Binary Logistic Regression

Checking Model Performance

Contents

Receiver Operating Characteristic Curve

- The **R**eceiver **O**perating **C**haracteristic (ROC) curve is

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

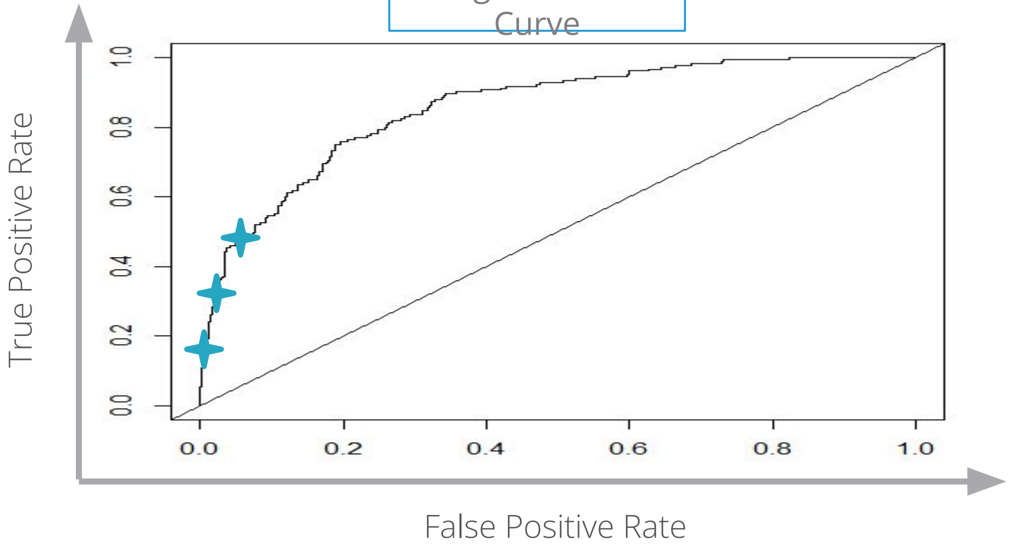
Y- axis: Sensitivity (true positive rate)

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

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

ROC Curve and Area Under ROC Curve

Fig. 01: ROC Curve



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

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

ROC in R

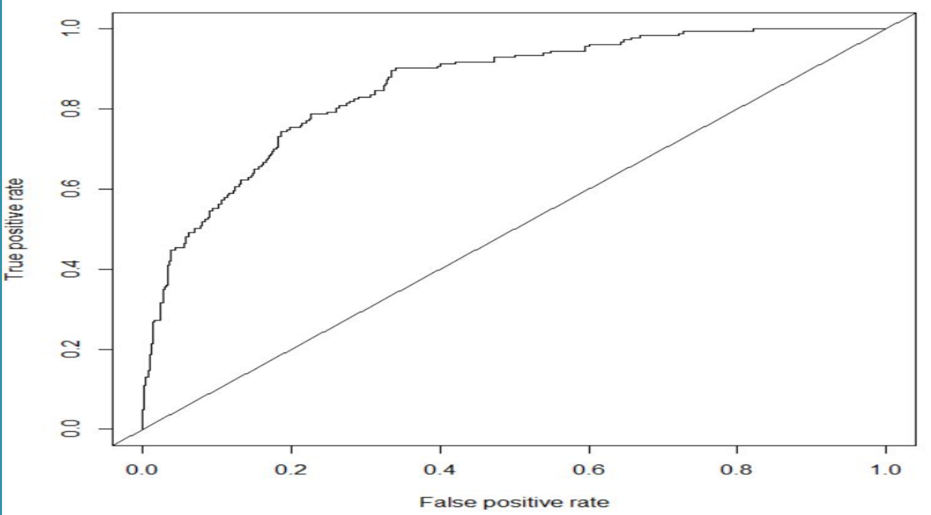
```
# Importing bank loan data & Fitting final Binary logistic model as
obtained in BLR02
data<-read.csv("BANK LOAN.csv",header=TRUE)
data$AGE<-factor(data$AGE)

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

```
# 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
- **abline()** adds a straight line to the plot.

