An Assignment on Logistic Regression

Course Name: Applied Regression Analysis

Course Code: PM-ASDS08

Submitted to:

Sultana Begum

Assistant Professor

Department of Statistics, JU

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in
Applied Statistics and Data Science (PM-ASDS)
JAHANGIRNAGAR UNIVERSITY

Class Performance 3

Problem Statement:

Suppose that an analyst is asked by the management of a company to conduct an analysis to determine how the factors of financial performance (sales), customer ratings, and performance ratings influence the likelihood of a given salesperson being promoted.

a) Test the goodness of fit of the model

b) Comment on the significance of the parameter estimates (use $\alpha = 0.05$)

c) Point out any issues with diagnostics by plotting residuals

d) Interpret the relationship between predictors and response of the final model

Dataset Description:

Description:

To determine the likelihood of a given salesperson being promoted based on the impact of three variables, which are how the factors of financial performance (sales), customer ratings, and performance ratings influence. Our dataset has 350 observations and 4 variables.

Format:

The variables are described as

- Promoted (y):1 if the individual was promoted and 0 if not
- Sales (x_1): the sales (in thousands of dollars) attributed to the individual in the period of the promotion
- customer rate (x_2): the average satisfaction rating from a survey of the customers during the promotion period

~ performance (x_3): the most recent performance rating prior to promotion, from 1 (lowest) to 4 (highest)

Descriptive Statistics:

	y	x_1	x_2	x_3
Min	0	151	1	1
1st Qu.	0	389.2	3	2
Median	0	475	3.620	3
Mean	0.3229	527	3.608	2.5
3rd qu.	1	667.2	4.290	3
Max	1	945	5	4

Logistic Regression Model:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4 + \epsilon$$

Where $x_1 =$ Sales ($1000)

$x_2 =$ Customer Rating

$$D_2 = \begin{cases} 1, & if\ performance\ rank\ 2 \\ 0, & Otherwise \end{cases}$$

$$D_3 = \begin{cases} 1, & if\ performance\ rank\ 3 \\ 0, & Otherwise \end{cases}$$

$$D_4 = \begin{cases} 1, & if\ performance\ rank\ 4 \\ 0, & Otherwise \end{cases}$$

Assumptions:

~ The response variable is binary

~ There is no multicollinearity among explanatory variables

~ There is a linear relation between explanatory variables and logit of response

~ There should be no outliers or highly influential points

To fit logistic regression, we follow the below checklist

Checklist: Based on train dataset

i. Remove missing observation if any

ii. Check multicollinearity

iii. Run regression (primary step)

iv. Check influential observations

v. Test goodness of fit (Deviance statistic, Hosmer-Lemeshow statistic, Pseudo R-Squared, AIC)

vi. Perform individual test (Wald test)

vii. Perform other diagnostic test(s)

Now we do step by step following the above checklist:

i. Remove missing observation if any:

After importing the dataset, we checked that there are no missing values in the dataset. Below, we see the internal structure of the dataset.

```
# import data
data<-read.table("sales_Table1.2.txt",header=TRUE)
str(data)
```

```
## 'data.frame':    350 obs. of  4 variables:
##  $ promt   : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ sales   : int  594 446 674 525 657 918 318 364 342 387 ...
##  $ custrate: num  3.94 4.06 3.83 3.62 4.4 4.54 3.09 4.89 3.74 3 ...
##  $ perf    : int  2 3 4 2 3 2 3 1 3 3 ...
```

ii. Check multicollinearity:

Below scatter plot shows the relationship between variables. And from the correlation matrix,

~ Sales and Customer Rating: 33.8% (Weak Positive Relationship)

~ Sales and Performance: 28% (Weak Positive Relationship)

~ Customer Rating and Performance: 5.9% (Weak Positive Relationship)

The relationships show that there is no multicollinearity exists in our data.

```
## Checking multicollinearity
# scatter plot and correlation matrix
library(ggplot2)
library(GGally)
ggpairs(data = data, mapping = aes(color="red"))
```

Fig: Scatter Plot

iii. Run regression (primary step):

In this step, we fit logistic regression into our model to get an overall summary of the dataset. Here, we consider Rank 1 in performance variable is reference.

```
# Performance category declared as factor
perf1<-as.factor(data$perf)
# Reference Category: Consider Rank 1 in Performance is reference
per<-relevel(perf1,ref=1)
sales_n<- cbind(data[c("promt", "sales", "custrate")],per)
str(sales_n)
```

```
## 'data.frame':    350 obs. of  4 variables:
##  $ promt   : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ sales   : int  594 446 674 525 657 918 318 364 342 387 ...
##  $ custrate: num  3.94 4.06 3.83 3.62 4.4 4.54 3.09 4.89 3.74 3 ...
##  $ per     : Factor w/ 4 levels "1","2","3","4": 2 3 4 2 3 2 3 1 3 3 ...
```

