**Logistic Regression**

**Example- Bank dataset**

**Target Variable “y” is in categorical format.**

**Summary 🡺**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| age | balance | day | duration | campaign | pdays | previous |
| Min. :18.00 | Min. : -8019 | Min. : 1.00 | Min. : 0.0 | Min. : 1.000 | Min. : -1.0 | Min. : 0.0000 |
| 1st Qu.:33.00 | 1st Qu.: 72 | 1st Qu.: 8.00 | 1st Qu.: 103.0 | 1st Qu.: 1.000 | 1st Qu.: -1.0 | 1st Qu.: 0.0000 |
| Median :39.00 | Median : 448 | Median :16.00 | Median : 180.0 | Median : 2.000 | Median : -1.0 | Median : 0.0000 |
| Mean :40.94 | Mean : 1362 | Mean :15.81 | Mean : 258.2 | Mean : 2.764 | Mean : 40.2 | Mean : 0.5803 |
| 3rd Qu.:48.00 | 3rd Qu.: 1428 | 3rd Qu.:21.00 | 3rd Qu.: 319.0 | 3rd Qu.: 3.000 | 3rd Qu.: -1.0 | 3rd Qu.: 0.0000 |
| Max. :95.00 | Max. :102127 | Max. :31.00 | Max. :4918.0 | Max. :63.000 | Max. :871.0 | Max. :275.0000 |

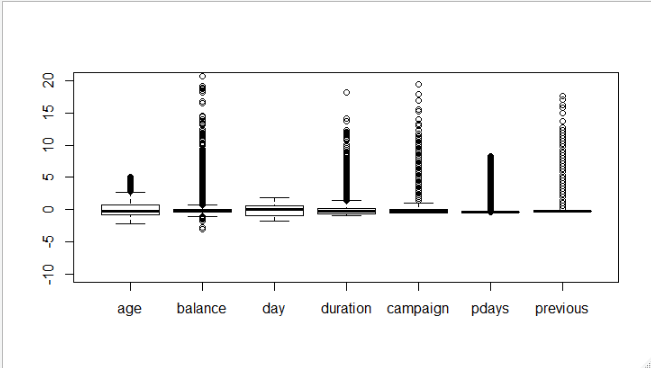
**There is significant difference between mean and median of some variables in the dataset.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| marital | education | default | housing | loan | contact | poutcome | y |
| divorced: 5207 | primary : 6851 | no :44396 | no :20081 | no :37967 | cellular :29285 | failure: 4901 | no :39922 |
| married :27214 | secondary:23202 | yes: 815 | yes:25130 | yes: 7244 | telephone: 2906 | other : 1840 | yes: 5289 |
| single :12790 | tertiary :13301 |  |  |  | unknown :13020 | success: 1511 |  |
|  | unknown : 1857 |  |  |  |  | unknown:36959 |  |

**From the above information default and y categories are not balanced.**

|  |  |
| --- | --- |
| job | month |
| blue-collar:9732 | may :13766 |
| management :9458 | jul : 6895 |
| technician :7597 | aug : 6247 |
| admin. :5171 | jun : 5341 |
| services :4154 | nov : 3970 |
| retired :2264 | apr : 2932 |
| (Other) :6835 | (Other): 6060 |

**Box Plot 🡺**



**Splitting of data into train and test**

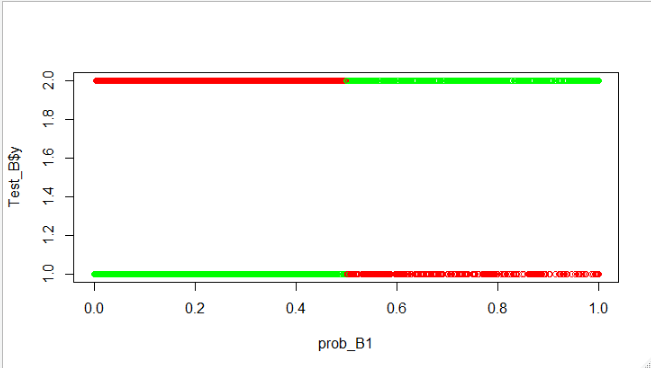
**Train = 31648 & Test = 13563**

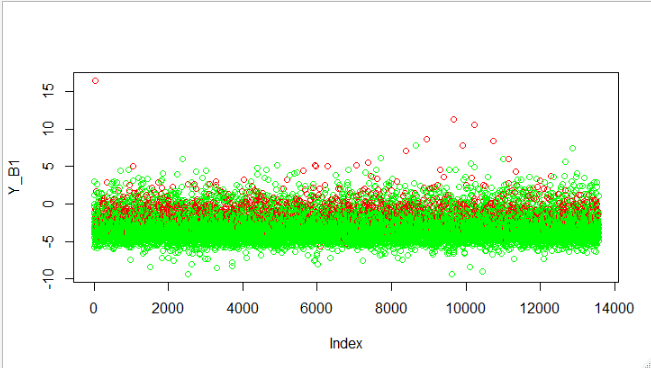
**Model-1 Building 🡺**

glm(formula = y ~ ., family = binomial(link = "logit"), data = Train\_B)