ROC in R

```
# Output:
```




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

❑ Gives area under curve (AUC)

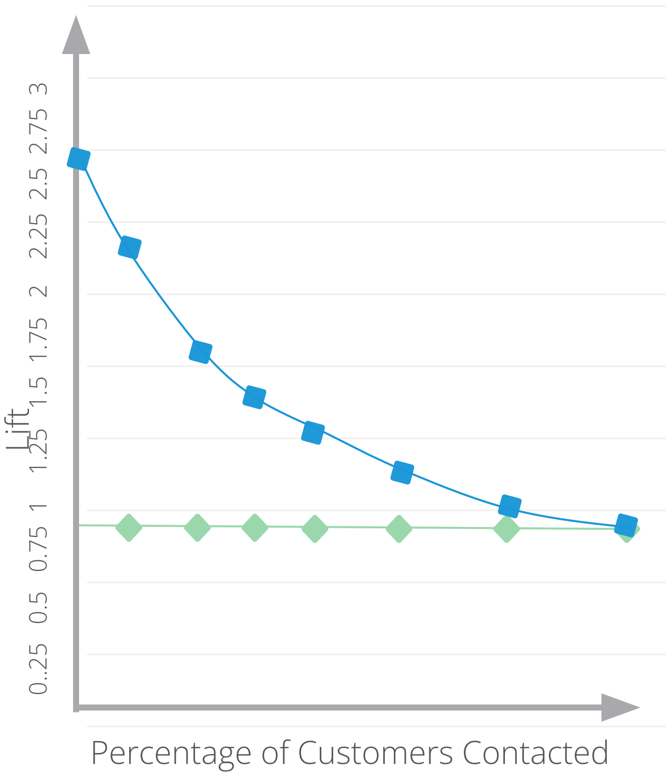
Interpretation :

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

Lift Curve

- The idea is to quantify and compare two scenarios- one uses the model to identify certain cases and second using random selection of cases for a specific purpose such as a marketing campaign.
- Lift is the ratio of results obtained **with and without a model**.
- Although primarily used in marketing analytics, the concept finds applicability in other domains as well, such as risk modeling, supply chain analytics, etc.

Lift Curve



Lift Curve: After contacting X% of customers, Y% of respondents will be identified if a statistical model is used.

Ratio Y/X is plotted

Baseline: After contacting X% of customers, X% of respondents will be identified if random method is used.

Ratio X/X is plotted

Lift Chart in R

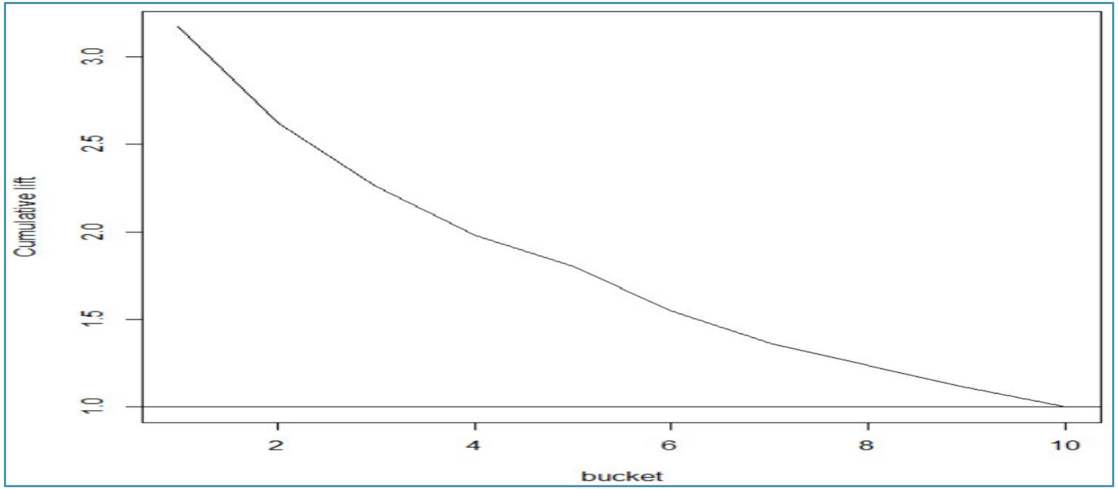
Install and load package “lift”