The summary of logistic model:

```
## Primary Step regression
modelfull <- glm( promt~.,data =sales_n, family = "binomial")
summary(modelfull)
## Call:
## glm(formula = promt ~ ., family = "binomial", data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11890  -0.08910  -0.01981   0.00867   2.98107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.858932   3.444079  -5.766 8.11e-09 ***
## sales         0.040124   0.006576   6.101 1.05e-09 ***
## custrate     -1.112131   0.482682  -2.304   0.0212 *
## per2          0.263000   1.021980   0.257   0.7969
## per3          0.684955   0.982167   0.697   0.4856
## per4          0.734493   1.071964   0.685   0.4932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  64.374  on 344  degrees of freedom
## AIC: 76.374
##
## Number of Fisher Scoring iterations: 8
```

iv. Check influential observations

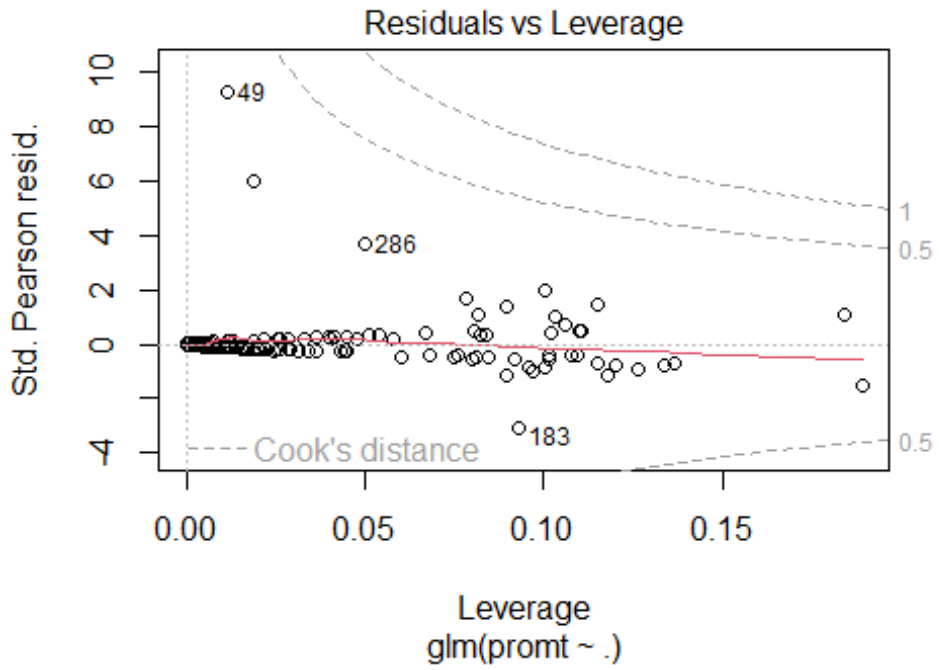
From the below analysis we get, there are no influence observations in the dataset.

And the Residuals vs Leverage plot shows that there are no points falling in the Cook's distance. Hence, there are no influence observations in the dataset.

```
## Identifying Influential Observations
inf.id = which(cooks.distance(modelfull)>1) ## Check for 0.5 or 1)
inf.id

## named integer(0)
```

```
plot_diag<-plot(modelfull)
```



The BIC of the model:

$$BIC = (n-p-q)\ln\left[\frac{n\hat{\sigma}^2}{(n-p-q)}\right] + n\left(1+ln\sqrt{2\pi}\right) + (p+q)ln\left[\frac{\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2}{p+q}\right]$$

```
BIC(modelfull)
```

```
## [1] 99.52193
```

v. Test goodness of fit:

H_0: Model fits well

H_a : Model does not fit well

Hosmer-Lemeshow Statistics: $G_{HL} = \sum_{j=1}^{g} \frac{(o_j - e_j)^2}{e_j}$, Where ***oj*** is the number of observed successes and failures and e_j is the number of expected successes and failures for a given group j If the model fits the data well $G_{HL} \sim X_{g-2}^2$

The below Hosmer-Lemeshow test gives p-value of 0.7126 which is above the level of 0.05 (alpha). We don't reject the null hypothesis. Hence, the model is adequate.

