Analysis of Categorical Data Project 1

Maya Elshweikhy

Installing packages

```
library(tinytex)
library(carData)
library(magrittr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

library(pscl)

## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.

library(car)

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(ROCR)
library(ggplot2)
```

Reading Loan Data Set

```r
# reading the data
x<-read.csv('Loan_Data.csv', header = TRUE)
head(x)

##    Loan_ID Gender Married Dependents    Education Self_Employed
ApplicantIncome
## 1 LP001002   Male      No          0     Graduate            No
5849
## 2 LP001003   Male     Yes          1     Graduate            No
4583
## 3 LP001005   Male     Yes          0     Graduate           Yes
3000
## 4 LP001006   Male     Yes          0 Not Graduate            No
2583
## 5 LP001008   Male      No          0     Graduate            No
6000
## 6 LP001011   Male     Yes          2     Graduate           Yes
5417
##   CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History
Property_Area
## 1                 0         NA              360              1
Urban
## 2              1508        128              360              1
Rural
## 3                 0         66              360              1
Urban
## 4              2358        120              360              1
Urban
## 5                 0        141              360              1
Urban
## 6              4196        267              360              1
Urban
##   Loan_Status
## 1           Y
## 2           N
## 3           Y
## 4           Y
## 5           Y
## 6           Y

# Check for missing values
sum(is.na(x))

## [1] 86

# Remove missing values
x <- na.omit(x)
```

```r
x <- x[x$Gender != "", ]
x <- x[x$Married != "", ]
x <- x[x$Dependents != "", ]
x <- x[x$Education != "", ]
x <- x[x$Self_Employed != "", ]
x <- x[x$Property_Area != "", ]
x <- x[x$Loan_Status != "", ]
x$Loan_Status <- ifelse(x$Loan_Status == "Y", 1, 0)

x$Gender <- as.factor(x$Gender)
x$Married <- as.factor(x$Married)
x$Dependents <- as.factor(x$Dependents)
x$Education <- as.factor(x$Education)
x$Self_Employed <- as.factor(x$Self_Employed)
x$Property_Area <- as.factor(x$Property_Area)
x$Credit_History <- as.factor(x$Credit_History)
prop.table(table(x$Loan_Status))
```

```
##
##         0         1
## 0.3083333 0.6916667
```

Checking for Multicollinearity

Fitting a Logistic Regression Model and Checking for Multicollinearity

```r
fit1 <- glm(factor(x$Loan_Status) ~
            factor(x$Gender) + factor(x$Married) + factor(x$Dependents) +
            factor(x$Education) + factor(x$Self_Employed) + x$ApplicantIncome
+
            x$CoapplicantIncome + x$LoanAmount + x$Loan_Amount_Term +
            factor(x$Credit_History) + factor(x$Property_Area),
            family = binomial, data = x)
vif(fit1)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## factor(x$Gender)         1.234064  1        1.110884
## factor(x$Married)        1.436086  1        1.198368
## factor(x$Dependents)     1.417540  3        1.059878
## factor(x$Education)      1.084292  1        1.041294
## factor(x$Self_Employed)  1.054886  1        1.027077
## x$ApplicantIncome        1.572246  1        1.253892
## x$CoapplicantIncome      1.143515  1        1.069352
## x$LoanAmount             1.666382  1        1.290884
## x$Loan_Amount_Term       1.058330  1        1.028752
## factor(x$Credit_History) 1.041659  1        1.020617
## factor(x$Property_Area)  1.123862  2        1.029623
```

As shown in the R output, all VIF values are less than 10, therefore there is no multicollinearity indicatig that there are no correlated variables.

