   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.3355429 0.0622390 -21.458 < 2e-16 \*\*\*  
## age.c 0.0037814 0.0023288 1.624 0.10443   
## campaign.c -0.0845965 0.0279693 -3.025 0.00249 \*\*   
## previous.c 0.1572276 0.0607350 2.589 0.00963 \*\*   
## cci.c -0.0291049 0.0115570 -2.518 0.01179 \*   
## ne.c -0.0115469 0.0007371 -15.666 < 2e-16 \*\*\*  
## age.c:cci.c -0.0006721 0.0003221 -2.086 0.03695 \*   
## campaign.c:ne.c 0.0006821 0.0002818 2.420 0.01552 \*   
## previous.c:cci.c 0.0241696 0.0089428 2.703 0.00688 \*\*   
## cci.c:ne.c -0.0004694 0.0001498 -3.133 0.00173 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3043.6 on 2594 degrees of freedom  
## Residual deviance: 2523.1 on 2585 degrees of freedom  
## AIC: 2543.1  
##   
## Number of Fisher Scoring iterations: 5

All the predictor variables are statistically significant in the model except the centered age predictor variable. The above model has an AIC value of 2543.1 whereas the model previous had an AIC value of 2566.2. Therefore, the mdoel with the interation effects and centered variables is the better model as indicated by the AIC values.

predpr = predict (m.both, type='response')  
predlogit = predict (m.both)  
plot (jitter (y, 0.2) ~ predlogit, xlab="Predicted Logit",  
 ylab="Term Deposit, Observed and Probability")  
pred.ord = order (predlogit)  
lines (predlogit[pred.ord], predpr[pred.ord])

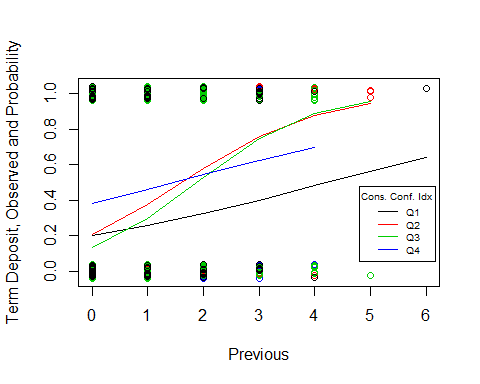


categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [which ((quartiles[1] < x) & (x <= quartiles [2]))] = "Q2"  
 result [which ((quartiles[2] < x) & (x <= quartiles [3]))] = "Q3"  
 result [which (quartiles[3] < x)] = "Q4"  
 return (result)  
}  
par (mfrow=c(1,1))  
cci.cat = categorize (cci)  
plot (jitter (y, 0.2) ~ ne, col=as.factor(cci.cat), xlab="Number of Employees",   
 ylab="Term Deposit, Observed and Probability")  
  
fit.q1 = glm (y[cci.cat=="Q1"] ~ ne[cci.cat=="Q1"], family=binomial)  
fit.q2 = glm (y[cci.cat=="Q2"] ~ ne[cci.cat=="Q2"], family=binomial)  
fit.q3 = glm (y[cci.cat=="Q3"] ~ ne[cci.cat=="Q3"], family=binomial)  
fit.q4 = glm (y[cci.cat=="Q4"] ~ ne[cci.cat=="Q4"], family=binomial)  
  
cci.ord.1 = order (ne [cci.cat=="Q1"])  
cci.ord.2 = order (ne [cci.cat=="Q2"])  
cci.ord.3 = order (ne [cci.cat=="Q3"])  
cci.ord.4 = order (ne [cci.cat=="Q4"])  
  
lines (ne[cci.cat=="Q1"][cci.ord.1], predict (fit.q1, type='response')[cci.ord.1], col=1)  
lines (ne[cci.cat=="Q2"][cci.ord.2], predict (fit.q2, type='response')[cci.ord.2], col=2)  
lines (ne[cci.cat=="Q3"][cci.ord.3], predict (fit.q3, type='response')[cci.ord.3], col=3)  
lines (ne[cci.cat=="Q4"][cci.ord.4], predict (fit.q4, type='response')[cci.ord.4], col=4)  
  
legend (5125, 0.9, list("Q1", "Q2", "Q3","Q4"), lty=rep(1, 4), col=1:4,   
 title="Cons. Conf. Idx", cex=0.8)



The above plot indicates the number of employees by consumer confidence index interaction effect. With consumer confidence index in Q1, the probablity of a term deposit decreases sigmoidally with the number of employees at the bank. With consumer confidence index in Q2, the probability of a term deposit decreases sigmoidally with the number of employees at the bank. With consumer confidence index in Q3, the probability of a term deposit decreases sigmoidally with the number of employees at the bank. With consumer confidence index in Q4, the probability of a term deposit decreases sigmoidally with the number of employees at the bank.

categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [which ((quartiles[1] < x) & (x <= quartiles [2]))] = "Q2"  
 result [which ((quartiles[2] < x) & (x <= quartiles [3]))] = "Q3"  
 result [which (quartiles[3] < x)] = "Q4"  
 return (result)  
}  
par (mfrow=c(1,1))  
cci.cat = categorize (cci)  
plot (jitter (y, 0.2) ~ previous, col=as.factor(cci.cat), xlab="Previous",   
 ylab="Term Deposit, Observed and Probability")  
  
fit.q1 = glm (y[cci.cat=="Q1"] ~ previous[cci.cat=="Q1"], family=binomial)  
fit.q2 = glm (y[cci.cat=="Q2"] ~ previous[cci.cat=="Q2"], family=binomial)  
fit.q3 = glm (y[cci.cat=="Q3"] ~ previous[cci.cat=="Q3"], family=binomial)  
fit.q4 = glm (y[cci.cat=="Q4"] ~ previous[cci.cat=="Q4"], family=binomial)  
  
cci.ord.1 = order (previous [cci.cat=="Q1"])  
cci.ord.2 = order (previous [cci.cat=="Q2"])  
cci.ord.3 = order (previous [cci.cat=="Q3"])  
cci.ord.4 = order (previous [cci.cat=="Q4"])  
  
lines (previous[cci.cat=="Q1"][cci.ord.1], predict (fit.q1, type='response')[cci.ord.1], col=1)  
lines (previous[cci.cat=="Q2"][cci.ord.2], predict (fit.q2, type='response')[cci.ord.2], col=2)  
lines (previous[cci.cat=="Q3"][cci.ord.3], predict (fit.q3, type='response')[cci.ord.3], col=3)  
lines (previous[cci.cat=="Q4"][cci.ord.4], predict (fit.q4, type='response')[cci.ord.4], col=4)  
  
legend (4.8, 0.475, list("Q1", "Q2", "Q3","Q4"), lty=rep(1, 4), col=1:4,   
 title="Cons. Conf. Idx", cex=0.65)



The above plot indicates the previous(number of contacts performed before this campaign and for this client) by consumer confidence index interaction effect. With consumer confidence index in Q2, the probablity of a term deposit increases sigmoidally with the previous phone calls made. With consumer confidence index in Q1, the probability of a term deposit increases steadily with the previous phone calls made. With consumer confidence index in Q3, the probability of a term deposit increases sigmoidally with the previous phone calls made. With consumer confidence index in Q4, the probability of a term deposit increases steadily with the previous phone calls made.

The table below shows us the odds ratios for the parameter estimates and their standard errors.

cbind.data.frame (exp.beta = exp (m.both$coefficients), exp (as.data.frame (confint(m.both))))

## Waiting for profiling to be done...

## exp.beta 2.5 % 97.5 %  
## (Intercept) 0.2630153 0.2322585 0.2964828  
## age.c 1.0037885 0.9992237 1.0083903  
## campaign.c 0.9188830 0.8679048 0.9686876  
## previous.c 1.1702619 1.0392532 1.3189718  
## cci.c 0.9713146 0.9493910 0.9934217  
## ne.c 0.9885195 0.9870697 0.9899276  
## age.c:cci.c 0.9993282 0.9986964 0.9999590  
## campaign.c:ne.c 1.0006823 1.0001109 1.0012258  
## previous.c:cci.c 1.0244640 1.0067623 1.0427861  
## cci.c:ne.c 0.9995307 0.9992349 0.9998225

We simply cannot interpret any of the effects of the predictor variables because they are all one way or another in an interaction.

# Model Diagnostics

#------------------------Deviance test of lack of fit--------------------  
  
# First model  
pchisq(deviance(y.logit), df.residual(y.logit), lower=F)

## [1] 0.6734763

# Final model  
pchisq(deviance(m.both), df.residual(m.both), lower=F)

## [1] 0.8046132

There is no significant lack of fit with both first and final model because the p value is greater than 0.05.

#Likelihood Ratio (LR) test statistic and P-value in R (multiple logistic regression):  
  
# First model  
  
1 - pchisq(y.logit$null.deviance - y.logit$deviance,   
 y.logit$df.null - y.logit$df.residual)

## [1] 0

# Final model  
  
1 - pchisq(m.both$null.deviance - m.both$deviance,   
 m.both$df.null - m.both$df.residual)

## [1] 0

The first model and the final model both have significant effects on term deposit because p is less than 0.05.