The value of AIC is also low (76.374). $AIC = -2l + 2q$

The value of $pseudo\ R^2$ is $\left(1 - \frac{64.374(\text{Residual deviance})}{440.303(\text{Null deviance})}\right)$, 0.845. 84.5% deviance is explained by the model.

```
## Goodness of fit test
library(ResourceSelection)

HL<-hoslem.test(sales_n$promt,fitted(modelfull),g=7) ## g= Number of parameters in the model+1
HL

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  sales_n$promt, fitted(modelfull)
## X-squared = 2.918, df = 5, p-value = 0.7126
```

vi. Perform individual test (Wald test):

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0\ (\beta_1 > 0\ or\ \beta_1 < 0)$

Test Statistic, $Z = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}$

Decision: Reject the null hypothesis at α level of significance if 0.05> p-value.

To find individual significant variables, here we do 4 tests which are-

~ Chi-square significant test

~ Fit logistic regression changing performance rank

~ Stepwise Method

~ Finding odds Ratio

Chi-square significant test:

Here, the Chi-square test shows that the p-value of sales and customer rating is less than 0.05 (alpha). And the p-value of the performance rating is greater

than 0.05. Hence, this test shows that sales and customer ratings are significant for a salesperson being promoted or not.

```
## Individual test
drop1(modelfull,test='Chisq')
```

```
## Single term deletions
##
## Model:
## promt ~ sales + custrate + per
##          Df Deviance    AIC    LRT  Pr(>Chi)
## <none>         64.37  76.37
## sales     1   402.89 412.89 338.52 < 2.2e-16 ***
## custrate  1    71.06  81.06   6.69  0.009693 **
## per       3    65.13  71.13   0.76  0.859693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit model changing rank in performance:

When Rank 1 in performance is referenced, the below results show that sales and customer rating are significant for predicting whether a salesperson is being promoted or not. Performance rating is not significant.

```
# When Rank 1 in Performance is reference
ind_rank1 <- glm( promt~.,data =sales_n, family = "binomial")
summary(ind_rank1)
```

```
##
## Call:
## glm(formula = promt ~ ., family = "binomial", data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11890  -0.08910  -0.01981   0.00867   2.98107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.858932   3.444079  -5.766 8.11e-09 ***
## sales         0.040124   0.006576   6.101 1.05e-09 ***
## custrate     -1.112131   0.482682  -2.304   0.0212 *
## per2          0.263000   1.021980   0.257   0.7969
## per3          0.684955   0.982167   0.697   0.4856
## per4          0.734493   1.071964   0.685   0.4932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  64.374  on 344  degrees of freedom
## AIC: 76.374
##
## Number of Fisher Scoring iterations: 8
```

When Rank 2 in performance is referenced, the below results show that sales and customer rating are significant for predicting whether a salesperson is being promoted or not. Performance rating is not significant.

```
# When Rank 2 in Performance is reference
per<-relevel(perf1,ref=2)
sales_n<- cbind(data[c("promt", "sales", "custrate")],per)
ind_rank2 <- glm( promt~.,data =sales_n, family = "binomial")
summary(ind_rank2)

##
## Call:
## glm(formula = promt ~ ., family = "binomial", data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11890  -0.08910  -0.01981   0.00867   2.98107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.595932   3.487000  -5.620 1.91e-08 ***
## sales         0.040124   0.006576   6.101 1.05e-09 ***
## custrate     -1.112131   0.482682  -2.304   0.0212 *
## per1         -0.263000   1.021980  -0.257   0.7969
## per3          0.421955   0.852319   0.495   0.6206
## per4          0.471494   0.933504   0.505   0.6135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  64.374  on 344  degrees of freedom
## AIC: 76.374
##
## Number of Fisher Scoring iterations: 8
```

When Rank 3 in performance is referenced, the below results show that sales and customer rating are significant for predicting whether a salesperson is being promoted or not. Performance rating is not significant.

```
# When Rank 3 in Performance is reference
per<-relevel(perf1,ref=3)
sales_n<- cbind(data[c("promt", "sales", "custrate")],per)
ind_rank3 <- glm( promt~.,data =sales_n, family = "binomial")
summary(ind_rank3)

##
## Call:
## glm(formula = promt ~ ., family = "binomial", data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11890  -0.08910  -0.01981   0.00867   2.98107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.173977   3.352543  -5.719 1.07e-08 ***
## sales         0.040124   0.006576   6.101 1.05e-09 ***
## custrate     -1.112131   0.482682  -2.304   0.0212 *
## per1         -0.684955   0.982167  -0.697   0.4856
## per2         -0.421955   0.852319  -0.495   0.6206
## per4          0.049539   0.911615   0.054   0.9567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  64.374  on 344  degrees of freedom
## AIC: 76.374
##
## Number of Fisher Scoring iterations: 8
```

When Rank 4 in performance is referenced, the below results show that sales and customer rating are significant for predicting whether a salesperson is being promoted or not. Performance rating is not significant.