1. Use the step() function to find the best logistic model without any interaction terms and write the model equation.

```
fit2=step(fit1, test="Chisq")

## Start:  AIC=465.72
## factor(x$Loan_Status) ~ factor(x$Gender) + factor(x$Married) +
##     factor(x$Dependents) + factor(x$Education) + factor(x$Self_Employed) +
##     x$ApplicantIncome + x$CoapplicantIncome + x$LoanAmount +
##     x$Loan_Amount_Term + factor(x$Credit_History) +
factor(x$Property_Area)
##
##                           Df Deviance    AIC     LRT  Pr(>Chi)
## - factor(x$Dependents)     3   438.36 462.36   2.636  0.451294
## - x$ApplicantIncome        1   435.78 463.78   0.057  0.810764
## - factor(x$Self_Employed)  1   435.90 463.90   0.177  0.673970
## - x$Loan_Amount_Term       1   435.93 463.93   0.213  0.644731
## - factor(x$Gender)         1   436.67 464.67   0.954  0.328710
## - x$CoapplicantIncome      1   437.14 465.14   1.423  0.232875
## - factor(x$Education)      1   437.60 465.60   1.885  0.169738
## <none>                         435.72 465.72
## - x$LoanAmount             1   438.07 466.07   2.349  0.125371
## - factor(x$Married)        1   439.59 467.59   3.873  0.049078 *
## - factor(x$Property_Area)  2   448.67 474.67  12.953  0.001539 **
## - factor(x$Credit_History) 1   561.73 589.73 126.010 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=462.36
## factor(x$Loan_Status) ~ factor(x$Gender) + factor(x$Married) +
##     factor(x$Education) + factor(x$Self_Employed) + x$ApplicantIncome +
##     x$CoapplicantIncome + x$LoanAmount + x$Loan_Amount_Term +
##     factor(x$Credit_History) + factor(x$Property_Area)
##
##                           Df Deviance    AIC     LRT  Pr(>Chi)
## - x$ApplicantIncome        1   438.40 460.40   0.045  0.832811
## - x$Loan_Amount_Term       1   438.48 460.48   0.125  0.723306
## - factor(x$Self_Employed)  1   438.52 460.52   0.164  0.685760
## - x$CoapplicantIncome      1   439.63 461.63   1.279  0.258144
## - factor(x$Gender)         1   439.70 461.70   1.342  0.246718
## - factor(x$Education)      1   440.23 462.23   1.873  0.171080
## <none>                         438.36 462.36
## - x$LoanAmount             1   440.75 462.75   2.397  0.121546
## - factor(x$Married)        1   443.07 465.07   4.710  0.029985 *
## - factor(x$Property_Area)  2   450.77 470.77  12.415  0.002014 **
## - factor(x$Credit_History) 1   564.04 586.04 125.683 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=460.4
```

```
## factor(x$Loan_Status) ~ factor(x$Gender) + factor(x$Married) +
##     factor(x$Education) + factor(x$Self_Employed) + x$CoapplicantIncome +
##     x$LoanAmount + x$Loan_Amount_Term + factor(x$Credit_History) +
##     factor(x$Property_Area)
##
##                             Df Deviance    AIC     LRT  Pr(>Chi)
## - x$Loan_Amount_Term        1   438.53 458.53   0.133  0.714965
## - factor(x$Self_Employed)   1   438.54 458.54   0.142  0.706488
## - factor(x$Gender)          1   439.75 459.75   1.349  0.245367
## - x$CoapplicantIncome       1   439.89 459.89   1.494  0.221520
## - factor(x$Education)       1   440.31 460.31   1.911  0.166892
## <none>                          438.40 460.40
## - x$LoanAmount              1   441.33 461.33   2.935  0.086693 .
