**Logistic Regression**

**Example-Affairs Dataset**

**Target Variable Extra Marital Affair (EMA) is categorial variable with values “yes” and “no”**

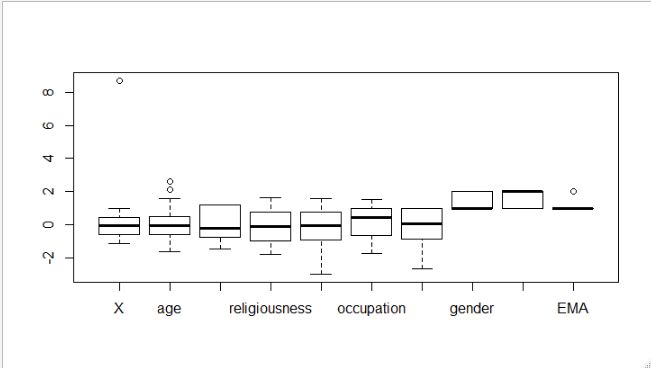
**Summary 🡺**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | age | yearsmarried | religiousness | education | occupation | rating |
| Min. : 4 | Min. :17.50 | Min. : 0.125 | Min. :1.000 | Min. : 9.00 | Min. :1.000 | Min. :1.000 |
| 1st Qu.: 528 | 1st Qu.:27.00 | 1st Qu.: 4.000 | 1st Qu.:2.000 | 1st Qu.:14.00 | 1st Qu.:3.000 | 1st Qu.:3.000 |
| Median:1009 | Median:32.00 | Median :7.000 | Median:3.000 | Median:16.00 | Median:5.000 | Median:4.000 |
| Mean :1060 | Mean :32.49 | Mean : 8.178 | Mean :3.116 | Mean :16.17 | Mean :4.195 | Mean :3.932 |
| 3rd Qu.:1453 | 3rd Qu.:37.00 | 3rdQu.:15.000 | 3rd Qu.:4.000 | 3rd Qu.:18.00 | 3rd Qu.:6.000 | 3rd Qu.:5.000 |
| Max. :9029 | Max. :57.00 | Max. :15.000 | Max. :5.000 | Max. :20.00 | Max. :7.000 | Max. :5.000 |

|  |  |  |
| --- | --- | --- |
| EMA | gender | children |
| no :451 | female:315 | no :171 |
| yes:150 | male :286 | yes:430 |

**From the above summary, there is negligible difference between mean and median, so possibly there are less numbers of outliers.**

**Box Plot 🡺**



**From above box plot, age variable contain outlier.**

**Splitting data into train and test 🡺**

**Train = 419 and Test = 179**

**Model-1 Building 🡺**

model\_A1 <- glm(EMA~.,data=Train\_A,family = 'binomial')

**AIC: 429.34**

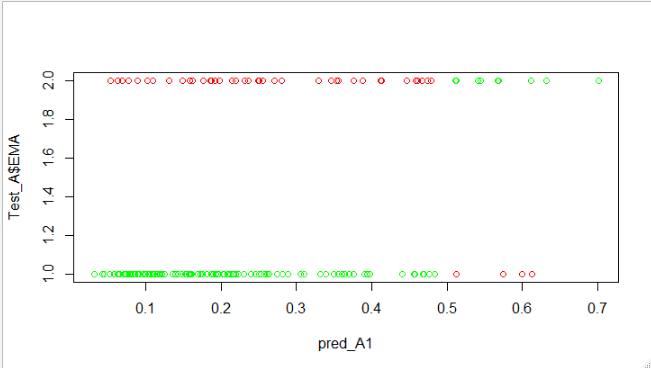
**Confusion Matrix 🡺**

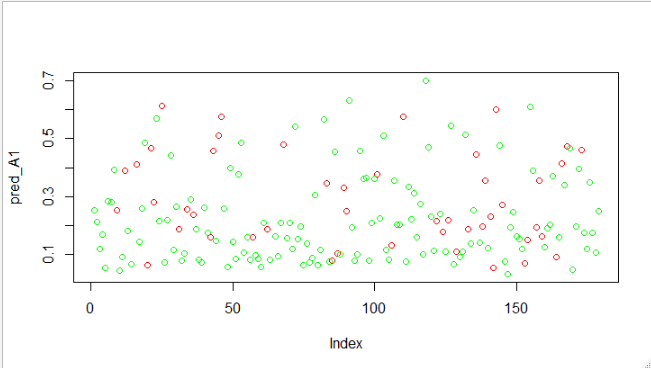
FALSE TRUE

no 122 5

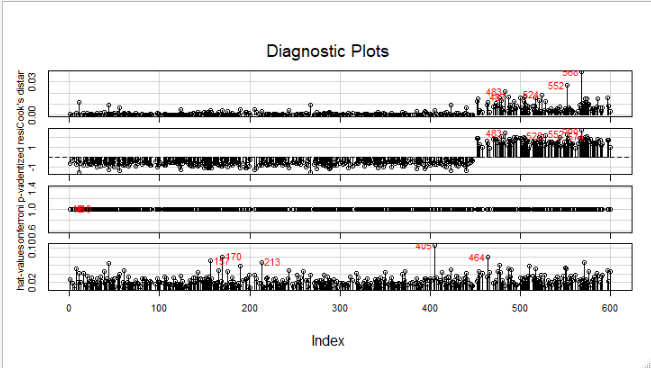
yes 41 9

**Efficiency 🡺 0.740113**

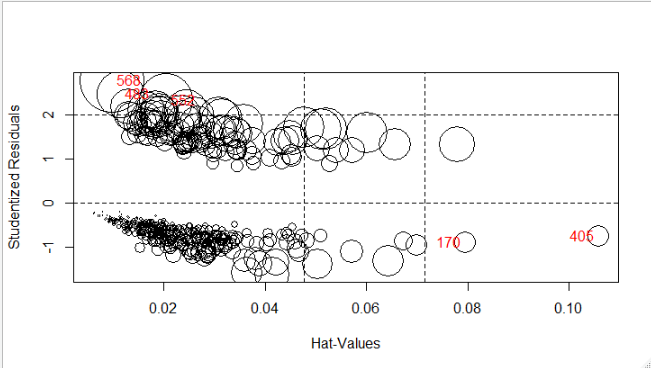




In the above plot, Red is wrong prediction and Green is actual prediction.



