BINARAY LOGISTIC REGRESSION MODEL PERFORMANCE

Data Snapshot

Bank Loan Data

Independent Variables

Dependent Variable

SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTE
1	3	17	12	9.3	11.36	5.01	1
2	1	10	6	17.3	1.36	4	0

Column	Description	Type	Measurement	Possible Values
SN	Serial Number		-	-
AGE	Age Groups	Categorical	1(<28 years), 2(28-40 years), 3(>40 years)	3
EMPLOY	Number of years customer working at current employer	Continuous	-	Positive value
ADDRESS	Number of years customer staying at current address	Continuous	-	Positive value
DEBTINC	Debt to Income Ratio	Continuous	-	Positive value
CREDDEBT	Credit Card Debt	Continuous	-	Positive value
OTHDEBT	Other Debt	Continuous	-	Positive value

DATA SCIENCE INSTITUTE

Binary Logistic Regression in R

```
# Import data and check data structure before running model
```

```
data<-read.csv("BANK LOAN.csv",header=TRUE)
str(data)
```

```
data$AGE<-factor(data$AGE)
str(data)
```

```
'data.frame':   700 obs. of  8 variables:
 $ SN      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE     : Factor w/ 3 levels "1","2","3": 3 1 2 3 1 3 2 3 1 2 ...
 $ EMPLOY  : int  17 10 15 15 2 5 20 12 3 0 ...
 $ ADDRESS : int  12 6 14 14 0 5 9 11 4 13 ...
 $ DEBTINC : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
 $ CREDDEBT : num  11.36 1.36 0.86 2.66 1.79 ...
 $ OTHDEBT  : num  5.01 4 2.17 0.82 3.06 ...
 $ DEFAULTER: int  1 0 0 0 1 0 0 0 1 0 ...
```

Binary Logistic Regression in R

```
riskmodel<-glm(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,family=binomial,data=data)
summary(riskmodel)
```

```
Call:
glm(formula = DEFAULTER ~ AGE + EMPLOY + ADDRESS + DEBTINC +
    CREDDEBT + OTHDEBT, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3495  -0.6601  -0.2974   0.2509   2.8583

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78821    0.26407  -2.985  0.00284 **
AGE2         0.25202    0.26651   0.946  0.34433
AGE3         0.62707    0.36056   1.739  0.08201 .
EMPLOY      -0.26172    0.03188  -8.211  < 2e-16 ***
ADDRESS     -0.09964    0.02234  -4.459 8.22e-06 ***
DEBTINC      0.08506    0.02212   3.845  0.00012 ***
CREDDEBT     0.56336    0.08877   6.347 2.20e-10 ***
OTHDEBT      0.02315    0.05709   0.405  0.68517
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 553.41  on 692  degrees of freedom
AIC: 569.41

Number of Fisher Scoring iterations: 6
```

Interpretation :

- Since p-value is <0.05 for Employ, Address, Debtinc, Creddebt, these independent variables are significant.

Re-run Model in R

Re-run the model with employ, address, debtinc, creddebt.

```
riskmodel<-glm(DEFAULTER~EMPLOY+ADDRESS+DEBTINC+CREDDEBT,
               family=binomial,data=data)

summary(riskmodel)
```

Re-run Model in R

```
# Output:
```

```
Call:
glm(formula = DEFAULTER ~ EMPLOY + ADDRESS + DEBTINC + CREDDEBT,
    family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4483  -0.6396  -0.3108   0.2583   2.8496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79107    0.25154  -3.145  0.00166 **
EMPLOY      -0.24258    0.02806  -8.646  < 2e-16 ***
ADDRESS     -0.08122    0.01960  -4.144 3.41e-05 ***
DEBTINC      0.08827    0.01854   4.760 1.93e-06 ***
CREDDEBT     0.57290    0.08725   6.566 5.17e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 556.74  on 695  degrees of freedom
AIC: 566.74

Number of Fisher Scoring iterations: 6
```

Interpretation :

- Since p-value is <0.05 for Employ, Address, Debtinc, Creddebt, these independent variables are significant and sign of the coefficients are also logical.

DATA SCIENCE INSTITUTE

Final Model

- Final Model is :

$$\log\,() = -0.79107 - 0.24258 * (\text{EMPLOY}) - 0.08122 * (\text{ADDRESS}) + 0.08827 * (\text{DEBTINC}) + 0.57290\,(\text{CREDDEBT})$$

- This model is used for predicting the probabilities.