## - factor(x$Married)         1   443.07 463.07   4.667  0.030751 *
## - factor(x$Property_Area)   2   450.82 468.82  12.424  0.002006 **
## - factor(x$Credit_History)  1   564.15 584.15 125.747 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=458.53
## factor(x$Loan_Status) ~ factor(x$Gender) + factor(x$Married) +
##     factor(x$Education) + factor(x$Self_Employed) + x$CoapplicantIncome +
##     x$LoanAmount + factor(x$Credit_History) + factor(x$Property_Area)
##
##                             Df Deviance    AIC     LRT  Pr(>Chi)
## - factor(x$Self_Employed)   1   438.67 456.67   0.133  0.715688
## - factor(x$Gender)          1   439.95 457.95   1.417  0.233892
## - x$CoapplicantIncome       1   440.03 458.03   1.499  0.220886
## - factor(x$Education)       1   440.37 458.37   1.833  0.175721
## <none>                          438.53 458.53
## - x$LoanAmount              1   441.51 459.51   2.981  0.084238 .
## - factor(x$Married)         1   443.38 461.38   4.850  0.027650 *
## - factor(x$Property_Area)   2   450.98 466.98  12.449  0.001981 **
## - factor(x$Credit_History)  1   564.15 582.15 125.617 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=456.67
## factor(x$Loan_Status) ~ factor(x$Gender) + factor(x$Married) +
##     factor(x$Education) + x$CoapplicantIncome + x$LoanAmount +
##     factor(x$Credit_History) + factor(x$Property_Area)
##
##                             Df Deviance    AIC     LRT  Pr(>Chi)
## - factor(x$Gender)          1   440.10 456.10   1.437  0.230572
## - x$CoapplicantIncome       1   440.16 456.16   1.495  0.221418
## - factor(x$Education)       1   440.52 456.52   1.856  0.173056
## <none>                          438.67 456.67
## - x$LoanAmount              1   441.84 457.84   3.177  0.074678 .
## - factor(x$Married)         1   443.49 459.49   4.823  0.028089 *
## - factor(x$Property_Area)   2   451.13 465.13  12.464  0.001965 **
```

```
## - factor(x$Credit_History)  1   564.35 580.35 125.680 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=456.1
## factor(x$Loan_Status) ~ factor(x$Married) + factor(x$Education) +
##     x$CoapplicantIncome + x$LoanAmount + factor(x$Credit_History) +
##     factor(x$Property_Area)
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## - x$CoapplicantIncome       1   441.20 455.20   1.100  0.294244
## - factor(x$Education)       1   441.77 455.77   1.671  0.196070
## <none>                          440.10 456.10
## - x$LoanAmount              1   443.13 457.13   3.023  0.082110 .
## - factor(x$Married)         1   447.88 461.88   7.778  0.005288 **
## - factor(x$Property_Area)   2   451.79 463.79  11.686  0.002900 **
## - factor(x$Credit_History)  1   566.33 580.33 126.226 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=455.2
## factor(x$Loan_Status) ~ factor(x$Married) + factor(x$Education) +
##     x$LoanAmount + factor(x$Credit_History) + factor(x$Property_Area)
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## - factor(x$Education)       1   442.72 454.72   1.513  0.218624
## <none>                          441.20 455.20
## - x$LoanAmount              1   444.85 456.85   3.648  0.056135 .
## - factor(x$Married)         1   448.72 460.72   7.515  0.006119 **
## - factor(x$Property_Area)   2   453.01 463.01  11.807  0.002729 **
## - factor(x$Credit_History)  1   567.39 579.39 126.189 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=454.72
## factor(x$Loan_Status) ~ factor(x$Married) + x$LoanAmount +
factor(x$Credit_History) +
##     factor(x$Property_Area)
##
##                            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                          442.72 454.72
## - x$LoanAmount              1   445.58 455.58   2.862  0.090721 .
## - factor(x$Married)         1   449.91 459.91   7.195  0.007312 **
## - factor(x$Property_Area)   2   454.75 462.75  12.029  0.002443 **
## - factor(x$Credit_History)  1   570.59 580.59 127.875 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the output we will fit the best model