# Compare the final model with the reduced model   
  
1 - pchisq(m2$deviance - m.both$deviance,   
 m2$df.residual - m.both$df.residual)

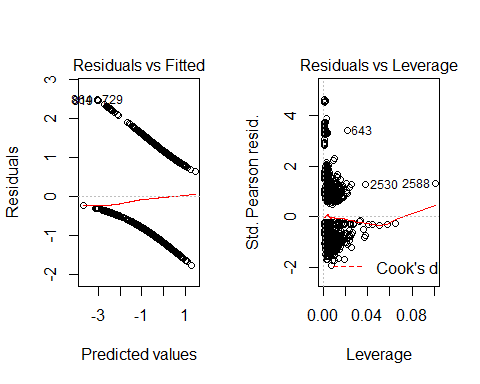
## [1] 4.09552e-06

m2 which is the reduced model is not significantly worse than the final model with the interaction effects.

# Residual Plots

Below are the residual plots of the final model.

# Residual plot  
par (mfrow=c(1,2))  
  
# The plot function applied to the fitted model object, selecting plots 1 and 5,  
# provides the necessary plots.  
plot (m.both, which = c(1,5))

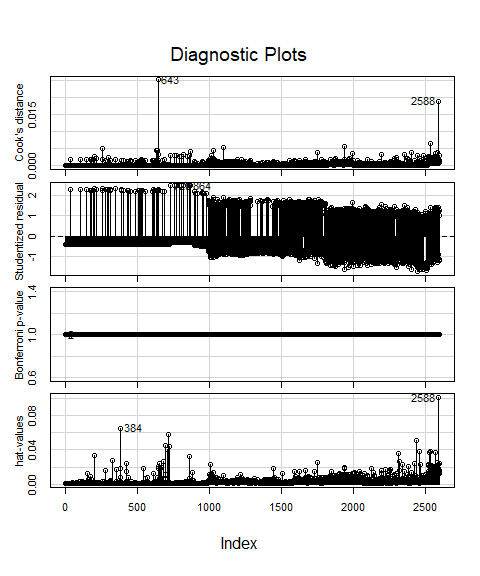


The Residuals vs Fitted graph indicates that there is a reasonable fit due to the red Lowess line being relatively flat and close to zero. The Residuals vs Leverage graph shows us that there is also a reasonable fit because Cook’s distance contour lines don’t appear on the graph. There are no high leverage values and no large standardized residuals.

# Influence diagnostics  
  
library(car)

## Loading required package: carData

influenceIndexPlot(m.both)



Row 643 and 2588 have high Cook’s distance but they are not above the 0.5 mark. Rows 384 and 2588 have the highest leverage values, but are not obvious outliers with respect to the rest of the leverage values.

# Conclusion

## ROC Curve

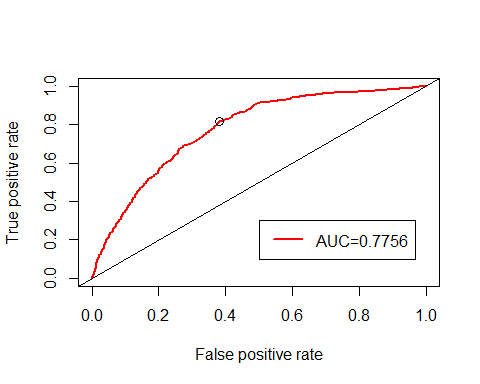
# ROC curve - install package ROCR  
par (mfrow=c(1,1))  
library(ROCR)  
pred1 <- prediction(m.both$fitted.values, m.both$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]  
auc1

## [1] 0.7755819

plot(perf1, lwd=2, col=2)  
abline(0,1)  
legend(0.5, 0.3, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)  
  
# Extract the X and Y values from the ROC plot, as well as the probability cutoffs  
roc.x = slot (perf1, "x.values") [[1]]  
roc.y = slot (perf1, "y.values") [[1]]  
cutoffs = slot (perf1, "alpha.values") [[1]]  
  
auc.table = cbind.data.frame(cutoff=pred1@cutoffs, tp=pred1@tp, fp=pred1@fp, tn=pred1@tn, fn=pred1@fn)  
names (auc.table) = c("Cutoff", "TP", "FP", "TN", "FN")  
auc.table$sensitivity = auc.table$TP / (auc.table$TP + auc.table$FN)  
auc.table$specificity = auc.table$TN / (auc.table$TN + auc.table$FP)  
auc.table$FalsePosRate = 1 - auc.table$specificity  
auc.table$sens\_spec = auc.table$sensitivity + auc.table$specificity  
  
# Find the row(s) in the AUC table where sensitivity + specificity is maximized  
  
auc.best = auc.table [auc.table$sens\_spec == max (auc.table$sens\_spec),]  
auc.best

## Cutoff TP FP TN FN sensitivity specificity FalsePosRate  
## 1256 0.2487238 578 720 1166 131 0.8152327 0.6182397 0.3817603  
## sens\_spec  
## 1256 1.433472

# Plot the maximum point(s) on the ROC plot  
  
points (auc.best$FalsePosRate, auc.best$sensitivity, cex=1.3)



The ROC curve indicates that the predictive nature of the model is much better than guessing because the AUC is larger than 0.5 at 0.7756. The optimal cutoff for classification is a fitted probability of 0.2487, which has a false positive rate of 0.38, and a true positive rate of 0.82. The true positive rate value of 0.82 is seen in the graph above as the black circle.