Predicting Probabilities in R

```
# Predicting Probabilities
```

```
data$predprob<-round(fitted(riskmodel),2)
head(data,n=10)
```

	SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTER	predprob
1	1	3	17	12	9.3	11.36	5.01	1	0.81
2	2	1	10	6	17.3	1.36	4.00	0	0.20
3	3	2	15	14	5.5	0.86	2.17	0	0.01
4	4	3	15	14	2.9	2.66	0.82	0	0.02
5	5	1	2	0	17.3	1.79	3.06	1	0.78
6	6	3	5	5	10.2	0.39	2.16	0	0.22
7	7	2	20	9	30.6	3.83	16.67	0	0.19
8	8	3	12	11	3.6	0.13	1.24	0	0.01
9	9	1	3	4	24.4	1.36	3.28	1	0.75
10	10	2	0	13	19.7	2.78	2.15	0	0.82

Last column in the data ‘predprob;’ is the probabilities generated using final model.

Classification Table

- Based on cut-off value of p, Y is estimated to be either 1 or 0

  Ex. $p>0.5$; $Y=1$

  $p\leq0.5$; $Y=0$

- Cross tabulation of observed values of Y and predicted values of Y is called as Classification Table.
- The predictive success of the logistic regression can be assessed by looking at the classification table, but classification table is not a good measure of goodness fit since it varies with the cut off value set.

- Accuracy Rate measures how accurate a model is in predicting outcomes.
- In the adjoining table, 479 times Y=0 was observed as well as predicted. Similarly, Y=1 was observed and predicted 92 times.

  Accuracy Rate = 479+92/700 = 81.571

		Expected	
		0	1
Observed	0	479	38
	1	91	92

DATA SCIENCE INSTITUTE

Misclassification

- Misclassification Rate ✉ Percentage of wrongly predicted observations
- Note that misclassification rate depends on cut off used for predictions

Suppose our classification table looks as follows:

		Expected	
		0	1
Observed	0	479	38
	1	91	92

- Here misclassification rate is : (38 +91) / 700=18.43%

Classification Table Terminology

Sensitivity	% of occurrences correctly predicted P(Ypred=1/Y=1)
Specificity	% of non occurrences correctly predicted P(Ypred=0/Y=0)
False Positive Rate (1 – Specificity)	% of non occurrences which are incorrectly predicted. P(Ypred=1/Y=0)
False Negative Rate (1- Sensitivity)	% of occurrences which are incorrectly predicted. P(Ypred=0/Y=1)

		Predicted	
		0	1
Observed	0	Specificity	False Positive (1-Specificity)
	1	False Negative (1-Sensitivity)	Sensitivity

DATA SCIENCE INSTITUTE

Classification and Sensitivity and Specificity table in R

```
# Predicting Probabilities
classificationtable<-table(data$DEFAULTER,data$predprob > 0.5)
classificationtable
```

- **table** function will create a cross table of observed Y (defaulter) vs. predicted Y

```
# Output:
   FALSE TRUE
 0   479   38
 1    91   92
```

Interpretation :

- True indicates predicted defaulters and False indicates predicted non-defaulters.
- There are 479 correctly predicted non-defaulters and 92 correctly predicted defaulters.
- There are 38 wrongly predicted as defaulters and 91 wrongly predicted as non-defaulters.

DATA SCIENCE INSTITUTE

Sensitivity and Specificity in R

```
# Sensitivity and Specificity
sensitivity<-(classificationtable[2,2]/
(classificationtable[2,2]+classificationtable[2,1]))*100
sensitivity

specificity<-(classificationtable[1,1]/
(classificationtable[1,1]+classificationtable[1,2]))*100
specificity
```

```
# Output:
sensitivity
[1] 50.27322

specificity
[1] 92.6499
```

Interpretation :

- The Sensitivity is at 50.3% and the Specificity is at 92.7% . This is when the cutoff was set at 0.5

Sensitivity and Specificity calculations

Cut-off Value		Accuracy	Sensitivity	Specificity
0.1	FALSE TRUE 0 252 265 1 12 171	(245+171)/700 = 60.4%	171/183=93.4%	245/517=48.7%
0.2	FALSE TRUE 0 352 165 1 28 155	(352+155)/700 = 72.4%	155/183=84.7%	352/517=68.1%
0.3	FALSE TRUE 0 415 102 1 46 137	(415+137)/700 = 78.9%	137/183=74.9%	415/517=80.3%
0.4	FALSE TRUE 0 449 68 1 70 113	(449+113)/700 = 80.14%	113/183=61.7%	449/517=86.8%
0.5	FALSE TRUE 0 479 38 1 91 92	(479+92)/700 =81. 57%	92/183=50.3%	479/517=92.6%

* Note : Here we are trying to find out the best cut-off value based on accuracy, sensitivity & specificity.