```
best_model<-
glm(factor(x$Loan_Status)~factor(x$Married)+factor(x$Credit_History)+factor(x
$Property_Area)+x$LoanAmount, family = binomial, data = x)
summary(best_model)

##
## Call:
## glm(formula = factor(x$Loan_Status) ~ factor(x$Married) +
factor(x$Credit_History) +
##     factor(x$Property_Area) + x$LoanAmount, family = binomial,
##     data = x)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2110  -0.4295   0.5169   0.6970   2.4761
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.696180   0.514478  -5.241  1.6e-07 ***
## factor(x$Married)Yes              0.667373   0.248585   2.685  0.00726 **
## factor(x$Credit_History)1         3.617154   0.425869   8.494  < 2e-16 ***
## factor(x$Property_Area)Semiurban  0.938358   0.297659   3.152  0.00162 **
## factor(x$Property_Area)Urban      0.147326   0.289297   0.509  0.61057
## x$LoanAmount                     -0.002474   0.001444  -1.713  0.08664 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.05  on 479  degrees of freedom
## Residual deviance: 442.72  on 474  degrees of freedom
## AIC: 454.72
##
## Number of Fisher Scoring iterations: 4
```

Model Equation

Log (π/1-π) =-2.696180+0.667373 Married (yes) + 3.617154 CreditHistory (1) + 0.938358 PropertyArea (Semiurban) + 0.147326 Property Area (Urban) - 0.002474 LoanAmount

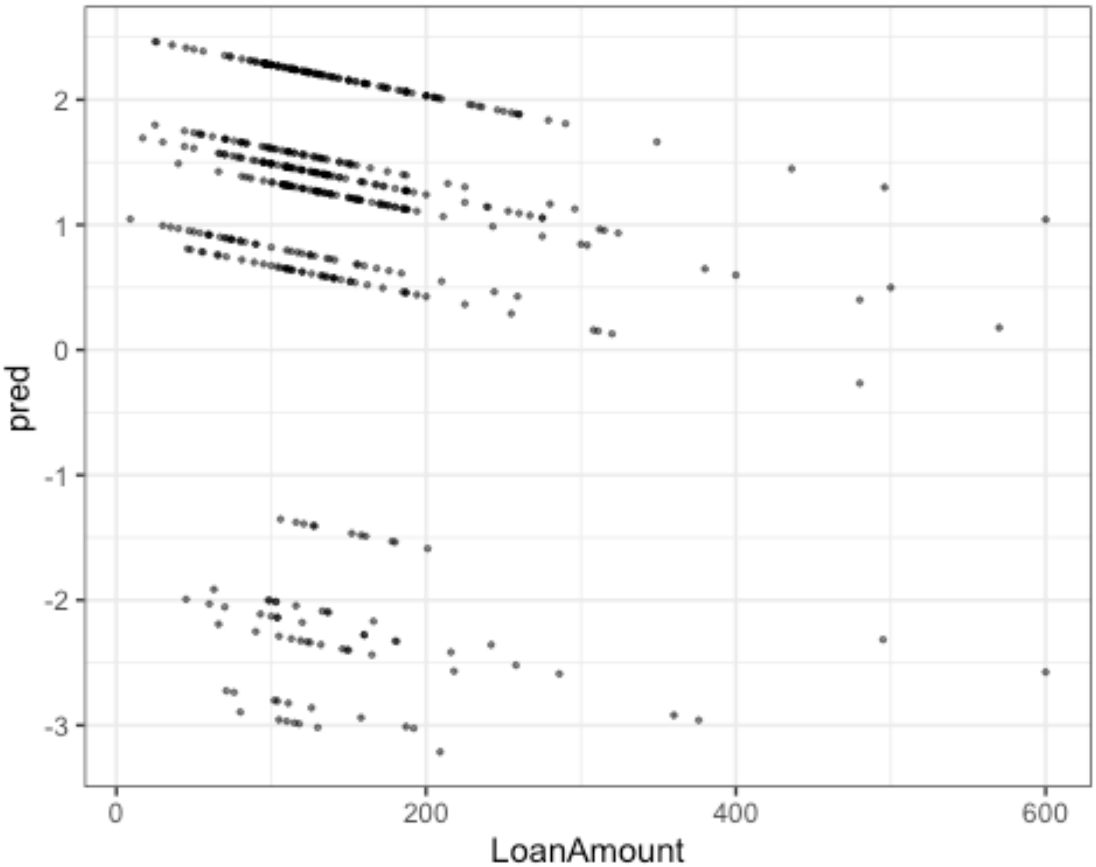
2. Run all model diagnostics and comment on them.

1) Checking for Linearity