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

data$predprob<-round(fitted(riskmodel),2)

plotLift(data$predprob,data$DEFAULTER, cumulative = TRUE,
n.buckets = 10)

abline(1,0)
```

- **fitted()** generates predicted probabilities.
- **plotLift()** plots a Lift curve by ordering the data by predicted probabilities and computing proportion of positives for each bucket.
- **cumulative=T** logical for specifying whether cumulative lift curve should be plotted
- **n.buckets=** how many buckets should be used
- **abline()** adds a straight line to the plot.

Lift Chart in R

```
# Output:
```



Interpretation :

- Model is performing better. As more defaulters identified in earlier buckets.

Kolmogorov-Smirnov Statistic

Kolmogorov-Smirnov (KS) Statistic is one of the most commonly used measures to assess predictive power for marketing or credit risk models.

KS is the maximum difference between % cumulative Goods and Bads distribution across probability bands.
The gains table typically has % cumulative Goods (or Non-Event) and % Cumulative Bads (Or Event) across 10 or 20 probability bands

- **KS is a point estimate**, which means it is only one value and indicates the probability band where separation between Goods (or Non-Event) and Bads (or Event) is maximum.
- Theoretically **K-S can range from 0-100**. **KS less than 25, may not indicate a good model**. Too high value should also be evaluated carefully.

Kolmogorov-Smirnov Statistic

BAND	Count	Percent	Count(bad)	%(bad)	Count(good)	%(good)	cum% bad	cum% good	KS
0.95-1	10	1.4%	9	4.9%	1	0.2%	4.9%	0.2%	4.7%
0.90-0.95	7	1.0%	7	3.8%	0	0.0%	8.7%	0.2%	8.5%
0.85-0.90	7	1.0%	6	3.3%	1	0.2%	12.0%	0.4%	11.6%
0.80-0.85	7	1.0%	5	2.7%	2	0.4%	14.8%	0.8%	14.0%
0.75-0.80	11	1.6%	9	4.9%	2	0.4%	19.7%	1.2%	18.5%
0.70-0.75	17	2.4%	14	7.7%	3	0.6%	27.3%	1.7%	25.6%
0.65-0.70	17	2.4%	12	6.6%	5	1.0%	33.9%	2.7%	31.2%
0.60-0.65	10	1.4%	7	3.8%	3	0.6%	37.7%	3.3%	34.4%
0.55-0.6	24	3.4%	14	7.7%	10	1.9%	45.4%	5.2%	40.1%
0.5-0.55	21	3.0%	9	4.9%	12	2.3%	50.3%	7.5%	42.7%
0.45-0.5	22	3.1%	9	4.9%	13	2.5%	55.2%	10.1%	45.1%
0.40-0.45	31	4.4%	13	7.1%	18	3.5%	62.3%	13.5%	48.8%
0.35-0.4	29	4.1%	11	6.0%	18	3.5%	68.3%	17.0%	51.3%
0.3-0.35	27	3.9%	13	7.1%	14	2.7%	75.4%	19.7%	55.7%
0.25-0.3	40	5.7%	7	3.8%	33	6.4%	79.2%	26.1%	53.1%
0.2-0.25	45	6.4%	12	6.6%	33	6.4%	85.8%	32.5%	53.3%
0.15-0.2	52	7.4%	10	5.5%	42	8.1%	91.3%	40.6%	50.6%
0.10-0.15	66	9.4%	4	2.2%	62	12.0%	93.4%	52.6%	40.8%
0.05-0.1	80	11.4%	8	4.4%	72	13.9%	97.8%	66.5%	31.3%
0-0.05	177	25.3%	4	2.2%	173	33.5%	100.0%	100.0%	0.0%
Total	**700**	**100%**	183	**100%**	517	**100%**			

Pearson Residuals

- The Pearson residual is defined as the standardized difference between observed and predicted frequency. It measures relative deviations between observed and fitted values. :

$$\boxed{r_j = \frac{(\mathbf{Y_j} - M_j\mathbf{p_j})}{\sqrt{M_j\mathbf{p_j}(1-\mathbf{p_j})}}}$$

where
Mj : number of observations with jth covariate pattern
Y_j : Observed value (1 or 0) for jth covariate pattern
p_j : Predicted probability for j^{th} covariate pattern

- Binary Logistic Regression does not require 'Normality' of residuals

Pearson Residuals in R