```
# When Rank 4 in Performance is reference
per<-relevel(perf1,ref=4)
sales_n<- cbind(data[c("promt", "sales", "custrate")],per)
```

```
ind_rank4 <- glm( promt~.,data =sales_n, family = "binomial")
summary(ind_rank4)
```

```
##
## Call:
## glm(formula = promt ~ ., family = "binomial", data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11890  -0.08910  -0.01981   0.00867   2.98107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.124439   3.501115  -5.462 4.70e-08 ***
## sales         0.040124   0.006576   6.101 1.05e-09 ***
## custrate     -1.112131   0.482682  -2.304   0.0212 *
## per1         -0.734493   1.071964  -0.685   0.4932
## per2         -0.471494   0.933504  -0.505   0.6135
## per3         -0.049539   0.911615  -0.054   0.9567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  64.374  on 344  degrees of freedom
## AIC: 76.374
##
## Number of Fisher Scoring iterations: 8
```

Stepwise Method:

The stepwise method shows that when we count sales, customer rating, and performance variables, the value of AIC is 76.37 and when we count sales and customer rating, the value of AIC is 71.13. The lowest value of AIC is acceptable. Hence, this test shows that sales and customer ratings are significant, and performance is not significant.

```
# Stepwise method
step(modelfull)
```

```
## Start:  AIC=76.37
## promt ~ sales + custrate + per
##
##            Df Deviance    AIC
## - per       3    65.13  71.13
```

```
## <none>           64.37  76.37
## - custrate  1    71.06  81.06
## - sales     1   402.89 412.89
##
## Step:  AIC=71.13
## promt ~ sales + custrate
##
##            Df Deviance    AIC
## <none>           65.13  71.13
## - custrate  1    72.50  76.50
## - sales     1   428.52 432.52

##
## Call:  glm(formula = promt ~ sales + custrate, family = "binomial",
##     data = sales_n)
##
## Coefficients:
## (Intercept)        sales     custrate
##   -19.51769      0.04039     -1.12206
##
## Degrees of Freedom: 349 Total (i.e. Null);  347 Residual
## Null Deviance:       440.3
## Residual Deviance: 65.13     AIC: 71.13
```

Odds Ratio:

$H_0: OR = 1$ (No relationship exists)

$H_a: OR \neq 1$ (Relationship exists)

An odds ratio equal to 1 indicates, no effects on the response variable and predictor variable. Here, the odds ratio of performance rank 2, 3, and 4 can be 1 because1 lies between their lower limit and upper limit of CI.

```
# Confidence interval for OR
exp(cbind(OR=coef(modelfull), confint(modelfull)))
```

```
##                        OR        2.5 %       97.5 %
## (Intercept) 2.373425e-09 7.879943e-13 7.385387e-07
## sales       1.040940e+00 1.029762e+00 1.057214e+00
## custrate    3.288573e-01 1.141645e-01 7.793018e-01
## per2        1.300826e+00 1.800447e-01 1.061602e+01
## per3        1.983682e+00 3.060299e-01 1.547188e+01
## per4        2.084426e+00 2.614852e-01 1.870827e+01
```

Variables	*Coefficients*	*Standards error of coefficients*	*p-value*	*Odds Ratio*		
				Value	Lower	Upper
Constant	-19.8589	3.4440	8.11e-09	-	-	-
Sales	0.0401	0.0065	1.05e-09	1.040	1.029	1.057
Customer Rating	-1.1121	0.4826	0.0212	0.328	0.1142	0.7793
Performance rank 2	0.2630	1.0219	0.7969	1.301	0.18004	1.0616
Performance rank 3	0.6849	0.9821	0.4856	1.984	0.30603	1.547
Performance rank 4	0.7344	1.0719	0.4932	2.084	0.2615	1.8708

vii. Perform other diagnostic test(s):

To check diagnostic test, we do

-Deviance residual

-Pearson residual

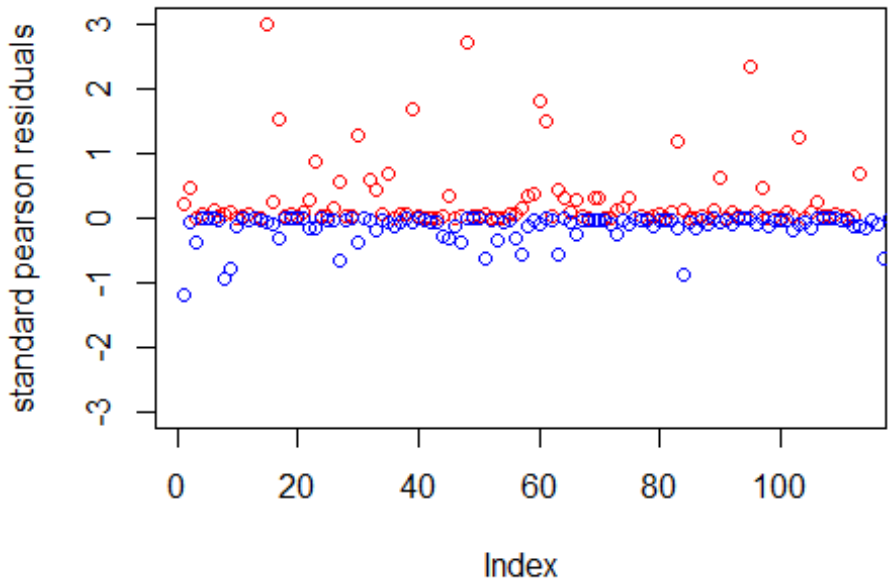
-Standardized person residual: $r_{si} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}}, i = 1,2,\dots,n$

This below figure presents parallel pattern. This is due to the feature of data (y = 0 and y =1). Hence the Bernoulli assumption is valid. Also, there is no systematic pattern in the plot, linearity assumption is also valid (link function is linear function of predictors).