```
pred=predict(best_model, type = "link")
p<-ggplot(data = data.frame(LoanAmount= x$LoanAmount, pred = pred),
aes(LoanAmount, pred)) +
geom_point(size = 0.5, alpha = 0.5) +
geom_smooth(method = "lowess") + theme_bw()
p

## `geom_smooth()` using formula = 'y ~ x'
```

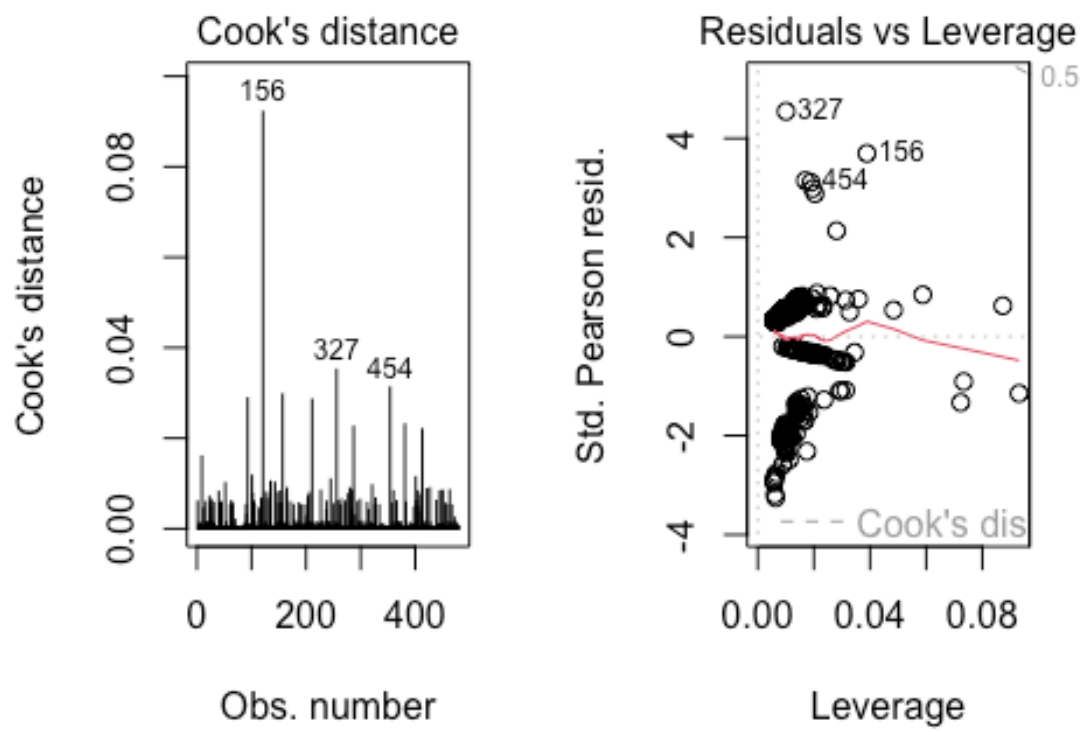
```
## Warning: Computation failed in `stat_smooth()`
## Caused by error in `method()`:
## ! unused arguments (data = data, weights = weight)
```



As shown in the R output, from the graph we can see that there a linear between the Loan Amount and the log of the odds ratio.

2) Checking for Influential Points

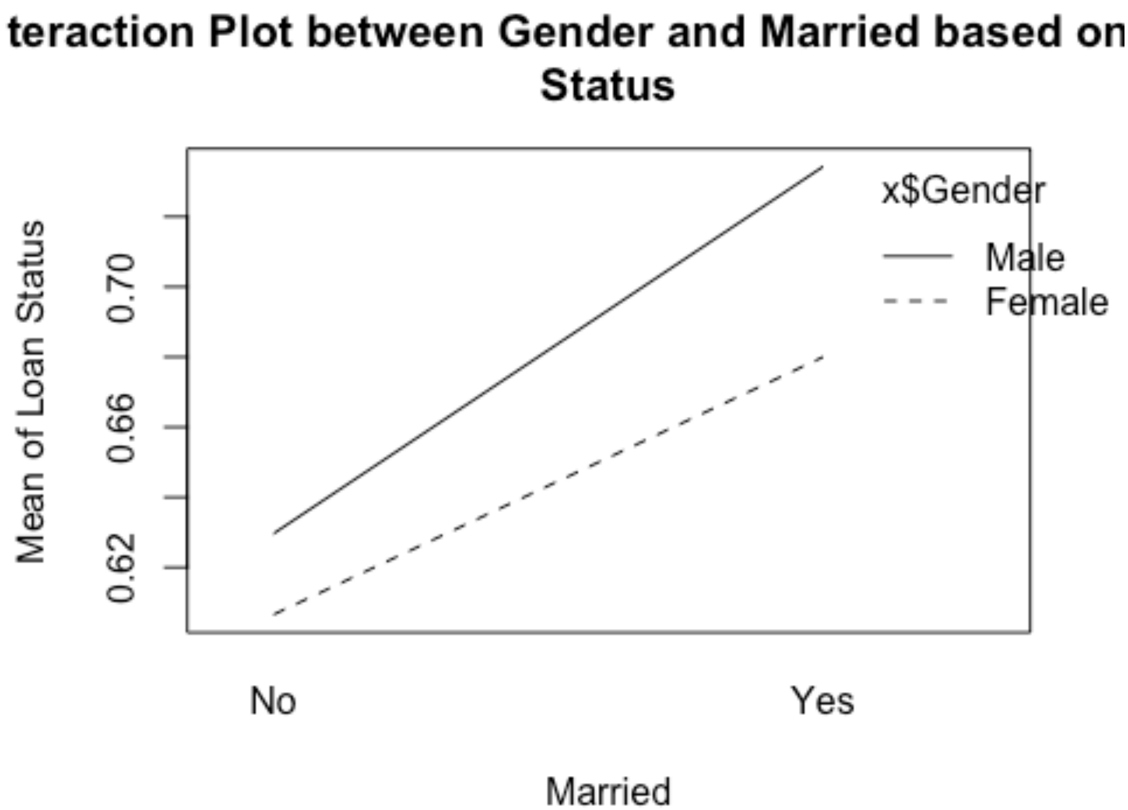
```
par(mfrow=c(1,2))
plot(best_model, which=4)
plot(best_model, which=5)
```



As shown in the R output,there are influential points in the data. To illustrate, from Cook's distance graph, there are three points identified as influential points which are observations 156, 327, and 454 respectively. From the Residuals vs. Leverage plot, we can see that we have outliers in x and in y. The points that are considered high residuals are observations 32, 454, and 156 and these points are also outliers in y. There are about five points that considered high leverage on the right and these points are outliers in x.

3. After reaching the best model in (1), include the interaction term between Married and Gender to your best model and interpret its coefficients.