**Influence Plot**



**From the above plot it is seen that influencing value is in the model. So will remove this in our next model.**

**Model-2 Building 🡺**

model\_A2 <- glm(EMA~.,data=Train\_A[-in\_1,-c(8,7,1,5)],family = 'binomial')

**AIC: 421.8**

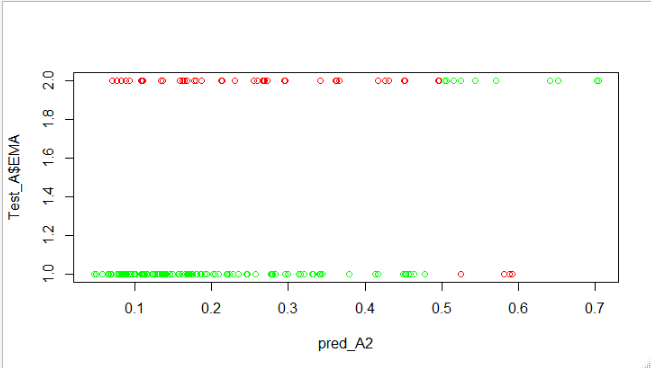
**Confusion Matrix 🡺**

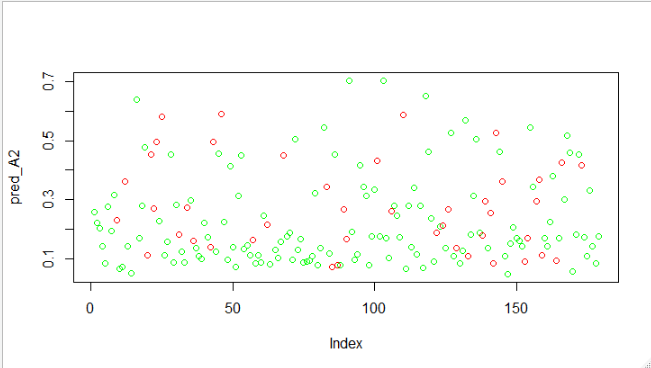
FALSE TRUE

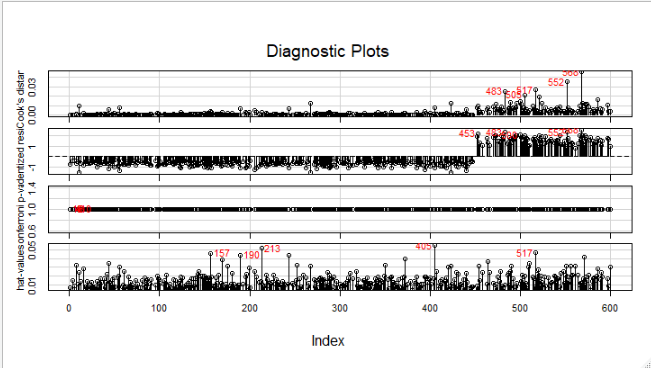
no 123 4

yes 39 11

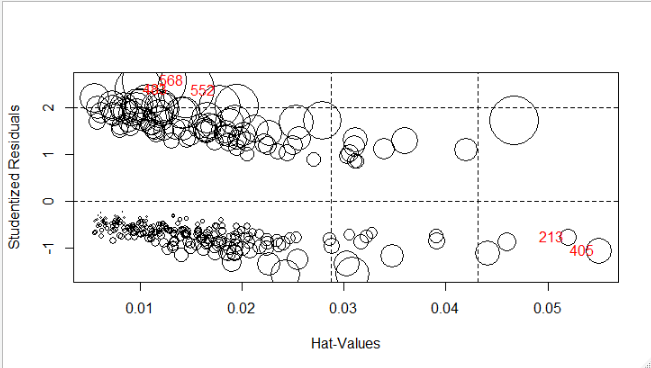
**Efficiency 🡺0.7570621**



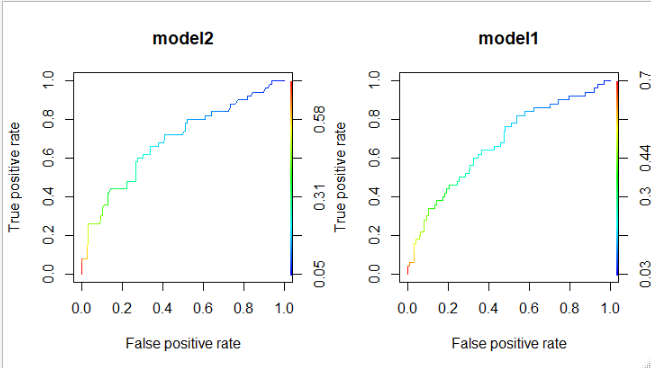




**Influence Plot 🡺**



**Comparision in Model-1 and Model-2 🡺**



**From the above curve we can infer that area under curve in increased in Model-2.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model No** | **AIC** | **Efficiency** | **F1 Scores** |
| **Model-1** | **429.34** | **0.740113** | **0.8413793** |
| **Model-2** | **421.8** | **0.7570621** | **0.8512111** |

**From above information we can conclude that Model-2 is final best model.**