```
# Getting Pearson Residuals:
```

```r
data$resi<-residuals(riskmodel,"pearson")
head(data)
```

```
# Output:
```

	SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTER	predprob	resi
1	1	3	17	12	9.3	11.36	5.01	1	0.80834673	0.4869219
2	2	1	10	6	17.3	1.36	4.00	0	0.19811470	-0.4970525
3	3	2	15	14	5.5	0.86	2.17	0	0.01006281	-0.1008221
4	4	3	15	14	2.9	2.66	0.82	0	0.02215972	-0.1505387
5	5	1	2	0	17.3	1.79	3.06	1	0.78180810	0.5282862
6	6	3	5	5	10.2	0.39	2.16	0	0.21646839	-0.5256165

 Residuals

- Pearson residuals are calculated by simply adding the argument “**pearson**” in the **residuals()** function.

Influence plot

- Influence plots are used to identify extreme values and their influence on a model.
- If removal of an observation causes substantial change in estimates of coefficients or predicted probabilities, then the observation is called an influential observation.
- Influential observations are analysed separately.

Influence plot in R

```
# Install and load "car" package.
```

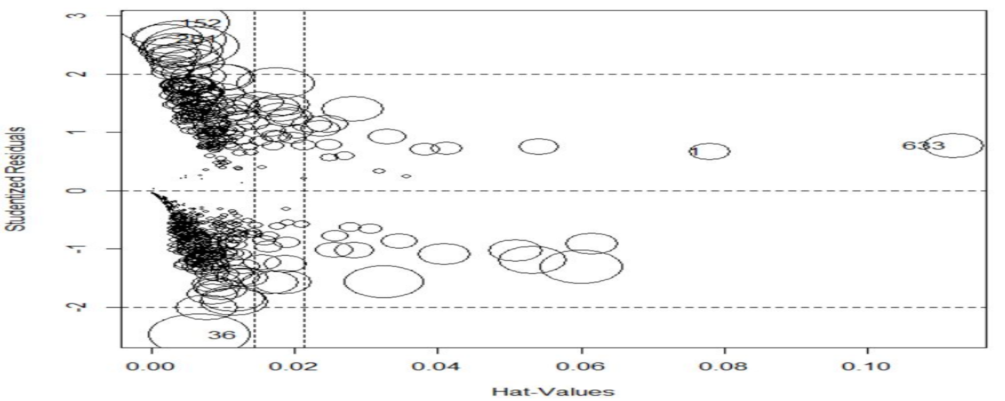
```
install.packages("car")
library(car)

influencePlot(riskmodel)
```

- **influencePlot()** creates a bubble plot of Studentised residuals by hat values, with the areas of the circles representing the observations proportional to Cook's distances.

Influence plot in R

```
# Output:
```

	StudRes	Hat	CookD
1	0.6675108	0.077944303	0.004347290
36	-2.4744534	0.006728529	0.025951104
152	2.8779760	0.002847547	0.032633681
281	2.6041504	0.002354813	0.013123240
633	0.7685420	0.112165052	0.009019769



- Large value of CookD indicates an influential observation
- Plot is for studentized residuals against hat-values, and the size of circle is proportional to Cook’s distance

Multicollinearity

- Multicollinearity exists if there is a strong linear relationship among the continuous independent variables.
- Do not ignore multicollinearity in Binary Logistic Regression .
- Use variance inflation factors to detect multicollinearity.

* Multicolinearity is explained in detail in MLR module.

Quick Recap

ROC Curve	• Graphical representation of the trade off between the false positive (FPR) and true positive (TPR) rates for various cut off values.
Lift Curve	• Compare model results with baseline without model
K-S statisitc	• KS is the maximum difference between % cumulative Goods (event/Y=1) and cumulative Bads (non events/Y=0) distribution across probability groups.
Residual	• Pearson's residual is used for binary logistic regression
Influence Plot	• Influence plots are used to identify the extreme values and their contribution to the model