DATA SCIENCE INSTITUTE

Receiver Operating Characteristic Curve

- The Receiver Operating Characteristic (ROC) curve is

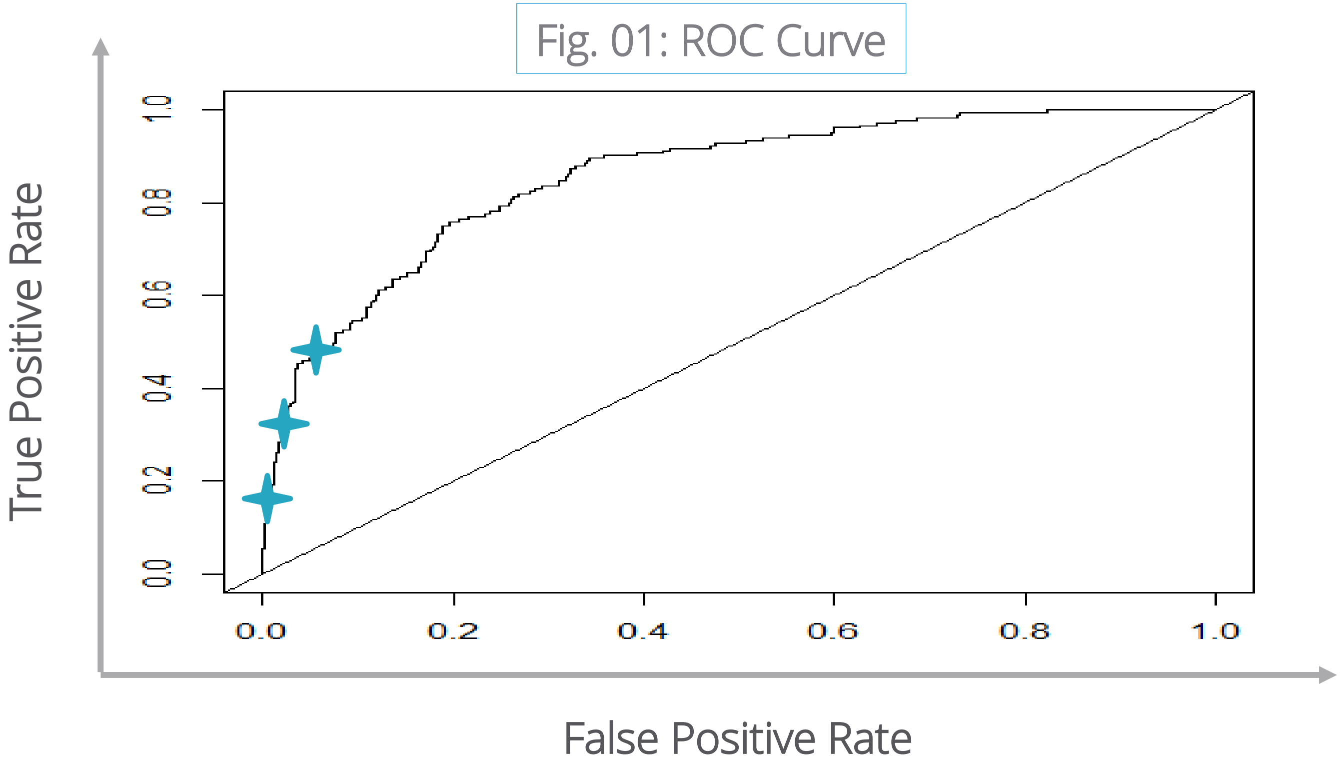
> A graphical representation of the trade off between the false positive and true positive rates for various cut off values

Y- axis: Sensitivity (true positive rate)

X-axis: 1-Specificity (false positive rate)

The performance of the classification model can be assesed by area under the ROC curve (C).