**AIC: 15017**

**Plot of wrong prediction( Red ) v/s actual prediction ( Green )**





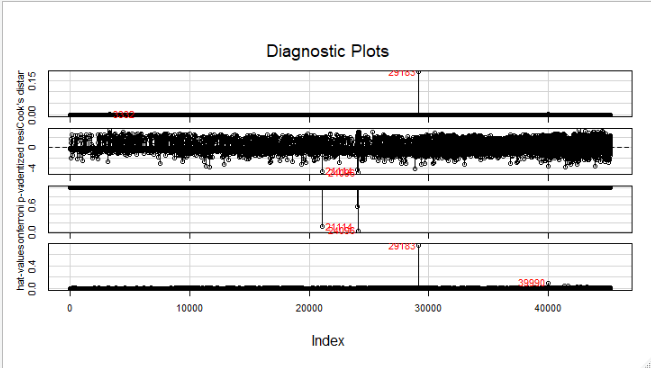
**Confusion Matrix 🡺**

no yes

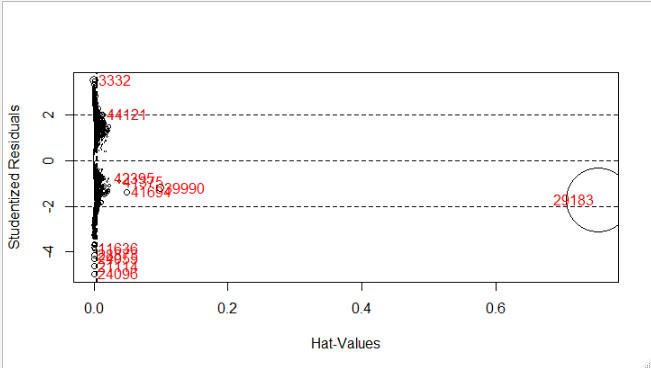
FALSE 11660 1065

TRUE 287 551

**Efficiency 🡺 0.900317**



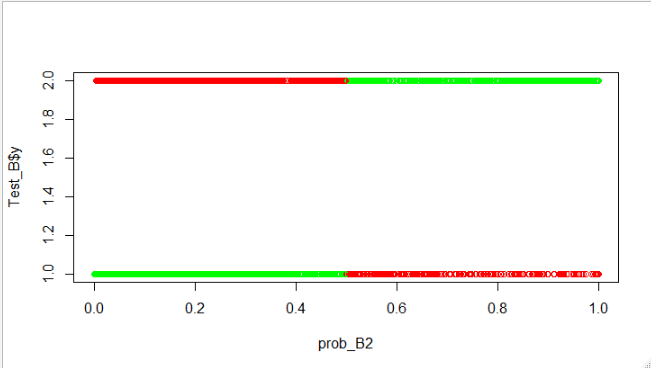
**Influence Plot**

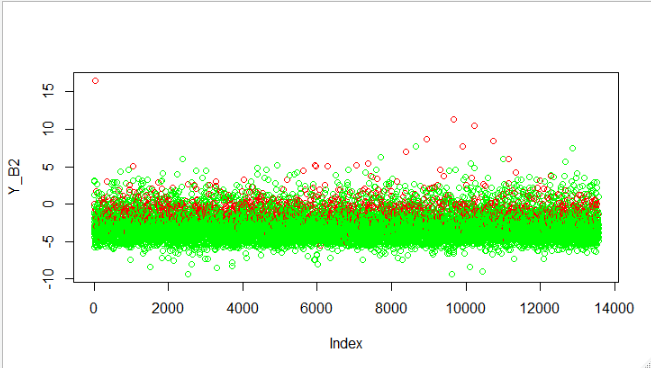


**Model-2 Building 🡺**

model\_B2 <- glm(y~.,data = Train\_B[-influence\_B1,-c(1,14,5)],family = "binomial")

**AIC: 15010**





**Confusion Matrix 🡺**

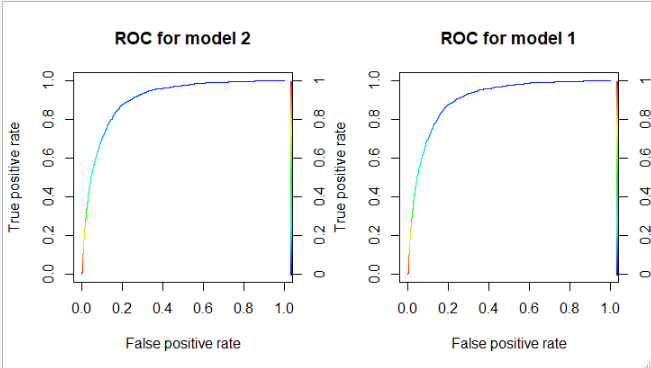
no yes

FALSE 11659 1067

TRUE 288 549

**Efficiency 🡺 0.9000958**

**Comparison between Model-1 and Model-2 🡺**



|  |  |  |  |
| --- | --- | --- | --- |
| **Model No** | **AIC** | **Efficiency** | **F1 Scores** |
| **Model-1** | **15017** | **0.900317** | **0.945201** |
| **Model-2** | **15010** | **0.9000958** | **0.9452817** |

**From the above information we can infer that there is no significant difference between Model-1 and Model-2. But we have considered many insignificant variables in Model-1 and only significant variables in Model-2.**

**So our Model-2 is final model as best model.**