```
interaction.plot(x$Married, x$Gender, x$Loan_Status,
main = "Interaction Plot between Gender and Married based on Loan
Status",
xlab = "Married", ylab = "Mean of Loan Status", legend =
TRUE)
```



Fitting the model with interaction term

```
fit3=glm(Loan_Status ~
factor(Gender)*factor(Married)+factor(Credit_History)+factor(Property_Area)+L
oanAmount, family = binomial,data = x)
summary(fit3)

##
## Call:
## glm(formula = Loan_Status ~ factor(Gender) * factor(Married) +
##     factor(Credit_History) + factor(Property_Area) + LoanAmount,
##     family = binomial, data = x)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2374  -0.4318   0.5410   0.6938   2.4478
##
## Coefficients:
##                                        Estimate Std. Error z value
Pr(>|z|)
## (Intercept)                           -2.855952   0.563798  -5.066 4.07e-
07 ***
## factor(Gender)Male                     0.234781   0.385779   0.609
```

```
0.54280
## factor(Married)Yes                        0.378948   0.618331   0.613
0.53997
## factor(Credit_History)1                  3.618809   0.426695   8.481  < 2e-
16 ***
## factor(Property_Area)Semiurban           0.974726   0.300649   3.242
0.00119 **
## factor(Property_Area)Urban               0.147816   0.289606   0.510
0.60977
## LoanAmount                              -0.002488   0.001433  -1.736
0.08250 .
## factor(Gender)Male:factor(Married)Yes  0.240438   0.677527   0.355
0.72268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.05  on 479  degrees of freedom
## Residual deviance: 441.67  on 472  degrees of freedom
## AIC: 457.67
##
## Number of Fisher Scoring iterations: 4
```

Interpretation of Interaction term coefficient The difference between the log-odds ratio comparing males vs females who are married and the log-odds ratio comparing males vs. females who are not married is 0.240438, holding all other variables constant.

4. Use the best model you reach whether with or without the interaction term to

(a) find the confusion matrix at a threshold of 0.5.

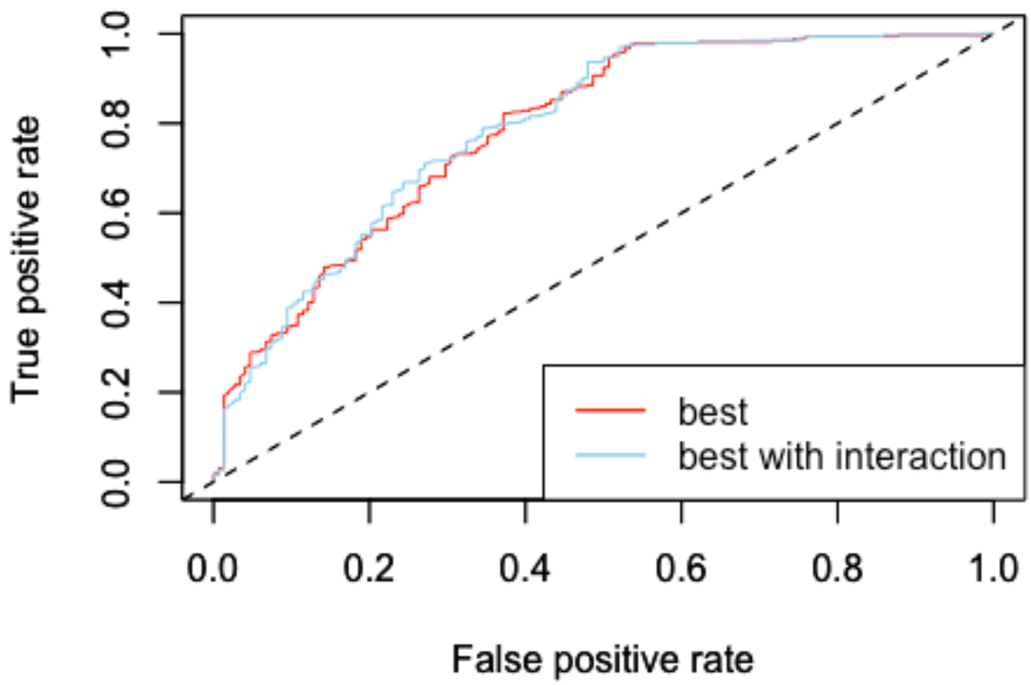
```
predict_1=predict(best_model, type="response")
table(best_model$y)