ROC Curve and Area Under ROC Curve



High TPR with low FPR is indicative of a good model. This will result in curve that is closer to the Y-axis and top left corner of the plot. It implies higher Area Under the ROC Curve.

DATA SCIENCE INSTITUTE

ROC Curve and Area Under ROC Curve

Interpreting different versions of an ROC curve

Critical Points	Interpretations
TPR = 0 and FPR = 0	Model predicts every instance to be Non-event
TPR = 1 and FPR = 1	Model predicts every instance to be Event
TPR = 1 and FPR = 0	The Perfect Model

- If the model is perfect, AUC = 1
- If the model is guessing randomly, AUC = 0.5
- Thumb rule: Area Under ROC Curve > 0.65 is considered acceptable

DATA SCIENCE INSTITUTE

ROC in R

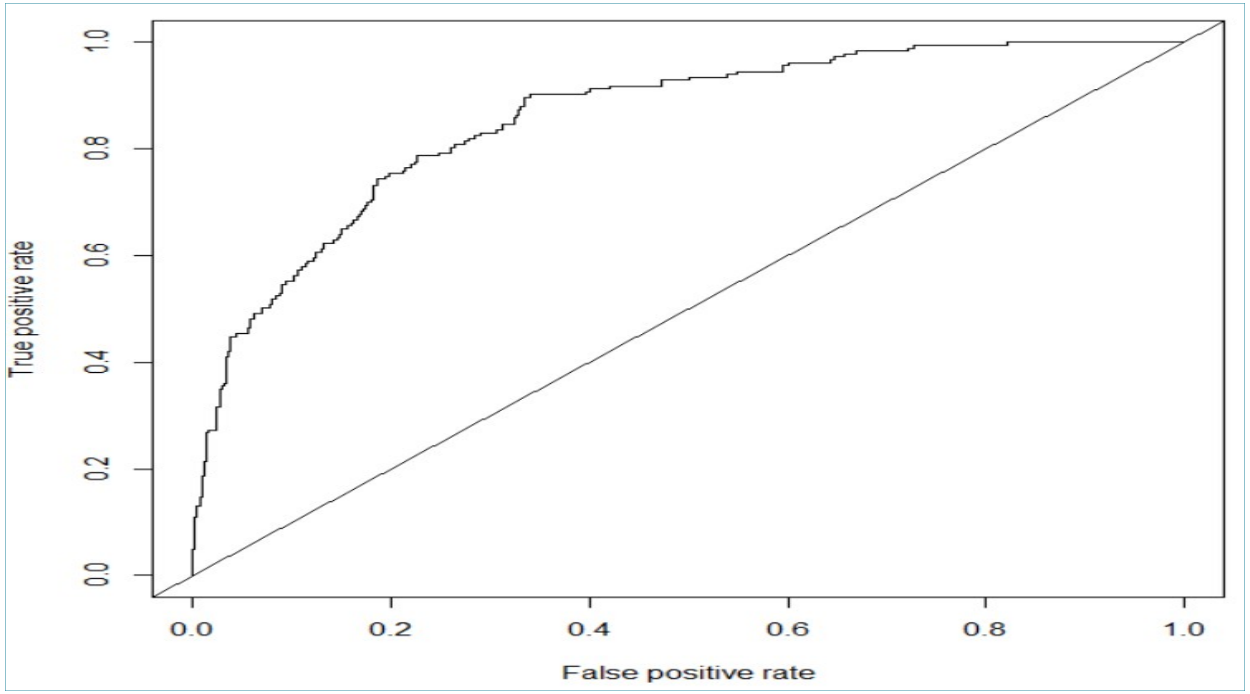
```
# Install and Load "ROCR" package.
```

```
install.packages("ROCR")
library(ROCR)

data$predprob<-fitted(riskmodel)
pred<-prediction(data$predprob,data$DEFAULTER)

perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)
```

- **prediction()** function prepares data required for ROC curve.
- **performance()** function creates performance objects, "tpr“ (True positive rate), "fpr” (False positive rate).
- **plot()** function plots the objects created using performance

ROC in R

```
# Output:
```




```
auc<-performance(pred,"auc")
auc@y.values
[1] 0.8556193
```

- ❑ Gives area under curve (AUC)

Interpretation :

- ➢ Area under the curve is 0.8556 which means model is performing well.

DATA SCIENCE INSTITUTE