Following this, we can make some predictions looking at different ages relative to the other predictor variables in the model.

# With logistic regression, the predict function does not provide confidence limits, even  
# with the interval= option. Instead, we request the se.fit=T option and calculate our own  
# limits on the logist scale, and then back-transform to the probability scale.  
  
preds = predict (m.both, se.fit = T)  
pred.df = cbind.data.frame (bm, as.data.frame (preds))  
  
pred.df$lwr = pred.df$fit - 1.96 \* pred.df$se.fit  
pred.df$upr = pred.df$fit + 1.96 \* pred.df$se.fit  
  
pred.df$fit.pr = round (exp (pred.df$fit) / (1 + exp (pred.df$fit)), 3)  
pred.df$lwr.pr = round (exp (pred.df$lwr) / (1 + exp (pred.df$lwr)), 3)  
pred.df$upr.pr = round (exp (pred.df$upr) / (1 + exp (pred.df$upr)), 3)  
  
pred.df [c(20,47,78,21,16,11,84, 48, 41, 51, 74, 58), c(2,3:7,12:20)]

## age campaign previous cci ne y cci.c fit se.fit  
## 20 57 1 0 -36.4 5191 0 2.717457 -2.486270 0.1467083  
## 47 59 6 0 -36.4 5191 0 2.717457 -2.629380 0.1731527  
## 78 57 1 0 -36.4 5191 0 2.717457 -2.486270 0.1467083  
## 21 28 1 0 -36.4 5191 0 2.717457 -2.542967 0.1541739  
## 16 59 2 0 -36.4 5191 0 2.717457 -2.511764 0.1403926  
## 11 58 1 0 -36.4 5191 0 2.717457 -2.484315 0.1470425  
## 84 58 1 0 -36.4 5191 0 2.717457 -2.484315 0.1470425  
## 48 42 1 0 -36.4 5191 0 2.717457 -2.515596 0.1464786  
## 41 47 2 0 -36.4 5191 0 2.717457 -2.535225 0.1385518  
## 51 32 1 0 -36.4 5191 0 2.717457 -2.535146 0.1512429  
## 74 28 1 0 -36.4 5191 0 2.717457 -2.542967 0.1541739  
## 58 28 2 0 -36.4 5191 0 2.717457 -2.572371 0.1477542  
## residual.scale lwr upr fit.pr lwr.pr upr.pr  
## 20 1 -2.773818 -2.198722 0.077 0.059 0.100  
## 47 1 -2.968759 -2.290000 0.067 0.049 0.092  
## 78 1 -2.773818 -2.198722 0.077 0.059 0.100  
## 21 1 -2.845148 -2.240786 0.073 0.055 0.096  
## 16 1 -2.786933 -2.236594 0.075 0.058 0.097  
## 11 1 -2.772518 -2.196112 0.077 0.059 0.100  
## 84 1 -2.772518 -2.196112 0.077 0.059 0.100  
## 48 1 -2.802694 -2.228498 0.075 0.057 0.097  
## 41 1 -2.806786 -2.263663 0.073 0.057 0.094  
## 51 1 -2.831583 -2.238710 0.073 0.056 0.096  
## 74 1 -2.845148 -2.240786 0.073 0.055 0.096  
## 58 1 -2.861969 -2.282772 0.071 0.054 0.093

The table above shows us some predicted probabilities as well as their confidence intervals for customers regarded as students or retired employees.

The table below summarizes the observed and predicted classifications of term deposit.

pred.df$pred.bm = ifelse (pred.df$fit.pr >= auc.best$Cutoff[1], "Pred.Yes", "Pred.No")  
table (pred.df$y, pred.df$pred.bm)

##   
## Pred.No Pred.Yes  
## 0 1165 721  
## 1 131 578

The model has shown us the probability of term deposit which can be predicted using Age, Campaign (number of contacts performed during this campaign and for this client), Previous (number of contacts performed before this campaign and for this client), CCI(Consumer Confidence Index), NE(Number of Employees). The binary logistic regression model could then be utilized for helping the bank determine which sectors are doing well so that they could target their marketing campaigns to those industries.

Some future questions to this analysis could include:

1. Are there other factors that could be measured, that could improve the predictability of the model?
2. . How would catergorical variables affect the success of the final model and its interactions effects?