##
##   0   1
## 148 332
```

(b) draw the ROC curves for the two models, compare them and interpret them.

```
pred1 = predict(best_model, type = "response")
pred2 = predict(fit3, type = "response")
rocr.pred1 = prediction(pred1, labels =x$Loan_Status)
rocr.pred2 = prediction(pred2, labels =x$Loan_Status) #ROCR prediction object
roc.perf1 = performance(rocr.pred1, measure = "tpr", x.measure = "fpr")
roc.perf2 = performance(rocr.pred2, measure = "tpr", x.measure = "fpr") #ROCR
performance object
plot(roc.perf1, col = "red")
abline(a = 0, b = 1, lty = 2)
par(new=TRUE)
```

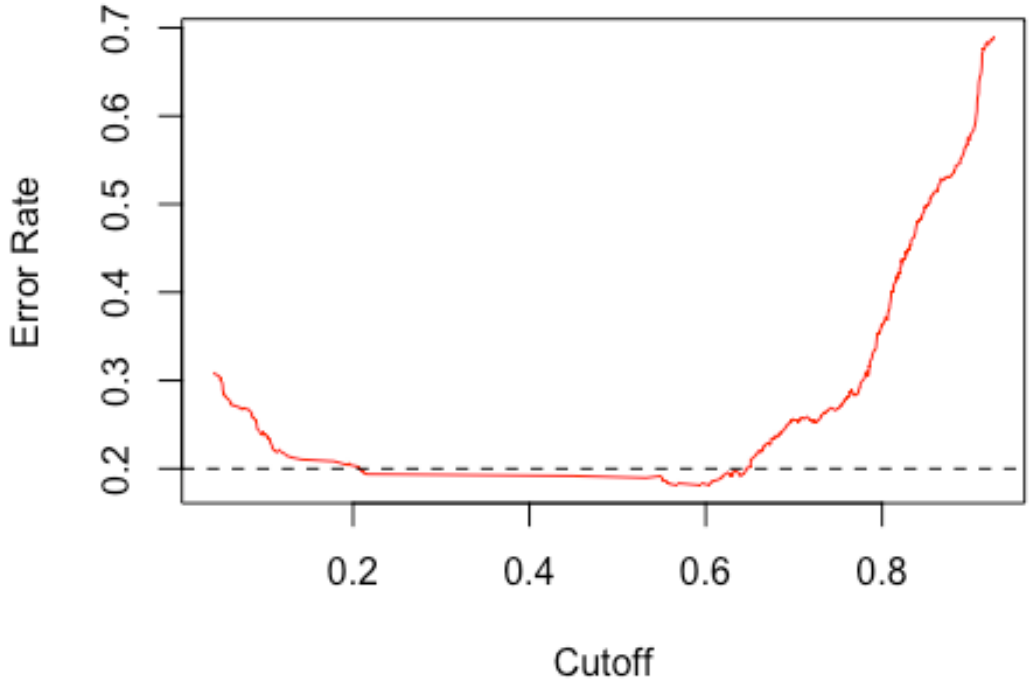
```r
plot(roc.perf2, col = "skyblue")
abline(a = 0, b = 1, lty = 2) #diagonal corresponding to a random assignment
legend("bottomright", leg = c("best","best with interaction"), col
=c("red","skyblue"), lwd = 1.5)
```



From the ROC plot, we can see that both curves are almost similar as they have the same shape. We can see that the best model with interaction term is better than the best model. The aim is to hae more values at the upper right so our model would have high TPR and Low FPR and sccorddigly our model correctly predit.

(c) what is the approximate optimum threshold to determine the loan eligibility using the better model.

```r
plot(performance(rocr.pred2, measure = "err"), col = "red")
abline(h = 0.2, lty = 2)
```



The approximate optimum threshold to determine the loan eligibility using the better model is approximatly 0.6 and this is the cut off point in which reach the minimum error rate.