```
## Residual analysis
res.p<-residuals(modelfull, type='pearson')
res.d<-residuals(modelfull, type='deviance')
```

```
sta.res.p<-residuals(modelfull, type='pearson')/sqrt(1-hatvalues(modelfull))
sta.res.d<-residuals(modelfull, type='deviance')/sqrt(1-hatvalues(modelfull))

plot(sta.res.d[data$promt==1],col='red',ylim=c(-3,3),ylab='standard pearson r
esiduals',xlab='Index')
points(sta.res.d[data$promt==0],col='blue',ylim=c(-3,3),ylab='standard pearso
n residuals',xlab='Index')
```



Best Fitted model:

The fitted model is

$$log\left(\frac{p}{1-p}\right) = -19.52 + 0.0404x_1 - 1.122x_2$$

Where, x_1 = Sales ($1000)

x_2 = Customer rating

~ For every one unit ($1,000) increase in sales, the odds of a salesman to be promoted increases by (exp(0.0404) = 1.0412) 4.12% holding other variable (customer rating) unchanged.

~ Similarly, for every five unit ($5,000) increase in sales, the odds of a salesman to be promoted increases by (exp(5 ∗ 0.0404) = 1.2238) 22.38% holding other variable (Customer rating) unchanged.

~ For a one unit increase in customer's rating, the odds of an individual to be promoted decrease by (exp(−1.122) = 0.33) 67% considering sales unchanged.

```
# Best fitted model
simpler_model <- glm( promt~sales+custrate,data =sales_n, family = "binomial"
)
summary(simpler_model)

##
## Call:
## glm(formula = promt ~ sales + custrate, family = "binomial",
##     data = sales_n)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.02984  -0.09256  -0.02070   0.00874   3.06380
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.517689   3.346762  -5.832 5.48e-09 ***
## sales         0.040389   0.006525   6.190 6.03e-10 ***
## custrate     -1.122064   0.466958  -2.403   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.303  on 349  degrees of freedom
## Residual deviance:  65.131  on 347  degrees of freedom
## AIC: 71.131
##
## Number of Fisher Scoring iterations: 8

BIC(simpler_model)

## [1] 82.70525
```

Prediction of New inputs:

Here, the new input is given to three salespersons being promoted or not based on their sales and customer rating.

1st salesperson's sales are $4200, and the customer rating is 3.4.

2nd salesperson's sales are $5100, and the customer rating is 2.3.

3rd salesperson's sales are $7100, and the customer rating is 4.2.

```r
# Prediction or Classification
new_data <- data.frame(sales = c(420, 510, 710),custrate = c(3.4, 2.3, 4.2))
predict<-round(predict(simpler_model, new_data, type = "response"))
x<-data.frame(predict)
x
```

```
##   predict
## 1       0
## 2       0
## 3       1
```

Here, 0: Not promoted

1: Promoted

The prediction result shows that 1st salesperson will not be promoted. 2nd salesperson will not be promoted. 3rd salesperson will be promoted

THE END